Causal inference with invalid instruments

Analysis of three different approaches for linear and non-linear models

by

Daniël Cohen

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on Wednesday August 14, 2024 at 14:00

Student number: 5128110
Project duration: December 1, 2023 – August 14, 2024
Thesis committee: Prof. dr. A. W. van der Vaart, TU Delft, supervisor
Dr. rer. nat. Ö. Şahin TU Delft

TU Delft

Abstract

Suppose that we want to infer the effect of a treatment on a certain outcome, where both the treatment and outcome are influenced by other variables. It has been well-established that in the linear setting, in case we know beforehand which of these other variables are instrumental (for the effect of the treatment on the outcome), we can infer the treatment effect in a consistent sense. This thesis analyses 3 methods that deals with the issue of unknown instrumental variables (IVs) and functional relationships in different ways to infer the treatment effect. The first method, Causal Inference with Invalid Instruments (CIII), assumes that we have a linear setting and a set with potential instrumental variables for whom a majority or plurality rule holds to obtain a robust confidence interval for the treatment effect. The second method, Anchor Regression (AR), only assumes a linear setting. By mediating between different methods, the AR-estimator turns out the be robust to changes in the distribution of the sampled data. Lastly, Two Stage Curvature Identification (TSCI), does not require a linear setting or information on the IVs. Instead, it relies on the difference in functional form between the effect of the variables on the treatment and the effect of the variables on the outcome for consistent estimation and asymptotic normality. TSCI also provides a test for IV presence in the non-linear setting. In this thesis, I will explain the workings of these 3 methods, analyse their theoretical foundation and do simulation studies. Based on these analyses, I make several additions and suggestions to expand the theoretical scope and improve practical efficacy.

Acknowledgements

I would, first of all, like to thank Aad van der Vaart and Özge Şahin for being on my exam committee. Aad has played a fundamental role in shaping my future mathematical career by recommending me courses to follow at the start of my Masters degree. These courses took me from Utrecht (Nonparametric Statistics) to Amsterdam (Asymptotic Statistics) and were crucial for my thesis. My thesis would have had much less depth without these courses and I am grateful to have had the opportunity to follow them. I would also like to thank Aad for his guidance during my thesis: he helped develop a "researcher's" mindset and this, I believe, will prove to be crucial as it is my intention to continue with a PhD.

I am also grateful to have had the support of friends and family. Some friends that come to mind now are Janic Bijlhout (who has been a true friend to me since high school), Joran Haasnoot (with whom I have enjoyed many in-depth conversations about human nature), Gideon Vissers (I always enjoy our 5 hour podcast-style conversations about a wide range of topics) and many others. I want to thank the crew of Delft Improv Group (DIG) for being great and thoughtful friends and for encouraging me to do improv shows and various other activities (like camping in the rain for 1 week with 7 others ;)). I would also like to thank Hana Hasanbegovic for volunteering to read parts of my thesis and give me pointers on the format. I have not noted many people that have impacted me here over the last 5 years. If you're still reading this, you are likely part of this group and I want to give you a profound thank you.

Last, but certainly not least, I want to thank God for giving me the right insights at the right time, connecting me with the people I need to meet and guiding me on my journey of life.

Again, thank you to all. It is my belief, that our best days are still ahead of us.

Daniël Cohen, July 2024, Delft.

Contents

Frequently used notation

$\|X\|_p$ denotes L^p-norm for random variable X.

$A \in \mathbb{R}^{m\times m}$: $\lambda_k(A)$ refers to the k-th ordered eigenvalue of A.

$A \in \mathbb{R}^{m\times m} : \|A\|_2 = \max_{x\in\mathbb{R}^m, \|x\|_2=1} \|Ax\|_2 = \sqrt{\lambda_{\max}(A^TA)}$

$A \in \mathbb{R}^{m\times m} : \|A\|_F = \sqrt{\mathrm{Tr}(A^TA)} = \sqrt{\sum_{i=1}^m \lambda_i(A^TA)} = \sum_{i,j=1}^m |A_{ij}|^2$

$A \in \mathbb{R}^{m\times m} : \|A\|_1 = \max_{1\le j\le m} \sum_{i=1}^m |A_{ij}|$

$A \in \mathbb{R}^{m\times m} : \|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^m |A_{ij}|$

For two real sequences $(s_n)_{n\ge1}, (t_n)_{n\ge1}$: $s_n \gg t_n \iff \limsup_{n\to\infty} \frac{t_n}{s_n} = 0$

$A, B \in \mathbb{R}^m : A \preccurlyeq B \iff B - A$ is positive semi-definite

1. Introduction

When I was younger, there was a hype for a while regarding new research about dark chocolate. According to various media sources, the research had suggested that eating dark chocolate would decrease the risk for particular types of cancer [1]. The overall implication: eat more dark chocolate as it might save your life. On an initial glance, from an observational study, this seems quite reasonable to conclude: presumably a lower percentage of people that ate dark chocolate developed various forms of cancer over a certain period of time compared to the group of people that didn't eat the dark chocolate. Well, before we all start eating our "healthy" daily share of dark chocolate for so called "cancer prevention", maybe we should first consider that the overall conclusion might sound a bit too good to be true (if you like dark chocolate). Could there be nuances to this research? Maybe it's not the dark chocolate doing the magic, but other factors related to it? Below I will list some examples of such potential related factors:

1. Dark chocolate contains a ton of magnesium, compared to other foods. Maybe it is not the dark chocolate in specific, but the magnesium that decreases the risk to cancer. If that is the case, it would hence be more effective to take supplements of magnesium instead (but yes, less fun).
2. It is well-known that dark chocolate is a snack with less sugar than its other chocolate counterparts. Maybe the people that eat dark chocolate in this study also consumed less sugar in their overall diet compared to the general population. It could be the case that a reduction of sugar in one's diet reduces the cancer risk rather than the consumption of dark chocolate.
3. Maybe the risk of cancer is not in our control at all. Perhaps, the people that like the taste of dark chocolate also prefer more bitter tastes in general. A preference for bitter tastes could indicate a genetic component that reduces the risk to various types of cancer.

The overall message of the three examples above: we could credit dark chocolate for things it isn't actually doing. In the cases of example (2) and (3): eating extra dark chocolate wouldn't benefit us at all! So, when we wonder whether eating dark chocolate benefits us in preventing cancer, what we specifically want to find out it whether the chocolate benefits us when we take other potential factors of influence into account.
To answer the question about dark chocolate properly, mathematicians tend to work in 3 steps in their problem-solving:

Step 1. We first need to come up with factors that could potentially influence the risk to cancer besides dark chocolate. We have already come up with three such factors, namely: magnesium, sugar intake and genetics. We also need to take into account that there are some unmentioned factors that might also influence the cancer rate.

Step 2. Next, we need to come up with possible relations between these factors and think about in what way they may influence each other. One such example can be found in figure 1.1 (where the caption explains the notation in the figure. An arrow from node X to f.e. Y means that X influences Y). Here, we can pay attention to two things in particular: firstly, how the arrows are pointed. Beforehand, we don't know this. In figure 1.1, makes the suggestion that sugar intake does not directly influence your risk to cancer but only suggests something about the amount of dark chocolate you eat. What is interesting

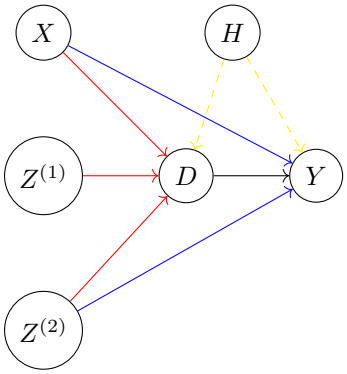

Figure 1.1.: An example causal representation of magnesium (X), sugar intake ($Z^{(1)}$), genetics ($Z^{(2)}$), dark chocolate intake (D), unmeasured (or hidden) factors (H) and risk of various cancer types (Y).

here is that magnesium and genetics influence both the chocolate intake and the cancer rate while sugar only influences the dark chocolate intake. In the context of figure 1.1, we call the sugar intake an instrumental variable (IV), while we call the magnesium, genetics and the unmeasured factors confounders. Magnesium and genetics could here also be classified as invalid instrumental variables. Once we are aware which variables are instrumental are IVs and which are confounders, there are well-established methods to infer the effect of dark chocolate intake on getting cancer (the black arrow in figure 1.1). The first method I'm going to discuss in my thesis (section 2) provides a method to distinguish between IVs and invalid IVs (for linear models). The second method (section 3 mediates the effects of IVs with the effects of confounders (again for linear models).

Step 3. Lastly, we need to ask ourselves what the arrows exactly mean as a means for a functional relation. In the previous step, it was already mentioned that we assume linear relations for the first and second method. For the third method (section 4), we don't need such linearity assumptions. We do, however, assume that we are able to distinguish between the functions that represent the direct effect of factors on the outcome (so the blue arrows in figure 1.1) from the functions that represent the influence on the treatment (so the red arrows in figure 1.1).

All the mentioned methods above will be specified in the next sections. There I will discuss the methods, provide my analysis of the theoretical justification behind these methods (including my additions) and I will also perform simulation studies and discuss what they revealed about said methods.

2. Causal inference with invalid instruments (CIII) using Searching & Sampling

2.1. General idea CIII

For the proposed CIII-method in this section (which originated from [2]), we work with linear models only. Coming back to figure 1.1, this could then be translated to mathematical terms as follows:

$$Y_i = \beta^* D_i + \pi_2^* Z_i^{(2)} + \phi^* X_i + e_i$$
$$D_i = \gamma_1^* Z_i^{(1)} + \gamma_2^* Z_i^{(2)} + \psi^* X_i + \delta_i$$

where $\pi_2^*, \phi^*, \gamma_1^*, \gamma_2^*, \psi^*$ are non-zero and $(Y_i, D_i, X_i, Z_i)_{i=1}^n$ represent n data points. In this model, we can also assume that (X_i, Z_i) are independent of confounding effects i.e. $(X_i, Z_i) \perp\!\!\!\perp (e_i, \delta_i)$, together with $\mathbb{E}(e_i) = \mathbb{E}(\delta_i) = 0$ and that D_i does not influence (e_i,δ_i) or (Z_i, X_i). In that case, β^* is (in a formal sense when we also assume consistency in the context of [11] i.e. $Y = Y^D$) the causal effect per unit intervention on D with respect to outcome Y. Due to the influence of confounders on D, we can't apply ordinary least squares (OLS) to the Y_i-equation and obtain a consistent estimator for β^*. We will be in a position to apply OLS once we substitute the D_i-equation into the Y_i-equation above:

$$Y_i = \Gamma_1^* Z^{(1)} + \Gamma_2^* Z^{(2)} + \Psi^* X + \epsilon_i$$
$$D_i = \gamma_1^* Z_i^{(1)} + \gamma_2^* Z_i^{(2)} + \psi^* X_i + \delta_i$$

Here: $\Gamma_1^* = \beta^*\gamma_1^*, \Gamma_2^* = \beta^*\gamma_2 + \pi_2^*, \Psi^* = \beta^*\psi^* + \phi^*$ and $\epsilon_i = \beta^*\delta + e_i$. Now observe that we can apply OLS to both the Y_i and D_i equation above so that we are in a position to obtain a consistent estimator for γ_1^* (let's call this estimator $\hat{\gamma}_1$) and for Γ_1^* (let's call that estimator $\hat{\Gamma}_1$). As $\Gamma_1^* = \beta^*\gamma_1^*$: $\hat{\Gamma}_1/\hat{\gamma}_1$ is a consistent estimator for β^*. As $\Gamma_2^*/\gamma_2^* = \beta^* + \pi_2^*/\gamma_2^*$ and π_2^* can't be estimated from the data (i.e. the data can't distinguish between the indirect effect of $Z^{(2)}$ through D and the direct effect of $Z^{(2)}$ on Y): the same trick won't work there. We call this special variable $Z^{(1)}$ an instrumental variable (as it influences Y only through D and not directly).

Once we know which variables are instrumental in a model we can obtain consistent estimators for β^*. The general problem is that we don't know (for certain) beforehand which variables are instrumental and which aren't. We might have an idea of which variables are potentially instrumental. The methods proposed in this section, provides a procedure for selecting IVs under the assumption of a majority and/or plurality rule on a subset of the potential instrumental variables. The set of potential IVs is selected beforehand by the user (using for example domain knowledge).

Instead of selecting IVs (which involves also checking this majority/plurality rule) and then computing the likely consistent estimators for β^* (which has been done in previous literature

[2, p.1-3 1.Introduction]), the method provides a (more robust) searching procedure where for $\beta \in \mathbb{R}$, under hypothesis that $\beta = \beta^*$ and i being instrumental $\hat{\Gamma}_i - \beta\hat{\gamma}_i \approx 0$ is checked (through asymptotic normality of the OLS-estimators). For the specific example derived from figure 1.1, it is presumed that for enough data points $\hat{\Gamma}_1 - \beta^*\hat{\gamma}_1 \approx 0$ is not rejected while $\hat{\Gamma}_2 - \beta^*\hat{\gamma}_2 \approx 0$ will be rejected.

2.2. Models, assumptions, goals

Consider iid observations with errors $(Y_i, D_i, X_i, Z_i, e_i, \delta_i)_{i=1}^n \sim (Y, D, X, Z, e, \delta)$, where:

- $Y_i \in \mathbb{R}$ outcome
- $D_i \in \mathbb{R}$ treatment
- $X_i = (X_i^{(1)}, ..., X_i^{(p_X)})^T \in \mathbb{R}^{p_X}$ baseline covariates.
- $Z_i = (Z_i^{(1)}, ..., Z_i^{(p_Z)})^T \in \mathbb{R}^{p_Z}$ candidate instrumental variables (with respect to the direct effect of the treatment on the outcome)

Below we define the model corresponding to the observations and what we require of an instrumental variable.

Definition 2.2.1. *The **outcome model** is defined as:*

$$Y_i = D_i\beta^* + Z_i^T\pi^* + X_i^T\phi^* + e_i$$
$$\mathbb{E}(e_i Z_i) = 0, \mathbb{E}(e_i X_i) = 0$$

*β^* is called the **treatment effect**.*

Definition 2.2.2. *The **association model** is defined as:*

$$D_i = Z_i^T\gamma^* + X_i^T\psi^* + \delta_i$$
$$\mathbb{E}(\delta_i Z_i) = 0, \mathbb{E}(\delta_i X_i) = 0$$

The main objective will be to infer the treatment effect β^*. Note that because we don't require that $\mathbb{E}(e_i D_i) = 0$, we can't obtain a (consistent) estimator of the treatment effect by directly applying the OLS-method to the outcome model.

We now define what it means for $Z^{(j)}$ to be an instrumental variable within the context of the outcome and association model. Note that beforehand, we don't know whether $Z^{(j)}$ is instrumental: it is a candidate.

Definition 2.2.3. *Within the context of the outcome and association model, we define $Z^{(j)}$ to be an instrumental variable if it satisfies the following 2 conditions:*

1. *The IV is associated with the treatment, i.e. $\gamma_j^* \neq 0$.*
2. *The IVs have no direct effect on the outcome, i.e. $\pi_j^* = 0$.*

Remark 2.2.4. *It could be that due to unmeasured/hidden confounders: e_i and δ_i are correlated. In that case we could have that $\mathbb{E}(e_i D_i) \neq 0$. In the context of causal inference, this raises a question about potential relations between (X, Z) and (e, δ). It is, in theory, not excluded for X and Z to influence the hidden confounders in this setting. For instance, in case $(X, Z, \tilde{e}) \sim \mathcal{N}_3(0, \mathrm{Id})$ with $e = Z^2 - 1 + \tilde{e}$, then it still holds that $\mathbb{E}(Ze) = \mathbb{E}(Xe) = 0$. In the same way, it is in theory not excluded for D to influence e and/or δ which would result in β^* not being the causal effect of the treatment of the outcome anymore (in the context of [11]).*

Based on what we require an instrumental variable to satisfy, we can categorise the candidate instrumental variables as follows:

Definition 2.2.5.

- *Set of relevant instruments:* $S = \{1 \leq j \leq p_Z : \gamma_j^* \neq 0\}$
- *Set of valid instruments:* $V = \{j \in S : \pi_j^* = 0\}$

Remark 2.2.6.

- *In case γ_j^*=0, we would see that any change in a potential instrumental variable would not impact the treatment. Hence we would not be able to apply instrumental methods to determine the treatment effect.*
- *Note that any valid instrument is also relevant and that every valid instrument is an instrumental variable as defined in definition 2.2.3.*

Lemma 2.2.7 (Reduced form equations)**.** *The outcome model and association model satisfy the following expressions:*

$$\begin{aligned} Y_i &= Z_i^T \Gamma^* + X_i^T \Psi^* + \epsilon_i, \quad \mathbb{E}(\epsilon_i Z_i) = 0, \mathbb{E}(\epsilon_i X_i) = 0 \\ D_i &= Z_i^T \gamma^* + X_i^T \psi^* + \delta_i, \quad \mathbb{E}(\delta_i Z_i) = 0, \mathbb{E}(\delta_i X_i) = 0 \end{aligned}$$

where:

$$\begin{aligned} \Gamma^* &= \beta^* \gamma^* + \pi^* \\ \Psi^* &= \beta^* \psi^* + \phi^* \\ \epsilon_i &= \beta^* \delta_i + e_i \end{aligned}$$

Proof. For the equations:

$$\begin{aligned} Y_i &= D_i \beta^* + Z_i^T \pi^* + X_i^T \phi^* + e_i \\ &= (Z_i^T \gamma^* + X_i^T \psi^* + \delta_i)\beta^* + Z_i^T \pi^* + X_i^T \phi^* + e_i \\ &= Z_i^T (\gamma^* \beta^* + \pi^*) + X_i^T (\psi^* + \phi^*) + \delta_i \beta^* + e_i \\ &\stackrel{\text{def}}{=} Z_i^T \Gamma^* + X_i^T \Psi^* + \epsilon_i \end{aligned}$$

For the error-term:

$$\mathbb{E}(\epsilon_i X_i) = \mathbb{E}(\beta^* \delta_i X_i + e_i X_i) = 0$$

□

It can be observed that OLS-estimation can be applied to the reduced form equations for consistent estimators for the coefficients.

2.3. β^* identification

Observe that for the j-th (candidate) IV, we have that: $\Gamma_j^* = \beta^* \gamma_j^* + \pi_j^*$. In case it this candidate IV valid, we have that $\beta^* = \frac{\Gamma_j^*}{\gamma_j^*}$. In case it is only relevant: $\frac{\Gamma_j^*}{\gamma_j^*} \neq \beta^* = \frac{\Gamma_j^*}{\gamma_j^*} - \frac{\pi_j^*}{\gamma_j^*}$.
The following condition, the population majority rule, requires that β^* can be identified using the majority of relevant instruments:

Condition 2.3.1. *The Population Majority Rule: More than half of the relevant IVs are valid i.e.* $|V| > \frac{|S|}{2}$.

A second, less restrictive, approach to identify β^* is through a plurality rule. For invalid (but relevant) IVs we have that $\beta^* = \frac{\Gamma_j^*}{\gamma_j^*} - \frac{\pi_j^*}{\gamma_j^*}$, $\frac{\pi_j^*}{\gamma_j^*} \neq 0$. We call the term $\frac{\pi_j^*}{\gamma_j^*}$ the invalidity level. In case an IV is valid, the invalidity level of that IV is 0. By assuming that the number of valid IVs is larger than the number of relevant but invalid IVs at any invalidity level $v \neq 0$, we can identify β^*.

Condition 2.3.2. *The Population Plurality Rule: the number of valid IVs is larger than the number of invalid IVs with any invalidity level* $v \neq 0$*, that is:*

$$|V| > \max_{v \neq 0} |I_v|, \; I_v = \{j \in S : \frac{\pi_j^*}{\gamma_j^*} = v\}$$

Remark 2.3.3. *As* $V = I_0$*, the condition above could be rewritten to requiring that* $|I_0| > \max_{v \neq 0} |I_v|$

The next lemma shows that the majority rule is indeed stronger than the plurality rule:

Lemma 2.3.4. *If the population majority rule holds, so does the population plurality rule.*

Proof. $|V| \overset{(1)}{>} |S/V| \overset{(2)}{\geq} I_v, \quad \forall v \neq 0$. Hence: $|V| > \max_{v \neq 0} |I_v|$.
(1): The majority of S is valid.
(2): $I_v \subseteq S/V$ for $v \neq 0$. □

2.4. Data-dependent estimators for γ^* and Γ^*

In both the identification of β^* through the majority rule and plurality rule, one needs expressions for Γ^* and γ^*. By the application of OLS, we can obtain consistent estimators of Γ^* and γ^* which has a rate of convergence of order $\sqrt{n}$. Using these estimators, we can also obtain asymptotically normal and consistent estimators for π^* if we are willing to assume validity of certain potential IVs. These notions are specified below.

2.4.1. Asymptotic normality OLS-estimators Γ^*, γ^*

First I will introduce some notation.

Notation 2.4.1. *For $1 \leq i \leq n$:*

- $W_i := \begin{pmatrix} Z_i \\ X_i \end{pmatrix}$, $W := \begin{pmatrix} W_1^T \\ \dots \\ W_n^T \end{pmatrix}$
- $\Sigma := \mathbb{E}(W_i W_i^T)$, $\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} W_i W_i^T$
- $Y := (Y_1, ..., Y_n)^T$, $D := (D_1, ..., D_n)^T$
- $\epsilon := (\epsilon_1, ..., \epsilon_n)^T, \delta := (\delta_1, ..., \delta_n)^T$

Definition 2.4.2 (OLS estimators from reduced form equations)**.** *Assume $W^T W$ is invertible. Then the OLS-estimator from outcome equation as seen in the reduced form equations is defined as:*

$$\begin{pmatrix} \hat{\Gamma} \\ \hat{\Psi} \end{pmatrix} = (W^T W)^{-1} W^T Y$$

and from the association model:

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\psi} \end{pmatrix} = (W^T W)^{-1} W^T D$$

The following theorem establishes asymptotic normality for $\sqrt{n}(\begin{pmatrix} \hat{\Gamma} \\ \hat{\gamma} \end{pmatrix} - \begin{pmatrix} \Gamma^* \\ \gamma^* \end{pmatrix})$. This property was mentioned in the original paper [2, p.9 C.Proofs] with more restrictive conditions and without proof. I came up with the proof myself.

Theorem 2.4.3. *Assume the following:*

- *Σ is finite and invertible*
- *$\mathbb{E}(\epsilon_i^2 W_i W_i^T)$, $\mathbb{E}(\epsilon_i \delta_i W_i W_i^T)$ and $\mathbb{E}(\delta_i^2 W_i W_i^T)$ are finite.*

Then:

$$\sqrt{n}(\begin{pmatrix} \hat{\Gamma} \\ \hat{\gamma} \end{pmatrix} - \begin{pmatrix} \Gamma^* \\ \gamma^* \end{pmatrix}) \xrightarrow{d} \mathcal{N}(0, \text{Cov})$$

Here,

$$\begin{aligned} \text{Cov} &= \begin{pmatrix} V^\Gamma & C \\ C & V^\gamma \end{pmatrix} \\ V^\Gamma &= [\Sigma^{-1} \mathbb{E}(\epsilon_i^2 W_i W_i^T) \Sigma^{-1}]_{1:p_Z, 1:p_Z} \\ V^\gamma &= [\Sigma^{-1} \mathbb{E}(\delta_i^2 W_i W_i^T) \Sigma^{-1}]_{1:p_Z, 1:p_Z} \\ C &= [\Sigma^{-1} \mathbb{E}(\epsilon_i \delta_i W_i W_i^T) \Sigma^{-1}]_{1:p_Z, 1:p_Z} \end{aligned}$$

Proof. The strategy here will be to first prove asymptotic normality for $(\hat{\Gamma}, \hat{\Psi}, \hat{\gamma}, \hat{\psi})^T$ and then to restrict ourselves to $(\hat{\Gamma}, \hat{\gamma})^T$ by multiplying with a linear mapping.

$$\begin{aligned}
\sqrt{n}(\begin{pmatrix}\hat{\Gamma}\\ \hat{\Psi}\\ \hat{\gamma}\\ \hat{\psi}\end{pmatrix} - \begin{pmatrix}\Gamma\\ \Psi\\ \gamma\\ \psi\end{pmatrix}) &= \sqrt{n}\begin{pmatrix}(W^TW)^{-1}W^T\epsilon\\ (W^TW)^{-1}W^T\delta\end{pmatrix}\\
&= \sqrt{n}\begin{pmatrix}(W^TW)^{-1} & 0\\ 0 & (W^TW)^{-1}\end{pmatrix}\begin{pmatrix}W^T\epsilon\\ W^T\delta\end{pmatrix}\\
&= \begin{pmatrix}\hat{\Sigma}^{-1} & 0\\ 0 & \hat{\Sigma}^{-1}\end{pmatrix}((\sqrt{n})^{-1}\sum_{i=1}^{n}\begin{pmatrix}W_i\epsilon_i\\ W_i\delta_i\end{pmatrix})\\
&\xrightarrow{\text{d}} \begin{pmatrix}\Sigma^{-1} & 0\\ 0 & \Sigma^{-1}\end{pmatrix}\mathcal{N}(0, \begin{pmatrix}\text{Cov}(W_i\epsilon_i) & \text{Cov}(W_i\epsilon_i, W_i\delta_i)\\ \text{Cov}(W_i\epsilon_i, W_i\delta_i) & \text{Cov}(W_i\delta_i\end{pmatrix})
\end{aligned}$$

In the last step I first applied the law of large numbers to see that $\hat{\Sigma} \xrightarrow{\text{d}} \Sigma$, then the continuous mapping theorem to see that $\hat{\Sigma}^{-1} \xrightarrow{\text{d}} \Sigma^{-1}$ (which we can do as Σ is invertible), then applied the continuous mapping theorem again to see that:

$$\begin{pmatrix}\hat{\Sigma}^{-1} & 0\\ 0 & \hat{\Sigma}^{-1}\end{pmatrix} \xrightarrow{\text{d}} \begin{pmatrix}\Sigma^{-1} & 0\\ 0 & \Sigma^{-1}\end{pmatrix}$$

After that, I applied the multivariate central limit theorem to

$$(\sqrt{n})^{-1}\sum_{i=1}^{n}\begin{pmatrix}W_i\epsilon_i\\ W_i\delta_i\end{pmatrix}$$

as we also have that $\mathbb{E}(W_i\epsilon_i) = \mathbb{E}(W_i\delta_i) = 0$ and then lastly I applied Slutsky's lemma for multiplication.
It holds that:

$$\begin{aligned}
\text{Cov}(W_i\epsilon_i) &= \mathbb{E}(\epsilon_i^2 W_i W_i^T)\\
\text{Cov}(W_i\epsilon_i, W_i\delta_i) &= \mathbb{E}(\epsilon_i\delta_i W_i W_i^T)\\
\text{Cov}(W_i\delta_i) &= \mathbb{E}(\delta_i^2 W_i W_i^T)
\end{aligned}$$

Hence, we end up with:

$$\begin{aligned}
\sqrt{n}(\begin{pmatrix}\hat{\Gamma}\\ \hat{\Psi}\\ \hat{\gamma}\\ \hat{\psi}\end{pmatrix} - \begin{pmatrix}\Gamma\\ \Psi\\ \gamma\\ \psi\end{pmatrix}) &\xrightarrow{\text{d}} \begin{pmatrix}\Sigma^{-1} & 0\\ 0 & \Sigma^{-1}\end{pmatrix}\mathcal{N}(0, \begin{pmatrix}\mathbb{E}(\epsilon_i^2 W_i W_i^T) & \mathbb{E}(\epsilon_i\delta_i W_i W_i^T)\\ \mathbb{E}(\epsilon_i\delta_i W_i W_i^T) & \mathbb{E}(\delta_i^2 W_i W_i^T))\end{pmatrix}\\
&\stackrel{\text{d}}{=} \mathcal{N}(0, \begin{pmatrix}\Sigma^{-1}\,\mathbb{E}(\epsilon_i^2 W_i W_i^T)\Sigma^{-1} & \Sigma^{-1}\,\mathbb{E}(\epsilon_i\delta_i W_i W_i^T)\Sigma^{-1}\\ \Sigma^{-1}\,\mathbb{E}(\epsilon_i\delta_i W_i W_i^T)\Sigma^{-1} & \Sigma^{-1}\,\mathbb{E}(\delta_i^2 W_i W_i^T)\Sigma^{-1}\end{pmatrix})\\
&=: \mathcal{N}(0, \text{Cov}^A)
\end{aligned}$$

Again using Slutsky's for multiplication, we obtain that:

$$\sqrt{n}(\begin{pmatrix}\hat{\Gamma}\\ \hat{\gamma}\end{pmatrix} - \begin{pmatrix}\Gamma^*\\ \gamma^*\end{pmatrix}) \xrightarrow{\text{d}} \mathcal{N}(0, \begin{pmatrix}V^\Gamma & C\\ C & V^\gamma\end{pmatrix})$$

□

As notation, I will write that $\text{Var}(\begin{pmatrix}\hat{\Gamma}\\ \hat{\gamma}\end{pmatrix}) = \begin{pmatrix}V^\Gamma & C\\ C & V^\gamma\end{pmatrix}$ throughout the text. As a consequence of theorem 2.4.3, we also obtain consistency of the OLS-estimators (apply Slutsky's lemma to the expression that converges in distribution to the normal and the sequence $\frac{1}{\sqrt{n}}$ by multiplying them).

2.4.2. Challenges with variance estimation

For this β^* identification method, estimators are used for the asymptotic covariance matrix of $(\hat{\Gamma}, \hat{\gamma})^T$. The following estimators are used for the errors terms:

$$\hat{\epsilon}_i = Y_i - Z_i^T\hat{\Gamma} - X_i^T\hat{\Psi}$$
$$\hat{\delta}_i = D_i - Z_i^T\hat{\gamma} - X_i^T\hat{\psi}$$

Consequently, we get the following estimators for the elements of $\text{Var}(\begin{pmatrix}\hat{\Gamma}\\ \hat{\gamma}\end{pmatrix})$:

$$\hat{V}^\Gamma = [\hat{\Sigma}^{-1}(\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 W_i W_i^T)\hat{\Sigma}^{-1}]_{1:p_Z,1:p_Z}$$
$$\hat{V}^\gamma = [\hat{\Sigma}^{-1}(\frac{1}{n}\sum_{i=1}^n \hat{\delta}_i^2 W_i W_i^T)\hat{\Sigma}^{-1}]_{1:p_Z,1:p_Z}$$
$$\hat{C} = [\hat{\Sigma}^{-1}(\frac{1}{n}\sum_{i=1}^n \hat{\delta}_i\hat{\epsilon}_i W_i W_i^T)\hat{\Sigma}^{-1}]_{1:p_Z,1:p_Z}$$

For the estimators above, we might hope to obtain convergence (in probability) without too many extra constraints. We might wonder whether for example:

$$\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 W_i W_i^T \xrightarrow[?]{\mathbb{P}} \mathbb{E}(\epsilon_i^2 W_i W_i^T) \tag{2.4.1}$$

Although it is true that by consistency of $\hat{\Gamma}$ and $\hat{\Psi}$:

$$\hat{\epsilon}_i^2 \xrightarrow{\mathbb{P}} \epsilon_i^2, \quad \forall i \geq 1$$

$\hat{\epsilon}_i^2$ are not, in general, iid samplings from the distribution of ϵ_i^2. Hence, the law of large numbers can't be directly applied. Despite this, under the same conditions as theorem 2.4.3 plus extra mixed moments assumptions (up to the 4-th moment), there is consistency for the variance estimators. These variance estimators were mentioned in [2, p.9 C.Proofs]. Here it was not alleged (or directly used) that these were consistent estimators under some conditions and I came up with the conditions and proof myself.

Theorem 2.4.4. *Under the same conditions as theorem 2.4.3 together with the assumption that the following list of mixed moments are finite* $\forall k, l = 1, ..., p$*:* $\mathbb{E}(\epsilon_i Z_i^T W_i^{(k)} W_i^{(l)})$, $\mathbb{E}(Z_i Z_i^T W_i^{(k)} W_i^{(l)})$, $\mathbb{E}(\epsilon_i W_i^{(k)} W_i^{(l)})$, $\mathbb{E}(X_i X_i^T W_i^{(k)} W_i^{(l)})$, $\mathbb{E}(Z_i X_i^T W_i^{(k)} W_i^{(l)})$, $\mathbb{E}(\delta_i Z_i^T W_i^{(k)} W_i^{(l)})$ *and* $\mathbb{E}(\delta_i X_i^T W_i^{(k)} W_i^{(l)})$*. Then:*

$$(\hat{V}^\Gamma, \hat{V}^\gamma, \hat{C}) \xrightarrow{\mathbb{P}} (V^\Gamma, V^\gamma, C)$$

Proof. I will show that $\hat{V}^\Gamma \xrightarrow{\mathbb{P}} V^\Gamma$ and the others can be shown in a similar manner.
By observing that: $\hat{\epsilon}_i = \epsilon_i + Z_i^T(\Gamma^* - \hat{\Gamma}) + X_i^T(\Psi^* - \hat{\Psi})$, one can also see that:

$$\begin{aligned}\hat{\epsilon}_i^2 = \epsilon_i^2 &+ 2\epsilon_i Z_i^T(\Gamma^* - \hat{\Gamma}) + 2\epsilon_i X_i^T(\Psi^* - \hat{\Psi}) + (Z_i^T(\Gamma^* - \hat{\Gamma}))^2 + (X_i^T(\Gamma^* - \hat{\Gamma}))^2 \\ &+ 2Z_i^T(\Gamma^* - \hat{\Gamma})X_i^T(\Psi^* - \hat{\Psi})\end{aligned}$$

Furthermore, it holds that (using definition 2.4.2 and lemma 2.2.7):

$$\Gamma^* - \hat{\Gamma} = [(W^TW)^{-1}]_{1:p_Z,\cdot} \sum_{j=1}^{n} \epsilon_j W_j$$

Now we will look at the terms of $\frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i^2 - \epsilon_i^2)W_iW_i^T$ related to Z_i. The X_i terms follow a similar argument.
We first look at:

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i Z_i^T(\Gamma^* - \hat{\Gamma})W_iW_i^T = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i Z_i^T([(W^TW)^{-1}]_{1:p_Z,\cdot}\sum_{j=1}^{n}\epsilon_j W_j)W_iW_i^T \tag{2.4.2}$$

Now consider the (k,l)-th element of (2.4.2):

$$\begin{aligned}[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i Z_i^T(\Gamma^* - \hat{\Gamma})W_iW_i^T]_{k,l} = \\ \frac{1}{n}\sum_{i=1}^{n}\epsilon_i Z_i^T([(W^TW)^{-1}]_{1:p_Z,\cdot}\sum_{j=1}^{n}\epsilon_j W_j)W_i^{(k)}W_i^{(l)} = \\ \{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i Z_i^T W_i^{(k)}W_i^{(l)}\}\{[(\frac{1}{n}W^TW)^{-1}]_{1:p_Z,\cdot}\frac{1}{n}\sum_{j=1}^{n}\epsilon_j W_j\}\end{aligned}$$

By the law of large numbers (and the assumptions of this theorem):

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i Z_i^T W_i^{(k)}W_i^{(l)} \xrightarrow{\mathbb{P}} \mathbb{E}(\epsilon_i Z_i^T W_i^{(k)}W_i^{(l)})$$

By the law of large numbers and continuous mapping theorem (together with the fact that Σ is finite):

$$(\frac{1}{n}W^TW)^{-1} \xrightarrow{\mathbb{P}} \Sigma^{-1}$$

hence also:

$$[(\frac{1}{n}W^TW)^{-1}]_{1:p_Z,\cdot} \xrightarrow{\mathbb{P}} [\Sigma^{-1}]_{1:p_Z,\cdot}$$

As $\mathbb{E}(\epsilon_j W_j) = 0$ by definition 2.2.1:

$$\frac{1}{n}\sum_{j=1}^{n}\epsilon_j W_j \xrightarrow{\mathbb{P}} \mathbb{E}(\epsilon_j W_j) = 0$$

Hence, overall $\forall k, l = 1, ..., p$:

$$[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i Z_i^T(\Gamma^* - \hat{\Gamma})W_i W_i^T]_{k,l} \xrightarrow{\mathbb{P}} 0$$

Which gives us:

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i Z_i^T(\Gamma^* - \hat{\Gamma})W_i W_i^T \xrightarrow{\mathbb{P}} 0$$

Next, consider:

$$[\frac{1}{n}\sum_{i=1}^{n}(Z_i^T(\Gamma^* - \hat{\Gamma}))^2 W_i W_i^T]_{k,l} =$$
$$\frac{1}{n}\sum_{i=1}^{n}(Z_i^T(\Gamma^* - \hat{\Gamma}))^2 W_i^{(k)} W_i^{(l)} =$$
$$\frac{1}{n}\sum_{i=1}^{n}(\Gamma^* - \hat{\Gamma})^T Z_i Z_i^T(\Gamma^* - \hat{\Gamma})W_i^{(k)} W_i^{(l)} =$$
$$(\Gamma^* - \hat{\Gamma})^T(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T W_i^{(k)} W_i^{(l)})(\Gamma^* - \hat{\Gamma}) \xrightarrow{\mathbb{P}} 0$$

Lastly:

$$[\frac{1}{n}\sum_{i=1}^{n} Z_i^T(\Gamma^* - \hat{\Gamma})X_i^T(\Psi^* - \hat{\Psi})W_i W_i^T]_{k,l} =$$
$$\frac{1}{n}\sum_{i=1}^{n} Z_i^T(\Gamma^* - \hat{\Gamma})X_i^T(\Psi^* - \hat{\Psi})W_i^{(k)} W_i^{(l)} =$$
$$(\Gamma^* - \hat{\Gamma})^T\{\frac{1}{n}\sum_{i=1}^{n} Z_i X_i^T W_i^{(k)} W_i^{(l)}\}(\Psi^* - \hat{\Psi}) \xrightarrow{\mathbb{P}} 0$$

□

2.4.3. Standard error estimator $\hat{\pi}$ based on $\hat{\Gamma}$, $\hat{\gamma}$

For the methods in the next part of the text, the following estimator for π_k^* is frequently used:

$$\text{For } k, j \in S : \pi_k^{[j]} := \hat{\Gamma}_k - \hat{\gamma}_k \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} =: \hat{\Gamma}_k - \hat{\gamma}_k \hat{\beta}^{[j]}$$

Observe that for $j \in V$: $\pi_k^{[j]} \xrightarrow{\mathbb{P}} \pi_k^*$. $\pi_k^{[j]}$ also has asymptotic normality properties by theorem 2.4.3. Next we will consider the standard error[1] (SE) of $\pi_k^{[j]}$. For $j \in V$ and by asymptotic normality: $\mathbb{P}(\pi_k^* \in (\pi_k^{[j]} - z_{\alpha/2}\,\mathrm{SE}, \pi_k^{[j]} + z_{\alpha/2}\,\mathrm{SE})) \to 1 - \alpha$. So using (an estimator of) SE, we can obtain a confidence interval for π_k^*, where $z_{\alpha/2}$ is the $\alpha/2$-th quantile of the standard normal distribution. Before providing the expression for the standard error, some notation is introduced.

[1] Given that $\sqrt{n}(x_n - x) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2), \sigma^2 \in \mathbb{R}$, the standard error is defined as $\sqrt{\frac{1}{n}\sigma^2}$

Notation 2.4.5. *For $j, k \in S$, define:*

$$\begin{aligned}
\beta^{[j]} &:= \frac{\Gamma_j^*}{\gamma_j^*} \\
R^{[j]} &:= V^\Gamma + (\beta^{[j]})^2 V^\gamma - 2\beta^{[j]} C \\
T_{j,k}^0 &:= \sqrt{\frac{1}{n}\left(\frac{R_{k,k}^{[j]}}{(\gamma_k^*)^2} + \frac{R_{j,j}^{[j]}}{(\gamma_j^*)^2} - 2\frac{R_{j,k}^{[j]}}{\gamma_k^*\gamma_j^*}\right)} \\
T_{j,k} &:= \min(T_{j,k}^0, T_{k,j}^0)
\end{aligned}$$

The interpretation of the $T_{j,k}^0$ term is intertwined with the SE, as can be seen in the next proposition. This result was first mentioned at [2, p.6 Section 3] (up to an absolute value). I came up with the proof myself.

Proposition 2.4.6. *For $j \in S$: $|\gamma_k^*| T_{j,k}^0$ is equal to the standard error of $\hat{\Gamma}_k - \hat{\gamma}_k \hat{\beta}^{[j]}$*

Proof. Consider the following mapping from $\mathbb{R}^{2p_Z}$ to $\mathbb{R}$:

$$\phi : (a_1, .., a_{p_Z}, b_1, .., b_{p_Z}) \mapsto a_k - b_k \frac{a_j}{b_j}$$

. Wlog assume that $k < j$. As $j \in S$, ϕ is differentiable at $(\Gamma^*, \gamma^*)^T$, where $\phi'((\Gamma^*)^T, (\gamma^*)^T) \in \mathbb{R}^{1 \times 2p_Z}$ has the following coordinates:

$$\begin{aligned}
[\phi'((\Gamma^*)^T, (\gamma^*)^T)]_k &= 1 \\
[\phi'((\Gamma^*)^T, (\gamma^*)^T)]_j &= \frac{-\gamma_k^*}{\gamma_j^*}, \\
[\phi'((\Gamma^*)^T, (\gamma^*)^T)]_{p_Z+k} &= \frac{-\Gamma_j^*}{\gamma_j^*}, \\
[\phi'((\Gamma^*)^T, (\gamma^*)^T)]_{p_Z+j} &= \frac{\Gamma_j^* \gamma_k^*}{(\gamma_j^*)^2}
\end{aligned}$$

For all the other coordinates, it is 0. By the Delta Rule, we obtain that:

$$\begin{aligned}
\sqrt{n}(\pi_k^{[k]} - (\Gamma_k^* - \gamma_k^* \frac{\Gamma_j^*}{\gamma_j^*})) &\xrightarrow{\text{d}} \mathcal{N}(0, \phi'((\Gamma^*)^T, (\gamma^*)^T) \begin{pmatrix} V^\Gamma & C \\ C & V^\gamma \end{pmatrix} [\phi'((\Gamma^*)^T, (\gamma^*)^T)]^T) \\
&=: \mathcal{N}(0, \Sigma')
\end{aligned}$$

Writing out Σ' and re-arranging terms gives:

$$\begin{aligned}
\Sigma' &= [V_{k,k}^\Gamma - 2\frac{\Gamma_j^*}{\gamma_j^*} C_{k,k} + (\frac{\Gamma_j^*}{\gamma_j^*})^2 V_{k,k}^\gamma] + [-2\frac{\gamma_k^*}{\gamma_j^*}(V_{j,k}^\Gamma - 2\frac{\Gamma_j^*}{\gamma_j^*} C_{j,k} + (\frac{\Gamma_j^*}{\gamma_j^*})^2 V_{k,j}^\gamma)] \\
&\quad + [(\frac{\gamma_k^*}{\gamma_j^*})^2 (V_{j,j}^\Gamma - 2\frac{\Gamma_j^*}{\gamma_j^*} C_{j,j} + (\frac{\Gamma_j^*}{\gamma_j^*})^2 V_{j,j}^\gamma)] \\
&=: [A1] + [A2] + [A3].
\end{aligned}$$

Per definition, we see that:

$$[A1] = R_{k,k}^{[j]}$$

$$[A2] = -2\frac{\gamma_k^*}{\gamma_j^*}R_{j,k}^{[j]}$$
$$[A3] = (\frac{\gamma_k^*}{\gamma_j^*})^2 R_{j,j}^{[j]}$$

Hence, the standard error is equivalent to:

$$\sqrt{\frac{1}{n}(R_{j,j}^{[j]} - 2\frac{\gamma_k^*}{\gamma_j^*}R_{j,k}^{[j]} + (\frac{\gamma_k^*}{\gamma_j^*})^2 R_{j,j}^{[j]})} = |\gamma_k^*|T_{j,k}^0$$

□

2.4.4. Locally invalid instrumental variables

Lastly, before introducing the proposed method (named Searching and Sampling), we will briefly consider the concept of locally invalid IVs. In the previous section it was mentioned that due to asymptotic normality: for $j \in V$, it holds that $\mathbb{P}(\pi_k^* \in (\pi_k^{[j]} - z_{\alpha/2}\,\mathrm{SE}, \pi_k^{[j]} + z_{\alpha/2}\,\mathrm{SE})) \to 1-\alpha$ or equivalently: $\mathbb{P}(\pi_k^{[j]} \in (\pi_k^* - z_{\alpha/2}|\gamma_k^*|T_{j,k}^0, \pi_k^* + z_{\alpha/2}|\gamma_k^*|T_{j,k}^0)) \to 1-\alpha$. Suppose that we want to determine whether π_k^* is 0 or not using the estimator $\pi_k^{[j]}$. In case $0 < |\pi_k^*| < z_{\alpha/2}|\gamma_k^*|T_{j,k}^0$ for a large n, we will have 0 in our confidence interval above even for a large n. This makes it hard to determine for any data-based method whether $Z^{(k)}$ is valid or invalid. We call these small π_k^* locally invalid and it is formally defined below:

Definition 2.4.7. *For $j \in S$, the j-th IV is locally invalid if:*

$$0 < |\frac{\pi_j^*}{\gamma_j^*}| < s_j(n),\ s_j(n) := 2\sqrt{\log(n)}\max_{k\in V}|T_{j,k}|$$

Remark 2.4.8.

- *Any locally invalid IV is also invalid.*
- *Later, it will be shown that if the j-th IV is invalid, then it can be separated from the valid IVs, using data-based methods, if $\frac{\pi_j^*}{\gamma_j^*} \geq s_j(n)$. It turns out that we in the other case don't have any (asymptotic) guarantees to whether we can separate the invalid from the valid IVs (see proposition 2.7.17 for exact details).*
- *As $\frac{\log(n)}{n} \xrightarrow{n\to\infty} 0$, using continuous mapping and the fact that we can express $\max_{k\in V}|T_{j,k}|$ as a constant (wrt n) divided by square-root n: $s_j(n) \xrightarrow{n\to\infty} 0$. Hence, the number of locally invalid IVs will decrease to 0 as $n \to \infty$*

Due to the difficulty in separating valid IVs from locally invalid IVs, it is a frequent occurrence that estimated sets for V will contain locally invalid IVs. A key difference between the Searching and Sampling method and other inference methods is that the Searching and Sampling method can correctly recover β^*, even with the presence of locally invalid IVs. This means that this method does not rely on 100 % correctly recovering V, but rather on recovering it "well enough" [2, p.2 1.Introduction]. In section 2.6, this procedure is explained in more detail.

2.5. Searching and Sampling: robust inference methods under majority rule

From the reduced-form equations, we know that the following identity holds:

$$\Gamma_j^* = \beta^* \gamma_j^* + \pi_j^*,\ 1 \leq j \leq p_Z$$

Our strategy to recover β^* is to estimate Γ^* and γ^* and then check the majority or plurality rule. Note that in case $\gamma_j^* = 0$, we can't recover β^* even when Γ_j^* is known as well. As we want to be able to exactly recover β^* with the proposed methods in an asymptotic sense ($\hat{\Gamma}$ and $\hat{\gamma}$ converge a.s. to Γ^* and γ^* respectively using the SLLNs), we want to exclude cases where the estimator for γ_j^* gives a non-zero value while the asymptotic value is 0 (because then we know that for $n \to \infty$ our inference is incorrect!). It would hence be nice if we beforehand could have a guarantee that when $|\hat{\gamma}_j|$ is "large enough", then γ_j^* won't be zero (with a high probability).

