Draft Nr.1: Parametric Inference Problems Tackled By *Nonparametric Statistical Learning Methods*

ALEXANDER KIEL[1,2,3]

[1] *Erasmus School of Economics, Erasmus University, Burgemeester Oudlaan 50, Rotterdam, 3062PA, Netherlands*
[2] *Bachelor Thesis, BSc² Econometrics/Economics*
[3] *Corresponding author: 386879ak@student.eur.nl*

Compiled June 3, 2018

We aim on providing a general modern framework for the application of *nonparametric statistical learning techniques* in the presence of a high-dimensional nuisance parameter set, potentially causing a distortion of the *treatment effect* θ_0, the parameter of interest in various modern economic research applications. Under this context, those branded *"machine learning techniques"* are compared to traditional regression results such as IV estimations. In order to empirically evaluate the aforementioned nonparametric framework we consult traditional results from the research paper "The Slave Trade and the Origins of Mistrust in Africa".

https://github.com/C-o-r/Inference_DML

1. INTRODUCTION

Our goal is to verify and potentially provide a general "machine learning" framework on doing inference about a low-dimensional causal parameter, framed in this context a policy or treatment variable, in the presence of a high-dimensional nuisance parameter. This high-dimensional nuisance parameter η_0 can potentially cause a distortion of the treatment effect θ_0.

Valid inferential statements about a low-dimensional parameter of interest are of high importance. Chernozhukov et al. apply in their recent research a generation of nonparametric statistical methods to do estimation over this exact high-dimensional nuisance parameter set up, targeting θ_0 [1]. The examined *regularization bias* of traditional machine learning techniques occurring in high-dimensional settings which they aim to overcome via their contribution of Neyman-Orthogonality and Sample Splitting (k-cross validation) will be thoroughly discussed in the *Methodology* of this paper and serves as the theoretical framework for our approach.

With the empirical validation of these debiased machine learning techniques we contribute to the findings of "traditional" nonparametric methods [2] such as kernel regression, IV-estimations or other techniques. The findings of this paper have different applications based on the underlying research field. This goes beyond the level of purely scientific relevance and can be applied in, for example, observational studies in the presence of a large amount of covariates or controls.

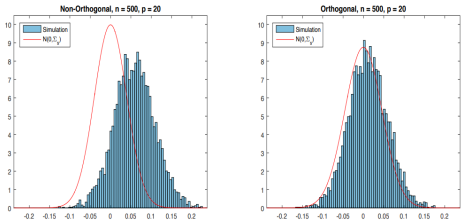

Fig. 1. Numerical illustration of the regularization bias phenomenon for a naive ML estimator based on a random forest in a simple computational experiment[1]

.

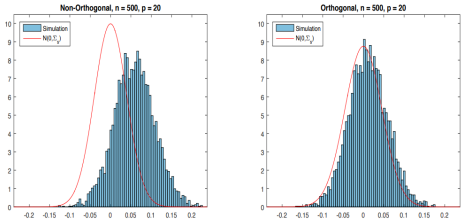

Fig. 2. WRONG PICTURE STILL: Numerical illustration of depiction on how sample splitting can eliminate overfitting based on the same computational experiment [1]

.

We aim on comparing traditional estimator findings of θ_0 to Chetverikov et al.'s techniques [1], by using equivalent underlying sample datasets. We analyze these "double biased machine learning" (DML) parameter estimates by constructing confidence intervals.

As our running example, we account hereby for regularization bias in the work of [3].While this paper adds to a new and growing literature in economics that seeks to better understand the role that culture, norms, and beliefs play in individual decision making, its result serve as a convenient object of investigation.

Chetverikov et al.'s double biased techniques are in this context compared with the traditional approaches (IV estimation) of the underlying research paper [3].

