Using Machine Learning Methods to Support Causal Inference in Econometrics

Chapter 1
Using Machine Learning Methods to Support Causal Inference in Econometrics*

Achim Ahrens
ETH Zürich, Switzerland

Christopher Aitken and Mark E. Schaffer
Heriot-Watt University, Edinburgh, United Kingdom

Email: m.e.schaffer@hw.ac.uk**

Abstract We provide an introduction to the use of machine learning methods in econometrics and how these methods can be employed to assist in causal inference. We begin with an extended presentation of the lasso (least absolute shrinkage and selection operator) of Tibshirani (1996). We then discuss the 'Post-Double-Selection' (PDS) estimator of (Belloni et al., 2012, 2014b) and show how it uses the lasso to address the omitted confounders problem. The PDS methodology is particularly powerful for the case where the researcher has a high-dimensional set of potential control variables, and needs to strike a balance between using enough controls to eliminate the omitted variable bias but not so many as to induce overfitting. The last part of the paper discusses recent developments in the field that go beyond the PDS approach.

Keywords: Causal inference, Lasso, Machine Learning

1.1 Introduction

Over the last 40 years, economic research has evolved significantly, and the pace of this change shows no sign of abating. Particularly notable is the discipline's increasingly empirical orientation. That shift is itself a reflection of two other strands of change: high-quality (administrative and private) data are now in plentiful supply and can be accessed for the purposes of research;

* Invited paper for the International Conference of the Thailand Econometric Society, 'Behavioral Predictive Modeling in Econometrics', Chiang Mai University, Thailand, 8-10 January 2020. Our exposition of the 'rigorous lasso' here draws in part on our paper Ahrens et al. (2019). All errors are our own.

** Corresponding author

and, methodologically, the profession has embraced 'credible' design-based research that exploits variation generated by quasi-experiments and randomised trials (Angrist and Pischke, 2010). These advances have stimulated interactions with other disciplines, and based on bibliometric analysis, they appear to have increased the influence of economics (Angrist et al, 2017).

Economists have also started to incorporate into their toolkits methods developed by researchers who work at the nexus of statistics and computer science—the field of machine learning. These methods have famously been used successfully for difficult, diverse tasks such as facial recognition and accurate language translation. Many problems that we now face in the conduct of economic research are very close in spirit to those, and have been tackled with machine learning methods. For example, economic historians have gained a deeper understanding of intergenerational mobility by employing those algorithmic models to link individuals across datasets that lack clean identifiers and which are rife with measurement and transcription issues (Feigenbaum, 2016). Similarly, in political economy, related methods have been used to quantify the 'partisanship' of congressional speeches (Gentzkow, Shapiro and Taddy, 2019). Machine learning methods have even been used to demonstrate in detail the changing nature of economic research (Angrist et al, 2017).[3]

These applications demonstrate the power of machine learning methods for a particular type of problem. However, their applicability is not bounded to that domain. The models can also be employed to make the econometric techniques we rely on more credible and more robust, even when applied to design-based studies. This aspect of the frontier will be the focus of the present review.

1.1.1 What is Machine Learning?

'Machine learning' (ML) is a relatively new discipline[4], and is newer still to many applied economists, so we begin with some definitions and by setting out how it relates to the familiar, more traditional field of econometrics. A caveat: as Humpty Dumpty told Alice in *Through the Looking Glass*, 'when I use a word, it means just what I choose it to mean'. Terminology in this area is still settling down.

Machine learning constructs algorithms that can learn from the data. **Statistical learning** is a branch of statistics that was born in response to machine learning. Statistical learning encompasses models which, naturally,

[3] Previous research on that topic had relied instead on anecdotal evidence, or restricted its attention to a subset of the literature that is not representative because it is not feasible to manually classify the full corpus of work (Hammermesh, 2013 and Backhouse and Cherrier, 2017).

[4] That said, traditional methods such as kmeans cluster analysis and ridge regression are often now associated with ML.

are more statistical in nature, and importantly, it emphasises the principled assessment of uncertainty. The distinction between these fields is subtle and narrowing.

Both fields are conventionally divided into two areas: unsupervised and supervised learning problems. **Unsupervised learning problems** are defined by tasks for which there is no output variable, only inputs. A key objective is dimension reduction: in other words, we wish to reduce the complexity of the data. Some of these methods, such as principal component analysis (PCA) and cluster analysis, are already well-known to economists. Unsupervised machine learning can be used to generate inputs (features) for supervised learning (e.g., principal component regression), as we suggested previously. Whilst these techniques are important, we do not consider them here further. We refer interested readers Gentzkow, Shapiro and Taddy (2017) or Athey and Imbens (2019) for primers.

In **supervised learning problems**, the researcher has an outcome for each individual y_i and predictors $\boldsymbol{x}_i$. The objective of the researcher is to fit a model using the (training) data which can be used to accurately predict y_i (or classify it if y_i is categorical) using additional ('held out') data $\{\boldsymbol{x}_i\}_{i \in H}$, where $H = \{i : i > n\}$.

1.1.2 Econometrics and Machine Learning

We can think of applied econometrics – econometrics as practiced by economists – as consisting in large part of two different but related and overlapping activities: predictive inference and causal inference.

Prediction is often done by economists in the context of forecasting using time-series data. The typical forecasting question: how can we reliably forecast y_{t+s} (GDP growth, inflation, etc.) based on information available up to time t? 'Nowcasting' is a variation on forecasting, where the nowcast takes place at time t but y_t becomes available only in the future after a lag of s periods. Prediction of outcomes is also done in cross-section and spatial settings; see Bansak et al. (2018) for an example (assignment of refugees across resettlement locations) and Mullainathan and Spiess (2017) for a general discussion.

Causal inference is fundamentally different, or alternatively, is a very special form of predictive inference. The typical causal inference question: what is the predicted policy impact on y of a change in policy d, and how can we estimate that impact using data $\{(y_i, \boldsymbol{x}_i, d_i)\}_{i=1}^n$? (Note that we have separated out the 'treatment' variable d from the other covariates x. We return to this point shortly.) The general framework used by applied economists allows for causal inference with respect to the treatment d. Usually the researcher specifies a model using theory and perhaps diagnostic or specification tests; the model is estimated using the full dataset; and parameter estimates

and confidence intervals are obtained based either on large-sample asymptotic theory or small-sample theory.

Typical examples from labour economics would be the impact of changing the school-leaving age (and hence education levels) on wages, and the impact of minimum wages on employment levels. In this framework, the causal variable d is commonly thought of as a policy level, i.e., something that can be set or influenced by economic policymakers. The framework is, of course, used widely outside of economics as well, and much of the work currently being done in causal inference is cross-disciplinary. For example, it is now common for econometrics textbooks to explain causal inference in terms of the 'experimental ideal', often referred to as the 'gold standard' for causal inference: if researchers could conduct a randomised control trial (RCT) and set d to have different values in two random samples of subjects, what would be the (mean) difference in outcomes y between the two groups? And indeed, RCTs and field experiments are now part of the standard armoury of applied economists; applied development economics in particular has been revolutionised by this approach.

Curiously, although the distinction between predictive and causal inference is fairly straightforward to explain and fundamental to what applied economists do, it is not often treated clearly as such in textbooks at either the undergraduate or graduate level. The standard approach in econometrics has been to teach both using the same toolbox: specify a parametric model, show how it can be estimated using various methods (Least Squares, Instrumental Variables (IV), Generalized Methods of Moments (GMM), Maximum Likelihood, etc.), and discuss the conditions under which a causal parameter can be consistently estimated. Prediction and forecasting is typically covered separately in detail as part of time series analysis.

The connection between prediction in econometrics and machine learning is fairly obvious. Economists and econometricians who work in forecasting have taken great interest in machine learning methods, and are importing these techniques into their work. We do not discuss the work in this area here.

This paper, instead, looks at how machine learning methods can be used in estimating **causal** impacts, and this is where the distinction between d and x comes in. We focus in particular on the confounder or omitted variable bias problem. Omitted variable bias means that standard methods for estimating the treatment effect of d will yield coefficient estimates that are biased and inconsistent. The standard textbook remedy is to include 'controls' $\boldsymbol{x}$. The practical problem facing researchers is that often the choice of controls is very difficult, and in particular the set of potential controls may be **high-dimensional**. The standard framework for estimating causal effects assumes that both d and x are low-dimensional. If x is high-dimensional, the research has a problem: if all controls are inserted, overfitting means the model estimates will be badly biased; the researcher selects a small number of controls but they are the wrong ones, the model estimates will again be bi-

ased. Machine learning methods can address this problem, and in this paper we show how to employ one such method: the 'post-double-selection' (PDS) and related methods introduced by (Belloni et al., 2012, 2014b) that use a popular machine-learning estimator, the lasso or Least Absolute Shrinkage and Selection Operator (Tibshirani, 1996).

The structure of the paper is as follows. In the next section we discuss in detail the lasso estimator and a particular version with a theory-driven penalty, the 'rigorous lasso'. We then discuss how this version of the lasso is used for causal inference in the PDS method of Belloni et al. PDS method. We illustrate its use with what is literally a textbook example: the impact of a disamenity on housing prices, employed by Wooldridge in his widely-used undergraduate and graduate texbooks (Wooldridge 2009, 2010) to illustrate the importance of including controls to address omitted variable bias. The last section briefly surveys the current literature and advances in this area.