2.5.1. Hard thresholding

Define the following two sets:

$$\hat{S} := \{1 \leq j \leq p_Z : |\hat{\gamma}_j| \geq \sqrt{\log(n)\frac{\hat{V}_{j,j}^{\gamma}}{n}}\}$$

$$S_{\text{str}} := \{1 \leq j \leq p_Z : |\gamma_j^*| \geq 2\sqrt{\log(n)\frac{V_{j,j}^{\gamma}}{n}}\} \text{ (set of strongly relevant IVs)}$$

It will turn out that with high probability (for n large enough):

$$S_{\text{str}} \subseteq \hat{S} \subseteq S \tag{2.5.1}$$

Hence, with high probability, $\hat{S}$ contains only relevant potential IVs and at least all the strongly relevant potential IVs! This means that we (with high probability) don't have to worry about the scenario described at the start of the section.

As we established that we can have certain asymptotic guarantees with strongly relevant IVs, we want to assume that we have at least a sufficient number of them as to be able to do β^* inference. For this, I will now establish a finite sample adjusted majority rule:

Condition 2.5.1. *The finite sample majority rule: More than half of the relevant IVs are strongly relevant and valid i,e.:* $|V \cap S_{\text{str}}| > \frac{|S|}{2}$

Observe that with condition 2.5.1 and (2.5.1), the majority of $\hat{S}$ is valid. Another property for the finite sample majority rule is that as $S_{\text{str}} = S$ for $n \to \infty$, it is the same as the population majority rule asymptotically.

2.5.2. Inference of the treatment effect

Define $\pi_j(\beta) := \Gamma_j^* - \beta\gamma_j^*$ and $\hat{\pi}_j(\beta) := \hat{\Gamma}_j - \beta\hat{\gamma}_j$ for $\beta \in \mathbb{R}$. Note that for $\beta = \beta^*$: $\pi_j(\beta) = \pi_j^*$ and $\hat{\pi}_j(\beta) \xrightarrow{\mathbb{P}} \pi_j^*$. Our goal will be to test for $\beta \in \mathbb{R}$ and $j \in \hat{S}$ whether $\beta = \beta^*$ and $j \in V$ both hold. For this, we consider the difference $|\pi_j(\beta) - \hat{\pi}_j(\beta)|$ which, under the hypothesis that $\beta = \beta^*$ and $j \in V$, equals $|\hat{\pi}_j(\beta)|$ (which is computable from the data). Due to the asymptotic normality as seen in lemma 2.4.3, the next lemma shows an asymptotic normality property used for the hypothesis testing. The outcome of this lemma was first stated in the original paper [2, p.9 4.2 The searching confidence interval] and was there later proven under stronger conditions than the ones presented here [2, p.18 6.Theoretical Justification Theorem 1]. I came up with the weaker conditions and the proof of the following lemma.

Lemma 2.5.2. *Let $\beta \in \mathbb{R}$. Under the same conditions as theorem 2.4.3 and the assumptions that*

$$(\hat{V}_{j,j}^{\Gamma}, \hat{V}_{j,j}^{\gamma}, \hat{C}_{j,j}) \xrightarrow{\mathbb{P}} (V_{j,j}^{\Gamma}, V_{j,j}^{\gamma}, C_{j,j}) \tag{2.5.2}$$

together with

$$\lim_{n\to\infty} \mathbb{P}(S_{\text{str}} \subseteq \hat{S} \subseteq S) = 1 \tag{2.5.3}$$

we have that:

$$\liminf_{n\to\infty} \mathbb{P}(\max_{j\in\hat{S}} \frac{|\pi_j(\beta) - \hat{\pi}_j(\beta)|}{\sqrt{\frac{1}{n}(\hat{V}_{j,j}^{\Gamma} + \beta^2\hat{V}_{j,j}^{\gamma} - 2\beta\hat{C}_{j,j})}} \leq \Phi^{-1}(1 - \frac{\alpha}{2|\hat{S}|})) \geq 1 - \alpha$$

Proof. By theorem 2.4.3, in combination with multiplying by a linear mapping, we obtain that:

$$\sqrt{n}(\hat{\pi}_j(\beta) - \pi_j(\beta)) = \sqrt{n}(\hat{\Gamma}_j - \Gamma_j^* - \beta(\hat{\gamma}_j - \gamma_j^*)) \xrightarrow{\text{d}} \mathcal{N}(0, V_{j,j}^{\Gamma} + \beta^2 V_{j,j}^{\gamma} - 2\beta C_{j,j})$$

Furthermore by assumption 2.5.2 and the continuous mapping theorem:

$$\sigma_{n,j} := \sqrt{\hat{V}_{j,j}^{\Gamma} + \beta^2\hat{V}_{j,j}^{\gamma} - 2\beta\hat{C}_{j,j}} \xrightarrow{\mathbb{P}} \sqrt{V_{j,j}^{\Gamma} + \beta^2 V_{j,j}^{\gamma} - 2\beta C_{j,j}} =: \sigma_j$$

and so by Slutsky's lemma (and the continuous mapping theorem):

$$\sqrt{n}\sigma_{n,j}^{-1}(\hat{\pi}_j(\beta) - \pi_j(\beta)) \xrightarrow{\text{d}} \mathcal{N}(0,1) \tag{2.5.4}$$

Observe that:

$$\begin{aligned}
&\mathbb{P}(\max_{j\in\hat{S}} \sqrt{n}\sigma_{n,j}^{-1}|\pi_j(\beta) - \hat{\pi}_j(\beta)| \leq \Phi^{-1}(1 - \frac{\alpha}{2|\hat{S}|}), S_{\text{str}} \subseteq \hat{S} \subseteq S) \geq \\
&\mathbb{P}(\max_{j\in S} \sqrt{n}\sigma_{n,j}^{-1}|\pi_j(\beta) - \hat{\pi}_j(\beta)| \leq \Phi^{-1}(1 - \frac{\alpha}{2|\hat{S}|}), S_{\text{str}} \subseteq \hat{S} \subseteq S) = \\
&1 - \mathbb{P}(\max_{j\in S} \sqrt{n}\sigma_{n,j}^{-1}|\pi_j(\beta) - \hat{\pi}_j(\beta)| > \Phi^{-1}(1 - \frac{\alpha}{2|\hat{S}|}) \vee (S_{\text{str}} \subseteq \hat{S} \subseteq S)^c) \geq \\
&1 - \mathbb{P}(\max_{j\in S} \sqrt{n}\sigma_{n,j}^{-1}|\pi_j(\beta) - \hat{\pi}_j(\beta)| > \Phi^{-1}(1 - \frac{\alpha}{2|\hat{S}|})) - \mathbb{P}((S_{\text{str}} \subseteq \hat{S} \subseteq S)^c) \geq \\
&1 - \sum_{j\in S} \mathbb{P}(\sqrt{n}\sigma_{n,j}^{-1}|\pi_j(\beta) - \hat{\pi}_j(\beta)| > \Phi^{-1}(1 - \frac{\alpha}{2|\hat{S}|})) - \mathbb{P}((S_{\text{str}} \subseteq \hat{S} \subseteq S)^c)
\end{aligned}$$

Observe that:

$$-\mathbb{P}(\sqrt{n}\sigma_{n,j}^{-1}|\pi_j(\beta)-\hat{\pi}_j(\beta)|>\Phi^{-1}(1-\frac{\alpha}{2|\hat{S}|}),S_{\text{str}}\subseteq\hat{S}\subseteq S)\geq$$
$$-\mathbb{P}(\sqrt{n}\sigma_{n,j}^{-1}|\pi_j(\beta)-\hat{\pi}_j(\beta)|>\Phi^{-1}(1-\frac{\alpha}{2|S_{\text{str}}|}),S_{\text{str}}\subseteq\hat{S}\subseteq S)$$

Using that $|S_{\text{str}}|\to|S|$ as $n\to\infty$, Φ^{-1} being continuous, (2.5.4), $|S|$ being a deterministic finite set (independent of n) and assumption (2.5.3), we obtain the desired result. □

Using the lemma 2.5.2 and the finite sample majority rule, we can search for β^* by repeatedly testing the hypothesis that $\beta=\beta^*$ and $j\in V$. The output of this method (described below) is a set with potential β^* values.

Method 2.5.3. *(Searching algorithm,[2, p.9,10 Section 4.2])*

1. *For $\beta\in\mathbb{R}$, $j\in\hat{S}$ compute:*

$$\hat{p}_j(\beta):=\Phi^{-1}(1-\frac{\alpha}{2|\hat{S}|})\sqrt{\frac{1}{n}(\hat{V}^{\Gamma}_{j,j}+\beta^2\hat{V}^{\gamma}_{j,j}-2\beta\hat{C}_{j,j})}$$
$$\bar{\pi}_j(\beta):=\hat{\pi}_j(\beta)1(|\hat{\pi}_j|\geq\hat{p}_j(\beta))$$

2. $\text{CI}^{\text{sear}}:=\{\beta\in\mathbb{R}:\left\|\bar{\pi}_{\hat{S}}(\beta)\right\|_0<\frac{|\hat{S}|}{2}\}$[2]

Remark 2.5.4.

- CI^{sear} *is not guaranteed to be an interval.*
- *For large n the probability will be high that $\bar{\pi}_j(\beta^*)=0$ for $j\in V\cap S_{\text{str}}$. In case all such j are correctly classified, the majority rule applied to $\hat{S}$ is met (with high probability). Hence (with high probability), it wouldn't matter if it happens that for some $j\notin V$ (like locally invalid IVs) we get $\bar{\pi}_j(\beta^*)=0$.*

2.5.3. Efficient implementation of searching CI

For the method above, we need to consider every $\beta\in\mathbb{R}$ in order to construct CI^{sear}. To improve the computational efficienty of this method, we would prefer to have a (well-chosen) grid of $\beta's$. Preferably this would result in an interval rather than a range of values. For this we can do the following [2, p.10,11 Section 4.3]:

- Our strategy is to first obtain an interval $[L,U]$ s.t. $P(\beta^*\in[L,U])\xrightarrow{n\to\infty}1$ and from there construct a grid set with step size n^{-a} with $a>\frac{1}{2}$ (i.e. smaller than the parametric rate).
- Define $\mathcal{B}=\{\beta_1,..,\beta_k\}$ as the n^{-a}-step sized grid for $[L,U]$.
- For $j\in S$, β^* is estimated by the ratio $\frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$ under the hypothesis that $\pi_j^*=0$. The variance is estimated by:

$$\hat{\text{Var}}(\frac{\hat{\Gamma}_j}{\hat{\gamma}_j})=\frac{\hat{V}^{\Gamma}_{jj}}{\hat{\gamma}_j^2}+\frac{\hat{V}^{\gamma}_{j,j}\hat{\Gamma}_j^2}{\hat{\gamma}_j^4}-2\frac{\hat{C}_{j,j}\hat{\Gamma}_j}{\hat{\gamma}_j^3}$$

[2]Here: $\bar{\pi}_{\hat{S}}=(\bar{\pi}_j)_{j\in\hat{S}}$ and $\|.\|_0$ denotes the amount of non-zero elements in a vector

- Then we define L, U as follows:

$$L = \min_{j \in \hat{S}} \{ \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} - \sqrt{\frac{\log(n)}{n} \hat{\text{Var}}(\frac{\hat{\Gamma}_j}{\hat{\gamma}_j})} \}$$
$$U = \max_{j \in \hat{S}} \{ \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} + \sqrt{\frac{\log(n)}{n} \hat{\text{Var}}(\frac{\hat{\Gamma}_j}{\hat{\gamma}_j})} \}$$

- We obtain the (approximated) confidence interval as follows:

$$\hat{\text{CI}}^{\text{sear}} = [\min_{\beta \in \mathcal{B}: \left\| \bar{\pi}_{\hat{S}}(\beta) \right\|_0 < \frac{|\hat{S}|}{2}} \beta, \max_{\beta \in \mathcal{B}: \left\| \bar{\pi}_{\hat{S}}(\beta) \right\|_0 < \frac{|\hat{S}|}{2}} \beta]$$

The next lemma shows that indeed (under some conditions) $\mathbb{P}(\beta^* \in [L, U]) \to 1$ for $n \to \infty$. The end result of this next lemma was first mentioned at [2, p.11 4.3 Efficient implementation of the searching CI] and later proven under stronger conditions [2, p.18 6.Theoretical Justification Theorem 1]. I came up with the conditions and proof myself:

Lemma 2.5.5. *Under the same conditions as theorem 2.4.3 together with*

$$(\hat{V}^{\Gamma}_{jj}, \hat{V}^{\gamma}_{jj}, \hat{C}_{jj})^T \xrightarrow{\mathbb{P}} (V^{\Gamma}_{jj}, V^{\gamma}_{jj}, C_{jj})^T \tag{2.5.5}$$
$$\lim_{n \to \infty} \mathbb{P}(S_{\text{str}} \subseteq S) = 1 \tag{2.5.6}$$

and the finite sample majority rule we have that for

$$L = \min_{j \in \hat{S}} \{ \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} - \sqrt{\frac{\log(n)}{n} \hat{\text{Var}}(\frac{\hat{\Gamma}_j}{\hat{\gamma}_j})} \}$$
$$U = \max_{j \in \hat{S}} \{ \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} + \sqrt{\frac{\log(n)}{n} \hat{\text{Var}}(\frac{\hat{\Gamma}_j}{\hat{\gamma}_j})} \}$$

it holds that: $\lim_{n \to \infty} \mathbb{P}(\beta^* \in [L, U]) = 1$

Proof. Let $n \geq 3$. Then the following holds:

$$\mathbb{P}(\beta^* \in [L, U]) \geq \mathbb{P}(\beta^* \in [L, U], S_{\text{str}} \subseteq \hat{S}) =: (\blacksquare)$$

$$(\blacksquare) \geq \mathbb{P}(\beta^* \in [\min_{j \in S_{\text{str}}} \{ \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} - \sqrt{\frac{\log(n)}{n} \hat{\text{Var}}(\frac{\hat{\Gamma}_j}{\hat{\gamma}_j})} \}, \max_{j \in S_{\text{str}}} \{ \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} + \sqrt{\frac{\log(n)}{n} \hat{\text{Var}}(\frac{\hat{\Gamma}_j}{\hat{\gamma}_j})} \}], S_{\text{str}} \subseteq \hat{S})$$
$$\geq \mathbb{P}(\beta^* \in [\frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}} - \sqrt{\frac{\log(n)}{n} \hat{\text{Var}}(\frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}})}, \frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}} + \sqrt{\frac{\log(n)}{n} \hat{\text{Var}}(\frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}})}], S_{\text{str}} \subseteq \hat{S}) =: (\blacktriangle)$$

In the last step, j_1 denotes a valid and strongly relevant IV in S_{str} for $n = 3$. We know j_1 exists due to the finite sample majority rule and that $j_1 \in S_{\text{str}}$ for $n \geq 3$ because of $\frac{\log(n)}{n}$ decreasing.
As

$$(\blacktriangle) = \mathbb{P}(\frac{|\beta^* - \frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}}|}{\sqrt{\hat{\text{Var}}(\frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}})}} \leq \sqrt{\frac{\log(n)}{n}}, S_{\text{str}} \subseteq \hat{S})$$

with $\hat{\text{Var}}(\frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}}) \xrightarrow{\mathbb{P}} \text{Var}(\frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}})$ (by (2.5.5) and continuous mapping theorem) and $\frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}} \xrightarrow{\mathbb{P}} \beta^*$ (consequence of theorem 2.4.3), we obtain that:

$$\frac{|\beta^* - \frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}}|}{\sqrt{\hat{\text{Var}}(\frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}})}} \xrightarrow{\mathbb{P}} 0$$

and consequently for $n \to \infty$:

$$\mathbb{P}(\frac{|\beta^* - \frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}}|}{\sqrt{\hat{\text{Var}}(\frac{\hat{\Gamma}_{j_1}}{\hat{\gamma}_{j_1}})}} \leq \sqrt{\frac{\log(n)}{n}}) \to 1$$

Together with (2.5.6), we can hence conclude that:

$$\lim_{n\to\infty} \mathbb{P}(\beta^* \in [L, U]) = 1$$

□

Remark 2.5.6. *When $\nexists \beta \in \mathcal{B} : \left\|\bar{\pi}_{\hat{S}}(\beta)\right\|_0 < \frac{|\hat{S}|}{2}$, $\hat{\text{CI}}^{\text{sear}}$ is empty. Then (with high likelihood if we have enough data), the (sample) majority rule is violated.*

2.5.4. Sampling CI

Next, we will consider an extension to the searching method above. It involves resampling $(\hat{\Gamma}, \hat{\gamma})^T$ and then applying the searching method multiple times. The definition of the resampled coefficients is specified below.

Definition 2.5.7. *The resampled coefficients $\{\hat{\Gamma}, \hat{\gamma}\}_{1\leq m\leq M}$ are defined as follows: conditioned on the observed data:*

$$\begin{pmatrix}\hat{\Gamma}^{[m]} \\ \hat{\gamma}^{[m]}\end{pmatrix} \sim N(\begin{pmatrix}\hat{\Gamma} \\ \hat{\gamma}\end{pmatrix}, \begin{pmatrix}\frac{\hat{V}^{\Gamma}}{n}, \frac{\hat{C}}{n} \\ \frac{\hat{C}}{n}, \frac{\hat{V}^{\gamma}}{n}\end{pmatrix})$$

$1 \leq m \leq M$. *M is called the resampling size.*

The idea of obtaining a confidence interval via sampling is as follows:

- Compared to the inference method of the previous section, we use $(\hat{\Gamma}^{[m]}, \hat{\gamma}^{[m]})$ instead of $(\hat{\Gamma}, \hat{\gamma})$ to construct a searching CI. This new CI, that uses the m-th resampled coefficient, will be referred to as the m-th sampled searching CI.
- Using the sampled gamma's, we can decrease the hard thresholding level in the estimation of π_j^* step (see section 2.7). This will lead to shorter resulting CIs for the β.

The method is summarised below [2, p.12 Section 4.4]:

- For each m between 1 and M, we estimate π_j^* for the m-th sample:

  $$\hat{\pi}_j^{[m]} = (\hat{\Gamma}_j^{[m]} - \beta\hat{\gamma}_j^{[m]})1(|\hat{\Gamma}_j^{[m]} - \beta\hat{\gamma}_j^{[m]}| \geq \lambda\hat{p}_j(\beta))$$

  for $j \in \hat{S}$. $\lambda = c_*(\frac{\log(n)}{M})^{\frac{1}{|\hat{S}|}}$, c_* is chosen by the user.

- Then for the m-th sample we define the following CI for β:

$$\beta^{[m]}_{\min}(\lambda) = \min_{\beta\in\mathcal{B}:\left\|\hat{\pi}^{[m]}_{\hat{S}}(\beta,\lambda)\right\|_0 < \frac{|\hat{S}|}{2}} \beta.$$
$$\beta^{[m]}_{\max}(\lambda) = \max_{\beta\in\mathcal{B}:\left\|\hat{\pi}^{[m]}_{\hat{S}}(\beta,\lambda)\right\|_0 < \frac{|\hat{S}|}{2}} \beta.$$

- Then we set: $\text{CI}^{\text{samp}} = [\min_{m\in\mathcal{M}} \beta^{[m]}_{\min}(\lambda), \max_{m\in\mathcal{M}} \beta^{[m]}_{\max}(\lambda)]$ Here, $\mathcal{M} = \{1 \leq m \leq M : [\beta^{[m]}_{\min}(\lambda), \beta^{[m]}_{\max}(\lambda)] \neq \emptyset\}$

From simulations, one can see that many of the M sampled searching CIs are empty and that the non-empty ones are much shorter than the searching CI counterpart. As a result, the CI^{samp} is in general shorter than CI^{sear}. See section 2.8 for simulations.

If the value of λ is too small, very few of the resampled reduced-form estimators will pass the majority rule and most of the M (=1000, f.e.) will be empty (the theorem on which the thresholding relies states that that there exists an m^* for which the high-thresholding assumption is met under the assumption of $\beta = \beta^*$ and j being a valid IV. Hence, for small λ it is likely to happen that outside of m^* the intervals will be empty because the thresholding becomes more strict, see section 2.8 for more). As it is desirable for the method to be robust to errors, we wouldn't want to rely on too few intervals. Hence, the proportion of the non-empty intervals indicates whether λ is large enough. One could start with a small value of λ f.e. $\lambda = \frac{1}{6}\frac{\log(n)}{M}\frac{1}{2|\hat{S}|}$ and increase the value of λ by a factor of 1.25 until more than f.e. 10% of the M intervals are non-empty. One can then choose the smallest value of λ achieving this to implement the algorithm. More on this, see section 2.8.

Remark 2.5.8.

- Instead of choosing the largest possible estimated interval of the M sample CIs, we could also choose the union. However, this is not guaranteed to be an interval.
- It turns out that when we try to filter out boundary cases of the sampled $(\hat{\Gamma}^{[m]}, \hat{\gamma}^{[m]})$, the resulting CI will be nearly the same as CI^{samp} [2, p.14 Remark 4] (see section 2.8 for such an example).

2.6. Uniform inference methods under plurality rule

Under the plurality rule, in order to estimate β^*, we want to determine whether $\frac{\pi_k^*}{\gamma_k^*}$ and $\frac{\pi_j^*}{\gamma_j^*}$ have the same validity level (are equal to each other) from the data. For this, we need a level of separation based on the data i.e. if the difference $|\frac{\pi_k^*}{\gamma_k^*} - \frac{\pi_j^*}{\gamma_j^*}|$ is larger than that level, then for large n, with high probability we can detect from the data that they don't have the same invalidity level.

The separation level used for the methods, is defined below:

Definition 2.6.1. *The separation level is defined as:*

$$sep(n) = 2\sqrt{\log(n)} \max_{j,s \in \hat{S}} T_{j,k}$$

See notation 2.4.5 for the definition of $T_{j,k}$

The $sep(n)$ term can be interpreted as follows:

- Let $j, k \in S$. Then for $\beta^{[j]} (= \frac{\Gamma_j^*}{\gamma_j^*} = \beta^* + \frac{\pi_j^*}{\gamma_j^*})$ we have that: $\hat{\beta}^{[j]} = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$ is an consistent estimator for $\beta^{[j]}$ and consequently $\hat{\beta}^{[j]} - \hat{\beta}^{[k]}$ is a consistent estimator for $\frac{\pi_j^*}{\gamma_j^*} - \frac{\pi_k^*}{\gamma_k^*}$. Consider the following 2 oracle (i.e. not-computable from the data) consistent estimators for $|\frac{\pi_j^*}{\gamma_j^*} - \frac{\pi_k^*}{\gamma_k^*}|$:
  - Estimator 1: $|\frac{\hat{\gamma}_k}{\gamma_k^*}(\hat{\beta}^{[k]} - \hat{\beta}^{[j]})|$
  - Estimator 2: $|\frac{\hat{\gamma}_j}{\gamma_j^*}(\hat{\beta}^{[j]} - \hat{\beta}^{[k]})|$

  Both estimators take into account the rate of convergence for $\hat{\gamma}_k$ and $\hat{\gamma}_j$ respectively. From proposition 2.4.6, it can be observed that $\frac{\hat{\gamma}_k}{\gamma_k^*}(\hat{\beta}^{[k]} - \hat{\beta}^{[j]})$ has a standard error of $T_{j,k}^0$ and $\frac{\hat{\gamma}_j}{\gamma_j^*}(\hat{\beta}^{[j]} - \hat{\beta}^{[k]})$ of $T_{k,j}^0$. As $T_{j,k} = \min(T_{j,k}^0, T_{k,j}^0)$, $\max_{j,k \in \hat{S}} T_{j,k}$ can be seen as some sort of maximum standard error of the validity difference when using estimator 1 and estimator 2.
- The $2\sqrt{\log(n)}$ term is introduced to adjust for the multiplicity for testing multiple hypothesis in the coming tests. In case we test multiple hypothesis, we are increasing the chance that at least one of them is wrongly rejecting the null-hypothesis. The $2\sqrt{\log(n)}$ term tries to combat these errors by making the bounds less strict for especially small n. It doesn't have asymptotic implications due to the fact that $T_{j,k}$ is of order n^{-1} and we have that $\frac{\log(n)}{n} \to 0$ for $n \to \infty$.

Again, as with the inference using the majority rule, we have to make sure we can do that with enough data. Using the definition that $\mathcal{I}(v, \tau) := \{j \in S : |\frac{\pi_j^*}{\gamma_j^*} - v| \leq \tau\}$, the finite sample plurality rule is defined as follows:

Condition 2.6.2. *(Finite sample plurality rule)*
Under the finite sample plurality rule, the following holds for $\tau_n = 3sep(n)$*:*

$$|V \cap S_{\text{str}}| > \max_{v \in \mathbb{R}} |\mathcal{I}(v, \tau_n)/V|$$

Remark 2.6.3.

- $\mathcal{I}(v, \tau_n)/V$ *consists of the invalid IVs with invalidity level around* v*. As* $n \to \infty$*, the latter becomes exact.*
- *The condition requires that there are more valid and strongly relevant IVs then invalid IVs of approximately similar invalidity level*
- *Observe that the finite sample plurality rule is less restrictive then the requirement that:* $|V \cap S_{\text{str}}| \geq \max_{v \neq 0} |\mathcal{I}(v, \tau_n)|$.

Under condition 2.6.2, I will now establish a 2 step inference procedure for β^* as introduced in [2, p.14-17 5.Uniform Inference methods under plurality rule]. It has the following general idea:

- Step 1: Construct a set $\hat{V}$ that satisfies (with high probability):

$$V \cap S_{\text{str}} \subset \hat{V} \subset \mathcal{I}(0, \tau_n)$$

- Step 2: Restrict our attention to $\hat{V}$ and generalize the methods of the previous section.

The second step of the inference procedure will rely on the fact that (asymptotically), $V \cap S_{\text{str}}$ will become the majority of $\hat{V}$ if we manage to construct $\hat{V}$ as outlined in the first step. Using this fact, we can then apply the methods from the previous sections (which rely on a majority rule). The notion that $V \cap S_{\text{str}}$ becomes the majority of $\hat{V}$ is proven below. This was not proven in the original paper.

Lemma 2.6.4. *Under condition 2.6.2,* $V \cap S_{\text{str}}$ *will become the majority of* $\hat{V}$ *(defined as in the first step of the inference method above) for* n *large enough.*

Proof. Under condition 2.6.2, we have that:

$$|V \cap S_{\text{str}}| > \max_{v \in \mathbb{R}} |\mathcal{I}(v, \tau_n)/V| \geq |\mathcal{I}(0, \tau_n)/V|$$

As $V \subseteq \mathcal{I}(0, \tau_n)$ we hence have that:

$$|V \cap S_{\text{str}}| > |\mathcal{I}(0, \tau_n)/V| = |\mathcal{I}(0, \tau_n)| - |V|$$

So that:

$$|V \cap S_{\text{str}}| + |V| > |\mathcal{I}(0, \tau_n)|$$

Hence, either $|V \cap S_{\text{str}}| > \frac{|\mathcal{I}(0,\tau_n)|}{2}$ or $|V| > \frac{|\mathcal{I}(0,\tau_n)|}{2}$. In either case: $|V| > \frac{|\mathcal{I}(0,\tau_n)|}{2}$. As $|V \cap S_{\text{str}}| \to |V|$ for $n \to \infty$, for n large enough $V \cap S_{\text{str}}$ will become the majority of $\mathcal{I}(0, \tau_n)$ consequently, by the definition of $\hat{V}$, it will become the majority of $\hat{V}$ □

Now it will be shown how the $\hat{V}$ is constructed:
Constructing $\hat{V}$:

- Set wlog $\hat{S} = \{1, 2, .., |\hat{S}|\}$. For any $j, k \in \hat{S}$ the following estimators for β^* and π^* are defined:

$$\hat{\beta}^{[j]} = \hat{\Gamma}_j / \hat{\gamma}_j, \hat{\pi}_k^{[j]} = \hat{\Gamma}_k - \hat{\beta}^{[j]} \hat{\gamma}_k$$

- $\hat{R}^{[j]} = \hat{V}^{\Gamma} + (\hat{\beta}^{[j]})^2 \hat{V}^{\gamma} - 2\hat{\beta}^{[j]} \hat{C}$ is defined and the standard error of $\hat{\pi}_k^{[j]}$ is estimated by:

$$\hat{\text{SE}}(\hat{\pi}_k^{[j]}) = [(\hat{R}_{k,k}^{[j]} + (\hat{\gamma}_k / \hat{\gamma}_j)^2 \hat{R}_{j,j}^{[j]} - 2(\hat{\gamma}_k / \hat{\gamma}_j) \hat{R}_{j,j}^{[j]})/n]^{0.5}$$

- For $1 \leq k, j \leq |\hat{S}|$, hard thresholding is applied and the corresponding (k, j)-th entry for the voting matrix $\hat{\Pi} \in \mathbb{R}^{|\hat{S}| \times |\hat{S}|}$ is defined as:

$$\hat{\Pi}_{k,j} = 1(|\hat{\pi}_k^{[j]}| \leq \hat{\text{SE}}(\hat{\pi}_k^{[j]}) \sqrt{\log(n)}, |\hat{\pi}_j^{[k]}| \leq \hat{\text{SE}}(\hat{\pi}_j^{[k]}) \sqrt{\log(n)})$$

- Define $\hat{W} = \arg \max_{1 \leq j \leq |\hat{S}|} \left\| \hat{\Pi}_{j.} \right\|_0$ i.e. the set of IVs to support the most number of IVs to be valid. Construct the following initial set:

$$\hat{V}^{\text{TSHT}} = \{1 \leq l \leq |\hat{S}| : \exists k \in \hat{S} \text{ and } j \in \hat{W} \text{ s.t. } \hat{\Pi}_{j,k} \hat{\Pi}_{k,l} = 1\}$$

Remark 2.6.5. *In case $\pi_k^* = 0$ we expect (with high probability) that: $|\hat{\pi}_k^{[j]}| = |\hat{\pi}_k^{[j]} - \pi_k^*| \leq \hat{\text{SE}}(\hat{\pi}_k^{[j]})$*

To get a feeling on how the proposed voting-matrix works, consider the case that the used estimators in the voting matrix are perfect i.e. $\hat{\pi}_k^{[j]} = \pi_k^* - \gamma_k^* \frac{\pi_j^*}{\gamma_j^*}$ and $\hat{\text{SE}}(\hat{\pi}_k^{[j]}) = \text{SE}(\hat{\pi}_k^{[j]}) = |\gamma_k^*| T_{j,k}^0$. It then holds that:

$$|\hat{\pi}_k^{[j]}| \leq \sqrt{\log(n)} \hat{\text{SE}} \iff |\frac{\pi_k^*}{\gamma_k^*} - \frac{\pi_j^*}{\gamma_j^*}| \leq T_{j,k}^0 \sqrt{\log(n)}$$

From here, we see that $\hat{\Pi}_{k,j} = 1 \iff |\frac{\pi_k^*}{\gamma_k^*} - \frac{\pi_j^*}{\gamma_j^*}| \leq T_{j,k} \sqrt{\log(n)}$. For this case, we see that if the k-th and j-th IVs have the same invalidity levels, they will vote for each other. However, it might also happen that a valid IV votes for a locally invalid IV. These observations also hold true asymptotically for the general case, see Proposition 2.7.17.

After computing the voting matrix, $\hat{W}$ is computed. Under the finite plurality rule, we know that:

$$|V \cap S_{\text{str}}| > \max_{v \in \mathbb{R}} |\mathcal{I}(v, \tau_n) \backslash V|$$

As $V \cap S_{\text{str}} \subseteq \mathcal{I}(0, \tau_n)$ (with high probability), we obtain that:

$$|\mathcal{I}(0, \tau_n)| > \max_{v \in \mathbb{R}} |\mathcal{I}(v, \tau_n) \backslash V|$$

This motivates the definition of $\hat{W}$: for $n \to \infty$ it holds that $\tau_n \to 0$ and we see that the number of locally invalid IVs (including valid IVs) is greater than the number of invalid IVs around the same level, no matter the invalidity level.

To see how the $\hat{V}^{\text{TSHT}}$ set works together with the finite sample plurality rule, consider a model with $p_Z = 6$ where $\{1,2,3\} \in V \cap S_{\text{str}}$, $\{4,5\}$ have the same locally invalid invalidity level v and $\{6\}$ with a much bigger invalidity level compared to $\{1,2,3,4,5\}$. For this model, the population plurality rule is satisfied. For n large enough, the finite sample plurality rule will also be satisfied as for n large enough 4,5 will sit in the same set $\mathcal{I}(v, \tau_n)$ separated from 6. Due to the locally invalid nature of 4 and 5, it could be that (a subset of) $\{1,2,3\}$ are thus also in this $\mathcal{I}(v, \tau_n)$. It would hence require more data points for $|V \cap S_{\text{str}}| > \max_{v \neq 0} |\mathcal{I}(v, \tau_n)|$ to be satisfied.

Table 2.1 shows what a resulting $\hat{\Pi}$ could look like. We would in this case end up with $\hat{W} = \{4\}$: a locally invalid IV! The second step firstly includes all IVs that $Z^{(4)}$ voted for, so we have that: $\{2,3,5\} \subseteq \hat{V}^{\text{TSHT}}$. It then also includes all $Z^{(i)\prime s}$ $Z^{(2)}, Z^{(3)}$ and $Z^{(5)}$ voted for, so we end up with $\hat{V}^{\text{TSHT}} = \{1,2,3,4,5\}$. In the end, all valid and locally invalid IVs have been selected. Observe that the majority rule indeed applies to $\hat{V}^{\text{TSHT}}$. See section 2.8 for simulated examples of the choice of $\hat{V}^{\text{TSHT}}$ for different models.

$\hat{\Pi}$ example	1	2	3	4	5	6
1	✓	✓	✓			
2	✓	✓		✓		
3	✓		✓	✓		
4		✓	✓	✓	✓	
5				✓	✓	
6						✓

Table 2.1.: An example of the voting matrix based on the model described above. A ✓ in position (i, j) means that $Z^{(i)}$ and $Z^{(j)}$ voted for each other

2.7. Theoretical Justification

In this section, we will explore the theorems and some proofs regarding the searching and sampling methods. We will first take a look into the assumptions made on the models for the proofs and what kind of properties they establish (needed for the proofs). After that, the formal (asymptotic) mathematical statements on the methods are explored and the implications they have.

Hence, first the assumptions on the models are considered. In total two types of assumptions are made: one regarding sub-Gaussian vectors and one regarding the covariance matrix of random variables in the model. I will first give a brief introduction to sub-Gaussian vectors to give a feeling as to what they entail.

2.7.1. Sub-Gaussian vectors

A sub-Gaussian assumption on random vectors ensure that tail probabilities and moments are bounded in a controlled manner. Below, the definition of a sub-Gaussian random variable to given:

Definition 2.7.1. *A random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called sub-Gaussian if there exists a constant $K_1 > 0$ s.t. for all $t \geq 0$:*

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/K_1^2)$$

It turns out that requiring a random variable to be sub-Gaussian is equivalent with putting constraints on its moments. The following lemma provides a link between the definition of sub-Gaussian and moments:

Lemma 2.7.2. *[3, p.8 Exercise 1.2.3] For a random variable $Y \geq 0$, it holds that $\forall p \geq 1$:*

$$\mathbb{E}(Y^p) = \int_0^\infty pt^{p-1}\mathbb{P}(Y \geq t)dt$$

Proof. For any $w \in \Omega$, we have that: $Y^p(w) = \int_0^{Y(w)} pt^{p-1}dt$. Hence it holds that:

$$\mathbb{E}(Y^p) = \mathbb{E}(\int_0^{Y(w)} pt^{p-1}dt) = \int_\Omega \int_0^{Y(w)} pt^{p-1}dtd\mathbb{P} =: (1)$$

The domain of integration of (1) can be written as:

$$\{(w,t) : w \in \Omega, 0 \leq t \leq Y(w)\} = \{(w,t) : t \geq 0, t \leq Y(w)\}$$

Hence by Tonelli's theorem:

$$\begin{aligned}(1) &= \int_0^\infty \int_{Y(w)\geq t} pt^{p-1}d\mathbb{P}dt \\ &= \int_0^\infty pt^{p-1} \int_{Y(w)\geq t} 1d\mathbb{P}dt = \int_0^\infty pt^{p-1}\mathbb{P}(Y \geq t)dt\end{aligned}$$

□

Now, we can prove the connection to the moments. This proof is a slightly adapted version compared the one found at [3, p.24 Prop. 2.5.2].

Theorem 2.7.3. *[3, p.24 Prop. 2.5.2] For a random variable X, the following are equivalent:*

1. $\exists K_1 \geq 0$: $\forall t \geq 0$: $\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/K_1^2)$
2. $\exists K_2 \geq 0$: $\forall p \geq 1$: $\|X\|_p \leq K_2\sqrt{p}$
3. $\exists K_3 > 0 : \mathbb{E}(\exp(X^2/K_3^2)) \leq 2$

Proof. $(1 \implies 2)$: Note that if we prove for the random variable $Y := \frac{X}{K_1}$ that $\|Y\|_p \leq K_2' \sqrt{p}$ for some $K_2' > 0$, then: $\|X\|_p \leq K_1 K_2' \sqrt{p}$. Hence, wlog assume that $K_1 = 1$. Then we have that:

$$\begin{aligned}
\|X\|_p^p &= \mathbb{E}(|X|^p) \\
&= \int_0^\infty p t^{p-1} \mathbb{P}(|Y| \geq t) dt && \text{(Lemma 2.7.2)} \\
&\leq \int_0^\infty p t^{p-1} 2 \exp(-t^2) dt && (X \text{ is sub-Gaussian, } K_1 = 1) \\
&= p\Gamma(p/2) && (\Gamma(x) := \int_0^\infty t^{x-1} \exp(-x) dt) \\
&\leq 3p(p/2)^{p/2} && (\text{Using that } \Gamma(x) \leq 3x^x \text{ for } x \geq \frac{1}{2})
\end{aligned}$$

Hence we end up with:

$$\|X\|_p \leq 3^{1/p} p^{1/p} p^{1/2} 2^{-1/2} \leq \frac{3^2}{\sqrt{2}} \sqrt{p}$$

In the last step it was used that $3^{1/p} \leq 3$ and $p^{1/p} \leq 3$.
$(2 \implies 3)$ As before, wlog $K_2 = 1$.

$$\begin{aligned}
\mathbb{E}(\exp(X^2/K_4^2)) &\overset{\text{MCT}}{=} 1 + \sum_{p=1}^\infty \mathbb{E}((\frac{X^2}{K_4^2})^p / p!) \\
&\leq 1 + \sum_{p=1}^\infty \frac{(2p)^{p/2}}{K_4^{2p} p!} && \text{(Using assumption (2))} \\
&\leq 1 + \sum_{p=1}^\infty \frac{2^p}{K_4^{2p}} (\frac{p^p}{p!}) \\
&\leq 1 + \sum_{p=1}^\infty (\frac{2e}{K_4})^p = \frac{1}{1 - \frac{2e}{K_4}} \leq 2 && (\text{Using that } \frac{p^p}{p!} \leq \exp(p))
\end{aligned}$$

In the last step holds provided that K_4 is chosen in a proper manner.
$(3 \implies 1)$ $\mathbb{P}(|X| \geq t) = \mathbb{P}(\exp(X^2/K_3^2)) \geq \exp(t^2/K_3)$. Now apply Markov's inequality. □

Now that we have some properties established for a sub-Gaussian random variable, we can define a sub-Gaussian random vector:

Definition 2.7.4. *A random vector $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^d, \mathcal{B}^d(\mathbb{R}^d), \lambda)$ is called sub-Gaussian if $\forall x \in \mathbb{R}^d : \langle x, Y \rangle = x^T Y$ is a sub-Gaussian random variable.*

It turns out that one can always create sub-Gaussian vectors out of sub-Gaussian random variables. I formulated and proved the following result myself:

Corollary 2.7.5. *Let $\epsilon_1, ..., \epsilon_n$ be sub-Gaussian random variables. Then $(\epsilon_1, ..., \epsilon_n)^T$ is a sub-Gaussian vector.*

Proof. Let $a = (a_1, ..., a_n)^T \in \mathbb{R}^n$ and let $p \geq 1$. Observe that any ϵ_i is an element of $L^p(\Omega, \mathcal{F}, \mathbb{P})$ by theorem 2.7.3. Then it holds that (using theorem 2.7.3 (2)):

$$\left\| \sum_{i=1}^{n} a_i \epsilon_i \right\|_p \leq \sum_{i=1}^{n} |a_i| \, \|\epsilon_i\|_p \leq (\sum_{i=1}^{n} |a_i| K_2^{(i)}) \sqrt{p} =: K_2' \sqrt{p}$$

Hence by theorem 2.7.3 and the definition of a sub-Gaussian vector the result is proven. □

Example 2.7.6. *Some notable examples of sub-Gaussian random variables are normal distributed random variables and bounded random variables.*

2.7.2. Assumptions on model and usage

The following conditions are put on model 2.2.7 for the proofs by the original paper [2]:
Assume that $(W_i, \epsilon_i, \delta_i)^T$ are iid for $1 \leq i \leq n$ with $W_i := (X_i^T, Z_i^T)^T \in \mathbb{R}^p$

(C1) For $i \leq i \leq n$, W_i are sub-Gaussian random vectors with $\Sigma := \mathbb{E}(W_i W_i^T)$ satisfying $0 < c_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_0$.

(C2) For $1 \leq i \leq n$, the errors $(\epsilon_i, \delta_i)^T$ from model 2.2.7 are sub-Gaussian random vectors satisfying:
$0 < c_1 \leq \lambda_{\min}(\mathbb{E}((\epsilon_i, \delta_i)^T(\epsilon_i, \delta_i)|W_i) \leq \lambda_{\max}(\mathbb{E}((\epsilon_i, \delta_i)^T(\epsilon_i, \delta_i)|W_i) \leq C_1$

Here, c_0, C_0, c_1 and C_1 are constants.

Remark 2.7.7.

- *Note that (C1) implies that Σ is positive definite hence invertible. Without this $(\hat{\Gamma}, \hat{\gamma})^T$ as defined in definition 2.4.2 wouldn't well-conditioned.*
- *As $e_i = \epsilon_i - \beta^* \delta_i$, condition (C2) implies that e_i is sub-Gaussian as well.*

(C2) implies that Cov^A (see proof of theorem 2.4.3, it is the asymptotic variance of $(\sqrt{n}((\hat{\Gamma}^T, \hat{\Psi}^T, \hat{\gamma}^T, \hat{\psi}^T)^T - (\Gamma^T, \Psi^T, \gamma^T, \psi^T)^T))$ is positive definite. Before I go about proving that statement, I will first prove a lemma which will be used during the proof. I proved the following lemma myself.

Lemma 2.7.8. *For a real symmetric matrix $S \in \mathbb{R}^{k \times k}$ with orthogonal decomposition $S = PDP^T$, where P has k columns $p_1, ..., p_k$, $D = \mathrm{diag}(\lambda_1, ..., \lambda_k)$ it holds that for any $x \in \mathbb{R}^k$:*

$$\lambda_{\min}(S) \, \|x\|_2^2 \leq x^T S x \leq \lambda_{\max} \, \|x\|_2^2$$

Proof.

$$\begin{aligned} x^T S x &= x^T P D P^T x \\ &= \sum_{i=1}^{k} \lambda_k x^T p_i p_i^T x \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{k} \lambda_i \left\| p_i^T x \right\|^2 \\
&\geq \lambda_{\min}(S) \sum_{i=1}^{k} x^T p_i p_i^T x \\
&= \lambda_{\min}(S) x^T P P^T x \\
&= \lambda_{\min}(S) \left\| P^T x \right\|^2 \\
&= \lambda_{\min} \left\| x \right\|^2
\end{aligned}
$$

In the last step it's used that P is orthogonal. The upper-bound follows analogously. □

Using the lemma above we can prove positive definiteness for Cov^A. This following lemma was stated at [2, p.9 C.Proofs] and I filled in the details of the corresponding proof. I also made slight adjustments to the argument.

Lemma 2.7.9. *Under (C1) and (C2),* Cov^A *is positive definite. In particular:*

$$
\begin{aligned}
\lambda_{\min}(\mathrm{Cov}^A) &\geq c_1 \lambda_{\min}(\Sigma^{-1}) \geq \frac{c_1}{C_0} \\
\lambda_{\max}(\mathrm{Cov}^A) &\leq C_1 \lambda_{\max}(\Sigma^{-1}) \leq \frac{C_1}{c_0}
\end{aligned}
$$

Proof. Recall that:

$$
\mathrm{Cov}^A = \begin{pmatrix} \Sigma^{-1}\,\mathbb{E}(\epsilon_i^2 W_i W_i^T)\Sigma^{-1} & \Sigma^{-1}\,\mathbb{E}(\epsilon_i \delta_i W_i W_i^T)\Sigma^{-1} \\ \Sigma^{-1}\,\mathbb{E}(\epsilon_i \delta_i W_i W_i^T)\Sigma^{-1} & \Sigma^{-1}\,\mathbb{E}(\delta_i^2 W_i W_i^T)\Sigma^{-1} \end{pmatrix}
$$

Let $u, v \in \mathbb{R}^p$, then $(u^T, v^T)\,\mathrm{Cov}^A (u^T, v^T)^T$ can be rewritten as:

$$
\begin{aligned}
u^T \Sigma^{-1}\,\mathbb{E}(\epsilon^2 W_i W_i^T)\Sigma^{-1} u + 2u^T \Sigma^{-1}\,\mathbb{E}(\epsilon_i \delta_i W_i W_i^T)\Sigma^{-1} v + v^T \Sigma^{-1}\,\mathbb{E}(\delta_i^2 W_i W_i^T)\Sigma^{-1} v = \\
\mathbb{E}((u^T \Sigma^{-1} W_i, v^T \Sigma^{-1} W_i)(\epsilon_i, \delta_i)^T (\epsilon_i, \delta_i)(u^T \Sigma^{-1} W_i, v^T \Sigma^{-1} W_i)^T) =: \\
\mathbb{E}((u_i, v_i)(\epsilon_i, \delta_i)^T (\epsilon_i, \delta_i)(u_i, v_i)^T) = \\
\mathbb{E}((u_i, v_i)\,\mathbb{E}((\epsilon_i, \delta_i)^T (\epsilon_i, \delta_i)|W_i)(u_i, v_i)^T) =: (1)
\end{aligned}
$$

Here, $u_i = W_i^T \Sigma^{-1} u, v_i = W_i^T \Sigma^{-1} v$(which are both W_i-measurable).
Since $\lambda_{\min}(\mathbb{E}((\epsilon_i, \delta_i)^T (\epsilon_i, \delta_i)|W_i)) \geq c_1 > 0$ and $\mathbb{E}(u_i^2) = u^T \Sigma^{-1} u$, $\mathbb{E}(v_i^2) = v^T \Sigma^{-1} v$, by (an extension of) lemma 2.7.8:

$$
\begin{aligned}
(1) &\geq \mathbb{E}(\lambda_{\min}((\epsilon_i, \delta_i)^T (\epsilon_i, \delta_i)|W_i)(u_i, v_i)(u_i, v_i)^T) \\
&\geq c_1 (u^T \Sigma^{-1} u + v^T \Sigma^{-1} v) \\
&\geq c_1 \lambda_{\min}(\Sigma^{-1})(\|(u, v)\|_2^2)
\end{aligned}
$$

In the last step, I applied lemma 2.7.8 again by observing that:

$$
u^T \Sigma^{-1} u + v^T \Sigma^{-1} v = (u^T v^T)^T \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & \Sigma^{-1} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}
$$

and that the matrix $\begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & \Sigma^{-1} \end{pmatrix}$ is symmetric and, when we define the orthogonal diagonalisation of Σ^{-1} as $\Sigma^{-1} = PDP^T$, has orthogonal diagonalisation:

$$\begin{pmatrix} P & 0 \\ 0 & P \end{pmatrix} \begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} P^T & 0 \\ 0 & P^T \end{pmatrix}$$

Thus we have established that:

$$(u^T, v^T)\,\mathrm{Cov}^A(u^T, v^T)^T \geq c_1 \lambda_{\min}(\Sigma^{-1})(\|(u,v)\|_2^2) \tag{2.7.1}$$

which proves positive definiteness.
Choosing $(u^T, v^T)^T$ in the eigenspace of $\lambda_{\min}(\mathrm{Cov}^A)$ with length 1 we obtain using (2.7.1):

$$\lambda_{\min}(\mathrm{Cov}^A) \geq c_1 \lambda_{\min}(\Sigma^{-1}) \geq c_1 / C_0 > 0.$$

The argument for $\lambda_{\max}(\mathrm{Cov}^A)$ goes analogous. □

I will now show how the sub-Gaussian assumption can be used to obtain finite sample results. Define the following sets ($\hat{\Omega} := \hat{\Sigma}^{-1}$ from notation 2.4.1):

$$\mathcal{G}_1 := \{\left\| \frac{1}{n} W^T \delta \right\|_\infty \leq C \frac{(\log(n))^{1/4}}{\sqrt{n}}\}$$
$$\mathcal{G}_3 := \{\left\| \hat{\Omega} - \Sigma^{-1} \right\|_2 \leq C \frac{\sqrt{\log(n)}}{\sqrt{n}}\}$$
$$\mathcal{G}_2' := \{\max_{1 \leq j \leq p_Z} \sqrt{n} \frac{|\hat{\gamma}_j - \gamma_j^*|}{\sqrt{V_{j,j}^\gamma}} \leq \tilde{C} (\log(n))^{1/4}\}$$

For convenience, I will use the notation $\Omega := \Sigma^{-1}$ Here, $C > 0$ (does not depend on n and is deterministic) and $\tilde{C}$ satisfies: $\tilde{C} > \max_{1 \leq j \leq p_Z} \frac{\|\Omega_{j.}\|_1}{\sqrt{V_{jj}^\gamma}} C$. (also deterministic and not dependent on n)
Using the sub-Gaussian assumptions (C1) and (C2), one can prove that for all $n \geq 1$ and some constant $c > 0$ [2, p.27,28 D.Proofs of Lemma's]:

$$\mathbb{P}(\mathcal{G}_1) \geq 1 - \exp(-c\sqrt{\log(n)}) \tag{2.7.2}$$
$$\mathbb{P}(\mathcal{G}_3) \geq 1 - n^{-c} \tag{2.7.3}$$

The next result was obtained in the original paper to prove a larger lemma [2, p.27,28 D.Proofs of Lemma's]. I adapted the proof by specifying $\tilde{C} > 0$ for the set $\mathcal{G}_2'$ above (namely satisfies: $\tilde{C} > \max_{1 \leq j \leq p_Z} \frac{\|\Omega_{j.}\|_1}{\sqrt{V_{jj}^\gamma}} C$.). In the original paper, this was unspecified. I filled in further details of the proof provided by [2].