Chetverikov et al.'s research both builds upon and extends

the classic work in semiparametric estimation which focused on "traditional" nonparametric methods for estimating θ_0 [2]. Under certain conditions, these "traditional" methods require the estimators to take values in the Donsker set, a set whose complexity is bounded so the complexity of the functions does not increase with the sample size i.e. it limits the complexity of the function space. This vital assumption rules out most of the high-dimensional techniques. Chetverikov et al.'s nonparametric statistical learning methods extend this framework by going beyond the constraining Donsker set for the high-dimensional parameter η_0.

The confounder interaction caused by for example more than 100 covariates causes trouble in original OLS testing frameworks. Traditional semi-parametric techniques applying kernel functions, etc. can handle maximally 10-11 covariates and have gained throughout the concept of IV testing various popularity in research. Instead, the discussed DML methods are able to handle an enormous amount of covariates.

A high-dimensional nuisance parameter space allows to go beyond prior work done on the field of inference between nuisance parameters and treatment effect parameters. We extend this framework, contemporary framed as a "naive" or prediction-based machine learning approach, where we naively plug in ML estimators of η_0 into the estimating equation of θ_0. Despite the remarkably effectiveness of these approaches methods in predicting, the estimation and inference about "causal" parameters could be misleading. This regularized estimator has a non-trivial effect on the estimation of θ_0 caused by a regularization bias which originates from keeping the variance in this highly complex setting reasonably small.The naive estimator fails to be $N^{\frac{1}{2}}$ consistent. A canonical running example in [1] is the partially linear model.

We apply Neyman-orthogonal scores, as well as sample splitting (cross-fitting) in order to construct high quality points that concentrate in a $N^{-\frac{1}{2}}$ neighborhood of the true parameter value. With respect to the compelling motion of sample-splitting, we examine whether parameter estimates are robust to a particular amount of sample splits.

Secondly, we construct valid confidence statements of the "causal" parameters obtained from an IV-estimation with additional control variables, outlined in [3].

Furthermore, we compare the estimation results and give further insights into the algorithm structure of these machine learning techniques. The question of interest arises whether we obtain broadly consistent results regardless of which DML method we employ. Moreover, we analyze the complexity of these varying methods.

Under this framework, the following nonparametric statistical methods are consulted:

- Random Forests
- Boosted Trees
- Lasso
- Ridge
- Deep and Standard Neural Nets
- Aggregations and Cross-Hybrids
-

and outlined in the *Methodology*. In the Data section we introduce our dataset as well as the consulted 'traditional' models. In the *Methodology* we additionally outline the machine learning techniques, as well as the corresponding regularization bias + sample splitting we aim to overcome. Finally, we present in the result section the comparison of parameter estimators and outline the strength of each machine learning technique with respect to the runtime complexity, etc.. We conclude with potential extensions.

2. DATA

The dataset, replications files and most importantly results of "The Slave Trade and the Origins of Mistrust in Africa." [1] serve in this context as the running example which we aim compare with DML techniques. Therefore, we use replications files of "Double/Debiased Machine Learning for Treatment and Structural Parameters" [2]

To demonstrate DML estimation of IV equation models with instrumental variables we consider in a first step the work of [3] and the underlying dataset. The applied models in this paper contain a large amount of covariates which could lead to an erroneous inference on the "causal" parameter of interest. The paper establishes a causal relationship of less trusting individuals nowadays whose ancestors were heavily raided during transatlantic slave trade. Their results obtained via three different strategies are illustrated in Table 1-10 which we aim to analytically verify via our DML methods.

