Double/Debiased/Neyman Machine Learning of Treatment Effects

By Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey*

Chernozhukov et al. (2016) provide a generic double/de-biased machine learning (ML) approach for obtaining valid inferential statements about focal parameters, using Neyman-orthogonal scores and cross-fitting, in settings where nuisance parameters are estimated using ML methods. In this note, we illustrate the application of this method in the context of estimating average treatment effects (ATE) and average treatment effects on the treated (ATTE) using observational data. Empirical illustrations and code are available as supplementary material to this paper, and more general discussion and references to the existing literature are available in Chernozhukov et al. (2016).

I. Scores for Average Treatment Effects

We consider estimation of ATE and ATTE under the unconfoundedness assumption of Rosenbaum and Rubin (1983). We consider the case where treatment effects are fully heterogeneous and the treatment variable, D, is binary, $D \in \{0, 1\}$. We let Y denote the outcome variable of interest and Z denote a set of control variables.

* Chernozhukov: Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02142 (email: vchern@mit.edu); Chetverikov: University of California Los Angeles, 315 Portola Plaza, Los Angeles, CA 90095 (email: chetverikov@econ.ucla.edu). Demirer: Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02142 (email: mdemirer@mit.edu); Duflo: Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02142 (email: duflo@mit.edu); Hansen: University of Chicago, 5807 S. Woodlawn Ave., Chicago, IL 60637 (email: chansen1@chicagobooth.edu); Newey: Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02142 (email: wnewey@mit.edu). This material is based upon work supported by the National Science Foundation under Grant No. 1558636.

We then model random vector (Y, D, Z) as

$$(1) \quad Y = g_0(D, Z) + \zeta, \quad \mathrm{E}[\zeta \mid Z, D] = 0,$$
$$(2) \quad D = m_0(Z) + \nu, \quad \mathrm{E}[\nu \mid Z] = 0.$$

Since D is not additively separable, this model allows for very general heterogeneity in treatment effects. Common target parameters θ_0 in this model are the ATE,

$$\theta_0 = \mathrm{E}[g_0(1, Z) - g_0(0, Z)],$$

and the ATTE,

$$\theta_0 = \mathrm{E}[g_0(1, Z) - g_0(0, Z)|D = 1].$$

The confounding factors Z affect the treatment variable D via the propensity score, $m_0(Z) := \mathrm{E}[D|Z]$, and the outcome variable via the function $g_0(D, Z)$. Both of these functions are unknown and potentially complicated, and we consider estimating these functions via the use of ML methods.

We proceed to set up moment conditions with scores that obey a type of orthogonality with respect to nuisance functions. Specifically, we make use of scores $\psi(W, \theta, \eta)$ that satisfy the identification condition

$$(3) \qquad \mathrm{E}\psi(W, \theta_0, \eta_0) = 0,$$

and the *Neyman orthogonality condition*

$$(4) \qquad \partial_\eta \mathrm{E}\psi(W, \theta_0, \eta)\Big|_{\eta=\eta_0} = 0$$

where $W = (Y, D, Z)$, θ_0 is the parameter of interest, and η denotes nuisance functions with population value η_0.

Using moment conditions that satisfy (4) to construct estimators and inference procedures that are robust to small mistakes in

nuisance parameters has a long history in statistics, e.g. Neyman (1959). Using moment conditions that satisfy (4) is also crucial to developing valid inference procedures for θ_0 after using ML methods to produce estimators $\hat{\eta}$ as discussed, e.g., in Chernozhukov, Hansen and Spindler (2015). In practice, estimation of θ_0 will be based on the empirical analog of (3) with η_0 replaced by $\hat{\eta}_0$, and the Neyman orthogonality condition (4) ensures sufficient insensitivity to this replacement that high-quality inference for θ_0 may be obtained. The second *critical ingredient*, that enables the use of wide array of modern ML estimators is *data splitting*, as discussed in the next section.

