Analysis of Testing-based Model Selection with Graphical Considerations

WORKING PAPER

Jorge de la Cal Medina

Supervised by Damian Kozbur

A thesis presented for the degree of
Master of Arts in Economics

Department of Economics
University of Zurich
Switzerland

1 Introduction

Estimating the causal effect or treatment effect of a certain policy on an outcome of interest, such as an economic indicator, is an important challenge that policy advisors face. If data from an experimental setting is available, where treatment assignment is randomized, the treatment effect can be estimated by taking the difference in outcomes between the treated and non-treated individuals. However, it is often not possible to conduct such an experiment to evaluate economic policies for obvious reasons. Data from observational settings, on the other hand, are readily available but also prone to endogeneity in the treatment assignment. That aggravates the difficulty of identifying a causal effect considerably. Taking differences between outcomes would in this case lead to confounding bias, or omitted variable bias, which can be severe and completely distort the actual effect. An extreme example of this is described in Simpson's paradox, where controlling for a confounding variable completely reverses the effect (Simpson, 1951).

Let treatment D be a putative cause of outcome Y and consider a confounder X, which is a cause of both, D and Y. In order to adjust for confounding bias, we must orthogonalize D and Y with respect to X. Formally, D and Y must be projected onto the orthogonal complement of the column space of X. According to the Frisch-Waughn-Lovell theorem, this can be done by ordinary least squares (OLS) while controlling for X. The treatment effect of D on Y, can then be estimated by regressing Y on D and X (Lovell, 2008). The resulting estimator is unbiased, given that all confounders have been effectively controlled for. When evaluating treatment effect on observational data, it is thus necessary to select a set of controls that adjusts for any lurking confounder.

In order to do that there has to be a set of covariates available in the first place, which is *causally sufficient* for the treatment effect in question. That is, the set of observed covariates should include all common causes of D and Y (Spirtes *et al.*, 2000). This assumption about the available data is a prerequisite for successful covariate adjustment, and will I hereafter be taken as given.

However, even if this assumption is granted, it is nonetheless paramount to select the control set in a principled manner and forego the backwards credo of controlling for any observable covariate. Otherwise, other sources of bias than confounder bias might be introduced instead, such as collider bias or M-bias (Pearl, 2009b). Moreover, controlling for all covariates regardless would at best produce an inefficient estimate, since too much of the variance in the explanatory variables would be reduced unnecessarily. This is only aggravated in the case of high-dimensional data. Conducting OLS while controlling for any available covariate might be entirely unfeasible if the number of covariates is larger than n.

High-dimensional data becomes more accessible as the number of observed characteristics grows with increasing data availability and computing power. It is therefore important to formulate new methods that address the inherent issue. Expert knowledge may suggest a set of variables, but this clearly lacks rigour. Economic studies will typically do a sensitivity

analysis reporting results for different control sets. This, however, is insofar not helpful as it does not explain how the differences in using different control sets come about (Belloni *et al.*, 2014b).

To address this problem, I develop a method that uses estimated causal structures to conduct root-n consistent estimation, where n is the number of observations, in high-dimensional models which can be approximated by a low-dimensional, *imperfect* model. Specifically, I use a causal structure-learning algorithm to conduct covariate-selection. The causal effect can then be estimated using the selected model applying the post-double selection framework. In the following, I will clarify these concepts and lay out the structure of thesis in more detail.

I will assume that the underlying data generating process, or model, can be described by structural equation models - a common assumption in policy evaluation (Heckman & Vytlacil, 2005). This allows me to describe the model as a directed acyclical graphs (DAG), which besides depicting causal relations in an intuitive manner, enables to draw from the rich literature on graphical causal models and algorithmic methods for identifying causal links (Pearl, 2012). A part of this literature is concerned with estimating causal effects from observational data through covariate adjustment (Shpitser *et al.*, 2012). This identification strategy relies on finding an *adjustment set* that is able to control for any confounders. An adjustment set is said to be *valid* if it blocks every *back-door path*, or in other words, controls for every confounder (see Pearl (2009a) for more details). There are several graphical criteria for validity of adjustment sets (Perkovi´c *et al.*, 2015). It has been shown that, given the Causal Markov property, the set of *immediate* causes in the model of Y, the so-called *parents*, is a valid adjustment set to estimate the treatment effect of D on Y (see Pearl (2009a), Spirtes *et al.* (2000)). Although the parental set might not necessarily be the most efficient adjustment set (Witte *et al.*, 2020), it requires less information about the causal structure than the entire DAG, as Vanderweele & Shpitser (2011) find. This is a desirable feature of the method I here develop.

The *PC* algorithm (Spirtes *et al.*, 2000) is a method for causal structure learning, which aims at learning the DAG from observational data. For our purpose, however, a slightly simplified version of *PC* is sufficient. Since we only search for the parental set, and assume our data does not contain any descendants of Y, we only need to check conditional independencies between each covariate and Y. An omniscient oracle, that knows about any conditional independencies between the variables, would produce exactly the set of parents of Y.

Alas, such an oracle is in practice unattainable and conditional independencies have to be checked by statistical testing instead. This brings an additional difficulty with itself. Belloni *et al.* (2014b) show that only checking for conditional independencies with respect to Y can lead to bias. I refer to this method of model selection as *post-single selection*. The reason this is problematic, is that the conditional independence test might fail to detect actual confounders. The result is an estimator that is not root-n consistent and non-regular. I motivate this issue with a simulation below, see section 3.

To prevent this mistake, the authors propose the *post-double selection* framework, which applies two selection steps instead of one. First, it selects a set $\mathbf{S}_Y$ of covariates that are predictive of outcome Y. Second, it selects a set $\mathbf{S}_D$ that is predictive of D. The exact interpretation of predictiveness depends on the particular model-selection method applied. The treatment effect is then estimated by OLS, regressing Y on D and all covariates indexed by the union of $\mathbf{S}_Y$ and $\mathbf{S}_D$. The authors show that the post-double selection estimator is root-n consistent. Post-double selection allows to draw sound conclusion on causal effects from observational with high confounding. Not only is it apt for high-dimensional settings, it also encourages data-mining, without necessarily losing interpretability of the model. It adds rigour to covariate selection since it operates on minimal ex-ante assumptions. It is thus a useful addition to the practitioner's toolkit for applied research, particularly for policy evaluation. Belloni *et al.* (2014b) use l_1-penalized methods, e.g. Lasso, for model selection, but highlight that the theoretical results are applicable to other methods as well.

My main contribution consists in developing a novel testing-based model-selection method that makes use of graphical structure learning. I call it *AdJ* - an abbreviation for adjustment[1]. Further, I analyze AdJ in its application for post-double selection in high-dimensional problems. First, I show the causal sufficiency of the model, selected using the population version of AdJ, which assumes an unattainable oracle to be known. Second, I provide partial results to show that AdJ allows for imperfect model selection, enabling it to handle high-dimensional data. These results, however, depend on a conjecture, that has yet to be proven formally. It remains a research component of the ongoing project.

I base my findings mainly on previous literature on policy evaluation in economics and on causal graph theory, as well as on machine learning. Economists have developed a range of methods to draw causal inference from observational data, see Angrist & Krueger (1999) or, for a more contemporary perspective, Athey & Imbens (2017). There has recently been an increased interest by economists to also apply machine learning techniques. Double machine learning is a good example of the fruitfulness this new research development can have, see Chernozhukov *et al.* (2018). It has been developed around the post-double selection framework, described above, and tackles regularization bias, confounder bias stemming from failure to detect confounders in the process of dimensionality-decrease, and overfitting bias, which is the case when the model captures noise terms rather than structural parameters. In this thesis, I focus exclusively on the former kind of bias. However, it would be interesting to analyze the problem of overfitting using AdJ as well.