The software used to implement the estimators used here is Stata and in particular the **lassopack** and **pdslasso** packages by Ahrens, Hansen, and Schaffer. See Ahrens et al. (2019) for a detailed discussion of lassopack.

1.2 Sparsity and the rigorous or plug-in lasso

1.2.1 High-dimensional data and sparsity

The high-dimensional linear model is:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{1.1}$$

We index observations by i and regressors by j. We have up to $p = \dim(\boldsymbol{\beta})$ potential regressors. p can be very large, potentially even larger than the number of observations n. For simplicity we assume that all variables have already been mean-centered and rescaled to have unit variance, i.e., $\sum_i y_i = 0$ and $\frac{1}{n}\sum_i y_i^2 = 1$, and similarly for the predictors x_{ij}.

If we simply use OLS to estimate the model and p is large, the result is disaster: we overfit badly and classical hypothesis testing leads to many false positives. If $p > n$, OLS is not even identified.

How to proceed depend on what we believe the 'true model' is. Does the model include only a small number of regressors or a very large number with little individual contribution. In other words, is the model 'sparse' or 'dense'?

In this paper, we focus primarily on the 'sparse' case and in particular an estimator that is particularly well-suited to the sparse setting, namely the *lasso* or 'Least Absolute Shrinkage and Selection Operator' introduced by Tibshirani (1996). One of the appealing features of the lasso is that it is both well-suited to the high-dimensional setting in terms of predictive

performance, and at the same time the lasso solution is sparse, with most coefficient estimates set exactly to zero, thus facilitating model interpretation.

In the **exact sparsity** case of the p potential regressors, **only s regressors belong in the model**, where

$$s := \sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\} \ll n. \tag{1.2}$$

In other words, most of the true coefficients β_j are actually zero. The problem facing the researcher is that which are zeros and which are not is unknown.

We can also use the weaker assumption of **approximate sparsity**: some of the β_j coefficients are well-approximated by zero, and the approximation error is sufficiently 'small'. The discussion and methods we present in this paper typically carry over to the approximately sparse case, and for the most part we will use the term 'sparse' to refer to either setting.

The sparse high-dimensional model accommodates situations that are very familiar to researchers and that typically presented them with difficult problems where traditional statistical methods would perform badly. These include both settings where the number p of observed potential predictors is very large and and the researcher does not know which ones to use, and settings where the the number of observed variables is small but the number of potential predictors in the model is large because of interactions and other non-linearities, model uncertainty, temporal & spatial effects, etc.

1.2.2 The penalisation approach and the lasso

There are various estimators available that can be used for regularisation in a high dimensional setting; the lasso is just one of these. The basic idea behind these estimators is *penalisation*: put a penalty or 'price' on the use of regressors in the objective function that the estimator minimizes.

One option is penalisation based on the number of predictors. This is the so-called ℓ_0 'norm'. For example, the estimator could minimize the residual sum of squares minus some 'price' λ for each nonzero coefficient:

$$\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 + \lambda \left\|\boldsymbol{\beta}\right\|_0 \tag{1.3}$$

where $\|\beta\|_0 = \sum_{j=1}^{p} 1\{\beta_j \neq 0\}$, i.e., the number of predictors. There is a cost to including many predictors, and minimisation of the objective function will include dropping predictors that contribute little to the fit. AIC and BIC are examples of this approach.

The problem with penalisation based on ℓ_0 'norm' is very simple: it is computationally infeasible if p is at all large (*NP*-hard). So we need another approach.

The lasso estimator minimizes the mean squared error subject to a penalty on the *absolute size* of coefficient estimates (i.e., using the ℓ_1 norm):

$$\hat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda) = \arg\min \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 + \frac{\lambda}{n}\sum_{j=1}^{p}\psi_j|\beta_j|. \tag{1.4}$$

The tuning parameter λ controls the overall penalty level and ψ_j are predictor-specific penalty loadings.

The intuition behind the lasso is straightforward: there is a cost to including predictors, the unit 'price' per regressor is λ, and we can reduce the value of the objective function by removing the ones that contribute little to the fit. The bigger the λ, the higher the 'price', and the more predictors are removed. The penalty loadings ψ_j introduce the additional flexibility of putting different prices on the different predictors x_{ij}. The natural base case for standardised predictors is to price them all equally, i.e., the individual penalty loadings $\psi_j = 1$ and they drop out of the problem (but we will see shortly that separate pricing for individual predictors is needed in some settings).

We can say 'remove' because in fact the effect of the penalisation with the ℓ_1 norm is that *the lasso sets the $\hat{\beta}_j$s for some variables to zero.* This is what makes the lasso so suitable to sparse problems: the estimator itself has a sparse solution.

In contrast to ℓ_0-'norm' penalisation, the lasso is computationally feasible: the path-wise coordinate descent ('shooting') algorithm allows for fast estimation.

It is also useful to compare the lasso to another commonly-used regularisation method, the Ridge estimator. The Ridge estimator uses the ℓ_2 norm:

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda) = \arg\min \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 + \frac{\lambda}{n}\sum_{j=1}^{p}\psi_j\beta_j^2. \tag{1.5}$$

The Ridge estimator, like the lasso, is computationally feasible. But it typically does not have a sparse solution: all predictors will appear, and the predictors that contribute little will have small but nonzero coefficients.

The lasso, like other penalized regression methods, is subject to an attenuation bias. This bias can be addressed by post-estimation using OLS, i.e., re-estimate the model using the variables selected by the first-stage lasso (Belloni and Chernozhukov, 2013):

$$\hat{\boldsymbol{\beta}}_{\text{post}} = \arg\min \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \qquad \text{subject to} \qquad \beta_j = 0 \;\text{ if }\; \tilde{\beta}_j = 0, \tag{1.6}$$

where $\tilde{\beta}_j$ is the first-step lasso estimator such as the lasso. In other words, the first-step lasso is used exclusively as a model selection technique, and OLS is used to estimate the selected model. This estimator is sometimes referred to as the 'Post-lasso' (Belloni and Chernozhukov, 2013).

In sum, the lasso yields sparse solutions. Thus, the lasso can be used for model selection. We have reduced a complex model selection problem into a one-dimensional problem. We 'only' need to choose the 'right' penalty level, i.e., λ. But what is the 'right' penalty?

1.2.3 The lasso: Choice of penalty level

The penalisation approach allows us to simplify the model selection problem to a one-dimensional problem, namely the choice of the penalty level λ. In this section we discuss two approaches: (1) *cross-validation* and (2) *'rigorous'* or *plugin* penalisation. We focus in particular on the latter.

The objective in *cross-validation* is to choose the lasso penalty parameter based on predictive performance. Typically, the dataset is repeatedly divided into a portion which is used to fit the model (the 'training' sample) and the remaining portion which is used to assess predictive performance (the 'validation' or 'holdout' sample), usually with mean squared prediction error (MSPE) as the criterion. Arlot and Celisse (2010) survey the theory and practice of cross-validation; we briefly summarise here how it can be used to choose the lasso tuning parameter with independent data. In the case of independent data, common approaches are 'leave-one-out' (LOO) cross-validation and the more general 'K-fold' cross-validation.

In 'K-fold' cross-validation, the dataset is split into K portions or 'folds'; each fold is used once as the validation sample and the remainder are used to fit the model for some value of λ. For example, in 10-fold cross-validation (a common choice of K) the MSPE for the chosen λ is the MSPE across the 10 different folds when used for validation. LOO cross-validation is a special case where $K = 1$, i.e., every observation is used once as the validation sample while the remaining $n - 1$ observations are used to fit the model.

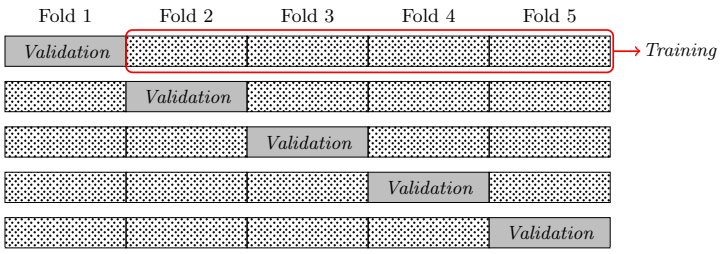

Fig. 1.1 This is cross-validation.

Cross-validation is computationally intensive because of the need to repeatedly estimate the model and check its performance across different folds and across a grid of values for λ. Standardisation of data adds to the computational cost because it needs to be done afresh for each training sample; standardising the entire dataset once up-front would violate a key principle of cross-validation, which is that a training dataset cannot contain any information from the corresponding validation dataset. LOO is a partial exception because the MSPE has a closed-form solution for a chosen λ, but a grid search across λ and repeated standardisation are still needed.

Cross-validation with dependent data adds further complications because we need to be careful that the validation data are independent of the training data. For example, one approach used with time-series data is 1-step-ahead cross-validation (Hyndman et al. 2018), where the predictive performance is based on a training sample with observations through time t and the forecast for time $t + 1$. The main setting for this paper is independent data so we do not discuss the dependent case further.

1.2.4 The Rigorous or Plug-in Lasso

Bickel et al. (2009) presented a theoretically-derived penalisation method for the lasso that assumed a known error variance. The method extended and feasible algorithms proposed in a series of papers by Belloni, Chernozhukov, Hansen, and coauthors to accommodate homoskedasticity with unknown variance, heteroskedasticity, non-Gaussian errors and clustered errors (e.g., Belloni et al. (2011), Belloni and Chernozhukov (2013), Belloni et al. (2016), Chernozhukov et al. (2015)). The approach is referred to in the literature as the 'rigorous' or 'plug-in' lasso. The rigorous lasso has several appealing features for our purposes. First, it has properties that enable it to be used straightforwardly in support of causal inference, the main topic of this paper. Second, it is theoretically and intuitively appealing, and a useful illustration of how theoretical approaches to high-dimensional problems can work. Lastly, it is computationally attractive compared to cross-validation, and hence of practical interest in its own right.

