Automatic Debiased Machine Learning for Dynamic Discrete Choice

Whitney K. Newey

Econometric Society Dynamic Structural Economics Summer School

2023
University of Lassaune

INTRODUCTION

Many interesting objects depend on a regression or other first step.

–Discrete dynamic structural economic models depend on conditional choice probabilities.

–Average equivalent variation bounds depend on average demand.

–Average treatment effect (ATE) depends on average outcome given covariates and treatment and/or propensity score.

–Average policy effect depends on average outcome.

The regression may be high dimensional.

Choice probability may depend on many state variables.

Average demand may depend on many prices.

There may be many observed covariates for the average treatment effect or policy effect.

Machine learning provides good prediction with many regressors.

Methods include neural nets (NN), random forests, Lasso, boosting.

Give excellent predictions but biased by regularization and/or model selection.

Regularization refers controlling the variance of the estimator; for NN regularization methods include ridge penalty and not iterating gradient descent until convergence.

If "plug-in" machine learner into formula for parameter of interest the regularization bias "passes through" and gives poorly centered confidence intervals for the parameter.

Also, if "plug-in" learner with model selection then local mistakes under root-n alternatives lead to invalid confidence intervals parameters in root-n neighborhood; Leeb and Potscher (2005).

A solution to regularization and model selection biases is Neyman orthogonal moment functions for GMM.

Orthogonality means first step has no effect, to first order, on average moment function.

Orthogonal moment functions can be constructed by adding to identifying moment functions the influence function of the expected moment functions evaluated at the plim of the first step when distribution is unrestricted (i.e. under general misspecification.

This orthogonality is model free in only depending on the plim of the machine learner and not on a model.

Model free orthogonality leads to standard errors that are robust to misspecification; not true with other debiasing (e.g. based on estimating propensity scores in treatment effects.)

We also use cross-fitting, where first steps are estimated from different observations than used to construct sample moments.

Cross-fitting removes some other biases, only requires mean-square convergence rates for the machine learners, which is all we know about NN, and leads to remainders of smaller order in some settings.

Exposition here is from "Locally Robust Semiparametric Estimation," (2016, 2022, LR), Chernozhukov et al. (2022).

Orthogonal moment function depends on another unknown function α_0 in addition to the regression.

Can use orthogonality property to estimate α_0.

Leads to "automatic" methods of estimating α_0 that only use formulas for object of interest and do not require knowing a formula for α_0.

These automatic methods often give estimator of parameter of interest with better properties than estimator based on plugging into a formula for α_0, perhaps because they do not use inverses of high dimensional estimators.

See V. Chernozhukov, W.K. Newey, V. Quintas-Martinez, V. Syrgkanis (2022) "RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests," *Proceedings of the 39th International Conference on Machine Learning* 162;

https://arxiv.org/abs/2110.03031;

AutoDML improves on state of the art NN estimator of Average Treatment Effect using inverse of NN estimator of propensity score in Monte Carlo.

See also V. Chernozhukov, W.K. Newey, V. Quintas-Martinez, V. Syrgkanis (2021) "Automatic Debiased Machine Learning via Neural Nets for Generalized Linear Regression,"

https://arxiv.org/abs/2104.14737;

AutoDML developed here; previous paper is an application of this one.

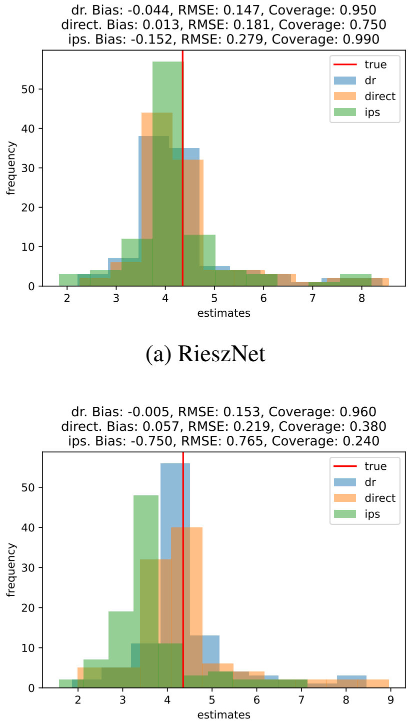
Table 1. RieszNet and ForestRiesz: Mean Absolute Error (MAE) and its standard error over 1000 semi-synthetic datasets based on the IHDP experiment.

