Machine Learning of Dynamic Discrete Choice

Whitney K. Newey,
MIT

Econometric Society Summer School in Dynamic Structural

Econometrics

University of Chicago
July 9, 2019

INTRODUCTION

Many interesting economic parameters depend on unknown functions.

–Dynamic structural parameters are functions of conditional choice probabilities, Hotz and Miller (1993).

–Bounds on average consumer surplus depend on expected demand, Hausman and Newey (2016).

The unknown functions can be high dimensional, e.g. for large state spaces or many prices.

Machine learning provides flexible, computationally feasible ways of learning functions needed to estimate structural parameters.

ML can be useful when there are many small coefficients in high dimensional regression.

ML methods include Lasso, neural nets, random forests, boosting.

None of these require that many true coefficients be zero, only that many true coefficients are small.

ML is different than standard nonparametric estimation; there is no need to order or group regressors prior to estimation, e.g. many more regressors than sample size can be included in Lasso algorithm.

The ability to potentially include many regressors distinguishes ML from standard methods (e.g. series estimation) in high dimensional settings where there are too many potential interaction terms to order or group effectively.

ML is good for prediction but is inherently biased.

ML sets bias $\approx$ standard deviation for good prediction but that gives bad confidence intervals.

Plugging in ML to continuous regression functionals can be even worse, with bias shrinking slower than standard deviation; the plug-in estimator may not be root-n consistent.

For Lasso the regularization bias has size $\sqrt{\ln(p)/n}$, where p is number of potential regressors.

$$\sqrt{n}Bias \approx \sqrt{n}\sqrt{\ln(p)/n} = \sqrt{\ln(p)} \longrightarrow \infty.$$

Plug-in estimator is not root-n consistent, see Chernozhukov et al (2019, "Locally Robust Semiparametric Estimation").

Plugging in post Lasso is also biased.

Post Lasso is unrestricted least squares from using regressors that have nonzero coefficients in Lasso.

Plugging in post Lasso suffers from usual problem of lack of uniform inference with model selection.

Plug in estimator is asymptotically equivalent to a sample average that changes depending on what regressors are important.

Consequently plug-in estimator is not locally regular, meaning that it is biased in some directions, so confidence intervals are incorrect in local neighborhoods; see Chernozhukov et al (2019, "Locally Robust Semiparametric Estimation").

Example is θ_0 in a partially linear model

$$y_i = \theta_0 D_i + h(V_i) + \varepsilon_i.$$

"Partialling out" $h(V)$ debiases to give $\hat{\theta}$ that is asymptotically unaffected by model selection on $h(V)$; Belloni, Chernozhukov, Hansen (2014).

Partialling out also makes regularization bias second order ($Bias = \mathsf{ln}(p)/n$), so $\hat{\theta}$ is root-n consistent and asymptotically unbiased with Lasso first step ($\sqrt{n}Bias = \mathsf{ln}(p)/\sqrt{n} \longrightarrow 0$).

The partialling out is like the Robinson (1988) regression of $y_i - \widehat{E[y_i|x_i]}$ on $D_i - \widehat{E[D_i|x_i]}$, where here "hats" are machine learners rather than kernel regression and do cross-fitting (sample splitting) as we describe.

Graphs compare distribution of plug-in estimator with debiased machine learner (DML) based on random forest.

A general way to construct moment functions where first step has no first order effect is to add to the identifying moment functions a learner of the influence function adjustment from Newey (1994); see Newey, Hsieh, Robins (1998, 2004) and Chernozhukov et al. (2016, "Locally Robust Semiparametric Estimation").

LR moment functions for GMM:

$$\text{LR moment functions}$$

$$= \text{identifying (original) moment functions}$$

$$+ \text{ first step influence adjustment.}$$

LR moment functions could also be thought of as the influence function of the moments when moment conditions hold; above is useful way to construct LR moments and decompose them for large sample theory.

The first step adjustment is available whenever the moment function limit is root-n consistently estimable, i.e. for "averages."

Adding adjustment term is a general way of "partialling out" the first step estimation.

Some formulas for the adjustment term are known; Newey (1994).

Often can be derived using straight forward influence function calculation; see below and Ichimura and Newey (2015).

Can also construct learners of the adjustment that do not require knowing its form; Chernozhukov, Newey, Robins (2018) and Chernozhukov, Newey, Singh (2018).

Advantages of LR moment conditions:

–Estimator of parameter of interest is root-n consistent with regularized first step; plug-in estimator is not.

–Fixes model selection problem; plug-in estimator has model selection bias.

–LR estimators have smaller MSE in many cases.

–LR estimators are doubly robust in many cases; i.e. they are consistent even when one first step is not.

–Much simpler regularity conditions for LR estimators than plug-in estimators.

Some history: Parametric MLE, Neyman C-alpha moments are LR.

Generalized to GMM by Wooldridge (1991), Lee (2005), Chernozhukov et al. (2010).

Decomposition here appears novel for GMM: LR moments = identifying moments + influence adjustment.

Haasminski and Ibragimov (1978) proposed LR estimators of functionals of a density.

Bickel and Ritov (1988) LR estimator of integrated squared density that is root-n consistent under minimal conditions.

Robinson (1988) partially linear estimator, Ichimura (1993) single index estimator, Newey (1994) derivative of objective function where first step is concentrated out.

Robins, Rotnitzky, and Zhao (1994, 1995) doubly robust estimators of average treatment effects that are now in widely thought to work well.

Debiased moment conditions in Newey, Hsieh, and Robins (1998, 2003) are LR.

Use of machine learning with LR moment conditions pioneered by, Belloni, Chen, Chernozhukov, Hansen (2012), Belloni, Chernozhukov, Hansen (2014), Farrell (2015).