The graphical approach to causal analysis, for whose development Pearl (2009a) and Spirtes *et al.* (2000) are co-responsible, has gained traction in statistics and computer science, see for instance Peters *et al.* (2017), Peters *et al.* (2014) or Hoyer *et al.* (2008). However, it has not been fully embraced by economists yet Imbens (2020). I believe nonetheless that there are important lessons to be learned from this literature, which could lead to new advancements

[1] Inspired by the PC algorithm's name which stands for its author's first names, Peter and Clark, I write AdJ with a upper case letter J, which stands for my first name, Jorge.

in the way causality is treated by economists. In that sense, this thesis represents an effort to bring both fields closer together. I am aware that although causal graph theorists and economists share common problems, they often use different language to address them. Therefore, I took care to express the ideas and concepts presented in an inclusive manner, such that hopefully members of either research community can easily follow them.

Furthermore, I contribute to the mentioned literatures in the following way. First, I contribute to the literature on causal identification, by providing a method to prevent confounder bias in policy evaluation, focusing on high-dimensional settings. Specifically, I complement the literature on post-double selection by another feasible model selection method. Second, I contribute to the literature on graphical causal models by finding an application of PC, or rather a version of it, to an applied problem in policy evaluation.

The focus of this thesis concerns lies in the development of AdJ and its basic properties that make it fit for usage. Further research could analyze its statistical properties in more detail. AdJ allows for different decision rules to be used in practice. It would be interesting to investigate how different tests affect the result.

This thesis is structured as follows. First I will give an intuition of the importance of post-double selection by means of a simulation. Then, I will define the AdJ Algorithm and analyze its theoretical properties. I then asses the performance of AdJ on simulated data. finally, I use actual data from Acemoglu *et al.* (2001) to exemplify the use of AdJ.

2 Imperfect Model Selection with Causal Considerations

In this section I will investigate the possibility of using causal discovery methods to conduct imperfect model selection. Causal structure learning is usually aimed at finding the underlying DAG from observational data. The AdJ algorithm is based on PC, but is aimed at finding only the adjacencies with respect to one target variable, rather than the entire graph. These adjacent variables will be the parental set of the target variable.

The primary framework of this thesis is a high-dimensional, sparse and linear model. The sparsity assumption requires that the data generating process can be approximated by a function over a small number of covariates, realtiv eto th enumber of observations. AdJ searches for the adjustment set of a given variable, consisting of the immediate causes of that variable. It starts by setting the adjustment set to the entire set of covariates $\mathbf{X}$ and proceeds by iteratively conducting conditional independence test between the variable of interest and any of the covariates. The conditioning set is initially set to the empty set and it increases in size by one with each iteration. If a test is positive, the covariate shown to be (conditionally) independent covariate is deleted from the set of covariates.

I will first define a version of AdJ, based on PC as it is described in Kalisch & Buhlmann (2007), which is able to find the parental set in the theoretical scenario where an oracle version of the conditional independence test is known. Then, I define a finite sample version of the algorithm that uses a feasible conditional independence and show that it is suited for

imperfect model selection. That is, it is able to find a approximate set of relevant covariate that suffice to predict the target variable, so that that the approximation error is sufficiently small. The framework I use is largely based on the described in Belloni *et al.* (2014b) and Kozbur (2020).

2.1 Preliminaries

It is useful to introduce some preliminary concepts and notation. I will make use of the following causal concepts. For definitions see Spirtes *et al.* (2000).

Consider a set of random variables $\mathbf{X} = \{X_1, ..., X_p\}$ with index set $\mathbf{V} = \{1, ..., p\}$. Let $\mathrm{P}(\mathbf{X})$ be a distribution over them. Let $\mathscr{G}$ be a graph consisting of nodes $\mathbf{V}$ and edges $\mathscr{E} \subseteq \mathbf{V}^2$ with $(v,v) \notin \mathscr{E}$. If an edge $(i,j) \in \mathscr{E}$ and $(j,i) \notin \mathscr{E}$ we say the edge is directed and write $i \to j$. For some subset $\mathbf{M} \subseteq \mathbf{V}$ graph $\mathscr{G}_{\mathbf{M}} = (\mathbf{V}, \mathscr{E}_{\mathbf{M}})$ is a subgraph of $\mathscr{G}$ with vertices $\mathbf{V}$ and edges $\mathscr{E}_{\mathbf{M}} \subseteq \mathbf{M}^2$. If $\mathscr{E}_{\mathbf{M}} \subseteq \mathscr{E}$we call it a proper subgraph.

A node i is called a parent of j if $(i,j) \in \mathscr{E}$. The set of parents of j is denoted by $\mathbf{PA}_j$.Two nodes (i,j) are ajacent if either $(i,j) \in \mathscr{E}$ or $(j,i) \in \mathscr{E}$ A path in $\mathscr{G}$ is a sequence of two or more nodes $i_1, ..., i_n$ such that there is an edge between $i_k, i_{k+1}, \forall k = 1, ..., n-1$. If $i_k \to i_{k+1}$ for all k then we call the path between i_1 and i_n directed and say i_n is a descendant of i_1. We denote the set of descendants of i by $\mathbf{DE}_i$. $\mathscr{G}$ is called a partially directed acyclical graph (PDAG) if it contains no directed cycles. $\mathscr{G}$ is called a DAG if it is a PDAG and all edges are directed. In $\mathscr{G}$ a path between i_1 and i_n is blocked by $M \subseteq V \setminus \{i_1, i_n\}$ whenever there is i_k such that one of the following hold: 1. $i_k \in \mathbf{M}$ and $i_{k-1} \to i_k \to i_{k+1}$ or $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ or $i_{k-1} \leftarrow i_k \to i_{k+1}$. Or 2. $i_{k-1} \to i_k \leftarrow i_{k+1}$ and neither i_k nor any of its descendants is in $\mathbf{M}$. If every path between disjoint sets $\mathbf{A}, \mathbf{B}$ is blocked by disjoint set $\mathbf{M}$, i.e. d-separated by $\mathbf{M}$, we write $\mathbf{A}$ *d-sep* $\mathbf{B}|\mathbf{M}$. $\mathrm{P}(\mathbf{X})$ is said to be Markov and Faithful wrt $\mathscr{G}$ if

$$\mathbf{A} \textit{ d-sep } \mathbf{B}|\mathbf{M} \Leftrightarrow A \perp\!\!\!\perp \mathbf{B}|\mathbf{M}. \tag{1}$$

Further, I will make use of the following notation. Let p denote the dimension of the parameter space and let $\mathbf{W} \subseteq \{1, ..., p\}$. Define submatrices $X_{\mathbf{W}} := [X_j : j \in X_{\mathbf{W}}]$ and $X_{\mathbf{W}^c} := [X_j : j \notin \mathbf{W}]$. Denote the matrix of variables spanning the entire parameter space $X_{\{1,...,p\}}$ as $\mathbf{X}$. Denote the projection onto the subspace that is orthogonal to $span(X_{\mathbf{W}})$ as $M_{\mathbf{W}} = I - X_{\mathbf{W}}(X'_{\mathbf{W}} X_{\mathbf{W}}^{-1}) X'_{\mathbf{W}}$.

2.2 The AdJ-algorithm: Population Version

To investigate the theoretical properties of AdJ, I first define the population version. Afterwards I turn to its finite-sample properties. It is important to stress that AdJ is designed to work under specific conditions on the data. Specifically, the following conditions need to be satisfied.

Let P be a distribution over variable tuple $(Z, \mathbf{X})$ with target variable Z and covariates $\mathbf{X} = \{X_1, ..., X_p\}$ with index set $\mathbf{V} = \{1, ..., p\}$. The goal is to find a subset of covariates that constitute the parental set of the target variable.

Condition Markov and Faithfulness. There exists a DAG $\mathscr{G}$ with vertex set $\{Z\} \cup \mathbf{V}$ that is Markov and Faithful wrt P, see (1).

Condition policy evaluation Setting (PES). We consider a setting in which $\mathbf{X}$ are potential causes of Z which itself is never a cause.

$$X_j \notin \mathbf{DE}_Z, \forall j \in \{1, ..., p\}. \tag{2}$$

PES is meant to describe a standard policy evaluation setting. It is possible to define a method which allows to relax this condition. This could be an extension of the AdJ as I describe it here.