Neyman-orthogonal scores are readily available for both the ATE and ATTE – they turn out to be the doubly robust/efficient scores of Robins and Rotnitzky (1995) and Hahn (1998). For estimating the ATE, we employ

$$\begin{aligned}\psi(W,\theta,\eta) &:= g(1,Z) - g(0,Z) \\ &+\tfrac{D(Y-g(1,Z))}{m(Z)} - \tfrac{(1-D)(Y-g(0,Z))}{1-m(Z)} \\ &-\theta, \quad \text{with}\end{aligned} \tag{5}$$

with

$$\begin{aligned}\eta(Z) &:= (g(0,Z), g(1,Z), m(Z))', \\ \eta_0(Z) &:= (g_0(0,Z), g_0(1,Z), m_0(Z))',\end{aligned}$$

where $\eta(Z)$ is the nuisance parameter with true value denoted by $\eta_0(Z)$ consisting of P-square integrable functions, for P defined in Assumption II.1, mapping the support of Z to $\mathbb{R}\times\mathbb{R}\times(\varepsilon, 1-\varepsilon)$ where $\varepsilon > 0$ is a constant. For estimation of ATTE, we use the score

$$\begin{aligned}\psi(W,\theta,\eta) &:= \tfrac{D(Y-g(0,Z))}{m} \\ &-\tfrac{m(Z)(1-D)(Y-g(0,Z))}{(1-m(Z))m} - \theta\tfrac{D}{m},\end{aligned} \tag{6}$$

with

$$\begin{aligned}\eta(Z) &:= (g(0,Z), g(1,Z), m(Z), m)', \\ \eta_0(Z) &:= (g_0(0,Z), g_0(1,Z), m_0(Z), \mathrm{E}[D])',\end{aligned}$$

where again $\eta(Z)$ is the nuisance parameter with true value denoted by $\eta_0(Z)$ consisting of three P-square integrable functions, for P defined in Assumption II.1, mapping the support of Z to $\mathbb{R}\times\mathbb{R}\times(\varepsilon, 1-\varepsilon)$ and a constant $m \in (\varepsilon, 1-\varepsilon)$. The respective scores for ATE and ATTE obey the identification condition (3) and the Neyman orthogonality property (4). Note that all semi-parametrically efficient scores share the orthogonality property (4). Moreover, the use of efficient scores could be considerably refined using the targeted maximum likelihood approach of van der Laan and Rubin (2006) in many contexts.

II. Algorithm and Result

We describe the estimator of θ_0 using random sample $(W_i)_{i=1}^N$. The algorithm makes use of a form of sample splitting, which we call cross-fitting. It builds on the ideas e.g. in Angrist and Krueger (1995). The use of sample-splitting is a *crucial ingredient* to the approach that helps avoid overfitting which can easily result from the application of complex, flexible methods such as boosted linear and tree models, random forests, and various ensemble and hybrid ML methods.

Algorithm: Estimation by K-fold Cross-Fitting

Step 1. Let K be a fixed integer. Form a K-fold random partition of $\{1, ..., N\}$ by dividing it into equal parts $(I_k)_{k=1}^K$ each of size $n := N/K$, assuming that N is a multiple of K. For each set I_k, let I_k^c denote all observation indices that are not in I_k.

Step 2. Construct K estimators

$$\check{\theta}_0(I_k, I_k^c), \quad k = 1, ..., K,$$

that employ the machine learning estimators

$$\begin{aligned}\hat{\eta}_0(I_k^c) = \Big(&\hat{g}_0(0, Z; I_k^c), \hat{g}_0(1, Z; I_k^c), \\ &\hat{m}_0(Z; I_k^c), \frac{1}{N-n}\sum_{i\in I_k^c} D_i\Big)',\end{aligned}$$

of the nuisance parameters

$$\eta_0(Z) = (g_0(0,Z), g_0(1,Z), m_0(Z), \mathrm{E}[D])',$$

and where each estimator $\check{\theta}_0(I_k, I_k^c)$ is de-

fined as the root θ of

$$\frac{1}{n}\sum_{i\in I_k}\psi(W,\theta,\hat{\eta}_0(I_k^c))=0,$$

for the score ψ defined in (5) for the ATE and in (6) for the ATTE.