Lemma 2.7.10. *Assume (C1) and (C2). Then* $\mathbb{P}(\mathcal{G}_2') \geq 1 - \exp(-c_1\sqrt{\log(n)})$ *for* n *large enough.* $c_1 > 0$ *is some constant.*

Proof. Observe that we can decompose $\hat{\gamma}_j$ as follows:

$$\hat{\gamma}_j = \hat{\Omega}_{j.} \frac{1}{n} W^T D$$

$$= \hat{\Omega}_{j.}\frac{1}{n}W^T(W\begin{pmatrix}\gamma^* \\ \psi^*\end{pmatrix} + \delta)$$
$$= \gamma_j^* + \hat{\Omega}_{j.}\frac{1}{n}W^T\delta$$

so that we have:

$$\hat{\gamma}_j - \gamma_j^* = \Omega_{j.}\frac{1}{n}W^T\delta + (\hat{\Omega}_{j.} - \Omega_{j.})\frac{1}{n}W^T\delta \tag{2.7.4}$$

This is used to establish the following:

$$\begin{aligned}
\mathbb{P}(\mathcal{G}_2') &\geq \mathbb{P}(\mathcal{G}_2' \cap \mathcal{G}_1 \cap \mathcal{G}_3) \\
&\geq \mathbb{P}(\{\forall j : \frac{|\Omega_{j.}\frac{1}{n}W^T\delta| + |(\hat{\Omega}_{j.} - \Omega_{j.})\frac{1}{n}W^T\delta|}{\sqrt{V_{j,j}^\gamma}} \leq \tilde{C}\frac{(\log(n))^{1/4}}{\sqrt{n}}\} \cap \mathcal{G}_1 \cap \mathcal{G}_3) \\
&\geq \mathbb{P}(\{\forall j : \frac{1}{\sqrt{V_{j,j}^\gamma}}(\|\Omega_{j.}\|_1 + \left\|\hat{\Omega}_{j.} - \Omega_{j.}\right\|_1)\left\|\frac{1}{n}W^T\delta\right\|_\infty \leq \tilde{C}\frac{(\log(n))^{1/4}}{\sqrt{n}}\} \cap \mathcal{G}_1 \cap \mathcal{G}_3) \\
&\geq \mathbb{P}(\{\forall j : \frac{1}{\sqrt{V_{j,j}^\gamma}}(\|\Omega_{j.}\|_1 + \left\|\hat{\Omega}_{j.} - \Omega_{j.}\right\|_1) \leq C'\} \cap \mathcal{G}_1 \cap \mathcal{G}_3) =: (1)
\end{aligned}$$

Here, $C' = \frac{\tilde{C}}{C}$ and the last step uses the fact that we are intersecting with $\mathcal{G}_1$.

$$\begin{aligned}
(1) &= \mathbb{P}(\{\forall j : \left\|\hat{\Omega}_{j.} - \Omega_{j.}\right\|_1 \leq (C'\sqrt{V_{j,j}^\gamma} - \|\Omega_{j.}\|_1)\} \cap \mathcal{G}_1 \cap \mathcal{G}_3) \\
&\geq \mathbb{P}(\{\forall j : \left\|\hat{\Omega}_{j.} - \Omega_{j.}\right\|_2 \leq (\sqrt{p})^{-1}(C'\sqrt{V_{j,j}^\gamma} - \|\Omega_{j.}\|_1)\} \cap \mathcal{G}_1 \cap \mathcal{G}_3) =: (2)
\end{aligned}$$

The last step uses that:

$$\left\|\hat{\Omega}_{j.} - \Omega_{j.}\right\|_1 \leq \sqrt{p}\left\|\hat{\Omega}_{j.} - \Omega_{j.}\right\|_2$$

which is an application of Cauchy-Schwarz.
Observe that (by the definition of C'): $C'\sqrt{V_{jj}^\gamma} - \|\Omega_{j.}\|_1 > 0$ and that this inequality does not depend on n. Furthermore, as $\hat{\Omega}$ and Ω are both symmetric we obtain that for any $j = 1, ..., p$:

$$\left\|\hat{\Omega}_{j.} - \Omega_{j.}\right\|_2 = \left\|\hat{\Omega}_{.j} - \Omega_{.j}\right\|_2 = \left\|(\hat{\Omega} - \Omega)e_j\right\|_2 \leq \left\|\hat{\Omega} - \Omega\right\|_2$$

Above, e_j denotes the j-th unit vector of $\mathbb{R}^p$. Hence, for n large enough, we obtain that:

$$\{\left\|\hat{\Omega} - \Omega\right\|_2 \leq C\sqrt{\frac{\log(n)}{n}}\} \subseteq \{\forall j : \left\|\hat{\Omega}_{j.} - \Omega_{j.}\right\|_2 \leq (\sqrt{p})^{-1}(C'\sqrt{V_{jj}^\gamma} - \|\Omega_{j.}\|_1)\}$$

And so, for n large enough:

$$\begin{aligned}
(2) &= \mathbb{P}(\mathcal{G}_1 \cap \mathcal{G}_3) \\
&= 1 - \mathbb{P}(\mathcal{G}_1^c \cup \mathcal{G}_3^c) \\
&\geq 1 - (\mathbb{P}(\mathcal{G}_1^c) + \mathbb{P}(\mathcal{G}_3^c)) \\
&= \mathbb{P}(\mathcal{G}_1) + \mathbb{P}(\mathcal{G}_3) - 1 \\
&\geq 1 - n^{-c} - \exp(-c\sqrt{\log(n)})
\end{aligned}$$

The last step uses the results established at (2.7.2).
As $\exp(-c\sqrt{\log(n)}) \geq \exp(-c\log(n)) = n^{-c}$, there exists a constant c_1 that satisfies: $0 < c_1 < c$ and $n^{-c} + \exp(-c\sqrt{\log(n)}) \leq \exp(-c_1\sqrt{\log(n)})$.
In conclusion: $\mathbb{P}(\mathcal{G}_2') \geq 1 - \exp(-c_1\sqrt{\log(n)})$ for n large enough. □

2.7.3. Relaxation of sub-Gaussian assumptions

This section is my contribution to the original paper as found at [2].

We already know, without any sub-Gaussian assumptions on model 2.2.7 that:
$(\frac{1}{n}W^TW)^{-1} \xrightarrow{\mathbb{P}} \Sigma^{-1}$ by the law of large numbers (if we assume that $\Sigma < \infty$ and invertibility) and consequently $\left\|(\frac{1}{n}W^TW)^{-1} - \Sigma^{-1}\right\|_2 \xrightarrow{\mathbb{P}} 0$. With the sub-Gaussian assumption, we furthermore know that $(\hat{\Omega} := (\frac{1}{n}W^TW)^{-1})$ [2, p.28 D.1.Proof of lemma 3] [4, p.25 Remark 5.40 (5.25)]

$$\mathbb{P}(\left\|\hat{\Omega} - \Sigma^{-1}\right\|_2 \leq C\sqrt{\frac{\log(n)}{n}}) \geq 1 - n^{-c}$$

with C and c positive constants (independent of n). But why is this necessary to know if we only consider asymptotic results $n \to \infty$ for the theorems? After all, a consequence of convergence in probability is that:

$$\mathbb{P}(\left\|\hat{\Omega} - \Sigma^{-1}\right\|_2 \leq C\sqrt{\frac{\log(n)}{n}}) \to 1$$

In my view, sub-Gaussian assumptions are not necessary for the asymptotic results if we are willing to assume some consistency properties. As seen in section 2.4.2, it does not directly follow from the law of large numbers that for example $\hat{V}^\gamma \xrightarrow{\mathbb{P}} V^\gamma$. This will follow as a consequence of assuming (C1) and (C2) (this will be shown below). Lemma 2.5.2 and lemma 2.5.5 are examples of how asymptotic properties will work out if just assume consistency rather than (the stronger) sub-Gaussian properties. A potential benefit of assuming sub-Gaussian properties, rather than only consistency properties, is that with the sub-Gaussian assumptions, we can establish data-dependent lower bounds for the probabilities for finite n. For example: $\hat{\Omega} \xrightarrow{\mathbb{P}} \Sigma^{-1}$ does not give clarity on from which n forwards $\hat{\Omega}$ is a good estimator for Σ^{-1}. On the other hand, $\mathbb{P}(\left\|\hat{\Omega} - \Sigma^{-1}\right\|_2 \leq C\sqrt{\frac{\log(n)}{n}}) \geq 1 - n^{-c}$, we can choose the n for which we are 'close enough'.

In the original paper, it was conjectured that one might be able to relax the sub-Gaussian assumptions to moment assumptions on $(W_i, \epsilon_i, \delta_i)^T$ to prove its main theorems [2, p.18, above theorem 1]. It is indeed the case that by assuming all the moment conditions of theorem 2.4.4, we obtain consistency which in turn gives us the main theoretical results. I will now show that the sub-Gaussian conditions are indeed stronger than the moment conditions:

Theorem 2.7.11. *If (C1) and (C2) hold, then all the conditions for consistency (and asymptotic normality) as seen in the statement of theorem 2.4.4 also hold.*

Proof. I will show that $\mathbb{E}(\epsilon_i Z_i^T W_i^{(k)} W_i^{(l)})$ is finite for $k, l = 1, ..., p$. All the others follow by a similar argument:
Take $a \in \{1, ..., p_Z\}$. We know that $|\mathbb{E}(\epsilon_i Z_i^{(a)} W_i^{(k)} W_i^{(l)})| \leq \mathbb{E}(|\epsilon_i Z_i^{(a)} W_i^{(k)} W_i^{(l)}|) =: (1)$. By Hölder's (generalised) inequality, it holds that:

$$(1) \leq \|\epsilon_i\|_4 \left\|Z_i^{(a)}\right\|_4 \left\|W_i^{(k)}\right\|_4 \left\|W_i^{(l)}\right\|_4 \leq C2^4$$

for some $C > 0$. The last step follows from theorem 2.7.3 and corollary 2.7.5 . In conclusion: $\mathbb{E}(\epsilon_i Z_i^T W_i^{(k)} W_i^{(l)})$ is finite. □

As a consequence of the above lemma, in case we assume that $(W_i, \epsilon_i, \delta_i)$ all have finite 4-th moments (coordinate-wise): all the conditions of theorem 2.4.4 will hold.

In summary, once we prove the theorems provided in the text assuming consistency properties, rather than the sub-Gaussian assumptions, the theorems will automatically hold for the sub-Gaussian cases in my view. These consistency assumptions are sufficient for the theoretical results to hold and will also reduce the length of proofs (for example consider the proof done at [2, D.1 Proof of Lemma 3 p.28] which first establishes the data-dependent lower bounds before doing $n \to \infty$ to establish the asymptotic properties for the set $\mathcal{G}_3$ defined on page 10 of [2] in Lemma 2).

2.7.4. Asymptotic justification methods

In this section, I will consider the formal asymptotic results for searching and sampling methods previously described.

I start with the searching method:

Theorem 2.7.12. *[2, p.18 6.Theoretical Justification Theorem 1] Consider the reduced form model 2.2.7. Suppose that the finite sample majority rule (condition 2.6.2), (C1) and (C2) all hold. Let $\alpha \in (0, 1/4)$, then CI^{sear} and $\hat{\text{CI}}^{\text{sear}}$ as seen at method 2.5.3 and subsection 2.5.3 satisfy:*

$$\liminf_{n\to\infty} \mathbb{P}(\beta^* \in \text{CI}^{\text{sear}}) \geq 1-\alpha \text{ and } \liminf_{n\to\infty} \mathbb{P}(\beta^* \in \hat{\text{CI}}^{\text{sear}}) \geq 1-\alpha$$

Also, there exists a positive, deterministic $C > 0$ that satisfies:

$$\liminf_{n\to\infty} \mathbb{P}(\max(\text{L}(\text{CI}^{\text{sear}}), \text{L}(\hat{\text{CI}}^{\text{sear}})) \leq \frac{C}{\min\limits_{j\in\hat{S}\cap V} |\gamma_j^*|\sqrt{n}}) \geq 1-\alpha$$

Here, $\text{L}(.)$ denotes the length of a set (largest element - smallest element)

Remark 2.7.13.

- *Theorem 2.7.12 does not rely on being able to perfectly separate valid and invalid IVs but instead relies on the (finite sample) majority rule. This contrasts previous work in this area [2, p.6 Section 3].*
- *In case $\mathbb{P}(\min\limits_{j\in\hat{S}\cap V} |\gamma_j^*| \geq K) \to 1$ for some $K > 0$, we have that the length of the searching CI is of the parametric length $1/\sqrt{n}$. As $\mathbb{P}(\hat{S} \subseteq S) \to 1$ under the same conditions as theorem 2.7.12, this will indeed be the case for $K = \min\limits_{j\in S\cap V} |\gamma_j^*|$.*

For the sampling CI, its asymptotic results hold for a decreased threshold level compared to what's introduced in section 2.5.4. I will first introduce this decreased threshold:
For $\alpha_0 \in (0, 1/4)$, define:

$$c^*(\alpha_0) = \frac{1}{[3\pi\lambda_{\min}(\text{Cov})]^{|\hat{S}|}} \exp(-\frac{|\hat{S}|3\lambda_{\max}(\text{Cov})}{\lambda_{\min}(\text{Cov})}[\Phi^{-1}(1-\frac{\alpha_0}{4|\hat{S}|})]^2)$$

$$= \exp(-\frac{|\hat{S}|3\lambda_{\max}(\text{Cov})}{\lambda_{\min}(\text{Cov})}[\Phi^{-1}(1-\frac{\alpha_0}{4|\hat{S}|})]^2 - |\hat{S}|\log(3\pi\lambda_{\min}(\text{Cov})))$$

See theorem 2.4.3 for the definition of Cov above. Note that $c^*(\alpha_0)$ does not directly depend on n. It only depends on n through $\hat{S}$. $c^*(\alpha_0)$ also depends in a non-trivial way on $|\hat{S}|$ due to the Φ^{-1} term which doesn't have an analytical expression. If $|\hat{S}|$ is large, $c^*(\alpha_0)$ will be small. I will now introduce the theoretical threshold term for the testing of $\pi_j^* = 0$ in the sampling case:

$$\text{err}_n(M, \alpha_0) := [\frac{2\log(n)}{c^*(\alpha_0)M}]^{\frac{1}{2|\hat{S}|}}.$$

Here, it can be noted that $\text{err}_n(M, \alpha_0) \to 0$ in case $M \to \infty$ and M goes faster to ∞ then $\log(n)$.

Proposition 2.7.14. *[2, p.19 6.Theoretical Justification Proposition 1][2, p.13,17 C.3. Proof of Prop 1] Suppose (C1) and (C2) hold and $\alpha_0 \in (0, 1/4)$. Then there exists a deterministic $C > 0$ for which if we have that:*

$$\text{err}_n(M, \alpha_0) < \min\{0.1 \min_{j\in\hat{S}}\{\hat{V}_{j,j}^{\Gamma}, \hat{V}_{j,j}^{\gamma}\}\Phi^{-1}(1-\frac{\alpha_0}{4|\hat{S}|}), \frac{1}{2}\frac{c^*(\alpha_0)}{C\sqrt{2|\hat{S}|}}\}$$

then:

$$\liminf_{n\to\infty} \mathbb{P}(\min_{1\le m\le M}\max_{\beta\in\mathcal{U}(a)}\max_{j\in\hat{S}}\frac{|\hat{\Gamma}_j^{[m]} - \Gamma_j^* - \beta(\hat{\gamma}_j^{[m]} - \gamma_j^*)|}{\sqrt{(\hat{V}_{j,j}^{\Gamma} + \beta^2\hat{V}_{j,j}^{\gamma} - 2\beta\hat{C}_{j,j})/n}} \le C\text{err}_n(M, \alpha_0)) \ge 1 - \alpha_0$$

with $\mathcal{U}(a) = \{\beta \in \mathbb{R} : |\beta - \beta^*| \le n^{-a}\}$ for any $a > \frac{1}{2}$.

Observe that, when fixing n, $\text{err}_n(M, \alpha_0) \to 0$ for $M \to \infty$ and the required upper-bound for it in proposition 2.7.14 does not depend on M. Hence, it holds that (when fixing n):

$$\lim_{M\to\infty} \mathbb{P}(\text{err}_n(M, \alpha_0) < \min\{0.1 \min_{j\in\hat{S}}\{V_{j,j}^{\Gamma}, V_{j,j}^{\gamma}\}\Phi^{-1}(1-\frac{\alpha_0}{4|\hat{S}|}), \frac{1}{2}\frac{c^*(\alpha_0)}{C\sqrt{2|\hat{S}|}})\} = 1$$

And so, translating the statement proposition 2.7.14 for large n and M (where M is much larger than $\log(n)$) gives that, with high probability, there exists an m between 1 and M for which:

$$\max_{\beta\in\mathcal{U}(a)}\max_{j\in\hat{S}}\frac{|\hat{\Gamma}_j^{[m]} - \Gamma_j^* - \beta(\hat{\gamma}_j^{[m]} - \gamma_j^*)|}{\sqrt{(\hat{V}_{j,j}^{\Gamma} + \beta^2\hat{V}_{j,j}^{\gamma} - 2\beta\hat{C}_{j,j})/n}} \le C\text{err}_n(M, \alpha_0)$$

This means that for any $\beta \in \mathcal{U}(a)$ (which includes β^*), we have a high chance that it ends in CI^{samp}.

The following theorem specifies the asymptotic probability that $\beta^* \in \text{CI}^{\text{samp}}$

Theorem 2.7.15. *[2, p.19 6.Theoretical Justification Theorem 2] Suppose that the conditions of proposition 2.7.14 hold,* $\alpha_0 \in (0, 1/4)$ *and* λ *as seen in subsection 2.5.4 satisfies:* $\lambda \geq 2C\text{err}_n(M, \alpha_0)/\Phi^{-1}(1 - \frac{\alpha}{2|\hat{S}|})$ *and* $\lambda \gg n^{\frac{1}{2}-a}$ *with* $a > \frac{1}{2}$*. Then* CI^{samp} *satisfies:*

$$\liminf_{n\to\infty} \mathbb{P}(\beta^* \in \text{CI}^{\text{samp}}) \geq 1 - \alpha_0$$

Furthermore there exists a deterministic $C > 0$ *that satisfies:*

$$\liminf_{n\to\infty} \mathbb{P}\left(\frac{\text{L}(\text{CI}^{\text{samp}})}{\sqrt{\log(|\mathcal{M}|)}} \leq \frac{C}{\min\limits_{j\in\hat{S}\cap V} |\gamma_j^*|\sqrt{n}}\right) \geq 1 - \alpha_0$$

Here, $\mathcal{M}$ *is defined as in section 2.5.4.*

Remark 2.7.16.

- *In the sampling method as seen in subsection 2.5.4* λ *of the form* $c_*(\frac{\log(n)}{M})^{\frac{1}{|\hat{S}|}}$ *is used. The requirement of theorem 2.7.15 can be written as* $\lambda \geq \hat{c}(\frac{\log(n)}{M})^{\frac{1}{2|\hat{S}|}}$ *with* $\hat{c} = [\frac{2}{c^*(\alpha_0)}]^{\frac{1}{2|\hat{S}|}}$
- *From theorem 2.7.15 one can not observe that the interval* CI^{samp} *is shorter than* CI^{sear} *(with high probability for a fixed, large enough* n*) due to the* $\log(|\mathcal{M}|)$ *term. However, simulation studies show that it is shorter most of the time, see section 2.8.*

The next two results relate to the β^* identification through the plurality rule. First the effect of the voting matrix and constructed $\hat{V}$ is displayed:

Proposition 2.7.17. *[2, p.20 6.Theoretical Justification Proposition 2] Suppose that (C1), (C2) and the finite sample plurality rule (condition 2.6.2) hold. Let* $j, k \in \hat{S}$*. Then:*

(a) If $\frac{\pi_k^*}{\gamma_k^*} = \frac{\pi_j^*}{\gamma_j^*}$*, then* $\liminf\limits_{n\to\infty} \mathbb{P}(\hat{\Pi}_{k,j} = \hat{\Pi}_{j,k} = 1) = 1$.

(b) If $|\frac{\pi_k^*}{\gamma_k^*} - \frac{\pi_j^*}{\gamma_j^*}| \geq 2\sqrt{\log(n)}T_{j,k}$*, then* $\liminf\limits_{n\to\infty} \mathbb{P}(\hat{\Pi}_{k,j} = \hat{\Pi}_{j,k} = 0) = 1$.

In addition, $\hat{V}$ *satisfies:*

$$\liminf_{n\to\infty} \mathbb{P}(V \cap S_{\text{str}} \subseteq \hat{V} \subseteq \mathcal{I}(0, 3\,\text{sep}(n)) = 1$$

Hence, if k and j have the same invalidity level, they are likely to vote for each other for large n. In case their invalidity level is large enough, they will vote against each other. For $0 < |\frac{\pi_k^*}{\gamma_k^*} - \frac{\pi_j^*}{\gamma_j^*}| < 2\sqrt{\log(n)}T_{j,k}$ (f.e. if k is valid while j is locally invalid) we have no guarantees. Thus, $\hat{V}$ will likely contain some locally invalid IVs as well (even for large n).

The last result confirms that if we apply the methods designed for the majority rule to $\hat{V}$, we will get the same asymptotic results:

Theorem 2.7.18. *[2, p.20 6.Theoretical Justification Theorem 3] Assume the finite sample plurality rule (instead of the finite sample majority rule). Then replacing* $\hat{S}$ *by* $\hat{V}$ *will yield the same previously established results for* CI^{sear} *as for* CI^{samp}*.*

2.8. Simulation studies

For the identification of β^* under the finite sample plurality rule we have four steps: first construct $\hat{S}$, then construct $\hat{W}$, construct $\hat{V}$ and then apply the searching or sampling method to $\hat{V}$ (see section 2.5 and section 2.6). The argument for using these steps is that for n large enough the valid instrumental variables will vote for each other and so if the finite sample plurality rule is satisfied the valid instrumental variables will be contained in $\hat{W}$. Then $V \cap S_{\text{str}}$ will be the majority of $\hat{V}$, hence making applying the searching or sampling method from section 2.5 a valid strategy to identify β^*. In this section I will analyse how different models perform for different n's and what type of models the different methods will struggle with (i.e. which models need a large number of data point to identify β^* reliably with a small interval length and the finite sample plurality rule satisfied?). I will analyse this by considering models (who will all satisfy the plurality rule) with varying invalidity levels and see how they perform for different n's in comparison and why, if applicable, the method struggles with particular models.

2.8.1. Set-up

I will use the same data distributions as seen in the original paper [2, p.20 7.Simulation Studies]. For $1 \leq i \leq n$, for each model in the next section we will have that the outcome model and association model (see definition 2.2.1 and definition 2.2.2) satisfy: $(X_i^T, Z_i^T)^T =: W_i \overset{\text{iid}}{\sim} \mathcal{N}_p(0, \Sigma)$ with $\Sigma_{i,j} = 2^{-|i-j|}$ and $(e_i, \delta_i)^T \overset{\text{iid}}{\sim} \mathcal{N}_2(0, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix})$. We also assume that W_i and (e_i, δ_i) are sampled independently. Furthermore, I fix $\beta^* = 1$, $p_X = p_Z = 10, \phi^* = (0.6, 0.7, ..., 1.4, 1.5)^T$ and $\psi^* = (1.1, 1.2, ..., 1.9, 2)^T$. Observe that using all this information Y_i and D_i can be computed using the definitions of the outcome and association model. It can furthermore be observed that this data satisfies the conditions from section 2.7.2.

For the method from section 2.6, I use the corresponding R-studio implementation named "SearchingSampling" as found at https://rdrr.io/cran/RobustIV/man/SearchingSampling.html. Unless stated else, I have the following settings to resemble section 2.6 as much as possible: method="OLS", intercept="False" and filtering="False". The settings "Sampling" (choosing searching or sampling in the second stage) and "M" (resampling number) are varied in this section. The unmentioned settings (such as the α value) are set to their default options. These default options include: Robust="True" (we don't assume homoskedastic errors in the model), $\alpha = 0.05$, $a = 0.6$ (this a is from the grid size of $[L, U]$, see section 2.5.3) and prop=0.1 (prop refers to proportion of non-empty interval for the sampling method).
In case a model does not meet certain finite sample assumptions (like the finite sample plurality rule), SearchingSampling has certain guardrails to come up with a result anyways. These guardrails are specified below:

1. In case $\hat{S} = \varnothing$, the method will continue with $\hat{S} = \{i : 1 \leq i \leq p_Z\}$, i.e. any potential instrumental variable is relevant.
2. SearchingSampling works with the efficient implementation of the searching CI (as seen in section 2.5.3). In case no $\beta \in \mathcal{B}$ satisfies the majority rule for the set $\hat{V}$, SearchingSampling will return the $[L, U]$ interval as its answer for the confidence interval of β^*.

I will vary between 4 different models for the simulation studies:

(S1) $\pi^* = (0 \cdot 1_6, \tau \cdot \gamma_0, \tau \cdot \gamma_0, -\frac{1}{2}, -1)$
$\gamma^* = \gamma_0 \cdot 1_{10}$

(S2) $\pi^* = (0, 0, 0, \tau \cdot \gamma_0, \tau \cdot \gamma_0 + 0.1, \tau \cdot \gamma_0, -\frac{1}{2}, -1, -\frac{2}{3}, -\frac{1}{2})$
$\gamma^* = \gamma_0 \cdot 1_{10}$

(S3) $\pi^* = (0, 0, 0, 0, \tau \cdot \gamma_0, \tau \cdot \gamma_0 + 0.1, -\frac{1}{6}, -\frac{1}{3}, -\frac{1}{2}, -\frac{2}{3})$
$\gamma^* = \gamma_0 \cdot 1_{10}$

(S4) $\pi^* = (0, 0, 0, 0, \tau \cdot \gamma_0, \tau \cdot \gamma_0, \tau \cdot \gamma_0, \tau \cdot \gamma_0 + 0.1, -\frac{1}{3}, -\frac{1}{2})$
$\gamma^* = \gamma_0 \cdot 1_{10}$

τ and γ_0 will be varied. Each model is going to be evaluated on three points: firstly the coverage (is β^* in the resulting CI?), secondly the length of the resulting CI and thirdly whether the finite sample plurality rule was satisfied in the $\hat{V}$ step (for at least some $\beta \in \mathcal{B}$). For each of the (S1)-(S4) with corresponding n and (τ, γ_0) values, 100 simulations are applied and the 100 corresponding results are averaged.

2.8.2. Fixing τ, γ_0, varying n

I will first fix $\tau = 0.2$ and $\gamma_0 = \frac{1}{2}$ and vary the number of data points generated for each model (S1)-(S4). I first consider the SearchingSampling method with the searching algorithm (see section 2.5.3) and then with the sampling algorithm (see section 2.5.4).
Searching algorithm: The results can be found in table 2.2.

Set	n	Cov	Len	Check
S1	50	0.95	47.30	1
	100	0.98	2.37	1
	500	1	0.61	1
	1000	1	0.40	1
	5000	1	0.17	1
S2	50	0.93	66.46	1
	100	0.90	2.92	1
	500	0.97	0.59	0.97
	1000	0.89	0.33	0.99
	5000	0.96	0.62	0.53
S3	50	0.95	49.77	1
	100	0.97	2.61	1
	500	0.95	0.68	1
	1000	0.98	0.39	0.98
	5000	1	0.17	1
S4	50	1	38.75	1
	100	0.99	2.14	1
	500	1	0.60	1
	1000	1	0.36	1
	5000	0.94	0.17	0.93

Table 2.2.: Results for Cov (coverage), Len (length) and Check (plurality rule) corresponding to models (S1)-(S4) with $\tau = 0.2$ and $\gamma_0 = \frac{1}{2}$ and $n = 50, 100, 500, 1000, 5000$ using searching

(S1) and (S3) follow a largely 'predictable' pattern of improvement across the categories as n increases. What is noteworthy is the collapse of the Check from $n = 1000$ to $n = 5000$ for (S2) with an increase in length and increase in coverage. A second noteworthy observation is that (S4) starts to (slightly) undercover and have a decreased Check for $n = 5000$. For (S2) with $n = 5000$, the most common $\hat{V}$'s are $\hat{V} = \{1, 2, 3, 4, 5, 6, 7\}$ and $\hat{V} = \{1, 2, 3, 4, 6\}$ across the 100 simulations. Observe that this first $\hat{V}$ from the previous sentence does not satisfy the majority rule we are testing, which explains a Check value around 50%. As in many cases the finite sample plurality rule won't be satisfied here, the algorithm will return the interval $[L, U]$ as explained before. We know that it holds that: $\hat{\text{CI}}^{\text{sear}} \subseteq [L, U]$ which explains the increase in length from $n = 1000$ to $n = 5000$. The simulation study showed that once $[L, U]$ is chosen by the algorithm, it (almost always) contains β^* hence justifying the improvement in the coverage level. This last observation is not surprising as we already established that $\mathbb{P}(\beta^* \in [L, U]) \to 1$ for $n \to \infty$ (see section 2.5.3) and that $[L, U]$ is in general a larger interval. Another thing that can be noticed is that the Cov and Check perform quite well for small n across the board. In reality, the finite sample plurality rule is not satisfied most of the time for these smaller n cases. For instance, (S4) with $n = 1000$, has as its most common $\hat{V}$: $\{1, 2, 3, 4, 5, 6, 7, 8\}$, which does not satisfy the majority rule. However, π_1^* to π_8^* may be small enough for $\hat{\Gamma}_1$ to $\hat{\Gamma}_8$ and $\hat{\gamma}_1$ to $\hat{\gamma}_8$ to be close to each other (for small n). Consequently $\hat{\Gamma}_j - \beta\hat{\gamma}_j$ will be close to each other for $j = 1, ..., 8$ and meet the majority rule for certain $\beta \in \mathcal{B}$. The difference between the chosen β's also tends to be larger as seen from the interval length when comparing it to interval lengths for larger n's.

When we apply SearchingSampling to (S2) for $n = 10000, 25000, 50000, 75000$ and 100000, we get the results as seen in table 2.3. Hence, only after $n = 25000$ we can see that the method stabilizes and reaches its asymptotic properties for the Cov and Check. For $n = 10000$ the most common $\hat{V}$'s were $\hat{V} = \{1, 2, 3\}$ and $\hat{V} = \{1, 2, 3, 7, 8, 9\}$. The latter $\hat{V}$ does not satisfy the majority rule which explains the still relatively low Check value. See section 2.8.3 for a more in-depth analysis of the voting procedures that lead to these finite sample plurality rule violations "even" for $n = 10000$.

Set	n	Cov	Len	Check
S2	10 000	0.98	0.31	0.75
	25 000	1	0.1	0.99
	50 000	1	0.06	1
	75 000	1	0.05	1
	100 000	1	0.04	1

Table 2.3.: Results for Cov (coverage), Len (length) and Check (plurality rule) corresponding to model (S2) with $\tau = 0.2$ and $\gamma_0 = \frac{1}{2}$ and $n = 10000, 25000, 50000, 75000, 100000$

The second observation about the under coverage and decreased Check of (S4) as seen in table 2.3 at $n = 5000$ will be addressed in more detail in section 2.8.3.

Sampling algorithm: We will now apply the sampling method (as seen in section 2.5.4) to the $\hat{V}$'s which are constructed from the same datasets as in the searching section before. $M = 100$ is chosen here as the number of resamplings. The results can be found in table 2.4.

Set	n	Cov	Len	Check
S1	50	0.90	35.15	1
	100	0.94	0.95	1
	500	0.97	0.25	1
	1000	0.98	0.18	1
	5000	0.98	0.07	1
S2	50	0.73	81.96	1
	100	0.72	1.50	0.99
	500	0.83	0.33	0.98
	1000	0.69	0.22	0.96
	5000	0.91	0.44	0.59
S3	50	0.83	72.74	1
	100	0.91	1.20	0.99
	500	0.94	0.44	0.99
	1000	0.91	0.22	0.98
	5000	0.99	0.08	1
S4	50	0.87	66.72	1
	100	0.92	0.85	1
	500	0.85	0.29	1
	1000	0.88	0.22	1
	5000	0.95	0.09	0.94

Table 2.4.: Results for Cov (coverage), Len (length) and Check (plurality rule) corresponding to models S1-S4 with $\tau = 0.2$ and $\gamma_0 = \frac{1}{2}$ and $n = 50, 100, 500, 1000, 5000$ using the sampling method.

Compared to the searching case, one can see a quite decisive improvement in the length of the intervals as n increases. The other categories remain quite steady or even slightly worse for smaller n's. The latter is not surprising as the intervals are now smaller compared to the searching case. Again, as for the searching case, (S2) has this decline in the Check category from $n = 1000$ to $n = 5000$.
Again increasing the n's as before leads to the conclusion that the method stabilizes at $n = 25000$, see table 2.5. So no improvement for (S2) compared to searching using these metrics.

Set	n	Cov	Len	Check
S2	10000	1	0.21	0.85
	25000	1	0.06	1
	50000	1	0.04	1
	75000	1	0.03	1
	100000	0.99	0.03	1

Table 2.5.: Results for Cov (coverage), Len (length) and Check (plurality rule) corresponding to model (S2) with $\tau = 0.2$ and $\gamma_0 = \frac{1}{2}$ and $n = 10000, 25000, 50000, 75000, 100000$ using sampling

Cov-(S2)	γ_0=0.05	0.075	0.1	0.5
τ=0.025	1	1	1	1
0.05	1	1	1	1
0.075	1	1	1	1
0.1	1	1	1	0.99
0.2	1	1	1	1
0.3	1	0.97	0.92	1
0.4	0.92	0.79	0.88	0.99
0.5	0.83	0.65	0.81	1

Table 2.7.: Results for Cov (coverage) corresponding to model (S2) with $\tau \in \{0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\gamma_0 \in \{0.05, 0.075, 0.1, 0.5\}$ with $n = 25000$ using sampling

Two alternative strategies to improve the performance of the method recommended by [2] are to apply "filtering" and to increase the M. Filtering essentially removes all the outliers from the M samples, see [2, p.14 Remark 4] for the exact details. Applying both concepts to (S2) for $n \in \{50, 100, 500, 1000, 5000\}$ did not change any underlying dynamics. Though, $M = 500$ did lead to quite a big improvement among smaller n's regarding the Cov but has bigger average interval lengths compared to $M = 100$. See table 2.6 for these results.

Set	n	Cov	Len	Check
S2 - filtering	50	0.73	81.94	1
	100	0.73	1.53	0.99
	500	0.84	0.34	0.98
	1000	0.71	0.22	0.96
	5000	0.92	0.44	0.59
S2- M=500	50	0.84	56.33	1
	100	0.93	1.75	1
	500	0.96	0.38	1
	1000	0.90	0.27	0.98
	5000	0.98	0.55	0.52

Table 2.6.: Results for Cov (coverage), Len (length) and Check (plurality rule) corresponding to model (S2) with $\tau = 0.2$ and $\gamma_0 = \frac{1}{2}$ and $n = 10000, 25000, 50000, 75000, 100000$ using sampling together with filtering and $M = 500$

2.8.3. Varying γ_0,τ with searching

In this final part of the simulation studies section, I will fix the n and consider the results for 2 different models for different values of τ and γ_0. I will consider
$\tau \in \{0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\gamma_0 \in \{0.05, 0.075, 0.1, 0.5\}$. The effect of these different values will be investigated for (S2) and (S4), where the n is put to $n = 25000$ and $n = 5000$ respectively. Only the searching algorithm will be considered here.

The results for (S2) can be found in tables 2.7-2.9

Len-(S2)	γ_0=0.05	0.075	0.1	0.5
τ=0.025	1.23	0.64	0.47	0.09
0.05	1.24	0.62	0.46	0.09
0.075	1.23	0.64	0.46	0.08
0.1	1.01	0.62	0.45	0.08
0.2	0.99	0.69	0.42	0.09
0.3	1.31	0.67	0.85	0.09
0.4	1.49	2.23	3.63	0.08
0.5	4.15	5.25	4.77	0.09

Table 2.8.: Results for Len (length) corresponding to model (S2) with $\tau \in \{0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\gamma_0 \in \{0.05, 0.075, 0.1, 0.5\}$ with $n = 25000$ using sampling

Check-(S2)	γ_0=0.05	0.075	0.1	0.5
τ=0.025	0.99	1	1	1
0.05	0.99	1	1	1
0.075	0.99	1	1	1
0.1	1	1	1	0.99
0.2	1	0.99	1	1
0.3	0.98	0.99	0.92	0.98
0.4	0.96	0.79	0.46	0.99
0.5	0.82	0.55	0.33	0.99

Table 2.9.: Results for Check (finite sample plurality rule) corresponding to model (S2) with $\tau \in \{0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\gamma_0 \in \{0.05, 0.075, 0.1, 0.5\}$ with $n = 25000$ using sampling

For this n, $\hat{S} = S$ across the board, so there is no difficulty here in the first stage (i.e. $\hat{S}$ selection). What can be noted is that the length of the confidence intervals decreases as γ_0 increases. This displays that it is likely that less data is needed for larger values of γ_0 to accurately estimate it. Another observation from the table is that, unlike previous simulations, the the lower Check cases, as seen for the rows of $\tau = 0.4$ and $\tau = 0.5$, don't result into a higher coverage rate. The simulation showed that this is (almost) purely due to the cases where the Check was satisfied: the $[L, U]$ interval contained β^* for virtually all cases for this n. The reason that even though the Check was satisfied, the resulting interval did not contain β^* will be discussed next.
For this next discussion, I will pick out two cases which look similar on paper but produce wildly different results. Two such instances are $(\tau, \gamma_0) = (0.5, 0.075)$ and $(\tau, \gamma_0) = (0.3, 0.075)$ which result in the following invalidity levels:

$$\text{M1: } \tau = \frac{1}{2}, \gamma_0 = 0.075: \qquad (\frac{\pi_j^*}{\gamma_j^*})_{1 \le j \le 10} = (0, 0, 0, \frac{1}{2}, \frac{11}{6}, \frac{1}{2}, -\frac{20}{3}, -\frac{40}{3}, -\frac{80}{9}, -\frac{20}{3})$$

$$\text{M2: } \tau = \frac{3}{10}, \gamma_0 = 0.075: \qquad (\frac{\pi_j^*}{\gamma_j^*})_{1 \le j \le 10} = (0, 0, 0, \frac{3}{10}, \frac{11}{6}, \frac{3}{10}, -\frac{20}{3}, -\frac{40}{3}, -\frac{80}{9}, -\frac{20}{3})$$

For M1 above, the most common $\hat{V}$'s were: $\{1, 2, 3, 4, 6, 7, 8, 9, 10\}, \{1, 2, 3, 4, 6\}$ and $\{7, 8, 9, 10\}$. For M2, by far, the most common $\hat{V}$ was $\{1, 2, 3, 4, 6\}$ and in rare instances it chose $\{1, 2, 3, 4, 6, 7, 8, 9, 10\}$. The ongoing 'battle' for both cases seems to be about the amount of

internal voting within the group $Z^(1) - Z^{(3)}, Z^{(4)}, Z^{(6)}$ versus $Z^{(7)} - Z^{(10)}$. The only differences between these two groups when comparing M1 with M2 with respect to invalidity levels are for $Z^{(4)}$ and $Z^{(6)}$. Due to 0.3 being smaller than 0.5, there seems to be more interval voting within $Z^{(1)} - Z^{(3)}, Z^{(4)}, Z^{(6)}$ which leads to (a part of this group) being selected for $\hat{W}$ and is in the following step expanded to $Z^{(1)} - Z^{(3)}, Z^{(4)}, Z^{(6)}$. In the case that there is as much internal voting within $Z^{(1)} - Z^{(3)}, Z^{(4)}, Z^{(6)}$ as in $Z^{(7)} - Z^{(9)}$: $Z^{(1)} - Z^{(10)}/Z^{(5)}$ is chosen as $\hat{W}$ (or a subset of that which is then extended in the next step). For M1, it is likely that the invalidity level of $Z^{(4)}$ is too different from the levels of $Z^{(1)} - Z^{(3)}$ which leads to $Z^{(7)} - Z^{(10)}$ being chosen as a result of the $\hat{W}$ and $\hat{V}$ step. M1 and M2 primarily show, in my view, that even for a relatively large number of data points ($n = 25000$), large invalidity levels (even if they are not particularly close) will vote for each other while smaller invalidity levels are more selective in this sense.
An interesting final remark on M1 and M2 is on the effect of the chosen covariance matrix of W on the above selection process. Repeating the same study above but then for $\Sigma = \text{Id}$ shows that the 'fight' for the most interval voting is between $Z^{(1)} - Z^{(3)}$ and $Z^{(7)}, Z^{(8)}, Z^{(10)}$. All the most outlying invalidity levels from both groups have been removed compared to before. This is not particularly shocking as we, in that case, have less standard error for the estimations of the invalidity levels. The interesting part is that $Z^{(8)}$ is still supported by the quite different $Z^{(7)}$ and $Z^{(10)}$. This is in-line the previously established notion that large invalidity levels tend to vote for each other.

We lastly turn to the results for (S4) and $n = 5000$ for the different (τ, γ_0) values, which can be found in table 2.10-2.12

Cov-(S4)	γ_0=0.05	0.075	0.1	0.5
τ=0.025	0.95	1	1	1
0.05	0.97	1	1	1
0.075	0.98	1	1	1
0.1	0.99	1	1	0.97
0.2	0.93	1	1	0.96
0.3	0.97	1	1	1
0.4	0.96	1	1	1
0.5	0.92	0.99	0.99	1

Table 2.10.: Results for Cov (coverage) corresponding to model (S4) with $\tau \in \{0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\gamma_0 \in \{0.05, 0.075, 0.1, 0.5\}$ with $n = 5000$ using sampling

Len-(S4)	γ_0=0.05	0.075	0.1	0.5
τ=0.025	1.96	1.55	1.22	0.19
0.05	1.73	1.60	1.25	0.19
0.075	2.40	1.54	1.21	0.19
0.1	2.37	1.60	1.19	0.18
0.2	2.10	1.51	1.15	0.16
0.3	1.77	1.47	1.11	0.17
0.4	2.45	1.45	1.07	0.17
0.5	2.33	1.40	0.98	0.17

Table 2.11.: Results for Len (length) corresponding to model (S4) with $\tau \in \{0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\gamma_0 \in \{0.05, 0.075, 0.1, 0.5\}$ with $n = 5000$ using sampling

Check-(S4)	γ_0=0.05	0.075	0.1	0.5
τ=0.025	0.99	1	1	1
0.05	0.99	1	1	1
0.075	0.94	1	1	1
0.1	0.96	1	1	1
0.2	0.93	1	1	0.97
0.3	0.97	1	1	1
0.4	0.93	1	1	0.99
0.5	0.94	1	1	0.98

Table 2.12.: Results for Check (finite plurality rule) corresponding to model (S4) with $\tau \in \{0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\gamma_0 \in \{0.05, 0.075, 0.1, 0.5\}$ with $n = 5000$ using sampling

With the Len-table, one can see a decreasing trend as γ_0 gets larger while there are no big differences when fixing γ_0 and varying τ. Related to this is the observation that $\gamma_0 = 0.05$ gives the worst results across the board, while in general τ does not seem to influence the results. To understand why this might be happening, I will compare the case $(\tau, \gamma_0) = (0.2, 0.05)$ with $(\tau, \gamma_0) = (0.2, 0.075)$. Which have the following invalidity levels:

$$\text{M3: } \tau = 0.2, \gamma_0 = 0.05: \qquad (\frac{\pi_j^*}{\gamma_j^*})_{1\leq j\leq 10} = (0, 0, 0, 0, 0.2, 0.2, 0.2, 2.2, -\frac{1}{3}, -\frac{1}{2})$$

$$\text{M4: } \tau = 0.3, \gamma_0 = 0.075: \qquad (\frac{\pi_j^*}{\gamma_j^*})_{1\leq j\leq 10} = (0, 0, 0, 0, 0.2, 0.2, 0.2, 1\frac{8}{15}, -\frac{1}{3}, -\frac{1}{2})$$

In this case, the main differences are due to different $\hat{S}$'s: for M3 chosen $\hat{S}$'s are for example $\{1, 3, 5, 7, 8, 9\}$ or $\{2, 5, 6, 9, 10\}$ where the common pattern among the chosen $\hat{S}$'s is that usually $Z^{(1)} - Z^{(4)}$ are not selected at the same time which can be seen in the corresponding Check values. This results in $\hat{V}$'s that don't satisfy the majority rule like for instance $\{7, 8, 9\}$ or $\{2, 6, 9, 10\}$. For M4 $Z^{(1)} - Z^{(4)}$ are, more often than for M3, selected at the same time (for instance selecting $Z^{(1)} - Z^{(9)}$). The most common $\hat{V}$ for M4 is $Z^{(1)} - Z^{(7)}$, which satisfies the majority rule. In short: for $n = 5000$ the models have some difficulty selecting relevant instrumental variables for too small a value of γ_0 which results into worse Check and Cov values.

