Double Machine Learning and Automated Model Selection: A Cautionary Tale

Paul Hünermund[†] Beyers Louw[‡] Itamar Caspi[*]

[†] *Copenhagen Business School, Kilevej 14A, Frederiksberg, 2000, DK.*
E-mail: phu.si@cbs.dk
[‡] *Maastricht University, Tongersestraat 53, 6211 LM Maastricht, NL.*
E-mail: jb.louw@maastrichtuniversity.nl
[*] *Bank of Israel, P.O.Box 780, 91007, Jerusalem, IL*
E-mail: itamar.caspi@boi.org.il

This version: May 25, 2023
First version: August 26, 2021

Summary Double machine learning (DML) has become an increasingly popular tool for automated variable selection in high-dimensional settings. Even though the ability to deal with a large number of potential covariates can render selection-on-observables assumptions more plausible, there is at the same time a growing risk that endogenous variables are included, which would lead to the violation of conditional independence. This paper demonstrates that DML is very sensitive to the inclusion of only a few "bad controls" in the covariate space. The resulting bias varies with the nature of the theoretical causal model, which raises concerns about the feasibility of selecting control variables in a data-driven way.

Keywords: *Double/Debiased Machine Learning, Directed Acyclic Graphs, Bad Controls, Backdoor Adjustment, Collider Bias, Causal Hierarchy*

"No causes in, no causes out."
— Nancy Cartwright

1. INTRODUCTION

Machine learning approaches for selecting suitable control variables to establish causal identification in high-dimensional settings are gaining increasing attention (Belloni et al., 2014b; Chernozhukov et al., 2018). Besides the evident benefits of automation for the analysis of high-dimensional data, this rising popularity can be explained by two specific advantages that applied researchers attribute to these methods. First, a mostly data-driven, automated procedure of model selection allows to systematize the research process and make it more transparent (Athey, 2019). And second, the ability to consider a large number of covariates — possibly larger than the sample size — could render selection-on-observables types of identification assumptions more plausible (Belloni et al., 2014a). For these reasons, automated variable selection has seen several recent applications in economics (Jones et al., 2019; Chang, 2020; Angrist and Frandsen, 2022), finance (Feng et al., 2020), political science (Dutt and Tsetlin, 2018; Blackwell and Olson, 2021), and organizational studies (Vanneste and Gulati, 2021), as well as as the introduction of

dedicated open source software libraries in *R* and *Python* (Chernozhukov et al., 2019; Bach et al., 2022).[1]

Double/debiased machine learning (DML) is a method developed to use regularized regression techniques, such as LASSO (Tibshirani, 1996) or l_2-boosting (Bühlmann and Yu, 2003), for variable selection in a high-dimensional causal inference setting (Belloni et al., 2014a). Compared to standard regularization on a single outcome equation, it seeks variables that are highly correlated with both treatment *and* outcome, which immunizes the procedure against small approximation errors that inevitably arise when selecting among a large set of covariates. Consider the following system of partially linear equations

$$y = \theta_0 d + g_0(x) + u, \tag{1.1}$$

$$d = m_0(x) + v, \tag{1.2}$$

with primary interest in the causal effect θ_0 of a treatment D on outcome Y. The vector $X = (X_1, \ldots, X_p)$ consists of a set of covariates and (U, V) are two disturbances with zero conditional mean. In settings where X is high-dimensional and $g_0(\cdot)$ and $m_0(\cdot)$ are approximately linear and sparse, meaning that only a few elements of X are important for predicting the treatment and outcome, regularization can be applied to automatically select the most suitable among a large set of potential control variables.

Yet, a naïve application of regularization to equation 1.1 can lead to substantial omitted variable bias (OVB), as it only selects variables that are highly correlated with the outcome Y, but not with the treatment D. The naïve approach therefore generally does not result in a root-N consistent estimator for the structural parameter θ_0 (Chernozhukov et al., 2018). Two main solutions to this problem are proposed in the literature: (a) partialling out, and (b) double selection, which both take into account the strength of association between D and X. The former uses regularization to estimate the residuals of the outcome equation, $\rho^y = y - x'\pi_0^y$, and of the treatment equation, $\rho^d = d - x'\pi_0^d$, with π_0^y and π_0^d being the respective coefficient vectors. It then finds the causal effect of interest $\hat{\theta}$ by regressing ρ^y on ρ^d (Robinson, 1988). The latter solution first determines suitable predictors for Y, then similarly finds predictors for D, and finally regresses Y on the union of the selected controls. It can be shown that both approaches rely on doubly-robust moment conditions and are thus insensitive to approximation errors stemming from regularization (Belloni et al., 2017; Chernozhukov et al., 2018).

To causal inference scholars it is generally well known that model-free covariate selection is a theoretical impossibility — a fact which was conceptualized by Pearl and Mackenzie under the rubric of the *ladder of causation* (Pearl and Mackenzie, 2018) and recently proven by Bareinboim et al. (2022). From this vantage point, the DML research program appears puzzling. If the starting point is a standard textbook regression equation, in which each variable X_k is exogenous and the number of parameters p is allowed to grow large, then variable selection is obviously feasible. Identification is achieved by assumption and the only task left for the machine learning algorithm is to pick the covariates with non-zero coefficients. But this ignores the problem that in reality not all covariates will be suitable controls.