(a) RieszNet

	MAE ± std. err.
Direct	0.123 ± 0.004
IPS	0.122 ± 0.037
DR	0.110 ± 0.003
Benchmark:	
Dragonnet (Shi et al., 2019)	0.146 ± 0.010

(b) ForestRiesz

	MAE ± std. err.
Direct	0.197 ± 0.007
IPS	0.669 ± 0.004
DR	0.126 ± 0.004
Benchmark:	
RF Plug-in (see text)	0.389 ± 0.024
CausalForest (Athey et al., 2019)	0.728 ± 0.028



(a) RieszNet

(b) ForestRiesz

Figure 2. RieszNet and ForestRiesz: Bias, RMSE, coverage and distribution of estimates over 100 semi-synthetic datasets based on the IHDP experiment, where we redraw T.

RMSE. On the other hand, the direct method (which does not use the debiasing term) seems to have lower bias for the RieszNet estimator, although in both cases its coverage is very poor. This is because the standard errors without the debiasing term greatly underestimate the true variance of the estimator.

5.2. Average Derivative in the BHP Gasoline Demand Data

To evaluate the performance of our estimators for average marginal effects of a continuous treatment, we conduct a semi-synthetic experiment based on gasoline demand data from Blundell et al. (2017) [BHP]. The dataset is constructed from the 2001 National Household Travel Survey, and contains 3,640 observations at the household level. The outcome of interest Y is (log) gasoline consumption. We want to estimate the effects of changing (log) price T, adjusting for differences in confounders X, including (log) household income, (log) number of drivers, (log) household respondent age, and a battery of geographic controls.

We generate our semi-synthetic data as follows. First, we estimate $\mu(X) := \mathbb{E}[T \mid X]$ and $\sigma^2(X) := \mathrm{Var}(T \mid X)$ by a Random Forest of T and $(T - \widehat{\mu}(X))^2$ on X, respectively. We then draw 3,640 observations of $T \sim \mathcal{N}(\widehat{\mu}(X), \widehat{\sigma}^2(X))$, and generate $Y = f(T, X) + \varepsilon$, for six different choices of $f(\cdot)$. The error term ϵ is drawn from a $\mathcal{N}(0, \sigma^2)$, with σ^2 chosen to guarantee that the simulated regression R^2 matches the one in the true data.

The exact form of f in each design is detailed in Appendix A.2. In the "simple f" designs we have a constant, homogeneous marginal effect of -0.6 (within the range of estimates in Blundell et al., 2012, using the real survey data). In the "complex f" designs, we have a regression function that is cubic in T, and where there are heterogeneous marginal effects by income (built to average approximately -0.6). In both cases, we evaluate the performance of the estimators without confounders X, and with confounders entering the regression function linearly and non-linearly.

Table 2 presents the results for the most challenging design: a complex regression function with linear and non-linear confounders (see Tables A1 and A2 in the Appendix for the full set of results in all designs). ForestRiesz with the doubly-robust moment combined with the post-processing TMLE adjustment (in which we use a corrected regression $\widetilde{g}(Z) = \widehat{g}(Z) + \epsilon \cdot \widehat{\alpha}(Z)$, where ϵ is the OLS coefficient of $Y - \widehat{g}(Z)$ on $\widehat{\alpha}(Z)$) seems to have the best performance in cases with many linear and non-linear confounders, with coverage close to or above the nominal confidence level (95%), and biases of around one order of magnitude lower than the true effect. As in the binary treatment case, the

Large sample theory only requires the machine learner converges faster than $n^{-1/4}$ in mean square and the product of mean-square convergence rates for the learner of α_0 and the regression learner is faster than $n^{-1/2}$.

LR gives regularity conditions for dynamic discrete choice with Lasso regression but not automatic estimator of α_0.