Condition Local Causal Sufficiency (LCS). There is no hidden common cause in $\mathscr{G}$ that is causing Z and $X_j, \forall j \in \{1, ..., p\}$.

Note that LCS is a weaker form of Causal Sufficiency, see Spirtes (2010). This condition depends on the covariate set being sufficient, which although hard to justify, is necessary for covariate adjustment.

The population version of Algorithm AdJ_{pop} is a modification of the PC algorithm, see 2.2.1. in Kalisch and Bühlmann (2007). It finds any adjacent nodes to Z in $\mathbf{X}$.

Algorithm 1: AdJ_{pop}

input : Data $(Z, \mathbf{X})$.
output : Set of indices $\mathbf{S}_Z$.
$\mathbf{S}_Z \leftarrow \mathbf{V}$;
l = -1;
repeat
  l = l+1 **repeat**
    foreach $j \in \mathbf{S}_Z$ **do**
      Choose $\mathbf{K} \subseteq \mathbf{S}_Z$ with $|\mathbf{K}| = l$;
      if $Z \perp\!\!\!\perp X_j | \mathbf{K})$ **then**
        $\mathbf{S}_Z \leftarrow \mathbf{S}_Z \setminus \{j\}$
      end
    end
  until *until all j in* $\mathbf{S}_Z$ *have been tested*;
until $|\mathbf{S}_Z| < l$;

In the worst case, when there are no independencies in the data, no variable is excluded and AdJ has $(p+1)$ steps. In step k there are $\binom{p+1}{k}$ independence done. The complexity in the worst case is therefore given by $\sum_{k=1}^{p+1} \binom{p+1}{k}$ which is equal to 2^{p+1}. This corresponds to exponential complexity. We will see in the next section, however, that it is alleviated given sparsity conditions.

Theorem 1. If AdJ$_{pop}$ is run on $(Z, \mathbf{X})$, it returns a set $\mathbf{S}_Z$ such that Z is independent of $X_{\mathbf{S}^c}$ given $X_{\mathbf{S}}$:

$$Z \perp\!\!\!\perp X_{\mathbf{S}_Z^c} | X_{\mathbf{S}_Z} \tag{3}$$

$$\nexists j \in \mathbf{S}_Z : Z \perp\!\!\!\perp X_j | X_{\mathbf{S}_Z \setminus j}. \tag{4}$$

Theorem 1 states that the result of AdJ$_{pop}$ gives a set of indices, conditional on which the target variable Z is independent of the remaining covariates $\mathbf{X}$. Furthermore, it states that this set is minimal. Therefore, the resulting set of indices given by AdJ corresponds to the parental set $\mathbf{PA}_Z$. This is useful, since causal graph theory tells us that the parental set of a variable is a valid adjustment set (Vanderweele & Shpitser, 2011; Van Der Zander *et al.*, 2019). See the proof for theorem 1 in the appendix.

2.3 The AdJ-algorithm: Sample Version

We now turn to the sample version of AdJ, which can be applied to a finite sample of observational data. The observed data are given by $\mathscr{D}_n = \{(Z_i, \mathbf{X}_i)\}_{i=1}^n$, with $Z \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n \times p}$, and are generated by a distribution P. We will allow for high-dimensional settings - for simplicity assume that $n = O(\log p)$. The model satisfies the following structural condition.

$$X_j = f_j(X_{\mathbf{PA}_j}) + \eta_j, j = 1, ..., p; \qquad \mathbb{E}[\eta_j | X_k] = 0, k \in \mathbf{PA}_j \tag{5}$$

$$Z = \mathbf{X}\theta + \varepsilon; \qquad \mathbb{E}[\varepsilon | \mathbf{X}] = 0 \tag{6}$$

with Gaussian noise terms η_j, ε and linear f_j. Note that this implies P is a multivariate Gaussian, which follows from the fact that all noise terms are Gaussian, and the model is a linear additive noise model (ANM). This is an elemental property of Gaussian distributions, see Lauritzen (1996). The model assumptions of linearity and Gaussianity are done for the sake of simplicity in the latter proof. However, it is arguably possible to generalize the results to partially linear models and non-Gaussian models. Further research is required for this.

The following conditions on P are regularity conditions necessary for the main result to hold, see 3.1. in Belloni et al (2014).

Condition Sparse Model (SM). The model is sparse in the sense that there is a sparse approximation with $s = s_n << n$ and some $\mathbf{S} \subseteq \mathbf{V}$ so that

$$\mathbf{X}\theta = X_{\mathbf{S}}\theta_{\mathbf{S}} + r \tag{7}$$

$$|\mathbf{S}| \leq s \tag{8}$$

$$\|r\|_2 = O(\sqrt{s/n}). \tag{9}$$

Definition 1. The empirical Gram matrix is defined $\mathbb{E}[\mathbf{X}'\mathbf{X}]$. Let the maximal and minimal m-sparse eigenvalues of the Gram matrix given by,

$$\phi_{max}(m)(\mathbb{E}[\mathbf{X}'\mathbf{X}]) = max_{\mathbf{M} \subseteq \{1,...,p\} : |\mathbf{M}| \leq m} \lambda_{max}(\mathbb{E}[X_{\mathbf{M}}' X_{\mathbf{M}}]) \tag{10}$$

$$\phi_{min}(m)(\mathbb{E}[\mathbf{X}'\mathbf{X}]) = min_{\mathbf{M} \subseteq \{1,...,p\} : |\mathbf{M}| \leq m} \lambda_{min}(\mathbb{E}[X_{\mathbf{M}}' X_{\mathbf{M}}]). \tag{11}$$

Condition Sparse Eigenvalues (SE). For sequence $l_n \to \infty, m \geq 1$, such that with a high probability the maximal and minimal $l_n s$-sparse eigenvalues are bounded from above and away from zero.

$$\kappa' \leq \phi_{min}(l_n s)(\mathbb{E}[\mathbf{X}'\mathbf{X}]) \leq \phi_{max}(l_n s)(\mathbb{E}[\mathbf{X}'\mathbf{X}]) \leq \kappa''; \qquad 0 < \kappa' < \kappa'' < \infty. \tag{12}$$

This implies sparseness of submatrix $X_{\mathbf{S}}$, which in turn implies a sparse subgraph $\mathscr{G}_{\mathbf{S}}$ over $X_{\mathbf{S}}$. Note that SE implies that the eigenvalues of the $l_n s$-sparse Gram matrix, $\mathbb{E}[X'_{\mathbf{M}^c} M_{\mathbf{M}} X_{\mathbf{M}^c}]$ with $\mathbf{M} \subseteq \{1,...,p\}$, are bounded in the same manner, since $M_{\mathbf{M}}$ is idempotent.

The following test will be used as decision rule in the algorithm. This test is apt for the given model as it is described. In order to be able to handle other data models this test can be altered, see Shah & Peters (2018), Zhang *et al.* (2017), Zhang *et al.* (2012) and Li & Fan (2020) for other conditional independence tests.

Definition 2. Let *CondIndTest* with significance level α be defined as the following rule. For $j \in \{1,...,p\}, \mathbf{K} \subseteq \{1,...,p\} \setminus \{j\}$ reject the null-hypothesis $r(Z, X_j | X_{\mathbf{K}}) = 0$ against the two-sided alternative $r(Z, X_j | X_{\mathbf{K}}) \neq 0$ if

$$\sqrt{n - |\mathbf{K}| - 3}\mathrm{Z}(Z, X_j | X_{\mathbf{K}}) > \Phi^{-1}(1 - \alpha/2), \tag{13}$$

where $\mathrm{Z}(.)$ denotes the Fisher's z transform

$$\mathrm{Z}(Z, X_j | X_{\mathbf{K}}) = \frac{1}{2}\log\left(\frac{1 + \hat{r}(Z, X_j | X_{\mathbf{K}})}{1 - \hat{r}(Z, X_j | X_{\mathbf{K}})}\right) \tag{14}$$

where $\Phi(.)$ denotes the cdf of the standard Gaussian. Note that the *CondIndTest* implies a threshold $t_\alpha := tanh\left(\frac{\Phi^{-1}(1-\alpha/2)}{n-|\mathbf{K}|-3}\right)$ for the estimated conditional correlation, which satisfies $t_\alpha = O(\sqrt{1/n})$ for $|\mathbf{K}| << n, \alpha > 0$.