The rigorous lasso is consistent in terms of prediction and parameter estimation under assumptions about three important model characteristics:

- **Sparsity**
- **Restricted sparse eigenvalue condition**
- **The 'regularisation event'**

We consider each of these in turn.

We have already discussed *exact sparsity*: there is a large set of potentially relevant variables, but the true model contains only a small number of them. Exact sparsity is a strong assumption, and in fact it is stronger than is needed for the rigorous lasso. Instead, we assume *approximate sparsity*. Intuitively,

some true coefficients may be non-zero but small enough in absolute size that the lasso performs well even if the corresponding predictors are not selected.

Belloni et al. (2012) define the *approximate sparse model (ASM)*,

$$y_i = f(\boldsymbol{w}_i) + \varepsilon_i = \boldsymbol{x}_i'\boldsymbol{\beta}_0 + r_i + \varepsilon_i. \tag{1.7}$$

where ε_i are independently distributed, but possibly heteroskedastic and non-Gaussian errors. The elementary predictors $\boldsymbol{w}_i$ are linked to the dependent variable through the unknown and possibly non-linear function $f(\cdot)$. The objective is to approximate $f(\boldsymbol{w}_i)$ using the target parameter vector $\boldsymbol{\beta}_0$ and the transformations $\boldsymbol{x}_i := P(\boldsymbol{w}_i)$, where $P(\cdot)$ is a set of transformations. The vector of predictors $\boldsymbol{x}_i$ may be large relative to the sample size. In particular, the setup accommodates the case where a large number of transformations (polynomials, dummies, etc.) approximate $f(\boldsymbol{w}_i)$.

Approximate sparsity requires that $f(\boldsymbol{w}_i)$ can be approximated sufficiently well using only a small number of non-zero coefficients. Specifically, the target vector $\boldsymbol{\beta}_0$ and the sparsity index s need to satisfy

$$\|\boldsymbol{\beta}_0\|_0 := s \ll n \qquad \text{with} \quad \frac{s^2 \log^2(p \vee n)}{n} \to 0 \tag{1.8}$$

and the resulting approximation error $r_i = f(\boldsymbol{w}_i) - \boldsymbol{x}_i'\boldsymbol{\beta}_0$ satisfied the bound

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} r_i^2} \leq C\sqrt{\frac{s}{n}}, \tag{1.9}$$

where C is a positive constant.

For example, consider the case where $f(\boldsymbol{w}_i)$ is linear with $f(\boldsymbol{w}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}^\star$, but the true parameter vector $\boldsymbol{\beta}^\star$ is high-dimensional: $\|\boldsymbol{\beta}^\star\|_0 > n$. Approximate sparsity means we can still approximate $\boldsymbol{\beta}^\star$ using the sparse target vector $\boldsymbol{\beta}_0$ as long as $r_i = \boldsymbol{x}_i'(\boldsymbol{\beta}^\star - \boldsymbol{\beta}_0)$ is sufficiently small as specified in (1.9).

The *Restricted sparse eigenvalue condition* (RSEC) relates to the Gram matrix, $n^{-1}\boldsymbol{X}'\boldsymbol{X}$. The RSEC condition specifies that sub-matrices of the Gram matrix of size m are well-behaved (Belloni et al. 2012). Formally, the RSEC requires that the minimum sparse eigenvalues

$$\phi_{\min}(m) = \min_{1 \leq \|\boldsymbol{\delta}\|_0 \leq m} \frac{\boldsymbol{\delta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2^2} \qquad \text{and} \qquad \phi_{\max}(m) = \max_{1 \leq \|\boldsymbol{\delta}\|_0 \leq m} \frac{\boldsymbol{\delta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2^2}$$

are bounded away from zero and from above. The requirement that $\phi_{\min}(m)$ is positive means that all sub-matrices of size m have to be positive definite.[5]

The *regularisation event* is the third central condition required for the consistency of the rigorous lasso. Denote by $\boldsymbol{S} = \nabla\hat{Q}(\boldsymbol{\beta})$, the gradient of

[5] Bickel et al. (2009) use instead the weaker *restricted eigenvalue condition (REC)*. The RSEC implies the REC and has the advantage of being sufficient for both the lasso and the post-lasso.

the objective function $\hat{Q}$ at the true value $\boldsymbol{\beta}$. $S_j = \frac{2}{n}\sum_{i=1}^{n} x_{ij}\varepsilon_i$ is the jth element of the score vector. The idea is to select the lasso penalty level(s) to control the estimation noise as summarised by the score vector. Specifically, the overall penalty level λ and the predictor-specific penalty loadings ψ_j are chosen so that the 'regularisation event'

$$\frac{\lambda}{n} \geq c \max_{1\leq j\leq p} \left|\psi_j^{-1} S_j\right| \tag{1.10}$$

occurs with high probability, where $c > 1$ is a constant slack parameter.

Denote by $\Lambda = \max_j |\psi_j^{-1} S_j|$ the maximal element of the score vector scaled by the predictor-specific penalty loadings ψ_j, and denote by $q_\Lambda(\cdot)$ the quantile function for Λ, i.e., the probability that Λ is at most a is $q_\Lambda(a)$. In the rigorous lasso, we choose the penalty parameters λ and ψ_j and confidence level γ so that

$$\frac{\lambda}{n} \geq c q_\Lambda(1-\gamma). \tag{1.11}$$

The intuition behind this approach is clear from a very simple example. Say that no predictors appear in the true model ($\beta_j = 0 \;\forall\; j = 1, \ldots, p$). For the lasso to select no variables, the penalty parameters λ and ψ_j need to satisfy $\lambda \geq 2\max_j |\sum_i \psi_j^{-1} x_{ij} y_i|$.[6] Because none of the regressors appear in the true model, $y_i = \varepsilon_i$, and the requirement for the lasso to correctly identify the model without regressors is therefore $\lambda \geq 2\max_j |\sum_i \psi_j^{-1} x_{ij}\varepsilon_i|$. Since $x_{ij}\varepsilon_i$ is the score for observation i and predictor j, this is equivalent to requiring $\lambda \geq n\max_j |\psi_j^{-1} S_j|$, which is the regularisation event in (1.10). We want this event to occur with high probability of at least $(1-\gamma)$. We therefore choose values for λ and ψ_j such that $\frac{\lambda}{n} \geq q_\Lambda(1-\gamma)$. Since $q_\Lambda(\cdot)$ is a quantile function, by definition we will choose the correct model (no predictors) with probability of at least $(1-\gamma)$. This yields (1.11), the rule for choosing penalty parameters.[7]

The procedure for choosing λ is not yet feasible, because the quantile function $q_\Lambda(\cdot)$ for the maximal element of the score vector is unknown, as is the predictor-specific penalty loadings ψ_j. We discuss how these issues are addressed in practice in the next subsection.

If the sparsity and restricted sparse eigevalue assumptions ASM and RSEC are satisfied, if certain other technical conditions are satisfied,[8] and if λ and ψ_j are estimated as described below, then Belloni et al. (2012) show the lasso and post-lasso obey:

---

[6] See, for example, Hastie et al. (2015, Ch. 2).

[7] In this special case, the requirement of the slack is loosened and $c = 1$.

[8] These conditions relate to the use of the moderate deviation theory of self-normalized sums (Jing et al., 2003) that allows the extension of the theory to cover non-Gaussianity. See Belloni et al. (2012).

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}-\boldsymbol{x_i'}\boldsymbol{\beta}\right)^2}=O\left(\sqrt{\frac{s\log(p\vee n)}{n}}\right), \tag{1.12}$$

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1=O\left(\sqrt{\frac{s^2\log(p\vee n)}{n}}\right), \tag{1.13}$$

$$\|\hat{\boldsymbol{\beta}}\|_0=O\left(s\right) \tag{1.14}$$

Equation (1.12) provides an asymptotic bound for the prediction error. Equation (1.13) provides an asymptotic bound for the bias in the estimated $\hat{\boldsymbol{\beta}}$. Equation (1.14) provides a sparsity bound; the number of selected predictors in the estimated model does not diverge relative to the true model.

The 'oracle' estimator is the least squares estimator obtained if the s predictors in the model were actually known. This provides a useful theoretical benchmark for comparison. Here, if the s predictors in the model were known, the prediction error would converge at the oracle rate $\sqrt{s/n}$. Thus, the logarithmic term $\log(p \vee n)$ in (1.12)-(1.13) can be interpreted as the cost of not knowing the true model. For this reason, Belloni et al. (2012) describe these rates of convergence as *near-oracle* rates.

For the case of the lasso with theory-driven regularisation, Belloni and Chernozhukov (2013) have shown that post-estimation OLS, also referred to as post-lasso, achieves the same convergence rates as the lasso and can outperform the lasso in situations where consistent model selection is feasible (see also Belloni et al., 2012).