The key identification assumption within the DML framework is *ignorability* (Imbens, 2004; Belloni et al., 2014b). Given the high-dimensional vector of control variables, treat-

[1] See, for example, the vignette of the *R*-package *hdm*, which presents automated variable selection as a main application for illustrating the usefulness of double machine learning approaches.

ment status is required to be conditionally independent of potential outcomes

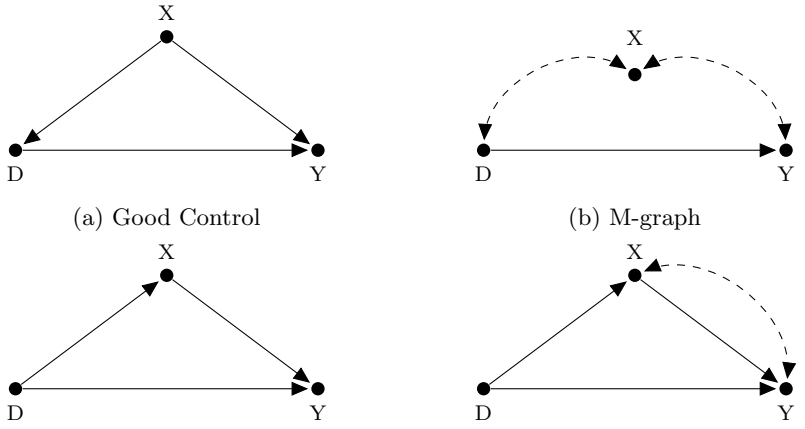
$$Y_{D=d} \perp\!\!\!\perp D|X, \tag{1.3}$$

with $Y_{D=d}$ denoting the potential outcome of Y given treatment status $D = d$. This assumption can easily be violated, if X includes variables that are not fully exogenous. In the following, we explore the consequences of violations of ignorability due to the presence of *bad controls* in the conditioning set of the DML algorithm (Angrist and Pischke, 2009; Cinelli et al., 2022). We focus on the LASSO case, which has received most attention so far (Jones et al., 2019; Vanneste and Gulati, 2021; Angrist and Frandsen, 2022; Knaus, 2021), presumably because of its appealing combination of interpretability and accuracy. However, as we will show, our arguments apply more broadly, also to the use of other machine learning algorithms for automated variable selection in a causal inference setting.

In a first step, we make precise the notion of bad controls in regression analyses by building on the *backdoor criterion* from the graphical causal models literature (Pearl, 1995; Cinelli et al., 2022). We then show in simulations that DML is very sensitive to minor violations of the ignorability assumption. Depending on the exact source of endogeneity, the advantage of DML over naïve LASSO — which was one of the main motivations for developing the method — vanishes completely. This is because bad controls, although they do not necessarily exert a causal influence, are often highly correlated with the treatment or the outcome (since they are related to unobservables that affect D or Y). Therefore, bad controls are very likely to be picked by DML, which has quantitative implications even if only a few endogenous variables are present in the conditioning set. We demonstrate this in an application of DML to the estimation of the gender wage gap using the data provided by Blau and Kahn (2017). We find that the estimation results obtained by the original study differ in non-negligible ways compared to when marital status, which the literature identifies as being likely endogenous with respect to women's labor-force decisions, is included in the covariate space.

Our study is related to a growing literature studying the performance of DML under various practically relevant data generating processes, of which most work has been focused on the omitted variable bias case. Wüthrich and Zhu (2021) show that double selection LASSO can exhibit substantial OVB as a result of variable under-selection in finite samples, even in favorable settings such as — most relevant for this manuscript — with uncorrelated, exogenous controls. Their findings render an application of the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ derived in Belloni et al. (2014b) potentially problematic. Moreover, Chernozhukov et al. (2022) derive sharp bounds on the OVB in the presence of unobserved confounders, which can be used to perform sensitivity analysis. Instead, we focus on the case with endogenous, bad controls in the conditioning set.

Our results highlight significant pitfalls of automated, data-driven variable selection in high-dimensional settings. In particular, if numerous potential controls are considered, in an attempt to justify selection-on-observables, without theoretical background knowledge to guide the choice, the likelihood that some bad controls are accidentally included in the algorithm is high. Therefore, dealing with a large covariate space in an automated fashion might do little to approximate the ignorability assumption and is instead more useful to determine a suitable functional-form specification for a *small* set of covariates, e.g., by considering higher-order polynomial terms (Belloni et al., 2014b; Athey and Imbens, 2017). We show that this problem is not only prevalent for post-treatment variables or variables that are themselves considered outcome variables (Vanneste and Gulati, 2021),



(a) Good Control (b) M-graph

(c) Mediator (d) Confounded Mediator

Figure 1: Directed acyclic graphs representing different structural causal models.

so that researchers cannot rely on simple rules of thumb for variable inclusion. By contrast, each potential control requires its own careful identification argument based on domain knowledge[2], which is difficult to provide if the feature space is large and ultimately undermines the purpose of automated variable selection. We stress, however, that DML has broader applications, e.g., for the estimation of high-dimensional instrumental variable models (Belloni et al., 2017) and arbitrary do-calculus objects (Jung et al., 2021), as well as for data-splitting to reduce over-fitting. Our argument therefore specifically applies to the case when machine learning tools are used for the purpose of confounder selection.

