Master's thesis
presented to the Department of Economics
of the University of Zurich
for the degree of
Master of Arts UZH in Business and Economics

Model Selection with Graphical Considerations

Author: Jorge de la Cal Medina
Student ID Nr.: 13755541

Supervised by Professor Damian Kozbur

Submission date: 12.01.2022

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 problem faced by policy advisors. If data from an experimental setting are available, where the 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 often readily available. The difficulty here lies in their proneness to endogeneity in the treatment assignment. This aggravates the problem of identifying causal effects considerably. Taking differences between outcomes could lead to confounding bias, or omitted variable bias, which can completely distort the actual effect. An extreme example of this distortion is described in Simpson's paradox, where not controlling for a confounding variable completely reverses the true effect when it is estimated (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) controlling for X. The treatment effect of D on Y, can thus be estimated by regressing Y on D and X (Lovell, 2008). The resulting estimator is unbiased, provided that any confounder has been effectively controlled for. Hence, when evaluating treatment effects with observational data, it is necessary to select a set of controls that adjusts for confounding.

In order to estimate causal effects via covariate adjustment 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 must include all common causes of D and Y (Spirtes *et al.*, 2000). This assumption concerns the available data; it is a prerequisite for successful covariate adjustment and will hereafter be taken as given.

However, even if the observed covariate set is causally sufficient for estimating the treatment effect, it is 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, such as collider bias or M-bias (Pearl, 2009b). Moreover, controlling for all covariates regardless, would at best produce an inefficient estimate, even if it turns out to be unbiased, since too much of the variance in the explanatory variables would be unnecessarily reduced.

This issue is only aggravated with high-dimensional data. In this case, conducting regression methods might be entirely unfeasible when the number of covariates is larger than the number of observations. Nonetheless, it is important to develop new methods that address the problems inherent to high-dimensional data. As the number of observed characteristics

has grown in recent years with increasing data availability and computing power, it has become more accessible to empirical researchers.

The task of model selection is to select a subset of covariates out of the full set of available ones, that is able to explain the data. Expert knowledge may suggest such a subset. But only relying on experts lacks the scientific rigour we would wish for. 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 from using different control sets come about (Belloni *et al.*, 2014b).

To address this problem more rigorously, I develop a testing-based model selection method that uses estimated causal relations to conduct root-n consistent estimation, where n is the number of observations. Specifically, I use a causal structure-learning algorithm to select a sparse model. Since mistakes might occur when selecting the model, the treatment effect can then be estimated using the post-double selection framework. In the following, I will clarify these concepts and lay out the aim of the paper in more detail.

I will assume that the underlying data generating process (dgp) can be described by structural equation - a common assumption in policy evaluation (Heckman & Vytlacil, 2005). This allows me to describe the dgp as a directed acyclical graph (DAG). Besides depicting causal relations in an intuitive manner, using DAGs 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, provided the Causal Markov property holds, the set of *immediate* causes 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) for a more detailed discussion of this result. 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 (Vanderweele & Shpitser, 2011) - a desideratum when search algorithms are applied.

The *PC* algorithm (Spirtes *et al.*, 2000) is a method for causal structure learning, which aims at learning the DAG from observational data. As stated above, for our purpose 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 for conditional independencies between each covariate and Y. An omniscient oracle, that knows about any conditional independencies between the variables, would this way 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*. This can be problematic, since the conditional independence test might fail to detect actual confounders. The result is an estimator that is not root-n consistent and non-regular. In this paper, I motivate this issue with a simulation, see section 3.

To prevent this, the authors propose the *post-double selection* framework, which applies two selection steps instead of one. First, a set $\mathbf{S}_Y$ of covariates that are predictive of outcome Y is selected by the model-selection method of choice. Second, a set $\mathbf{S}_D$ that is predictive of D is similarly selected. The exact meaning 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 $\mathbf{S}_D$. Belloni *et al.* (2014b) show that the post-double selection estimator is root-n consistent. They use l_1-penalized methods, e.g. Lasso, for model selection, but highlight that the theoretical results are applicable to other methods as well.

Post-double selection allows to draw sound conclusion on causal effects from observational data with high confounding. It is apt for high-dimensional settings and encourages data-mining, without necessarily losing interpretability of the model. Further, post-double selection adds rigour, 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.

The main contribution of this paper 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]. The AdJ algorithm is based on PC, but unlike the latter, it is aimed at finding only the adjacencies with respect to one target variable, rather than the entire graph. Since in the setting that I focus on only pre-treatment variables are considered, the adjacent variables will be the parental set. Hence, the AdJ algorithm is aimed at producing the parental set of the target variable.

AdJ 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 is deleted from the adjustment set. This approach differs from other machine learning model-selection methods that optimize the model relative to an objective function. Instead, AdJ applies causal theory to uncover the putative causal relations.

I analyze AdJ in its application for high-dimensional problems. First, I show the causal sufficiency of the model selected using the population version of AdJ. The population version assumes an unattainable oracle to be known. Second, I provide partial results to show that the sample version of AdJ allows for imperfect model selection, enabling it to handle high-dimensional data. This is a sufficient condition for using post-double selection with AdJ.

[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.