The rigorous lasso has recently been shown to have certain appealing properties *vis-a-vis* the K-fold cross-validated lasso. The rates of convergence of the rigorous lasso are faster than those for the K-fold cross-validated lasso derived in Chetverikov et al. (2019). More importantly for our purposes – using the lasso to assist in causal inference – the sparsity bound for the K-fold cross-validated lasso derived in Chetverikov et al. (2019) does not exclude situations where (1.14) fails badly, in the sense that the number of predictors selected via cross-validation is much larger than s. One of the implications is that cross-validation will select a penalty level λ that is 'too small' in the sense that the regularisation event (1.10) will no longer be guaranteed to occur with high probability. While it is possible to use the cross-validated lasso, and indeed other machine learning estimators, to address the confounder problem in causal inference, it should not be used in the basic 'post-double selection' framework we discuss below; other techniques are needed. We return to this point later.

1.2.5 Implementing the Rigorous Lasso

The quantile function $q_A(\cdot)$ for the maximal element of the score vector is unknown. The most common approach to addressing this is to use a theoretically-derived upper bound that guarantees that the regularisation event (1.10) holds asymptotically.[9] Specifically, Belloni et al. (2012) show that

$$\mathrm{P}\left(\max_{1\leq j\leq p} c\left|S_j\right| \leq \frac{\lambda\psi_j}{n}\right) \to 1 \quad \text{as} \quad n\to\infty,\ \gamma\to 0 \tag{1.15}$$

if the penalty levels and loadings are set to

$$\lambda = 2c\sqrt{n}\Phi^{-1}(1-\gamma/(2p)) \qquad \psi_j = \sqrt{\frac{1}{n}\sum_i x_{ij}^2\varepsilon_i^2} \tag{1.16}$$

c is the slack parameter from above and $\gamma\to 0$ means the probability of the regularisation event converges towards 1. Common settings for c and γ, based on Monte Carlo studies are $c=1.1$ and $\gamma = 0.1/\log(n)$, respectively.

The only remaining element is estimation of the ideal penalty loadings ψ_j. Belloni et al. (2012), Belloni et al. (2014b) recommend an iterative procedure based on some initial set of residuals $\hat{\varepsilon}_{0,i}$. One choice is to use the d predictors that have the highest correlation with y_i and regress y_i on these using OLS; $d=5$ is their suggestion. The residuals from this OLS regression can be used to obtain an initial set of penalty loadings $\hat{\psi}_j$ according to (1.16). These initial penalty loadings and the penalty level from (1.16) are used to obtain the lasso or post-lasso estimator $\hat{\boldsymbol{\beta}}$. This estimator is then used to obtain a updated set of residuals and penalty loadings according to (1.16), and then an updated lasso estimator. The procedure can be iterated further if desired.

The framework set out above requires only independence across observations; heteroskedasticity, a common issue facing empirical researchers, is automatically accommodated. The reason is that heteroskedasticity is captured in the penalty loadings for the score vector.[10] Intuitively, heteroskedasticity affects the probability that the term $\max_j|\sum_i x_{ij}\varepsilon_i|$ takes on extreme values, and this needs to be captured via the penalty loadings.

In the special case of homoskedasticity, the ideal penalisation in (1.16) simplifies:

$$\lambda = 2c\sigma\sqrt{n}\Phi^{-1}(1-\gamma/(2p)), \qquad \psi_j = 1. \tag{1.17}$$

This follows from the fact that we have standardised the predictors to have unit variance and hence homoskedasticity implies $E(x_{ij}^2\varepsilon_i^2) = \sigma^2 E(x_{ij}^2) = \sigma^2$.

[9] The alternative is to simulate the distribution of the score vector. This is known as the 'exact' or *X-dependent* approach. See Belloni and Chernozhukov (2011) for details and Ahrens et al. (2019) for a summary discussion and an implementation in Stata.

[10] The formula in (1.16) for the penalty loading is familiar from the standard Eicker-Huber-White heteroskedasticty-robust covariance estimator.

The iterative procedure above is used to obtain residuals to form an estimate $\hat{\sigma}^2$ of the error variance σ^2.

The rigorous lasso has also been extended to cover a special case of dependent data, namely panel data. In the cluster-lasso (Belloni et al., 2016), arbitrary within-panel correlation is accommodated. We do not discuss it further here and refer the reader to Belloni et al. (2016) for a presentation and Ahrens et al. (2019) for a summary discussion and implementation in Stata.

1.3 Using machine learning to assist causal inference: The lasso and 'post-double-selection'

1.3.1 The confounder problem

Programme evaluation is a fundamental part of the modern empirical social sciences. This is particularly true in economics, where the profession is frequently relied on to give sound advice about the structure and scope of large scale government policies. For theorists too, these studies are important, as they allow them to test the explanatory power of their models. As a result, it is imperative that these evaluations be as credible as possible.

In this literature, it is common to assume that unconfoundedness holds—in other words, conditioning on observable factors is sufficient to make the treatment assignment as good as random. For much of the literature's history, there was little substantive advice given to practitioners to help them satisfy this condition. Unfortunately, the issue is not a simple one.

Many of the most important questions posed in the empirical social sciences revolve around the causal effect of a programme or policy. However, estimating and drawing inferences about such effects is often difficult to achieve in a credible manner. If a researcher is fortunate, they may be able to conduct a randomised control trial (RCT), which would (in a number of important respects) simplify the task, since it allows one to calculate an unbiased estimate of the average effect of the treatment (that is, the policy, project or programme to which the 'units' were exposed). This is important, as *unbiasedness* imposes no requirement on the researcher to know anything specific about relevant covariates or confounders Deaton and Cartwright (2018). However, even in experimental setups it may not be possible for researchers to rely entirely on randomisation of treatment. In such circumstances the confounder problem reappears and researchers face the problem of choosing controls to address it.

In any case, in the discipline of economics researchers are often extremely limited in the extent to which they can employ randomised control trials Athey and Imbens (2017). RCTs are expensive, not always feasible and in some cases are unethical.

For instance, consider the evaluation of policy on minimum wages: it is hardly fair or politically feasible to randomly assign people or places to be subject to different minimum wage regulations. Studies evaluating the impact of minimum wage laws, as well as a myriad of other important policies, have often utilised other techniques which rely instead on observational data, where the assignment to treatment is not random. And because the assignment to treatment is not random, researchers are immediately confronted by the problem of omitted confounders, i.e., omitted variable bias.

Successfully addressing the omitted variable bias problem is challenging, even when confounders or proxies for them are observable. 'Just insert some controls' is not adequate advice.

- The dimensionality of the controls may be large, immediately posing a problem for the researcher if, as is often the case, they do not have a strong theoretical or other prior basis for reducing the number of controls.
- Include too many controls, and the estimated treatment effect will suffer from overfitting.
- Include too few controls, and it will suffer from bias.
- Typically the researcher also lacks information about whether interactions and/or polynomials would be required to adequately address the problem. A low-dimensional problem can easily become a high-dimensional problem this way.
- Classical hypothesis testing to reduce the number of controls are poorly suited to addressing the problem because of the resulting pre-test bias. False discovery control (multiple testing procedures) is problematic and rarely done.
- Perhaps most worrisome of all is the 'research degrees of freedom' (Simmons et al., 2011) or 'garden of forking paths' (Gelman and Loken, 2013) problem. Researchers may try many combinations of controls, looking for statistical significance in their results, and then report only the results that 'work'.[11]

Recent work by various authors has shown how machine learning methods can be used to address this problem. In the next subsection, we look at one such method: the 'post-double-selection' (PDS) and related methods of Belloni, Chernozhukov, and Hansen (2014a). These authors show that the 'rigorous lasso' can be used to select controls in a theory-driven, parsimonious and semi-automated way. 'Theory-driven' means that asymptotic properties of PDS estimators are known. 'Parsimonious' means that the controls selected can address the omitted variable bias problem and at the same time avoid gross overfitting. 'Semi-automated' means that researcher degrees of freedom in the selection of controls are reduced and hence 'p-hacking' is restrained.

[11] These authors are careful to note that the problem readily arises when researchers make decisions contingent on their data analysis; no conscious attempt to deceive is needed. Deliberate falsification, sometimes called 'p-hacking', is special case and likely much rarer.

1.3.2 The lasso and causal inference

The main strength of the lasso is prediction rather than model selection. But the lasso's strength as a prediction technique can also be used to aid causal inference.

In the basic setup causal inference setup, we *already know* the causal variable of interest. No variable selection is needed for this. The lasso is used instead to **select controls** used in the estimation. These other variables are not themselves subject to causal inference. But using them means we can obtained improved causal inference for the variable we **are** interested in.

Why can we use the lasso to select controls even though the lasso is (in most scenarios) not model selection consistent? There are two ways to look at this:

- **Immunisation property:** The moderate model selection mistakes of the lasso do not affect the asymptotic distribution of the estimator of the low-dimensional parameters of interest (Belloni et al., 2012, 2014b). We can treat modelling the the nuisance component of our structural model as a prediction problem.
- The **irrepresentable condition** states that the lasso will fail to distinguish between two variables (one in the active set, the other not) if they are highly correlated. These type of variable selection mistakes are not a problem if the aim is to control for confounding factors.

We note here in passing that the PDS lasso methodology can also be used to select instruments from a high-dimensional set in order to address endogeneity in instrumental variables (IV) estimation. We focus on the 'selection of controls' problem here; for discussion of the IV application, see Belloni et al. (2012) for the development of the theory.