2. PRELIMINARIES

An SCM is a 4-tuple $\langle V, U, F, P(u)\rangle$, where $V = \{V_1, \ldots, V_m\}$ is a set of endogenous variables that are determined in the model and U denotes a set of (exogenous) background factors. F is a set of functions $\{f_1, \ldots, f_m\}$ that assign values to the corresponding $V_i \in V$, such that $v_i \leftarrow f_i(pa_i, u_i)$, for $i = 1, \ldots, m$, and $PA_i \subseteq V \setminus V_i$.[3] Finally, $P(u)$ is a probability function defined over the domain of U.

Every SCM defines a directed graph $\mathcal{G} \equiv (V, E)$, where V is the set of endogenous variables, denoted as nodes (vertices) in the graph, and E is a set of edges (links) pointing from PA_i (the set of parent nodes) to V_i. An example is given by Fig. 1a, which corresponds to the SCM

$$\begin{aligned} x &\leftarrow f_1(u_1), \\ d &\leftarrow f_2(x, u_2), \\ y &\leftarrow f_3(d, x, u_3). \end{aligned} \tag{2.4}$$

[2]Early econometricians such as Tjalling Koopmans were of course well aware of this fact: "Without resort to theory [...] conclusions relevant to the guidance of economic policies cannot be drawn" (Koopmans, 1947).

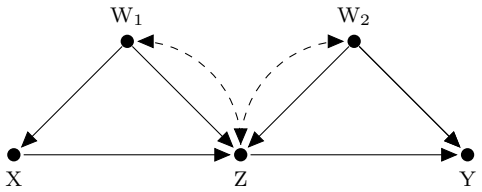
[3]The SCM literature uses assignment operators instead of equations to capture the asymmetric nature of causal relationships (Hünermund and Bareinboim, 2023).

Unobserved parents nodes induce a correlation between background factors in U. This is depicted by bidirected dashed arcs in the graph, which render the causal model *semi-Markovian* (Pearl, 2009, pp. 30). Fig. 1b depicts an example where the background factors of X and D, as well as X and Y are correlated due to the presence of common influence factors that remain unobservable to the analyst.

A sequence of edges connecting two nodes in $\mathcal{G}$ is called a *path*. Paths can be either undirected or directed (i.e., following the direction of arrowheads). Since edges correspond to stimulus-response relations between variables in the underlying SCM (Strotz and Wold, 1960), directed paths represent the direction of causal influence in the graph. Due to the notion of causality being asymmetric (Woodward, 2003; Cartwright, 2007), directed cycles (i.e., loops from a node back to itself) are excluded, to rule out that a variable can be an (instantaneous) cause of itself. This assumption renders $\mathcal{G}$ acyclic.

A semi-Markovian causal graph $\mathcal{G}$ allows to decompose the distribution of the observed variables according to the factorization: $P(v) = \sum_u \prod_i P(v_i|pa_i, u_i)P(u)$ (Pearl, 2009). The close connection between the topology of $\mathcal{G}$ and the probabilistic relationships — in particular conditional independence relations — between the variables that represent its nodes is further exemplified by the *d-separation* criterion (Pearl, 1988). Consider three disjoint sets of variables, X, Y, and Z in a DAG. These sets can either be connected via a (causal) chain, $X \to Z \to Y$, or a fork, $X \leftarrow Z \to Y$, where Z acts as a common parent of X and Y. A third possible configuration is the collider, $X \to Z \leftarrow Y$. In a chain and fork, conditioning on Z renders X and Y conditionally independent, such that $X \perp\!\!\!\perp Y|Z$. Z is then said to "d-separate" or "block the path between" X and Y. By contrast, in the collider structure, X and Y are independent from the outset, $X \perp\!\!\!\perp Y|\emptyset$, whereas conditioning on Z (or a descendant of Z; see (Pearl, 2009, def. 1.2.3), would unblock the path, such that $X \not\perp\!\!\!\perp Y|Z$.[4]

D-separation gives rise to testable implications of graphical causal models (Pearl, 2000). Consider the following DAG:



This graph implies four d-separation relations between observed variables in the model: $W_1 \perp\!\!\!\perp W_2$, $X \perp\!\!\!\perp W_2$, $X \perp\!\!\!\perp Y|W_2, Z$, and $Y \perp\!\!\!\perp W_1|W_2, Z$. They can be tested in the data with the help of a suitable conditional independence test, and if rejected, the hypothesized causal model can be discarded and refined.

Causal effects are defined in terms of interventions in the SCM, denoted by the $do(\cdot)$-operator (Haavelmo, 1943; Strotz and Wold, 1960; Pearl, 1995). For example, the intervention $do(D = d')$ in eq. 2.4 entails to delete the function $f_2(\cdot)$, which normally assigns values to D, from the model and to replace it with the constant value d'. The target is then to estimate the post-intervention distribution of the outcome variable, $P(Y =$

[4]Note that these d-separation relations hold for any distribution $P(v)$ over the variables in the model, in particular irrespective of any specific functional-form assumptions for f_i and any distributional assumptions for $P(u)$ (Hünermund and Bareinboim, 2023).

$y|do(D = d'))$, that results from this manipulation. Other quantities, such as the average causal effect (ACE) of a discrete change in treatment from d' to d'', can then be computed by taking the difference in expected values: $E(Y|do(D = d'')) - E(Y|do(D = d'))$. However, since $P(y|do(d))$ is not directly observable in non-experimental data, it first needs to be transformed into a probability object that does not contain any do-operator before estimation can proceed (Bareinboim and Pearl, 2016; Hünermund and Bareinboim, 2023). This constitutes the *identification* step in the graphical causal models literature (Koopmans, 1950; Pearl, 2009).