Proposition 1. Since P is multivariate gaussian, we have that $r(Z, X_j | X_{\mathbf{K}}) = 0$ if and only if $Z \perp X_j | X_{\mathbf{K}}$.

Proof. Since Z is given by a linear combination of gaussian variables, Z is also gaussian. The claim is an elementary property of the multivariate gaussian distribution, see Lauritzen (1996, Prop.5.2.).

■

Corollary 1.1. The conditional correlation $r(Z, X_j | X_{\mathbf{K}})$ is estimated by

$$\hat{r}(Z, X_j | X_{\mathbf{K}}) = \hat{r}(M_{\mathbf{K}} Z, M_{\mathbf{K}} X_j) \tag{15}$$

This holds, since by assumption P has additive noise, see Kalisch & Buhlmann (2007), or, for a more detailed explanation see Li & Fan (2020).

AdJ is based of the PC algorithm as described in 2.2.2. in Kalisch & Buhlmann (2007). It is identical on the population version except for the decision rule, which in this case is the conditional independence test as defined in definition 2. Note that given the stated conditions on the data, the expected complexity is bounded by a constant. By SM we have that there exists a set $\mathbf{S}_Z$ such that $\mathbb{E}[\hat{r}(Z, X_j|X_{\mathbf{S}_Z})] \leq t$, where t is the threshold as described in Definition 2. In expectation, AdJ is not going to have more than $s+1$ steps, in which case the size of the conditioning set $\mathbf{K}$ is set to s (see line 8 in Algorithm 2). The complexity is then given by 2^{s+1}. Note that s is dependent of n, but since $n = O(\log p)$, we have that the complexity is smaller than $O(2^{\log p})$, which simplifies to $O(p)$. Hence, in expectation the complexity of AdJ utilized on sparse data is linear in p.

Algorithm 2: AdJ

input : Data $(Z, \mathbf{X})$.
output : Set of indices $\hat{\mathbf{S}}_Z$.
$\hat{\mathbf{S}}_Z \leftarrow \{1, ..., p\}$;
l = -1;
repeat
  l = l+1 **repeat**
    foreach $j \in \hat{\mathbf{S}}_Z$ **do**
      Choose (new) $\mathbf{K} \subseteq \hat{\mathbf{S}}_Z$ with $|\mathbf{K}| = l$;
      if *CondIndTest*$(Z, X_j|\mathbf{K})$ *true, i.e. not rejected* **then**
        $\hat{\mathbf{S}}_Z \leftarrow \hat{\mathbf{S}}_Z \setminus \{j\}$
      end
    end
  until *until all j in* $\hat{\mathbf{S}}_Z$ *have been tested*;
until $|\hat{\mathbf{S}}_Z| < l$;

Conjecture 1. $\|\theta_{\hat{\mathbf{S}}_Z^c}\|_2 = O(\sqrt{s \log p / n})$.

Conjecture 1 suggests that the combined effect of any variable that has not been selected by AdJ on Z is small enough, so that imperfect model selection is feasible. It remains to be proven formally.

Theorem 2. Given the above stated conditions on P, if AdJ is run on $(Z, \mathbf{X})$, it returns an index set $\hat{\mathbf{S}}_Z \subseteq \{1, ..., p\}$ that defines an estimator $\hat{\beta} := argmin_{\beta: supp(\beta) \subset \hat{\mathbf{S}}_Z} \|Z - \mathbf{X}\beta\|_2^2$ such that

$$\mathbb{E}\left[\|\mathbf{X}\theta - \mathbf{X}\hat{\beta}\|_2^2\right]^{1/2} = O\left(\sqrt{\frac{s \log p}{n}}\right). \tag{16}$$

Thus, the rate for the loss is $\sqrt{s/n}$, the root of the number of parameters in the true model, s, divided by n, times a logarithmic factor $\sqrt{\log p}$, which can be thought of the price of not knowing the true underlying model.

Theorem 2 implies that the model using as covariates X_j, with $j \in \hat{\mathbf{S}}_Z$ as given by AdJ, is sparse and gives a good approximation of the actual model. In other words, it implies that AdJ allows for imperfect selection of covariates. That is, there might be covariates X_k with $k \in \hat{\mathbf{S}}_Z^c$ but $\theta_k \neq 0$. So, although X_k actually has an effect on Z, it is be statistically independent of Z conditional on $X_{\hat{\mathbf{S}}_Z}$. Therefore, we can do well enough with a model that includes $X_{\hat{\mathbf{S}}_Z}$ and excludes X_k.

We will see in the following section, that applying AdJ solely on the outcome variable - which effectively is post-single selection - might in practice lead to bias due to statistical error. Instead, AdJ has to be applied also on the treatment variable.

3 Intuition for the importance of double selection

In this section I present simulated data to show how post-single selection with AdJ might fail to control for confounders while double-selection overcomes this problem. I follow the simple example presented by Belloni *et al.* (2014b), section 2.4.

Let the data be given by the i.i.d. sample $\{(Y_i, D_i, X_i)\}_{i=1}^n$ distributed after P_n and let the model be an additive noise model with standard Gaussian noise

$$X \sim N(0,1) \tag{17}$$

$$Y = D\alpha + X\beta + \zeta; \qquad \zeta \sim N(0, \sigma_\zeta) \tag{18}$$

$$D = X\gamma + \nu; \qquad \nu \sim N(0, \sigma_\nu) \tag{19}$$

The corresponding DAG is depicted in Figure 4. The model has only one control X that confounds the effect between D and Y. The treatment effect α can be estimated consistently by OLS if X is controlled for.

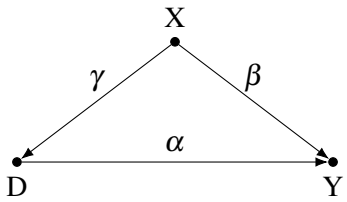

Figure 1: X acts as confounder for the effect of D on Y

Under the stated model we have that $\sigma_D = \gamma^2 + \sigma_\nu^2$ and correlation $\rho_{D,X} := r(D,X) = \gamma\sigma_X/\sigma_D$. If we conduct post-single selection applying AdJ only to (Y,X), the coefficient β will be statistically indistinguishable from 0, and X is omitted with probability $\to 1$, if

$$|\beta| \leq \frac{l_n}{\sqrt{n}} \left(\frac{\sigma_\zeta}{\sigma_X \sqrt{1-\rho_{D,X}}} \right) \tag{20}$$

where $l_n \to \infty$ is a slowly varying sequence depending on P_n. The result is that the asymptotic properties of the post-single-selection estimator depend strongly on P_n. This might lead the estimator to behave badly, that is, have a non-regular distribution.

post-double selection is able to undergo this problem. When running AdJ on both (Y,X) and (D,X), X is only omitted if both coefficients β and γ, which in turn leads to small confounder bias. X is omitted with positive probability only if

$$|\beta| \leq \frac{l_n}{\sqrt{n}} \left(\frac{\sigma_\zeta}{\sigma_X \sqrt{1-\rho_{D,X}}} \right) \qquad \text{and} \qquad |\gamma| \leq \frac{l_n}{\sqrt{n}} \left(\frac{\sigma_v}{\sigma_X} \right). \tag{21}$$

Given this, the post-double-estimator $\hat{\alpha}$ satisfies

$$\sigma_n^{-1} \sqrt{n}(\hat{\alpha} - \alpha) \to_d N(0,1) \tag{22}$$

I simulated data drawn independently 1000 times from the model as described in (18), (19) with $\sigma_\zeta = \sigma_v = 1$ and conducting conditional independence tests with significance level of 0.05. The left panel of Figure 3 depicts the pdf of the standard Gaussian over the the empirical distribution of the post-single estimator. Clearly, it does not satisfy (22) and is badly behaved. The right panel of Figure 3 similarly shows the result of the post-double selection estimator, which does satisfy (22), and is well-behaved.

