Improved Double Machine Learning Estimation for Multivariate Outcomes

Kyle Colangelo

University of California Irvine

July 27, 2020

Abstract

I propose a new approach to improve the double machine learning (DML) estimator for the case of multi-variate outcomes. The new approach takes advantage of the relationships between outcomes to improve efficiency. This is accomplished in two ways, first through the use of multi-task learning and transfer learning algorithms to learn the nuisance parameters η_0 more precisely. Second, for a given nuisance estimator we propose an averaging estimator which takes advantage of similarities between causal objects of interest. Various approaches are implemented in Monte Carlo simulations and empirical application which demonstrate the effectiveness of the new method.

Keywords:

JEL Classification: C14, C21, C31, C55

1 Introduction

Over the past few years it has been shown increasingly that a wide array of machine learning algorithms have applications not just in predictive modeling, but also in matters of causal inference. Of special note is the double machine learning (DML) estimator developed in Chernozhukov et al. (2018) which provides a general framework for the estimation of a variety of causal parameters in a high-dimensional setting via machine learning. There has been a growing literature surrounding DML, however no attention has been given to the case of multi-variate outcomes.

The primary goal of this paper can be stated as to investigate how the DML estimator can be refined to improve efficiency when considering multiple outcomes of interest. Two approaches and their combination are considered: First, by estimating the nuisance parameters using machine learning algorithms which take advantage of the relationships between outcome equations, namely transfer learning and multi-task learning. Second, by implementing a stein-type averaging estimator similar to Hansen (2016), which was extended to Seemingly Unrelated Regression (SUR) in Mehrabani and Ullah (2020).

Multitask Learning: In Caruana (1997) "Multitask Learning is an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias. It does this by learning tasks in parallel while using a shared representation; what is learned for each task can help other tasks be learned better."

Transfer Learning: In the machine learning literature Pratt (1993). In 2016, Andrew Ng, an industry leading machine learning and AI researcher, stated that transfer learning will be the next driver of industry success. Maurer et al. (2016) show the benefit of considering multiple outcomes concurrently with multi-task learning.

Naturally the potential gains will be widely varied depending on a number of factors.

In this paper I investigate extensions and usage of the DML estimator for the case of multi-variate outcomes of interest. Specifically, the non-parametric estimation of the

average treatment effect (ATE) and the slope coefficient in the partial linear model (PLM) are considered.

The new approach is analogous to the Seemingly Unrelated Regression (SUR) model in Zellner (1962).

As a further refinement, an averaging estimator similar to Mehrabani and Ullah (2020) is also derived an implemented.

Recent developments have shown that machine learning methods can be used effectively for a variety of causal inference problems. In particular, Chernozhukov et al. (2018) developed the Double/Debiased machine learning (DML) estimator, which is capable of estimating a variety of causal objects of interest with the proper construction of moment conditions. For example, the DML estimator allows for the estimation of the slope coefficient in the partial linear model, and also the average treatment effect (ATE) for the binary case (the ATE estimation has since been extended to the continuous case in Su et al. (2019) and Colangelo and Lee (2020)).

Consider an outcome of interest Y, treatment T (allowed to be multi-dimensional), covariate set X, and error ε. We may be interested in a nonparametric outcome equation

$$Y = g(T, X) + \varepsilon, \tag{1}$$

which also encompasses parametric equations as special cases (such as the partial linear model from Robinson (1988)).