2.1. Backdoor Adjustment

One popular strategy to identify the ACE is to control for confounding influence factors via covariate adjustment. This strategy can be rationalized with the help of the *backdoor criterion* (Pearl, 1995).

Definition 2.1. *Given an ordered pair of treatment and outcome variables (D, Y) in a causal graph $\mathcal{G}$, a set X is backdoor admissible if it blocks (in the d-separation sense) every path between D and Y in the subgraph $\mathcal{G}_{\underline{D}}$, which is formed by deleting all edges from $\mathcal{G}$ that are emitted by D.*

Deleting edges emitted by D from $\mathcal{G}$ ensures that all directed, causal paths between D and Y are kept open. The remaining paths are non-causal and thus create a spurious correlation between the treatment and outcome.[5] Consequently, a backdoor admissible set X blocks all non-causal paths between D and Y, while leaving the causal paths intact. The post-intervention distribution is then identifiable via the adjustment formula (Pearl, 2009)

$$P(y|do(d)) = \sum_x P(y|d, x)P(x). \tag{2.5}$$

Since the right-hand side expression does not contain any do-operator, it can be estimated from observational data either by nonparametric methods, such as matching and inverse probability weighting, or, under additional functional-form assumptions, by parametric regression methods such as OLS.

However, following the d-separation criterion, correctly blocking backdoor paths via covariate adjustment can be intricate. Take Fig. 1 as an example. In 1a there is one causal path, $D \to Y$, and one backdoor path, $D \leftarrow X \to Y$ (with X being possibly vector-valued). Following the d-separation criterion, the backdoor path can be blocked by conditioning on X so that only the causal influence of D remains. By contrast, in the other depicted cases, controlling for X would induce rather than reduce bias, thus, rendering X a *bad control* in these models. In Fig. 1b, which is known under the name of *m-graph* in the epidemiology literature (Greenland, 2003), X exerts no causal influence on any variable in the graph. Still, there are unobserved confounders that result in a backdoor path, $D \dashleftarrow\dashrightarrow X \dashleftarrow\dashrightarrow Y$, which is already blocked however, since X acts as a collider on this path. At the same time, since X is a collider, conditioning on it (or any of its descendants) would unblock the path and therefore produce a spurious correlation. By contrast, X does not lie on a backdoor path in Fig. 1c, but acts as a

[5] Since these paths point into D, they are said to "enter through the backdoor".

mediator between D and Y. Controlling for X would allow to filter out the direct effect of the treatment, $D \to Y$, from its mediated portion, $D \to X \to Y$, (Imai et al., 2010). However, this direct effect is generally different from the ACE, which has to be kept in mind for interpretation of results.[6] Moreover, such an approach is risky, because if there are unobserved confounders between X and Y, as depicted in Fig. 1d, X becomes a collider on the path $D \to X \dashleftarrow\dashrightarrow Y$ and would thus lead to bias if conditioned on.[7]

3. SIMULATION RESULTS

In the following, we present a variety of simulations results to assess the magnitude of the bias introduced by including bad controls in the DML algorithm. We focus on the high-dimensional linear setting and apply double selection DML based on l_1-regularization to automatically select covariates. However, our argument is not specific to the LASSO case. In the online supplement, we present additional simulation results using l_2-boosting, which show very similar patterns.

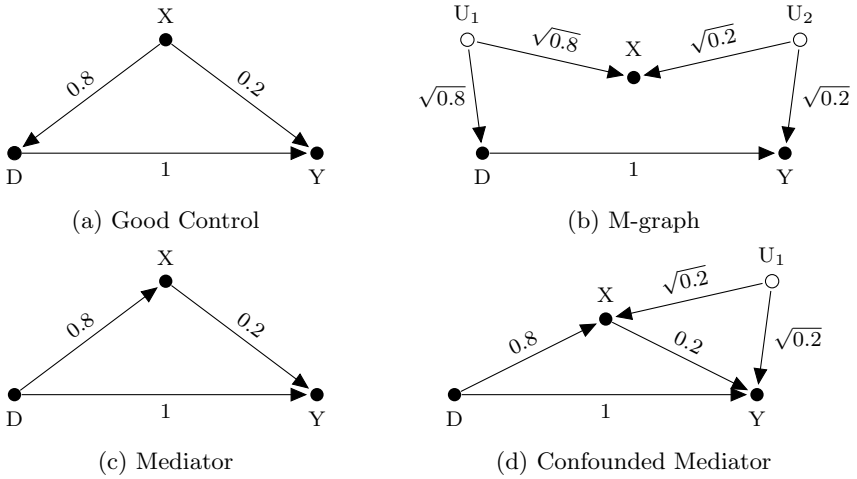
Since DML is specifically designed to spot variables that are mainly correlated with the treatment, which is the reason for its superior performance compared to naïve LASSO, for our baseline specification, we set a higher correlation between the controls and the treatment than with the outcome. We fix the sample size at $n = 1,000$ and number of covariates at $p = 100$. To introduce sparsity, only $q = 10$ out of these variables are specified as having non-zero coefficients. The treatment effect θ_0 is constant and set equal to one. All exogenous nodes (which do not receive any incoming arrows) are specified as standard normal. In the baseline, parameters are chosen in such a way that the strength, measured as the product of structural coefficients, of each path connecting the (non-zero) covariates and the treatment is equal to $b_1 = 0.8$. Similarly, the strength of paths connecting the covariates and the outcome is set to $b_2 = 0.2$ (Fig. 2 depicts the baseline parametrization in form of a path diagram with associated coefficients as edge labels, hollow circles indicate unobserved variables).