3. Anchor Regression (AR)

3.1. General idea behind Anchor Regression

Just as in the previous section, we assume linear models. For the CIII-method, we assumed a certain number of IVs present in our model. In case we don't have enough information to assume a majority/plurality rule or don't think it's reasonable to do so, CIII won't be a good option anymore to obtain an estimator for treatment effect and/or confidence interval for it.

Anchor Regression (AR) does not require prior information on the number of IVs or hidden confounders present. Instead of trying to obtain the exact treatment effect, AR tries to mediate between different assumptions on the model. In particular, AR mediates between two assumptions: the first one is that all covariates X and potential IVs Z are valid instrumental variables and the second one is that there are no hidden confounders.

In case we would already know beforehand that either all (X, Z) are instrumental, we'll see that there is a method that (asymptotically) infers the treatment and that there is a different method for the case there are no hidden confounders to obtain the treatment effect. AR combines the objective functions of both of these methods above to obtain a mediated estimator for the treatment effect. It will turn out that this mediated estimator will have properties such as distributional robustness (how will this mediated parameter perform on perturbed/shifted data?), replicability (if we try to obtain this parameter from a new dataset, will we get the same parameter?) and stability (can we obtain the same parameter from perturbed as well as unperturbed data?). This will all be further explained in this section. Anchor Regression originates from 3.

3.2. General setting

In this subsection, the general setting for section 3 will be outlined.

Setting: Assume that the data are generated from the linear structural equation model (SEM) as follows: $(X, Y, H, A)^T \sim \mathbb{P}_{\text{train}}$ satisfy (for every event):

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \epsilon + MA$$

- $X \in \mathbb{R}^d, Y \in \mathbb{R}$ are random and represent the covariates and outcome respectively.
- M and B are unknown deterministic matrices. M is also called the *shift matrix*.
- $A \in \mathbb{R}^q$, $H \in \mathbb{R}^r$ and $\epsilon \in \mathbb{R}^{d+1+r}$ are random and represent the anchor variables, hidden variables and noise respectively.
- $A \perp\!\!\!\perp \epsilon$, $\mathbb{E}_{\text{train}}(X) = 0, \mathbb{E}_{\text{train}}(Y) = 0, \mathbb{E}_{\text{train}}(\epsilon) = 0, \epsilon$ as well as A have finite second moments and the components of ϵ are independent of each other.
- $\text{Id} - B$ is invertible.

Remark 3.2.1.

- $\mathrm{Id} - B$ *being invertible guarantees that the distribution of* (X, Y, H) *under* $\mathbb{P}_{\text{train}}$ *can be uniquely defined in terms of* B, ϵ, M *and* A*, namely:*

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = (\mathrm{Id} - B)^{-1}(\epsilon + MA)$$

- *This model induces a directed graph* G *where in case* $M_{k,l} \neq 0$ *an (directed) edge is drawn from the* l*-th variable of* A *to the* k*-variable of* $(X, Y, H)^T$ *and in case* $B_{k,l} \neq 0$ *an edge is drawn from the* l*-th variable in* $(X, Y, H)^T$ *to the* k*-th variable in* $(X, Y, H)^T$*. As, with this construction,* A *is a source node* A *is also called the anchor (i.e.* A *influences* (X, H, Y) *(directly or indirectly) but is itself not influences by another random variable) . Note that* G *is allowed to be cyclic in this context.*

Remark 3.2.2 (Connection to setting CIII)**.** *The general setting of section 3 as defined above also includes the general setting of section 2. Here,* $(X, Z)^T$ *from the CIII section should be viewed as the anchor and* D *as the covariates in the sense of setting of Anchor Regression as described before.*

Under the setting as defined above, we can establish equivalent expressions for $\mathbb{E}(X|A)$, $\mathbb{E}(Y|A)$ and $\mathbb{E}(H|A)$ due the models linearity in A. This linearity property was used throughout the original paper but never explicitly mentioned (or proven therefore). I came up with the proof myself and added a consequential property for M and B.

Lemma 3.2.3. *Assume the setting of section 3.2 and that* $\mathbb{E}_{\text{train}}(AA^T)$ *is invertible. Then:*

$$\begin{aligned} \mathbb{E}_{\text{train}}(X^T|A) &= A^T[\mathbb{E}_{\text{train}}(AA^T)]^{-1}\,\mathbb{E}_{\text{train}}(AX^T) \\ \mathbb{E}_{\text{train}}(Y|A) &= A^T[\mathbb{E}_{\text{train}}(AA^T)]^{-1}\,\mathbb{E}_{\text{train}}(AY) \end{aligned}$$

Consequently:

$$\begin{aligned} M^T([(\mathrm{Id} - B)^{-1}]_{1:d,\cdot})^T &= [\mathbb{E}_{\text{train}}(AA^T)]^{-1}\,\mathbb{E}_{\text{train}}(AX^T) \\ M^T([(\mathrm{Id} - B)^{-1}]_{d+1,\cdot})^T &= [\mathbb{E}_{\text{train}}(AA^T)]^{-1}\,\mathbb{E}_{\text{train}}(AY) \end{aligned}$$

Proof. From the model it can be observed that (using that $\epsilon \perp\!\!\!\perp A$ and $\mathbb{E}_{\text{train}}(\epsilon) = 0$):

$$\begin{aligned} \mathbb{E}_{\text{train}}(X^T|A) &= \mathbb{E}_{\text{train}}((A^T M^T + \epsilon^T)([(\mathrm{Id} - B)^{-1}]_{1:d,\cdot})^T|A) \\ &= A^T M^T([(\mathrm{Id} - B)^{-1}]_{1:d,\cdot})^T \end{aligned}$$

Hence, $\mathbb{E}_{\text{train}}(X^T|A)$ is of the form $A^T C$, where C is a deterministic matrix. As $\mathbb{E}_{\text{train}}(X^T|A)$ is the orthogonal projection of X^T onto the L^2 space of $\sigma(A)$-measurable functions, we hence know that: $\mathbb{E}_{\text{train}}(X^T|A) = A^T\tilde{C}$, where:

$$\begin{aligned} \tilde{C} &= \underset{C \in \mathbb{R}^{q \times d}}{\arg\min}\, \mathbb{E}_{\text{train}}(X^T X - 2A^T CX + A^T CC^T A) \\ &\overset{(1)}{=} \underset{C \in \mathbb{R}^{q \times d}}{\arg\min}\, \mathbb{E}_{\text{train}}(X^T X - 2\,\mathrm{Tr}(A^T CX) + \mathrm{Tr}(A^T CC^T A)) \end{aligned}$$

$$\overset{(2)}{=} \underset{C\in\mathbb{R}^{q\times d}}{\arg\min}\, \mathbb{E}_{\text{train}}(X^TX - 2\,\text{Tr}(XA^TC) + \text{Tr}(C^TAA^TC))$$
$$\overset{(3)}{=} \underset{C\in\mathbb{R}^{q\times d}}{\arg\min}\, \mathbb{E}_{\text{train}}(X^TX) - 2\,\text{Tr}(\mathbb{E}_{\text{train}}(XA^T)C) + \text{Tr}(C^T\,\mathbb{E}_{\text{train}}(AA^T)C)$$
$$=: \underset{C\in\mathbb{R}^{q\times d}}{\arg\min}\, f(C)$$

$\partial_C f(C) = -2\,\mathbb{E}_{\text{train}}(XA^T) + C^T(2\,\mathbb{E}_{\text{train}}(AA^T))$[1]. Hence, setting this equation to zero and seeing that the second derivative is positive definite we can conclude that $\tilde{C} = [\mathbb{E}_{\text{train}}(AA^T)]^{-1}\,\mathbb{E}_{\text{train}}(AX^T)$. We also know that:

$$\begin{aligned} A^T\tilde{C} &= A^TM^T(([\text{Id}-B]^{-1})_{1:d,\cdot})^T \implies \\ AA^T\tilde{C} &= AA^TM^T(([\text{Id}-B]^{-1})_{1:d,\cdot})^T \implies \\ \mathbb{E}_{\text{train}}(AA^T)\tilde{C} &= \mathbb{E}_{\text{train}}(AA^T)M^T(([\text{Id}-B]^{-1})_{1:d,\cdot})^T \implies \\ \tilde{C} &= M^T(([\text{Id}-B]^{-1})_{1:d,\cdot})^T \end{aligned}$$

where in the last step $\mathbb{E}_{\text{train}}(AA^T)$ being invertible is used.
(1): For any real number $x \in \mathbb{R}$: $\text{Tr}(x) = x$.
(2): $\text{Tr}(B_1B_2) = \text{Tr}(B_2B_1)$ (where sizes of B_1 and B_2 are appropriate).
(3): $\mathbb{E}_{\text{train}}(\text{Tr}(B_1)) = \text{Tr}(\mathbb{E}_{train}(B_1))$ □

For the setting above, we assumed that $\text{Id}-B$ is invertible. In case G is acyclic we automatically have that $\text{Id}-B$ is invertible. This property was mentioned at [9, p.6 Section 2.1] but not proven.

Lemma 3.2.4. *Let G be the graph as derived in remark 3.2.1 where we assume it to be acyclic. Then $\text{Id}-B$ is invertible with $\det(\text{Id}-B) = 1$.*

Proof. It suffices to show that the echelon-form of $\text{Id}-B$ is a upper-triangular matrix with one's on the diagonal. As G is acyclic: $B_{i,i} = 0$ for all i, hence $(\text{Id}-B)_{i,i} = 1$ for all i. Consider the first column of $(\text{Id}-B)$ and suppose that $(\text{Id}-B)_{2,1} \neq 0$ Then we subtract the first row from the second row and as $(\text{Id}-B)_{2,1} \neq 0$ implies $(\text{Id}-B)_{1,2} = 0$ (G is acyclic), after the row is subtracted the second diagonal element is still equal to 1. Continuing this process of row-reducing, we end up with the echelon form which is an upper-triangular matrix with only one's on the diagonal: hence invertible with $\det(\text{Id}-B) = 1$. □

3.2.1. Partialling out and instrumental variable

As explained in the introduction, AR provides a trade-off between an instrumental variable method and a so called "partialling-out" method, who provides the causal effect under different conditions. To make this more precise, I will now introduce the IV and partialling-out methods. Below: $\text{P}_A(\cdot) := \mathbb{E}_{\text{train}}(\cdot|A)$.

Definition 3.2.5. *Under the assumption of existence and uniqueness:*

- *Partialling out:* $b_{\text{PA}} = \underset{b\in\mathbb{R}^d}{\arg\min}\, \mathbb{E}_{\text{train}}([(\text{Id}-\text{P}_A)(Y - X^Tb)]^2)$

[1] https://en.wikipedia.org/wiki/Matrix_calculus

- *Instrumental variable:* $b_{\text{IV}} = \arg\min_{b \in \mathbb{R}^d} \mathbb{E}_{\text{train}}([\text{P}_A(Y - X^T b)]^2)$

To illustrate the properties of b_{PA} and b_{IV}, I will show a series of examples of different models according to section 3.2 and their resulting b_{PA} and b_{IV}.
For convenience, assume that X, A and $Y \in \mathbb{R}$. In the linear structural equation setting, b_{PA} will be equal to the causal effect (under consistency) of X on Y in case A is a confounder to X and Y with no other (hidden) confounders. In that sense, b_{PA} computes the effect of other on Y by removing the effect of A on Y and X. For example:

$$\begin{aligned} Y &= \beta_1 X + \alpha_1 A + \epsilon_1 \\ X &= \alpha_2 A + \epsilon_2 \end{aligned}$$

results into:

$$b_{\text{PA}} = \arg\min_{b \in \mathbb{R}} \mathbb{E}([\epsilon_2 \beta_1 + \epsilon_1 - b\epsilon_2]^2) = \beta_1$$

b_{IV} equals the causal effect if A is an instrumental variable. For example:

$$\begin{aligned} Y &= \beta_1 X + \epsilon_1 \\ X &= \alpha_1 A + \epsilon_2 \end{aligned}$$

gives the result:

$$\beta_{IV} = \arg\min_{b \in \mathbb{R}} \mathbb{E}([\alpha_1 \beta_1 A - b\alpha_1 A]^2) = \beta_1$$

I will now point out some strengths and weaknesses in both b_{IV} and b_{PA} through 2 examples. Suppose that we have a confounding (hidden) variable $H \in \mathbb{R}$ but still have an instrumental setting for A:

$$\begin{aligned} Y &= \beta_1 X + \beta_2 H + \epsilon_1 \\ X &= \alpha_1 H + \alpha_2 A + \epsilon_2 \\ H &= \epsilon_3 \end{aligned}$$

Then:

$$\begin{aligned} b_{\text{IV}} &= \arg\min_{b \in \mathbb{R}} \mathbb{E}([A(\beta_1 \alpha_2 - \alpha_2 b)]^2) = \beta_1 \\ b_{\text{PA}} &= \arg\max_{b \in \mathbb{R}} \mathbb{E}([\epsilon_1 + \beta_1 \epsilon_2 + (\beta_1 \alpha_1 + \beta_2)\epsilon_3 - b\alpha_1 \epsilon_3 - b\epsilon_2]^2) \end{aligned}$$

Here, b_{PA} takes β_2 (effect of confounder H on Y) into account. Hence, b_{IV} is more robust, for obtaining the causal effect, to confounders if the instrumental setting is achieved.
In case the instrumental setting is not achieved and there are no hidden confounders, the roles switch. For for example:

$$\begin{aligned} Y &= \beta_1 X + \beta_2 A + \epsilon_1 \\ X &= \alpha_1 A + \epsilon_2 \end{aligned}$$

it holds that:

$$b_{\text{PA}} = \arg\min_{b \in \mathbb{R}} \mathbb{E}([\beta_1 \epsilon_2 - b\epsilon_2]^2) = \beta_1$$

$$b_{\mathrm{IV}} = \arg\min_{b\in\mathbb{R}} \mathbb{E}([(\beta_1\alpha_1 + \beta_2 - b\alpha_1]^2)$$

Hence β_2 is taken into account for b_{IV}.

To summarise this discussion about b_{IV} and b_{PA} above, the main challenge for b_{PA} in obtaining the causal effect are hidden confounders while for b_{IV} it is achieving the instrumental variable setting. In practise, for the multivariate $A = (A_1, .., A_q)^T$ we don't know beforehand which coordinates are confounders and which are instrumental and what role H plays. Hence usually a trade-off between b_{IV} and b_{PA} is considered. Later, it is shown that this trade-off is connected to certain predictive stability guarantees (i.e. how well will the model perform on particular data with a different distribution than the training data?)

3.3. Population Anchor Regression

Combining the objective functions for b_{IV} and b_{PA}, we obtain the objective function for Anchor Regression:

Definition 3.3.1. *In the same setting as section 3.2, the parameter of the population version of Anchor Regression is defined as follows (assuming that it exists and is unique) for $\gamma \geq 0$:*

$$b^\gamma = \arg\min_b \mathbb{E}_{\mathrm{train}}([(\mathrm{Id} - \mathrm{P}_A)(Y - X^T b)]^2) + \gamma\, \mathbb{E}_{\mathrm{train}}([\mathrm{P}_A(Y - X^T b)]^2)$$

Here, $\mathrm{P}_A(\cdot) = \mathbb{E}_{\mathrm{train}}(\cdot|A)$

We will first consider conditions for which b^γ indeed exists and is unique. Existence and uniqueness properties were unmentioned in the original paper. Instead, existence and uniqueness was assumed throughout the paper.
The following lemma shows that b^γ is the OLS-estimator on perturbed data of (X, Y).

Lemma 3.3.2. *For $\gamma \geq 0$: $b^\gamma = \arg\min_{b\in\mathbb{R}^d} \mathbb{E}_{\mathrm{train}}([\tilde{Y} - \tilde{X}b]^2)$*
Where:

$$\tilde{Y} = (\mathrm{Id} - \mathrm{P}_A)Y + \sqrt{\gamma}\mathrm{P}_A Y, \quad \tilde{X} = (\mathrm{Id} - \mathrm{P}_A)X^T + \sqrt{\gamma}\mathrm{P}_A X^T$$

The proof is analogous to the proof shown for the estimated Anchor Regression in lemma 3.4.3. The only differences are that we now use that for any random variable $Z \in \mathbb{R}$: $\mathbb{E}_{\mathrm{train}}([([\mathrm{Id} - \mathrm{P}_A])Z](\mathrm{P}_A Z)]) = 0$ and $\mathbb{E}_{\mathrm{train}}(X^T b|A) = \mathbb{E}_{\mathrm{train}}(X^T|A)b$ (where the conditional expectation over a vector given some σ-algebra is defined coordinate-wise.).
The next lemma gives a sufficient condition for an existing and unique b^γ.

Lemma 3.3.3. *Let $\gamma \in [0, \infty)$. Using the same notation as in lemma 3.3.2, in case $\mathbb{E}_{\mathrm{train}}(\tilde{X}\tilde{X}^T)$ is positive definite, then b^γ exists and is unique with the form:*

$$b^\gamma = (\mathbb{E}_{\mathrm{train}}(\tilde{X}^T\tilde{X}))^{-1}\, \mathbb{E}_{\mathrm{train}}(\tilde{Y}\tilde{X})$$

Proof. Define $f(b) = \mathbb{E}_{\text{train}}((\tilde{Y} - \tilde{X}b)^2)$. Then it holds that:

$$f(b) = \mathbb{E}_{\text{train}}(\tilde{Y}^2) - 2\,\mathbb{E}_{\text{train}}(\tilde{Y}\tilde{X}^T))b + b^T\,\mathbb{E}_{\text{train}}(\tilde{X}^T\tilde{X})b.$$

Hence:

$$\partial_b f(b) = -2\,\mathbb{E}_{\text{train}}(\tilde{Y}\tilde{X}^T) + 2b^T\,\mathbb{E}_{\text{train}}(\tilde{X}^T\tilde{X})$$

So, solving $\partial_b f(b) = 0$ gives the unique solution (as $\mathbb{E}_{\text{train}}(\tilde{X}^T\tilde{X})$ positive definite implies invertibility):

$$b = (\mathbb{E}_{\text{train}}(\tilde{X}^T\tilde{X}))^{-1}\,\mathbb{E}_{\text{train}}(\tilde{Y}\tilde{X})$$

As:

$$\partial_b^2 f(b) = 2\,\mathbb{E}_{\text{train}}(\tilde{X}^T\tilde{X})$$

which is positive definite, the b found is the unique global minimizer. □

Remark 3.3.4. *In case we don't assume $\mathbb{E}_{\text{train}}(\tilde{X}^T\tilde{X})$ is invertible, we might get into the situation that there are infinite minimizers. Without the condition it might still be that we have 1 unique global minimizer and infinite local minimizers.*

Due to $E(X|A)$ being the projection of the random variable X onto a closed subspace of L^2 (namely onto the L^2 with respect to the sigma algebra generated by A), we have that:

$$\begin{aligned}\mathbb{E}_{\text{train}}([Y - X^T b]^2) &= \mathbb{E}_{\text{train}}([(\mathrm{P}_A + (\mathrm{Id} - \mathrm{P}_A))(Y - X^T b)]^2) \\ &= \mathbb{E}_{\text{train}}([\mathrm{P}_A(Y - X^T b)]^2) + \mathbb{E}_{\text{train}}([(\mathrm{Id} - \mathrm{P}_A)(Y - X^T b)]^2)\end{aligned}$$

From this (under some assumptions), one can observe:

$$b^1 = b_{\text{OLS}}, b^0 = b_{\text{PA}} \tag{3.3.1}$$

I also want to define: $b^\infty := \lim_{\gamma \to \infty} b^\gamma$. The lemma below gives conditions for existence and different characterisations of this b^∞. This was not mentioned in the original paper (as existence and uniqueness was directly assumed when needed).

Lemma 3.3.5. *Define $F_{\text{PA}}(b) := \mathbb{E}_{\text{train}}([(\mathrm{Id} - \mathrm{P}_A)(Y - X^T b)]^2)$ and $F_{\text{IV}}(b) := \mathbb{E}_{\text{train}}([\mathrm{P}_A(Y - X^T b)]^2)$. Then:*

- *In case b_{IV} exists and is unique: $b^\infty = b_{\text{IV}}$.*
- *In case all $b \in \mathcal{B} \subseteq \mathbb{R}^d$ minimize $F_{\text{IV}}(b)$ where $|\mathcal{B}| \geq 2$: $b^\infty = \arg\min_{b \in \mathcal{B}} F_{\text{PA}}(b)$ (assuming that this minimum exists and is unique).*

Proof.

- $\forall b \in \mathbb{R}^d / \{b_{\text{IV}}\}$ we know that: $\gamma F_{\text{IV}}(b_{\text{IV}}) < \gamma F_{\text{IV}}(b)$. It suffices to establish that for γ large enough:

$$F_{\text{PA}}(b_{\text{IV}}) + \gamma F_{\text{IV}}(b_{\text{IV}}) < F_{\text{PA}}(b) + \gamma F_{\text{IV}}(b)$$

  This turns out to be the case for:

$$\frac{F_{\text{PA}}(b_{\text{IV}}) - F_{\text{PA}}(b)}{F_{\text{IV}}(b) - F_{\text{IV}}(b_{\text{IV}})} < \gamma$$

  which is a well-defined notion: nominator can either be any number in $\mathbb{R}$ and the denominator is strictly greater than zero by assumption.

- By the reasoning above and the fact that $\gamma F_{\text{IV}}(\tilde{b})$ is equal for any $\tilde{b} \in \mathcal{B}$ and for all $\gamma > 0$, the overall minimization problem reduces to:

$$\arg\min_{b \in \mathcal{B}} F_{\text{PA}}(b)$$

  (which exists and is unique by assumption).

□

Hence by (3.3.1) and the lemma above: Anchor Regression interpolates between b_{PA} and b_{OLS} for $\gamma \in [0, 1]$ and between b_{OLS} and b_{IV} for $\gamma \in [1, \infty]$. As mentioned in the introduction, with Anchor Regression we aim for distributional robustness and distributional replicability. In the next sections these properties will be established with respect to "interventions" on A.

Remark 3.3.6. *In lemma 3.3.3, the condition that* $\mathbb{E}_{\text{train}}(\tilde{X}\tilde{X}^T)$ *is invertible is shown to be sufficient for a unique global minimizer to exist. This shows that for partialling out* ($\gamma = 0$) $\mathbb{E}_{\text{train}}(XX^T) - 2\,\mathbb{E}_{\text{train}}(X\,\mathbb{E}_{\text{train}}(X^T|A))$ *being positive definite is sufficient for existence, for OLS* ($\gamma = 1$) $\mathbb{E}_{\text{train}}(XX^T)$ *should be positive definite and for IV* ($\gamma = \infty$) $\mathbb{E}_{\text{train}}(X\,\mathbb{E}_{\text{train}}(X^T|A)|A)$.

3.3.1. Perturbations

We will now study the distribution of $(X, Y, H)^T$ under perturbations (i.e. "interventions" on MA) and its connection to Anchor Regression.

Definition 3.3.7. *The new perturbed ("intervened on* MA*") distribution with respect to* v *is denoted by* $\mathbb{P}_v$*. The distribution of the variables* $(X, Y, H)^T$ *under* $\mathbb{P}_v$ *is defined as the solution of:*

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \epsilon + v$$

Here, $v \in \mathbb{R}^{d+1+q}$ *is a random vector for which* $v \perp\!\!\!\perp \epsilon$*. We assume that the distribution of* ϵ *is the same under* $\mathbb{P}_v$ *as under* $\mathbb{P}_{\text{train}}$*.* v *is also called a shift.*

3.3.2. Distributional robustness under perturbations

In this section, the connection between the population version of Anchor Regression and the worst-case risk over a class of shift perturbations will be established. This property displays the distributional robustness property of Anchor Regression. This result was mentioned at [9, p.11 Section 2.4 Theorem 1]. Here, I present a shorter proof than the originally proposed one [9, p.31 Section 8.6]:

Theorem 3.3.8. *Let the assumptions of 3.2 hold. Then* $\forall b \in \mathbb{R}^d$*:*

$$\mathbb{E}_{\text{train}}([(\text{Id} - \text{P}_A)(Y - X^T b)]^2) + \gamma\, \mathbb{E}_{\text{train}}([\text{P}_A(Y - X^T b)]^2) = \sup_{v \in C^\gamma} \mathbb{E}_v([Y - X^T b]^2)$$

Here: $C^\gamma = \{v \in \mathbb{R}^{d+1+r} \text{ random variable} : \mathbb{E}_v(vv^T) \preccurlyeq \gamma M\, \mathbb{E}_{\text{train}}(AA^T)M^T\}$

Proof. Under $\mathbb{P}_v$, it holds that:

$$Y - X^T b = ([(\text{Id} - B)^{-1}]_{d+1,\cdot} - b^T[(\text{Id} - B)^{-1}]_{1:d,\cdot})(\epsilon + v) =: w_b^T(\epsilon + v)$$

Similarly under $\mathbb{P}_{\text{train}}$: $Y - X^T b = w_b^T(\epsilon + MA)$. Hence, it also holds that:

$$\begin{aligned}\mathbb{E}_v([Y - X^T b]^2) &= \mathbb{E}_v([w_b^T(\epsilon + v)]^2) \\ &= \mathbb{E}_v([w_b^T \epsilon]^2) + \mathbb{E}_v([w_b^T v]^2) \\ &= \mathbb{E}_{\text{train}}([w_b^T \epsilon]^2) + \mathbb{E}_v([w_b^T v]^2)\end{aligned}$$

In the second line above it was used that $v \perp\!\!\!\perp \epsilon$ and $\mathbb{E}_v(\epsilon) = \mathbb{E}_{\text{train}}(\epsilon) = 0$ because ϵ has the same distribution under $\mathbb{P}_v$ as under $\mathbb{P}_{\text{train}}$. This last fact was also used in the last step. Observe that under $\mathbb{P}_{\text{train}}$: $w_b^T \epsilon = (\text{Id} - \text{P}_A)(Y - X^T b)$, and consequently:

$$\mathbb{E}_{\text{train}}([w_b^T \epsilon]^2) = \mathbb{E}_{\text{train}}([(\text{Id} - \text{P}_A)(Y - X^T b)]^2)$$

So it now just suffices to show that:

$$\sup_{v \in C^\gamma} \mathbb{E}_v([w_b^T v]^2) = \gamma\, \mathbb{E}_{\text{train}}([\text{P}_A(Y - X^T b)]^2)$$

Write: $\mathbb{E}_v([w_b^T v]^2) = w_b^T\, \mathbb{E}_v(vv^T) w_b$. For the optimal $v \in C^\gamma$ (if existent), it has to hold that $C := \gamma M\, \mathbb{E}_{\text{train}}(AA^T)M^T - \mathbb{E}_v(vv^T)$ is positive semi-definite and that $w_b^T\, \mathbb{E}_v(vv^T) w_b$ is maximized. Rewriting the expression of C and multiplying by w_b^T and w_b gives:

$$w_b^T\, \mathbb{E}_v(vv^T) w_b = \gamma w_b^T\, \mathbb{E}_{\text{train}}(AA^T)M^T w_b - w_b^T C w_b$$

As the first term on the right hand side does not depend on v and C is positive semi definite, the left hand side is optimized in case $C = 0$, which corresponds to the choice of $v = \sqrt{\gamma} MA$ ($\sim \mathbb{P}_{\text{train}}$).
As $\text{P}_A(Y - X^T b) = w_b^T MA$, we obtain that: $\sup_{v \in C^\gamma} \mathbb{E}_v([w_b^T v]^2) = \gamma\, \mathbb{E}_{\text{train}}([\text{P}_A(Y - X^T b)]^2)$ □

Suppose that we would only allow for deterministic perturbations $v \in \mathbb{R}^{d+r+1}$ in the general setting of this section. Then the analogous theorem would hold where now:

$$C^\gamma = \{v \in \mathbb{R}^{d+1+r} \text{ deterministic } : vv^T \preccurlyeq \gamma M\, \mathbb{E}_{\text{train}}(AA^T)M^T\}$$

In this setting, we can make statements on other (more digestible) expressions for C^γ. The following result was mentioned at [9, p.11 Section 2.4] but was not proven. I want to credit my thesis supervisor Aad v/d Vaart here as he helped me construct the argument.

Lemma 3.3.9. *Suppose that we only allow for deterministic perturbations in definition 3.3.7.*

- *For $\gamma > 0$: $C^\gamma \subseteq \text{span}(M)$*
- *Under the assumptions that $\mathbb{E}_{\text{train}}(AA^T)$ is positive definite, M has full rank and $\gamma \to \infty$: $C^\gamma = \text{span}(M)$*

Proof.

- Define $C := \gamma M \, \mathbb{E}_{\text{train}}(AA^T) M^T$. Note that since $\text{span}(C) \subseteq \text{span}(M)$, it is sufficient to prove:

$$\{v : vv^T \preccurlyeq C\} \subseteq \text{span}(C)$$

As C is symmetric we can write:

$$C = \sum_{i=1}^{d+r+1} \lambda_i q_i q_i^T$$

where q_i are orthogonal eigenvectors that span $\mathbb{R}^{d+r+1}$. Hence, we can also write:

$$v = \sum_{i=1}^{d+r+1} x_i q_i$$

which gives

$$vv^T = \sum_{i=1}^{d+r+1} x_i^2 q_i q_i^T .$$

As $C - vv^T$ is positive semi-definite (by assumption), we can write:

$$C - vv^T = \sum_{i=1}^{d+r+1} (\lambda_i - x_i^2) q_i q_i^T = Q \, \text{diag}(\lambda_j - x_j^2 : j = 1, .., d+r+1) Q^T .$$

where $Q_{.i} = q_i$. This implies that (by assumption): $\lambda_i - x_i^2 \geq 0 \; \forall i$ so $\lambda_i = 0 \implies x_i = 0$. Hence, we can rewrite v and C as follows:

$$v = \sum_{j=1}^{N} x'_j q'_j$$
$$C = \sum_{j=1}^{N} \lambda'_j q'_j q'^T_j$$

here, $\lambda'_j \neq 0 \; \forall j$ and q'_j is the eigenvector corresponding to λ'_j. To finish the proof, we would hence have to show that there exists a vector a for which: $Ca = v$. This equation we can rewrite to:

$$\sum_{j=1}^{N} \lambda'_j q'_j \langle q'_j, a \rangle = v$$

Which would be solved by find a for which:

$$\langle q'_j, a \rangle = \frac{x'_j}{\lambda'_j}, j = 1, ..., N$$

Which is equivalent to:

$$\begin{pmatrix} q_1'^T \\ \vdots \\ q_N'^T \end{pmatrix} a = \begin{pmatrix} x'_1/\lambda'_1 \\ \vdots \\ x'_N/\lambda'_N \end{pmatrix}$$

We know that the dimension of the row space of matrix in front of a above is N (eigenvectors form eigenbasis for symmetric matrices), hence the dimension of the column space is also N. This means that the matrix equation above has a solution (as $d+r+1 \geq N$), which finishes the proof.

- Consider Mx (an element from the column space of M). Then for $a \in \mathbb{R}^{d+r+1}$:

$$a^T(\gamma[M\,\mathbb{E}_{\text{train}}(AA^T)M^T] - Mxx^TM^T)a = a^TM\gamma\,\mathbb{E}_{\text{train}}(AA^T)M^Ta - a^TMxx^TM^Ta$$

As $M\,\mathbb{E}_{\text{train}}(AA^T)M^T$ positive definite (M has full rank and $\mathbb{E}_{\text{train}}(AA^T)$ is positive definite), $a^TM\,\mathbb{E}_{\text{train}}(AA^T)M^Ta > 0$ for $a \neq 0$. Multiplying this term by γ: it will be larger for γ, large enough than the non-negative term $a^TMxx^TM^Ta$.

□

The most important take-away from theorem 3.3.8 is that Anchor regression minimizes the worst-case MSE (Mean Squared Error) under shift perturbations up to a given strength in certain directions.

For different values of $\gamma \geq 0$, we have that theorem 3.3.8 gives us that:

- For $\gamma = 0$, b^0 is not guaranteed to work well on shifted data as $C^0 = \{0\}$.
- For $\infty > \gamma > 0$ we obtain predictive guarantees for b^γ on both shifted and unshifted data
- b^γ for $\gamma \to \infty$ works increasingly well under strong perturbations while it can increasingly perform worse on unshifted or moderately shifted data. After all, we minimize with respect to the worst case of the perturbations and for this case (under some other assumptions) C^γ consists all possible "interventions" on MA (see lemma 3.3.9).

For lemma 3.3.9, we assumed that v is deterministic. In case we allow for C^γ to have random vectors, by following the same steps as in the proof (with now the coefficients x_j of v being random), we find that for any $v \in C^\gamma$ there exists a random vector $a \in \mathbb{R}^q$ for which holds: $v = Ma$ a.s. It will also hold, that under the conditions of the second statement of lemma 3.3.9, $\{a : a \overset{\text{a.s.}}{=} Mx \text{ for some } x\} \subseteq C^\gamma$.

3.3.3. Simulated examples performance b^γ on perturbed model

In this section, 2 examples on how b^γ for $\gamma = 5$ performs on $b \mapsto \mathbb{E}_v([Y - X^T b]^2)$ (the Mean Squared Error (MSE) with respect to $\mathbb{P}_v$) for different perturbations v compared to $b_{\text{IV}} (= b^\infty)$, $b_{\text{OLS}} (= b^1)$ and $b_{\text{PA}} (= b^0)$ are displayed.

Example 3.3.10. *[9, p.11 Section 2.3] Let $v_t = (t, 0, 0)^T$, where $t \geq 0$. Under $\mathbb{P}_{\text{train}}$, we have the following linear structural equation:*

$$\begin{aligned} A &\sim \mathbb{P}_{\text{train}}(A = 1) = \mathbb{P}_{\text{train}}(A = -1) = \frac{1}{2} \\ \epsilon_H, \epsilon_X, \epsilon_Y &\overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(0, 1) \\ H &= \epsilon_H \\ X &= A + H + \epsilon_X \\ Y &= X + 2H + \epsilon_Y \end{aligned}$$

So, under $\mathbb{P}_{v_t}$ (the intervention $A = v_t$):

$$\begin{aligned} \epsilon_H, \epsilon_X, \epsilon_Y &\overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(0, 1) \\ H &= \epsilon_H \\ X &= t + H + \epsilon_X \\ Y &= X + 2H + \epsilon_Y \end{aligned}$$

From the examples in section 3.2.1, one can directly observe that $b_{\text{IV}} = 1$ (and equals the causal effect of X on Y with consistency). Hence we obtain that $\forall t \geq 0$:

$$\mathbb{E}_{v_t}([Y - X b_{\text{IV}}]^2) = \mathbb{E}_{v_t}([2\epsilon_H + \epsilon_Y]^2) = 4\,\mathbb{E}_{v_t}([\epsilon_H]^2) + \mathbb{E}_{v_t}([\epsilon_Y]^2) = 5$$

The other MSE's as a function of t are seen in figure 3.1. From this figure, it can be seen that for large perturbation strengths, b_{IV} out performs b^5. An interesting observation is that b_{PA} performs very well for small interventions and performs very poorly for larger interventions compared to the others. This is not particularly surprising in the light of section 3.2.1: setting $A = 0$ gives us the setting of only confounding effects by H (on X and Y) and no other variables affecting X and Y (besides independent errors). Computing b_{PA}, we obtain that:

$$b_{\text{PA}} = \arg\max_{b \in \mathbb{R}} \{(3 - b)^2\,\mathbb{E}_{\text{train}}(\epsilon_H^2) + (1 - b)^2\,\mathbb{E}_{\text{train}}(\epsilon_X^2) + \mathbb{E}_{\text{train}}(\epsilon_Y^2)\} = 2$$

From this, it follows that:

$$\mathbb{E}_{v_t}([Y - X b_{\text{PA}}]^2) = \mathbb{E}_{v_t}([\epsilon_Y - \epsilon_X + \epsilon_H - t]^2) = 3 + t^2$$

b^5 has, compared to $b_{\text{IV}}, b_{\text{PA}}$ and b_{OLS} a mediated performance.

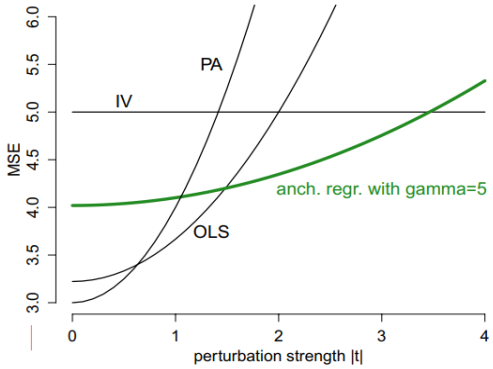

Figure 3.1.: Plot corresponding to example 3.3.10 which computes $\mathbb{E}_{v_t}([Y - Xb]^2)$ for $b = b_{\text{IV}}, b_{\text{PA}}, b_{\text{OLS}}$ and b^5 for increasing $|t|$

In example 3.3.10, we were in a instrumental variable setting for A with respect to the effect of X on Y. In case we are not in a instrumental variable setting nor a confounding A setting (as described in section 3.2.1), we can see the strength of b^γ for large perturbations. This is shown in the next example:

Example 3.3.11. *[9, p.11 Section 2.5] Consider $v_t = (0, 0, t)^T$ with the following model:*

$$\begin{aligned} A \sim \mathbb{P}(A = -1) = \mathbb{P}(A = 1) &= \frac{1}{2} \\ \epsilon_H, \epsilon_X, \epsilon_Y &\overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(0, 1) \\ H &= A + \epsilon_H \\ X &= H + \epsilon_X \\ Y &= X + 2H + \epsilon_Y \end{aligned}$$

See figure 3.2 for the results for intervention v_t as $|t|$ increases. In figure 3.2 "direct causal effect" refers to b_{IV}, however it is important to note that b_{IV} is not equal to the causal effect of X on Y (i.e. 1) but rather $b_{\text{IV}} = 3$. As $|t|$ grows, b^5 starts to outperform the others quite considerably.

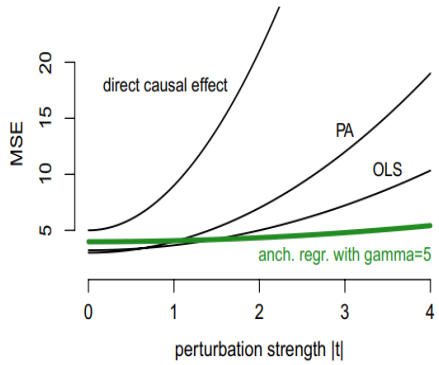

Figure 3.2.: Plot corresponding to example 3.3.11 which computes $\mathbb{E}_{v_t}([Y - Xb]^2)$ for $b = b_{\text{IV}}, b_{\text{PA}}, b_{\text{OLS}}$ and b^5 for increasing $|t|$

3.3.4. Interpretation of Anchor Regression via quantiles

In this subsection, a quantile interpretation of Anchor Regression is provided without the use of the linear assumptions from section 3.2. For the remainder, it is assumed that the anchors are continuous (similar results hold for discrete random variables [9, p.32 Section 8.8]). The linearity assumptions of section 3.2 are dropped and instead it is assumed that $(X, Y, A)^T$ is distributed as a centered multivariate normal.

First, some notation and a recall to the definition of the quantile of $\mathbb{E}((Y - X^T b)^2|A)$ are written below:

Definition 3.3.12. *$Q(\alpha)$ is defined as the α-th quantile of $\mathbb{E}((Y - X^T b)^2|A)$ i.e.:*

$$Q(\alpha) = \inf_{x \in \mathbb{R}} [\alpha \leq \mathbb{P}(\mathbb{E}([Y - X^T b]^2|A) \leq x)]$$

The following lemma connects the quantile of $\mathbb{E}((Y - X^T b)^2|A)$ with Anchor Regression. This result was first stated at [9, p.13 Section 2.6 Lemma 1] and I worked out the details of the corresponding proof given at [9, p.32 Section 8.2]:

Lemma 3.3.13. *Assume $(X, Y, A)^T$ is centered normally distributed under $\mathbb{P}$. Then for $\alpha \in [0, 1]$:*

$$Q(\alpha) = \mathbb{E}([(\mathrm{Id} - \mathrm{P}_A)(Y - X^T b)]^2) + \gamma_\alpha \, \mathbb{E}((\mathrm{P}_A(Y - X^T b))^2)$$

Here, γ_α is the α-th quantile of χ^2_1

Proof. We can write:

$$\begin{aligned}\mathbb{E}([Y - X^T b]^2|A) &= \mathbb{E}([Y - X^T b - \mathbb{E}(Y - X^T b|A) + \mathbb{E}(Y - X^T b|A)]^2|A) \\ &= \mathbb{E}([Y - X^T b - \mathbb{E}(Y - X^T b|A)]^2|A) + [\mathbb{E}(Y - X^T b|A)]^2 =: (*)\end{aligned}$$

As $(X, Y, A)^T$ is a centered multivariate normal:

$$\mathbb{E}(Y - X^T b|A) \sim N(0, \mathbb{E}([\mathbb{E}(Y - X^T b|A)]^2))$$

$$\text{and: } \mathbb{E}([Y - X^T b - \mathbb{E}(Y - X^T b|A)]^2|A) \overset{(1)}{=} \mathbb{E}([Y - X^T b - \mathbb{E}(Y - X^T b|A)]^2)$$

Hence by $(*)$, we obtain that the α-th quantile of $\mathbb{E}([Y - X^T b]^2|A)$ has the form:

$$\mathbb{E}([Y - X^T b - \mathbb{E}(Y - X^T b|A)]^2) + \chi^2_1(\alpha) \, \mathbb{E}([\mathbb{E}(Y - X^T b|A)]^2)$$

Here, $\chi^2_1(\alpha)$ is the α-th quantile of the chi-squared distribution with 1 degree of freedom.
(1) : As $(X, Y, A)^T$ is a centered multivariate normal, so is $(Y - X^T b, A)^T$ (linear transformation of $(X, Y, A)^T$). Consequently, we have that $\mathbb{E}(Y - X^T b|A) = M_1 A$ for some deterministic matrix M_1 (due to it also being a centered normal) which establishes that: $(Y - X^T b - \mathbb{E}(Y - X^T b|A), A)^T$ is also multivariate normal. As:

$$\begin{aligned}\mathrm{Cov}(Y - X^T b - \mathbb{E}(Y - X^T b|A), A) &= \mathbb{E}([Y - X^T b - \mathbb{E}(Y - X^T b|A)][A]) \\ &= \mathbb{E}(A[Y - X^T b] - \mathbb{E}(A[Y - X^T b]|A) = 0\end{aligned}$$

We can conclude that:

$$Y - X^T b - \mathbb{E}(Y - X^T b|A) \perp\!\!\!\perp A.$$

□

Remark 3.3.14. *This shows that Anchor Regression can be used to optimize quantiles of* $\mathbb{E}((Y - X^T b)^2|A)$ *(optimize with respect to* b*, where* α *is chosen beforehand).*

3.3.5. Replicability

From here on, I again assume the setting of section 3.2. We will now consider the question of replicability. This question is as follows: Suppose that we have training and test data where the training and test data are from different distributions. Under what conditions can b^γ perform well on the training data as well as the test data?

We will need the following property for later to ensure that the instrumental variable loss part of the Anchor Regression equation (the "penalty"-term) is set to zero by b_{IV}.

Definition 3.3.15. *The projectability condition is fulfilled if:*

$$\text{rank}(\text{Cov}_{\text{train}}(A, X)) = \text{rank}(\text{Cov}_{\text{train}}(A, X)|\, \text{Cov}_{\text{train}}(A, Y))$$

Here, $\text{Cov}_{\text{train}}(A, X)|\, \text{Cov}_{\text{train}}(A, Y)$ *is a* $q \times (d + 1)$ *matrix (it is* $\text{Cov}_{\text{train}}(A, X)$ *extended by a column consisting of* $\text{Cov}_{\text{train}}(A, Y)$

Remark 3.3.16. *In imprecise language, the projectability condition states that once we have information on the correlation (covariance) between* A *and* X*, information on the correlation (covariance) between* A *and* Y *is already contained in the information we have on the correlation (covariance) between* A *and* X*.*

Now I will show that it's indeed the case that b_{IV} sets its objective function to 0 in case the projectability condition holds. The next lemma is an extension from the mentioned result in the original paper [9, p.13 Section Section 3 Lemma 2]. All the extensions followed from the proof as found at [9, p.32 Section 8.9]. I worked out its details in the proof below:

Lemma 3.3.17. *The following are equivalent:*

1. $\min_{b \in \mathbb{R}^d} \mathbb{E}_{\text{train}}([\text{P}_A(Y - X^T b)]^2) = 0$
2. $\exists \tilde{b} \in \mathbb{R}^d$ *with* $\mathbb{E}_{\text{train}}(Y - X^T \tilde{b}|A) \overset{a.s.}{=} 0$
3. $\exists \tilde{b} \in \mathbb{R}^d$ *with* $\text{Cov}_{\text{train}}(A, Y) = \text{Cov}_{\text{train}}(A, X)\tilde{b}$
4. $\text{rank}(\text{Cov}_{\text{train}}(A, X)) = \text{rank}(\text{Cov}_{\text{train}}(A, X)|\, \text{Cov}_{\text{train}}(A, Y))$

Proof. (1) $\iff$ (2): Follows directly.
(2) $\iff$ (3): By lemma 3.2.3 we have that:

$$\mathbb{E}_{\text{train}}(Y|A) = A^T(\mathbb{E}_{\text{train}}(AA^T))^{-1}\, \mathbb{E}_{\text{train}}(AY)$$
$$\mathbb{E}_{\text{train}}(X^T|b) = A^T(\mathbb{E}_{\text{train}}(AA^T))^{-1}\, \mathbb{E}_{\text{train}}(AX^T)$$

This means that we write:

$$\mathbb{E}_{\text{train}}(Y - X^T \tilde{b}|A) = A^T(\mathbb{E}_{\text{train}}(AA^T))^{-1}[\mathbb{E}_{\text{train}}(AY) - \mathbb{E}_{\text{train}}(AX^T)\tilde{b}] \tag{3.3.2}$$

Assuming (2), we can hence write (above equation multiplied by A):

$$AA^T(\mathbb{E}_{\text{train}}(AA^T))^{-1}[\mathbb{E}_{\text{train}}(AY) - \mathbb{E}_{\text{train}}(AX^T)\tilde{b}] \overset{\text{a.s.}}{=} 0$$

Which implies :

$$\mathbb{E}_{\text{train}}(AA^T)(\mathbb{E}_{\text{train}}(AA^T))^{-1}[\mathbb{E}_{\text{train}}(AY) - \mathbb{E}_{\text{train}}(AX^T)\tilde{b}] = 0$$

So that:

$$\mathbb{E}_{\text{train}}(AY) = \mathbb{E}_{\text{train}}(AX^T)\tilde{b}$$

As, by model assumption X and Y are centered around 0, we thus obtain:

$$\text{Cov}_{\text{train}}(A, X)\tilde{b} = \text{Cov}_{\text{train}}(A, Y)$$

(3) implies (2) follows from expression 3.3.2 and using again that $\mathbb{E}_{\text{train}}(X)$ and $\mathbb{E}_{\text{train}}(Y)$ are both zero.
(3) $\iff$ (4): Is a straight-forward property of the column space of a matrix. □

Remark 3.3.18. *This lemma also fits in the (imprecise) logic of remark 3.3.16: when we look at $\mathbb{E}_{\text{train}}(X^T|A)$ it should (by the projectability property) already contain all the information about $\mathbb{E}_{\text{train}}(Y|A)$: hence we can find a b that fits perfectly.*

I will first make an analysis for the $\gamma = \infty$ case regarding replicability. For this we will assume the following setting below.