These results, however, depend on a conjecture, that has yet to be proven formally. It remains a research component of the ongoing project. Further, I show its applicability by conducting a simulation study and an empirical example.

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 for 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 selected model captures noise terms rather than structural parameters. In this paper, 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 in the way causality is treated by economists. In that sense, this paper 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. Namely, 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 paper 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 results in order to determine which are more appropriate given the specificities of the data.

This paper 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

This section investigates the possibility of using causal discovery methods to conduct imperfect model selection. It states the desired result and the conditions under which it holds. The appendix provides proofs for these results. Note that theorem 2 has not been proven entirely here; instead I provide an outline of a possible way to prove it.

The primary framework is a high-dimensional, sparse and linear dgp. The sparsity assumption requires that the data-generating process can be approximated by a function over a small number of covariates - small relative to the number of observations. Linearity is assumed so that known properties of additive node models (ANM) can be used. However, it is arguably possible to extend the here presented results of AdJ extent to nonlinear settings.

I will first define a population version of AdJ 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 AdJ that uses feasible conditional independence testing and show that this algorithm 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 ones described in Belloni *et al.* (2014b) and Kozbur (2020).

2.1 Preliminaries

It is useful to first 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 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 adjacent 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 and $i_{k+1}, \forall k = 1, ..., n-1$. 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 holds: 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* $\mathbf{B}|\mathbf{M}$. P($\mathbf{X}$) is said to be Markov and faithful with respect to $\mathscr{G}$ 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 $X_{\mathbf{W}} := [X_j : j \in \mathbf{W}]$ 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. Let P be a distribution over variable tuple $(Z, \mathbf{X})$, where Z represents the target variable and let $\mathbf{X} = \{X_1,...,X_p\}$ be the covariates with index set $\mathbf{V} = \{1,...,p\}$. Further, let P satisfy the following conditions.

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).

A Markovian and faithful graph is necessary for AdJ to find a valid adjustment set. It allows to exploit conditional independences in the data to detect adjacencies in the graph. This will be enlightened in the proof for theorem 1, see the appendix.

Condition policy evaluation Setting (PES). All covariates $\mathbf{X}$ are potential causes of Z. Z 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 where only pre-treatment covariates are observed. It implies that any vertices adjacent to Z are its parents. It might be possible to define an extension of AdJ which allows to relax this condition, but the algorithm would turn out to be more complex. Although this assumption requires justification in practice, I will here take it as given considering the simplification it entails for proving the results.

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.

Algorithm 1 describes AdJ$_{pop}$, the population version of AdJ. It is a modification of the PC algorithm, see 2.2.1. in Kalisch and Bühlmann (2007). It will be useful to show that in principle AdJ is able to select a valid adjustment set.

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 the complexity 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 a useful result, 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 $\mathscr{D}_n = (Z, \mathbf{X})$, with $Z \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n \times p}$, and are generated by a distribution P. Since, I will focus on high-dimensional data, assume

henceforth without loss of generality that $n = O(\log p)$. P satisfies the following structural equations.

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

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

with Gaussian noise terms η_j, ε and linear function f_j. Note that Gaussianity in the error terms implies P is a multivariate Gaussian, since the dgp is a linear additive noise model. This is an elemental property of Gaussian distributions, see Lauritzen (1996). The assumptions of linearity and Gaussianity are done for the sake of simplicity in the proof of theorem 2. However, it is arguably possible to generalize the results to partially linear and non-Gaussian dgps. Further research could address this.

The following conditions are necessary for the main result to hold. Let P satisfy the conditions described in the previous subsection 2.2. Additionally, let P also satisfy the following regularity conditions, which are similar to the regularity conditions described in 3.1. in Belloni *et al.* (2014b).

The main condition states that there exist a sparse model approximation of size s, such that regression methods can be applied and which entails a small approximation error compared to the estimated statistical error.

Condition Sparse Model (SM).

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

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

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

with $s = s_n << n$ and some $\mathbf{S} \subseteq \mathbf{V}$.

The next conditions concern the behaviour of the gram matrix. Whenever we have that $p > n$, the empirical Gram matrix is singular and not well-behaved. However, since we are selecting a subset of covariates, we only need that the corresponding empirical Gram submatrix is well-behaved. To ensure this is fulfilled, it is required that all submatrices formed by m covariates are well-behaved. The following definition will be useful to formulate this condition.

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

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

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

To ensure that all empirical submatrices are positive definite, i.e. their minimal eigenvalue is greater zero, $\phi_{min}(m)(\mathbb{E}[\mathbf{X}'\mathbf{X}]) > 0$, the following condition is sufficient. It is formulated after condition SE in Belloni *et al.* (2014b).

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

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

Further, it is required that the eigenvalues of the gram matrix of the covariates that have not been selected, are bounded from above. For this we need that, if m covariates have been selected for the model, the eigenvalues of any submatrix formed by the remaining $(p-m)$ covariates are bounded from above.