ORTHOGONAL MOMENT FUNCTIONS

θ : finite dimensional parameter vector of interest;

γ : unknown first step function, as possible realization of a machine learner;

W : data observation with unknown cumulative distribution function (CDF) F_0;

$g(w, \gamma, \theta)$: vector moment functions, w a possible realization w of W;

Moment condition

$$E[g(W, \gamma_0, \theta)] = 0,$$

θ_0 is assumed to be the unique solution this equation, i.e. θ_0 is identified by the moment condition, where $E[\cdot]$ is the expectation under the true distribution, and γ_0 is the probability limit of a first step estimator (machine learner) of γ.

Example 1: Simple and illustrative;

$$\theta_0 = E[Z\gamma_0(X)],\ \gamma_0(X) = E[Y|X].$$

Here the identifying moment function is

$$g(W,\gamma,\theta) = Z\gamma(X) - \theta.$$

In general the γ in $g(W,\gamma,\theta)$ allows the moment conditions to depend on the whole function γ.

In this example the moment function depends on γ through $\gamma(X)$.

Here the limit γ_0 of the machine learner is assumed to be the conditional expectation $E[Y|X]$ of Y given X.

Construct orthogonal moment functions using the plim $\gamma(F)$ of $\hat{\gamma}$ when F is distribution of a single observation W;

$\gamma(F)$ is plim of $\hat{\gamma}$ under general misspecification where F is unrestricted except for regularity conditions;

$\gamma_0 = \gamma(F_0)$.

Example 1 is for nonparametric regression $\hat{\gamma}$ where $\gamma(F)(x) = E_F[Y|X = x]$.

Let H be some alternative distribution that is unrestricted except for regularity conditions, and $F_\tau = (1-\tau)F_0 + \tau H$ for $\tau \in [0, 1]$.

Assume $\gamma(F_\tau)$ exists for τ small enough and possibly other regularity conditions are satisfied.

Key assumption is existence of unknown functions α and $\phi(w, \gamma, \alpha, \theta)$ such that for all H and θ,

$$\frac{d}{d\tau} E[g(W, \gamma(F_\tau), \theta)] = \int \phi(w, \gamma_0, \alpha_0, \theta) H(dw),$$
$$E[\phi(W, \gamma_0, \alpha_0, \theta)] = 0, \;\; E[\phi(W, \gamma_0, \alpha_0, \theta)^2] < \infty,$$

$$\frac{d}{d\tau}E[g(W,\gamma(F_\tau),\theta)] = \int \phi(w,\gamma_0,\alpha_0,\theta)H(dw),$$
$$E[\phi(W,\gamma_0,\alpha_0,\theta)] = 0,\ E[\phi(W,\gamma_0,\alpha_0,\theta)^2] < \infty,$$

α_0 is the α such that these equations hold; and $d/d\tau$ is the derivative from the right (i.e. for nonnegative values of τ) at $\tau = 0$.

The equations define $\phi(w,\gamma_0,\alpha_0,\theta)$ to be the *influence function* of $\mu(F) = E[g(W,\gamma(F),\theta)]$, as in Von Mises (1947), Hampel (1974), and Huber (1981).

$\phi(w,\gamma_0,\alpha_0,\theta)$ is unique because H is unrestricted.

Existence of $\phi(w,\gamma_0,\alpha_0,\theta)$ is equivalent to finite semiparametric variance bound for $\mu(F)$.

Here γ_0 and α_0 can depend on θ; equations hold for each $\theta \in \Theta$; F_0 and H do not depend on θ.

We refer to $\phi(w,\gamma,\alpha,\theta)$ as the *first step influence function* (FSIF); is "adjustment term" of Newey (1994).

Example 1: Here $g(W, \gamma, \theta) = Z\gamma(X) - \theta$ and we assume that $\gamma(F)$ is $E_F[Y|X]$, so that

$$E[g(W, \gamma(F), \theta)] = E[ZE_F[Y|X]] - \theta.$$

In this example

$$\phi(W, \gamma, \alpha, \theta) = \alpha(X)\{Y - \gamma(X)\},\ \alpha_0(X) = E[Z|X].$$

Follows from Newey (1994, "Asymptotic Variance of Semiparametric Estimators," EMA).