1.3.3 The Post-Double Selection (PDS) estimator

Our model is

$$y_i = \underbrace{\alpha d_i}_{\text{aim}} + \underbrace{\beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}}_{\text{nuisance}} + \varepsilon_i.$$

The causal variable of interest or "treatment" is d_i. The xs are the set of potential controls and not directly of interest. We want to obtain an estimate of the parameter α. The problem is the controls. We want to include controls because we are worried about omitted variable bias – the usual reason for including controls. But which ones do we use?

The naive approach does not work. If estimate the model using the rigorous lasso–but imposing that d_i is not subject to selection–and use the controls selected by the lasso, the estimated $\hat{\alpha}$ will be badly biased. The reason is that

we might miss controls that have a strong predictive power for d_i, but only small effect on y_i. Similarly, if we only consider the regression of d_i against the controls, we might miss controls that have a strong predictive power for y_i, but only a moderately sized effect on d_i. See Belloni et al. (2014a).

Instead, we use the **Post-Double-Selection lasso**:

- Step 1: Use the lasso to estimate

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_j x_{i,j} + \ldots + \beta_p x_{i,p} + \varepsilon_i,$$

  i.e., without d_i as a regressor. Denote the set of lasso-selected controls by A.
- Step 2: Use the lasso to estimate

$$d_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_j x_{i,j} + \ldots + \beta_p x_{i,p} + \varepsilon_i,$$

  i.e., where the causal variable of interest is the dependent variable. Denote the set of lasso-selected controls by B.
- Step 3: Estimate using OLS

$$y_i = \alpha d_i + \mathbf{w}_i' \beta + \varepsilon_i$$

  where $\mathbf{w}_i = A \cup B$, i.e., the union of the selected controls from Steps 1 and 2.

The PDS method is easily extended to cover the case of multiple causal variables: simply repeat Step 2 for each causal variable and add the selected controls to the set B. We illustrate in the example of housing prices below.

Belloni et al. (2012, 2014b) show that the estimate of α using the above methodology is consistent and asymptotically normal, under fairly general conditions. The restricted sparse eigenvalue is the same as in Section 1.2.4. There are further technical conditions relating to technical conditions; see (Belloni et al., 2014b, Condition SM). As above we need sparsity, but this approximate sparsity has to hold in both equations:

$$m(\boldsymbol{w}_i) = \boldsymbol{x}_i' \boldsymbol{\beta}_{m0} + r_{mi}, \qquad \|\boldsymbol{\beta}_{m0}\|_0 \leq s, \qquad \sqrt{\frac{1}{n}\sum_{i=1}^{n} r_{mi}^2} \leq C\sqrt{\frac{s}{n}} \tag{1.18}$$

$$g(\boldsymbol{w}_i) = \boldsymbol{x}_i' \boldsymbol{\beta}_{g0} + r_{gi}, \qquad \|\boldsymbol{\beta}_{g0}\|_0 \leq s, \qquad \sqrt{\frac{1}{n}\sum_{i=1}^{n} r_{gi}^2} \leq C\sqrt{\frac{s}{n}} \tag{1.19}$$

where $\frac{s^2 \log^2(p \vee n)}{n} \to 0$. C is constant like above. An important caveat is that justifies inference on the causal variable(s), but not on the selected high-dimensional controls.

The intuition behind the PDS lasso is that the procedure approximately orthogonalizes y_i and d_i with respect to the disturbance ε_i. A comparison with the Frisch-Waugh-Lovell (FWL) Theorem is perhaps helpful here.

Consider the simple regression model

$$Y = X_1\beta_1 + X_2\beta_2 + u$$

and the researcher is interested in β_1. The OLS estimate $\hat{\beta}_1$ can be obtained simply by regression Y against X_1 *and* X_2. Leaving X_2 out of the regression would cause the estimate $\hat{\beta}_1$ to suffer from omitted variable bias should the omitted X_2 be correlated with the included X_1.

The FWL Theorem states that the following procedure

1. Regress Y against X_2 (call the residuals $\tilde{Y}$)
2. Regress X_1 against X_2 (call the residuals $\tilde{X_1}$)
3. Regress $\tilde{Y}$ against $\tilde{X_1}$

generates the *numerically* same estimate $\hat{\beta}_1$. By regressing against X_2, we derive orthogonalized versions of Y and X_1 and, this way, account for the omitted variable bias.

The PDS methodology, in effect, finds a parsimonious set of controls that can be used to orthogonalize the outcome and treatment variables.

A closely related alternative to lasso PDS that is asymptotically equivalent is *Double-Orthogonalisation* (DO), proposed by Chernozhukov-Hansen-Spindler 2015. The PDS method is equivalent to FWL partialling-out all selected controls from both y_i and d_i. The DO method essentially partials out from y_i only the controls in set A (selected in Step 1, using the lasso with y_i on the LHS), and partials out from d_i only the controls in set B (selected in Step 2, using the lasso with d_i on the LHS). DO partialling-out can use either the lasso or Post-lasso coefficients. DO and PDS are asymptotically equivalent.

1.3.4 An example: The impact of a disamenity on house prices

The example we use is literally a textbook example, taken from Jeff Wooldridge's undergraduate and advanced graduate textbooks.[12] The original study is Kiel and McClain (1995), who look at the impact of a new incinerator (local waste disposal) on the prices of houses near the incinerator. The location is North Andover, Massachusetts. Starting in 1979, rumours about building a new incinerator start to circulate. Construction of the incinerator starts in 1981. By 1981, information about the new incinerator should be reflected in house

[12] Wooldridge (2009), pp. 450-3, 474 and Wooldridge (2010), pp. 153-4.

prices. We expect that houses that are closer to the site of the new incinerator should be negatively affected because of the perceived cost of the disamenity.

The example is used by Wooldridge to illustrate a 'differences-in-differences' estimation strategy: we compare the change in sales prices 1978-81 of houses near the incinerator vs the prices of houses far from the incinerator. We pool data for the years 1978 and 1981, and specify the model to estimate as

$$\log(rprice_i) = \beta_1 + \beta_2 \log(dist_i) + \beta_3 y81_i + \beta_4(\log(dist_i) \times y81_i) + \varepsilon_i \quad (1.20)$$

where $rprice_i$ is the sales price of house i in 1978 dollars, $dist_i$ is the distance in miles to the incinerator, and $y81_i$ is a dummy variable = 1 if the house sale took place in 1981 and = 0 if the house sale took place in 1978.

The omitted variables bias problem is obvious: we expect the choice of location of the incinerator to be related to the quality of housing nearby and to the nature of the location. It is likely the incinerator is built near low-quality housing and in an undesirable location. This will affect the estimate of the pure distance effect β_2 and possibly also the interaction effect β_3.

The potential set of controls includes variables that measure characteristics of the house itself (age, number of rooms, number of bathrooms, log house size in square feet, log land area in square feet), location characteristics (log distance to nearest interstate highway, log distance to the central business district). The textbook exposition cited above suggests using square terms of age and log(distance to highway) but not the central business district measures. Why higher-order terms should be included for some measures but not others, and why some measures should be included and not others, is difficult to justify or specify *a priori*.

In principle, including all possible levels and interactions of controls is appealing – in effect, a second-order Taylor approximation to an arbitrary control function – but in the traditional approach, where all these variables are included without penalisation, we run the risk of overfitting. The PDS methodology addresses this very easily; we include the full set of levels, squares and interactions (34 controls in all), and only those that contribute substantially to addressing the omitted variable bias are retained.

Below we report the result of estimating with no controls at all, the PDS estimates, and the DO estimates using the lasso and post-lasso coefficients, using the Wooldridge textbook dataset of 321 observations from the Kiel-McClain study. Heteroskedasticity-robust penalty loadings are used in the PDS and DO estimations; heteroskedasticy-robust standard errors are reported for all four sets of results.

Table 1.1 reports the lasso-selected controls based on the separate estimations for $log(rprice)$, $log(distance)$ and $log(distance) \times y81$. The lasso estimation for the dependent variable in the structural equation $log(rprice)$ selects 5 out of the 34 possible penalized controls; all 5 are interactions. The lasso estimations for the explanatory variables in the structural equation, $log(distance)$ and $log(distance) \times y81$, both select only the square of the log

Dependent variable:	**Selected:**
log($rprice$)	$baths \times \log(area)$, $\log(land) \times rooms$, $\log(land) \times \log(area)$, $log(area) \times rooms$, $\log(area) \times \log(CBD)$
log($distance$)	$\log(CBD)^2$
log($distance$) $\times\ y81$	$\log(CBD)^2$
Selected from:	Levels, squares and cross-products (34) of age, number of rooms, number of baths, log land area, log house area, log distance to interstate highway, log distance to central business district (CBD)
Unpenalised controls:	$y81$

Table 1.1 Lasso-selected variables for Kiel-McClain example

of the distance to the central business district. In all these estimations, the dummy for 1981 $y81$ is always included; this is done by specifying that the variable has a zero penalty in the lasso estimations. An unpenalised intercept is also always included. The union of these selected predictors plus the unpenalised dummy $y81$ yields 7 controls for the PDS estimation. The selected predictors for the 3 separate lassos are separately partialled out from $log(rprice)$, $log(distance)$ and $log(distance) \times y81$ and then used in the DO estimations. Either the lasso or post-lasso OLS coefficients can be used for the partialling-out, yielding two different sets of DO estimation results.

Regressor	**No controls**	**PDS**	**DO-lasso**	**DO-post-lasso**
log($distance$)	0.317	0.060	0.003	0.046
	(0.038)	(0.065)	(0.057)	(0.062)
log($distance$) $\times\ y81$	0.048	0.022	0.041	0.017
	(0.077)	(0.050)	(0.047)	(0.048)
$y81$	-0.275	-0.075	(n.a.)	(n.a.)
	(0.762)	(n.a.)		