Condition Bounded Eigenvalues (BE). For some sequence $l_n \to \infty$ we have that the maximal $(p-s)$-sparse eigenvalues from the residual Gram Matrix are bounded from above, with $s \geq 1$,

$$\phi_{max}(p - l_n s)(\mathbb{E}[\mathbf{X}'\mathbf{X}]) \leq \kappa; \qquad 0 < \kappa < \infty. \tag{13}$$

For the sample version of AdJ we need, as decision, a test for conditional independence. The following test is apt for Gaussian linear models. In order to be able to handle other dgps, which can feature non-linearities and non-Gaussian errors, other tests can be implemented instead. I illustrate this possibility in section 5 and 6, where I apply AdJ, as it is here described, to simulated, respectively empirical data; and additionally apply AdJ with an alternative conditional independence test. See Shah & Peters (2018), Zhang *et al.* (2017), Zhang *et al.* (2012) and Li & Fan (2020) for different feasible 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

$$\sqrt{n - |\mathbf{K}| - 3}\mathrm{Z}(Z, X_j|X_\mathbf{K}) > \Phi^{-1}(1 - \alpha/2), \tag{14}$$

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

$$\mathrm{Z}(Z, X_j|X_\mathbf{K}) = \frac{1}{2}\log\left(\frac{1 + \hat{r}(Z, X_j|X_\mathbf{K})}{1 - \hat{r}(Z, X_j|X_\mathbf{K})}\right), \tag{15}$$

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 $Z \perp X_j|X_\mathbf{K}$.

> 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 2. The conditional correlation $r(Z, X_j|X_{\mathbf{K}})$ is estimated by

$$\hat{r}(Z, X_j|X_{\mathbf{K}}) = \hat{r}(M_{\mathbf{K}}Z, M_{\mathbf{K}}X_j) \tag{16}$$

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

Algorithm 2 describes AdJ. It is based on the PC algorithm as described in 2.2.2. in Kalisch & Buhlmann (2007). It is identical to AdJ$_{pop}$, 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 not exponential but is bounded by a constant. By SM we have that there exists a set $\mathbf{S}$ such that $\mathbb{E}[\hat{r}(Z, X_j|X_{\mathbf{S}})] \leq t, j \in \mathbf{S}^c$, 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 $O(2^{\log p})$, which simplifies to $O(p)$. Hence, in expectation the complexity of AdJ applied on a sparse dataset 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 $|\hat{\mathbf{S}}_Z| < l$;

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

$$\mathbb{E}\left[\|\mathbf{X}\theta - \mathbf{X}\hat{\beta}\|_2^2\right]^{1/2} = O\left(\sqrt{\frac{s \log p}{n}}\right). \tag{17}$$

Thus, the rate for the loss is $\sqrt{s/n}$, the root of the number of parameters in the true approximation model divided by the number of observations, times a logarithmic factor $\sqrt{\log p}$, which can be thought of the price of not knowing the true underlying dgp (Belloni *et al.*, 2011).

Theorem 2 implies that the model using as covariates $X_{\hat{\mathbf{S}}_Z}$, is sparse and gives a good approximation, where $\hat{\mathbf{S}}_Z$ is the result of running AdJ on Z. 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 any X_k actually has an effect on Z, it is statistically independent of Z conditional on $X_{\hat{\mathbf{S}}_Z}$. Therefore, we can do well enough with a model that only includes $X_{\hat{\mathbf{S}}_Z}$ as covariates. To sum up, AdJ is able to find a sparse approximation model, when such an approximation exists.

We will see in the following section, that, for policy evaluation, applying AdJ solely on the outcome variable might in practice lead to bias due to statistical error. Instead, AdJ should be applied additionally on the treatment variable, as will be explained in more detail.

3 Intuition for the importance of double selection

This section discusses the application of AdJ on simulated data that imitates a simple policy-evaluation scenario where we have a treatment, an outcome and an observed covariate. It shows how post-single selection can fail to control for confounders while double-selection overcomes this problem. It follows the simulation presented in 2.4 Belloni *et al.* (2014b).

Let the data be given by the i.i.d. sample, that includes variables outcome variable $Y \in \mathbb{R}^n$, treatment variable $D \in \mathbb{R}^n$ and covariate, $X \in \mathbb{R}^n$. Let the data generating process be an additive noise model with standard Gaussian noise.

$$X \sim N(0, \sigma_X) \tag{18}$$

$$Y = D\alpha + X\beta + \zeta; \qquad \zeta \sim N(0, \sigma_\zeta) \tag{19}$$

$$D = X\gamma + v; \qquad v \sim N(0, \sigma_v) \tag{20}$$

The corresponding DAG is depicted in Figure 4. X confounds the treatment effect α of D on Y. Hence, α is estimated consistently by OLS if X is controlled for.

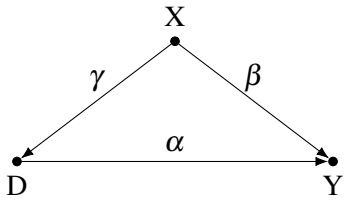

Figure 1: X acts as confounder for the effect of D on Y

We have that $\sigma_D = \gamma^2 + \sigma_v^2$ 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. X is omitted from the selected model with probability $\to 1$, if

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

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 can lead the estimator to behave badly and have a non-regular distribution.