Construct orthogonal moment functions by adding FSIF to identifying moment functions

$$\psi(W, \gamma, \alpha, \theta) = g(W, \gamma, \theta) + \phi(W, \gamma, \alpha, \theta).$$

LR shows two key orthogonality properties:

Property I : For a linear set $\Gamma \supseteq \{\gamma(F)\}$ of possible first steps,

$$\frac{d}{d\delta} E[\psi(W, \gamma_0 + \delta(\gamma - \gamma_0), \alpha_0, \theta)] = 0 \text{ for all } \gamma \in \Gamma \text{ and } \theta \in \Theta$$

where $\delta \in \mathbb{R}$ is a scalar and the derivative is evaluated at $\delta = 0$.

Varying γ has no effect (locally to γ_0) on $E[\psi(W, \gamma, \alpha_0, \theta)]$ for $\gamma \in \Gamma$.

Property II : For the set $\mathcal{A}$ of α_0 such that influence function equation is satisfied

$$E[\phi(W, \gamma_0, \alpha, \theta)] = 0 \text{ for all } \theta \in \Theta \text{ and } \alpha \in \mathcal{A}.$$

Implies $E[\psi(W, \gamma_0, \alpha, \theta_0)] = 0$ for all $\alpha \in \mathcal{A}$, so orthogonal moment function $\psi(W, \gamma_0, \alpha, \theta_0)$ is globally robust in α.

In general there are distinct $\phi(W, \gamma, \alpha, \theta)$ and α for each element of $g(W, \gamma, \theta)$.

Example 1: Orthogonal moment function is

$$\psi(W,\gamma,\alpha,\theta) = Z\gamma(X) - \theta + \alpha(X)\{Y - \gamma(X)\}.$$

Property II follows by iterated expectations, which gives

$$E[\phi(W,\gamma_0,\alpha,\theta)] = E[\alpha(X)\{Y - E[Y|X\}] = 0.$$

To show Property I, note that by iterated expectations,

$$\begin{aligned} E[\psi(W,\gamma,\alpha_0,\theta)] &= E[Z\gamma(X)] - \theta + E[\alpha_0(X)\{Y - \gamma(X)\}] \\ &= E[\alpha_0(X)\gamma(X)] - \theta + E[\alpha_0(X)Y] - E[\alpha_0(X)\gamma(X)] \\ &= E[\alpha_0(X)Y] - \theta. \end{aligned}$$

Here $E[\psi(W,\gamma,\alpha_0,\theta)]$ does not depend on γ, i.e. it is globally robust to γ, and hence directional derivative is zero and have property I.

Here $\psi(W,\gamma,\alpha,\theta)$ is an example of *doubly robust* moment function, meaning global robustness in γ as well as α.

Necessary and sufficient condition for double robustness is $E[\psi(W,\gamma,\alpha_0,\theta)]$ is linear (affine) in γ; see LR for more discussion and many examples.

Orthogonal moment functions are

$$\psi(W,\gamma,\alpha,\theta) = g(W,\gamma,\theta) + \phi(W,\gamma,\alpha,\theta).$$

Orthogonality sometimes equated with semiparametric efficient influence functions.

$\psi(W,\gamma,\alpha,\theta)$ is nonparametric efficient influence function for $E_F[g(W,\gamma(F),\theta)]$;

This moment function accounts fully for behavior of first step through $\gamma(F)$;

Orthogonality characterization here useful because $g(W,\gamma,\theta)$ and $\phi(W,\gamma,\alpha,\theta)$ have distinct roles as identifying moment functions and bias correction, leading to Property II: $E[\phi(W,\gamma_0,\alpha,\theta)] = 0$ for all α, θ.

Orthogonality property here is model free and robust to misspecification.

Standard errors robust to misspecification for $g(W,\gamma,\theta)$ with same dimension as θ.

$$\psi(W,\gamma,\alpha,\theta) = g(W,\gamma,\theta) + \phi(W,\gamma,\alpha,\theta).$$

Many formulae for $\phi(W,\gamma,\alpha,\theta)$ are available.

Newey (1994) gives formulae when γ is

–conditional means other nonparametric least squares projections.

–conditional pdfs.