Following the double selection method, we then regress Y on X using LASSO and record the variables with estimated non-zero coefficients. We do the same for a LASSO regression of D on X. Finally, we regress Y on the union of variables in X that have been picked in the preceding two LASSO regressions, this time using standard OLS. We record the estimated coefficients for the treatment effect of interest $\hat{\theta}$ across 10,000 simulation runs. In addition, we compare double selection with the naïve (post)LASSO method, in which we repeat the previous protocol but without the second step of regressing D on X. I.e., in naïve LASSO variables are only selected once for the outcome regression, disregarding their correlation with the treatment. To summarize the estimation algorithms:

- **DML (double selection)**
  1 Regress Y on X via LASSO and record all X_k with nonzero coefficients
  2 Regress D on X via LASSO and record all $X_{k'}$ with nonzero coefficients
  3 Regress Y on the union of all X_k and $X_{k'}$

[6] Additionally, following Imai et al. (2010), identifying direct and indirect effects in a mediation setting requires the assumption of sequential ignorability, which is fulfilled in linear models with constant effects, but does not need to hold for every SCM.

[7] See Cinelli et al. (2022) for a more comprehensive discussion of bad controls in graphical causal models that goes beyond the scope of this paper.



(a) Good Control (b) M-graph

(c) Mediator (d) Confounded Mediator

Figure 2: Baseline parametrization of the simulations.

- **Naïve (post-)LASSO**

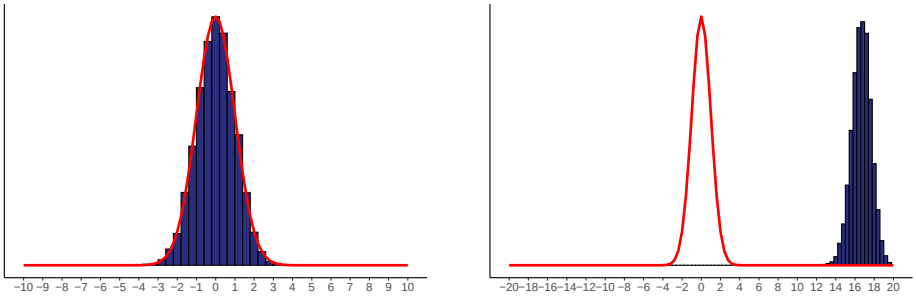
  1 Regress Y on X via LASSO and record all X_k with nonzero coeffcients
  2 Regress Y on all X_k

Fig. 3 shows simulation results using centered and studentized quantities, next to their theoretical (standard normal) distribution. In panel (a) we observe the familiar picture from Belloni et al. (2014b). DML is able to reliably filter out the good controls from irrelevant covariates, which leads to a distribution that closely matches the theoretical one. By contrast, naïve LASSO fails to pick relevant control variables that are only weakly correlated with the outcome, translating into substantial bias. However, this result reverses for the m-graph in panel (b). Here, the covariates are bad controls, due to the collider structure, and should not be included in the regression. They are nonetheless highly correlated with the treatment and thus get picked by the DML, leading to biased causal effect estimates. In fact, the advantage that DML had over naïve LASSO in (a) vanishes completely ($bias^{DML} = -0.120$, and $bias^{LASSO} = -0.119$). Interestingly, given the chosen parameterization with only a moderately high correlation between the covariates and the outcome, the naïve approach consistently selects fewer bad controls than DML. The mode of the number of controls selected across simulations is 5 for the naïve LASSO and 10 for DML.

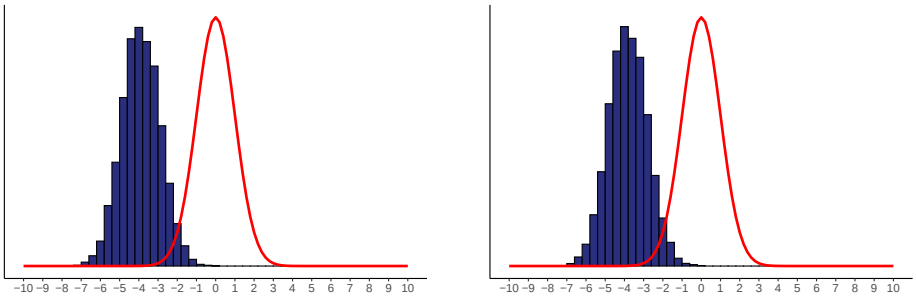
In panel (c) we investigate the mediator case. Now, the covariates are post-treatment variables, which nonetheless end up getting selected as controls by both the naïve LASSO and DML. According to the discussion in Sec. 2, this allows to consistently estimate the direct effect of the treatment. However, the researcher needs to keep this change of target parameter in mind for interpretation, since both naïve LASSO and DML are unable to consistently estimate the total effect of treatment. Moreover, once we introduce a confounded mediator in panel (d), both DML and naïve LASSO perform equally poorly. The direct effect cannot be consistently estimated in this model, as neither controlling for the mediators nor leaving them out would be sufficient for identification. The total effect