Fortunately, post-double selection is able to overcome this problem. When running AdJ on both (Y,X) and (D,X), X is only omitted in case both coefficients β and γ are small, which in turn would lead to small confounder bias. X is omitted with positive probability whenever

$$|\beta| \leq \frac{l_n}{\sqrt{n}} \left(\frac{\sigma_\zeta}{\sigma_X \sqrt{1-\rho_{D,X}}} \right) \qquad \text{and} \qquad |\gamma| \leq \frac{l_n}{\sqrt{n}} \left(\frac{\sigma_v}{\sigma_X} \right). \tag{22}$$

Let the post-double-estimator $\hat{\alpha}$ be the coefficient on D when regressing Y on D and the, if at all, selected covariate. It follows that it satisfies

$$\sigma_n^{-1} \sqrt{n}(\hat{\alpha} - \alpha) \to_d N(0,1), \tag{23}$$

where $\sigma_n^2 = \sigma_\zeta (\sigma_v^2)^{-1}$.

Data is simulated with number of obervations $n = 100$, drawn independently 10000 times from the dgp as it is described in (18) to (20) with $\sigma_X = \sigma_\zeta = \sigma_v = 1$. The parameters are chosen in a way, such that post-single selection leads to badly behaved estimator, with $\alpha = 1, \beta = 0.1, \gamma = -4$. AdJ is then applied to the simulated data as described in section 2.3 with significance level of the conditional independence test set to 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 (23) and is badly behaved. The right panel of Figure 3 analogically shows the result of the post-double selection estimator, which does satisfy (23), and is well-behaved.

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

The purpose of this section is to present the main result and describe the setting in which it holds. This result is sufficient for the applicability of AdJ within the post-double selection framework. Specifically, it is stated that the post-double selection estimator produced with AdJ is root-n consistent and asymptotically normal. It thus has the same beneficial properties as the post-double selection estimator described by Belloni *et al.* (2011). The conditions on the data and the required properties of AdJ for the main results are described in section 2.

The main result implies that AdJ can be applied for imperfect model selection and is thus a feasible method for causal analysis in high dimensional settings. Arguably, this finding might lead to new developments and extent the application of graphical methods, such as AdJ, to policy evaluation and other applied problems.

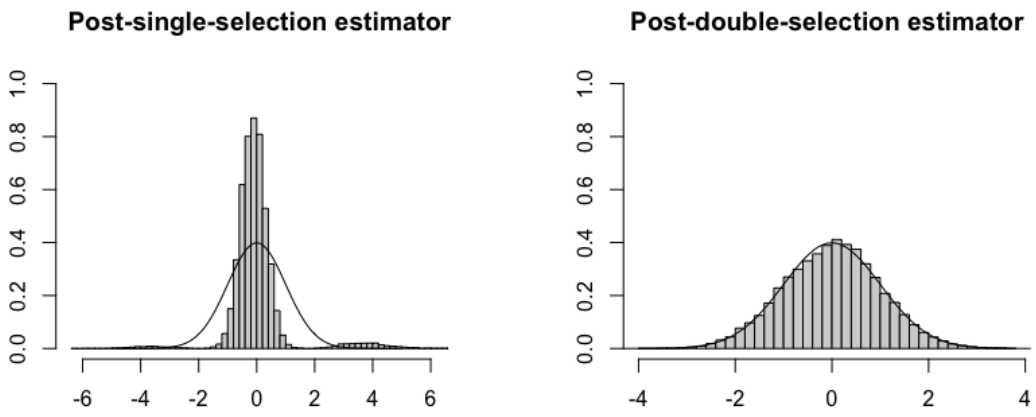

Figure 2: The empirical distributions (densities) of the studentized estimators.
The lines indicate the density curve for the standard Gaussian.

4.1 Framework

The following will be the framework on which the main result holds. The observed data $\mathscr{D}_n = (Y, D, \mathbf{X})$ are generated by P, and consist of outcome $Y \in \mathbb{R}^n$, a treatment $D \in \mathbb{R}^n$ and covariates $\mathbf{X} \in \mathbb{R}^{n \times p}$. X_j, $j \in \{1, ..., p\}$ is a confounding factor whenever $\theta_g \neq 0$ and $\theta_m \neq 0$. P satisfies conditions SE and SM, as described in section 2.3. Furthermore, P satisfies the following structural equations.

$$X_j = f_j(X_{\mathbf{PA}_j}) + \eta_j, j = 1, ..., p; \qquad \mathbb{E}[\eta_j | X_k] = 0, k \in \mathbf{PA}_j \tag{24}$$

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

$$D = \mathbf{X}\theta_m + \nu; \qquad \mathbb{E}[\nu | X_j] = 0, \forall j \in \{1, ..., p\} \tag{26}$$

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.

$$\hat{\beta}_g := argmin_{\beta : supp(\beta) \subset \hat{\mathbf{S}}_Y} \|Y - \mathbf{X}\beta\|_2^2 \tag{27}$$

$$\hat{\beta}_m := argmin_{\beta : supp(\beta) \subset \hat{\mathbf{S}}_D} \|D - \mathbf{X}\beta\|_2^2 \tag{28}$$

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

$$\mathbb{E}\left[\|X\theta_g - \mathbf{X}\hat{\beta}_g\|_2^2\right]^{1/2} = O\left(\sqrt{\frac{s \log p}{n}}\right), \quad \mathbb{E}\left[\|X\theta_m - \mathbf{X}\hat{\beta}_m\|_2^2\right]^{1/2} = O\left(\sqrt{\frac{s \log p}{n}}\right) \tag{29}$$