Ichimura and Newey (2022, "The Influence Function of Semiparamtric Estimators", Quantitative Economics) gives formulae when γ is

–conditional quantiles;

–conditional expectiles;

–other conditional location and scale measures;

–projection versions of conditional location and scale;

–solutions to conditional moment restrictions-

CROSS-FITTING

Combine orthogonal moment functions with cross-fitting, a form of sample splitting, to construct debiased sample moment functions; e.g. see Bickel (1982), Schick (1986), Klaassen (1987), and Chernozhukov et al. (2018).

Partition the observation indices $(i = 1, ..., n)$ into L groups I_ℓ, $(\ell = 1, ..., L)$.

Let $\hat{\gamma}_\ell$, $\hat{\alpha}_\ell$, and $\tilde{\theta}_\ell$ be estimators that are constructed using all observations *not* in I_ℓ. Debiased sample moment functions are

$$\hat{\psi}(\theta) = \hat{g}(\theta) + \hat{\phi},\ \hat{g}(\theta) = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} g(W_i, \hat{\gamma}_\ell, \theta),\ \hat{\phi} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} \phi(W_i, \hat{\gamma}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell).$$

$L = 5$ works well based on a variety of empirical examples and in simulations, for medium sized data sets; see Chernozhukov et al. (2018).

$L = 10$ works well for small data sets.

This cross-fitting a) eliminates "own observation" bias, like jackknife instrumental variables; b) helps remainders converge faster to zero, e.g. Newey and Robins (2017); c) eliminates need for Donsker conditions not satisfied by many machine learners

Debiased GMM is

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \hat{\psi}(\theta)' \hat{\Upsilon} \hat{\psi}(\theta),$$

where $\hat{\Upsilon}$ is a positive semi-definite weighting

As usual $\hat{\Upsilon}$ that minimizes the asymptotic variance of $\hat{\theta}$ is $\hat{\Upsilon} = \hat{\mathbf{\Psi}}^{-1}$, for

$$\hat{\mathbf{\Psi}} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_{i\ell} \hat{\psi}_{i\ell}', \ \hat{\psi}_{i\ell} = g(W_i, \hat{\gamma}_\ell, \tilde{\theta}_\ell) + \phi(W_i, \hat{\gamma}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell).$$

No need to account for the presence of $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ in $\hat{\psi}_{i\ell}$ because of orthogonality. See paper for initial estimator $\tilde{\theta}_\ell$.

An estimator $\hat{V}$ of the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$ is

$$\hat{V} = (\hat{G}' \hat{\Upsilon} \hat{G})^{-1} \hat{G}' \hat{\Upsilon} \hat{\mathbf{\Psi}} \hat{\Upsilon} \hat{G} (\hat{G}' \hat{\Upsilon} \hat{G})^{-1}, \ \hat{G} = \frac{\partial \hat{g}(\hat{\theta})}{\partial \theta}.$$

AUTOMATIC ESTIMATION of α_0

Need $\hat{\alpha}_\ell$ with plim α_0 for debiased GMM.

Can use Property I, orthogonality of $\psi(w, \gamma, \alpha_0, \theta)$ with respect to γ, to construct estimators of α_0 without knowing the form of α_0.

Chernozhukov, Newey, and Singh (2022) give AutoDML for Lasso regression.

Chernozhukov, Newey, Quintas-Martinez, Syrgkanis (2021, "Automatic Debiased Machine Learning via Neural Nets for Generalized Linear Regression," developed AutoDML for NN and other first steps.

Based on the observation that the orthogonality in Property I can sometimes be interpreted as first order conditions for minimizing an objective function.

An important case of this is where $\gamma_F(X) = E[Y|X]$.

In this case

$$\phi(W, \gamma, \alpha, \theta) = \alpha(X, \theta)\{Y - \gamma(X)\}.$$

Can use orthogonality to get objective function to which NN can be applied for estimating α_0.