Double Machine Learning | Naïve Lasso

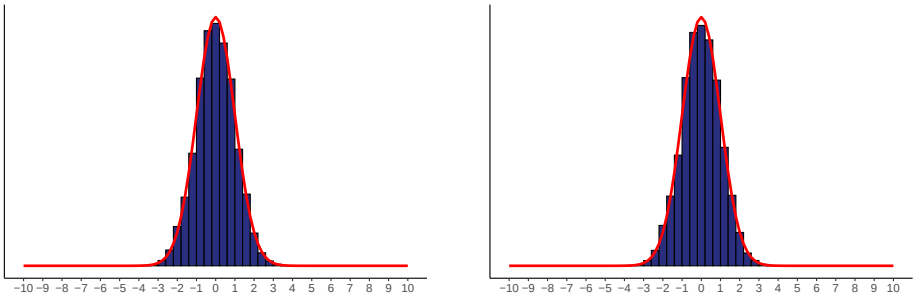
(a) Good Control



(b) M-graph



(c) Mediator (direct effect)



(d) Confounded Mediator (direct effect)

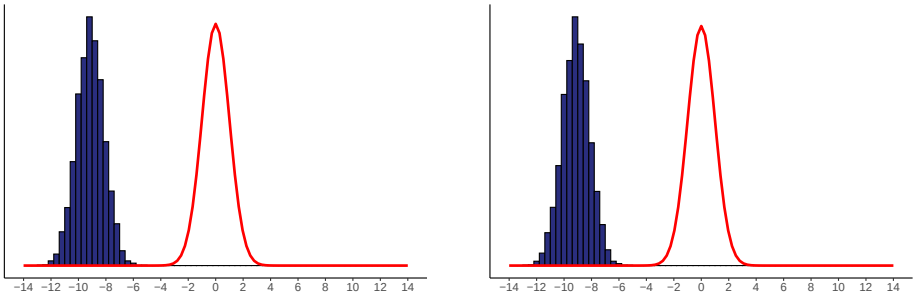


Figure 3: Performance of DML compared to naïve LASSO for different causal models.

Table 1: Bias obtained from DML under various parameter constellations ($\theta_0 = 1$).

$(b_1, b_2) =$	$(0.8, 0.2)$	$(0.6, 0.4)$	$(0.5, 0.5)$	$(0.4, 0.6)$	$(0.2, 0.8)$
Good Control	0.000	0.000	0.000	0.000	0.000
M-graph	-0.120	-0.172	-0.179	-0.174	-0.126
Mediator	-0.001	-0.001	-0.001	0.000	0.000
Confounded Mediator	-0.534	-0.480	-0.417	-0.343	-0.178
$q =$	1	5	10	20	50
Good Control	0.000	0.000	0.000	0.000	0.000
M-graph	-0.054	-0.105	-0.120	-0.128	-0.134
Mediator	0.000	-0.001	-0.001	-0.001	-0.001
Confounded Mediator	-0.134	-0.401	-0.534	-0.641	-0.728

of treatment is likewise not estimable via DML (but would be by a simple regression of Y on D).

Table 1 depicts the bias obtained from DML for varying parameter constellations. In the top panel, we study performance depending on whether there is a higher strength of association between the covariates and the treatment or the outcome. For the two bad control cases, i.e., the m-graph and confounded mediator, substantial bias arises regardless of the chosen parametrization. When taking into account the change of target parameter from the total to the direct effect, bias is low for the simple mediator model across all setups. Moreover, the DML generally performs well in the good control case, although bias becomes slightly larger when the strength of association is stronger with the outcome than the treatment (see also the $n = 100$ case in the supplemental material).

In the bottom panel of Table 1, we vary the number of covariates with non-zero coefficients q (with $b_1 = 0.8, b_2 = 0.2$, as before), while the total number of variables considered in the conditioning set remains fixed at $p = 100$.[8] Interestingly, a noticeable bias (around 5 percent for the m-graph and 13 percent for the confounded mediator) arises already with one bad control out of a hundred, and increases monotonically in q. The bias for the direct effect remains low for the simple mediator model.

In the online supplement we present additional simulations with varying p and n. In particular, we explore the case of $n = 100$ since DML is often proposed as a technique for dealing with a large number of predictors in relatively small data sets ($p \gg n$). We find results that are in line with the ones presented here in the main text.

4. APPLICATION

For an application to real-world data, we make use of the *Panel Study of Income Dynamics* (PSID) microdata provided by Blau and Kahn (2017). They estimate the extent of the gender wage gap in six waves of the PSID between 1981 and 2011. For their full specification, they employ a rich set of 50 control variables (as described in Section IV of

[8]In unreported analyses, we find similar results if bad controls are mixed with good controls instead of irrelevant (zero-coefficient) ones. The two cases are conceptually similar since the DML either picks the good or leaves out the irrelevant controls, resulting in a zero bias baseline, which then gets distorted by the selected bad controls.

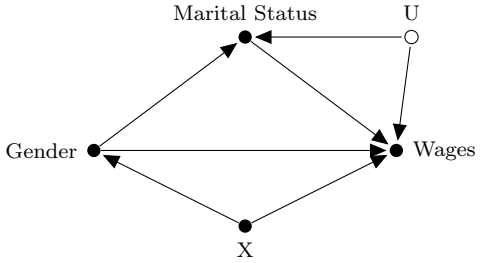


Figure 4: Causal diagram for the gender wage gap study in Blau and Kahn (2017).

their online appendix), including individual-level information on education, experience, race, occupation, unionization, as well as regional and industry characteristics. However, Blau and Kahn deliberately decide to exclude marital status and number of children from their regressions, because these variables "are likely to be endogenous with respect to women's labor-force decisions" (p. 797). Although the source of this endogeneity is not further discussed, we find it plausible that marital status acts as a confounded mediator, since it is likely influenced by the same unobserved background factors that also affect wages (see Fig. 4).