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

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

$$(\hat{\alpha}, \hat{\theta}_g) := argmin_{\alpha \in \mathbb{R}, \theta_g : supp(\theta_g) \subset \hat{\mathbf{S}}_Y \cup \hat{\mathbf{S}}_D} \|Y - D\alpha - \mathbf{X}\theta_g\|_2^2. \tag{30}$$

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

$$\hat{\sigma}_\alpha^{-1} \sqrt{n}(\hat{\alpha} - \alpha) \rightarrow_d N(0,1), \tag{31}$$

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 simulations follow a dgp based on Belloni *et al.* (2011) section 4.2. The simulation design is high-dimensional and has high confounding, while allowing for a sparse approximation, such as is described in condition SM, in section 2.3. This design is thus appropriate to test the qualities of AdJ. The following structural equations describe the dgp the simulation is based on.

$$Y = D\alpha + \mathbf{X}\theta_g + \sigma_Y \zeta, \qquad \zeta \sim N(0,1) \tag{32}$$

$$D = \mathbf{X}\theta_m + \sigma_D \nu, \qquad \nu \sim N(0,1) \tag{33}$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}, D \in R^n, Y \in R^n$. 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 following specifications are implemented. The treatment effect α is set to 0.5. The coefficients θ_m, θ_g are defined so that a chosen R^2 can be fixed for each of the structural equations (33), denoted Structure R^2, and (32), denoted First Stage R^2.

$$\theta_{m,j} = c_D \beta_j \tag{34}$$

$$\theta_{g,j} = c_Y \beta_j, \tag{35}$$

where $\beta_j = (1/j)^2$ for $j = 1, ..., p$ and tuning parameters c_D, c_Y.

I apply two different simulation designs. Design A uses homoscedastic error terms, with $\sigma_Y = \sigma_D = 1$. Design B uses heteroscedastic error terms, with $\sigma_Y = \sqrt{\frac{(1+D\alpha+\mathbf{X}\beta)^2}{\mathbb{E}[1+D\alpha+\mathbf{X}\beta]^2}}$ and $\sigma_D = \sqrt{\frac{(1+\mathbf{X}\beta)^2}{\mathbb{E}[1+\mathbf{X}\beta]^2}}$. Tuning parameters c_D and c_Y are chosen to determine the desired R^2 of the structural equations for design A, but are then used for the simulation of B without adjusting.

Note that the dgp is defined such that there is no exact sparse model for determining either Y or D. In the corresponding DAG two arrows emanate from each covariate X_j and

point to the treatment, D, and outcome variable, Y, respectively. Additionally, there are double-headed arrows between every pair of covariates X_j and $X_k, j \neq k$, which indicate the correlation between them. This convoluted causal structure represents the high degree of confounding in the data. Hence, actually knowing the DAG would be of little use to choose an adjustment set, since every covariate would be selected - they are all confounders. This selection, however, would not reduce the dimensionality of the covariate set, which of course defeats our purpose, since then regression methods could not be applied for the estimation of the treatment effect.

Instead, the selection of an adjustment set can be done based on the strength of the coefficients on the different covariates - essentially what model selection methods such as Lasso regression or AdJ do. In this particular simulation design, the coefficients decay fast; it therefore allows for a sparse, approximate model to be selected based on the strength of the coefficients.

I conduct 10000 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 regresses $(Y - \mathbf{X}\theta)g$ on D by OLS. This gives an unfeasible benchmark estimator, since it uses information about the true dgp, which in practice is unknown. The remaining three methods are feasible.

The second method uses post-double selection with Lasso regression (Tibshirani, 1996), as it is proposed by Belloni *et al.* (2014b). Lasso regression is run of Y on $\mathbf{X}$ to select a subset of covariates that are good predictors for Y; and of D on $\mathbf{X}$ to select a subset of good predictors for D. Subsequently the union of these two subsets is taken, which represents the adjustment set selected by Lasso.

To choose the penalty, I use the following equation from Belloni & Chernozhukov (2011).

$$\lambda := 2c\sigma\sqrt{n}\Phi^{-1}(1 - \alpha_\lambda/2p), \tag{36}$$

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 a decent benchmark for feasible methods.

The third and fourth methods are testing-based methods, namely post-double selection with AdJ, each using a different conditional independence tests. The third method uses the Fisher-Z transformation to test for vanishing conditional correlation, as described in section 2.3.

The fourth method 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). To check whether two random variables A, B are independent, $A \perp\!\!\!\perp B|C$, in a first step A and B are separately regressed on C and in a second step the residuals of each regression are tested for vanishing covariance. This method performs well on ANM. Moreover, this test, unlike the Fisher-Z test, is not limited to Gaussianity.

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	0.859	0	0.387	0
Post-double-sel. w/ Lasso	1.145	0.002	1.114	0
Post-double-sel. w/ AdJ (Fisher-Z)	1.129	0.005	1.072	0.001
Post-double-sel. w/ AdJ (dCov)	1.156	0.004	1.081	0
B. Heteroscedastic design				
Oracle	1.728	0.156	1.509	0.095
Post-double-sel. w/ Lasso	1.82	0.184	2.38	0.424
Post-double-sel. w/ AdJ (Fisher-Z)	1.891	0.195	2.527	0.401
Post-double-sel. w/ AdJ (dCov)	1.898	0.191	2.561	0.402

The significance level for the testing-based methods is set to 0.01, which is chosen following the advice of Kalisch & Buhlmann (2007) for conditional independence tests in PC. For each AdJ-method the adjustment set is selected by taking the union of the two resulting sets of covariates when running AdJ on Y and on D separately.