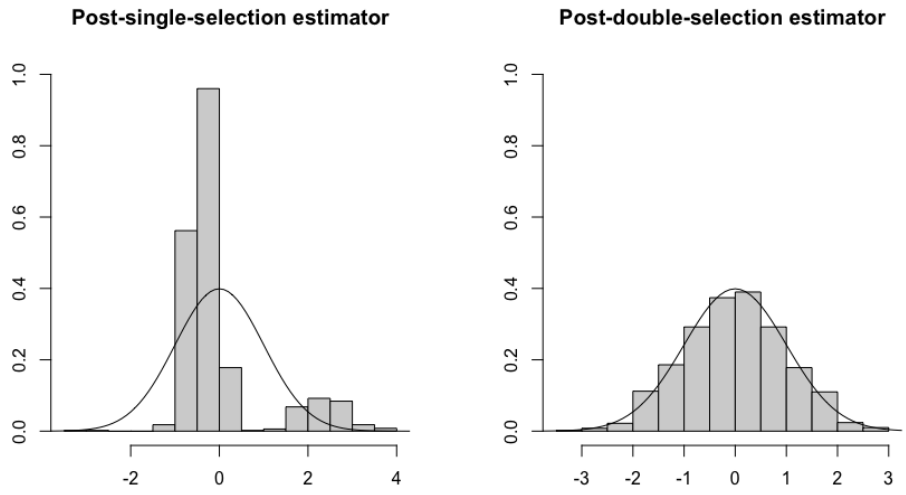

Figure 2: The empirical distributions (densities) of the studentized estimators. The lines indicate the density curve for the standard Gaussian.

4 Estimation and Inference on the Treatment Effect controlling for observable covariates

4.1 Framework

The following will be the framework on which the main result holds. The observed data $\mathscr{D}_n = \{Y_i, D_i, \mathbf{X}_i\}_{i=1}^n$ consist of an outcome $Y \in \mathbb{R}^n$, a treatment $D \in \mathbb{R}^n$ and covariates

$\mathbf{X} \in \mathbb{R}^{n \times p}$, and are generated by P. P satisfies conditions SE and SM. Furthermore, the data generating process is described by the following equations.

$$X_j = f_j(X_{\mathbf{PA}_j}) + \eta_j, j = 1, ..., p; \qquad \mathbb{E}[\eta_j | X_k] = 0, k \in \mathbf{PA}_j \tag{23}$$

$$Y = D\alpha + \mathbf{X}\theta_g + \zeta; \qquad \mathbb{E}[\zeta | D,\ X_j] = 0, \forall j \in \{1, ..., p\} \tag{24}$$

$$D = \mathbf{X}\theta_m + \nu; \qquad \mathbb{E}[\nu | X_j] = 0, \forall j \in \{1, ..., p\} \tag{25}$$

with Gaussian noise terms η_j, ε, ν and linear f_j.

4.2 Main result

Given Theorem 2 holds, the post-double selection estimator with AdJ is root-n consistent and asymptotically normal. Running AdJ on $(D, \mathbf{X})$ and on $(Y, \mathbf{X})$ returns the index sets $\hat{\mathbf{S}}_D \subseteq \{1, ..., p\}$ and $\hat{\mathbf{S}}_Y \subseteq \{1, ..., p\}$. Hence, we can define approximation estimators

$$\hat{\beta}_g := argmin_{\beta : supp(\beta) \subset \hat{\mathbf{S}}_Y} \|Y - \mathbf{X}\beta\|_2^2 \tag{26}$$

$$\hat{\beta}_m := argmin_{\beta : supp(\beta) \subset \hat{\mathbf{S}}_D} \|D - \mathbf{X}\beta\|_2^2. \tag{27}$$

By Theorem 2 the size of the resulting approximation errors is small in probability compared to the estimation error

$$\mathbb{E}\left[\|X\theta_g - \mathbf{X}\hat{\beta}_g\|_2^2\right]^{1/2} = O\left(\sqrt{\frac{s \log p}{n}}\right), \quad \mathbb{E}\left[\|X\theta_m - \mathbf{X}\hat{\beta}_m\|_2^2\right]^{1/2} = O\left(\sqrt{\frac{s \log p}{n}}\right). \tag{28}$$

for some constant $0 < C < \infty$.

We can now define the post-double selection estimator using AdJ as follows

$$(\hat{\alpha}, \hat{\theta}_g) := argmin_{\alpha \in \mathbb{R}, \theta_g : supp(\theta_g) \subset \hat{\mathbf{S}}_Y \cup \hat{\mathbf{S}}_D} \|Y - D\alpha - \mathbf{X}\theta_g\|_2^2. \tag{29}$$

By Theorem 1 of Belloni *et al.* (2014b) it follows that the post-double selection estimator for the treatment effect obeys

$$\hat{\sigma}_\alpha^{-1} \sqrt{n}(\hat{\alpha} - \alpha) \to_d N(0, 1), \tag{30}$$

where $\hat{\sigma}_\alpha$ is the sample standard deviation of $\hat{\alpha}$.

5 Simulation Study

In this section I will study properties of the AdJ algorithm on finite sample data. The simulation are based on the following structural model, which is based on the Monte-Carlo Experiments in Belloni et al (2014),

$$Y = D\alpha + \mathbf{X}\theta_g + \sigma_Y \zeta, \qquad \zeta \sim N(0, 1) \tag{31}$$

$$D = \mathbf{X}\theta_m + \sigma_D \nu, \qquad \nu \sim N(0, 1) \tag{32}$$

where $X \in \mathbb{R}^{n \times p}, D \in R^n, Y \in R^n$. The covariates $\mathbf{X}$ are jointly Gaussian, $\mathbf{X} \sim N(0, \Sigma)$ with $\Sigma_{kj} = (0.5)^{|j-k|}$. The number of observations n is set to 100 and the number of covariates p to 200.

The treatment effect α is set to 0.5. The coefficients θ_m, θ_g are generated so that a specific R^2 is fixed for each of the structural equations (32), denoted Structure R^2, and (31), denoted First Stage R^2.

$$\theta_{m,j} = c_D \beta_j \tag{33}$$

$$\theta_{g,j} = c_Y \beta_j, \tag{34}$$

where $\beta_j = (1/j)^2$ for $j = 1, ..., p$ and c_D, c_Y are tuning parameters. I apply two different simulation designs, A and B. Design A uses homoscedastic error terms, where $\sigma_Y = \sigma_D = 1$. Design B uses heteroscedastic error terms, where $\sigma_Y = \sqrt{\frac{(1+D\alpha+\mathbf{X}\beta)^2}{\mathbb{E}[1+D\alpha+\mathbf{X}\beta]^2}}$ and $\sigma_D = \sqrt{\frac{(1+\mathbf{X}\beta)^2}{\mathbb{E}[1+\mathbf{X}\beta]^2}}$. Note that c_D and c_Y are chosen to determine the R^2 of the structural equation for design A, but are then used also for the simulation of B.

The designs of the simulation data model are such that there is no exact sparse model for either Y or D. In the DAG representing the model we would have that from each the X_j there would be emanating two arrows pointing at each of the treatment and outcome variable. Additionally, between each X_j and X_k there would be a double-header arrows indicating the correlation. If this DAG was known it would not bring us very far in the selection of a section set, since all of X_j would be selected since they are all confounders. This is not helpful since it would not mean any reduction in dimensionality. However, the coefficients decay fast enough such that sparse approximate model exists. This setting thus nicely exemplifies how AdJ is apt for imperfect model selection.