[] extend with their findings the studies of [**?**]. Nunn found that slavery had a "significant negative effect on long-term economic development". Even though he identifies a negative causal relationship between slave trade and income today, causal mechanisms remain unknown to the reader. [] stresses in the research their main finding, mainly that slave trade caused a culture of mistrust to develop within Africa based on heuristic decision- making strategies [3] describe slavery as an environment of ubiquitous insecurity which caused individuals to both turn on others including friends and family members and to kidnap, trick. Convincingly, [3] establishes the hypothesis that in this environment, a culture of mistrust may have evolved, which may persist to this day. The underlying heuristic decision- making strategy reasoning seems fairly one-sided and far-fetched. The paper does not discuss potentially other emotional feelings attached with the state of slavery such as potential protectiveness and care-taking among the slaves. Based and the results and conclusion, the question of concern arises whether slaves really become mistrusting solely due to the fact that they were betrayed and whether they pass this attribute among generations. The control variable have a vial function in this context which we aim to better account for via our new modern methods. It seems rather doubtful that the emotional state of the survey correspondent is largely based on the ancestors betrayal. The replication results, including the treatment effect coefficient estimators under question, can be found in the appendix and serve as a natural comparison for our DML techniques and results. EXPLAIN SPECIFIC PURPOSE OF EACH TABLE

[1] available via https://www.aeaweb.org/articles?id=10.1257/aer.101.7.3221
[2] available via https://github.com/VC2015/DMLonGitHub/

3. METHODOLOGY

The Methodology mainly builds upon [1] results and findings. We try to extend the framework by giving a deeper insight into the structure of the underlying machine learning algorithms. At first, we depict however the Neyman-orthogonality/ sample splitting approach.

A. Overcoming Regularization Bias

[1] depart from the classical setting by allowing η_0 to be so high-dimensional that the Donsker properties break down. Modern analyses aims to model p, the amount of confounding factors as increasing with the sample size. This causes traditional assumptions, limiting the complexity of the parameter space for the nuisance parameters to fail. The bias causing naive estimators failing to be $N_1/2$ consistent is the result of both the regularization bias and overfitting for estimators η_0 into estimating equations for θ_0.
Rephrase "Here, highly complex formally means that the entropy of the parameter space for the nuisance parameter is increasing with the sample size in a way that moves us outside of the traditional framework considered in the classical semi-parametric literature where the complexity of the nuisance parameter space is taken to be sufficiently small. Offering a general and simple procedure for estimating and doing inference on 0 that is formally valid in these highly complex settings is the main contribution of this paper."

The following two sub points outline the key ingredients for DML. Those rely on only weak theoretical requirements and can be applied to a list of modern ML methods, which shall be discussed in this section as well.

A.1. Neyman-orthogoanl moments/scores

In the partially linear model outlined by [1] the treatment effect estimator's ($\hat{\theta}_0$) rate of convergence is smaller than $n^{-1/2}$. This behavior is caused by the bias of the parameter coefficients g_0. Accordingly, the estimator of θ_0 does not converge

$$| \sqrt{n}(\hat{\theta}_0 - \theta_0) | \to \infty \tag{1}$$

Under this context, $\hat{\theta}_0$ is estimated in the following manner

$$\hat{\theta}_0 = (\frac{1}{n}\sum_{i\in I} D_i^2)^{-1} \frac{1}{n}\sum_{i\in I} D_i(Y_i - \hat{g}_0(X_i)) \tag{2}$$

However, by applying double prediction i.e. partialling out the effect of X on D where we make use of Neymann orthogonality and obtain the estimator

$$\check{\theta} = (\frac{1}{n}\sum_{i\in I} \hat{V}_i D_i)^{-1} \frac{1}{n}\sum_{i\in I} \hat{V}_i(Y_i - \hat{g}_0(X_i)) \tag{3}$$

Explain much more

A.2. cross-fitting as an efficient form of data-splitting

Trough sample splitting our remainder term will contrary to the inital estimator approach 0 in probability

$$\frac{1}{\sqrt{n}}\sum_{i\in I}(\hat{g}_0(X_i) - g_0(X_i))^2 \to 0 \tag{4}$$

Explain much more

B. Debiased Machine Learning techniques

In this section we shortly depict the underlying idea of each machine learning method both mathematically and algorithmically:

B.1. Random Forests: library(randomForest)

First proposed by Tin Kam Ho, random forests came into the spotlight in 2001 after their description by Breiman and serve as an extension to the original decision trees. These supervised ensemble-learning begging models based on low-bias decision trees effectively work via bootstrap samples. By finding an optimal threshold and covariance for the d arbitrarily selected regressors in each step, we construct trees or models that are not correlated with each other (bootstrapping). The node splitting criteria is based on the a search for the best feature among a random subset of features.