For each of the four methods, the treatment effect α is finally estimated by running OLS of Y on D and the respective adjustment set for each feasible method. Inference on α is conducted by heteroscedasticity-robust OLS.

The simulation is conducted for both simulation designs and for different combinations of Structure R^2 and First Stage R^2. The Structure R^2 is fixed at 0.8, while the First Stage R^2 is set at two values, 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 also apparent that higher First Stage R^2, as well as homoscedastic errors lead to a lower RMSE, hence, to a more precise estimator. The Rejection Rate is also always lowest for the Oracle and is equally influenced by structural R^2 and design type as the RMSE.

The results for the three post-double selection methods are in the same ball-park, with minor differences depending on the simulation specifications. Based on the RMSE, AdJ seems to work slightly better in the homoscedastic design, while Lasso does so in the heteroscedastic one. On the other hand, the Rejection Rate is in most cases higher when using AdJ. The results also show that there is little difference between using Fisher-Z or using dCov for conditional independence test in AdJ. The results show in most cases a slightly better RMSE for AdJ with Fisher-Z. Arguably, these results are likely to be reversed when errors are defined as non-Gaussian instead, since dCov is better suited to handle this setting.

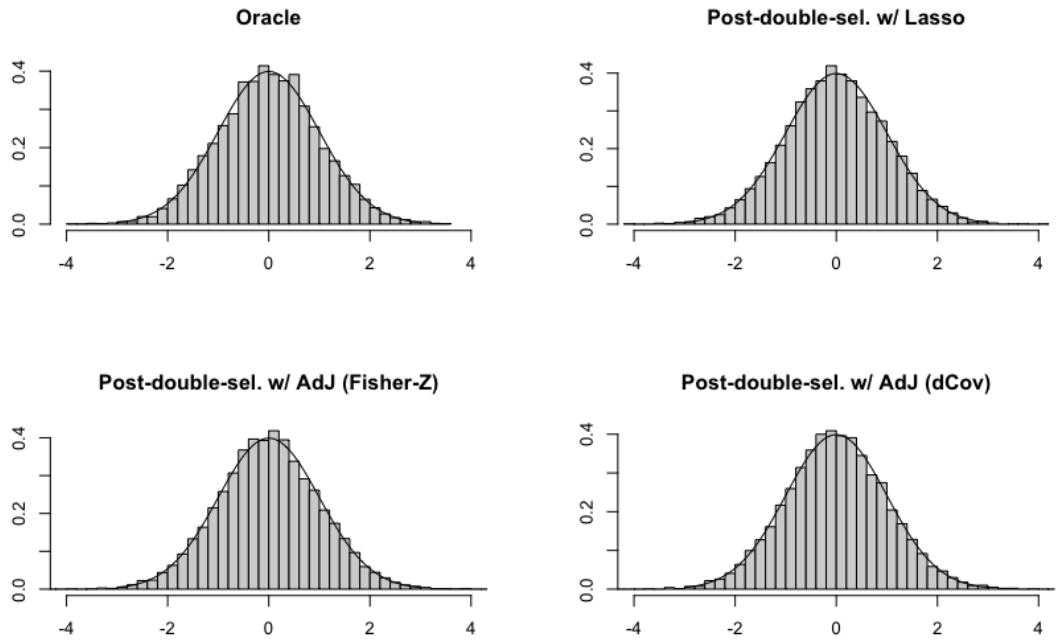

Figure 3: The empirical distributions (densities) of the studentized estimators for simulation design A with First Stage $R^2 = 0.8$, Structure $R^2 = 0.8$. The lines indicate the density curve for the standard Gaussian.

Figure 3 shows the empirical studentized distribution of the simulated estimators for Design A with First Stage $R^2 = 0.8$. It is apparent from the figure that all four estimators are root-n consistent. In fact, their convergence behaviour seems to be quite similar to each other. This serves as an empirical confirmation of the main result in section 4.2.

Overall, the results of this simulation study show that the estimator resulting from post-double selection with AdJ is well-behaved in this data setting, which features high-dimensional and high confounded data. Further, its performance is comparable with that of the feasible benchmark, the post-double selection with Lasso. Arguably, this makes it a fit for broader usage.

6 Empirical Example: Estimating the effect of Institutions on Output

In this section I will exemplify the use of post-double selection with AdJ on a real data set. Namely, I will work through the causal analysis in Acemoglu, Johnson and Robinson (2001), henceforth referred to as AJR. I choose this study for exemplification of AdJ for several

reasons; first it uses a high-dimensional data set with a clearly defines treatment and outcome. It is also widely known and the methodology is relatively simple.

AJR argue that better institutions are more able to create favourable conditions for economic development, for instance by effectively enforcing property rights, and that this in turn leads to higher income. The aim of the study is to estimate this putative causal effect which institutions have on income. The difficulty of analyzing this mechanism comes from possible simultaneous causality: Richer countries can afford, and might therefore be more likely, to choose better institutions. Additionally, there might be confounding factors that lead to both, better institutions and higher output.