In many applications we may wish to study the relationship between treatments and multiple different outcomes of interest. If we consider m outcomes of interest, we can extend

equation 1 for all m outcomes:

$$
\begin{aligned}
Y_1 &= g_1(T, X) + \varepsilon_1, \\
Y_2 &= g_2(T, X) + \varepsilon_2, \\
&\vdots \\
Y_m &= g_m(T, X) + \varepsilon_m,
\end{aligned}
$$

where each g_j is allowed to be a different function which may be dependent on a different subset of covariates. The m equations can be stacked together as

$$\mathbf{Y} = \mathbf{g}(T, X) + \varepsilon, \tag{2}$$

Recent developments in the machine learning literature have demonstrated the effectiveness of analyzing multiple outcomes of interest jointly. (see Borchani et al. (2015)) for a review

This paper not only contributes to the DML literature but also to the non-parametric and semi-parametric estimation literature. While there is a wealth of literature on Seemingly Unrelated Regression (SUR) and multiple equation models, there is very little research on non-parametric estimation of these models.

When the relationships between outcome equations are non-existent then performing DML equation by equation is optimal. However, in the case where the relationships between outcome equations are strong, efficiency gains can be had by considering them jointly, especially when there is substantial missing data for particular outcomes.

This paper is organized as follows: I introduce the framework, estimation procedure. Section 3 presents the relevant theoretical results. Section 4 discusses the Monte Carlo simulation results which show the effectiveness of the new estimator. Section 4 provides an empirical demonstration of the new method. Proofs of all results are provided in the appendix.

2 Framework

Assumption 1 (Random Sample)**.** (Y_i, T_i, X_i) *are i.i.d.*

Assumption 2 (Exogeneity)**.** $E(\varepsilon|X) = 0$

Assumption 3 (Error Variance)**.** $E(\varepsilon_i\varepsilon_j') = \sigma_{ij}I_m$ *(*$E(\varepsilon\varepsilon') = \Sigma \otimes I_m$*)*

where $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ & & \vdots & \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{bmatrix}$

Assumption 4 (Unconfoundedness)**.** $Y_{1i}, Y_{0i} \perp T_i|X_i$

3 Theory

Theorem 1 (Joint Normality-ATE))**.** *Under assumptions 1-3, and the assumptions in Chernozhukov et al. (2018) we have:*

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, V)$$

Where

$$V = \begin{bmatrix} E[\psi_{1i}^2] & E[\psi_{1i}\psi_{2i}] & \cdots & E[\psi_{1i}\psi_{mi}] \\ E[\psi_{2i}\psi_{1i}] & E[\psi_{2i}^2] & \cdots & E[\psi_{2i}\psi_{mi}] \\ & & \vdots & \\ E[\psi_{mi}\psi_{1i}] & E[\psi_{mi}\psi_{2i}] & \cdots & E[\psi_{mi}^2] \end{bmatrix}$$

$$\hat{V} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{1i}^2 & \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{1i}\hat{\psi}_{2i} & \cdots & \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{1i}\hat{\psi}_{mi} \\ \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{1i}\hat{\psi}_{2i} & \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{2i}^2 & \cdots & \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{2i}\hat{\psi}_{mi} \\ & & \vdots & \\ \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{1i}\hat{\psi}_{mi} & \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{2i}\hat{\psi}_{mi} & \cdots & \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{mi}^2 \end{bmatrix}$$

4 Simulations

In this section we evaluate the effectiveness of the previously discussed approaches to double machine learning for multiple outcomes. Two experiments are considered

Experiment 1: We consider two outcomes Y_1 and Y_2 which are fully observed. Nuisance parameters are estimated with multi-task learning.

Experiment 2: We consider the same two outcomes, but we allow for Y_2 to be missing for 80% of observations.

For both experiments we compare efficiency for the estimation of the average treatment

effect for Y_2. The following data generating process is used for both experiments:

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim N(0, \Sigma)$$

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

$$X \sim N(\mu, V_x)$$

$$\mu = (1, ..., 1)'$$

$$diag(V_x) = 1$$

$$V_{x,ij} = 0.5 \text{ for } |i - j| = 1$$

$$\theta_1 = \mathbf{1}$$

$$\theta_2 = \theta_1 + [\delta, \delta/2, \cdots, \delta/k]$$

$$\beta_1 = 1$$

$$\beta_2 = 1 + \delta$$

$$P(T = 1|X) = \Phi\left(\frac{X\theta_1 - \theta_1'\mu}{\theta_1' V_x \theta_1}\right)$$