Majority voting beneficially adds to the idea of overfitting and reduces the effect of outliers, etc. Averaging our results over decorrelated trees improves the predictive accuracy. Also used in the context of classification, we apply this procedure, instead, for our specific regression context. Pre-pruning options should be specified e.g. depth (vital for the complexity) or to select the smallest subset that can be split. Recent research emphasizes the use of random forest and even suggests that asymptotic normality may hold
Potential advantages: - beneficial effects of majority voting Potential disadvantages: -Not easily interpretable

EXPLAIN MORE; MATH

Algorithm 1. Random Forrest Pseudocode

```
procedure RANDOM FORREST(t, n...)          ▷ Input variables
    i ← 0                                  ▷ # trees
    j ← 0                                  ▷ # nodes
    while i ≠ t do
        while j ≠ n do
            Randomly select k regressors from total m
            Among all k, find node d with best cov + treshold
            Split node using the best split
    Assess results by averaging trees over bootstrap samples
    return estimator
```

B.2. Boosted Trees/Gradient Boosting

EXPLAIN

B.3. Lasso

EXPLAIN

B.4. Ridge

EXPLAIN

B.5. Deep and Standard Neural Nets

Hidden layers in this method aim on capturing the relations between...

$$Z_{i,m} = \sigma(\alpha_{0m} + \alpha'_{1m} X_i), \quad for\ m = 1, ..., M \tag{5}$$

$$Y_i = \beta_0 + \beta'_1 Z_i + \varepsilon_i \tag{6}$$

$$\sum_{i=1}^{N}(Y_i - g(X_i, \alpha, \beta))^2 \tag{7}$$

$$\lambda(\sum_{k,m}\alpha_{k,m}^2 + \sum_k \beta_k^2) \tag{8}$$

EXPLAIN A LOT

B.6. Aggregations and Cross-Hybrids

C. Algorithm Complexity

D. Comparison Procedure

We examine results from both 2-fold cross-fitting and 5-fold cross-fitting. Moreover, we construct standard errors both across the splits, as well as calculated by using the median method.

4. RESULTS

5. PAPER 1

A. Table 1-3

B. Table 5-6

C. Table 9-10

6. PAPER 2

7. COMPLEXITY ANALYSIS

8. CONCLUSION

REFERENCES

1. V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, The Econom. J. **21**, C1 (2018).
2. A. Van Der Vaart, The Annals Stat. pp. 178–204 (1991).
3. N. Nunn and L. Wantchekon, Am. Econ. Rev. **101**, 3221 (2011).

9. APPENDIX

Table 1. Table 1- OLS Estimates of the Determinants of Trust in Neighbors

Dependent variable: Trust of neighbors	Slave exports (thousands) (1)	Exports/ area (2)	Exports/ historical pop (3)	ln(1+exports) (4)	ln(1+exports/area) (5)	ln(1+exports/historical pop) (6)
Estimated coeffient	-0.00068	-0.019	-0.531	-0.037	-0.159	-0.743
std adj. clust	0.00014	0.005	0.147	0.014	0.034	0.187
std adj. 2-clust	0.00015	0.005	0.147	0.014	0.034	0.187
std adj. spatial	0.00013	0.005	0.165	0.015	0.034	0.212
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
District controls	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	20,027	20,027	17,644	20,027	20,027	17,644
Number of ethnicities	185	185	157	185	185	157
Number of districts	1,257	1,257	1,214	1,257	1,257	1,214
R^2	0.16	0.16	0.15	0.15	0.16	0.15