In order to overcome these issues, AJR use an instrumental variable regression. The authors take into consideration data from ex-colonial states. As exogenous variation in institutions they use settler mortality in colonial times. Their theory states that in places with high settler mortality European, colonizers implemented rather extractive colonisation policies, which did not need property-rights enforcement to protect from expropriation. On the other hand, in places where European migrants could settle more easily, better institutions were erected to protect their property. These colonial institutions persisted after the states gained independence, so that their influence persisted. Furthermore, they argue that settler morality in colonial times is unlikely to influence GDP directly.

However, it is possible that there exist confounding factors between settler mortality and output, which would be problematic since it would compromise the validity of the instrument. To counter this, 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 indicating the continent is sufficient for possible confounding.

Note that AJR make assumptions on the data to back their claim of sufficiency of the adjustment set, which in turn grants validity to the instrument. Such assumptions, however, can be debated so it might be good to circumvent them if possible. Applying post double-selection with AdJ allows to relax such assumptions formulated by researchers. For, which controls among the available geographical factors will be selected, depends now only on the given data. Arguably, this impartiality awards to the results a higher degree of objectivity and possibly also more robustness.

I use the same data as AJR, which has 64 country-observations and includes the following characteristics. The treatment strength of institutions is measured by the proxy $PropertyProtection$, an index for extent to which property rights are enforced against expropriation. The instrument, $SettlerMortality$, is the logarithmized mortality rates of settlers. The outcome, GDP, is the logarithmized GDP per capital. the controls, $GeographicalFactors$, consist of a set that include four dummies for Africa, Asia, North America and South America and a control for the latitude. I follow the approach taken by Belloni *et al.* (2014a) and add to the set of controls a cubic spline for latitude with nodes at 0.08, 0.16 and 0.24. The result is a control set with dimension of 16. Different subsets of this set will be used as adjustment sets

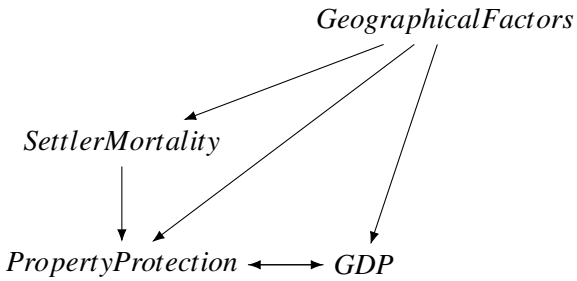

Figure 4: The bidirected arrow indicates that the direction of the causal effect might go both ways. *SettlerMortality* is a valid instrument for the effect of *PropertyProtection* on *GDP*, given we control for putatively confounding *GeographicalFactors*.

for estimation of the treatment effect. Figure 4 depicts the described variables in the DAG which describes the causal relations as they are assumed in AJR.

After the methodology by AJR, 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. In the second stage *GDP* is regressed on the predicted *PropertyProtection* and the selected adjustment set.

I apply three different methods to select the adjustment set. The three methods are denominated as Latitude, All controls and Post-double selection. Latitude refers to the estimator including only the control for latitude in the adjustment set. For All controls, all 16 controls are used. For Post-double selection, AdJ is run for each of *GDP*, *SettlerMortality* and *PropertyProtection* with respect to *GeographicalFactors* and the union of the resulting controls is selected as adjustment set. The resulting adjustment set consists of the Africa dummy and the variable $(\text{latitude} - 0.24)_+$, a component of the cubic spline[2]. Optimally, this last adjustment set is able to capture efficiently, using the fewest possible controls, any possibly lurking confounders between any pair of those three variables. The IV estimator, that is, the coefficient on predicted *PropertyProtection*, and the corresponding standard error are reported for the three methods in Table 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 effect on income. The All-controls-estimator, on the other hand, is not statistically significant, which suggests that it is inefficient. Both, the first and second stage show rather large standard errors.

The Post-double selection-estimator is significant only at the 0.05-level, 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 post-double selection using Lasso for model selection.

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

The results come to show that the model selected with post-double selection represents a trade-off in variance between the first two models. This indicates the usefulness high-dimensional selection methods for causal inference can have for practitioners, either by themselves or as an addition to sensitivity analyses. Such methods allow to relax any assumptions on the underlying causal structure, which determine whether a variable is a confounder or not. The decision is instead dependent on the data.

This empirical example also illustrates how the structural assumptions made by researchers can be nicely accommodated within the post-double framework. While some assumptions are necessary - at the very least the definition of the baseline set of controls is - using post-double selection allows the data decide which controls to use. This is, I would argue, a more rigorous approach to covariate selection than merely relying on expert judgement. Therefore, in my view its findings should therefore be granted greater credibility.

Table 2: Effect of Institutions on Output

	Latitude	All controls	Post-double-selection
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

The aim of this paper is to establish, that graphical methods are a useful addition to the toolkit of an empirical researcher. It shows its applicability by means of theoretical results and practical examples.