Example 1: Orthogonal moment function is

$$\psi(W,\gamma,\alpha,\theta) = Z\gamma(X) - \theta + \alpha(X)\{Y - \gamma(X)\}.$$

Property I is that for every function $\gamma(X)$ and $\Delta(X) = \gamma(X) - \gamma_0(X)$ of X,

$$0 = \frac{dE[\psi(W,\gamma_0 + \delta\Delta, \alpha_0, \theta)]}{d\delta} = E[\{Z - \alpha_0(X)\}\Delta(X)] = 0.$$

This is the first order condition for α_0 given by

$$\begin{aligned} \alpha_0 &= \arg\min_{\alpha} E[-2Z\alpha(X) + \alpha(X)^2] = \arg\min_{a} E[-2g(W,\alpha,\theta) + \alpha(X)^2] \\ &= \arg\min_{a} E[\{Z - \alpha(X)\}^2] = E[Z|X]. \end{aligned}$$

Here Neyman orthogonality implies that $\alpha_0(X)$ minimizes the objective function $E[-2g(W,\alpha,\theta)+\alpha(X)^2]$ that only depends on the identifying moment function.

Can then use NN to estimate $\alpha_0(X)$ by minimizing a sample version of this objective function over a neural net choice of α; here this is neural net regression of Z on space of neural nets.

Can do this in general when identifying moment $g(W,\gamma,\theta)$ is smooth in γ.

Suppose that there is $D(W,\Delta;\gamma,\theta)$ that is linear in Δ such that for any $\Delta(X)$,

$$\frac{dg(W,\gamma+\delta\Delta,\theta)}{d\delta} = D(W,\Delta;\gamma,\theta).$$

Then Property I of Neyman orthogonality is that for all Δ,

$$\begin{aligned} 0 &= \frac{dE[\psi(W,\gamma_0+\delta\Delta,\alpha_0,\theta_0)]}{d\delta} = \frac{E[dg(W,\gamma_0+\delta\Delta,\theta_0)]}{d\delta} \\ &\quad + \frac{E[d\alpha_0(X)\{Y-\gamma_0(X)-\delta\Delta(X)\}]}{d\delta} \\ &= E[D(W,\Delta;\gamma_0,\theta_0)-\alpha_0(X)\Delta(X)]. \end{aligned}$$

This is the first order condition for

$$\alpha_0 = \arg\min_{\alpha} E[-2D(W,\alpha;\gamma_0,\theta_0)+\alpha(X)^2].$$

An NN estimator $\hat{\alpha}(X)$ of $\alpha_0(X)$ can then be formed from minimizing a sample analog of this objective function

$$\hat{\alpha} = \arg\min_{\alpha\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\{-2D(W_i,\alpha;\hat{\gamma},\hat{\theta})+\alpha(X_i)^2\},$$

where $\mathcal{A}$ is set of NN.

EXAMPLE 2: DYNAMIC BINARY CHOICE

Estimate structural parameters via learners of conditional choice probabilities.

Use Hotz and Miller approach (1993) that replaces computation of expected value functions with nonparametric estimation of choice probabilities.

Individuals choose between two alternatives $j = 1$ and $j = 2$ to maximize the expected present discounted value of per period utility $U_{tj} = D_j(X_t)'\theta_0 + \varepsilon_{tj}$, $(j = 1, 2; t = 1, ..., T)$, where ε_{jt} is i.i.d. with known CDF independent of the entire history $\{X_s\}_{s=1}^{\infty}$ of a state variable vector X, and X_t is Markov of order 1 and stationary.

The parameter vector of interest is θ_0.

Assume that choice 1 is a renewal choice where the conditional distribution of X_{t+1} given X_t and choice 1 does not depend on X_{it} and $D_1(X_t) = (-1, 0')'$ and $D_{21}(X_t) = 0$.

Three first steps:

$$\begin{aligned} \gamma_{10}(X_t) &= \mathsf{Pr}(Y_{2t} = 1 | X_t),\ \gamma_{20}(X_t) = E[H(\gamma_{10}(X_{t+1}))|X_t, Y_{2t} = 1], \\ \gamma_{30} &= E[H(\gamma_{10}(X_{t+1}))|Y_{1t} = 1] \end{aligned}$$

Here $H(\gamma_1)$ is the known function from Hotz and Miller such that for the expected value function

$$E[V(X_{t+1})|X_t, Y_{2t} = 1] - E[V(X_{t+1})|Y_{1t} = 1] = \gamma_{20}(X_t) - \gamma_{30}.$$