From the PSID data, we can infer a woman's marital status based on whether she is recorded as "legally married wife" in her relation to the household head (men are by default indicated as household heads). Our goal is to test the sensitivity of the estimated (adjusted) gender wage gap to the inclusion of this potentially bad control. As a benchmark, we regress log wages on a female dummy and the original set of controls for each wave separately. We then employ DML using the double selection method, which allows us to include all interactions of the control variables up to degree 2. In a last step, we add marital status and its interactions to the model matrix in the DML.

Results are shown in Table 2. The estimated gender wage gaps in the OLS specifications range from $(1 - \exp(-0.249)) \approx 22$ percentage points in 1981 to approximately 13.5 p.p. in 2011. Most of the convergence between male and female wages happens in the 1980s, which coincides with the results in Blau and Kahn (2017). Although the DML relies on a much larger set of covariates, the results are very similar to OLS. We find greater discrepancies, however, when marital status is included in the feature space. Across all six waves, marital status (as well as several interactions) ends up getting picked as control by the double selection DML. This has non-negligible impact on the estimated gender wage gaps, which are 10.6% larger on average, in absolute terms, compared to the benchmark OLS. Under the assumption that marital status is a confounded mediator, larger gaps might be the result of a negative correlation between wages and the decision to get married, induced by unobservables. The respective path gets activated when marital status, as a collider, is conditioned on. Thus, the example demonstrates how having only one endogenous control within a large covariate space, paired with a flexible DML approach, can substantially affect the quantitative conclusions drawn from a study.[9]

[9]We find even larger differences if marital status is included as a single regressor, without interacting it with other covariates; see Table S4 in the online supplemental material.

Table 2: Effect of gender on log wages using PSID data from Blau and Kahn (2017) (standard errors in parentheses).

Wave =	1981	1990	1999	2007	2009	2011
OLS	-0.249	-0.137	-0.158	-0.168	-0.157	-0.145
	(0.016)	(0.014)	(0.016)	(0.015)	(0.015)	(0.016)
DML	-0.268	-0.139	-0.158	-0.164	-0.157	-0.136
	(0.017)	(0.015)	(0.016)	(0.016)	(0.016)	(0.017)
DML incl.	-0.270	-0.154	-0.173	-0.190	-0.179	-0.163
marital status	(0.022)	(0.019)	(0.020)	(0.019)	(0.020)	(0.021)

5. DISCUSSION

In this paper, we demonstrate the sensitivity of automated confounder selection using double machine learning approaches to the inclusion of bad controls in the conditioning set. In our simulations, only when covariates are strictly exogenous, DML shows superior performance to naïve LASSO. In all other cases it performs equally poorly or worse. Furthermore, our empirical application illustrates that a non-negligible bias can already occur with a small number of endogenous variables in an otherwise much larger covariate space.

These results highlight why it may be problematic to use machine learning techniques for the automatic selection of control variables in regression settings. While the ability to deal with a large set of potential controls in an automated fashion can add to the plausibility of selection-on-observable assumptions, there is an increasing chance that bad controls might be included unintentionally if the covariate space grows large. Automated approaches thus turn out to be a double-edged sword, in particular if the number of control variables becomes so large that the researcher is unable to provide a sufficient theoretical discussion for each of them. We show that simple rules of thumb, such as restricting the conditioning set to only pre-treatment variables, do not offer adequate safeguards against this problem. Indeed, as Figure 1b shows, our results are not limited to post-treatment variables. The intricacies of the backdoor criterion (recall, e.g., the implications of subtle differences between Figures 1c and 1d) imply that a vague intuition, without the the guidance of a causal model, will likely be insufficient to ensure causal identification.

Because DML already assumes unconfounded covariates (Chernozhukov et al., 2018, sec. 5), using its ability to handle a large feature space in order to justify unconfoundedness, ultimately leads to a circular argument. As long as causal inference is the goal, the analyst needs to provide a theoretical justification for the exogeneity of each of the considered control variables individually, which echoes Cartwright (1989)'s familiar adage: "no causes in, no causes out." Since this is difficult to achieve in high-dimensional settings, from a practical standpoint, smaller models that focus only on the most relevant covariates for a given context might actually be preferable.

For the purpose of automated model selection, causal discovery algorithms from the artificial intelligence literature could represent a viable alternative (Spirtes et al., 2000; Peters et al., 2017). These methods do not rely on unconfoundedness and clarify the possibilities for data-driven causal learning based on a minimal set of assumptions. A

key insight from this literature is that causal structures can only be learned up to a certain equivalence class from data. As a result, the ultimate justification for a particular causal model needs to come from theoretical background knowledge (Bareinboim et al., 2022). The same applies to DML, which is a highly effective tool, e.g., for selecting suitable functional specifications involving a small set of controls in a data-driven way. In big data settings with a large number of potential covariates, however, DML needs to be applied carefully to avoid bad controls and to ensure robust results.

ACKNOWLEDGEMENTS

The authors are grateful to Elias Bareinboim, Victor Chernozhukov, Jevgenij Gamper, Daniel Millimet, Judea Pearl, and seminar participants at Booking.com, Microsoft, RWTH Aachen, and Vinted for useful comments and suggestions.