Table 1.2 Structural equation estimations for the Kiel-McClain example; heteroskedastic-robust standard errors in parentheses

Table 1.2 reports the no-controls, PDS, and DO estimation results. Coefficients and heteroskedasticity-robust standard errors are shown for all four estimations. The coefficient on the year dummy $y81$ is also shown, as is standard for a differences-in-differences estimation, but no standard error is displayed in the PDS estimation because the variable is treated as an unpenalised control rather than as a causal variable.[13] No coefficient on $y81$ is reported for the DO estimations as it is partialled-out along with the other controls.

Without controls, there is no incinerator effect but a strong positive distance effect. The impact of the inclusion of the controls is to make the dis-

[13] To treat it as a causal variable and obtain a valid standard error, we would have to estimate an additional lasso regression with $y81$ as the dependent variable etc.

tance effect much smaller and less precisely estimated; the incinerator effect remains small and becomes slightly more precisely estimated. These results are similar to those in the examples and discussions in Wooldridge's textbooks, and illustrate the same point - the apparent distance effect is spurious and driven by omitted variable bias. But the controls are drawn from a more flexible functional form, selected in such a way that overfitting is avoided at the same time that omitted variable bias is addressed, and because the selection is automated it is relatively immune to suspicions of p-hacking.

1.3.5 Caveats

Probably the most important caveat to bear in mind when using the PDS methodology is the requirement that the confounder dimensionality is sufficiently sparse. The Kiel-McClain example is a good one in the sense that this assumption seems reasonable: the original authors set out the dimensions in which confounding could be an issue, provided proxies for these dimensions, and it is plausible that levels and interactions of these proxies is enough to approximate the problem.

In other applications, the sparsity assumption will be less plausible. For example, we might have employee data with codes for occupation, or sales data disaggregated by product code. It is natural to code characteristics using dummy variables, and it would be tempting to use this large set of dummies along with the PDS methodology to address the counfounder issue. But it is problematic to assume sparsity here, because it amounts to the assumption that most jobs or products are very similar and a few are very different.

1.4 Heterogeneous Treatment Effects

Up to this point, we have concerned ourselves with tools, still in development, that have made it possible to estimate causal effects using machine learning methods. The advantage of these techniques is that they allow us to address arbitrary and auxiliary assumptions which have the potential to weaken the validity of empirical work.

One of the most influential lines of current research in this area has focused on heterogeneity in treatment effects: it is unrealistic to assume that the effect of a policy, intervention or treatment (generally defined) does not vary for each individual to whom it was applied. The methods employed by researchers should be robust to that fact, and should explicitly target some element of the distribution of effects to ensure that the results can be reliably understood and interpreted.

Much of the modern literature on heterogeneous effects builds from a common framework, the Rubin Causal Model. Its roots lie with the foundational work conducted by Fisher (1925; 1935) and Neyman (1990) on agricultural experiments, and it is now the dominant structure used in the analysis of causality in many fields, including statistics and econometrics.

The framework is simple: suppose we wish to evaluate the effect of a policy or decision, or something similar, on a number of individuals' outcomes, which we denote separately by $Y_i \in \mathbb{R}$. We record the treatment status of each individual with $W_i \in \{0, 1\}$, which takes the value 1 for individuals who were affected explicitly by the policy, and is 0 otherwise. The settings that this describes are intentionally and necessarily restricted; the treatment status of an individual is binary. We do not allow for varying treatment intensity.

Alongside this information, we have available for each individual additional data, contained in the p-dimensional column vector X_i, which we require to address confounding. This information can be (and often is) large in scale. There are few practical constraints on the nature of the data that we can include, but there is clear guidance about the inclusion of one particular type of variable: the vector of covariates must not contain any series which is directly affected by the treatment (Wooldridge, 2005). For instance, if one were interested in evaluating the effects of a past labour market programme, incorporating as a control participants' current employment status would be inappropriate.

Finally but perhaps most importantly, the realised outcome for each individual, Y_i, is assumed to be a function of their own 'potential' outcomes $(Y_i(0), Y_i(1))$, where $Y_i(0)$ represents the outcome that individual i would have experienced had they not received the treatment, and $Y_i(1)$ is defined analogously. When the treatment is assigned, we can only observe one of these values per person. The remaining half has become 'missing' by definition, which is the fundamental problem of causal inference. We can summarize this argument by writing $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$.

This framework is extremely flexible. The causal effect of the treatment is allowed to vary at the level of the individual, as $Y_i(1) - Y_i(0)$, and is effectively left unrestricted. As a result, it is typically the case that researchers target a summary of these effects. In econometrics, the targets (or 'estimands') most commonly selected are the average treatment effect (ATE),

$$\tau = \tau_{\text{ATE}} = E\left[Y_i(1) - Y_i(0)\right],$$

and the average treatment effect on the treated (ATET)

$$\tau_{\text{ATET}} = E\left[Y_i(1) - Y_i(0) \,|\, W_i = 1\right].$$

This is at least partly due to convenience, and some prominent researchers (e.g., the Nobelist econometrician James Heckman) have criticised others for not motivating their targets more carefully. However, as a result of advances at the intersection of machine learning and causal inference (and related

programmes of research), this practice is beginning to change. Instead of simple summaries, it is now not uncommon for researchers to report the pattern of heterogeneity in effects. This is a topic we briefly return to in the conclusion.

The model is flexible in another important regard: it places no assumptions over the distribution of the potential outcomes. One could proceed to make such assumptions. However, it is not clear that one could justify such an assumption without a rigorous theoretical underpinning for it; and so it is likely that the imposition of such a condition would introduce a source of fragility into the model. As such we maintain the looser framework. Nevertheless we must impose a minimal degree of structure on the model to ensure that the estimands are identified. That structure is delivered by two assumptions.

The first, and most important, is **unconfoundedness**. Intuitively, it says that the assignment to treatment is 'as good as random', *conditional on the information contained within the covariates*. Formally, we express this as

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i. \tag{1.21}$$

Notice the burden placed on the observed covariates: they need to be sufficiently comprehensive to ensure that this condition holds. But how does one know for a given set of circumstances that this assumption is credible? Earlier we discussed the difficult process of selecting controls, and it is important to emphasise that the same dilemma arises here. However, the solution that we presented is also appropriate: machine learning methods, which are designed for variable selection and high dimensional data, can naturally address this issue for us by shifting assumptions over the relevance (or otherwise) of covariates onto an assumption of sparsity, which encodes our prior belief that they are not all required—indeed *many* should be left out of the model.

Unfortunately, whilst machine learning methods allow us to remain agnostic about the inclusion and form of covariates, and consequently make unconfoundedness more palatable, their use is also likely to increase the frequency with which researchers encounter issues related to the second identifying assumption, **overlap**:

$$\kappa < P(W_i = 1 \mid X_i = x) < 1 - \kappa, \quad \kappa \in \left(0\,,\,\frac{1}{2}\right), \quad \forall\, x. \tag{1.22}$$

D'Amour et al. (2019) demonstrate that there is an unfortunate link between the dimension of the covariates, p, and the overlap assumption. Strict overlap, which is the form presented above and the form required to ensure that the estimators we are interested in are $\sqrt{n}$-consistent, implies that the average imbalance in covariate means between the treated and control units converges to zero as the dimension p grows. This suggests that there are some instances in which it would be unsuitable to employ estimators that make use of machine learning methods without first pruning the set of covariates

to be included. Trimming outliers may help to alleviate this problem, but D'Amour et al. (2019) highlight that there are subtle issues with this fix.

To construct estimators in a setting such as this, one must turn to semiparametric theory. Semiparametric methods make it possible to leave unrestricted important components of the data generating process. This is particularly useful if we lack information about those components of the model, or in the event that those components involve nuisance functions that are not of direct interest. A semiparametric model is simply one that is indexed by an infinite-dimensional parameter (Kennedy, 2016). That definition is quite wide: it encompasses everything from nonparametric models (which in no way restrict the set of possible probability distributions for the data), to regression models with parametric functional forms but errors that have an unspecified distribution. More specifically, one can view semiparametric models as possessing a finite-dimensional parameter of interest (which is parametric) and an infinite-dimensional nuisance parameter (which is the nonparametric component) (e.g. Begun et al., 1983). For example, in the context of causal inference it is often the case that researchers place assumptions on the treatment mechanism (in other words, they model it parametrically) but they leave the outcome model unrestricted. That reflects the relative strength of the information we can garner about the two models.

Many semiparametric models can be characterised and analysed through their influence function. An estimator $\hat{\Phi}$ for Φ is asymptotically linear if it can be approximated by a sample average, with i.i.d. data $D_i = (Y_i, X_i, W_i)$, in the following manner

$$(\hat{\Phi} - \Phi_0) = \frac{1}{n}\sum_{i=1}^{n} \phi_i(D_i) + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{1.23}$$

where ϕ has mean zero and finite variance (that is, $E\left[\phi(D)\right] = 0$ and $V = E\left[\phi(D)\,\phi(D)'\right] < \infty$). Then by the Lindeberg-Levy central limit theorem, the estimator $\hat{\Phi}$ with influence function ϕ is asymptotically normal:

$$\sqrt{n}(\hat{\Phi} - \Phi_0) \rightsquigarrow N(0\,, V), \tag{1.24}$$

where $\rightsquigarrow$ denotes convergence in distribution. Thus, the estimator's asymptotic behaviour, up to a small degree of error, can be fully described by its influence function, which renders simple the construction of confidence regions and test statistics that have approximately correct coverage and size in large samples.