Step 3. Average the K estimators to obtain the final estimator:

$$\tilde{\theta}_0=\frac{1}{K}\sum_{k=1}^{K}\check{\theta}_0(I_k,I_k^c). \tag{7}$$

An approximate standard error for this estimator is $\hat{\sigma}/\sqrt{N}$, where

$$\hat{\sigma}^2=\frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_i^2,$$

$\hat{\psi}_i:=\psi(W_i,\tilde{\theta}_0,\hat{\eta}_0(I^c_{k(i)}))$, and $k(i):=\{k\in\{1,...,K\}:i\in I_k\}$. An approximate $(1-\alpha)\times 100\%$ confidence interval is

$$\mathrm{CI}_n:=[\tilde{\theta}_0\pm\Phi^{-1}(1-\alpha/2)\hat{\sigma}/\sqrt{N}].$$

We now state a formal result that provides the asymptotic properties of $\tilde{\theta}_0$. Let $(\delta_n)_{n=1}^{\infty}$ and $(\Delta_n)_{n=1}^{\infty}$ be sequences of positive constants approaching 0. Let c,ε,C and $q>4$ be fixed positive constants, and let K be a fixed integer.

ASSUMPTION II.1: *Let $\mathcal{P}$ be the set of probability distributions P for (Y,D,Z) such that (i) equations (1)-(2) hold, with $D\in\{0,1\}$, (ii) the following conditions on moments hold for all N and $d\in\{0,1\}$: $\|g(d,Z)\|_{P,q}\leq C$, $\|Y\|_{P,q}\leq C$, $P(\varepsilon\leq m_0(Z)\leq 1-\varepsilon)=1$, and $\|\zeta^2\|_{P,2}\geq c$, and (ii) the ML estimators of the nuisance parameters based upon a random subset I_k^c of $\{1,...,N\}$ of size $N-n$, obey the condition for all $N\geq 2K$ and $d\in\{0,1\}$: $\|\hat{g}_0(d,Z;I_k^c)-g_0(d,Z)\|_{P,2}\cdot\|\hat{m}_0(Z;I_k^c)-m_0(Z)\|_{P,2}\leq\delta_n n^{-1/2}$, $\|\hat{g}_0(d,Z;I_k^c)-g_0(d,Z)\|_{P,2}+\|\hat{m}_0(Z;I_k^c)-m_0(Z)\|_{P,2}\leq\delta_n$, and $P(\varepsilon\leq\hat{m}_0(Z;I_k^c)\leq 1-\varepsilon)=1$, with probability no less than $1-\Delta_n$.*

The assumption on the rate of estimating the nuisance parameters is a non-primitive condition. These rates of convergence are available for most often used ML methods and are case-specific, so we do not restate conditions that are needed to reach these rates. The conditions are not the tightest possible but are chosen for simplicity.

THEOREM II.1: *Suppose that the ATE, $\theta_0=E[g_0(1,Z)-g_0(0,Z)]$, is the target parameter and we use the estimator $\tilde{\theta}_0$ and other notations defined above. Alternatively, suppose that the ATTE, $\theta_0=E[g_0(1,Z)-g_0(0,Z)\mid D=1]$, is the target parameter and we use the estimator $\tilde{\theta}_0$ and other notations above. Consider the set $\mathcal{P}$ of probability distributions P defined in Assumption II.1. Then, uniformly in $P\in\mathcal{P}$, the estimator $\tilde{\theta}_0$ concentrates around θ_0 with the rate $1/\sqrt{N}$ and is approximately unbiased and normally distributed:*

$$\sigma^{-1}\sqrt{N}(\tilde{\theta}_0-\theta_0)\rightsquigarrow N(0,1),$$
$$\sigma^2=E[\psi^2(W,\theta_0,\eta_0(Z))],$$

and the result continues to hold if σ^2 is replaced by $\hat{\sigma}^2$. Moreover, confidence regions based upon $\tilde{\theta}_0$ have uniform asymptotic validity:

$$\sup_{P\in\mathcal{P}}|P(\theta_0\in CI_n)-(1-\alpha)|\to 0.$$

The scores ψ are the efficient scores, so both estimators are asymptotically efficient, in the sense of reaching the semi-parametric efficiency bound of Hahn (1998).

The proof, given in the online appendix, relies on the application of Chebyshev inequality and the central limit theorem.