Then for the CDF $\Lambda(a)$ of $\varepsilon_{t1} - \varepsilon_{t2}$, $D(X_t) = D_2(X_t) - D_1(X_t)$, and δ the discount factor (assumed known) the conditional choice probability for $j = 2$ is

$$\mathsf{Pr}(Y_{2t} = 1|X_t) = \Lambda(a(X_t, \theta_0, \gamma_{20}, \gamma_{30})),\ a(x, \theta, \gamma_2, \gamma_3) = D(x)'\theta + \delta\{\gamma_2(x) - \gamma_3\}.$$

Data of i.i.d. observations on individuals each followed for T time periods, that also includes the $T+1$ observation X_{T+1} of the state variables, where $W = (X_1', Y_{21}, ..., X_T', Y_{2T}, X_{T+1}')'$.

First step $\hat{\gamma}_1(x)$ choice probability can be NN or anything with fast enough mean square convergence rate.

First step $\hat{\gamma}_2(x)$ obtained from NN or other regression of $H(\hat{\gamma}_1(X_{it+1}))$ on functions of X_{it}, for $Y_{i2t} = 1$.

For relative simplicity take identifying moment functions to be instrumental variables first order conditions for quasi maximum likelihood, where

$$g(W, \gamma, \theta) = \frac{1}{T} \sum_{t=1}^{T} Z(X_t)[Y_{2t} - \Lambda(a(X_t, \theta, \gamma_2, \gamma_3))].$$

For each element $g_j(W, \gamma, \theta)$ of $g(W, \gamma, \theta)$ there are three bias correction terms ϕ_{1j}, ϕ_{2j}, ϕ_{3j} corresponding to $\hat{\gamma}_1$, $\hat{\gamma}_2$, and $\hat{\gamma}_3$.

Each term is obtained treating the other γ functions as equal to the true ones.

Have for each $g_j(W,\gamma,\theta)$,

$$\begin{aligned}
\phi_{1j}(W,\gamma_1,\gamma_2,\gamma_3,\alpha_{1j},\theta) &= \frac{1}{T}\sum_{t=1}^{T}\alpha_{1j}(X_t)\left\{Y_{2t}-\gamma_1(X_t)\right\},\\
\phi_{2j}(W,\gamma_1,\gamma_2,\gamma_3,\alpha_{2j},\theta) &= \frac{1}{T}\sum_{t=1}^{T}\alpha_{2j}(X_t)Y_{2t}\left\{H(\gamma_1(X_{t+1}))-\gamma_2(X_t)\right\}
\end{aligned}$$

ϕ_3 given in LR paper.

Describe here NN estimator of $\alpha_{1j0}(X_t)$; let

$$\begin{aligned}
D_{1j}(W,\alpha;\hat{\gamma},\hat{\theta}) &= \frac{dg_j(W,\hat{\gamma}_1+\delta\alpha;\hat{\gamma}_2,\hat{\gamma}_3,\hat{\theta})}{d\delta}\\
&= \frac{-\delta}{T}\sum_{t=1}^{T}Z_j(X_t)\Lambda_a(a(X_t,\hat{\theta},\hat{\gamma}_2,\hat{\gamma}_3))[\hat{\gamma}_{21}(\alpha,X_t)-\hat{\gamma}_{31}(\alpha)],\\
\hat{\gamma}_{21}(\alpha,X_t) &= \hat{E}[H'(\hat{\gamma}_1(X_{t+1}))\alpha(X_{t+1})|X_t,Y_{2t}=1],\\
\hat{\gamma}_{31}(\alpha,X_t) &= \hat{E}[H'(\hat{\gamma}_1(X_{t+1}))\alpha(X_{t+1})|Y_{2t}=1].
\end{aligned}$$

A NN is given by

$$\hat{\alpha}_{1j}=\arg\min_{\alpha\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}\{-2D_{1j}(W_i,\alpha;\hat{\gamma},\hat{\theta})+\frac{1}{T}\sum_{t=1}^{T}\alpha(X_{it})^2\}$$

SUMMARY

Described here how to do automatic debiased machine learning for estimators with moment functions that depend on NN and other estimators of conditional expectations.

These results are preliminary; still need to formulate regularity conditions for dynamic discrete choice and other moment functions with automatic estimation of bias correction functions α.

Many more problems to for which these methods can be used.