Setting 3.3.19. *We consider two different data-generating distributions. The training data (under $\mathbb{P}_{\text{train}}$) is distributed as follows:*

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \epsilon + v$$
$$v = M\delta, \delta = \kappa A + \xi$$

Here: $\mathbb{E}_{\text{train}}(\xi) = 0, \mathbb{E}_{\text{train}}(X) = 0, \mathbb{E}_{\text{train}}(Y) = 0$, $\xi \perp\!\!\!\perp \epsilon$, $\kappa \neq 0$ and $A \perp\!\!\!\perp (\epsilon, \xi)$
The test data (under $\mathbb{P}_{\text{train}}$) is distributed as follows:

$$\begin{pmatrix} X' \\ Y' \\ H' \end{pmatrix} = B \begin{pmatrix} X' \\ Y' \\ H' \end{pmatrix} + \epsilon' + v'$$
$$v' = M\delta', \delta' = \kappa' A' + \xi'$$

Here: $\mathbb{E}_{\text{test}}(\xi') = 0, \mathbb{E}_{\text{test}}(X') = 0, \mathbb{E}_{\text{test}}(Y') = 0$, $\xi' \perp\!\!\!\perp \epsilon'$, $\kappa' \neq 0$ and $A' \perp\!\!\!\perp (\epsilon', \xi')$

- *v' and A' can have arbitrarily different distributions from v and A but we do assume that the dimensions remain the same.*
- *Note that B and M (deterministic) are the same in both models and we furthermore assume:*

$$\text{Cov}_{\text{test}}(\epsilon') = L\,\text{Cov}_{\text{train}}(\epsilon), L > 0$$
$$\mathbb{E}_{\text{test}}(\epsilon') = \mathbb{E}_{\text{train}}(\epsilon) = 0$$

- *Assume* $\text{Id} - B$ *is invertible.*

Remark 3.3.20. *Roughly speaking, the models in the training and test data differ by arbitrary shifts in* $\text{span}(M)$ *and a scalar factor in the noise distribution as* $(X, Y, H)^T = (\text{Id} - B)^{-1}(\epsilon + v)$.

Observe that setting 3.3.19 both the training as well as the test data can be written in the form of setting 3.2 (the general setting for Anchor Regression as described at the start). Hence every (previous) result for setting 3.2 proven before also holds for setting 3.3.19.

In this setting we will consider the notion of replicability with respect to the training and test data specifically for the cases $\gamma = \infty$ and $\gamma = 0$.
The next lemma will come in handy for proving a statement on replicability later.

Lemma 3.3.21. *Assume the projectability condition. Then*

$$b^\infty = \arg\min_{b \in I} \mathbb{E}_{\text{train}}([Y - X^T b]^2).$$

Here, $I = \{b \in \mathbb{R}^d : \mathbb{E}_{\text{train}}(Y - X^T b | A) \overset{a.s.}{=} 0\}$

Proof. By lemma 3.3.17, $J = \{b \in \mathbb{R}^d : \mathbb{E}_{\text{train}}([\text{P}_A(Y - X^T b)]^2) = 0\}$ is non-empty. Hence, as $I = J$, I is also non-empty (I defined as in the statement of the current lemma). By lemma 3.3.5 we thus obtain:

$$\begin{aligned} b^\infty &= \arg\min_{b \in J} \mathbb{E}_{\text{train}}([(\text{Id} - \text{P}_A)(Y - X^T b)]^2) \\ &\overset{(1)}{=} \arg\min_{b \in J} \mathbb{E}_{\text{train}}([Y - X^T b]^2) \\ &= \arg\min_{b \in I} \mathbb{E}_{\text{train}}([Y - X^T b]^2) \end{aligned}$$

(1): $\forall b \in J$ it holds that $\mathbb{E}_{\text{train}}([\text{P}_A(Y - X^T b)]^2) = 0$. Hence:
$\mathbb{E}_{\text{train}}([(\text{Id} - \text{P}_A)(Y - X^T b)]^2) = \mathbb{E}_{\text{train}}([(\text{Id} - \text{P}_A)(Y - X^T b)]^2) + \mathbb{E}_{\text{train}}([\text{P}_A(Y - X^T b)]^2) = \mathbb{E}_{\text{train}}([Y - X^T b]^2)$ □

The following result was first mentioned at [9, p.14 Section 3.1]. I present here my own proof which is shorter than the one in the original paper [9, p.33 Section 8.10]

Theorem 3.3.22. *Assume the setting as established above,* $\mathbb{E}_{\text{train}}(AA^T), \mathbb{E}_{\text{test}}(A'(A')^T)$ *are both invertible and the projectability condition holds (on the training data). In that case:*

$$b^\infty = b'^\infty$$

Proof. This proof will consist of 3 steps:

1. Show that the projectability property holds for the test data
2. From lemma 3.3.21, show that $I = \{b \in \mathbb{R}^d : \mathbb{E}_{\text{test}}(Y' - X'^T b | A') \overset{\text{a.s.}}{=} 0\} =: I'$
3. In the context of lemma 3.3.21, show that we are optimizing over the same objective function for b^∞ and for b'^∞.

As the projectability condition holds for the training data, we know that lemma 3.3.21 holds with $I \neq \emptyset$. In case I show that $I = I'$: $I' \neq \emptyset$ which means that the projectability condition also holds for the test data by lemma 3.3.17. To that end, showing (2) implies (1) as well.
Proof of (2): Let $b \in I$. Then following remark 3.3.20, we have:

$$0 \overset{\text{a.s.}}{=} \mathbb{E}(Y - X^T b|A) = w_b \kappa M A$$

Here: $w_b = [(\mathrm{Id} - B)^{-1}]_{d+1,\cdot} - b^T[(\mathrm{Id} - B)^{-1}]_{1:d,\cdot}$. From this we obtain:

$$\begin{aligned} w_b \kappa M A &\overset{\text{a.s.}}{=} 0 \implies \\ w_b \kappa M A A^T &\overset{\text{a.s.}}{=} 0 \implies \\ w_b \kappa M \, \mathbb{E}_{\text{train}}(AA^T) &= 0 \iff \\ w_b M &= 0 \end{aligned}$$

The last step above uses the assumption that $\mathbb{E}_{\text{train}}(AA^T)$ is invertible and $\kappa \neq 0$. Hence it holds that $w_b M A' = 0$, which implies that $b \in I'$. The reverse direction is analogous.
Proof of (3): By lemma 3.3.21, it suffices to show that:

$$\arg\min_{b \in I} \mathbb{E}_{\text{train}}([Y - X^T b]^2) = \arg\min_{b \in I} \mathbb{E}_{\text{test}}([Y' - X'^T b]^2)$$

We know that:

$$\begin{aligned} Y - X^T b &= w_b((\epsilon + M\xi) + \kappa M A) \\ Y' - X'^T b &= w_b((\epsilon' + M\xi) + \kappa M A') \end{aligned}$$

As $w_b M = 0$ for any $b \in I$, we see that:

$$\arg\min_{b \in I} \mathbb{E}_{\text{train}}([Y - X^T b]^2) = \arg\min_{b \in I} \mathbb{E}_{\text{train}}([w_b \epsilon]^2) = \arg\min_{b \in I} w_b \, \mathrm{Cov}_{\text{train}}(\epsilon) w_b^T$$

In a similar fashion:

$$\arg\min_{b \in I} \mathbb{E}_{\text{test}}([Y' - X'^T b]^2) = \arg\min_{b \in I} \mathbb{E}_{\text{test}}([w_b \epsilon']^2) = \arg\min_{b \in I} w_b L \, \mathrm{Cov}_{\text{train}}(\epsilon) w_b^T$$

In the last step above, it was used that $\mathrm{Cov}_{\text{test}}(\epsilon') = L \, \mathrm{Cov}_{\text{train}}(\epsilon)$. □

A similar result can be established for b^0 with fewer extra restrictions on the training and test data. This is result was not stated in the original paper and serves as an addition.

Theorem 3.3.23. *Under the assumption that* $\mathrm{Cov}_{\text{train}}(\xi) = L \, \mathrm{Cov}_{\text{test}}(\xi)$, *(where* $L > 0$ *is the same* L *as seen in the setting 3.3.19), it holds that:*

$$b^0 = b'^0$$

Proof. As seen in the proof of theorem 3.3.22: $Y - X^T b = w_b(\epsilon + M\xi + \kappa M A)$. From this it follows that:

$$(\mathrm{Id} - \mathrm{P}_A)(Y - X^T b) = w_b(\epsilon + M\xi)$$

And so it holds that:

$$\mathbb{E}_{\text{train}}([(\text{Id} - \text{P}_A)(Y - X^T b)]^2) = w_b(\text{Cov}_{\text{train}}(\epsilon) + M\,\text{Cov}_{\text{train}}(\xi)M^T)$$

Above, it was used that $\mathbb{E}_{\text{train}}(\epsilon) = 0$, $\mathbb{E}_{\text{train}}(\xi)$ and $\epsilon \perp\!\!\!\perp \xi$. Similarly:

$$\begin{aligned}\mathbb{E}_{\text{test}}([(\text{Id} - \text{P}_{A'})(Y' - X'^T b)]^2) &= w_b(\text{Cov}_{\text{test}}(\epsilon') + M\,\text{Cov}_{\text{test}}(\xi')M^T) \\ &= L w_b(\text{Cov}_{\text{train}}(\epsilon) + M\,\text{Cov}_{\text{train}}(\xi)M^T)\end{aligned}$$

As $L > 0$, we can conclude (using the definition of b^0): $b^0 = b'^0$ □

3.3.6. Anchor stability

For the remainder of this subsection, we use setting 3.2 and assume existence and uniqueness for b^γ, $\gamma \in [0, \infty]$. See section 3.2 for sufficient conditions.

We will now consider the notion of anchor stability. This is defined below

Definition 3.3.24. *We will call the data $(X, Y, A, H)^T$ (under $\mathbb{P}_{\text{train}}$) anchor stable in case $b^\infty = b^\gamma \ \forall \gamma \in [0, \infty)$*

Anchor stability will give us certain predictive stability and replicability properties. More specifically, it will turn out that, under certain conditions, we can use perturbed data instead of the training data to compute b^γ.

The first result shows that we have anchor stability if the two endpoints of $(b^\gamma)_{\gamma \in [0,\infty]}$ agree. This result (including proof) originates from [9, p.15 Section 3.2][9, p.34 Section 8.11]

Proposition 3.3.25. *If $b^0 = b^\infty$, then $b^0 = b^\gamma \ \forall \gamma \in (0, \infty)$.*

Proof. Define $f(b) = \mathbb{E}_{\text{train}}([\text{P}_A(Y - X^T b)]^2)$ and $g(b) = \mathbb{E}_{\text{train}}([(\text{Id} - \text{P}_A)(Y - X^T b)]^2)$. By the assumption of b^∞ existing and $b^\infty = b^0$ we have that:

$$\partial_b f(b^\infty) = \partial_b f(b^0) = \partial_b g(b^0) = \partial_b g(b^\infty) = 0$$

The objective function for Anchor Regression for $\gamma \geq 0$ can be expressed as:

$$g(b) + \gamma f(b)$$

hence, it has a zero derivative at b^0. As we assumed that b^γ is the unique global minimum and the objective function is convex in b: $b^\gamma = b^0$. □

In the case we have anchor stability, and some other conditions such as projectability, it turns out that optimizing over the training data (OLS-estimator) is equivalent to optimizing over certain shifted data. This next result can also be found at [9, p.15 Section 3.2]. Below, I present my own proof.

Theorem 3.3.26. *(Anchor stability implies predictive stability and replicability)*
Assume the setting as described in section 3.2, the projectability condition and $\mathbb{E}_{\text{train}}(AA^T)$ is invertible. If $b^0 = b^\infty$, then for all random vectors v for which it holds that: $v = Mx$ (where x is random) and $\mathbb{E}_v(\epsilon v^T) = 0$, we have that:

1. *(Predictive stability)* $\mathbb{E}_{\text{train}}((Y - X^T b^0)^2) = \mathbb{E}_v((Y - X^T b^0)^2)$
2. *(Replicability)* $b^0 = \arg\min_{b \in \mathbb{R}^d} \mathbb{E}_v((Y - X^T b)^2)$

Proof. From the projectability condition and $\mathbb{E}_{\text{train}}(AA^T)$ being invertible, we have in previous results established that from lemma 3.3.21, it follows that: $w_{b^\infty} M = 0$ (where $w_b = [(\text{Id} - B)^{-1}]_{d+1,\cdot} - b^T[(\text{Id} - B)^{-1}]_{1:d,\cdot}$). Hence, as $b^0 = b^\infty$: $w_{b^0} M = 0$. It follows that under $\mathbb{P}_{\text{train}}$:

$$Y - X^T b^0 = w_{b^0}(\epsilon + MA) = w_{b^0}\epsilon$$

and that under $\mathbb{P}_{\text{test}}$:

$$Y - X^T b^0 = w_{b^0}(\epsilon + v) = w_{b^0}(\epsilon + Mx) = w_{b^0}\epsilon$$

As, by assumption, ϵ has the same distribution under $\mathbb{P}_{\text{train}}$ as under $\mathbb{P}_{\text{test}}$, which establishes the (Predictive stability) result.
The (Replicability) result can be seen as follows:

$$\begin{aligned}
\mathbb{E}_v([Y - X^T b]^2) &= \mathbb{E}_v([w_b(\epsilon + v)]^2) \\
&= \mathbb{E}_v([w_b\epsilon]^2) + \mathbb{E}_v([w_b v]^2) \\
&\geq \mathbb{E}_v([w_b\epsilon]^2) \\
&= \mathbb{E}_{\text{train}}([(\text{Id} - \text{P}_A)(Y - X^T b)]^2) \\
&\geq \mathbb{E}_{\text{train}}([(\text{Id} - \text{P}_A)(Y - X^T b^0)]^2) \\
&= \mathbb{E}_{\text{train}}([w_{b^0}\epsilon]^2) \\
&= \mathbb{E}_v([w_{b^0}\epsilon]^2) \\
&= \mathbb{E}_v([w_{b^0}(\epsilon + v)]^2) = \mathbb{E}_v([Y - X^T b^0]^2)
\end{aligned}$$

Above, it was used that $\epsilon \perp\!\!\!\perp v$, $\mathbb{E}_{\text{train}}(\epsilon) = 0$, ϵ has the same distribution under $\mathbb{P}_{\text{train}}$ as under $\mathbb{P}_v$, $(\text{Id} - \text{P}_A)(Y - X^T b) = w_b\epsilon$ and that $w_{b^0} v = 0$ (as $v = Mx$). □

Remark 3.3.27.

- *Part 1 of theorem 3.3.26 implies that the risk of b^γ is constant across various $\mathbb{P}_v$ distributions as long as $v \in \text{span}(M)$. This can be seen as a form of predictive stability across a range of distributions.*
- *Part 2 of theorem 3.3.26 together with proposition 3.3.25 imply that running a regression on perturbed data sets in the population case returns the same coefficients as the ones computed in the training data as long as v lies in $\text{span}(M)$. This can be seen as a form of replicability.*
- *At surface level, the first part is not enough to prove the second part. There could namely still exist b in $\mathbb{R}^d$ where the risk over the perturbed data is unequal to the risk over the training data and could hence outperform b^γ. By the second statement it doesn't turn out to be the case.*

We will now consider a special case of setting 3.2 where we can establish that the anchor coefficients are equal to the causal effect per unit intervention on X in case we have anchor stability. For this, the following linear structural equation model is considered:

Setting 3.3.28.

$$\begin{aligned}
X_1 &= \epsilon_1 + \sum_{j=1}^{q} M_{1,j} A_j \\
X_2 &= B_{2,1} X_1 + \epsilon_2 + \sum_{j=1}^{q} M_{2,j} A_j \\
&\vdots \\
X_d &= \sum_{j=1}^{d-1} B_{d,i} X_i + \epsilon_d + \sum_{j=1}^{q} M_{d,j} A_j \\
Y &= \sum_{j=1}^{d} B_{d+1,i} X_i + \epsilon_{d+1}
\end{aligned}$$

Here, $B_{i,j}$ and $M_{i,j}$ are constants in $\mathbb{R}$, $A := (A_1, ..., A_q)^T \in \mathbb{R}^q, \epsilon := (\epsilon_1, ..., \epsilon_{d+1})^T \in \mathbb{R}^{d+1}$ with $A \perp\!\!\!\perp \epsilon$ and $\epsilon_i \perp\!\!\!\perp \epsilon_j \quad \forall i \neq j$. $\mathbb{E}(\epsilon) = 0$, $Y \in \mathbb{R}$. Also assume consistency, i.e. $Y = Y^X$.

Corollary 3.3.29. *Under the setting as established above, for $x := (x_1, ..., x_d)$:*

$$b^0 = b^\infty \implies b^0 = b^\infty = \partial_x \mathbb{E}(Y^x)$$

Proof. As the setting assumed for this corollary is a special case of the one seen in section 3.2, all previously established theorems also hold for this setting. Hence by proposition 3.3.25, $b^0 = b^\infty$ gives that $b^0 = b^\infty = b^1$ where b^1 is equivalent to the population OLS estimator of the effect of X on Y. Hence it suffices to show: $b^1 = \partial_x \mathbb{E}(Y^x)$.
For the current setting it holds that:

$$Y^x = \sum_{i=1}^{d} B_{d+1,i} x_i + \epsilon_{d+1}$$

which gives:

$$\mathbb{E}(Y^x) = \sum_{i=1}^{d} B_{d+1,i} x_i =: x^T \beta$$

Hence: $\partial_x \mathbb{E}(Y^x) = \beta$. From lemma 3.3.3 we know that:

$$\begin{aligned}
b^1 &= [\mathbb{E}(XX^T)]^{-1} \mathbb{E}(XY) \\
&= [\mathbb{E}(XX^T)]^{-1} \mathbb{E}(X(X^T\beta + \epsilon_{d+1})) \\
&= \beta + [\mathbb{E}(XX^T)]^{-1} \mathbb{E}(X\epsilon_{d+1}) \\
&\overset{(1)}{=} \beta
\end{aligned}$$

Hence in conclusion: $\partial_x \mathbb{E}(Y^x) = \beta = b^1$.
(1): Using repeated substitution of X_i in X_{i+1}, one can see that X can be written as a function of A plus a function of $(\epsilon_1, ..., \epsilon_d)$. Hence by independence: $[\mathbb{E}(X\epsilon_{d+1})]_i = \mathbb{E}(X_i \epsilon_{d+1}) = 0$. □

3.4. Anchor regression estimators

Thus far we have only considered the population version of Anchor Regression. In this section, we will consider a finite sample estimator for b^γ.

Notation: In this section we will consider n observations iid with respect to $(X, Y, A)^T = (X^{(1)}, ..., X^{(d)}, Y, A^{(1)}, ..., A^{(q)})^T$. $\mathbf{X} : n \times d$, $\mathbf{Y} : n \times 1$, $\mathbf{A} : n \times q$ denote the n observations stored as rows. We will also assume in this section that b^γ is exists and is unique (as established in lemma 3.3.3) and that A is continuous unless indicated otherwise.

We will use the following estimator for b^γ (where I assume that $d < n$):

Definition 3.4.1. $\hat{b}^\gamma = \underset{b \in \mathbb{R}^d}{\arg\min}(\|(\mathrm{Id} - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2 + \gamma \|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2)$.
Here, $\Pi_{\mathbf{A}} = \mathrm{P}_{\mathrm{col}(\mathbf{A})}$

Remark 3.4.2. *When pre-processing the data, it is recommended to center* $\mathbf{X}$ *and* $\mathbf{Y}$ *as we have previously assumed* X *and* Y *to have mean 0 in the population version.*

It turns out that one can solve for $\hat{b}^\gamma$ via a substitution and then application of OLS. This result was first mentioned at [9, p.15 Section 4.1] without proof.

Lemma 3.4.3. *Define* $\tilde{X} := (\mathrm{Id} - \Pi_{\mathbf{A}})\mathbf{X} + \sqrt{\gamma}\Pi_{\mathbf{A}}\mathbf{X}$ *and* $\tilde{Y} := (\mathrm{Id} - \Pi_{\mathbf{A}})\mathbf{Y} + \sqrt{\gamma}\Pi_{\mathbf{A}}\mathbf{Y}$. *Then:*

$$\hat{b}^\gamma = \underset{b \in \mathbb{R}^d}{\arg\min} \left\| \tilde{Y} - \tilde{X}b \right\|_2^2$$

Proof.

$$\begin{aligned} \|(\mathrm{Id} - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2 = \|(\mathrm{Id} - \Pi_{\mathbf{A}})\mathbf{Y} - (\mathrm{Id} - \Pi_{\mathbf{A}})\mathbf{X}b\|_2^2 = \\ \|(\mathrm{Id} - \Pi_{\mathbf{A}})\mathbf{Y} + \sqrt{\gamma}\Pi_{\mathbf{A}}\mathbf{Y} - \sqrt{\gamma}\Pi_{\mathbf{A}}\mathbf{Y} - [(\mathrm{Id} - \Pi_{\mathbf{A}})\mathbf{X} + \sqrt{\gamma}\Pi_{\mathbf{A}}\mathbf{X} - \sqrt{\gamma}\Pi_{\mathbf{A}}\mathbf{X}b]\|_2^2 = \\ \left\| \tilde{Y} - \tilde{X}b + \sqrt{\gamma}\Pi_{\mathbf{A}}(\mathbf{X}b - \mathbf{Y}) \right\|_2^2 = \end{aligned}$$

Hence:

$$\hat{b}^\gamma = \underset{b \in \mathbb{R}^d}{\arg\min}(\left\| \tilde{Y} - \tilde{X}b - \sqrt{\gamma}\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b) \right\|_2^2 + \|\sqrt{\gamma}\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2)$$

Using that $\|\cdot\|_2^2 = \langle \cdot, \cdot \rangle$:

$$\hat{b}^\gamma = \underset{b \in \mathbb{R}^d}{\arg\min}(\left\| \tilde{Y} - \tilde{X}b \right\|_2^2 - 2\langle \tilde{Y} - \tilde{X}b, \sqrt{\gamma}\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b) \rangle + 2\|\sqrt{\gamma}\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2)$$

Now using that for any x: $\langle \Pi_{\mathbf{A}}x, (\mathrm{Id} - \Pi_{\mathbf{A}})x \rangle = 0$ together with the $\tilde{Y}$ and $\tilde{X}$ expressed in terms of $\mathbf{Y}$ and $\mathbf{X}$ respectively, we obtain the identity as stated in the lemma. □

Remark 3.4.4.

- *In this sense Anchor Regression can be seen as a two-step procedure:*
  1. *Generate perturbed data* $(\tilde{X}, \tilde{Y})^T$ *given perturbation strength* γ.

2. Run OLS on perturbed data set

As for population Anchor Regression, we can impose conditions as to ensure existence and uniqueness of $\hat{b}^\gamma$. The proof of the lemma below is analogous to the one seen in lemma 3.3.3.

Lemma 3.4.5. *In case $\tilde{X}^T\tilde{X}$ (as defined in lemma 3.4.3) is positive definite: $\hat{b}^\gamma$ exists and is unique with*

$$\hat{b}^\gamma = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y}$$

The following lemma shows consistency for $\hat{b}^\gamma$. This result was first mentioned at [9, p.16 Section 4.1] but was not proven there.

Lemma 3.4.6. *Assume that $\mathbb{E}_{\text{train}}(AA^T)$ is invertible. Then, for any $\gamma \in [0, \infty)$:*

$$\hat{b}^\gamma \xrightarrow{\mathbb{P}} b^\gamma$$

Proof. As $(X_i, Y_i, A_i)^T \overset{\text{iid}}{\sim} (X, Y, A)^T$ for $1 \leq i \leq n$ by assumption and X, Y and A all have finite second moments as a consequence of setting 3.2, the law of large numbers gives:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n X_iY_i &\xrightarrow{\mathbb{P}} \mathbb{E}_{\text{train}}(X_iY_i) \\
\frac{1}{n}\sum_{i=1}^n X_iA_i^T &\xrightarrow{\mathbb{P}} \mathbb{E}_{\text{train}}(X_iA_i^T) \\
\frac{1}{n}\sum_{i=1}^n A_iA_i^T &\xrightarrow{\mathbb{P}} \mathbb{E}_{\text{train}}(A_iA_i^T) \\
\frac{1}{n}\sum_{i=1}^n A_iY_i &\xrightarrow{\mathbb{P}} \mathbb{E}_{\text{train}}(A_iY_i) \\
\frac{1}{n}\sum_{i=1}^n X_iX_i^T &\xrightarrow{\mathbb{P}} \mathbb{E}_{\text{train}}(X_iX_i^T)
\end{aligned}$$

And as

$$\begin{aligned}
\tilde{X}^T\tilde{Y} = \sum_{i=1}^n X_iY_i - 2(1-\sqrt{\gamma})(\sum_{i=1}^n X_iA_i^T)(\sum_{i=1}^n A_iA_i^T)^{-1}\sum_{i=1}^n A_iY_i \\
+ (1-\sqrt{\gamma})^2(\sum_{i=1}^n X_iA_i^T)(\sum_{i=1}^n A_iA_i^T)^{-1}\sum_{i=1}^n A_iY_i
\end{aligned}$$

It holds that (using continuous mapping):

$$\begin{aligned}
\frac{1}{n}\tilde{X}^T\tilde{Y} \xrightarrow{\mathbb{P}} \mathbb{E}_{\text{train}}(X_iY_i) - 2(1-\sqrt{\gamma})\,\mathbb{E}_{\text{train}}(X_iA_i^T)(\mathbb{E}_{\text{train}}(A_iA_i^T))^{-1}\,\mathbb{E}_{\text{train}}(A_iY_i) \\
+ (1-\sqrt{\gamma})^2\,\mathbb{E}_{\text{train}}(X_iA_i^T)(\mathbb{E}_{\text{train}}(A_iA_i^T))^{-1}\,\mathbb{E}_{\text{train}}(A_iY_i)
\end{aligned}$$

Similarly, we can establish:

$$\frac{1}{n}\tilde{X}^T\tilde{X} \xrightarrow{\mathbb{P}} \mathbb{E}_{\text{train}}(X_iX_i^T) - 2(1-\sqrt{\gamma})\,\mathbb{E}_{\text{train}}(X_iA_i^T)(\mathbb{E}_{\text{train}}(A_iA_i^T))^{-1}\,\mathbb{E}_{\text{train}}(A_iX_i^T) + (1-\sqrt{\gamma})^2\,\mathbb{E}_{\text{train}}(X_iA_i^T)(\mathbb{E}_{\text{train}}(A_iA_i^T))^{-1}\,\mathbb{E}_{\text{train}}(A_iX_i^T)$$

By lemma 3.2.3 and then lemma 3.3.3 it follows that: $\hat{b}^\gamma \xrightarrow{\mathbb{P}} b^\gamma$ □

Remark 3.4.7. *The original paper warns that confounding effects may result in no asymptotic normality properties for $\hat{b}^\gamma$ [9, p.16 Section 4.1].*

Remark 3.4.8. *In case $d > n$ (high-dimensional Anchor Regression), $\hat{b}^\gamma$ will not have an unique solution. To obtain an unique solution and obtain a solution for $\hat{b}^\gamma$ that sets coefficients to 0 if they have too little impact on the data, one might add a LASSO penalty term to the objective function as seen in definition 3.4.1.*

3.5. Finite sample bound for discrete anchors

In the previous section we assumed that A is a continuous random variable. We will now consider Anchor Regression in case A is discrete and can only take a finite number of values. Let $\mathcal{A}$ be the levels of the A (the possible values A can take). Assume that all levels are given equal weight (i.e. $\forall a_1, a_2 \in \mathcal{A} : \mathbb{P}(A = a_1) = \mathbb{P}(A = a_2) = \frac{1}{|\mathcal{A}|}$). We still assume setting 3.2.

Lemma 3.5.1. *The objective function for the discrete population Anchor Regression can be expressed as follows:*

$$R(b) := \mathbb{E}_{\text{train}}([Y - X^Tb - \mathbb{E}_{\text{train}}(Y - X^T|A)]^2) + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\mathbb{E}_{\text{train}}(Y - X^Tb|A = a))^2$$

Proof. For the expression of $R(b)$ only the final term differs from the population Anchor Regression expression as defined in 3.3.1.

$$\begin{aligned}\mathbb{E}_{\text{train}}([\mathrm{P}_A(Y - X^Tb)]^2) &= \\ \mathbb{E}_{\text{train}}([\mathbb{E}(Y - X^Tb|A)]^2) &= \\ \sum_{a\in\mathcal{A}}(\mathbb{E}_{\text{train}}(Y - X^Tb|A = a))^2\mathbb{P}(A = a) &= \\ \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\mathbb{E}_{\text{train}}(Y - X^Tb|A = a))^2 &\end{aligned}$$

Here, the last equality is due to the fact that all levels are given equal weight. □

Remark 3.5.2.

- *It still holds that by theorem 3.3.8: $R(b) = \sup_{v\in C^\gamma} \mathbb{E}_v((Y - X^Tb)^2)$, where now C^γ can be reduced to v's with a discrete and uniform distribution [9, p.17 Section 4.3][9, p.32 Section 8.8].*
- *Anchor regression for A discrete has a quantiles interpretation under again some normality assumptions [9, p.32 Section 8.8].*

It will turn out that we have theoretical guarantees for the performance of the Anchor Regression estimator in comparison to the population Anchor Regression coefficients counterpart. For this, I will first introduce some notation and then define the Anchor Regression estimator for the A being discrete case.

Notation:

- n_a=number of observations at level $A = a$
- $n_{\min}$=$\min_{a \in \mathcal{A}} n_a$
- $\mathbf{X}^{(a)} : \mathbb{R}^{n_a} \times d$ denotes the observations at level $A = a$ (in rows).
- $\bar{\mathbf{X}}^{(a)} = \frac{1}{n_a} \sum_{i=0}^{n_a} \mathbf{X}_i^{(a)}$
- $\mathbf{Y}^{(a)}$ and $\bar{\mathbf{Y}}^{(a)}$ are similarly defined as above.

With this notation we can define the Anchor Regression estimator, $\hat{b}^{\gamma,\lambda}$, as follows:[2]:

$$\arg\min_{b \in \mathbb{R}^d} \left(\frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{Y}_i^{(a)} - \bar{\mathbf{Y}}_i^{(a)} - (\mathbf{X}_{i.}^{(a)} - \bar{\mathbf{X}}_i^{(a)})b)^2 + \frac{\gamma}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} (\bar{\mathbf{Y}}^{(a)} - \bar{\mathbf{X}}^{(a)}b)^2 + 2\lambda \|b\|_1 \right.$$

We will now define the anchor compatibility constant. In the next theorem, it will become clear that it plays the role of a threshold for which, if met with large enough probability, we can (with a certain probability) guarantee properties for the Anchor Regression estimator in connection with population Anchor Regression.

Definition 3.5.3. *Let $S \subset \{1, .., d\}$ and $L > 0$. Then the anchor compatibility constant with respect to S and L is defined as:*

$$\hat{\phi}^2(L, S) = \min_{\|b_S\|_1 = 1, \|b_{-S}\|_1 \leq L} \left(|S| \sum_{a \in \mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} ((\mathbf{X}_{i.}^{(a)} - \bar{\mathbf{X}}^{(a)})b)^2 + \frac{\gamma}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} (\bar{\mathbf{X}}^{(a)}b)^2\right)$$

For the following result we allow for $d = d_n > n \to \infty$, $\mathcal{A} = \mathcal{A}_n$, $M = M_n$ and $b^\gamma = b_n^\gamma$ (i.e. they can have dependence on n, and so the model distribution from section 3.2 now depends on n).

Theorem 3.5.4. *[9, p.18 Section 4.3 Theorem 5](Connection finite sample estimator with population Anchor Regression)*
Assume the setting from section 3.2 and ϵ is centered multivariate normal. Moreover, assume that $(\mathbf{X}_{i,.}^{(a)}, \mathbf{Y}_i^{(a)}) \overset{iid}{\sim} (X, Y)|A = a$ under $\mathbb{P}_{\text{train}}$. Fix $\gamma > 0$ and assume that $\hat{\phi}^2(8, S^*) \geq C$ for some constant $C > 0$ with probability $1 - \delta$ and that $S^* \neq \emptyset$. Choose $t \geq 0$ such that: $|S^*|^2 \frac{t + \log(d) + \log(|\mathcal{A}|)}{n_{\min}} \leq C'$ for some $C' > 0$. Then for $\lambda \geq C\sqrt{\frac{t + \log(d) + \log(|\mathcal{A}|))}{n_{\min}}}$ with probability exceeding $1 - 10\exp(-t) - \delta$:,

$$R(\hat{b}^{\gamma,\lambda}) \leq R(b^\gamma) + C'\lambda^2|S^*|$$

Here, $R(b)$ is as defined in lemma 3.5.1 (i.e. the population Anchor Regression objective function). Furthermore, the constants $C, C' < \infty$ depend on: $\max_{k \in \{1,..d\}} \text{Var}_{\text{train}}(X^{(k)})$, $\text{Var}_{\text{train}}(Y - X^T b^\gamma), \max_{a \in \mathcal{A}} \|\mathbb{E}_{\text{train}}(X|A = a)\|_\infty$, $\max_{a \in \mathcal{A}} |\mathbb{E}_{\text{train}}(Y - X^T b^\gamma | A = a)$ and γ.

[2]Note that we are now in the case that A is discrete. As we can't work with $\Pi_{\mathbf{A}}$ (which was initially defined as the projection onto the column space of $\mathbf{A}$), instead of this projection, we average over all observations on a certain level $A = a$.

Remark 3.5.5.

- *The 8 has no theoretical value: just chosen to simplify result [9, p.17 Section 4.3 below Theorem 5].*
- *For $\gamma = 1$, $b^{\lambda,\gamma}$ coincides with LASSO. The result of theorem 3.5.4 is comparable to established risk bounds for LASSO (under similar settings) in case $\frac{n}{n_{\min}}$ is bounded [9, p.17 Section 4.3 below Theorem 5].*

4. Two Stage Curvature Identification (TSCI)

4.1. General idea behind TSCI

We again reconsider the very first example from this paper (see figure 1.1) where we now ignore the effect of X (magnesium) for simplicity. As in section 2, we have $(Y_i, Z_i^{(1)}, Z_i^{(2)}, D_i)_{i=1}^n$ sampled and assume a linear effect of the treatment. This time, we don't want to assume a linear model in $(Z^{(1)}, Z^{(2)})$ and/or a certain number of present valid IVs. It turns out that if we have prior information about the functional forms of the effect of $(Z^{(1)}, Z^{(2)}$ on Y and D (so the blue and red arrows respectively), we can still infer the treatment effect exactly (in an asymptotic sense).

For instance, suppose Y_i and D_i satisfy the following expressions:

$$\begin{aligned} Y_i &= \beta D_i + Z_i^{(2)} + e_i =: \beta D_i + g_1(Z_i^{(1)}, Z_i^{(2)}) + e_i & \mathbb{E}(e_i | Z_i^{(1)}, Z_i^{(2)}) = 0 \\ D_i &= (Z_i^{(1)})^5 + Z_i^{(2)} + \delta_i =: f_1(Z_i^{(1)}, Z_i^{(2)}) + \delta_i & \mathbb{E}(\delta_i | Z_i^{(1)}, Z_i^{(2)}) = 0 \end{aligned}$$

which implies that:

$$\mathbb{E}(Y_i | Z_i^{(1)}, Z_i^{(2)}) = \beta((Z_i^{(1)})^5 + Z_i^{(2)}) + Z_i^{(2)} \tag{4.1.1}$$

Suppose that we already had a suspicion beforehand (through for example expert knowledge) that the effect of $(Z_i^{(1)}, Z_i^{(2)})$ on Y (i.e. the effect of sugar and genetics on various cancer types), which are represented by the blue arrows in figure 1.1, might be a second degree polynomial or lower i.e. $g_1(z_1, z_2) \in \text{span}(z_1, z_2, z_1 z_2, z_1^2, z_2^2) =: \text{span}(\mathcal{V})$ and that $f_1(z_1, z_2)$ is a higher degree polynomial than degree 2. Then by applying the projection onto $\text{span}^\perp(\mathcal{V})$to (4.1.1) gives us:

$$\mathrm{P}_{\mathcal{V}}^\perp \, \mathbb{E}(Y_i | Z_i^{(1)}, Z_i^{(2)}) = \beta \mathrm{P}_{\mathcal{V}}^\perp f_1(Z^{(1)}, Z^{(2)})$$

which uses that $\mathrm{P}_{\mathcal{V}}^\perp g_1 = 0$. As $f_1 \notin \text{span}(\mathcal{V})$, it follows from (4.1.1) that:

$$\beta = \frac{\langle \mathrm{P}_{\mathcal{V}}^\perp \, \mathbb{E}(Y_i | Z_i^{(1)}, Z_i^{(2)}), \mathrm{P}_{\mathcal{V}}^\perp f_1(Z_i^{(1)}, Z_i^{(2)}) \rangle}{\langle \mathrm{P}_{\mathcal{V}}^\perp f_1(Z_i^{(1)}, Z_i^{(2)}), \mathrm{P}_{\mathcal{V}}^\perp f_1(Z_i^{(1)}, Z_i^{(2)}) \rangle}$$

Above, we can estimate f_1 by applying a machine learning (ML) method to the treatment equation.

Observe that the argument above would also have worked if we had suspected a first degree polynomial or lower for g_1. For finite sample estimations of the above process, there is a faster convergence (in probability) to β the more precise our suspicions are. Hence, along with an identification method for β, there is also a test introduced that provides us with the basis (which is a subset of our initial set provided by the user) that gives the most efficient estimation of g_1. This test will, by design, also give us an idea about which $Z^{(j)}$'s could be instrumental variables, by means of showing us which $Z^{(j)}$'s are in the most efficient basis for g_1.

4.2. General setting

Consider $(Y_i, D_i, Z_i, X_i) \overset{\text{iid}}{\sim} (Y, D, Z, X)$ for $i = 1, ..., n$ where:

- $Y_i \in \mathbb{R}$, outcome
- $D_i \in \mathbb{R}$, treatment
- $Z_i = (Z_i^{(j)})_{j=1,...,p_Z} \in \mathbb{R}^{p_Z}$, (possibly invalid) instrumental variables.
- $X_i = (X_i^{(j)})_{j=1,...,p_X} \in \mathbb{R}^{p_X}$, measured covariates.

$(e_i, \delta_i)_{i=1}^n$ (the errors) are sampled independently.

Definition 4.2.1.

- *The outcome model is defined as follows:*

$$Y_i = \beta D_i + g(Z_i, X_i) + \epsilon_i, \quad \mathbb{E}(\epsilon_i | X_i, Z_i) = 0$$

  $\beta \in \mathbb{R}$ *is the constant treatment effect on the outcome.*

- *Define* $h(Z_i, X_i) = g(Z_i, X_i) - \phi(X_i)$ *with* $\phi(X_i) = \mathbb{E}(g(Z_i, X_i)|X_i)$. *$h$ is called the violation function.*

Observe that we can write the outcome model in terms of the violation function:

$$Y_i = \beta D_i + h(Z_i, X_i) + \phi(X_i) + \epsilon_i, \quad \mathbb{E}(\epsilon_i | X_i, Z_i) = 0.$$

Definition 4.2.2. *The treatment model is defined as follows:*

$$D_i = f(Z_i, X_i) + \delta_i, \quad \mathbb{E}(\delta_i | Z_i, X_i) = 0$$

Remark 4.2.3. *Note that the treatment model is not a condition on the treatment as D_i can always be written as: $D_i = \mathbb{E}(D_i|Z_i, X_i) + (D_i - \mathbb{E}(D_i|Z_i, X_i))$.*

Definition 4.2.4. *In this context, $Z^{(j)}$ is called an instrumental variable if it is present in the expression of $f(Z_i, X_i)$ and not present in the expression of $g(Z_i, X_i)$.*

Remark 4.2.5. *When all $Z^{(j)}$ for $j = 1, ..., p_Z$ are valid IVs (i.e. $f(Z, X)$ depends on $Z^{(j)}$ while $g(Z, X)$ does not depend on $Z^{(j)}$ for all $j = 1, ..., p_Z$), we have that $h(Z, X) = 0$ i.e. no violation.*

Substituting the treatment model into the outcome model gives us the reduced form model as can be seen below:

Lemma 4.2.6 (Reduced form model)**.**
$Y_i = F(Z_i, X_i) + \epsilon_i + \beta\delta_i$, $F(Z_i, X_i) = \beta f(Z_i, X_i) + g(Z_i, X_i)$

4.3. Identification

The TSCI-method will not rely on which $Z^{(j)\prime}s$ are valid instrumental variables but rather on the difference in functional form between $g(z,x)$ and $f(z,x)$. This will be explained in the next bit.
Let (Ω,$\mathcal{F}$,$\mathbb{P}$) be a probability space and the domain for (Y,D,Z,X) which satisfies the outcome and treatment model. Define the Hilbert Space $\mathcal{S} := \{w(Z,X) : w : \Omega \to \mathbb{R} \text{ under some conditions}\}$ with (real) inner product $\langle \cdot,\cdot \rangle : \Omega \times \Omega \to \mathbb{R}$. Define $\mathcal{V} := \text{span}(v_1(Z,X),...,v_L(Z,X)) \subseteq \mathcal{S}$. Assume that $g(Z,X) \in \mathcal{V}$ while $f \in \mathcal{S}\backslash\mathcal{V}$. From the outcome model it follows that:

$$F(X,Z) = \mathbb{E}(Y|X,Z) = \beta f(X,Z) + g(X,Z).$$

Let $\mathrm{P}_{\mathcal{V}}$ be the projection matrix onto $\mathcal{V}$. Then it holds that:[1]

$$\mathrm{P}_{\mathcal{V}}^{\perp}\,\mathbb{E}(Y|X,Z) = \beta \mathrm{P}_{\mathcal{V}}^{\perp} f(X,Z)$$

As $f \notin \mathcal{V}$: $\left\|\mathrm{P}_{\mathcal{V}}^{\perp} f(X,Z)\right\| > 0$, it holds that:

$$\beta = \frac{\langle \mathrm{P}_{\mathcal{V}}^{\perp} F(X,Z), \mathrm{P}_{\mathcal{V}}^{\perp} f(X,Z)\rangle}{\langle \mathrm{P}_{\mathcal{V}}^{\perp} f(X,Z), \mathrm{P}_{\mathcal{V}}^{\perp} f(X,Z)\rangle}$$

Observe that in case that $f \in \mathcal{V}$ as well as $g \in \mathcal{V}$, β becomes unidentifiable as we would now need information on the exact form of g. Hence, this identification method relies on the difference in functional form between f and g.

4.4. TSCI with random forests

A key component of TSCI is to estimate $f(X_i,Z_i) = \mathbb{E}(D_i|X_i,Z_i)$. In this section, we will estimate the conditional mean function using random forests. Other approaches to estimate conditional mean function will be referenced in section 4.5.

We will first split the data $\{(X_i,Z_i,D_i)\}_{1\le i\le n}$ into two disjoint sets $\mathcal{A}_1$ and $\mathcal{A}_2$ with $|\mathcal{A}_1| = n_1 = \lfloor \frac{2n}{3} \rfloor$ and $|\mathcal{A}_2| = n - n_1$. Wlog write: $\mathcal{A}_1 = \{1,...,n_1\}$. To estimate $f(z,x)$ for the prediction of the conditional mean, we first use data from $\mathcal{A}_2$ to construct partitions of the covariate space $\mathbb{R}^{p_X+p_Z}$ using random forests. Using these partitions, we will make a prediction of $f(z,x)$ based on the data-points from $\mathcal{A}_1$ in the same partition as (z,x) by taking averages over the corresponding D_i observations in that partition. This is specified below:

Method 4.4.1. *[5, p.9 Section 3.1][Estimation $f(z,x), (z,x) \in \mathbb{R}^{p_Z+p_X}$ using random forests]*

1. *Take S bootstrap samples of $\mathcal{A}_2$.*
2. *To each bootstrap sample 1,...,S fit a tree (to predict D based on (X,Z) using the $\mathcal{A}_2$ data through f.e. taking averages over all observations in each subspace of the partition) for which for each leaf only m covariates (out of $(X^{(1)},...,X^{(p_X)},Z^{(1)},...,Z^{(p_Z)})$) at random are considered for the split. Denote the randomness by which each tree is grown by θ_s.*
3. *Each decision tree can be viewed as a partition of the whole covariate space $\mathbb{R}^{p_X+p_Z}$ into disjoint subspaces $\{\mathcal{R}_l^s\}_{1\le l\le J_s}$. For any given $(z^T,x^T)^T \in \mathbb{R}^{p_X+p_Z}$ and given a tree s there exists an unique leaf $l(z,x,\theta_s)$ with $1 \le l(z,x,\theta_s) \le J_s$ such that $\mathcal{R}^s_{l(z,x,\theta_s)}$ contains $(z^T,x^T)^T$.*

[1] $\mathrm{P}_{\mathcal{V}}^{\perp} = \mathrm{Id} - \mathrm{P}_{\mathcal{V}}$

4. *For each decision tree, we predict $f(z,x)$ by:*

$$\hat{f}_{\theta_s}(z,x) = \sum_{j\in\mathcal{A}_1} w_j(z,x,\theta_s)D_j,$$
$$w_j(z,x,\theta_s) = \frac{1((Z_j^T, X_j^T)^T \in \mathcal{R}^s_{l(x,z,\theta_s)})}{\sum_{k\in\mathcal{A}_1} 1((Z_k^T, X_k^T)^T \in \mathcal{R}^s_{l(x,z,\theta_s)})}$$

5. *For the whole random forest, we obtain:*

$$\hat{f}(z,x) = \frac{1}{S}\sum_{s=1}^{S} \hat{f}_{\theta_s}(z,x) =: \sum_{j\in\mathcal{A}_1} w_j(z,x)D_j$$
$$w_j(z,x) := \frac{1}{S}\sum_{s=1}^{S} w_j(z,x,\theta_s)$$

Remark 4.4.2. *[6, p.587-603 Section 15 Random Forests] The random forests can, for example, use recursive binary splitting for tree building. The decorrelation of the trees will lead to an overall lower variance than when fitting a single tree or using out-of-bag sample. One downside to using random forests is that they will not perform well if the number of relevant covariates is small, hence the user should be aware of that. One potential way to select only relevant covariates is through a variable importance measure.*

The weights $(w_j(z,x))_{j\in\mathcal{A}_1}$ as defined in method 4.4.1 have the following properties:

Lemma 4.4.3. $\{w_j(z,x)\}_{j\in\mathcal{A}_1}$ *satisfies:*

- $w_j(z,x) \geq 0$
- $\sum_{j\in\mathcal{A}_1} w_j(z,x) = 1$

Proof.