Table 2. Table 2— OLS Estimates of the Determinants of the Trust of Others

	Trust of relatives	Trust of neighbors	Trust of local council	Intra group trust	Inter-group trust
	(1)	(2)	(3)	(4)	(5)
ln(1+exports/area)	-0.133	-0.159	-0.111	-0.144	-0.097
std adj. 2-clust	0.037	0.034	0.021	0.032	0.028
Individual controls	Yes	Yes	Yes	Yes	Yes
District controls	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Number of observations	20,062	20,027	19,733	19,952	19,765
Number of ethnicity clusters	185	185	185	185	185
Number of district clusters	1,257	1,257	1,283	1,257	1,255
R^2	0.13	0.16	0.20	0.14	0.11

Table 3. Table 3— OLS Estimates of the Determinants of the Trust of Others, with Additional Controls

	Trust of relatives	Trust of neighbors	Trust of local council	Intra group trust	Inter-group trust
	(1)	(2)	(3)	(4)	(5)
ln(1+exports/area)	-0.178	-0.202	-0.129	-0.188	-0.115
std adj. 2-clust	0.032	0.031	0.022	0.033	0.030
Colonial population density	Yes	Yes	Yes	Yes	Yes
Ethnicity-level colonial controls	Yes	Yes	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes	Yes	Yes
District controls	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Number of observations	16,709	16,679	15,905	16,636	16,473
Number of ethnicity clusters	147	147	146	147	147
Number of district clusters	1,187	1,187	1,194	1,186	1,184
R^2	0.13	0.16	0.21	0.16	0.12

Table 4. Table 5— IV Estimates of the Effect of the Slave Trade on Trust

	Trust of relatives	Trust of neighbors	Trust of local council	Intragroup trust
	(1)	(2)	(3)	(4)
Second stage: Dependent variable is an individual's trust				
ln(1+exports/area)	-0.190	-0.245	-0.221	-0.251
std adj. 2-clust	0.067	0.070	0.060	0.088
Hausman test (p-value)	0.88	0.53	0.09	0.44
R^2	0.13	0.16	0.20	0.15
First stage: Dependent variable is ln(1+exports/area)				
Historical distance of ethnic group from coast	-0.0014	-0.0014	-0.0014	-0.0014
std	0.0003	0.0003	0.0003	0.0003
Colonial population density	Yes	Yes	Yes	Yes
Ethnicity-level colonial controls	Yes	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes	Yes
District controls	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes
Number of observations	16,709	16,679	15,905	16,636
Number of (ethnicity) clusters	147/1,187	147/1,187	146/1,194	147/1,186
F-stat of excl. instrument	26.9	26.8	27.4	27.1
R^2	0.81	0.81	0.81	0.81

Table 5. Table 6— IV Estimates of the Effect of the Slave Trade on Trust, with Additional Controls

	Trust of relatives	Trust of neighbors	Trust of local council	Intragroup trust
	(1)	(2)	(3)	(4)
Second stage: Dependent variable is an individual's trust				
ln(1+exports/area)	-0.172	-0.271	-0.262	-0.254
std adj. 2-clust	0.076	0.088	0.075	0.109
Hausman test (p-value)	0.98	0.42	0.05	0.53
R^2	0.13	0.16	0.20	0.15
First stage: Dependent variable is ln(1+exports/area)				
Historical distance of ethnic group from coast	-0.0015	-0.0015	-0.0015	-0.0015
std	0.0003	0.0003	0.0003	0.0003
Colonial population density	Yes	Yes	Yes	Yes
Ethnicity-level colonial controls	Yes	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes	Yes
District controls	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes
Number of observations	16,709	16,679	15,905	16,636
Number of (ethnicity) clusters	147/1,187	147/1,187	146/1,194	147/1,186
F-stat of excl. instrument	21.7	21.6	22.2	21.8
R^2	0.81	0.81	0.81	0.81