Equally, one can reverse this logic to create estimators given a candidate influence function. The end of this process is pleasingly simple: one solves the estimating equation formed from the sample analogue of the moment

condition the influence function satisfies.[14]

$$\frac{1}{n}\sum_{i=1}^{n}\phi_i(D_i;\Phi,\hat{\eta})=0. \tag{1.25}$$

Note here the explicit dependence of the function on a nuisance parameter η and that the estimating equation is evaluated with $\hat{\eta}$, which is produced by another model.

Before we proceed it will be useful to mildly restrict and clarify the framework we are working with. It has two basic components: the first is an equation for the outcome, Y_i, as usual; the second describes the model for the assignment to treatment. Let $\mu(w,x)=E\left[Y_i(w)\,|\,X_i=x\right]$ for some $w\in\{0,1\}$ and $e(x)=P\left(W_i=1\,|\,X_i=x\right)$, which is called the propensity score in the literature. Then

$$Y_i=\mu(W_i,X_i)+\varepsilon_i, \qquad E\left[\varepsilon_i\,|\,X_i,W_i\right]=0, \tag{1.26}$$

$$W_i=e(X_i)+\nu_i, \quad \text{and} \quad E\left[\nu_i\,|\,X_i\right]=0. \tag{1.27}$$

The target we focus on is the ATE, which in this context is

$$\tau=E\left[\mu(1,X_i)-\mu(0,X_i)\right]. \tag{1.28}$$

Note that the confounders X_i are related to both the treatment *and* the outcome variables, and hence comparing the raw outcomes of the treated and control units would produce a biased estimate. The manner in which Y_i and W_i depend on X_i (through $\mu(W_i,X_i)$ and $e(X_i)$) is left unrestricted, and it accommodates general heterogeneity in the effect of the treatment. As described in (1.7), we tackle the problem of the unknown and complex form of the functions by using a linear combination of transformations of the covariates and the treatment indicator to approximate them. As we have said before, the number of terms required for this task may be considerable, so to obtain $\hat{\mu}(W_i,X_i)$ and $\hat{e}(X_i)$, we use methods from the machine learning literature, and particularly the lasso.

How do we link these estimates to τ? One approach is to construct an influence function which identifies the target parameter through its associated moment condition. The problem with this approach is that the nuisance parameters $\eta=(\mu,e)$ are estimated imperfectly. Again, as mentioned previously, machine learning methods address high-dimensionality through regularisation. That controls, and in fact substantially reduces, the variance of the predicted values produced by the model, but achieves that at the cost of the introduction of bias in its coefficient estimates. As a result, the influence function must be designed carefully to ensure that its associated moment con-

[14] Of course, this discussion raises a more fundamental question: given a specification like that above, how does one construct such a function? The answer is a little technical, unfortunately. Interested readers can find a detailed discussion in van der Vaart (1998).

dition is locally insensitive to the value of the nuisance parameter η around η_0. Formally, we want this orthogonality condition to hold:[15]

$$\partial_\eta E\big[\phi(D;\tau_0,\eta)\big]_{\eta=\eta_0} = 0.$$

It turns out that there is an interesting and useful connection between functions which possess this property and semiparametric efficiency. The efficient influence function is the (almost surely) unique function which satisfies the semiparametric efficiency bound (analogous to the Cramer-Rao lower bound) (Powell, 1994). Aside from being efficient, it possesses a number of other interesting properties—one of which is orthogonality.[16,17] Given the target and the assumptions of unconfoundedness and overlap, its form (derived by Hahn (1998)) is as follows:

$$\phi(y,w,x;\tau,\eta) = \frac{w(y-\mu(1,x))}{e(x)} - \frac{(1-w)(y-\mu(0,x))}{1-e(x)} + (\mu(1,x)-\mu(0,x)) - \tau. \tag{1.29}$$

We recover $\hat{\tau}$, as we explained, from the finite-sample analogue of the moment condition

$$\frac{1}{n}\sum_{i=1}^{n}\phi(Y_i,W_i,X_i;\hat{\tau},\hat{\eta}) = 0, \tag{1.30}$$

The Double Machine Learning (DML) estimator of Chernozhukov et al. (2018) extends this framework in one important direction. Machine learning methods have an inbuilt tendency to overfit the data. Regularisation attenuates that but, *generally speaking*, it does not completely address it. Across equations, when the true structural errors, ϵ_i and ν_i, are correlated with the estimation errors of the nuisance parameters, such as $\hat{e}(X_i) - e(X_i)$, poor performance can result. The dependence between these components, however, can be addressed fairly easily using sample splitting.[18]

Sample splitting begins with partitioning the data into two sets, an auxiliary fold and a main fold. With the auxiliary fold, one obtains $\hat{\mu}(W_i, X_i)$ and $\hat{e}(X_i)$, that is, estimates of the nuisance parameters. Then with the main fold, the treatment effect is estimated by plugging the results of the first step, $\hat{\tau}$, into the estimating equation defined by (1.30). This process carries a significant disadvantage: the final estimate is produced using only a portion of the data, and thus it is not efficient. The DML estimator makes use of

[15] ∂_η is shorthand for $\partial/\partial\eta'$. This version of the condition is actually more stringent than required. A more general definition of it, based on Gateaux derivatives, can be found in Belloni et al. (2017) and Chernozhukov et al. (2018).

[16] The estimators that possess this property need not be semiparametrically efficient, but because they can be, we restrict our focus to those that are.

[17] Estimators based on the efficient influence function are also double-robust (Robins and Rotznizcky, 1995).

[18] Sample splitting was originally introduced by Angrist and Krueger (1995) and Altonji and Segal (1996) in the context of bias reduction of IV and GMM estimators.

an adapted form of the process called cross-fitting, where the procedure is repeated with the roles for the samples reversed. The resulting treatment effect estimates are averaged together, which ensures full efficiency. Finally, to guarantee that the performance of the estimator is not adversely affected by an unusual partition of the data, the procedure is repeated a 'large' number of times, say 100, and the median of the final set of values forms the estimate for the treatment effect. This procedure can be generalised to allow for splits of unequal size, which in finite samples may improve the performance of the estimator, as it allows more data to be used for the first stage, when the machine learning methods learn the structure of the nuisance functions.

This aspect of the estimator's design is crucial. In more conventional work, the bias induced by the data-adaptive nature of the machine learning methods would be analysed and controlled using empirical process theory. In particular, that approach would proceed by imposing constraints, called Donsker conditions, on the function class that contains the values of the nuisance estimator (Vaart and Wellner, 1996). These conditions make it possible to conclude that terms responsible for the estimator's bias vanish asymptotically. However, the size—or complexity—of that function class must be suitably bounded for it to be Donsker.[19] Unfortunately, in the environment we are interested in, where the dimension of the covariates, p, is allowed to grow with the sample size, that requirement will not hold. If the complexity of the function class grows but at a rate that is sufficiently slow relative to n, it is possible to show that the estimator's bias, caused by overfitting, will tend to zero (Chernozhukov et al., 2018). However, the assumptions required to demonstrate that result are typically restrictive in practice: for example, if one were to use the lasso to estimate the nuisance functions, the model would need to be very sparse for the bias to vanish.

With sample-splitting, those conditions become substantially weaker. The model is allowed to be more complex with that adaptation, and thus, the number of non-zero coefficients can be larger. Specifically, for the DML, if the outcome and propensity score models are estimated with the lasso, we require that $s_\mu s_e \ll n$, where s_μ and s_e denote their respective sparsity indices. Without the procedure (that is, relying solely on orthogonalisation), that condition is instead $s_\mu^2 + s_e^2 \ll n$ (Chernozhukov et al., 2018). The former is clearly weaker than the latter, and embedded within the first condition is a useful trade-off. Say the model for the propensity score was expected to be very sparse, such that $s_e \ll \sqrt{n}$, perhaps because the assignment procedure is well-understood and is contingent on only a small number of factors (as it may be if one were looking at, for instance, a treatment prescribed by clinicians). In that case, the outcome model can be relatively dense and there is scope to include many covariates in the model. Thus, we can use external information to reason about the relative complexity of the assignment model and the

[19] There are a number of conditions that are intimately related to the size of the function class, as measured by its bracketing and covering numbers, which if satisfied are sufficient for it to be Donsker. See Vaart and Wellner (1996) for a full statement.

process that determines the outcome, and then we can balance one against the other.

There are further, more general benefits to sample-splitting. The loosened requirements over the complexity of the model allow one to use, with little adjustment to the underlying theory, a large variety of machine learning methods for the estimation of η. Without the procedure, one would have to verify that those methods individually satisfied relevant complexity restrictions. The methods do not have to rely on approximate sparsity. Instead, they need only satisfy conditions on the quality of the approximations they provide for the nuisance functions. Crudely, for instance, we could require that both nuisance parameters are estimated at the $o_p\big(n^{-1/4}\big)$ rate, such that their product

$$E\Big[\big(\hat{\mu}(w, X_i) - \mu(w, X_i)\big)^2\Big]^{1/2} E\Big[\big(\hat{e}(X_i) - e(X_i)\big)^2\Big]^{1/2} = o_p\Big(n^{-1/2}\Big) \quad (1.31)$$

is asymptotically negligible (Athey and Imbens, 2019). Notice that each estimator is allowed to converge at a rate that is substantially slower than that of a correctly specified parametric model.[20] Random forests, neural nets and L_2 boosting all satisfy this condition (see, e.g., Wager and Walther (2019) and Luo and Spindler (2019)), as do more traditional, flexible estimators, such as generalised additive models. Of course, ensembles of these are also permissible, and they are likely to perform at least as well as the best of the individual models.[21]

1.5 Related Developments

The pace of development in this field is rapid, so it would not be possible to comprehensively cover its related literature in the space we have. Instead, we will simply point to a number of new techniques, sampled from this body of work, which are tied to the methods we have discussed.