I conduct 1000 simulation runs, each time drawing new $\mathbf{X}, \zeta, \nu$. For each run I estimate α using 4 different methods: The first method is the oracle, which estimates α by regressing $Y - \mathbf{X}\theta)g$ on D by OLS. This is an unfeasible benchmark, since it uses information about the true model, which in practice is unknown. The remaining three methods are feasible. The second method is post-double selection with Lasso regression (Tibshirani, 1996). This method is proposed by Belloni *et al.* (2014b), the paper from which I also largely adopt the simulation designs. Lasso is run of Y on $\mathbf{X}$ to select a subset $S_Y^{Lasso} \subseteq \mathbf{X}$ of predictors for Y and of D on $\mathbf{X}$ to select $S_D^{Lasso} \subseteq \mathbf{X}$ of predictors for Y the union of those sets is denoted S^{Lasso}. I use the following equation from Belloni & Chernozhukov (2011) to choose the penalty

$$\lambda := 2c\sigma\sqrt{n}\Phi^{-1}(1 - \alpha_\lambda/2p), \tag{35}$$

with $c = 1.1$ and $\alpha_\lambda = 0.05$. This penalty is independent of $\mathbf{X}$ unlike the penalty Belloni *et al.* (2014b) use for their simulation. This is important to note, since an $\mathbf{X}$- dependent penalty is likely to improve the post-double selection estimator with Lasso. Nonetheless, using the $\mathbf{X}$-independent penalty will give us an appropriate benchmark for feasible methods.

The third and fourth methods are Post-double selection with AdJ. Their difference lies in the conditional independence test used. One uses the Fisher-Z transformation to test for

Table 1: Simulation Results for selected R^2 values

	First stage R^2= 0.2 Structure R^2= 0.8		First stage R^2= 0.8 Structure R^2= 0.8	
Estimation method	RMSE	Rej. Rate	RMSE	Rej. Rate
A. Homoscedastic design				
Oracle	0.273	0.001	0.141	0
Post-double-sel. w/ Lasso	0.353	0	0.326	0.001
Post-double-sel. w/ AdJ (Fisher-Z)	0.34	0.002	0.32	0.001
Post-double-sel. w/ AdJ (dCov)	0.361	0	0.319	0.001
B. Heteroscedastic design				
Oracle	0.521	0.14	0.371	0.024
Post-double-sel. w/ Lasso	0.564	0.153	0.694	0.322
Post-double-sel. w/ AdJ (Fisher-Z)	0.586	0.173	0.709	0.323
Post-double-sel. w/ AdJ (dCov)	0.586	0.172	0.72	0.324

vanishing conditional correlation. The other uses a regression-based test using covariance distance test (dcov), as advocated in (Shah & Peters, 2018) (see Li & Fan (2020), Zhang *et al.* (2017), Zhang *et al.* (2012) for Regression-based tests). Following this method, to check whether $A \perp\!\!\!\perp B|C$, in a first step A and B are separately regressed on C. In a second step the residuals of each regression are tested for vanishing covariance. This method performs well on ANM, like the model we are using for our simulation. The significance level for both methods is set at 0.01, following the advice of Kalisch & Buhlmann (2007) for conditional independence tests in PC. The treatment effect α is finally estimated by running OLS of Y on D and the post-double adjustment set for each method respectively and inference on α is done by heteroscedasticity robust OLS inference.

The simulation is conducted for both simulation designs and for different combinations of Structure and First Stage R^2. The Structure R^2 is kept fixed at 0.8, while the First Stage R^2 varies between 0.2 and 0.8. Table 1 reports the root-mean-squared error (RMSE) and the rate that the significance of the treatment effect α is rejected at the 0.01 significance level (Rejection Rate). As would be expected, the oracle estimator shows the lowest RMSE in any case. It is apparent higher First Stage R^2, as well as homoscedastic errors, lead to lower RMSE. Surprisingly, The Rejection Rate is not always lowest for the Oracle. Further, the results for the three post-double selection methods are in the same ball-park. AdJ seems to work slightly better in the homoscedastic design, according to the RMSE, while Lasso does so in the heteroscedastic one. The Rejection Rate, however, is usually higher when using AdJ. The results also show that tehre is little difference in using Fisher-Z or dcov for conditional; independence test. Overall, the results show that the performance of the post-double selection estimator with AdJ is comparable with that of Lasso for post-double selection.

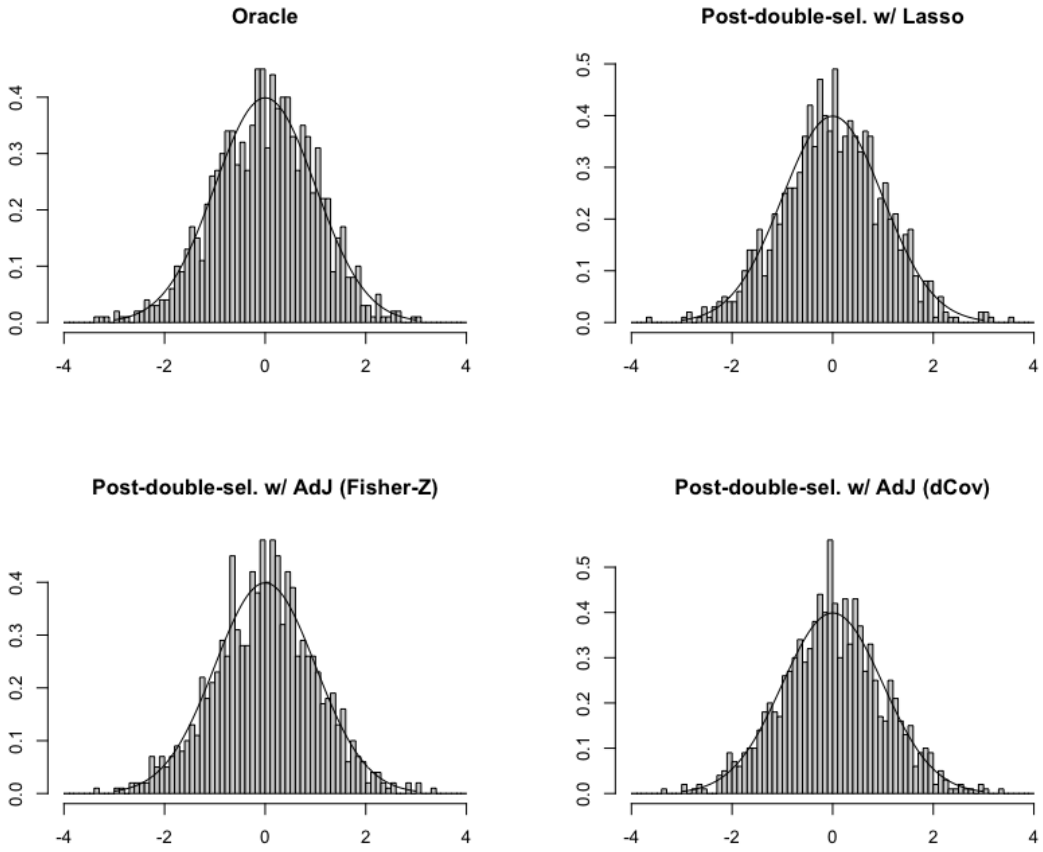

Figure 3: The empirical distributions (densities) of the studentized estimators for simulation design A with First Stage $R^2 = 0.8$, Structure $R^2 = 0.8$. The lines indicate the density curve for the standard Gaussian.

I further provide details about the studentized distribution of the simulated estimators for Design A with First Stage $R^2 = 0.8$ and Structure $R^2 = 0.8$. It is apparent that all the estimators are root-n consistent.

6 Empirical Example: Estimating the effect of Institutions on Output

In this section I will exemplify the use of post-double with AdJ on a real data set. I will work through the causal analysis by Acemoglu, Johnson and Robinson (2001), henceforth referred to as AJR, using the AdJ-post-double selection estimator. The aim is to estimate the causal effect institutions have on income in ex-colonial states. The argument goes that better institutions are more able to create favourable conditions for economic development,

for instance by effectively enforcing property rights, and this lead to higher income. The difficulty comes from the possible simultaneous causality: Richer countries can afford and might be more likely to choose better institutions. Additionally there might be confounding factors that lead a country to have both, better institutions and higher output.