- Clear from the definition.
- $$\begin{aligned}\sum_{j\in\mathcal{A}_1} w_j(z,x) &= \sum_{j\in\mathcal{A}_1} \frac{1}{S}\sum_{s=1}^{S} w_j(z,x,\theta_s)\\ &= \frac{1}{S}\sum_{s=1}^{S}\sum_{j\in\mathcal{A}_1} w_j(z,x,\theta_s)\\ &= \frac{1}{S}\sum_{s=1}^{S} \frac{\sum_{j\in\mathcal{A}_1} \mathbb{1}((Z_j^T, X_j^T)^T \in \mathcal{R}_{l(z,x,\theta_s)})}{\sum_{j\in\mathcal{A}_1} \mathbb{1}((Z_j^T, X_j^T)^T \in \mathcal{R}_{l(z,x,\theta_s)})}\\ &= 1\end{aligned}$$

□

Remark 4.4.4. *The second bullet point from lemma 4.4.3 suggests that if $D_j = 1$ for all $j \in \mathcal{A}_1$, the random forests will always predict 1 no matter the input of (z, x).*

The TSCI method will try to predict β using $\hat{f}$. First, some new notation is introduced for f, D and $\hat{f}$ evaluated on $\mathcal{A}_1$:

Notation 4.4.5.

- $f_{\mathcal{A}_1} := (f(Z_1, X_1), ..., f(Z_{n_1}, X_{n_1}))^T$, $D_{\mathcal{A}_1} = (D_1, ..., D_{n_1})^T$
- $\hat{f}_{\mathcal{A}_1} = \Omega D_{\mathcal{A}_1}$, $\Omega_{i,j} = w_j(Z_i, X_i)$ *with* $i, j \in \mathcal{A}_1$.

The second main objective of this method, after identifying β, is to give an invalidity test for the potential instrumental variables (IVs) Z. The test will provide an answer as to which $Z = (Z^{(1)}, ..., Z^{(p_Z)})^T$ are (potentially) instrumental. To be able to do this, we will now split g, from the outcome model, into $g = h + \phi$ (as defined in section 4.2). Then we proceed to estimate h and ϕ separately as follows:
$\phi(X_i) = \phi(X_i^{(1)}, ..., X_i^{(p_X)})$ is estimated by a linear combination of $(b_{1,1}(X_i^{(1)}), ..., b_{1,m_1}(X_i^{(1)}), b_{2,1}(X_i^{(2)}), ..., b_{p_X,m_{p_X}}(X_i^{(p_X)}))^T =: W_i^T$ (the functions $b_{i,j}$ are chosen by the user).
Next, $h(Z_i, X_i)$ is estimated by a linear combination of $(v_1(Z_i, X_i), ..., v_L(Z_i, X_i))^T =: V_i$ (where again the $v_i's$ are chosen by the user). Hence, we end up with the following uncomputable (as g is generally unknown) least squares estimate of g: $\hat{g}(Z_i, X_i) = V_i^T\pi + W_i^T\psi$ for $i = 1, ..., n$.

Based on the above approximation of g, the following definitions are introduced:

Definition 4.4.6.

- *The violation matrix $V \in \mathbb{R}^{n \times L}$ is defined as:*

$$V = \begin{pmatrix} V_1^T \\ \vdots \\ V_n^T \end{pmatrix}$$

- *The approximation error vector $R(V) = (R_1(V), ..., R_n(V))^T \in \mathbb{R}^n$ is defined as:*

$$R_i(V) = g(Z_i, X_i) - V_i^T\pi - W_i^T\psi, \quad 1 \leq i \leq n.$$

With the above basis approximation for g, we can write the outcome model as:

$$Y_i = D_i\beta + V_i^T\pi + W_i^T\psi + R_i + \epsilon_i$$

Multiplying both sides with the transformation matrix Ω with data in $\mathcal{A}_1$ we obtain:

$$\hat{Y}_{\mathcal{A}_1} := \hat{f}_{\mathcal{A}_1}\beta + \hat{V}_{\mathcal{A}_1}\pi + \hat{W}_{\mathcal{A}_1}\psi + \hat{R}_{\mathcal{A}_1} + \hat{\epsilon}_{\mathcal{A}_1} \tag{4.4.1}$$

Here, $\hat{Y}_{\mathcal{A}_1} = \Omega Y_{\mathcal{A}_1}$ etc.
As π and ψ can in no way be derived from the data, like in the identification section, we want to remove them by multiplying with a projection matrix and then isolate β. This procedure is shown in the following lemma, which was not mentioned in the original paper [5]:

Lemma 4.4.7. *Assuming that* $\hat{f}_{\mathcal{A}_1}^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}} \hat{f}_{\mathcal{A}_1} \neq 0$ [2]*:*

$$\beta = \frac{\hat{Y}_{\mathcal{A}_1}^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}} \hat{f}_{\mathcal{A}_1}}{\hat{f}_{\mathcal{A}_1}^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}} \hat{f}_{\mathcal{A}_1}} - \frac{\hat{f}_{\mathcal{A}_1}^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}} \hat{R}_{\mathcal{A}_1}}{\hat{f}_{\mathcal{A}_1}^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}} \hat{f}_{\mathcal{A}_1}} - \frac{\hat{f}_{\mathcal{A}_1}^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}} \hat{\epsilon}_{\mathcal{A}_1}}{\hat{f}_{\mathcal{A}_1}^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}} \hat{f}_{\mathcal{A}_1}}$$

Proof. Multiply equation 4.4.1 by $\hat{f}^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}}$ and use that for $a \in \mathbb{R}$: $a = a^T$ to get the order of the first term of the right-hand side of β as alleged in the lemma. □

Observe that from the expression of β in lemma 4.4.7 only the first term from the right of the "=" is computable from the data. In the identification section, β was identified through two main assumptions: first that $g \in \mathrm{span}(\mathcal{V})$ and secondly that $\left\|\mathrm{P}^{\perp}_{\mathcal{V}} f\right\| > 0$. Compared to the expression of β in section 4.3, we now have $\hat{R}_{\mathcal{A}_1}$ and $\hat{\epsilon}_{\mathcal{A}_1}$ involved. In an asymptotic sense, we want to get rid of those terms to have consistency for the computable term. Observe that $\hat{f}_{\mathcal{A}_1}^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}} \hat{f}_{\mathcal{A}_1} = \left\|\mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}} \hat{f}_{\mathcal{A}_1}\right\|^2$. Building on the idea that f shouldn't be well-estimated by $\mathcal{V}$, a natural translation would be to require that: $\left\|\mathrm{P}^{\perp}_{\hat{V},\hat{W}} \hat{f}_{\mathcal{A}_1}\right\|^2 \to \infty$. This would also have to mean that $\hat{R}'_{\mathcal{A}_1}s$ growth rate should be slower than that of $\hat{f}_{\mathcal{A}_1}^T \mathrm{P}^{\perp}_{\hat{V},\hat{W}} \hat{f}_{\mathcal{A}_1}$. In the perfect case that $\left\|\hat{R}_{\mathcal{A}_1}\right\| = 0$ $(\forall n \geq 1)$ this is satisfied. Observe that it not required that $\mathrm{span}(V_i, W_i)$ suddenly perfectly estimates $g(Z_i, X_i)$ for $n \to \infty$ (which would also be not viable as the amount of basis functions is generally fixed and does not increase with n), but rather is requires that the approximation is well enough. These ideas are used for the first estimator of β defined below:

Definition 4.4.8.

$$\hat{\beta}_{\mathrm{init}}(V) := \frac{\hat{Y}_{\mathcal{A}_1}^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}} \hat{f}_{\mathcal{A}_1}}{\hat{f}_{\mathcal{A}_1}^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}} \hat{f}_{\mathcal{A}_1}} := \frac{Y_{\mathcal{A}_1}^T \mathrm{M}_{\mathrm{RF}}(V) D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^T \mathrm{M}_{\mathrm{RF}}(V) D_{\mathcal{A}_1}}$$
$$\mathrm{M}_{\mathrm{RF}}(V) := \Omega^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}} \Omega$$

As seen from lemma 4.4.7, $\hat{\beta}_{\mathrm{init}}$ suffers from a bias for finite n. In order to correct for that error for finite n the following bias-corrected estimator is introduced:

Definition 4.4.9.

$$\hat{\beta}_{\mathrm{RF}}(V) = \hat{\beta}_{\mathrm{init}}(V) - \frac{\sum_{i=1}^{n_1} [\mathrm{M}_{\mathrm{RF}}(V)]_{i,i} \hat{\delta}_i [\hat{\epsilon}(V)]_i}{D_{\mathcal{A}_1}^T \mathrm{M}_{\mathrm{RF}}(V) D_{\mathcal{A}_1}}$$

with $\hat{\delta}_{\mathcal{A}_1} = D_{\mathcal{A}_1} - \hat{f}_{\mathcal{A}_1}$, $\hat{\epsilon}(V) = \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},W_{\mathcal{A}_1}} [Y_{\mathcal{A}_1} - D_{\mathcal{A}_1} \hat{\beta}_{\mathrm{init}}(V)]$.

Then the following estimated confidence interval and standard error can be constructed for $\hat{\beta}_{\mathrm{RF}}(V)$:

[2] Here, $\mathrm{P}_{A,B}(\cdot) := \mathrm{P}_{\mathrm{col}(A,B)}(\cdot)$

Definition 4.4.10.

$$\begin{aligned}\text{CI}_{\text{RF}}(V) &= (\hat{\beta}_{\text{RF}}(V) - z_{\alpha/2}\hat{\text{SE}}(V), \hat{\beta}_{\text{RF}}(V) + z_{\alpha/2}\hat{\text{SE}}(V)) \\ \hat{\text{SE}}(V) &= \frac{\sqrt{\sum_{i=1}^{n_1}[\hat{\epsilon}(V)]_i^2[\text{M}_{\text{RF}}(V)D_{\mathcal{A}_1}]_i^2}}{D_{\mathcal{A}_1}^T \text{M}_{\text{RF}}(V)D_{\mathcal{A}_1}} \\ \hat{\epsilon}(V) &= \text{P}^{\perp}_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}[Y_{\mathcal{A}_1} - D_{\mathcal{A}_1}\hat{\beta}_{\text{init}}(V)]\end{aligned}$$

See section 4.6.1 for further technical details on $\hat{\beta}_{\text{RF}}$ and the corresponding confidence interval.

4.4.1. Generalized IV strength

As mentioned in the previous section, the identification of β relies upon that $\left\|\text{P}^{\perp}_{\mathcal{V}} f\right\| > 0$ i.e. there is enough difference in the functional form of f and g. In the formal theorems (see section 4.6.1) that provide properties such as asymptotic normality and consistency for $\hat{\beta}_{\text{RF}}(V)$, this notion of $\left\|\text{P}^{\perp}_{\mathcal{V}} f\right\| > 0$ is translated to $f_{\mathcal{A}_1}^T \text{M}_{\text{RF}}(V) f_{\mathcal{A}_1} \to \infty$ (for $n \to \infty$ a.s.), which is equivalent to requiring that $\left\|\text{P}^{\perp}_{\hat{V}_{\mathcal{A}_1}, \hat{W}_{\mathcal{A}_1}} \Omega f_{\mathcal{A}_1}\right\| \to \infty$ (for $n \to \infty$ a.s.). Another imposed assumption is that $f_{\mathcal{A}_1}^T \text{M}_{\text{RF}}(V) f_{\mathcal{A}_1} \gg \text{Tr}(\text{M}_{\text{RF}}(V))$ (see section 4.6.1 for further details on these conditions). To check whether these two conditions are reasonable to assume for our data and choice of V, a test is introduced which checks whether a standardized estimator of $f_{\mathcal{A}_1}^T \text{M}_{\text{RF}}(V) f_{\mathcal{A}_1}$ is big enough. If the test is passed, we say that the generalised IV strength is strong enough. This notion is introduced below:

Definition 4.4.11. *Given a set of basis functions $\mathcal{V}$, the generalised IV strength is defined as:*

$$\mu(V) = \frac{f_{\mathcal{A}_1}^T \text{M}_{\text{RF}}(V) f_{\mathcal{A}_1}}{\sum_{i \in \mathcal{A}_1} \text{Var}(\delta_i | X_i, Z_i)(|\mathcal{A}_1|)^{-1}}$$

Remark 4.4.12. *If $\text{Var}(\delta_i | X_i, Z_i) = \sigma_\delta^2$, then:*

$$\mu(V) = \frac{f_{\mathcal{A}_1}^T \text{M}_{\text{RF}}(V) f_{\mathcal{A}_1}}{\sigma_\delta^2}$$

Here, we see directly that if $f_{\mathcal{A}_1}^T \text{M}_{\text{RF}}(V) f_{\mathcal{A}_1} \to \infty$ that then $\mu(V) \to \infty$.

As hinted at before, a sufficiently large $\mu(V)$ will give us hope that $\hat{\beta}(V)$ and $\text{CI}_{\text{RF}}(V)$ are good estimators for β. As f is generally unknown beforehand, we'll have to estimate $\mu(V)$:

Definition 4.4.13.

$$\hat{\mu}(V) = \frac{D_{\mathcal{A}_1}^T \text{M}_{\text{RF}}(V) D_{\mathcal{A}_1}}{\left\|D_{\mathcal{A}_1} - \hat{f}_{\mathcal{A}_1}\right\|^2 / n_1}$$

Now I will describe a test to check whether $\hat{\mu}(V)$ is large enough. Of course, for the theoretical properties the restrictions are put on $\mu(V)$ and not on $\hat{\mu}(V)$. Hence, this test will also try to capture the error between $\mu(V)$ and $\hat{\mu}(V)$ by using that $D_{\mathcal{A}_1} = f_{\mathcal{A}_1} + \delta_{\mathcal{A}_1}$ with:

$$D_{\mathcal{A}_1}^T \mathrm{M_{RF}}(V) D_{\mathcal{A}_1} - f_{\mathcal{A}_1}^T \mathrm{M_{RF}}(V) f_{\mathcal{A}_1} = 2 f_{\mathcal{A}_1}^T \mathrm{M_{RF}}(V) \delta_{\mathcal{A}_1} + \delta_{\mathcal{A}_1}^T \mathrm{M_{RF}}(V) \delta_{\mathcal{A}_1}.$$

Method 4.4.14 (Testing IV strength $\mathcal{V}$). *[5, p.13,14 Section 3.3]*

- *For $1 \leq i \leq n_1$, we define $\hat{\delta}_i = D_i - \hat{f}_i$ and compute $\tilde{\delta}_i = \hat{\delta}_i - \bar{\mu}_\delta$, where $\bar{\mu}_\delta = \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{\delta}_i$. Hence, we are considering a centered error.*
- *For $1 \leq l \leq L$, we generate $\delta_i^{[l]} = U_i^{[l]} \tilde{\delta}_i$, for $1 \leq i \leq n_1$, where $\{U_i^{[l]}\}_{1 \leq i \leq n_1}$ are iid standard normal.*
- *For $1 \leq l \leq L$, we compute:*

$$S^{[l]} = \frac{2 \hat{f}_{\mathcal{A}_1}^T \mathrm{M_{RF}}(V) \delta^{[l]} + (\delta^{[l]})^T \mathrm{M_{RF}}(V) \delta^{[l]}}{\left\| D_{\mathcal{A}_1} - \hat{f}_{\mathcal{A}_1} \right\|^2 / n_1}$$

  and we use $S_{\alpha_0}(V)$ to denote the upper α_0 empirical quantile of $\{|S^{[l]}|\}_{1 \leq l \leq L}$[3]
- *We conduct the generalized IV strength test:*

$$\hat{\mu}(V) \geq \max(2 \operatorname{Tr}(\mathrm{M_{RF}}(V)), 10) + S_{\alpha_0}(V)$$

  Here, S_{α_0} captures the estimation error of $\hat{\mu}(V) - \mu(V)$

In case V passes the test, we have hope for that there is enough difference in functional form between f and g to obtain consistency (among other properties) for our estimators.

4.4.2. Data dependent selection of $\mathcal{V}$ and IV validity test

Suppose that we have a model with, in theory, enough functional difference between f and g to obtain a good estimator of β from the data (with the methods described before). We need some prior knowledge beforehand on the basis functions $\mathcal{V}$. There are two important considerations when choosing $\mathcal{V}$: firstly, if we choose $\mathcal{V}$ with too little basis functions the difference between the estimated g from the basis and f might be large enough, but the difference between g and the estimated $\hat{g}$ will be too big. Secondly, if we choose $\mathcal{V}$ too large, the difference between g and the estimated g will be small enough but now f is likely to be well-estimated by the basis functions in $\mathcal{V}$. To strike a balance between these two considerations, we'll consider a nested set of basis functions:

$$\{0\} =: \mathcal{V}_0 \subset ... \subset \mathcal{V}_Q, Q \in \mathbb{N}_{>0}$$

and devise a data dependent way to choose the best among $\{V_q\}_{0 \leq q \leq Q}$.

To choose between these different bases we first want to know up until which $m \in \mathbb{N}$ f is not well-approximated by the basis $\mathcal{V}_m$. $Q_{\max}$, as defined below, represents the largest basis for which this still holds.

[3] z is the upper α_0 empirical quantile of $Z_1, ..., Z_N$ if: $z = \min\{z \in \mathbb{R} : \frac{1}{N} \sum_{i=1}^{N} 1(Z_i \leq z) \geq 1 - \alpha_0\}$

Definition 4.4.15. $Q_{\max} = \arg\max_{q\geq 0}\{\hat{\mu}(V_q) \geq \max(2\,\mathrm{Tr}(\mathrm{M}_{\mathrm{RF}}(V_q)), 10) + S_{\alpha_0}(V_q)\}$

With $Q_{\max}$, we shall now choose among $\{V_q\}_{0\leq q\leq Q_{\max}}$. Though $Q_{\max}$ will give the smallest R vector (i.e. difference between $\hat{g}$ and g, see definition 4.4.6) that also has enough difference between f and $\hat{g}$, it could be that a smaller $0 \leq q \leq Q_{\max}$ will also give a small enough R vector and even more difference between f and $\hat{g}$. That choice V_q would lead to less conservative CI's for finite n [5, p.26 5.Theoretical Justification Remark 4].

The main idea behind the test for choosing among $\{V_q\}_{0\leq q\leq Q_{\max}}$ is that if both V_q and $V_{q'}$ are good basis choices, then for $n \to \infty$:

$$\frac{\hat{\beta}_{\mathrm{RF}}(V_q) - \hat{\beta}_{\mathrm{RF}}(V_{q'})}{\sqrt{H(V_q, V_{q'})}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

where $H(V_q, V_{q'})$ represent a variance term (see section 4.6.1 for more technical details). The test is described below:

Method 4.4.16 (Choosing among $\{V_q\}_{0\leq q\leq Q_{\max}}$)**.** *[5, p.15,16 Section 3.4]*

1. *For any given $0 \leq q \leq Q_{\max}$, compute:*

$$\hat{\beta}_{\mathrm{RF}}(V_q) = \hat{\beta}_{\mathrm{init}}(V_q) - \frac{\sum_{i=1}^{n_1}[\mathrm{M}_{\mathrm{RF}}(V_q)]_{ii}\hat{\delta}_i[\hat{\epsilon}(V_{Q_{\max}})]_i}{D_{\mathcal{A}_1}^T\,\mathrm{M}_{\mathrm{RF}}(V_q)D_{\mathcal{A}_1}}$$

2. *We estimate the (asymptotic) variance of $\hat{\beta}_{\mathrm{RF}}(V_q) - \hat{\beta}_{\mathrm{RF}}(V_{q'})$ by:*

$$\begin{aligned}\hat{H}(V_q, V_{q'}) = &\frac{\sum_{i=1}^{n_1}[\hat{\epsilon}(V_{Q_{\max}})]_i^2[\mathrm{M}_{\mathrm{RF}}(V_{q'})D_{\mathcal{A}_1}]_i^2}{[D_{\mathcal{A}_1}^T\,\mathrm{M}_{\mathrm{RF}}(V_{q'})D_{\mathcal{A}_1}]^2}\\ &+\frac{\sum_{i=1}^{n_1}[\hat{\epsilon}(V_{Q_{\max}})]_i^2[\mathrm{M}_{\mathrm{RF}}(V_q)D_{\mathcal{A}_1}]_i^2}{[D_{\mathcal{A}_1}^T\,\mathrm{M}_{\mathrm{RF}}(V_q)D_{\mathcal{A}_1}]^2}\\ &-2\frac{\sum_{i=1}^{n_1}[\hat{\epsilon}(V_{Q_{\max}})]_i^2[\mathrm{M}_{\mathrm{RF}}(V_{q'})D_{\mathcal{A}_1}]_i[\mathrm{M}_{\mathrm{RF}}(V_q)D_{\mathcal{A}_1}]_i}{[D_{\mathcal{A}_1}^T\,\mathrm{M}_{\mathrm{RF}}(V_{q'})D_{\mathcal{A}_1}][D_{\mathcal{A}_1}^T\,\mathrm{M}_{\mathrm{RF}}(V_q)D_{\mathcal{A}_1}]}\end{aligned}$$

   and define the following data-dependent test statistic for choosing between V_q and $V_{q'}$:

$$C^{RF}(V_q, V_{q'}) = 1(|\hat{\beta}_{\mathrm{RF}}(V_q) - \hat{\beta}_{\mathrm{RF}}(V_{q'})|/\sqrt{\hat{H}(V_q, V_{q'})} \geq z_{\alpha_0})$$

   where z_{α_0} is the α_0 upper quantile of the standard normal (i.e. $P(N(0,1) \leq z_{\alpha_0}) = 1-\alpha_0$)

3. *For $Q_{\max} \geq 2$, we generalize pairwise comparisons to the following:*

$$C^{RF}(V_q) = 1(\max_{q+1\leq q'\leq Q_{\max}}[|\hat{\beta}_{\mathrm{RF}}(V_q) - \hat{\beta}_{\mathrm{RF}}(V_{q'})|/\sqrt{\hat{H}(V_q, V_{q'})}] \geq \hat{\phi})$$

   where $\hat{\phi} > 0$ is to be decided. We define $C^{RF}(V_{Q_{\max}}) = 0$.

4. *We choose the smallest q for which holds that $C^{RF}(V_q) = 0$.*

Besides choosing among $\{V_q\}_{0\leq q\leq Q_{\max}}$, it also tests which $Z = (Z^{(1)}, ..., Z^{(pz)})^T$ are (potentially) valid instrumental variables depending on our choice of $(V_i)_{1\leq i\leq Q_{\max}}$. This is done as follows: in case V_0 comes out as the best basis choice, it would mean that the violation function h is best approximated by 0, so no violation function. Then we conclude that all $Z^{(j)}$ (involved in $(V_i)_{1\leq i\leq Q_{\max}}$) are (candidate) valid instrumental variables.

We now consider a choice for $\hat{\phi} > 0$ as seen in method 4.4.16. For this, we use a bootstrap sample of $\hat{\beta}_{\mathrm{RF}}(V_q) - \hat{\beta}_{\mathrm{RF}}(V_{q'})$. Here we use that if $\|R_q\|$ and $\|R_{q'}\|$ are small it holds that (see section 4.6.1 for the technical details):

$$\hat{\beta}_{\mathrm{RF}}(V_{q'}) - \hat{\beta}_{\mathrm{RF}}(V_q) \approx \frac{f_{\mathcal{A}_1}^T \mathrm{M}_{\mathrm{RF}}(V_{q'})\epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathrm{M}_{\mathrm{RF}}(V_{q'})f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^T \mathrm{M}_{\mathrm{RF}}(V_q)\epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathrm{M}_{\mathrm{RF}}(V_q)f_{\mathcal{A}_1}}$$

Method 4.4.17 (Choosing $\hat{\phi} > 0$ using bootstrap)**.** *[5, p.15,16 Section 3.4]*

1. *For $1 \leq i \leq n_1$, we compute $\tilde{\epsilon}_i = [\hat{\epsilon}(V_{Q_{\max}})]_i - \bar{\mu}_\epsilon$, $\bar{\mu}_\epsilon = \frac{1}{n_1}\sum_{i=1}^{n_1}[\hat{\epsilon}(V_{Q_{\max}})]_i$*
2. *For $1 \leq l \leq L$, we generate: $\epsilon_i^{[l]} = U_i^{[l]}\tilde{\epsilon}_i$ with for $1 \leq i \leq n_1$, $\{U_i^{[l]}\}_{1\leq i\leq n_1} \overset{iid}{\sim} N(0,1)$.*
3. *For $1 \leq l \leq L$, we compute:*

$$T^{[l]} = \max_{0\leq q<q'\leq Q_{\max}} \frac{1}{\sqrt{\hat{H}(V_q, V_{q'})}}\left[\frac{D_{\mathcal{A}_1}^T \mathrm{M}_{\mathrm{RF}}(V_{q'})\epsilon^{[l]}}{D_{\mathcal{A}_1}^T \mathrm{M}_{\mathrm{RF}}(V_{q'})D_{\mathcal{A}_1}} - \frac{D_{\mathcal{A}_1}^T \mathrm{M}_{\mathrm{RF}}(V_q)\epsilon^{[l]}}{D_{\mathcal{A}_1}^T \mathrm{M}_{\mathrm{RF}}(V_q)D_{\mathcal{A}_1}}\right]$$

4. *We set $\hat{\phi} = \hat{\phi}(\alpha_0)$ to be the α_0 upper empirical quantile of $\{|T^{[l]}|\}_{1\leq l\leq L}$*

We can now establish a choice of $0 \leq q \leq Q_{\max}$ using method 4.4.16 and method 4.4.17:

Definition 4.4.18. $\hat{q}_c = \underset{0\leq q\leq Q_{\max}}{\arg\min}\{C^{RF}(V_q) = 0\}$

Observe that as, per definition, $C^{RF}(V_{Q_{\max}}) = 0$, $\hat{q}_c$ always exists.

Remark 4.4.19. *The c of $\hat{q}_c$ stands for comparison, as we choose among the best violation matrices.*

For small sample sizes, a more robust choice for the best violation matrix is proposed (in order to counter that certain too small violations can't be detected): $\hat{q}_r = \min(\hat{q}_c + 1, Q_{\max})$

4.4.3. Finite-sample adjustment of uncertainty from data splitting

Even though the asymptotic theory is valid for any random sample splitting, the constructed point estimators and confidence intervals do vary with different sample splittings in finite samples. In the following: a confidence interval that aggregates multiple intervals due to different sample splittings is introduced.

Definition 4.4.20. *(Aggregate multiple splittings)*

- *Consider S random sample splittings and for the s-th splitting, we use $\hat{\beta}^s$ and $\hat{\mathrm{SE}}^s$*
- *Define the median estimators:*

$$\hat{\beta}^{\mathrm{Med}} := \mathrm{Med}\{\hat{\beta}^s\}_{1\le s\le S}$$
$$\hat{\mathrm{SE}}^{\mathrm{Med}} := \mathrm{Med}\{\sqrt{(\hat{\mathrm{SE}}^s)^2 + (\hat{\beta}^s - \hat{\beta}^{\mathrm{Med}})^2}\}_{1\le s\le S}$$

- *We then obtain the following median CI:*

$$(\hat{\beta}^{\mathrm{Med}} - z_{\alpha/2}\hat{\mathrm{SE}}^{\mathrm{Med}}, \hat{\beta}^{\mathrm{Med}} + z_{\alpha/2}\hat{\mathrm{SE}}^{\mathrm{Med}})$$

Observe that aggregation is especially important when dealing with imbalanced data. The implementation of the TSCI method (with random forests) provided by the original paper (https://github.com/zijguo/TSCI-Replication) does not do aggregation.

4.5. TSCI with general machine learning methods

In the previous section, we have seen that random forests were used to construct an estimator for f. We can generalize this to general machine learning methods as follows:

1. First, we write the first stage machine learning estimator of $f_{\mathcal{A}_1}$ as a linear transformation of $D_{\mathcal{A}_1}$: $\hat{f}_{\mathcal{A}_1} = \Omega D_{\mathcal{A}_1}$, where $\Omega \in \mathbb{R}^{n_1 \times n_1}$ is allowed to be stochastic.
2. We define a generalized transform matrix:

$$\mathrm{M}(V) = \Omega^T \mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1}, \hat{W}_{\mathcal{A}_1}} \Omega, \hat{V}_{\mathcal{A}_1} = \Omega V_{\mathcal{A}_1}, \hat{W}_{\mathcal{A}_1} = \Omega W_{\mathcal{A}_1}$$

3. Compared to TSCI with random forests: replace $\mathrm{M}_{\mathrm{RF}}(\cdot)$ with $\mathrm{M}(\cdot)$
4. Notation wise: $\hat{\beta}_{\mathrm{RF}}(V)$ becomes $\hat{\beta}(V)$.

Alternatives to random forests include boosting, deep neural networks and basis approximation, see [5, p.20,21 4.TSCI with general machine learning methods].

4.6. Theoretical justification

4.6.1. Main results

We start with a first set of required conditions on the models for the proofs. For the rest of this section, if there is an inequality or convergence of a stochastic identity, without the mode specified, it will be used in a sure sense (so it will hold for any event). This can be extended to almost sure conditions which would use analogous arguments.

Define $\mathcal{O} := \{(X_i, Z_i)_{i=1}^n, (D_i)_{i\in\mathcal{A}_2}\}$. For the remainder of this section, assume that $(X_i, Z_i) \overset{\text{iid}}{\sim} (X, Z)$ for $i = 1, ..., n$ and (ϵ_i, δ_i) are independent for $i = 1, ..., n$ with $(X_i, Z_i) \perp\!\!\!\perp (\epsilon_j, \delta_j)$ for $i \neq j$. The data satisfies model 4.2.1.

Conditions 4.6.1 (R1)**.**

1. *For $i = 1, ..., n$, conditioning on Z_i, X_i: ϵ_i and δ_i are sub-Gaussian random random variables i.e.:*

$$\sup_{X_i, Z_i} \max(\mathbb{P}(|\epsilon_i| > t|Z_i, X_i), \mathbb{P}(|\delta_i| > t|Z_i, X_i)) \leq \exp(-K^2t^2/2)$$

   Here, $\sup_{X_i,Z_i}$ denotes the sup *taken over the support of the density of (X_i, Z_i) and $K > 0$ is deterministic.*

2. *The random vectors $\{\Psi_i, f_i\}_{1\leq i\leq n_1}$ with $\Psi_i := (V_i^T, W_i^T)^T$ and $f_i := f(Z_i, X_i)$ satisfy $\forall n_1 \in \mathbb{N}$:*

   a) $\lambda_{\min}(\sum_{i=1}^{n_1}(\Psi_i\Psi_i^T)/n_1) \geq c$

   b) $\left\|\sum_{i=1}^{n_1} \Psi_i f_i/n_1\right\|_2 \leq C$

   c) $\max_{1\leq i\leq n_1} \{|f_i|, \|\Psi_i\|_2\} \leq C\sqrt{\log(n_1)}$

   d) $\left\|\sum_{i=1}^{n_1} \Psi_i[R(V)]_i/n_1\right\|_2 \leq C \|R(V)\|_\infty$

3. *For Ω from $\hat{f}_{\mathcal{A}_1} = \Omega D_{\mathcal{A}_1}$ we have: $\lambda_{\max}(\Omega) \leq C$*

Here, $c, C > 0$ are deterministic constants which don't depend on n, p_X, p_Z.

A few observations regarding (R1):

- Observe that we assume that ϵ_i and δ_i are sub-Gaussian conditioning on Z_i and X_i which is different (and weaker) from what is assumed in section 2.7.2. This is due to the fact that for the TSCI model we assume a weaker condition on the errors, namely: $\mathbb{E}(\epsilon_i|X_i, Z_i) = \mathbb{E}(\delta_i|X_i, Z_i) = 0$ which implies $\mathbb{E}(\epsilon_i X_i) = \mathbb{E}(\delta_i X_i) = 0, \mathbb{E}(\epsilon_i Z_i) = \mathbb{E}(\delta_i Z_i) = 0$, the conditions as seen in model 2.2.1.
- Secondly, note that condition 2a implies that $\mathbb{E}(\Psi_i\Psi_i^T)$ is positive definite (if it is finite) see lemma 4.6.2 below.

- Conditions 2b and 2c denote assumptions on the basis functions of the approximation of g and f. 2c shows that both f and the basis functions should have a growth rate with order less or equal to $\sqrt{\log(n_1)}$ when evaluated on the data. This does put some restrictions on the form of f, v_i and $b_{i,j}$. For instance, $f(z,x) = \exp(zx)$ would not be allowed if (Z_i, X_i) takes too large values and n_1 is not large enough to compensate for that.
- 2d indicates that $\frac{1}{n_1}\sum_{i=1}^{n_1} \Psi_i[R(V)]_i$ its growth rate is dominated by the error between g and its basis approximation.
- It will be shown that for Random Forests: $\lambda_{\max}(\Omega_{\text{RF}}) \leq 1$ (see section 4.6.2), hence satisfying the condition (R1) (3).
- Observe that all conditions of (R1) (2a) and (2c) are put on the dataset $\mathcal{A}_1$ with no mention of $\mathcal{A}_2$. Ω is the only stochastic expression in of (R1) (2a) to (2c) that depends on $\mathcal{A}_2$. So putting restrictions on Ω, we are indirectly putting restrictions on $\mathcal{A}_2$.

Lemma 4.6.2. *Assume that* $\mathbb{E}(\Psi_i\Psi_i^T) < \infty$ *and* $\forall n_1 \in \mathbb{N}$*:*

$$\mathbb{P}(\lambda_{\min}(\frac{1}{n_1}\sum_{i=1}^{n_1}\Psi_i\Psi_i^T) \geq c > 0) = 1 \tag{4.6.1}$$

Then $\mathbb{E}(\Psi_i\Psi_i^T)$ *is positive definite.*

Proof. Let $x \in \mathbb{R}^{L+m_1+\ldots+m_{pX}}$ be a vector of length 1. Then by lemma 2.7.8 together with assumption (4.6.1):

$$\mathbb{P}(x^T(\frac{1}{n_1}\sum_{i=1}^{n_1}\Psi_i\Psi_i^T)x \geq c) = 1$$

Consider two cases:

1. $x^T\,\mathbb{E}(\Psi_i\Psi_i^T)x = c$. In that case we can conclude that $\mathbb{E}(\Psi_i\Psi_i^T)$ is positive definite.
2. In case $x^T\,\mathbb{E}(\Psi_i\Psi_i^T)x \neq c$, c is a continuity point of $\mathbb{R} \ni a \mapsto \mathbb{P}(x^T\,\mathbb{E}(\Psi_i\Psi_i^T)x \leq a)$ and so by the law of large numbers:

$$\mathbb{P}(x^T(\frac{1}{n_1}\sum_{i=1}^{n_1}\Psi_i\Psi_i^T)x \geq c) \overset{n\to\infty}{\to} \mathbb{P}(x^T\,\mathbb{E}(\Psi_i\Psi_i^T)x \geq c) = 1(x^T\,\mathbb{E}(\Psi_i\Psi_i^T)x \geq c)$$

As $\mathbb{P}(x^T(\frac{1}{n_1}\sum_{i=1}^{n_1}\Psi_i\Psi_i^T)x \geq c) = 1$ for any n_1, we hence obtain that: $1(x^T\,\mathbb{E}(\Psi_i\Psi_i^T)x \geq c) = 1$. In conclusion: $\mathbb{E}(\Psi_i\Psi_i^T)$ is positive definite.

□

The second set of conditions is imposed on the generalised IV strength (i.e. $\mu(V)$). Throughout the next text, the asymptotics are taken $n \to \infty$.

Conditions 4.6.3 (R2)**.** $f_{\mathcal{A}_1}^T\,\mathrm{M}(V)f_{\mathcal{A}_1}$ *satisfies:*

- $f_{\mathcal{A}_1}^T\,\mathrm{M}(V)f_{\mathcal{A}_1} \to \infty$

- $f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1} \gg \mathrm{Tr}(\mathrm{M}(V))$

Remark 4.6.4. *Per definition, in case $c \leq \mathrm{Var}(\delta_i | X_i, Z_i) \leq C$ for some $c, C > 0$: $\mu(V)$ is proportional to $f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1}$.*

The following proposition establishes that $\hat{\beta}_{\text{init}}$ is consistent if (R1) and (R2) are satisfied together with the assumption that $\{R_i(V)\}_{1 \leq i \leq n}$ are small enough:

Proposition 4.6.5. *[2, p.22 Section 5 Proposition 2](Consistency $\beta_{\text{init}}(V)$)*
Suppose that (R1) and (R2) are satisfied together with $f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1} \gg \|R_{\mathcal{A}_1}\|_2^2$. Then: $\hat{\beta}_{\text{init}}(V) \xrightarrow{\mathbb{P}} \beta$

Remark 4.6.6. *The condition $f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1} \gg \|R_{\mathcal{A}_1}\|_2^2$ will be satisfied in case g is well-approximated enough by the column space of (V, W). In the extreme case, $R(V) = 0$, this condition is automatically satisfied.*

Next, we will consider the steps the original paper takes to prove proposition 4.6.5. To prove proposition 4.6.5, lemma 4.4.7 in combination with the fact that $D_{\mathcal{A}_1} = f_{\mathcal{A}_1} + \delta_{\mathcal{A}_1}$ is used to obtain the following decomposition:

$$\hat{\beta}_{\text{init}}(V) - \beta = \frac{\epsilon_{\mathcal{A}_1}^T \mathrm{M}(V) \delta_{\mathcal{A}_1} + \epsilon_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1} + R_{\mathcal{A}_1}^T \mathrm{M}(V) D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^T \mathrm{M}(V) D_{\mathcal{A}_1}} \tag{4.6.2}$$

Next, we want to "take control" over the terms in the numerator above and connect them to condition (R2) to be able to prove consistency for $n \to \infty$. How this is done, will be displayed in the following part. It will make use of a lemma about the concentration of quadratic forms of sub-Gaussian random vectors, which is stated below:

Lemma 4.6.7. *(Hanson-Wright inequality)[7, p.2 Theorem 1.1]*
Let $\epsilon \in \mathbb{R}^n$ be a sub-Gaussian vector with independent components ϵ_i with mean 0. Then there exists a constant $K > 0$ such that for any $A : n \times n$ deterministic matrix it holds that $\forall t \geq 0$:

$$\mathbb{P}(|\epsilon^T A \epsilon - \mathbb{E}(\epsilon^T A \epsilon)| > t) \leq 2 \exp(-c \min(\frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|_2^2}))$$

Here, c does not depend on ϵ, A or t while K does depend on ϵ (but not on A or t).

As a consequence, we also have a concentration inequality for $\epsilon^T A \delta$. This result was established in the original paper [5, p.35 Section 9 below Lemma 1] and I worked out the details of the provided proof.

Corollary 4.6.8. *Let $\epsilon \in \mathbb{R}^n$ and $\delta \in \mathbb{R}^n$ both be sub-Gaussian vectors with independent components and mean 0 (as in lemma 4.6.7). Then $\exists K' > 0$ (constant) such that $\forall A : n \times n$ deterministic and $t \geq 0$:*

$$\mathbb{P}(|\epsilon^T A \delta - \mathbb{E}(\epsilon^T A \delta)| > t) \leq 6 \exp(-c \min(\frac{t^2}{K'^4 \|A\|_F^2}, \frac{t}{K'^2 \|A\|_2^2}))$$

Here, c does not depend on ϵ, δ, A or t while K' does depend on ϵ and δ (but not on A or t)

Proof. Observe that $\epsilon^T A\delta = \frac{1}{2}[(\epsilon+\delta)^T A(\epsilon+\delta) - \epsilon^T A\epsilon - \delta^T A\delta]$. Consequently, it also holds that $\mathbb{P}(|\epsilon^T A\delta - \mathbb{E}(\epsilon^T A\delta)| > t)$ can be written as:

$$\mathbb{P}(|(\epsilon+\delta)^T A(\epsilon+\delta) - \epsilon^T A\epsilon - \delta^T A\delta - \mathbb{E}[(\epsilon+\delta)^T A(\epsilon+\delta) - \epsilon^T A\epsilon - \delta^T A\delta]| > 2t) \leq$$
$$\mathbb{P}(|(\epsilon+\delta)^T A(\epsilon+\delta) - \mathbb{E}[(\epsilon+\delta)^T A(\epsilon+\delta)]| + |\epsilon^T A\epsilon - \mathbb{E}(\epsilon^T A\epsilon)| + |\delta^T A\delta - \mathbb{E}(\delta^T A\delta)| > 2t)$$

This last expression can in turn be upper-bounded by:

$$\mathbb{P}(|(\epsilon+\delta)^T A(\epsilon+\delta) - \mathbb{E}[(\epsilon+\delta)^T A(\epsilon+\delta)]| > \frac{2}{3}t \vee |\epsilon^T A\epsilon - \mathbb{E}(\epsilon^T A\epsilon)| > \frac{2}{3}t \vee |\delta^T A\delta - \mathbb{E}(\delta^T A\delta)| > \frac{2}{3}t)$$

which, in turn, can be upper-bounded by:

$$\mathbb{P}(|(\epsilon+\delta)^T A(\epsilon+\delta) - \mathbb{E}[(\epsilon+\delta)^T A(\epsilon+\delta)]| > \frac{2}{3}t) + \mathbb{P}(|\epsilon^T A\epsilon - \mathbb{E}(\epsilon^T A\epsilon)| > \frac{2}{3}t) + \mathbb{P}(|\delta^T A\delta - \mathbb{E}(\delta^T A\delta)| > \frac{2}{3}t)$$

Which, as a last step, can be upper-bounded by:

$$6\exp(-c\min(\frac{t^2}{K'^4\|A\|_F^2}, \frac{t}{K'^2\|A\|_2^2}))$$

In the last step the Hanson-Wright inequality was applied to the 3 terms and then upper-bounded. □

Remark 4.6.9. *For the current setting where $\epsilon_{\mathcal{A}_1}$ and $\delta_{\mathcal{A}_1}$ are conditional sub-Gaussian random errors, the corollary above can be extended to A and t being non-deterministic but $\mathcal{O}$-measurable. In that case:*

$$\mathbb{P}(|\epsilon^T A\delta - \mathbb{E}(\epsilon^T A\delta|\mathcal{O})| > t|\mathcal{O}) \leq 6\exp(-c\min(\frac{t^2}{K'^4\|A\|_F^2}, \frac{t}{K'^2\|A\|_2^2}))$$

The following lemma and subsequent corollary show how the first term from the error decomposition 4.6.2 can be controlled. The other terms follow a similar reasoning. Here, I worked out the proof as seen at [5, p.35 9.Proofs Lemma 2]

Lemma 4.6.10. *Assuming condition (R1), for $t_0 > 0$ (which is allowed to be an $\mathcal{O}$-measurable random variable):*

$$\mathbb{P}(|\epsilon_{\mathcal{A}_1}^T \mathrm{M}(V)\delta_{\mathcal{A}_1} - \mathrm{Tr}(\mathrm{M}(V)\Lambda)| \leq t_0 K^2\sqrt{\mathrm{Tr}([\mathrm{M}(V)]^2)}|\mathcal{O}) \geq 1 - 6\exp(-c\min(t_0^2, t_0))$$

where $\Lambda = \mathbb{E}(\delta_{\mathcal{A}_1}\epsilon_{\mathcal{A}_1}^T|X_{\mathcal{A}_1}, Z_{\mathcal{A}_1})$

Proof. Observe that:

$$\mathbb{E}(\epsilon_{\mathcal{A}_1}^T \mathrm{M}(V)\delta_{\mathcal{A}_1}|\mathcal{O}) = \mathbb{E}(\mathrm{Tr}(\epsilon_{\mathcal{A}_1}^T \mathrm{M}(V)\delta_{\mathcal{A}_1})|\mathcal{O}) = \mathbb{E}(\mathrm{Tr}(\mathrm{M}(V)\delta_{\mathcal{A}_1}\epsilon_{\mathcal{A}_1}^T)|\mathcal{O}) = \mathrm{Tr}(\mathrm{M}(V)\Lambda)$$

where it is used that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, $\mathrm{M}(V)$ is $\mathcal{O}-$ measurable and $(\epsilon_i, \delta_i) \perp\!\!\!\perp (X_j, Z_j)$ for $i \neq j$. Let $t_0 > 0$. Then apply corollary 4.6.8 (in combination with remark 4.6.9) to $\epsilon_{\mathcal{A}_1}$ and $\delta_{\mathcal{A}_1}$ with $A = \mathrm{M}(V)$ and $t = t_0 K^2\|\mathrm{M}(V)\|_F$ to see that:

$$\begin{aligned}\mathbb{P}(|\epsilon_{\mathcal{A}_1}^T \mathrm{M}(V)\delta_{\mathcal{A}_1} - \mathrm{Tr}(\mathrm{M}(V)\Lambda)| \geq t_0 K^2\|\mathrm{M}(V)\|_F|\mathcal{O}) &\leq 6\exp(-c\min(t_0^2, t_0\frac{\|\mathrm{M}(V)\|_F}{\|\mathrm{M}(V)\|_2})) \\ &\leq 6\exp(-c\min(t_0, t_0^2))\end{aligned}$$

where in the last step it is used that $\|\mathrm{M}(V)\|_F \geq \|\mathrm{M}(V)\|_2$ □

Consider $\Lambda = \mathbb{E}(\delta_{\mathcal{A}_1}\epsilon_{\mathcal{A}_1}^T|X_{\mathcal{A}_1}, Z_{\mathcal{A}_1})$. By independence of (ϵ_i, δ_i) and (X_j, Z_j) for $i \neq j$, it holds that for $i \neq j : \Lambda_{ij} = \mathbb{E}(\delta_i\epsilon_j|X_{\mathcal{A}_1}, Z_{\mathcal{A}_1}) = \mathbb{E}_{\delta_i}(\delta_i\,\mathbb{E}(\epsilon_j|X_j, Z_j, \delta_i)) = \mathbb{E}_{\delta_i}(\delta_i\,\mathbb{E}(\epsilon_j|X_j, Z_j)) = 0$. It also holds that by the sub-Gaussian assumption of (R1), we have that [5, p.35 5.Theoretical Justification]:

$$\max_{1\leq j\leq n_1} |\Lambda_{jj}| \leq K^2 \tag{4.6.3}$$

Using all the tools above we can now prove that the first error term from (4.6.2) will vanish in probability assuming (R1) and (R2). This will be established in the next corollary, which was not mentioned in the original paper. I came up with the proof myself.