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. However, in contrast to structure-learning methods, AdJ's goal is not to learn about the underlying causal structure *per se*. It is limited to finding the set of parents of the target variable, which usually requires less causal information necessary than the entire DAG. This turns out sufficient for evaluating treatment effects. AdJ exploits this property.

Further, the paper shows that AdJ can handle high-dimensional data, by being able to conduct imperfect model selection. The algorithm is aimed at finding the most relevant causal relations, to find an approximate, sparse model.

I provide partial results to show that AdJ satisfies the required properties so that it is able to conduct post-double selection. The results are insofar partial, as the proof for theorem 2, which is necessary for the main result to hold, is incomplete. To complete this proof is the current focus of the still ongoing project. To illustrate the use and the mentioned properties,

post-double selection with AdJ is applied on simulated data. AdJ's application in practice is exemplified by an empirical example using data from Acemoglu *et al.* (2001).

Further research is necessary to extend my results to other, more general dgps. Arguably, different conditional independence tests can be taken into account to relax the assumptions on the data of linearity and Gaussianity. This paper is limited to showing the basic use of AdJ for approximated model selection. The statistical properties of AdJ can be investigated in more detail, for instance with respect to the efficiency of the resulting estimator. Moreover, possible overfitting of the model selected with AdJ could also be investigated in detail. Nonetheless, in the here presented simulations and the empirical example, this latter point not been an issue.

8 Appendix

8.1 Proof for Theorem 1

Proof. By Faithfulness the decision rule in Algorithm 1 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 Sketch Proof of Theorem 2

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

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

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

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

The predictions of Z are then given by

$$X_{\hat{\mathbf{S}}_Z} \hat{\beta}_{\hat{\mathbf{S}}_Z} = X_{\hat{\mathbf{S}}_Z} \theta_{\hat{\mathbf{S}}_Z} + X_{\hat{\mathbf{S}}_Z} (X'_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z})^{-1} X'_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z^c} \theta_{\hat{\mathbf{S}}_Z^c}. \tag{40}$$

And the expected approximation error is given by

$$\mathbb{E}\left[X\theta - X_{\hat{\mathbf{S}}_Z} \hat{\beta}_{\hat{\mathbf{S}}_Z} \right]. \tag{41}$$

Since the underlying true 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

$$\mathbb{E}\left[M_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z^c}\right] \theta_{\hat{\mathbf{S}}_Z^c}. \tag{42}$$

Take the squared $L2$ norm of the expected approximation error:

$$\mathbb{E}\left[\|X\theta - X_{\hat{\mathbf{S}}_Z}\hat{\beta}_{\hat{\mathbf{S}}_Z}\|_2^2\right] \tag{43}$$

$$= \theta'_{\hat{\mathbf{S}}_Z^c}\mathbb{E}\left[X'_{\hat{\mathbf{S}}_Z^c} M_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z^c}\right] \theta_{\hat{\mathbf{S}}_Z^c}. \tag{44}$$

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

$$\mathbb{E}\left[\|X\theta - X_{\hat{\mathbf{S}}_Z}\hat{\beta}_{\hat{\mathbf{S}}_Z}\|_2^2\right] \leq \lambda_{max}\left(\mathbb{E}\left[X'_{\hat{\mathbf{S}}_Z^c} M_{\hat{\mathbf{S}}_Z} X_{\hat{\mathbf{S}}_Z^c}\right]\right) \|\theta_{\hat{\mathbf{S}}_Z^c}\|_2^2. \tag{45}$$

We need that the right-hand side is of order $O(s\log p/n)$. By BE we have that the first term is bounded by some constant, since $M_{\hat{\mathbf{S}}_Z}$ is idempotent. So, we need that

$$\|\theta_{\hat{\mathbf{S}}_Z^c}\|_2^2 = O\left(\frac{s\log p}{n}\right). \tag{46}$$

We can write the squared $L2$ norm of the approximation error, r, as described in SM in matrix notation similar to (44). By SM we have that

$$\theta'_{\mathbf{S}^c}\mathbb{E}\left[X'_{\mathbf{S}^c} M_{\hat{\mathbf{S}}} X_{\mathbf{S}^c}\right] \theta_{\mathbf{S}^c} = O\left(\frac{s}{n}\right). \tag{47}$$

Similarly to the reasoning before, by symmetry and BE it follows that

$$\|\theta_{\mathbf{S}^c}\|_2^2 = O\left(\frac{s}{n}\right). \tag{48}$$

Equation (48) states that the desired property is fulfilled for the true sparse approximation model with covariates indexed by S. However, we are unlikely to get the sparse model exactly right through model selection. Although, the more observations we have the lower is the probability of errors. In the case of AdJ an error happens either when a true independence which remains undetected, or an independence falsely asserted. Hence, $\hat{\mathbf{S}}_Z$ approximates $\mathbf{S}$ with a certain rate and therefore does $\|\theta_{\hat{\mathbf{S}}_Z^c}\|_2^2$ approximate $\|\theta_{\mathbf{S}^c}\|_2^2$ with that same rate. It remains to be proven that it converges quick enough such that equation (46) holds. Kalisch & Buhlmann (2007) describe the convergence rate of the PC, which can be useful for the remaining proof.

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. 2011. Inference on Treatment Effects After Selection Amongst High-Dimensional Controls. *arXiv:1201.0224*, 12.

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*.