In order to estimate the causal effect of institutions on income, AJR use an instrumental variable regression. As exogenous variation in institutions they choose settler mortality. Their theory states that in places with higher settler mortality, European colonizers implemented extractive colonisation policies, which did not need property-rights enforcement to protect from expropriation. On the other hand, where European migrants could settle, better institutions were erected to protect their property. Further, these colonial institutions persisted after the states gained independence. Finally, they argue that settler morality in colonial times is unlikely to influence GDP directly. Settler mortality is thus a valid instrument. However, it is possible that there exist confounding factors between settler mortality and output, which would invalidate the instrumental variable. AJR argue that geographical factors are likely to act as confounders. They assume that controlling for distance from the equator and a set of dummies is sufficient to control for confounding. As an addition to their method, I apply double-selection with AdJ. This relaxes the assumption about the sufficiency of the adjustment set done by the researchers, which controls, among the available geographical factors, will be selected depends now only on the given data.

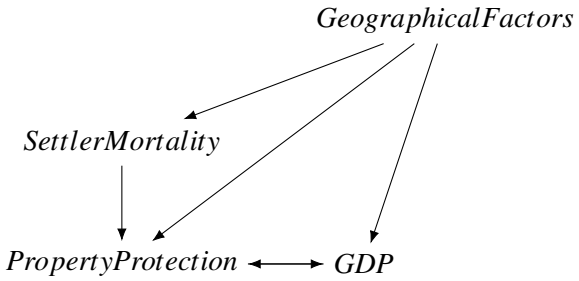

Figure 4: The bidirected arrow indicates that the direction of the causal effect might go both ways. *SettlerMortality* is a valid instrument for the effect of *PropertyProtection* on *GDP*, given we control for putatively confounding *GeographicalFactors*.

I use the same data as AJR, 64 country-observations. Strength of institutions is measured by the proxy *PropertyProtection*, an index for extent to which property rights are enforced against expropriation. *SettlerMortality* stands for the logarithmized mortality rates of settlers and *GDP* is the logarithmized GDP per capita. Our set of controls *GeographicalFactors* consists of a set of 16 variables, namely 4 dummies for Africa, Asia, North America and South America, the latitude and a cubic spline for latitude with nodes at 0.08, 0.16 and 0.24. The result is a dimension of the control set of 16. 4 depicts the underlaying DAg as it is assumed in AJR.

Following AJR's methodology I estimate the effect of institutions on output by IV estimation. In the first stage *PropertyProtection* is regressed on *SettlerMortality* and the selected adjustment set. The resulting coefficient on *SettlerMortality* and its standard error are reported. In the second stage *GDP* is regressed on the predicted *PropertyProtection* and the selected adjustment set. The resulting coefficient on predicted *PropertyProtection* and its standard error are reported.

Table 2 reports the results of the IV estimator for three different methods to select the adjustment set, Latitude, All controls and post-double selection with AdJ. Latitude refers to the estimator using only Latitude as adjustment set, All controls uses all 16 controls, and for post-double selection with AdJ the adjustment set is selected by running AdJ for each of *GDP*, *SettlerMortality* and *PropertyProtection* with respect to *GeographicalFactors* and taking the union of the resulting index sets. This way, I should be able to control for any possibly lurking confounders between every pair of those three variables. The result is an adjustment set consisting of the Africa dummy and the variable $(\text{latitude} - 0.24)_+$, a component of the cubic spline[2].

The Latitude-estimator corresponds to the result of AJR. It is significant at the 0.001-level, which suggests that institutions have a strong positive on income. The All-controls-estimator, on the other hand, is not statistically significant, which results in a very imprecise estimator. Both first and second stage show rather large standard errors.

The post-double selection-estimator is significant only at the 0.05-level and although it is a bit attenuated in size it is still strong. The first stage is also weaker than the Latitude-estimator. These findings are very much in line with the ones obtained by Belloni *et al.* (2014a), where the authors conduct a similar exercise but using Lasso instead of AdJ for model selection.

The results come to show that the model selected by post-double selection represents a trade-off in variance between the first two models. This indicates the usefulness that high-dimensional selection methods for causal inference can have for practitioners, as an addition to sensitivity analyses. In this case it enabled us to relax any assumptions on whether a variable is a confounder or not, by letting the decision be dependant of the data.

This example also illustrates how the structural assumptions made by the practitioner can be nicely accommodated within the post-double framework. While some assumptions are necessary, namely the definition of the baseline set of controls, using post-double selection allows some space to let the data decide which controls to use. This is, I would argue, is a more rigorous approach to covariate adjustment than only relying on expert judgement, and its findings should therefore be granted greater credibility.

[2] $(a)_+$ denotes the function $f(a) = a * 1(a > 0)$, where $1(.)$ is the indicator function.

Table 2: Effect of Institutions on Output

	Latitude	All controls	Post-double-sel. w/ AdJ
First Stage	-0.5372	-0.2175	-0.3329
	(0.1545)	(0.2145)	(0.1840)
Second stage	0.9692	0.9839	0.8152
	(0.1546)	(0.7396)	(0.3655)

7 Conclusion

In this thesis, I present the AdJ algorithm, a novel method for covariate selection that is appropriate for post-double selection. It exploits conditional independencies, much like similar graphical methods aimed at structure-learning, such as the PC-algorithm. Note, however, that AdJ's goal is not to learn about the structure *per se*, but merely to find the set of parents of one target variable. This means that usually much less causal information than the entire DAG is retrieved. It is, however, sufficient for our purpose of evaluating treatment effects.

I provide partial results to show that AdJ is able to conduct post-double selection, as it is prescribed in Belloni *et al.* (2014b). This depends on a con Conjecture 1 in 2.3, which is the current focus of my still ongoing project. Further, I simulate data and apply post-double selection with AdJ to illustrate its beneficial properties. I also exemplify its use by means of an empirical example.

Further research is necessary to extend my results to other, more general models that do not make assumptions of linearity and Gaussianity. The statistical properties of AdJ can also be investigated in more detail. In this thesis I limit myself to the the necessary properties required for Theorem 2. Moreover, possible overfitting of AdJ-model selection could also be investigated in detail. It has, however, not shown to be an issue in the presented simulations or the empirical example.

8 Appendix

8.1 Proof for Theorem 1

Proof. By Faithfulness the decision rule in the algorithm implies that whenever $j \notin \mathbf{S}_Z$ we have that Z *d-sep* $j|\mathbf{K}$ for some $\mathbf{K}$. Therefore no node in $\mathbf{S}_Z^c$ is adjacent to Z.

Moreover, whenever $j \in \mathbf{S}_Z$ we have that there exists no $\mathbf{K}$ s.t. Z *d-sep* $j|\mathbf{K}$. Therefore, any node in $\mathbf{S}_Z$ is adjacent to Z. By LCS we have that any adjacent node of Z is also an ancestor of it (no hidden common causes) and by PES there are no colliders adjacent to Z. So every path going into Z is blocked by some subset of $\mathbf{S}_Z$

But this implies that any path going into Z is blocked by $\mathbf{S}_Z$. Therefore Z *d-sep* $\mathbf{S}_Z^c|\mathbf{S}_Z$. Which, by Markov, implies that $Z \perp\!\!\!\perp X_{\mathbf{S}_Z^c}|X_{\mathbf{S}_Z}$.

Furthermore, the decision rule implies that if $j \in \mathbf{S}_Z$ there is no $\mathbf{K}$ such that $Z \perp\!\!\!\perp X_j|X_{\mathbf{K}}$. Therefore $\mathbf{S}_Z$ is minimal and we have that $\nexists j \in \mathbf{S}_Z : Z \perp\!\!\!\perp X_j|X_{\mathbf{S}_Z \setminus j}$.