For the estimation of average treatment effects (and restricted versions thereof), Athey et al. (2018) introduce the Approximate Residual Balancing (ARB) estimator. The method is split into two stages: first, the lasso is used to estimate the outcome model, which is assumed to be linear; then, the residuals from that are weighted and added back to the result from the first stage. Those weights are chosen to ensure that the distribution of the covariates for the treated and control units match closely (in-sample). The functional

[20] Note that this observation can be connected to the discussion in the paragraph above: if one of the models is estimated parametrically based on a relationship that is known to be true, the other model need only be consistent, since the product of the two rates would be $o_p\big(n^{-1/2}\big)$.

[21] Laan et al. (2006) provide asymptotic justifications for weighted combinations of estimators, particularly those which use cross-validation to calculate the weights.

form restriction they impose allows them to obtain a tight guarantee on the model's finite-sample bias. Furthermore, the model is consistent even when the propensity score is very dense. In fact, their asymptotic results continue to hold when the propensity score cannot be consistently estimated. This requires strong sparsity of the outcome model ($s_\mu \ll \sqrt{n}$), though. Thus, it is a complement to the DML.

Ning et al. (2018) developed a related method, which they call the High-Dimensional Covariate Balancing Propensity Score estimator, which is a modified version of the Horvitz and Thompson (1952) method. Based on an adapted version of the lasso with a quasi-likelihood, they first estimate the coefficients in the propensity score model. Then, they use a weighted version of the lasso to estimate the outcome model for the treated and control groups separately. For the variables selected in the second step, they find calibrated new coefficients for the propensity score model using a quadratic programme which ensures the covariate distributions are balanced (as above). Those new coefficients are layered over (i.e., replace) the corresponding set from the original propensity score model, fit in the first step. Finally, with the modified model for $e(\hat{X}_i)$, the fitted probability of treatment is obtained, and that is used to weight the outcomes of the treated and control units. Aside from being $\sqrt{n}$-consistent, asymptotically normal and semiparametrically efficient (qualities it shares with the DML and ARB), it also possesses the sample-boundedness property: loosely, the estimated treatment effect is guaranteed to be reasonable because the estimate of each component of the target (e.g., $E[Y_i(1)]$) must lie within the range of outcomes observed in the data. Notably, the authors also incorporate sample-splitting and find that the level of sparsity required of the outcome and propensity score models matches that of the DML exactly.

Farrell et al. (2019) focus on deep neural networks and develop the theory necessary to justify their use for causal inference. Specifically, they demonstrate that the most popular variant of the models at present, multilayer perceptrons with rectified linear activation functions, can be used to estimate the nuisance parameters as set out above at the required rate under appropriate smoothness conditions, and subsequently, one can conduct valid inference on treatment effects with the estimator.

Finally, Wager and Athey (2018) developed new results for an adaptation of random forests (that they call causal forests) which show that they can consistently estimate conditional average treatment effects. That is, they can be used to estimate average treatment effects for a given value of x. Moreover, provided that the subsamples used for the estimation of each tree are large enough, the method is asymptotically normal and unbiased. Pointwise confidence intervals can thus be constructed: they develop results for an infinitesimal jackknife estimator for the variance of the forest which is also consistent. Results from a simulation study they conduct demonstrate the power of the theory, but they also highlight a number of features of the

method which require further study to improve performance, including the splitting rules used for the trees.

There is a vast pool of work which extends these methods and others in important and relevant directions. The activity in this field and rapid progress made clearly indicate its importance, and we expect that the insights gleaned from this research will become increasingly important for applied economic research.

References

Ahrens, A., C. B. Hansen, and M. E. Schaffer (2019): "lassopack: Model selection and prediction with regularized regression in Stata," .

Altonji, J. G. and L. M. Segal (1996): "Small-Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business Economic Statistics*, 14, 353–366.

Angrist, J. D. and A. B. Krueger (1995): "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics*, 13, 225–235.

Angrist, J. D. and J.-S. Pischke (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," *Journal of Economic Perspectives*, 24, 3–30.

Athey, S. and G. W. Imbens (2017): "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, 31, 3–32.

——— (2019): "Machine Learning Methods That Economists Should Know About," *Annual Review of Economics*, 11, 685–725.

Athey, S., G. W. Imbens, and S. Wager (2018): "Approximate residual balancing: debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 597–623.

Bansak, K., J. Ferwerda, J. Hainmueller, A. Dillon, D. Hangartner, D. Lawrence, and J. Weinstein (2018): "Improving refugee integration through data-driven algorithmic assignment," *Science*, 359, 325–329.

Begun, J. M., W. J. Hall, W.-M. Huang, and J. A. Wellner (1983): "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *The Annals of Statistics*, 11, 432–452.

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

Belloni, A. and V. Chernozhukov (2011): "High Dimensional Sparse Econometric Models: An Introduction," in *Inverse Problems and High-Dimensional Estimation SE - 3*, ed. by P. Alquier, E. Gautier, and G. Stoltz, Springer Berlin Heidelberg, Lecture Notes in Statistics, 121–156.

——— (2013): "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19, 521–547.

Belloni, A., V. Chernozhukov, I. Fernandez-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 (2011): "Inference for High-Dimensional Sparse Econometric Models," .

——— (2014a): "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28, 29–50.

——— (2014b): "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, 81, 608–650.

Belloni, A., V. Chernozhukov, C. Hansen, and D. Kozbur (2016): "Inference in High Dimensional Panel Models with an Application to Gun Control," *Journal of Business & Economic Statistics*, 34, 590–605.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732.

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," *The Econometrics Journal*, 21, C1–C68.

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

Chetverikov, D., Z. Liao, and V. Chernozhukov (2019): "On cross-validated Lasso," .

D'Amour, A., P. Ding, A. Feller, L. Lei, and J. Sekhon (2019): "A Gaussian Process Framework for Overlap and Causal Effect Estimation with High-Dimensional Covariates," *arXiv:1711.02582v3 [math.ST]*.

Deaton, A. and N. Cartwright (2018): "Understanding and misunderstanding randomized controlled trials," *Social Science & Medicine*, 210, 2–21.

Farrell, M. H., T. Liang, and S. Misra (2019): "Deep Neural Networks for Estimation and Inference," .

Fisher, R. A. (1925): *Statistical Methods for Research Workers*, Oliver and Boyd Ltd., 5 ed.

——— (1935): *The Design of Experiments*, Hafner Publishing Company, 8 ed.

Gelman, A. and E. Loken (2013): "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time," `http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf`.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315.

Hastie, T., R. Tibshirani, and M. J. Wainwright (2015): *Statistical Learning with Sparsity: The Lasso and Generalizations*, Monographs on Statistics & Applied Probability, Boca Raton: CRC Press, Taylor & Francis.

Horvitz, D. G. and D. J. Thompson (1952): "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663.

Jing, B.-Y., Q.-M. Shao, and Q. Wang (2003): "Self-normalized Cramér-type large deviations for independent random variables," *The Annals of Probability*, 31, 2167–2215.

Kennedy, E. H. (2016): "Semiparametric Theory and Empirical Processes in Causal Inference," *arXiv:1510.04740v3 [math.ST]*.

Kiel, K. and K. McClain (1995): "House Prices during Siting Decision Stages: The Case of an Incinerator from Rumor through Operation," *Journal of Environmental Economics and Management*, 28, 241–255.

Laan, M. J. v. d., S. Dudoit, and A. W. v. d. Vaart (2006): "The cross-validated adaptive epsilon-net estimator," *Statistics Decisions*, 24.

Luo, Y. and M. Spindler (2019): "High-Dimensional L2 Boosting: Rate of Convergence," *arXiv:1602.08927v2 [stat.ML]*.

Mullainathan, S. and J. Spiess (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106.

Neyman, J. (1990): "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Translated by D. M. Dabrowska and T. P. Speed," *Statistical Science*, 5, 465–472.

Ning, Y., S. Peng, and K. Imai (2018): "Robust Estimation of Causal Effects via High-Dimensional Covariate Balancing Propensity Score," *Unpublished*.

Powell, J. (1994): *Estimation of Semiparametric Models*, Elsevier Science B.V.

Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011): "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366.

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

Vaart, A. W. v. d. and J. A. Wellner (1996): *Weak convergence and empirical processes*, Springer.

Wager, S. and S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242.

Wager, S. and G. Walther (2019): "Adaptive concentration of regression trees, with application to random forests," *arXiv:1503.06388 [math.ST]*.

Wooldridge, J. M. (2005): "Violating Ignorability of Treatment by Controlling for Too Many Factors," *Econometric Theory*, 21, 1026–1028.

——— (2009): *Introductory Econometrics: A Modern Approach*, Cengage, 4 ed.

——— (2010): *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press, 2 ed.