REFERENCES

Angrist, J. D. and B. Frandsen (2022). Machine labor. *J Labor Econ 40*(S1), S97–S140.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Athey, S. (2019). The impact of machine learning on economics. In A. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*. Chicago, IL, USA: The University of Chicago Press.

Athey, S. and G. W. Imbens (2017). The state of applied econometrics: Causality and policy evaluation. *J Econ Perspect 31*(2), 3–32.

Bach, P., V. Chernozhukov, M. S. Kurz, and M. Spindler (2022). Doubleml - an object-oriented implementation of double machine learning in python. *J Mach Learn Res 23*(53), 1–6.

Bareinboim, E., J. D. Correa, D. Ibeling, and T. Icard (2022, February). On pearl's hierarchy and the foundations of causal inference. *Probabilistic and Causal Inference: The Works of Judea Pearl*, 507–556.

Bareinboim, E. and J. Pearl (2016). Causal inference and the data-fusion problem. *Proc Natl Acad Sci 113*, 7345–7352.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*, 233–298.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a). High-dimensional methods and inference on structural and treatment effects. *J Econ Perspect 28*(2), 29–50.

Belloni, A., V. Chernozhukov, and C. Hansen (2014b). Inference on treatment effects after selection among high-dimensional controls. *Rev Econ Stud 81*, 608–650.

Blackwell, M. and M. P. Olson (2021). Reducing model misspecification and bias in the estimation of interactions. *Polit Anal 30*(4), 495–514.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *J Econ Lit 55*, 789–865.

Bühlmann, P. and B. Yu (2003). Boosting with the l_2 loss: Regression and classification. *J Am Stat Assoc 98*, 324–339.

Cartwright, N. (1989). *Nature's Capacities and Their Measurement*. Oxford, UK: Clarendon Press.

Cartwright, N. (2007). *Hunting Causes and Using Them*. Cambridge, UK: Cambridge University Press.

Chang, N.-C. (2020). Double/debiased machine learning for difference-in-differences models. *Econom J 23*(2), 177–191.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econom J 21*, C1–C68.

Chernozhukov, V., C. Cinelli, W. Newey, A. Sharma, and V. Syrgkanis (2022). Long story short: Omitted variable bias in causal machine learning. `https://doi.org/10.48550/arXiv.2112.13398`.

Chernozhukov, V., C. Hansen, and M. Spindler (2019). *High-dimensional Metrics in R*.

Cinelli, C., A. Forney, and J. Pearl (2022). A crash course in good and bad controls. *Sociol Method Res*. forthcoming.

Dutt, P. and I. Tsetlin (2018). Income distribution and economic development: Insights from machine learning. *Econ Polit 33*(1), 1–36.

Feng, G., S. Giglio, and D. Xiu (2020). Taming the factor zoo: A test of new factors. *J Financ 75*(3), 1327–1370.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology 14*, 300–306.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica 11*, 1–12.

Hünermund, P. and E. Bareinboim (2023). Causal inference and data fusion in econometrics. *Econom J*. forthcoming.

Imai, K., L. Keele, and T. Yamamoto (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Stat Sci 25*, 51–71.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev Econ Stat 86*, 4–29.

Jones, D., D. Molitor, and J. Reif (2019). What do workplace wellness programs do? evidence from the illinois workplace wellness study. *Quart J Econ 134*(4), 1747–1791.

Jung, Y., J. Tian, and E. Bareinboim (2021). Estimating identifiable causal effects through double machine learning. *Proc AAAI Conf Artif Intell* (35).

Knaus, M. C. (2021). A double machine learning approach to estimate efffects of musical practice on student's skills. *J R Stat Soc Ser A Stat Soc 184*(1), 282–300.

Koopmans, T. C. (1947). Measurement without theory. *Rev Econ Stat 29*(3), 161–172.

Koopmans, T. C. (1950). *Cowles Foundation Monograph 10: Statistical Inference in Dynamic Economic Models*. John Wiley & Sons.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA, USA: Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*(4), 669–709.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference* (1st ed.). New York, NY, USA: Cambridge University Press.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). New York, NY, USA: Cambridge University Press.

Pearl, J. and D. Mackenzie (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Peters, J., D. Janzing, and B. Schölkopf (2017). *Elements of Causal Inference*. Cambridge, MA, USA: MIT Press.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Spirtes, P., C. N. Glymour, R. Scheines, and D. Heckerman (2000). *Causation, Prediction, and Search*. Cambridge, MA, USA: MIT Press.

Strotz, R. H. and H. O. A. Wold (1960). Recursive vs. nonrecursive systems: An attempt at synthesis (part i of a triptych on causal chain systems). *Econometrica 28*, 417–427.

Tibshirani, R. (1996). Regression shrinkage and selection via lasso. *J R Stat Soc Series B Stat Methodol 58*(1), 267–288.

Vanneste, B. S. and R. Gulati (2021). Generalized trust, external sourcing, and firm performance in economic downturns. *Organ Sci 33*(4), 1251–1699.

Woodward, J. (2003). *Making Things Happen*. Oxf Stud Philos Sci. Oxford, UK: Oxford University Press.

Wüthrich, K. and Y. Zhu (2021). Omitted variable bias of lasso-based inference methods: A finite sample analysis. *Rev Econ Stat*. forthcoming.