Corollary 4.6.11. *Under condition (R1) and (R2) together with* $\frac{D_{\mathcal{A}_1}^T \mathrm{M}(V) D_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1}} \xrightarrow{\mathbb{P}} 1$ *and* $\mathrm{Tr}([\mathrm{M}(V)]^2) \leq \mathrm{Tr}([\mathrm{M}(V)])$*, it holds that:*

$$\frac{\epsilon_{\mathcal{A}_1}^T \mathrm{M}(V)\delta_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^T \mathrm{M}(V) D_{\mathcal{A}_1}} \xrightarrow{\mathbb{P}} 0$$

Proof. From lemma 4.6.10, together with the assumption of this corollary on $\mathrm{Tr}([\mathrm{M}(V)]^2)$, it follows that the established upper bound can be written as (using the reverse triangle inequality):

$$\begin{aligned} |\epsilon_{\mathcal{A}_1}^T \mathrm{M}(V)\delta_{\mathcal{A}_1}| &\leq t_0 K^2 \sqrt{\mathrm{Tr}([\mathrm{M}(V)]^2)} + |\,\mathrm{Tr}(\mathrm{M}(V)\Lambda)| \\ &\leq t_0 K^2 \sqrt{\mathrm{Tr}(\mathrm{M}(V))} + |\,\mathrm{Tr}(\mathrm{M}(V)\Lambda)| \end{aligned}$$

with probability larger than $1-\exp(-c\min(t_0^2, t_0))$ conditional on $\mathcal{O}$. Choosing $t_0 = (f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1})^{1/4}$ and using (4.6.3), one yields:

$$\mathbb{P}(|\frac{\epsilon_{\mathcal{A}_1}^T \mathrm{M}(V)\delta_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1}}| \leq \frac{K^2\sqrt{\mathrm{Tr}(\mathrm{M}(V))}}{(f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1})^{3/4}} + \frac{K^2|\,\mathrm{Tr}(\mathrm{M}(V))|}{f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1}}|\mathcal{O}) \geq 1 - 6\exp(-c\min(t_0, t_0^2))$$

Hence by condition (R2): $\frac{\epsilon_{\mathcal{A}_1}^T \mathrm{M}(V)\delta_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1}}|\mathcal{O} \xrightarrow{\mathbb{P}} 0$ and consequently (application of Dominated Convergence Theorem): $\frac{\epsilon_{\mathcal{A}_1}^T \mathrm{M}(V)\delta_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1}} \xrightarrow{\mathbb{P}} 0$. The argument is finished by applying Slutsky's lemma using that $\frac{D_{\mathcal{A}_1}^T \mathrm{M}(V) D_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1}} \xrightarrow{\mathbb{P}} 1$.

□

Remark 4.6.12.

- $\mathrm{Tr}([\mathrm{M}(V)]^2) \leq \mathrm{Tr}([\mathrm{M}(V)])$ *holds for Random Forests, see section 4.6.2.*
- $\frac{D_{\mathcal{A}_1}^T \mathrm{M}(V) D_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1}} \xrightarrow{\mathbb{P}} 1$ *will hold under (R1) and (R2) (see [5, p.36 9.1 Proof of Proposition 2]). So it's not necessary to make it an extra assumption.*

Observe that for proving consistency, no explicit assumption is made on how good of an estimator $\hat{f}_{\mathcal{A}_1}$ should be of $f_{\mathcal{A}_1}$. Where $f_{\mathcal{A}_1}$ is linked to $\hat{f}_{\mathcal{A}_1}$ in the proof of corollary 4.6.11 is in the result that assuming (R1) and (R2), it holds that $\frac{D_{\mathcal{A}_1}^T \mathrm{M}(V) D_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1}} \xrightarrow{\mathbb{P}} 1$, which is equivalent

to: $\frac{\left\|\mathrm{P}^{\perp}_{\hat{V},\hat{W}}\hat{f}_{\mathcal{A}_1}\right\|^2}{\left\|\mathrm{P}^{\perp}_{\hat{V},\hat{W}}\Omega f_{\mathcal{A}_1}\right\|^2} \xrightarrow{\mathbb{P}} 1$. Hence, the comparison here is with respect to lengths which should be the same asymptotically speaking. This would be satisfied if $\hat{f}_{\mathcal{A}_1}$ is a "good enough" fit for $\Omega f_{\mathcal{A}_1}$.

The next condition is used to establish asymptotic normality of $\hat{\beta}(V)$ (as seen in definition 4.4.9):

Conditions 4.6.13 (R2-Inf)**.** $f_{\mathcal{A}_1}^T[\mathrm{M}(V)]^2 f_{\mathcal{A}_1}$ *satisfies:*

- $f_{\mathcal{A}_1}^T[\mathrm{M}(V)]^2 f_{\mathcal{A}_1} \to \infty$
- $f_{\mathcal{A}_1}^T[\mathrm{M}(V)]^2 f_{\mathcal{A}_1} \gg \|R(V)\|_2^2$
- $f_{\mathcal{A}_1}^T[\mathrm{M}(V)]^2 f_{\mathcal{A}_1} \gg \max\{(\mathrm{Tr}(\mathrm{M}(V)))^c, \log(n)\eta_n^2(V)[\mathrm{Tr}(\mathrm{M}(V))]^2\}$ *where* $c > 1$ *is some constant.*

Here:

$$\eta_n(V) = \left\|f_{\mathcal{A}_1} - \hat{f}_{\mathcal{A}_1}\right\|_\infty + (|\beta - \hat{\beta}_{\text{init}}(V)| + \|R(V)\|_\infty + \frac{\log(n)}{\sqrt{n}})(\sqrt{\log(n)} + \left\|f_{\mathcal{A}_1} - \hat{f}_{\mathcal{A}_1}\right\|_\infty)$$

Compared to (R2), it will turn out that for random forests this is a slightly stronger condition due to the fact that $f_{\mathcal{A}_1}^T[\mathrm{M}_{\mathrm{RF}}(V)]^2 f_{\mathcal{A}_1} \leq f_{\mathcal{A}_1}^T[\mathrm{M}_{\mathrm{RF}}(V)] f_{\mathcal{A}_1}$ (see section 4.6.2). Also observe that in case: $\left\|f_{\mathcal{A}_1} - \hat{f}_{\mathcal{A}_1}\right\|_\infty \xrightarrow{\mathbb{P}} 0$, $|\beta - \hat{\beta}_{\text{init}}| \xrightarrow{\mathbb{P}} 0$ and $\|R(V)\|_\infty \xrightarrow{\mathbb{P}} 0$ (at a fast enough rate), then $\eta_n(V) \xrightarrow{\mathbb{P}} 0$. This last condition of (R2-inf) essentially requires that all previously mentioned terms of $\eta_n(V)$ are "small enough".

(R2-Inf) is used to prove the following result on asymptotic normality of $\hat{\beta}(V)$:

Theorem 4.6.14. *[5, p.25 Section 5 Theorem 2](Asymptotic normality $\hat{\beta}(V)$) Suppose that (R1),(R2-Inf) holds. Furthermore assume that:*

$$\frac{\max\limits_{1\leq i\leq n_1} \sigma_i^2[\mathrm{M}(V)f_{\mathcal{A}_1}]_i^2}{\sum\limits_{i=1}^{n_1} \sigma_i^2[\mathrm{M}(V)f_{\mathcal{A}_1}]_i^2} \to 0, \sigma_i^2 = \mathbb{E}(\epsilon_i^2|Z_i, X_i) \tag{4.6.4}$$

Then:

$$\frac{1}{\mathrm{SE}(V)}(\hat{\beta}(V) - \beta) \xrightarrow{d} \mathcal{N}(0,1)$$

with

$$\mathrm{SE}(V) = \frac{\sqrt{\sum\limits_{i=1}^{n_1} \sigma_i^2[\mathrm{M}(V)f_{\mathcal{A}_1}]_i^2}}{f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1}}$$

and $\hat{\beta}(V)$ *defined as in definition 4.4.9.*
Furthermore, if $\hat{\mathrm{SE}}(V)$ *as in definition 4.4.10 satisfies:*

$$\frac{\hat{\mathrm{SE}}(V)}{\mathrm{SE}(V)} \xrightarrow{\mathbb{P}} 1$$

then:

$$\liminf_{n\to\infty} P(\beta \in \mathrm{CI}(V)) = 1 - \alpha$$

Remark 4.6.15. *Conditions for which* $\hat{\mathrm{SE}}(V)$ *is a consistent estimator for* $\mathrm{SE}(V)$ *can be found at [5, p.3 A.4 Consistency of variance estimators].*

To prove theorem 4.6.14, the original paper used the following error decomposition of $\hat{\beta}(V) - \beta$:

Lemma 4.6.16. *[5, p.36 Section 9.2]* $\hat{\beta}(V) - \beta = \mathcal{G}(V) + \mathcal{E}(V)$ *with:*

$$\mathcal{G}(V) = \frac{1}{\mathrm{SE}(V)} \frac{\epsilon_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^T \mathrm{M}(V) D_{\mathcal{A}_1}}$$
$$\mathcal{E}(V) = \frac{1}{\mathrm{SE}(V)} \frac{R_{\mathcal{A}_1}^T \mathrm{M}(V) D_{\mathcal{A}_1} - \mathrm{Err}_1(V) - \mathrm{Err}_2(V)}{D_{\mathcal{A}_1}^T \mathrm{M}(V) D_{\mathcal{A}_1}}$$

and with Err_1 *and* Err_2 *defined as:*

$$\mathrm{Err}_1(V) = \sum_{i,j=1}^{n_1} [\mathrm{M}(V)]_{ij} \delta_i \epsilon_j$$
$$\mathrm{Err}_2(V) = \sum_{i=1}^{n_1} [\mathrm{M}(V)]_{ii} (f_i - \hat{f}_i) \hat{\epsilon}_i(V) + \sum_{i=1}^{n_1} [\mathrm{M}(V)]_{ii} \delta_i (\epsilon_i - \hat{\epsilon}_i(V))$$

Proof. Using lemma 4.4.7, observe that:

$$\hat{\beta}(V) - \beta = \frac{D_{\mathcal{A}_1}^T \mathrm{M}(V) R_{\mathcal{A}_1} + D_{\mathcal{A}_1}^T \mathrm{M}(V) \epsilon_{\mathcal{A}_1} - \sum_{i=1}^{n_1} [\mathrm{M}(V)]_{ii} \hat{\delta}_i \hat{\epsilon}_i}{D_{\mathcal{A}_1}^T \mathrm{M}(V) D_{\mathcal{A}_1}}$$

It holds that:

$$D_{\mathcal{A}_1}^T \mathrm{M}(V) \epsilon_{\mathcal{A}_1} = f_{\mathcal{A}_1}^T \mathrm{M}(V) \epsilon_{\mathcal{A}_1} - \delta_{\mathcal{A}_1}^T \mathrm{M}(V) \epsilon_{\mathcal{A}_1}$$

Now consider:

$$\delta_{\mathcal{A}_1}^T \mathrm{M}(V) \epsilon_{\mathcal{A}_1} + \sum_{i=1}^{n_1} [\mathrm{M}(V)]_{ii} \hat{\delta}_i \hat{\epsilon}_i =$$
$$\delta_{\mathcal{A}_1}^T \mathrm{M}(V) \epsilon_{\mathcal{A}_1} + \sum_{i=1}^{n_1} [\mathrm{M}(V)]_{ii} (f_i - \hat{f}_i) \hat{\epsilon}_i - \sum_{i=1}^{n_1} [\mathrm{M}(V)]_{ii} \delta_i \hat{\epsilon}_i =$$
$$\delta_{\mathcal{A}_1}^T \mathrm{M}(V) \epsilon_{\mathcal{A}_1} + \sum_{i=1}^{n_1} [\mathrm{M}(V)]_{ii} (f_i - \hat{f}_i) \epsilon_i - \sum_{i=1}^{n_1} [\mathrm{M}(V)]_{ii} \delta_i (\hat{\epsilon}_i - \epsilon_i) - \sum_{i=1}^{n_1} [\mathrm{M}(V)]_{ii} \delta_i \epsilon_i$$

Lastly, using that $\delta_{\mathcal{A}_1}^T \,\mathrm{M}(V)\epsilon_{\mathcal{A}_1} = \sum_{i,j=1}^{n_1} \delta_i [\mathrm{M}(V)]_{ij}\epsilon_j$, the established identities above and some reordering we obtain the desired decomposition. □

Remark 4.6.17. *Observe that the $(f_i - \hat{f}_i)$ term (which can be observed in the Err_2 term above) comes back in assumption (R2-Inf), namely as $\left\|f_i - \hat{f}_i\right\|_\infty$ in the $\eta_n(V)$ term.*

Next, it is shown that $\mathcal{G}(V)$ is asymptotically normal. The statement is slightly different from the one presented in [5] (which is discussed after the proof). I constructed the proof myself.

Lemma 4.6.18. *Assuming (R1), (R2-inf) and*

$$\frac{1}{\sum_{k=1}^{n} [\mathrm{M}(V)f_{\mathcal{A}_1}]_k^2 \sigma_k^2} \sum_{k=1}^{n} \mathbb{E}(\epsilon_k^2 [\mathrm{M}(V)f_{\mathcal{A}_1}]_k^2 1(|\epsilon_k^2 [\mathrm{M}(V)f_{\mathcal{A}_1}]_k^2| > \epsilon \sqrt{\sum_{k=1}^{n} [\mathrm{M}(V)f_{\mathcal{A}_1}]_k^2 \sigma_k^2}) | \mathcal{O}) \to 0$$

where $\sigma_k^2 = \mathbb{E}(\epsilon_k^2 | X_k, Z_k)$ together with $\frac{D_{\mathcal{A}_1}^T \,\mathrm{M}(V) D_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \,\mathrm{M}(V) f_{\mathcal{A}_1}} \xrightarrow{\mathbb{P}} 1$, we obtain that:

$$\mathcal{G}(V) \xrightarrow{d} \mathcal{N}(0,1)$$

Proof. Observe that:

$$\epsilon_{\mathcal{A}_1}^T \,\mathrm{M}(V) f_{\mathcal{A}_1} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \epsilon_i \,\mathrm{M}(V)_{ij} f_j = \sum_{i=1}^{n_1} \epsilon_i [\mathrm{M}(V) f_{\mathcal{A}_1}]_i$$

Define $\tilde{X}_i := \epsilon_i [\mathrm{M}(V) f_{\mathcal{A}_1}]_i$ for $i = 1, ..., n_1$. Then it holds that $\tilde{X}_i | \mathcal{O}$ are independent together with: $\mathbb{E}(\tilde{X}_i | \mathcal{O}) = 0$ and $\mathrm{Var}(\tilde{X}_i | \mathcal{O}) = [\mathrm{M}(V) f_{\mathcal{A}_1}]_i^2 \sigma_i^2$. As (the conditional) Lindeberg's condition holds by assumption, we get:

$$\frac{\epsilon_{\mathcal{A}_1}^T \,\mathrm{M}(V) f_{\mathcal{A}_1}}{\sqrt{\sum_{i=1}^{n_1} [\mathrm{M}(V) f_{\mathcal{A}_1}]_i^2 \sigma_i^2}} | \mathcal{O} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

Hence, also:

$$\frac{\epsilon_{\mathcal{A}_1}^T \,\mathrm{M}(V) f_{\mathcal{A}_1}}{\sqrt{\sum_{i=1}^{n_1} [\mathrm{M}(V) f_{\mathcal{A}_1}]_i^2 \sigma_i^2}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

The argument is finished by applying Slutsky's lemma. □

When comparing the conditions of lemma 4.6.18 with the conditions of theorem 4.6.14, it can be readily noticed that I directly assumed Lindeberg's condition in lemma 4.6.18 while in theorem 4.6.14 (4.6.4) is assumed. In remark A.0.2 it can be seen that Lindeberg's condition implies (4.6.4). The original paper (which mistakenly refers to "Linderberg's condition" rather than "Lindeberg's condition) states that (4.6.4) implies Lindeberg's condition without further detail [5, p.36 Proof of Theorems 1 and 2]. In general, this is not the case and I don't see

evidence that (R1) and (R2-Inf) together with (4.6.4) imply Lindeberg's condition. Regardless, directly assuming Lindeberg's condition will give the desired end-result of asymptotic normality of $\mathcal{G}(V)$.

Thus far, we have assumed that V is a proper basis. In method 4.4.16, we choose among $(V_q)_{0\leq q\leq Q_{\max}}$ and in the next bit we will consider results relating to method 4.4.16.

When considering $\hat{\beta}(V_q) - \hat{\beta}(V_{q'})$ for $q \neq q'$, one can observe that:

$$S^T\epsilon_{\mathcal{A}_1} := (\frac{f_{\mathcal{A}_1}^T M(V_q)}{f_{\mathcal{A}_1}^T M(V_q) f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^T M(V_{q'})}{f_{\mathcal{A}_1}^T M(V_{q'}) f_{\mathcal{A}_1}})\epsilon_{\mathcal{A}_1}$$

is a dominating term (for large n_1) using the second part of remark 4.6.12, proposition 4.6.5 and assuming that $\|R(V_q)\|$ and $\|R(V_{q'})\|$ are small enough. For $S^T\epsilon_{\mathcal{A}_1}$ it holds that $\mathrm{Var}(S^T\epsilon_{\mathcal{A}_1}|\mathcal{O}) = H(V_q, V_{q'})$, where $H(V_q, V_{q'})$ is defined as:

$$H(V_q, V_{q'}) := \frac{\sum_{i=1}^{n_1}\sigma_i^2[M(V_{q'})f_{\mathcal{A}_1}]_i^2}{[f_{\mathcal{A}_1}^T M(V_{q'})f_{\mathcal{A}_1}]^2} + \frac{\sum_{i=1}^{n_1}\sigma_i^2[M(V_q)f_{\mathcal{A}_1}]_i^2}{[f_{\mathcal{A}_1}^T M(V_q)f_{\mathcal{A}_1}]^2} - 2\frac{\sum_{i=1}^{n_1}\sigma_i^2[M(V_{q'})f_{\mathcal{A}_1}]_i[M(V_q)f_{\mathcal{A}_1}]_i}{[f_{\mathcal{A}_1}^T M(V_q)f_{\mathcal{A}_1}][f_{\mathcal{A}_1}^T M(V_q)f_{\mathcal{A}_1}]}$$

The following condition requires that that this variance $H(V_q, V_{q'})$ dominates the approximation errors of estimating f by $\hat{f}$ and $g_{\mathcal{A}_1}$ by the column space of $(V_{Q_{\max}}, W)$:

Conditions 4.6.19 (R3)**.** *The variance $H(V_q, V_{q'})$ satisfies:*

- $\sqrt{H(V_q, V_{q'})} \gg \max_{V\in\{V_q, V_{q'}\}} \{\frac{1}{\mu(V)}[1 + (1 + \sqrt{\log(n)}\eta_n(V_{Q_{\max}})\,\mathrm{Tr}(\mathrm{M}(V))\}$
- $\exists C > 0 : \mathrm{Var}(\delta_i|Z_i, X_i) \geq C$

Remark 4.6.20. *With the second bullet point of (R3), we have that $\mu(V)$ is proportional to $f_{\mathcal{A}_1}^T\,\mathrm{M}(V)f_{\mathcal{A}_1}$.*

The following theorem is the basis as for why method 4.4.16 works (asymptotically):

Theorem 4.6.21. *[5, p.25 Section 5 Theorem 3](Asymptotic normality $\hat{\beta}(V_q) - \hat{\beta}(V_{q'})$) Suppose (R1) and (R2) hold for $V \in \{V_q, V_{q'}\}$, (R3) holds and S satisfies:*

$$\frac{\max_{i\in\mathcal{A}_1} S_i^2}{\sum_{i\in\mathcal{A}_1} S_i^2} \to 0$$

If

$$\sqrt{H(V_q, V_{q'})} \gg \max_{V\in\{V_q, V_{q'}\}} \|R(V)\|_2 / \sqrt{\mu(V)} \tag{4.6.5}$$

then we have:

$$\frac{\hat{\beta}(V_q) - \hat{\beta}(V_{q'})}{\sqrt{H(V_q, V_{q'})}} \xrightarrow{d} \mathcal{N}(0, 1)$$

To see that V_q and $V_{q'}$ from theorem 4.6.21 above are indeed "good" basis choices, one might wonder whether $\hat{\beta}_{\text{init}}(V_q)$ and $\hat{\beta}_{\text{init}}(V_{q'})$ are consistent estimators of β. In case $H(V_q, V_{q'})$ is bounded, this will be indeed the case (see proposition 4.6.5). When we compare theorem 4.6.21 to theorem 4.6.14, one can see that S_i looks at differences between individual components (related to the asymptotic variance $H(V_q, V_{q'})$) rather than just the individual components. So there is no obvious guarantee that under the setting of theorem 4.6.21, for example, $\frac{1}{\text{SE}(V_q)}(\beta(\hat{V}_q) - \beta)$ is asymptotically normal.

4.6.2. Properties of $\text{M}_{\text{RF}}(V)$

In this section, some of the (R1)-(R3) conditions from the previous section are proved for random forests in specific.

The following lemma summarises some (well-known) identities used in this section for the proofs:

Lemma 4.6.22. *For a (real) matrix $A : m \times m$ and $x : m \times 1$, the following identities hold:*

1. $\|Ax\|_2 \leq \|A\|_2 \|x\|_2$
2. $\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty$
3. $\lambda_k(A) = \max_{U \subseteq \mathbb{R}^m, \dim(U)=k} \min_{u \in U} \frac{u^T A u}{\|u\|_2^2}$, *$U$ is a linear subspace of $\mathbb{R}^m$.*
4. $\|\Omega_{\text{RF}}\|_1 = \|\Omega_{\text{RF}}\|_\infty = 1$

Proof. (4): $\|\Omega_{\text{RF}}\|_\infty = 1$ is a direct consequence of lemma 4.4.3. $\|\Omega_{\text{RF}}\|_1 = 1$ follows from the fact that: $w_j(X_i, Z_i, \theta_s) = w_i(X_j, Z_j, \theta_s) \forall i, j \in \{1, ..., n_1\}, s \in \{1, ..., S\}$ □

The resulting properties for $\text{M}_{\text{RF}}(V)$, as found in the next lemma, were first stated in [5, p.3 A.3 Lemma 5] where I filled in the details of the proof as seen at [5, p.13 C.4].

Lemma 4.6.23. *The transformation matrix $\text{M}_{\text{RF}}(V)$ satisfies*

$$\lambda_{\max}(\text{M}_{\text{RF}}(V)) \leq 1, \quad b^T[\text{M}_{\text{RF}}(V)]^2 b \leq b^T \text{M}_{\text{RF}}(V) b \quad \forall b \in \mathbb{R}^{n_1}$$

Consequently: $\text{Tr}([\text{M}_{\text{RF}}(V)]^2) \leq \text{Tr}(\text{M}_{\text{RF}}(V))$.

Proof. In the next bit, "L.4.6.23" refers to lemma 4.6.23.

1. Showing: $b^T[\text{M}_{\text{RF}}(V)]^2 b \leq b^T \text{M}_{\text{RF}}(V) b$
   - $b^T[\text{M}_{\text{RF}}(V)]^2 b = \left\|\Omega_{\text{RF}}^T \text{P}^\perp_{\hat{V}_{\mathcal{A}_1}, \hat{W}_{\mathcal{A}_1}} \Omega_{\text{RF}} b\right\|_2^2 \overset{\text{L.4.6.23(1)}}{\leq} \|\Omega_{\text{RF}}^T\|_2^2 \left\|\text{P}^\perp_{\hat{V}_{\mathcal{A}_1}, \hat{W}_{\mathcal{A}_1}} \Omega_{\text{RF}} b\right\|_2^2$
   - $\left\|\text{P}^\perp_{\hat{V}, \hat{W}} \Omega b\right\|_2^2 = b^T \text{M}_{\text{RF}}(V) b$ and $\|\Omega_{\text{RF}}\|_2^2 \overset{\text{L.4.6.23(2)}}{\leq} \|\Omega_{\text{RF}}\|_1 \|\Omega_{\text{RF}}\|_\infty \overset{\text{L.4.6.23(4)}}{=} 1$
   - As $\Omega_{\text{RF}}^T = \Omega_{\text{RF}}$, the argument is finished.
2. Showing: $\lambda_{\max}(\text{M}_{\text{RF}}(V)) \leq 1$:
   - $b^T \text{M}_{\text{RF}}(V) b = b^T \Omega_{\text{RF}}^T \text{P}^\perp_{\hat{V}_{\mathcal{A}_1}, \hat{W}_{\mathcal{A}_1}} \Omega_{\text{RF}} b = \left\|\text{P}^\perp_{\hat{V}_{\mathcal{A}_1}, \hat{W}_{\mathcal{A}_1}} \Omega_{\text{RF}} b\right\|_2^2$

- $\left\|\mathrm{P}^{\perp}_{\hat{V}_{\mathcal{A}_1},\hat{W}_{\mathcal{A}_1}}\Omega_{\mathrm{RF}}b\right\|_2^2 \overset{(*)}{\leq} \|\Omega_{\mathrm{RF}}b\|_2^2 \overset{\text{L.4.6.23(1)}}{\leq} \|\Omega_{\mathrm{RF}}\|_2^2\,\|b\|_2^2 \overset{\text{L.4.6.23(2),(4)}}{\leq} \|b\|_2^2$

where in (*) above it is used that for any vector $a \in \mathbb{R}^m$ and any projection matrix $Q \in \mathbb{R}^{m\times m}$: $\|a\|_2 = \|Qa\|_2 + \left\|Q^{\perp}a\right\|_2$.
Hence by the max-min theorem we obtain that $\forall k$:

$$\lambda_k(\mathrm{M}_{\mathrm{RF}}(V)) \leq 1$$

3. Showing: $\mathrm{Tr}([\mathrm{M}_{\mathrm{RF}}(V)]^2) \leq \mathrm{Tr}(\mathrm{M}_{\mathrm{RF}}(V))$
   - As the trace is the sum of the eigenvalues, it suffices to show $\lambda_k([\mathrm{M}_{\mathrm{RF}}(V)]^2) \leq \lambda_k(\mathrm{M}_{\mathrm{RF}}(V))$ for any k.
   - As we already established before that $u^T[\mathrm{M}_{\mathrm{RF}}(V)]^2u \leq u^T\,\mathrm{M}_{\mathrm{RF}}(V)u$ for any $u \in \mathbb{R}^{n_1}$: the max-min theorem give the desired result.

□

4.7. Simulation studies

4.7.1. Set-up

In the context of definition 4.2.1 and definition 4.2.2, the following data (which is a reflection of the data used in the original paper [5]) will be used :
Set $\beta = 1$, $p_X = 5$ and $p_Z = 1$. Define $p = p_X + p_Z$. Generate $X_i^* \sim \mathcal{N}_p(0, \Sigma)$ with $\Sigma_{ij} = 2^{-|i-j|}$. Then define $X_i = (\Phi(X_{i1}^*), ..., \Phi(X_{i(p-1)}^*)^T$ and $Z_i = 4(\Phi(X_{ip}^*) - \frac{1}{2})$. Independently from (X_i, Z_i), sample: $(\delta_i, \epsilon_i) \overset{\text{iid}}{\sim} \mathcal{N}_2(0, \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix})$.
As X_i and Z_i are bounded and (δ_i, ϵ_i) a sub-Gaussian vector (independent of (X_i, Z_i)), most conditions of (R1) from section 4.6.1 will be satisfied for the (V, W) and f considered in the next part.

I will be using that implementation of the TSCI method with Random Forests as found at: https://github.com/zijguo/TSCI-Replication. This implementation will output TSCI (with Random Forests by default) for two estimators used in the first stage: $\hat{\beta}_{\text{init}}$ defined as in definition 4.4.9 and $\tilde{\beta}_{\text{RF}}$. $\tilde{\beta}_{\text{RF}}$ has not been previously defined in this text. It is also a bias-correcting estimator (just as $\hat{\beta}_{\text{RF}}$ as seen at definition 4.4.9) but for the asymptotic properties (such as asymptotic normality) it requires the extra assumption that $\text{Cov}(\epsilon_i, \delta_i | Z_i, X_i) = \text{Cov}(\epsilon_i, \delta_i)$ (see [5, p.1 A. Additional Discussions A.1] for more details). It is therefore more restrictive then $\hat{\beta}_{\text{RF}}$. $\tilde{\beta}_{\text{RF}}$ is defined as follows:

Definition 4.7.1.

$$\tilde{\beta}_{\text{RF}}(V) = \hat{\beta}_{\text{init}}(V) - \frac{\hat{\text{Cov}}(\epsilon_i, \delta_i)\,\text{Tr}(\text{M}_{\text{RF}}(V))}{D_{\mathcal{A}_1}^T\,\text{M}_{\text{RF}}(V) D_{\mathcal{A}_1}}$$

$$\hat{\text{Cov}}(\delta_i, \epsilon_i) = \frac{1}{n_1 - r}(D_{\mathcal{A}_1} - \hat{f}_{\mathcal{A}_1})\text{P}^{\perp}_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}[Y_{\mathcal{A}_1} - D_{\mathcal{A}_1}\hat{\beta}_{\text{init}}(V)], r = \text{rank}((V, W))$$

For the simulations, I kept the method mostly to its default options as this is the direct translation from how the method was introduced in section 4.4. The one setting that was not put to the default mode was that number of trees per random forest which I put to 50. One other remark on the implementation is that by default $W_i = (X_i^{(1)}, X_i^{(2)}, ..., X_i^{(p_X)})^T$ and it does not allow for other inputs of W_i. The user can specify the input for V_i.
In this simulation study, 3 topics are considered: Firstly 2 cases where g could be perfectly estimated by (V, W) (where then the main question is whether the method indeed perfectly estimates them), secondly the IV-invalidity test is checked for a quadratic case and lastly a case of a more involved f term is considered. These models will mostly be evaluated based on the bias of $\hat{\beta}_{\text{init}}$, bias of $\tilde{\beta}_{\text{RF}}$, coverage based on $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$, CI-length based on $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$ and which $\hat{q}_c$ is chosen for different n. Every study is repeated 100 times and the corresponding outcomes are averaged (and rounded in the case of $\hat{q}_c$).

4.7.2. Models with possible perfect estimation g

Here, two models will be considered:

$$\text{(M1)}\quad g_1(Z_i, X_i) = 1 + \sum_{j=1}^{p_X} X_i^{(j)}$$
$$f_1(Z_i, X_i) = 1 + \sum_{j=1}^{p_X} X_i^{(j)} + b(\cos(2\pi Z_i) + Z_i \sum_{j=1}^{p_X} X_i^{(j)}), b \in \{0, \tfrac{1}{2}, 1, 5\}$$

$$\text{(M2)}\quad g_2(Z_i, X_i) = \sum_{q=0}^{10} Z_i^q + aZ_i^{11} + \sum_{j=1}^{p_X} X_i^{(j)},\ a \in \{\tfrac{1}{2}, 1\}$$
$$f_2(Z_i, X_i) = \tfrac{1}{2} Z_i^5 + b(\cos(2\pi Z_i) + Z_i \sum_{j=1}^{p_X} X_i^{(j)}), b \in \{\tfrac{1}{2}, 1\}$$

For (M1), $\mathcal{V} = \{1, z, z^2\}$ is chosen as its largest possible basis. In tables 4.1 to 4.4 the corresponding results can be found.

Bias β_{init} (M1)	$n = 100$	300	500	1000
$b = 0$	0.49	0.47	0.48	0.51
$\frac{1}{2}$	0.20	0.05	0.05	0.03
1	0.08	0.03	0.02	0.01
5	0.01	0.005	0.004	0.002

Table 4.1.: Results for the bias of $\beta_{\text{init}}(V_{\hat{q}_c})$ corresponding to model (M1) with $n \in \{100, 300, 500, 1000\}$ and $b \in \{0, \frac{1}{2}, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$

Bias $\tilde{\beta}_{\text{RF}}$ (M1)	$n = 100$	300	500	1000
$b = 0$	0.49	0.47	0.49	0.52
$\frac{1}{2}$	0.23	0.06	0.05	0.03
1	0.1	0.03	0.02	0.01
5	0.01	0.005	0.004	0.003

Table 4.2.: Results for the bias of $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$ corresponding to model (M1) with $n \in \{100, 300, 500, 1000\}$ and $b \in \{0, \frac{1}{2}, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$

Cov $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$ (M1)	$n = 100$	300	500	1000
$b = 0$	0.24	0.28	0.19	0.1
$\frac{1}{2}$	0.68	0.91	0.85	0.92
1	0.78	0.96	0.95	0.93
5	0.86	0.94	0.96	0.94

Table 4.3.: Results for the coverage of $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$ corresponding to model (M1) with $n \in \{100, 300, 500, 1000\}$ and $b \in \{0, \frac{1}{2}, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$

CI $\hat{\beta}_{\text{RF}}(V_{\hat{q}_c})$ (M1)	$n = 100$	300	500	1000
$b = 0$	0.62	0.62	0.56	0.46
$\frac{1}{2}$	0.45	0.23	0.18	0.12
1	0.23	0.12	0.08	0.06
5	0.05	0.02	0.02	0.01

Table 4.4.: Results for the CI-length of $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$ corresponding to model (M1) with $n \in \{100, 300, 500, 1000\}$ and $b \in \{0, \frac{1}{2}, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$

For all n and b, the most frequently chosen basis for the violation function h was $\{1\}$. For all n: $b = 0$ has, by far, the worst performance across the board. This is not surprising as there is no functional differences between g_1 and f_1 in this case and the chosen basis (namely $(1, x^{(1)}, ..., x^{(p_X)})$) can perfectly represent f as well as g. So, from the identification section we already expected the algorithm to perform poorly here. From a theoretical point of view this translates to $f_{\mathcal{A}_1}^T \mathrm{M}(V) f_{\mathcal{A}_1} = 0$, hence (R2) is not satisfied. For the other b's: there is in general an improvement as n increases.
For a fixed n: the performance improves as b increases. This means that the TSCI-algorithm here performs better in case g_1 and f_1 are more distinguishable. The performance at $b = 5$ is of a different order than the others across the categories (expect for the $b = 1$ coverage where it's the same order). This is not surprising as the algorithm is built on the idea that there are functional differences between f and g which are exploited as $n \to \infty$.
In section 4.7.5, I will have a generalised discussion on the results of the bias of $\hat{\beta}_{\text{init}}$ versus that of $\tilde{\beta}_{\text{RF}}$.

The results of model M2 can be found in tables 4.5 to 4.9. Here, $n = 300$ is used for all cases with $\mathcal{V} = \{1, z, ..., z^{12}\}$.

Bias $\beta_{\text{init}}(V_{\hat{q}_c})$ (M2)	$a = \frac{1}{2}$	1
$b = \frac{1}{2}$	84.49	144.61
1	27.88	58.47

Table 4.5.: Results for the bias of $\beta_{\text{init}}(V_{\hat{q}_c})$ corresponding to model (M2) with $a \in \{\frac{1}{2}, 1\}$ and $b \in \{\frac{1}{2}, 1\}$ using $\mathcal{V} = \{1, z, ..., z^{12}\}$ and $n = 300$

Bias $\tilde{\beta}(V_{\hat{q}_c})$ (M2)	$a = \frac{1}{2}$	1
$b = \frac{1}{2}$	90.13	152.23
1	29.55	61.55

Table 4.6.: Results for the bias of $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$ corresponding to model (M2) with $a \in \{\frac{1}{2}, 1\}$ and $b \in \{\frac{1}{2}, 1\}$ using $\mathcal{V} = \{1, z, ..., z^{12}\}$ and $n = 300$

Cov $\tilde{\beta}(V_{\hat{q}_c})$ (M2)	$a = \frac{1}{2}$	1
$b = \frac{1}{2}$	0	0
1	0.32	0.29

Table 4.7.: Results for the coverage of $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$ corresponding to model (M2) with $a \in \{\frac{1}{2}, 1\}$ and $b \in \{\frac{1}{2}, 1\}$ using $\mathcal{V} = \{1, z, ..., z^{12}\}$ and $n = 300$

CI $\tilde{\beta}(V_{\hat{q}_c})$ (M2)	$a = \frac{1}{2}$	1
$b = \frac{1}{2}$	33.27	47.33
1	12.62	23.16

Table 4.8.: Results for the CI-length of $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$ corresponding to model (M2) with $a \in \{\frac{1}{2}, 1\}$ and $b \in \{\frac{1}{2}, 1\}$ using $\mathcal{V} = \{1, z, ..., z^{12}\}$ and $n = 300$

$\hat{q}_c$ (M2)	$a = \frac{1}{2}$	1
$b = \frac{1}{2}$	3	3
1	7	6

Table 4.9.: Results for the $\hat{q}_c$ choice using $\tilde{\beta}_{\text{RF}}$ corresponding to model (M2) with $a \in \{\frac{1}{2}, 1\}$ and $b \in \{\frac{1}{2}, 1\}$ using $\mathcal{V} = \{1, z, ..., z^{12}\}$ and $n = 300$

For a fixed a: as b increases so does the performance. Increasing b will lead to more functional difference between g_2 and f_2, hence a better performance.
For a fixed b and as a increases the performance gets gradually worse. When we also consider that $\hat{q}_c$ does not choose the basis to perfectly estimate the violation function: increasing a exaggerates the error due to the "wrong" basis choice.
Compared to (M1), model (M2) is much more involved and thus needs more data to perform well. One of the difficulties of this model is the concern of underfitting f_2. In general, if we choose a large basis to fit the violation function the chance increases that this basis also approximates the estimated $\hat{f}$ well. This would result in a smaller $f_{\mathcal{A}_1}^T \text{M}(V) f_{\mathcal{A}_1}$ and hence a slower convergence. This under fitting concern leads to the "wrong" choice being made for $\hat{q}_c$. As the number of data points grows and $\hat{f}$ gets closer to f, the $\hat{q}_c$ likely become larger.
One last remark on this model is that its running time was very slow: it took multiple hours to run this for "just" $n = 300$. In general, this method takes a lot longer to run than the method used for causal inference for linear models in section 2. This has multiple reasons, but the notable ones are:

1. For each simulation we have to fit 50 decision trees (for each random forest) and it has an automatic parameter tuning process for the random forest. So not only do we compute one random forest: there is first parameter tuning where multiple random forests are computed beforehand.
2. For each possible basis, it has to fit a random forest in order to compute the generalised IV strength and then to compute different bases. So in the case of (M2), it has to test (at most) 13 different bases.
3. The whole process (with all the slowing factors above) is repeated 100 times for each model of which there are 4 for (M2).

4.7.3. IV-invalidity test

For this, the following model will be used:

$$\text{(M3)} \quad g_3(Z_i, X_i) = 1 + a(Z_i + Z_i^2) + \sum_{j=1}^{p_X} X_i^{(j)}, a \in \{\tfrac{1}{8}, \tfrac{1}{4}, \tfrac{1}{2}\}$$
$$f_3(Z_i, X_i) = -\tfrac{1}{2} + 0.2 \sum_{j=1}^{p_X} X_i^{(j)} + b(\cos(2\pi Z_i) + Z_i \sum_{j=1}^{p_X} X_i^{(j)}), b \in \{0, 1, 5\}$$

Its results can be found in tables 4.15 to 4.20. Here, $n = 500$ with $\mathcal{V} = \{1, z, z^2\}$ are used.

Bias $\beta_{\text{init}}(V_{\hat{q}_c})$ (M3)	$a = \frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$
$b = 0$	0.49	0.49	0.59
1	0.06	0.09	0.05
5	0.01	0.02	0.01

Table 4.10.: Results for the Bias of $\hat{\beta}_{\text{init}}(V_{\hat{q}_c})$ using corresponding to model (M3) with $a \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}\}$ and $b \in \{0, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$ and $n = 500$

Bias $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$ (M3)	$a = \frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$
$b = 0$	0.51	0.57	0.89
1	0.05	0.09	0.05
5	0.01	0.02	0.01

Table 4.11.: Results for the Bias of $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$ corresponding to model (M3) with $a \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}\}$ and $b \in \{0, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$ and $n = 500$

Cov $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$ (M3)	$a = \frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$
$b = 0$	0.28	0.32	0.28
1	0.37	0.33	0.89
5	0.37	0.32	0.91

Table 4.12.: Results for the coverage of $\tilde{\beta}_{\text{RF}}(V_{\hat{q}_c})$ corresponding to model (M3) with $a \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}\}$ and $b \in \{0, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$ and $n = 500$

$\hat{q}_c$ $\tilde{\beta}_{\text{RF}}$ (M3)	$a = \frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$
$b = 0$	1	1	1
1	0	1	2
5	0	1	2

Table 4.13.: Results for choice of $\hat{q}_c$ using $\tilde{\beta}_{\text{RF}}$ corresponding to model (M3) with $a \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}\}$ and $b \in \{0, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$ and $n = 500$

IV-invalidity $\tilde{\beta}_{\text{RF}}$ (M3)	$a = \frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$
$b = 0$	0.51	0.89	0.98
1	0.10	0.36	0.98
5	0.08	0.37	0.94

Table 4.14.: Results of the IV-invalidity test (i.e. if $\hat{q}_c \neq 0$ test is satisfied and conclude that Z is an invalid IV) using $\tilde{\beta}_{\text{RF}}$ corresponding to model (M3) with $a \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}\}$ and $b \in \{0, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$ and $n = 500$

One can readily observe that there is a drastic improvement in the invalid IV recognition as a becomes larger. Before $a = \frac{1}{2}$, there is too little data to notice the quadratic nature of the $a(Z_i + Z_i^2)$ term. For $a = \frac{1}{8}$,the most often chosen $\hat{q}_c$ is 0 (i.e. no IV in the model), for $a = \frac{1}{4}$, $\hat{q}_c = 1$ (i.e. there is an IV but it is linear) and then finally for $a = \frac{1}{2}$ $\hat{q}_c = 2$. Even for $b = 0$, the IV-invalidity test gives almost always the correct result for $a = \frac{1}{2}$ and $a = \frac{1}{4}$ even though it most often chose $\hat{q}_c = 1$. The coverage for $b = 1, 5$ and $a = \frac{1}{2}$ is, by far, the best. Here, the

term $a(Z_i + Z_i^2)$ is sufficiently large to be noticed by the test and $b(\cos(2\pi Z_i) + Z_i \sum_{j=1}^{n} X_i^{(j)})$ is also sufficiently large for the test to distinguish the functional forms of g_3 and f_3.

4.7.4. More complex form for f

For this part, I will involve a more complex term in the f function. The following model is used:

(M4) $g_4(Z_i, X_i) = Z_i + \frac{1}{2}\sum_{j=1}^{p_X} X_i^{(j)}$

$f_4(Z_i, X_i) = \cos(2\pi Z_i) + a\sin(2\pi Z_i)(\sum_{j=1}^{p_X} X_i^{(j)})\exp(Z_i) + b(Z_i + \frac{1}{2}Z_i^2)$

$a, b \in \{0, 1, 5\}$

Its results for the coverage, CI length and $\hat{q}_c$ choice can be found in tables 4.17 to **??**.

Bias $\beta_{\text{init}}(\hat{q}_c)$ (M4)	$a = 0$	1	5
$b = 0$	0.18	0.006	0.001
1	0.06	0.007	0.002
5	0.02	0.008	0.001

Table 4.15.: Results of the bias using $\beta_{\text{init}}(\hat{q}_c)$ corresponding to model (M4) with $a \in \{0, 1, 5\}$ and $b \in \{0, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$ and $n = 500$

Bias $\tilde{\beta}_{\text{RF}}(\hat{q}_c)$ (M4)	$a = 0$	1	5
$b = 0$	0.14	0.006	0.001
1	0.06	0.007	0.002
5	0.02	0.009	0.001

Table 4.16.: Results of the bias using $\tilde{\beta}_{\text{RF}}(\hat{q}_c)$ corresponding to model (M4) with $a \in \{0, 1, 5\}$ and $b \in \{0, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$ and $n = 500$

Cov $\tilde{\beta}_{\text{RF}}(\hat{q}_c)$ (M4)	$a = 0$	1	5
$b = 0$	0.75	0.94	0.94
1	0.93	0.94	0.88
5	0.92	0.93	0.93

Table 4.17.: Results of the coverage using $\tilde{\beta}_{\text{RF}}(\hat{q}_c)$ corresponding to model (M4) with $a \in \{0, 1, 5\}$ and $b \in \{0, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$ and $n = 500$

CI-length $\tilde{\beta}_{\text{RF}}(\hat{q}_c)$ (M4)	$a = 0$	1	5
$b = 0$	0.40	0.03	0.006
1	0.30	0.03	0.006
5	0.08	0.04	0.006

Table 4.18.: Results of the coverage using $\tilde{\beta}_{\text{RF}}(\hat{q}_c)$ corresponding to model (M4) with $a \in \{0, 1, 5\}$ and $b \in \{0, 1, 5\}$ using $\mathcal{V} = \{1, z, z^2\}$ and $n = 500$

$\hat{q}_c$ (M4)	$a=0$	1	5
$b=0$	1	2	2
1	1	2	2
5	1	2	2

Table 4.19.: Results of $\hat{q}_c$ corresponding to model (M4) with $a \in \{0,1,5\}$ and $b \in \{0,1,5\}$ using $\mathcal{V} = \{1, z, z^2\}$ and $n = 500$

IV-invalidity (M4)	$a=0$	1	5
$b=0$	1	1	1
1	1	1	1
5	1	1	1

Table 4.20.: Results of the IV-invalidity test (1 if there are invalid IVs present) using $\tilde{\beta}_{\hat{q}_c}$ corresponding to model (M4) with $a \in \{0,1,5\}$ and $b \in \{0,1,5\}$ using $\mathcal{V} = \{1, z, z^2\}$ and $n = 500$

It is noticeable that as a and b increase, the results across the board improve. One exception to this are the coverage levels which stay quite steady once either a or b are not 0 anymore. In the other categories, one can see the most drastic improvement when fixing b and increasing a. This is not unsurprising as with a larger a, the differences in functional form between g_4 and f_4 will become more exaggerated. This, in turn, will then result into more precise estimators and smaller confidence intervals for β while using the same number of data points.
One interesting observation is that once $a \neq 0$, $\{1, z, z^2\}$ is chosen for $\hat{q}_c$ as the best basis for the violation function, Though, it is true that $\{1, z\}$ is the most efficient basis to estimate the violation function of g_4, due to the $a \neq 0$ adding extra functional difference between g_4 and f_4 there is likely more leeway to choose bigger bases. There is less concern now for not noticing enough difference between g_4 and f_4 if $\hat{q}_c$ is chosen too big.

4.7.5. Bias $\hat{\beta}_{\text{init}}$ vs bias $\tilde{\beta}_{\text{RF}}$

For (M1)-(M4), it can be observed that the bias of $\tilde{\beta}_{\text{RF}}$ does not, in general, outperform the bias β_{init}. This could be due to the number of data points used being too small. Hence, doing simulations with more data is likely to lead to $\tilde{\beta}_{\text{RF}}$ outperforming β_{init} with respect to the bias. These observations about the bias do not by itself contradict the simulation studies of the original paper ([5, p.26 6.Simulation studies]). In the original paper they consider models for $n \geq 1000$ and do not directly compare the biases (rather they compare the biases when prior knowledge is given to the method about the bases that represents g best).

5. Future research

In this final section, I will give some possible directions and ideas I would consider if I were to expand upon my thesis.

5.1. CIII with non-linear models

I had the following idea on how to expand CIII to include (some) non-linear functions: Consider the setting of section 4 where f and g can be split in a linear and non-linear part. The linear parts of f and g satisfy the majority/plurality rule as required in section 2. If we already have some idea of the functional form of the non-linear parts of f and g, we may multiply the outcome and association model with a projection to get rid of the non-linear parts (like in the TSCI method from section 4). Now apply CIII to the remaining part (where we are now working with the baseline covariates and potential instrumental variables multiplied by a projection as the new baseline covariates and potential IVs). Here, we might consider different bases for the non-linear parts of f and g and compute $\hat{\text{CI}}^{\text{sear}}$ for each specific basis. In the end, we combine all resulting $\hat{\text{CI}}^{\text{sear}}$'s to end up with a confidence interval for β^*

5.2. Resolving voting issues CIII

In section 2.8, model (S2) for $n = 25000$ (see table 2.8-2.10) showed a low coverage rate for certain cases. A further analysis then showed that higher invalidity levels tended to vote for each other compared to lower invalidity levels, leading to choices for $\hat{V}$ which wouldn't satisfy the majority rule. As mentioned in section 2.8, the most common $\hat{V}'s$ for (M1) were $\{1,2,3,4,6,7,8,9,10\}, \{1,2,3,4,6\}$ and $\{7,8,9\}$ where only $\{1,2,3\}$ were valid. One idea to resolve this, is to first apply sampling to the plurality rule method as mentioned in section 2.6, rather than just computing the estimators from the given data. So, we sample $(\begin{pmatrix}\hat{\Gamma}^{[k]} \\ \hat{\gamma}^{[k]}\end{pmatrix})_{1\leq k\leq K}$ as seen in definition 2.5.7 and then apply the method from section 2.6 to each of the K estimators. Each of these will output their own $\hat{V}_k$. Consider all $(\hat{V}_k)_{1\leq k\leq K}$ together and substitute each $\hat{V}_k$ for the most conservative subset option available for these K sets. For instance, for $K = 2$ if $\hat{V}_1 = \{1,2,3,5,7\}$ and $\hat{V}_2 = \{1,2,3\}$, then set $\hat{V}_1 = \hat{V}_2$. After this selection procedure, let all K proceed to the second stage and combine all resulting K confidence intervals like in the sampling algorithm.