■

8.2 Proof of Theorem 2

Proof. We can write the expected value of the estimator in matrix notation and reformulate.

$$\hat{\beta} = (X'_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z})^{-1} X'_{\hat{\mathbf{S}}_Z} Z \tag{36}$$

$$= (X'_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z})^{-1} X'_{\hat{\mathbf{S}}_Z} X\theta \tag{37}$$

$$= \theta_{\hat{\mathbf{S}}_Z} + (X'_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z})^{-1} X'_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z^c} \theta_{\hat{\mathbf{S}}_Z^c} \tag{38}$$

The predictions of Z are then given by

$$X_{\hat{\mathbf{S}}_Z}\hat{\beta}_{\hat{\mathbf{S}}_Z} = X_{\hat{\mathbf{S}}_Z}\theta_{\hat{\mathbf{S}}_Z} + X_{\hat{\mathbf{S}}_Z}(X'_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z})^{-1} X'_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z^c} \theta_{\hat{\mathbf{S}}_Z^c}. \tag{39}$$

And the expected approximation error is given by

$$\mathbb{E}\left[X\theta - X_{\hat{\mathbf{S}}_Z}\hat{\beta}_{\hat{\mathbf{S}}_Z}\right]. \tag{40}$$

Since the model is additive, with $\mathbf{X}\theta = X_{\hat{\mathbf{S}}_Z} + X_{\hat{\mathbf{S}}_Z^c}$, the expected approximation error can be written as

$$\mathbb{E}\left[M_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z^c}\right] \theta_{\hat{\mathbf{S}}_Z^c}. \tag{41}$$

Take the $L2$ norm of the approximation error:

$$\mathbb{E}\left[\|X\theta - X_{\hat{\mathbf{S}}_Z}\hat{\beta}_{\hat{\mathbf{S}}_Z}\|_2^2\right] \tag{42}$$

$$= \theta'_{\hat{\mathbf{S}}_Z^c}\mathbb{E}\left[X'_{\hat{\mathbf{S}}_Z^c} M_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z^c}\right] \theta_{\hat{\mathbf{S}}_Z^c}. \tag{43}$$

And since $X'_{\hat{\mathbf{S}}_Z^c} M_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z^c}$ is symmetric it holds that

$$\mathbb{E}\left[\|X\theta - X_{\hat{\mathbf{S}}_Z}\hat{\beta}_{\hat{\mathbf{S}}_Z}\|_2^2\right] \leq \phi_{max}\left(\mathbb{E}\left[X'_{\hat{\mathbf{S}}_Z^c} M_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z^c}\right]\right) \|\theta_{\hat{\mathbf{S}}_Z^c}\|_2^2. \tag{44}$$

And have have that the right-hand side is of order $O(s\log p/n)$, since by SE we have that the first term is bounded by some constant and by 2.3 we have that

$$\|\theta_{\hat{\mathbf{S}}_Z^c}\|_2^2 = O\left(\frac{s\log p}{n}\right). \tag{45}$$

■

References

Acemoglu, Daron, Johnson, Simon, & Robinson, James A. 2001. The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review*, **91**(5), 1369–1401.

Angrist, Joshua D., & Krueger, Alan B. 1999. Empirical Strategies in Labor Economics. *Handbook of Labor Economics*, **3**(1), 1277–1366.

Athey, Susan, & Imbens, Guido W. 2017. The state of applied econometrics: Causality and policy evaluation. *Pages 3–32 of: Journal of Economic Perspectives*, vol. 31. American Economic Association.

Belloni, Alexandre, & Chernozhukov, Victor. 2011. High Dimensional Sparse Econometric Models: An Introduction. *arXiv:1106.5242*, 6.

Belloni, Alexandre, Chernozhukov, Victor, & Hansen, Christian. 2014a. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, **28**(2), 29–50.

Belloni, Alexandre, Chernozhukov, Victor, & Hansen, Christian. 2014b. Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, **81**(2), 608–650.

Chernozhukov, Victor, Chetverikov, Denis, Demirer, Mert, Duflo, Esther, Hansen, Christian, Newey, Whitney, & Robins, James. 2018. Double/Debiased Machine Learning for Treatment and Structural Parameters 1. *The Econometrics Journal*, **21**(1), C1–C68.

Heckman, James J, & Vytlacil, Edward. 2005. Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, **73**(3), 669–738.

Hoyer, Patrik O, Janzing, Dominik, Mooij, Joris, Peters, Jonas, & Schölkopf, Bernhard. 2008. Nonlinear causal discovery with additive noise models.

Imbens, Guido W. 2020. Potential Outcome and Directed Acyclic GraphApproaches to Causality: Relevance forEmpirical Practice in Economics. *arXiv:1907.07271v2*.

Kalisch, Markus, & Buhlmann, Peter. 2007. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm Peter Buhlmann. *Journal of Machine Learning Research*, **8**, 613–636.

Kozbur, Damian. 2020. Analysis of Testing-Based Forward Model Selection. *Econometrica*, **88**(5), 2147–2173.

Lauritzen, Steffen L. 1996. *Graphical Models*. Oxford University Press.

Li, Chun, & Fan, Xiaodan. 2020. On nonparametric conditional independence tests for continuous variables. *WIREs Computational Statistics*, **12**(3).

Lovell, Michael C. 2008. A Simple Proof of the FWL Theorem. *The Journal of Economic Education*, **39**(1), 88–91.

Pearl, Judea. 2009a. *Causality, Models Reasoning and Inference*. Second edn. New York, NY: Cambridge University Press.

Pearl, Judea. 2009b. Remarks on the method of propensity score. *Statistics in Medicine*, **28**(9), 1415–1416.

Pearl, Judea. 2012. The Causal Foundations of Structural Equation Modeling. *Chap. 5, pages 68–91 of:* Hoyle, R.H. (ed), *Handbook of Structural Equation Modeling*. New York: Guilford Press.

Perkovi´c, Emilija, Textor, Johannes, Kalisch, Markus, & Maathuis, Marloes H. 2015. A Complete Generalized Adjustment Criterion. *arXiv:1507.01524*.

Peters, Jonas, Mooij, Joris M, Janzing, Dominik, & Schölkopf, Bernhard. 2014. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research*, **15**, 2009–2053.

Peters, Jonas, Janzing, Dominik, & Schoelkopf, Bernhard. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: Adaptive Computation and Machine Learning MIT Press.

Shah, Rajen D., & Peters, Jonas. 2018. The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *arXiv:1507.01524*, 4.

Shpitser, Ilya, Vanderweele, Tyler, & Robins, James M. 2012. On the Validity of Covariate Adjustment for Estimating Causal Effects. *arXiv:1203.3515*.

Simpson, E H. 1951. The Interpretation of Interaction in Contingency Tables. *Source: Journal of the Royal Statistical Society. Series B (Methodological)*, **13**(2), 238–241.

Spirtes, Peter. 2010. Introduction to Causal Inference. *Journal of Machine Learning Research*, **11**, 1643–1662.

Spirtes, Peter, Glymour, Clark, & Scheines, Richard. 2000. *Causation, Prediction, and Search*. Second edn. Cambridge, Massachusetts: The MIT Press.

Tibshirani, Robert. 1996. Regression Shrinkage and Selection via the Lasso. *Source: Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Van Der Zander, Benito, Li´skiewiczli´skiewicz, Maciej, & Textor, Johannes. 2019. Separators and Adjustment Sets in Causal Graphs: Complete Criteria and an Algorithmic Framework. *arXiv:1803.00116v3*.

Vanderweele, Tyler J, & Shpitser, Ilya. 2011. A New Criterion for Confounder Selection. *Biometrics*, **67**, 1406–1413.

Witte, Janine, Henckel, Leonard, Maathuis, Marloes H, & Didelez, Vanessa. 2020. On Efficient Adjustment in Causal Graphs. *Journal of Machine Learning Research*, **21**, 1–45.

Zhang, Hao, Zhou, Shuigeng, Zhang, Kun, Guan, Jihong, & Key, Shanghai. 2017. Causal Discovery Using Regression-Based Conditional Independence Tests. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, **31**(1).

Zhang, Kun, Peters, Jonas, Janzing, Dominik, & Schölkopf, Bernhard. 2012. Kernel-based Conditional Independence Test and Application in Causal Discovery. *arXiv:1202.3775*.