III. Accounting for Uncertainty Due to Sample-Splitting

The method outlined in this note relies on subsampling to form auxiliary samples for estimating nuisance functions and main samples for estimating the parameter of interest. The specific sample partition has no impact on estimation results asymptotically but may be important in finite samples. Specifically, the dependence of the estimator on the particular split creates an additional source of variation. Incorporating a measure of this additional source of

variation into estimated standard errors of parameters of interest may be important for quantifying the true uncertainty of the parameter estimates.

Hence we suggest making a slight modification to the asymptotically valid estimation procedure detailed in Section II. Specifically, we propose repeating the main estimation procedure S times, for a large number S, repartitioning the data in each replication $s = 1, ..., S$. Within each partition, we then obtain an estimate of the parameter of interest, $\tilde{\theta}_0^s$. Rather than report point estimates and interval estimates based on a single replication, we may then report estimates that incorporate information from the distribution of the individual estimates obtained from the S different data partitions.

For point estimation, two natural quantities that could be reported are the sample average and the sample median of the estimates obtained across the S replications, $\tilde{\theta}_0^{\text{Mean}}$ and $\tilde{\theta}_0^{\text{Median}}$. Both of these reduce the sensitivity of the estimate for θ_0 to particular splits. $\tilde{\theta}_0^{\text{Mean}}$ could be strongly affected by any extreme point estimates obtained in the different random partitions of the data, and $\tilde{\theta}_0^{\text{Median}}$ is obviously much more robust. We note that asymptotically the specific random partition is irrelevant, and $\tilde{\theta}_0^{\text{Mean}}$ and $\tilde{\theta}_0^{\text{Median}}$ should be close to each other.

To quantify and incorporate the variation introduced by sample splitting, one might also compute standard errors that add an element to capture the spread of the estimates obtained across the S different sets of partitions. For $\tilde{\theta}_0^{\text{Mean}}$, we propose adding an element that captures the spread of the estimated $\tilde{\theta}_0^s$ around $\tilde{\theta}_0^{\text{Mean}}$. Specifically, we suggest

$$\hat{\sigma}^{\text{Mean}} = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\hat{\sigma}_s^2 + (\tilde{\theta}_0^s - \frac{1}{S}\sum_{j=1}^{S}\tilde{\theta}_0^j)^2\right)},$$

where $\hat{\sigma}_s$ is defined as in Section II. The second term in this formula takes into account the variation due to sample splitting which is added to a usual estimate of sampling uncertainty. Using this estimated standard error obviously results in more conservative inference than relying on the $\hat{\sigma}_s$ alone. We adopt a similar formulation for $\tilde{\theta}_0^{\text{Median}}$. Specifically, we propose a median deviation defined as

$$\hat{\sigma}^{\text{Median}} = \text{median}\Big\{\sqrt{\hat{\sigma}_i^2 + (\hat{\theta}_i - \hat{\theta}^{Median})^2}\Big\}_{i=1}^{S}.$$

This standard error is more robust to outliers than $\hat{\sigma}^{\text{Mean}}$.

REFERENCES

Angrist, J. D., and A. B. Krueger. 1995. "Split-Sample Instrumental Variables Estimates of the Return to Schooling." *Journal of Business and Economic Statistics*, 13(2): 225–235.

Chernozhukov, V., C. Hansen, and M. Spindler. 2015. "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach." *Annual Review of Economics*, 7: 649–688.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. K. Newey. 2016. "Double Machine Learning for Treatment and Causal Parameters." *ArXiv e-prints*.

Hahn, J. 1998. "On the role of the propensity score in efficient semiparametric estimation of average treatment effects." *Econometrica*, 66(2): 315–331.

Neyman, J. 1959. "Optimal asymptotic tests of composite statistical hypotheses." In *Probability and Statistics: The Harald Cramér Volume.* , ed. Ulf Grenander, 213–234. Almqvist and Wiksell.

Robins, James M., and Andrea Rotnitzky. 1995. "Semiparametric efficiency in multivariate regression models with missing data." *J. Amer. Statist. Assoc.*, 90: 122–129.

Rosenbaum, P. R., and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70(1): 41–55.

van der Laan, M. J., and D. Rubin. 2006. "Targeted maximum likelihood learning." *International Journal of Biostatistics*, 2(1): Article 11.