5.3. AR with non-linear data

In the original paper, it was mentioned that there was hope for non-linear extensions to Anchor Regression. By now, there are new papers out that address this topic [8, p.415-422 5.Nonlinear Anchor Regression] and these would be interesting to study.

5.4. Adding simulation studies

The following is a list of extra simulation studies I would consider to expand upon this current work.

1. In the simulation study for CIII (section 2.8), I only considered the low-dimensional setting of p_X and p_Z being smaller than n. The high-dimensional setting is also (for practical purposes) useful to study. For the high-dimensional setting, the OLS-estimation can for example be combined with LASSO. There is a discussion on this in the appendix of the original paper [2, p.2 A.3. High Dimensional IVs and covariates].
2. In my thesis, I only studied the population setting for Anchor Regression and not the finite sample estimator. For this, one could for example, compare the performance of $\hat{b}^\gamma$ on different perturbed data samplings by computing the finite sample mean squared error (MSE) rather than the population MSE as seen in section 3.3.3.
3. For section 4.7 (TSCI simulation studies) one could study more cases and higher n's: in particular higher dimensional covariates and IVs, different bases than just linear combinations of $(X, Z)^T$ and more involved f and g's.

5.5. Comparing CIII, AR and TSCI

In case the assumptions of CIII, AR and TSCI are not met (assumptions like linear models for CIII and AR or that we have a good initial idea for the basis for g for TSCI), we are in a grey-area where it is not clear which one method would outperform the other. This calls for an extensive simulation study that uses a wide-range of models (for instance, a non-linear setting for the treatment and association model where a majority/plurality rule is still met). It is almost guaranteed that such a simulation study has no one answer what to do in case a model does not clearly fall under the setting of either CIII, AR or TSCI. Hence, this would require us to test a lot of different models and study the value of each method regarding that model in detail and expand the study from there. This could be the topic of an entire paper.

A. Lindeberg's central limit theorem

The standard central limit theorem (CLT) requires iid random variables with a finite second moment. Lindeberg's CLT gives conditions for which we have asymptotic normality for independent random variables with a finite second moment (i.e. we can drop the 'identically distributed' condition here).

Theorem A.0.1. *[10, p.359 Theorem 27.2.] Let $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k)$ be a probability space for $X_k : \Omega \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ (i.e. probability space is allowed to change with k), $k \in \mathbb{N}$ be independent random variables. Assume that $\mathbb{E}(X_k) = \mu_k < \infty$, $\mathrm{Var}(X_k) = \sigma_k^2 < \infty$ and define $S_n^2 = \sum_{k=1}^{n} \sigma_k^2$. Then, in case $(X_k)_{k\geq 1}$ satisfies Lindeberg's condition i.e.:*

$$\lim_{n\to\infty} \frac{1}{S_n^2} \sum_{k=1}^{n} \mathbb{E}((X_k - \mu_k)^2 1(|X_k - \mu_k| > \epsilon S_n)) = 0 \quad \forall \epsilon > 0$$

We have:

$$Z_n := \frac{\sum_{k=1}^{n} (X_k - \mu_k)}{S_n} \xrightarrow{d} N(0,1)$$

Remark A.0.2. *Note that:*

$$\frac{1}{S_n^2} \sum_{k=1}^{n} \mathbb{E}((X_k - \mu_k)^2 1(|X_k - \mu_k| > \epsilon S_n)) \geq$$
$$\frac{1}{S_n^2} \max_{k=1,\dots,n} \{\mathbb{E}((X_k - \mu_k)^2 1(|X_k - \mu_k| > \epsilon S_n))\} =$$
$$\frac{1}{S_n^2} \max_{k=1,\dots,n} \{\sigma_k^2 - \mathbb{E}((X_k - \mu_k)^2 1(|X_k - \mu_k| \leq \epsilon S_n))\} \geq$$
$$\frac{1}{S_n^2} \max_{k=1,\dots,n} \{\sigma_k^2 - \epsilon^2 S_n^2\} = \frac{1}{S_n^2} \max_{k=1,\dots,n} \sigma_k^2 - \epsilon^2$$

Hence, in case Lindeberg's condition is satisfied, it holds that:

$$\lim_{n\to\infty} \frac{\max_{k=1,\dots,n} \sigma_k^2}{S_n^2} = 0$$

Example A.0.3. *Consider the following independent random variables:*

$$\mathbb{P}(X_k = k) = \frac{1}{2} = \mathbb{P}(X_k = -k)$$

Now we check Lindeberg's condition:

Observe that as $\sqrt{\sum_{i=1}^{n} i^2} = \mathcal{O}(n^{1.5})$, for n large enough (say $n \geq N_\epsilon$):

$$n \leq \epsilon \sqrt{\sum_{i=1}^{n} i^2}$$

and so for $n \geq N_\epsilon : \mathbb{P}(|X_k| > \epsilon\sqrt{\sum_{i=1}^{n} i^2}) = 0$ *for all* $k = 1, ..., n$. *As* $X_k^2 = k^2$ *for any event, we hence obtain:*

$$\lim_{n\to\infty} \frac{1}{\sum_{i=1}^{n} i^2} \sum_{k=1}^{n} \mathbb{E}(X_k^2 1(|X_k| > \epsilon\sqrt{\sum_{i=1}^{n} i^2})) = \lim_{n\to\infty} \frac{1}{\sum_{i=1}^{n} i^2} \sum_{k=1}^{n} k^2 \mathbb{P}(|X_k| > \epsilon\sqrt{\sum_{i=1}^{n} i^2}) = 0$$

And so by Lindeberg's CLT:

$$\frac{\sum_{k=1}^{n} X_k}{\sum_{k=1}^{n} k^2} \xrightarrow{d} N(0,1)$$

B. Simulation studies code

All the code in this section is implemented in R-studio.

B.1. CIII

```
##Author: Daniël Cohen
##Aim: Simulation study for searching and sampling method for master thesis

##First: Compare search and samp for 4 different settings, gamma_0, tau fixed
library(MASS)
library(RobustIV)
library(xtable)

set.seed(2024)

gamma0 = 0.5
tau = 0.2
nsim = 500
betastar = 1
px = 10
pz = 10
p = px+pz

#Simulate with normal data
simul_norm_W <- function(n){
  W_Cov_norm=matrix(0,nrow=p,ncol=p)
  for(i in 1:p) for(j in 1:p) W_Cov_norm[i,j]=2^(-abs(i-j))
  W_norm=mvrnorm(n,mu=matrix(0,nrow=1,ncol=p),Sigma=W_Cov_norm)
  return(W_norm)} #returning W (obsv in rows)

simul_normI_W <- function(n){
  W_Cov_norm=diag(1,nrow=p,ncol=p)
  W_norm=mvrnorm(n,mu=matrix(0,nrow=1,ncol=p),Sigma=W_Cov_norm)
  return(W_norm)} #returning W (obsv in rows)

simul_norm_e <- function(n){
  e_norm=mvrnorm(n,mu=matrix(0,nrow=1,ncol=2),Sigma=matrix(c(1,0.8,0.8,1)
                                                           ,nrow=2, ncol=2,
                                                          byrow = TRUE))
  return(e_norm)}

#Specify S1-S4
phistar = seq(0.6,1.5,by=0.1)
psistar = seq(1.1,2,by=0.1)
##S1##
S1_DYW <- function(n,tau,gamma0){
```

```
gammastar=gamma0*matrix(1,1,pz)
pistar=c(matrix(0,1,6),tau*gamma0,tau*gamma0,-0.5,-1)
D=matrix(0,n,1)
Y=matrix(0,n,1)
W=simul_norm_W(n)
e=simul_norm_e(n)
for(i in 1:n) D[i]=c(gammastar,psistar) %*% W[i,]+e[i,2]
for(i in 1:n) Y[i]=betastar * D[i] + c(pistar,phistar) %*% W[i,] + e[i,1]
DY=matrix(0,n,(2+p))
DY[,1]=D
DY[,2]=Y
DY[,3:(p+2)]=W
return(DY)} #Returns nx(p+2) matrix: First column=D, Second=Y, rest: W

##S2##
S2_DYW <- function(n,tau,gamma0){
  gammastar=gamma0*matrix(1,1,pz)
  pistar=c(matrix(0,1,3),tau*gamma0,(tau*gamma0)+0.1,tau*gamma0,-0.5,-1,-2/3,
           -0.5)
  D=matrix(0,n,1)
  Y=matrix(0,n,1)
  W=simul_norm_W(n)
  e=simul_norm_e(n)
  for(i in 1:n) D[i]=c(gammastar,psistar) %*% W[i,]+e[i,2]
  for(i in 1:n) Y[i]=betastar * D[i] + c(pistar,phistar) %*% W[i,] + e[i,1]
  DY=matrix(0,n,(2+p))
  DY[,1]=D
  DY[,2]=Y
  DY[,3:(p+2)]=W
  return(DY)}

S2I_DYW <- function(n,tau,gamma0){
  gammastar=gamma0*matrix(1,1,pz)
  pistar=c(matrix(0,1,3),tau*gamma0,(tau*gamma0)+0.1,tau*gamma0,-0.5,-1,-2/3,
           -0.5)
  D=matrix(0,n,1)
  Y=matrix(0,n,1)
  W=simul_normI_W(n)
  e=simul_norm_e(n)
  for(i in 1:n) D[i]=c(gammastar,psistar) %*% W[i,]+e[i,2]
  for(i in 1:n) Y[i]=betastar * D[i] + c(pistar,phistar) %*% W[i,] + e[i,1]
  DY=matrix(0,n,(2+p))
  DY[,1]=D
  DY[,2]=Y
  DY[,3:(p+2)]=W
  return(DY)}



##S3##
```

```
S3_DYW <- function(n,tau,gamma0){
  gammastar=gamma0*matrix(1,1,pz)
  pistar=c(matrix(0,1,4),tau*gamma0,tau*gamma0+0.1,-1/6,-1/3,-1/2,-2/3)
  D=matrix(0,n,1)
  Y=matrix(0,n,1)
  W=simul_norm_W(n)
  e=simul_norm_e(n)
  for(i in 1:n) D[i]=c(gammastar,psistar) %*% W[i,]+e[i,2]
  for(i in 1:n) Y[i]=betastar * D[i] + c(pistar,phistar) %*% W[i,] + e[i,1]
  DY=matrix(0,n,(2+p))
  DY[,1]=D
  DY[,2]=Y
  DY[,3:(p+2)]=W
  return(DY)}

##S4##
S4_DYW <- function(n,tau,gamma0){
  gammastar=gamma0*matrix(1,1,pz)
  pistar=c(matrix(0,1,4),tau*gamma0,tau*gamma0,tau*gamma0,tau*gamma0 + 0.1,
           -1/3,-1/2)
  D=matrix(0,n,1)
  Y=matrix(0,n,1)
  W=simul_norm_W(n)
  e=simul_norm_e(n)
  for(i in 1:n) D[i]=c(gammastar,psistar) %*% W[i,]+e[i,2]
  for(i in 1:n) Y[i]=betastar * D[i] + c(pistar,phistar) %*% W[i,] + e[i,1]
  DY=matrix(0,n,(p+2))
  DY[,1]=D
  DY[,2]=Y
  DY[,3:(p+2)]=W
  return(DY)}
##S5##
S5_DYW <- function(n,tau,gamma0){
  gammastar=c(gamma0,gamma0,1/2,gamma0,matrix(1/2,1,6))
  pistar=c(matrix(0,1,3),tau/2,tau/2,tau,-1/2,-1,-2/3,-1/2)
  D=matrix(0,n,1)
  Y=matrix(0,n,1)
  W=simul_norm_W(n)
  e=simul_norm_e(n)
  for(i in 1:n) D[i]=c(gammastar,psistar) %*% W[i,]+e[i,2]
  for(i in 1:n) Y[i]=betastar * D[i] + c(pistar,phistar) %*% W[i,] + e[i,1]
  DY=matrix(0,n,(p+2))
  DY[,1]=D
  DY[,2]=Y
  DY[,3:(p+2)]=W
  return(DY)}

##S6##
S6_DYW <- function(n,tau,gamma0){
  gammastar=c(gamma0,gamma0,1/2,gamma0,matrix(1/2,1,6))
```

```
  pistar=c(matrix(0,1,6),tau/2,tau/2,-1/2,-1)
  D=matrix(0,n,1)
  Y=matrix(0,n,1)
  W=simul_norm_W(n)
  e=simul_norm_e(n)
  for(i in 1:n) D[i]=c(gammastar,psistar) %*% W[i,]+e[i,2]
  for(i in 1:n) Y[i]=betastar * D[i] + c(pistar,phistar) %*% W[i,] + e[i,1]
  DY=matrix(0,n,(p+2))
  DY[,1]=D
  DY[,2]=Y
  DY[,3:(p+2)]=W
  return(DY)}



## Compare searching with sampling for different n ##
set.seed(2024)
##S1##
n_vec=c(50,100,500,1000,5000)
Cov_sear_S1=matrix(0,5,1)
Len_sear_S1=matrix(0,5,1)
Check_sear_S1=matrix(0,5,1)
#Searching
for(i in 1:5){
  Cov_sear_100=matrix(0,100,1)
  Len_sear_100=matrix(0,100,1)
  Check_sear_100=matrix(0,100,1)
  for(j in 1:100){
    Data=S1_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,method="OLS",Sampling=FALSE, filtering=FALSE
                         ,intercept=FALSE)
    Cov_sear_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
    Len_sear_100[j,]=SS$ci[,2]-SS$ci[,1]}
    Check_sear_100[j,]=SS$check
  Cov_sear_S1[i,]=mean(Cov_sear_100)
  Len_sear_S1[i,]=mean(Len_sear_100)
  Check_sear_S1[i,]=mean(SS$check)}


set.seed(2024)
tau=0.2
gamma0=0.5
##S2##
n_vec=c(50,100,500,1000,5000)
Cov_sear_S2=matrix(0,5,1)
Len_sear_S2=matrix(0,5,1)
Check_sear_S2=matrix(0,5,1)
#Searching
for(i in 1:5){
```

```
  Cov_sear_100=matrix(0,100,1)
  Len_sear_100=matrix(0,100,1)
  Check_sear_100=matrix(0,100,1)
  for(j in 1:100){
    Data=S2_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,method="OLS",Sampling=FALSE, filtering=FALSE
                         ,intercept=FALSE)
    Cov_sear_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
    Len_sear_100[j,]=SS$ci[,2]-SS$ci[,1]
    Check_sear_100[j,]=SS$check}
  Cov_sear_S2[i,]=mean(Cov_sear_100)
  Len_sear_S2[i,]=mean(Len_sear_100)
  Check_sear_S2[i,]=mean(Check_sear_100)}
#S2, Increased n's
set.seed(2024)
n_vec=c(10000,25000,50000,75000,100000)
Cov_sear_S2_n=matrix(0,5,1)
Len_sear_S2_n=matrix(0,5,1)
Check_sear_S2_n=matrix(0,5,1)
#Searching
for(i in 1:5){
  Cov_sear_100=matrix(0,100,1)
  Len_sear_100=matrix(0,100,1)
  Check_sear_100=matrix(0,100,1)
  for(j in 1:100){
    Data=S2_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,method="OLS",Sampling=FALSE, filtering=FALSE
                         ,intercept=FALSE)
    Cov_sear_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
    print(SS$VHat)
    Len_sear_100[j,]=SS$ci[,2]-SS$ci[,1]
    Check_sear_100[j,]=SS$check}
  Cov_sear_S2_n[i,]=mean(Cov_sear_100)
  Len_sear_S2_n[i,]=mean(Len_sear_100)
  Check_sear_S2_n[i,]=mean(Check_sear_100)}

set.seed(2024)
##S3##
n_vec=c(50,100,500,1000,5000)
Cov_sear_S3=matrix(0,5,1)
Len_sear_S3=matrix(0,5,1)
Check_sear_S3=matrix(0,5,1)
#Searching
for(i in 1:5){
  Cov_sear_100=matrix(0,100,1)
  Len_sear_100=matrix(0,100,1)
  Check_sear_100=matrix(0,100,1)
  for(j in 1:100){
```

```
    Data=S3_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,method="OLS",Sampling=FALSE, filtering=FALSE
                         ,intercept=FALSE)
    Cov_sear_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
    Len_sear_100[j,]=SS$ci[,2]-SS$ci[,1]
    Check_sear_100[j,]=SS$check}
  Cov_sear_S3[i,]=mean(Cov_sear_100)
  Len_sear_S3[i,]=mean(Len_sear_100)
  Check_sear_S3[i,]=mean(Check_sear_100)}

set.seed(2024)
##S4##
n_vec=c(50,100,500,1000,5000)
Cov_sear_S4=matrix(0,5,1)
Len_sear_S4=matrix(0,5,1)
Check_sear_S4=matrix(0,5,1)
for(i in 1:5){
  Cov_sear_100=matrix(0,100,1)
  Len_sear_100=matrix(0,100,1)
  Check_sear_100=matrix(0,100,1)
  for(j in 1:100){
    Data=S4_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,method="OLS",Sampling=FALSE,
                         filtering=FALSE,intercept=FALSE)
    Cov_sear_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
    Len_sear_100[j,]=SS$ci[,2]-SS$ci[,1]
    print(SS$VHat)
    Check_sear_100[j,]=SS$check}
  Cov_sear_S4[i,]=mean(Cov_sear_100)
  Len_sear_S4[i,]=mean(Len_sear_100)
  Check_sear_S4[i,]=mean(Check_sear_100)}

set.seed(2024)
##S4##
n_vec=c(10000,25000,50000,75000,100000)
Cov_sear_S4_n=matrix(0,5,1)
Len_sear_S4_n=matrix(0,5,1)
Check_sear_S4_n=matrix(0,5,1)
for(i in 1:5){
  Cov_sear_100=matrix(0,100,1)
  Len_sear_100=matrix(0,100,1)
  Check_sear_100=matrix(0,100,1)
  for(j in 1:100){
    Data=S4_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,method="OLS",Sampling=FALSE, filtering=FALSE
                         ,intercept=FALSE)
    Cov_sear_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
```

```
    Len_sear_100[j,]=SS$ci[,2]-SS$ci[,1]
    print(SS$VHat)
    Check_sear_100[j,]=SS$check}
  Cov_sear_S4_n[i,]=mean(Cov_sear_100)
  Len_sear_S4_n[i,]=mean(Len_sear_100)
  Check_sear_S4_n[i,]=mean(Check_sear_100)}
#Sampling

set.seed(2024)
##S1##
n_vec=c(50,100,500,1000,5000)
Cov_samp_S1=matrix(0,5,1)
Len_samp_S1=matrix(0,5,1)
Check_samp_S1=matrix(0,5,1)
for(i in 1:5){
  Cov_samp_100=matrix(0,100,1)
  Len_samp_100=matrix(0,100,1)
  Check_samp_100=matrix(0,100,1)
  for(j in 1:100){
    Data=S1_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,M=100,method="OLS",filtering=FALSE,intercept=FALSE)
    Cov_samp_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
    Len_samp_100[j,]=SS$ci[,2]-SS$ci[,1]
    Check_samp_100[j,]=SS$check}
  Cov_samp_S1[i,]=mean(Cov_samp_100)
  Len_samp_S1[i,]=mean(Len_samp_100)
  Check_samp_S1[i,]=mean(Check_samp_100)}

set.seed(2024)
tau=0.2
gamma0=0.5
##S2##
n_vec=c(50,100,500,1000,5000)
Cov_samp_S2=matrix(0,5,1)
Len_samp_S2=matrix(0,5,1)
Check_samp_S2=matrix(0,5,1)
for(i in 1:5){
  Cov_samp_100=matrix(0,100,1)
  Len_samp_100=matrix(0,100,1)
  Check_samp_100=matrix(0,100,1)
  for(j in 1:100){
    Data=S2_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,method="OLS",M=100,filtering=FALSE,intercept=FALSE)
    Cov_samp_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
    Len_samp_100[j,]=SS$ci[,2]-SS$ci[,1]
    Check_samp_100[j,]=SS$check}
  Cov_samp_S2[i,]=mean(Cov_samp_100)
  Len_samp_S2[i,]=mean(Len_samp_100)
```

```
  Check_samp_S2[i,]=mean(Check_samp_100)}

set.seed(2024)
##S3##
n_vec=c(50,100,500,1000,5000)
Cov_samp_S3=matrix(0,5,1)
Len_samp_S3=matrix(0,5,1)
Check_samp_S3=matrix(0,5,1)
for(i in 1:5){
  Cov_samp_100=matrix(0,100,1)
  Len_samp_100=matrix(0,100,1)
  Check_samp_100=matrix(0,100,1)
  for(j in 1:100){
    Data=S3_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,method="OLS",M=100,filtering=FALSE,intercept=FALSE)
    Cov_samp_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
    Len_samp_100[j,]=SS$ci[,2]-SS$ci[,1]
    Check_samp_100[j,]=SS$check}
  Cov_samp_S3[i,]=mean(Cov_samp_100)
  Len_samp_S3[i,]=mean(Len_samp_100)
  Check_samp_S3[i,]=mean(Check_samp_100)}

set.seed(2024)
##S4##
n_vec=c(50,100,500,1000,5000)
Cov_samp_S4=matrix(0,5,1)
Len_samp_S4=matrix(0,5,1)
Check_samp_S4=matrix(0,5,1)
for(i in 1:5){
  Cov_samp_100=matrix(0,100,1)
  Len_samp_100=matrix(0,100,1)
  matrix_samp_100=matrix(0,100,1)
  for(j in 1:100){
    Data=S4_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,method="OLS",M=100, filtering=FALSE,intercept=FALSE)
    Cov_samp_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
    Len_samp_100[j,]=SS$ci[,2]-SS$ci[,1]
    Check_samp_100[j,]=SS$check}
  Cov_samp_S4[i,]=mean(Cov_samp_100)
  Len_samp_S4[i,]=mean(Len_samp_100)
  Check_samp_S4[i,]=mean(Check_samp_100)}

set.seed(2024)
##S2 incr n##
n_vec=c(10000,25000,50000,75000,100000)
Cov_samp_S2_n=matrix(0,5,1)
Len_samp_S2_n=matrix(0,5,1)
Check_samp_S2_n=matrix(0,5,1)
```

```
for(i in 1:5){
  Cov_samp_100=matrix(0,100,1)
  Len_samp_100=matrix(0,100,1)
  Check_samp_100=matrix(0,100,1)
  for(j in 1:100){
    Data=S2_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,method="OLS",M=100,filtering=FALSE,intercept=FALSE)
    Cov_samp_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
    Len_samp_100[j,]=SS$ci[,2]-SS$ci[,1]
    print(SS$VHat)
    Check_samp_100[j,]=SS$check}
  Cov_samp_S2_n[i,]=mean(Cov_samp_100)
  Len_samp_S2_n[i,]=mean(Len_samp_100)
  Check_samp_S2_n[i,]=mean(Check_samp_100)}
```

```
set.seed(2024)
##S2, incl filtering##
n_vec=c(50,100,500,1000,5000)
Cov_samp_S2_f=matrix(0,5,1)
Len_samp_S2_f=matrix(0,5,1)
Check_samp_S2_f=matrix(0,5,1)
for(i in 1:5){
  Cov_samp_100=matrix(0,100,1)
  Len_samp_100=matrix(0,100,1)
  Check_samp_100=matrix(0,100,1)
  for(j in 1:100){
    Data=S2_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,method="OLS",M=100,intercept=FALSE)
    Cov_samp_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
    Len_samp_100[j,]=SS$ci[,2]-SS$ci[,1]
    Check_samp_100[j,]=SS$check}
  Cov_samp_S2_f[i,]=mean(Cov_samp_100)
  Len_samp_S2_f[i,]=mean(Len_samp_100)
  Check_samp_S2_f[i,]=mean(Check_samp_100)}

set.seed(2024)
##S2, M=500##
n_vec=c(50,100,500,1000,5000)
Cov_samp_S2_500=matrix(0,5,1)
Len_samp_S2_500=matrix(0,5,1)
Check_samp_S2_500=matrix(0,5,1)
for(i in 1:5){
  Cov_samp_100=matrix(0,100,1)
  Len_samp_100=matrix(0,100,1)
```

```
  Check_samp_100=matrix(0,100,1)
  for(j in 1:100){
    Data=S2_DYW(n_vec[i],tau,gamma0)
    SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                         ,method="OLS",M=500,intercept=FALSE)
    Cov_samp_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
    Len_samp_100[j,]=SS$ci[,2]-SS$ci[,1]
    Check_samp_100[j,]=SS$check}
  Cov_samp_S2_500[i,]=mean(Cov_samp_100)
  Len_samp_S2_500[i,]=mean(Len_samp_100)
  Check_samp_S2_500[i,]=mean(Check_samp_100)}

set.seed(2024)
### Different gamma0, tau for S5, n=5000 searching ##
Cov_samp_S5_gamma0tau=matrix(0,8,3)
Len_samp_S5_gamma0tau=matrix(0,8,3)
Check_samp_S5_gamma0tau=matrix(0,8,3)
gamma0=c(0.05,0.075,0.1)
tau=c(0.025,0.05,0.075,0.1,0.2,0.3,0.4,0.5)
Cov_samp_100=matrix(0,100,1)
Len_samp_100=matrix(0,100,1)
Check_samp_100=matrix(0,100,1)
for(i in 1:3){
  for(k in 1:8){
    Cov_samp_100=matrix(0,100,1)
    Len_samp_100=matrix(0,100,1)
    for(j in 1:100){
      Data=S5_DYW(5000,tau[k],gamma0[i])
      SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                           ,method="OLS",Sampling=FALSE,intercept=FALSE)
      Cov_samp_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
      Len_samp_100[j,]=SS$ci[,2]-SS$ci[,1]
      Check_samp_100[j,]=SS$check}
    Cov_samp_S5_gamma0tau[k,i]=mean(Cov_samp_100)
    Len_samp_S5_gamma0tau[k,i]=mean(Len_samp_100)
    Check_samp_S5_gamma0tau[k,i]=mean(Check_samp_100)}}

set.seed(2024)
### Different gamma0, tau for S4, n=5000 searching ##
Cov_samp_S4_gamma0tau=matrix(0,8,4)
Len_samp_S4_gamma0tau=matrix(0,8,4)
Check_samp_S4_gamma0tau=matrix(0,8,4)
gamma0=c(0.05,0.075,0.1,0.5)
tau=c(0.025,0.05,0.075,0.1,0.2,0.3,0.4,0.5)
Cov_samp_100=matrix(0,100,1)
Len_samp_100=matrix(0,100,1)
Check_samp_100=matrix(0,100,1)
for(i in 1:4){
  for(k in 1:8){
    Cov_samp_100=matrix(0,100,1)
```

```
    Len_samp_100=matrix(0,100,1)
    Check_samp_100=matrix(0,10)
    for(j in 1:100){
      Data=S4_DYW(5000,tau[k],gamma0[i])
      SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                           ,method="OLS",Sampling=FALSE,filtering=FALSE
                           ,intercept=FALSE)
      Cov_samp_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
      Len_samp_100[j,]=SS$ci[,2]-SS$ci[,1]
      Check_samp_100[j,]=SS$check}
    Cov_samp_S4_gamma0tau[k,i]=mean(Cov_samp_100)
    Len_samp_S4_gamma0tau[k,i]=mean(Len_samp_100)
    Check_samp_S4_gamma0tau[k,i]=mean(Check_samp_100)}}


set.seed(2024)
##S4 tau=0.2, gamma0=0.05 case Vhat
tau=0.5
gamma0=0.075
for(j in 1:100){
  Data=S4_DYW(25000,tau,gamma0)
  SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                       ,method="OLS",Sampling=FALSE,intercept=FALSE)
  print("Shat")
  print(SS$SHat)}

set.seed(2024)
##S4 tau=0.2, gamma0=0.075 case Vhat
tau=0.3
gamma0=0.075
for(j in 1:100){
  Data=S4_DYW(5000,tau,gamma0)
  SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                       ,method="OLS",Sampling=FALSE,intercept=FALSE)
  print("Shat")
  print(SS$SHat)}









set.seed(2024)
### Different gamma0, tau for S2, n=5000 searching ##
Cov_samp_S2_gamma0tau=matrix(0,8,4)
Len_samp_S2_gamma0tau=matrix(0,8,4)
Check_samp_S2_gamma0tau=matrix(0,8,4)
gamma0=c(0.05,0.075,0.1,0.5)
```

```
tau=c(0.025,0.05,0.075,0.1,0.2,0.3,0.4,0.5)
Cov_samp_100=matrix(0,100,1)
Len_samp_100=matrix(0,100,1)
Check_samp_100=matrix(0,100,1)
for(i in 1:4){
  for(k in 1:8){
    Cov_samp_100=matrix(0,100,1)
    Len_samp_100=matrix(0,100,1)
    for(j in 1:100){
      Data=S2_DYW(5000,tau[k],gamma0[i])
      SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                           ,method="OLS",Sampling=FALSE,intercept=FALSE)
      Cov_samp_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
      Len_samp_100[j,]=SS$ci[,2]-SS$ci[,1]
      Check_samp_100[j,]=SS$check}
    Cov_samp_S2_gamma0tau[k,i]=mean(Cov_samp_100)
    Len_samp_S2_gamma0tau[k,i]=mean(Len_samp_100)
    Check_samp_S2_gamma0tau[k,i]=mean(Check_samp_100)}}


set.seed(2024)
##S2 tau=0.5, gamma0=0.075
c=0
tau=0.5
gamma0=0.075
for(j in 1:100){
  Data=S2_DYW(5000,tau,gamma0)
  SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                       ,method="OLS",Sampling=FALSE,intercept=FALSE)
  print(SS$check)
  print(SS$VHat)
  print((SS$ci[,1]<=1)&(1<=SS$ci[,2]))}

set.seed(2024)
##S4 tau=0.2 gamma0=0.05 n=5000
tau=0.2
gamma0=0.05
for(j in 1:100){
  Data=S2_DYW(5000,tau,gamma0)
  SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                       ,method="OLS",Sampling=FALSE,intercept=FALSE)
  print("Shat")
  print(SS$SHat)
  print("Vhat")
  print(SS$VHat)}

set.seed(2024)
##S2_I tau=0.4, gamma0=0.075 case Vhat n=25000
tau=0.5
gamma0=0.075
```

```
for(j in 1:100){
  Data=S2I_DYW(25000,tau,gamma0)
  SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                       ,method="OLS",Sampling=FALSE,intercept=FALSE)
  print("Vhat")
  print(SS$VHat)}
```

```
set.seed(2024)
### Different gamma0, tau for S2, n=25 000 searching ##
Cov_samp_S2_gamma0tau=matrix(0,8,4)
Len_samp_S2_gamma0tau=matrix(0,8,4)
Check_samp_S2_gamma0tau=matrix(0,8,4)
gamma0=c(0.05,0.075,0.1,0.5)
tau=c(0.025,0.05,0.075,0.1,0.2,0.3,0.4,0.5)
Cov_samp_100=matrix(0,100,1)
Len_samp_100=matrix(0,100,1)
Check_samp_100=matrix(0,100,1)
for(i in 1:4){
  for(k in 1:8){
    Cov_samp_100=matrix(0,100,1)
    Len_samp_100=matrix(0,100,1)
    for(j in 1:100){
      Data=S2_DYW(25000,tau[k],gamma0[i])
      SS=SearchingSampling(Data[,2],Data[,1],Data[,3:(2+pz)],Data[,(3+pz):(2+p)]
                           ,method="OLS",Sampling=FALSE,intercept=FALSE)
      Cov_samp_100[j,]=(SS$ci[,1]<=1)&(1<=SS$ci[,2])
      Len_samp_100[j,]=SS$ci[,2]-SS$ci[,1]
      Check_samp_100[j,]=SS$check}
    Cov_samp_S2_gamma0tau[k,i]=mean(Cov_samp_100)
    Len_samp_S2_gamma0tau[k,i]=mean(Len_samp_100)
    Check_samp_S2_gamma0tau[k,i]=mean(Check_samp_100)}}
```

B.2. TSCI

```
##Author: Daniël Cohen
##Aim: Simulation study TSCI implementation

library(MASS)
library(RobustIV)
library(xtable)
source("C:/Users/Daniël Cohen/Desktop/TU Delft/Thesis/TSCI_code.R")
#see https://github.com/zijguo/TSCI/blob/main/R/Source-Random-Forest.R for
#source
```

```
library(Rcpp)

set.seed(2024)
beta=1
p_X=5
p_Z=1
p=p_X+p_Z

#Simulate with normal data Xstar
simul_norm_Xstar <- function(n){
  Xstar_Cov_norm=matrix(0,nrow=p,ncol=p)
  for(i in 1:p) for(j in 1:p) Xstar_Cov_norm[i,j]=2^(-abs(i-j))
  Xstar_norm=mvrnorm(n,mu=matrix(0,nrow=1,ncol=p),Sigma=Xstar_Cov_norm)
  return(Xstar_norm)} #returning W (obsv in rows)

#Simulate X and Z
simul_norm_XZ<- function(n){
  XZ=matrix(0,n,p)
  X_star=simul_norm_Xstar(n)
  for(i in 1:n){for(j in 1:p-1){XZ[i,j]=pnorm(X_star[i,j])}
  }
  for(i in 1:n){XZ[i,p]=4*(pnorm(X_star[i,p])-0.5)}
  return(XZ)
} #X=XZ[,1:p-1],Z=XZ[,p]

#Simulate errors
simul_norm_e <- function(n){
  e_norm=mvrnorm(n,mu=matrix(0,nrow=1,ncol=2),Sigma=matrix(c(1,0.5,0.5,1)
                                                           , nrow=2,
                                                           ncol=2,
                                                           byrow = TRUE))
  return(e_norm)}


M_1 <- function(n,Z,X,e,b){
  g=matrix(0,n,1)
  f=matrix(0,n,1)
  DY=matrix(0,n,2)
  for(i in 1:n){
    g[i]=1+sum(X[i,])
    f[i]=-0.5+0.2*sum(X[i,])+b*(cos(2*pi*Z[i])+Z[i]*sum(X[i,]))
    DY[i,1]=f[i]+e[i,1] #D expr
    DY[i,2]=beta*DY[i,1]+g[i]+e[i,2] #Y expr
  }
  return(DY) #D=DY[,1], Y=DY[,2]
}


M_2 <- function(n,Z,X,e,a,b){
  g=matrix(0,n,1)
```

```
  f=matrix(0,n,1)
  DY=matrix(0,n,2)
  for(i in 1:n){
    S=0
    for(q in 0:10){S=+Z[i]^q}
    g[i]=S+a*Z[i]^{11}+sum(X[i,])
    f[i]=0.5*Z[i]^{5} + b*(cos(2*pi*Z[i])+Z[i]*sum(X[i,]))
    DY[i,1]=f[i]+e[i,1] #D expr
    DY[i,2]=beta*DY[i,1]+g[i]+e[i,2] #Y expr
  }
  return(DY) #D=DY[,1], Y=DY[,2]
}

M_3 <- function(n,Z,X,e,a,b){
  g=matrix(0,n,1)
  f=matrix(0,n,1)
  DY=matrix(0,n,2)
  for(i in 1:n){
    g[i]=1+a*(Z[i]+Z[i]^2)+sum(X[i,])
    f[i]=-0.5 + 0.2*sum(X[i,]) + b*(cos(2*pi*Z[i])+Z[i]*sum(X[i,]))
    DY[i,1]=f[i]+e[i,1] #D expr
    DY[i,2]=beta*DY[i,1]+g[i]+e[i,2] #Y expr
  }
  return(DY) #D=DY[,1], Y=DY[,2]
}

M_4 <- function(n,Z,X,e,a,b){
  g=matrix(0,n,1)
  f=matrix(0,n,1)
  DY=matrix(0,n,2)
  for(i in 1:n){
    g[i]=Z[i]+0.5*sum(X[i,])
    f[i]=cos(2*pi*Z[i])+a*sin(2*pi*Z[i])*Z[i]*sum(X[i,])*exp(Z[i])+
    b*(Z[i]+0.5*Z[i]^2)
    DY[i,1]=f[i]+e[i,1] #D expr
    DY[i,2]=beta*DY[i,1]+g[i]+e[i,2] #Y expr
  }
  return(DY) #D=DY[,1], Y=DY[,2]
}

n=c(100,300,500)
b=c(0,0.5,1,5)
betainit_res=matrix(0,4,3)
betatilde_res=matrix(0,4,3)
Cov_res=matrix(0,4,3)
CIlen_res=matrix(0,4,3)
qcomp_res=matrix(0,4,3)
qrob_res=matrix(0,4,3)
#M_1 results
set.seed(2024)
```

```
for(i in 1:4){
  for(j in 1:3){
    M_data_biasinit=matrix(0,100,1)
    M_data_biastilde=matrix(0,100,1)
    M_data_Cov=matrix(0,100,1)
    M_data_qcomp=matrix(0,100,1)
    M_data_qrob=matrix(0,100,1)
    M_data_CIlen=matrix(0,100,1)
    for(M in 1:100){
      XZ=simul_norm_XZ(n[j])
      X=XZ[,1:p-1]
      Z=XZ[,p]
      e=simul_norm_e(n[j])
      DY=M_1(n[j],Z,X,e,b[i])
      D=DY[,1]
      Y=DY[,2]
      vio.space1 <- matrix(NA,n[j],0)
      for (q in 1:2) {
        vio.space1 <- cbind(Z^q,vio.space1)}
      TSCI=TSCI.RF(Y,D,Z,X,vio.space=vio.space1,num.trees=50)
      betainit=TSCI$Coef.robust[1] #comp
      betatilde=TSCI$Coef.robust[2] #comp
      CI=TSCI$CI.robust
      qcomp=TSCI$q.comp
      qrob=TSCI$q.robust

      M_data_biasinit[M]=abs(betainit - beta)
      M_data_biastilde[M]=abs(betatilde -beta)
      M_data_Cov[M]=(CI[1,2]<=beta)&(beta<=CI[2,2]) #tilde,comp
      M_data_CIlen[M]=CI[2,2]-CI[1,2]
      M_data_qcomp[M]=qcomp
      M_data_qrob[M]=qrob


    }
   betainit_res[i,j]=mean(M_data_biasinit)
   betatilde_res[i,j]=mean(M_data_biastilde)
   Cov_res[i,j]=mean(M_data_Cov)
   CIlen_res[i,j]=mean(M_data_CIlen)
   qcomp_res[i,j]=round(mean(M_data_qcomp))
   qrob_res[i,j]=round(mean(M_data_qrob))

  }
}

#M_2
a=c(0.5,1)
b=c(0.5,1)
betainit_resM2=matrix(0,3,2)
betatilde_resM2=matrix(0,3,2)
```

```
Cov_resM2=matrix(0,3,2)
CIlen_resM2=matrix(0,3,2)
qcomp_resM2=matrix(0,3,2)
qrob_resM2=matrix(0,3,2)
set.seed(2024)
for(i in 1:3){
  for(j in 1:2){
    M_data_biasinit=matrix(0,100,1)
    M_data_biastilde=matrix(0,100,1)
    M_data_Cov=matrix(0,100,1)
    M_data_qcomp=matrix(0,100,1)
    M_data_qrob=matrix(0,100,1)
    M_data_CIlen=matrix(0,100,1)
    for(M in 1:100){
      print(M)
      XZ=simul_norm_XZ(300)
      X=XZ[,1:p-1]
      Z=XZ[,p]
      e=simul_norm_e(300)
      DY=M_2(300,Z,X,e,a[j],b[i])
      D=DY[,1]
      Y=DY[,2]
      vio.space1 <- matrix(NA,300,0)
      for (q in 1:12) {
        vio.space1 <- cbind(Z^q,vio.space1)}
      TSCI=TSCI.RF(Y,D,Z,X,vio.space=vio.space1,num.trees = 50)
      betainit=TSCI$Coef.robust[1] #comp
      betatilde=TSCI$Coef.robust[2] #comp
      CI=TSCI$CI.robust
      qcomp=TSCI$q.comp
      qrob=TSCI$q.robust

      M_data_biasinit[M]=abs(betainit - beta)
      M_data_biastilde[M]=abs(betatilde -beta)
      M_data_Cov[M]=(CI[1,2]<=beta)&(beta<=CI[2,2]) #tilde,comp
      M_data_CIlen[M]=CI[2,2]-CI[1,2]
      M_data_qcomp[M]=qcomp
      M_data_qrob[M]=qrob


    }
    betainit_resM2[i,j]=mean(M_data_biasinit)
    betatilde_resM2[i,j]=mean(M_data_biastilde)
    Cov_resM2[i,j]=mean(M_data_Cov)
    CIlen_resM2[i,j]=mean(M_data_CIlen)
    qcomp_resM2[i,j]=round(mean(M_data_qcomp))
    qrob_resM2[i,j]=round(mean(M_data_qrob))

  }
}
```

```
#M_3
a=c(1/8,1/4,1/2)
b=c(0,1,5)
betainit_resIV1=matrix(0,3,3)
betatilde_resIV1=matrix(0,3,3)
Cov_resIV1=matrix(0,3,3)
CIlen_resIV1=matrix(0,3,3)
qcomp_resIV1=matrix(0,3,3)
qrob_resIV1=matrix(0,3,3)
IV_resIV1=matrix(0,3,3)
set.seed(2024)
for(i in 1:3){
  for(j in 1:3){
    M_data_biasinit=matrix(0,100,1)
    M_data_biastilde=matrix(0,100,1)
    M_data_Cov=matrix(0,100,1)
    M_data_qcomp=matrix(0,100,1)
    M_data_qrob=matrix(0,100,1)
    M_data_CIlen=matrix(0,100,1)
    M_data_IV=matrix(0,100,1)
    for(M in 1:100){
      XZ=simul_norm_XZ(500)
      X=XZ[,1:p-1]
      Z=XZ[,p]
      e=simul_norm_e(500)
      DY=M_3(500,Z,X,e,a[j],b[i])
      D=DY[,1]
      Y=DY[,2]
      vio.space1 <- matrix(NA,500,0)
      TSCI=TSCI.RF(Y,D,Z,X,num.trees = 50)
      betainit=TSCI$Coef.robust[1] #comp
      betatilde=TSCI$Coef.robust[2] #comp
      CI=TSCI$CI.robust
      qcomp=TSCI$q.comp
      qrob=TSCI$q.robust

      M_data_biasinit[M]=abs(betainit - beta)
      M_data_biastilde[M]=abs(betatilde -beta)
      M_data_Cov[M]=(CI[1,2]<=beta)&(beta<=CI[2,2]) #tilde,comp
      M_data_CIlen[M]=CI[2,2]-CI[1,2]
      M_data_qcomp[M]=qcomp
      M_data_qrob[M]=qrob
      M_data_IV[M]=TSCI$invalidity


    }
    betainit_resIV1[i,j]=mean(M_data_biasinit)
    betatilde_resIV1[i,j]=mean(M_data_biastilde)
    Cov_resIV1[i,j]=mean(M_data_Cov)
```

```
    CIlen_resIV1[i,j]=mean(M_data_CIlen)
    qcomp_resIV1[i,j]=round(mean(M_data_qcomp))
    qrob_resIV1[i,j]=round(mean(M_data_qrob))
    IV_resIV1[i,j]=mean(M_data_IV)

  }
}

#M_4
a=c(0,1,5)
b=c(0,1,5)
betainit_resf=matrix(0,3,3)
betatilde_resf=matrix(0,3,3)
Cov_resf=matrix(0,3,3)
CIlen_resf=matrix(0,3,3)
qcomp_resf=matrix(0,3,3)
qrob_resf=matrix(0,3,3)
IV_resf=matrix(0,3,3)
set.seed(2024)
for(i in 1:3){
  for(j in 1:3){
    M_data_biasinit=matrix(0,100,1)
    M_data_biastilde=matrix(0,100,1)
    M_data_Cov=matrix(0,100,1)
    M_data_qcomp=matrix(0,100,1)
    M_data_qrob=matrix(0,100,1)
    M_data_CIlen=matrix(0,100,1)
    M_data_IV=matrix(0,100,1)
    for(M in 1:100){
      XZ=simul_norm_XZ(500)
      X=XZ[,1:p-1]
      Z=XZ[,p]
      e=simul_norm_e(500)
      DY=M_4(500,Z,X,e,a[j],b[i])
      D=DY[,1]
      Y=DY[,2]
      TSCI=TSCI.RF(Y,D,Z,X,num.trees = 50)
      betainit=TSCI$Coef.robust[1] #comp
      betatilde=TSCI$Coef.robust[2] #comp
      CI=TSCI$CI.robust
      qcomp=TSCI$q.comp
      qrob=TSCI$q.robust

      M_data_biasinit[M]=abs(betainit - beta)
      M_data_biastilde[M]=abs(betatilde -beta)
      M_data_Cov[M]=(CI[1,2]<=beta)&(beta<=CI[2,2]) #tilde,comp
      M_data_CIlen[M]=CI[2,2]-CI[1,2]
      M_data_qcomp[M]=qcomp
      M_data_qrob[M]=qrob
      M_data_IV[M]=TSCI$invalidity
```

```
    }
    betainit_resf[i,j]=mean(M_data_biasinit)
    betatilde_resf[i,j]=mean(M_data_biastilde)
    Cov_resf[i,j]=mean(M_data_Cov)
    CIlen_resf[i,j]=mean(M_data_CIlen)
    qcomp_resf[i,j]=round(mean(M_data_qcomp))
    qrob_resf[i,j]=round(mean(M_data_qrob))
    IV_resf[i,j]=mean(M_data_IV)

  }
}
```

Bibliography

[1] (Jan 2016) Pancreatic Cancer UK https://www.pancreaticcancer.org.uk/news-and-blogs/does-dark-chocolate-protect-against-pancreatic-cancer/

[2] Guo, Z. (2023) Causal Inference with Invalid Instruments: Post-selection Problems and A Solution Using Searching and Sampling. arXiv:2104.06911

[3] Vershynin, R. High-Dimensional probability: An Introduction with Applications in Data Science. https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf

[4] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok (Eds.), Compressed Sensing: Theory and Applications, pp. 210–268. Cambridge University Press.

[5] Bühlmann, P., Guo Z. (2022) Two Stage Curvature identification with Machine Learning: Causal Inference with possibly Invalid Instrumental Variables, arXiv:2203.12808v2

[6] Hastie, T., Tibshirani, R., Friedman, J. (Corrected 12th printing- 13 Jan, 2017) The Element of Statistical Learning, Second Edition.

[7] Rudelson, M., Vershynin, R. (2013) Hanson-Wright inequality and sub-Gaussian concentration, Electronic Communications in Probability

[8] Bühlmann, P. Statistical Science 2020.Vol.35 No.3 404-426, Invariance,Causality and Robustness.

[9] Bühlmann, P. et al. (2020) Anchor regression: heterogeneous data meet causality

[10] Billingsley, P. Probability and Measure, Third edition

[11] van der Vaart, A. (17-6-2023) Causality and Graphical Models (lecture notes)