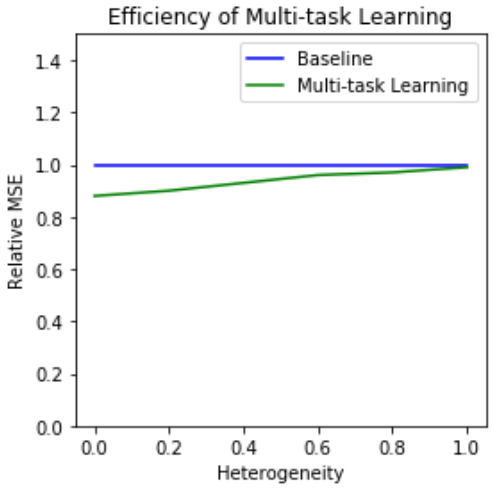
$$Y_1 = \beta_1 T + X\theta_1 + \varepsilon_1$$

$$Y_2 = \beta_2 T + X\theta_2 + \varepsilon_2$$

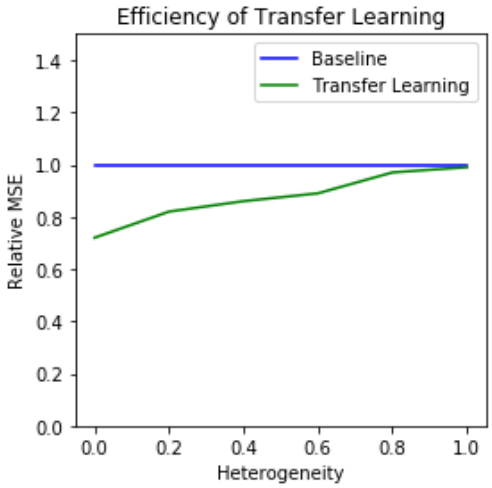
We allow δ to vary between $[0, 1]$

5 Empirical Application

6 Summary/Conclusion



(a) Relative efficiency of DML with multi-task Learning



(b) Relative efficiency of DML with transfer learning

Figure 1: Results for simulation experiments 1 and 2

References

Borchani, Hanen, Gherardo Varando, Concha Bielza, and Pedro Larrañaga, "A survey on multi-output regression," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2015, *5* (5), 216–233.

Caruana, Rich, "Multitask learning," *Machine learning*, 1997, *28* (1), 41–75.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/debiased machine learning for treatment and structural parameters," 2018.

Colangelo, Kyle and Ying-Ying Lee, "Double debiased machine learning nonparametric inference with continuous treatments," *arXiv preprint arXiv:2004.03036*, 2020.

Hansen, Bruce E, "Efficient shrinkage in parametric models," *Journal of Econometrics*, 2016, *190* (1), 115–132.

Maurer, Andreas, Massimiliano Pontil, and Bernardino Romera-Paredes, "The benefit of multitask representation learning," *The Journal of Machine Learning Research*, 2016, *17* (1), 2853–2884.

Mehrabani, Ali and Aman Ullah, "Improved Average Estimation in Seemingly Unrelated Regressions," *Econometrics*, 2020, *8* (2), 15.

Pratt, Lorien Y, "Discriminability-based transfer between neural networks," in "Advances in neural information processing systems" 1993, pp. 204–211.

Robinson, Peter M, "Root-N-consistent semiparametric regression," *Econometrica: Journal of the Econometric Society*, 1988, pp. 931–954.

Su, Liangjun, Takuya Ura, and Yichong Zhang, "Non-separable models with high-dimensional data," *Journal of Econometrics*, 2019, *212* (2), 646–677.

Zellner, Arnold, "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias," *Journal of the American statistical Association*, 1962, *57* (298), 348–368.