Contributions of Chernozhukov et al. (2019, Locally Robust Semiparametric Estimation)

–General construction of LR moment conditions from any moment condition and first step estimator that is estimator based, not model based.

–Show Lasso plug-in not root-n consistent.

–Show LR moment conditions are regular with model selection, unlike plug-ins.

–Derive LR moment conditions for dynamic discrete choice.

–General characterization of double robustness.

–Simple and general conditions for asymptotic normality.

SEMIPARAMETRIC GMM ESTIMATORS

The general class of estimators we consider are GMM where the moments depend on first step nonparametric estimators.

Estimation is based on a vector of functions $m(z, \beta, \gamma)$ depending on a data observation z, parameters of interest β, and a first step function γ satisfying an identifying moment condition

$$E[m(z_i, \beta_0, \gamma_0)] = 0.$$

The 0 subscript denotes the true values.

Let $\hat{\gamma}$ be the first step estimator of γ_0. Estimated sample moments can be formed as

$$\hat{m}(\beta) = \sum_{i=1}^{n} m(z_i, \beta, \hat{\gamma})/n.$$

For a positive semi-definite weight matrix $\hat{W}$ a GMM estimator is given by

$$\tilde{\beta} = \arg\min_{\beta \in B} \hat{m}(\beta)^T \hat{W} \hat{m}(\beta).$$

Form LR moments by adding an adjustment term to original (identifying) moments.

To describe let $\gamma(F)$ be the probability limit of $\hat{\gamma}$ when distribution of a single observation is F.

Let F_0 denote the true distribution, G some other distribution, and $F_\tau = (1-\tau)F_0 + \tau G$, such that $\gamma(F_\tau)$ is well defined.

Adjustment term $\phi(z, \beta, \gamma, \lambda)$ satisfies

$$\frac{d}{d\tau} E[m(z_i, \beta, \gamma(F_\tau))] = \int \phi(z, \beta, \gamma_0, \lambda_0) G(dz), \; E[\phi(z_i, \beta, \gamma_0, \lambda_0)] = 0.$$

Generally $\phi(z, \beta, \gamma, \lambda)$ depends on additional unknown function(s) λ.

$$\frac{d}{d\tau}E[m(z_i,\beta,\gamma(F_\tau))] = \int \phi(z,\beta,\gamma_0,\lambda_0)G(dz),\ E[\phi(z_i,\beta,\gamma_0,\lambda_0)] = 0.$$

First equation is Von Mises (1947), Hampel (1974), Huber (1981) Gateaux derivative definition of the influence function of

$$\mu(F) = E[m(z_i,\beta,\gamma(F))].$$

Can often find $\phi(z,\beta,\gamma,\lambda)$ by differentiating (before equality) for smooth G and evaluating (after equality) at G with $\Pr(z_i = z) = 1$; see Ichimura and Newey (2015).

$\phi(z,\beta,\gamma,\lambda)$ generally exists when $\mu(F)$ is root-n consistently estimable.

LR moment functions can be constructed by adding $\phi(z,\beta,\gamma,\lambda)$ to $m(z,\beta,\gamma)$ to obtain new moment functions

$$\psi(z,\beta,\gamma,\lambda) = m(z,\beta,\gamma) + \phi(z,\beta,\gamma,\lambda).$$

$$\psi(z, \beta, \gamma, \lambda) = m(z, \beta, \gamma) + \phi(z, \beta, \gamma, \lambda).$$

Partition observations into L groups $I_\ell, (\ell = 1, ...L)$, (e.g. $L = 5$ or $L = 10$). Let

$$\hat{\psi}(\beta) = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \psi(z_i, \beta, \hat{\gamma}_\ell, \hat{\lambda}_\ell).$$

where $\hat{\gamma}_\ell$ and $\hat{\lambda}_\ell$ are formed from observations not in I_ℓ.

A LR GMM estimator is

$$\hat{\beta} = \arg \min_{\beta \in B} \hat{\psi}(\beta)^T \hat{W} \hat{\psi}(\beta).$$

As usual $\hat{W} = \hat{\Omega}^{-1}$ minimizes asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_0)$ where

$$\Omega = E[\psi(z_i, \beta_0, \gamma_0, \lambda_0)\psi(z_i, \beta_0, \gamma_0, \lambda_0)^T].$$

Presence of first steps $\hat{\gamma}$ and $\hat{\lambda}$ has no effect on asymptotic variance because using LR moments.

$$\hat{\beta} = \arg\min_{\beta \in B} \hat{\psi}(\beta)^T \hat{W} \hat{\psi}(\beta), \quad \hat{\psi}(\beta) = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \psi(z_i, \beta, \hat{\gamma}_\ell, \hat{\lambda}_\ell).$$

Cross-fitting removes bias and avoids Donsker conditions which are not known to be satisfied for many machine learners.

Cross-fitting removes "own observation" bias that leads to large remainders.

$\sqrt{n}\hat{\psi}(\beta_0)$ and $\sqrt{n}\hat{m}(\beta_0)$ have the same asymptotic variance if $\sqrt{n}\hat{m}(\beta_0)$ is asymptotically normal because adjustment term has mean zero for all data distributions.

$\sqrt{n}\hat{m}(\beta_0)$ blows up for ML $\hat{\gamma}$ so LR moments are not just "higher-order" better, they are essential for root-n consistency.

It is easy to show that $\hat{\gamma}$ and $\hat{\lambda}$ have no first-order effect on $E[\psi(z, \beta, \gamma, \lambda)]$. Let $\gamma(F)$ and $\lambda(F)$ denote plim of $\hat{\gamma}$ and $\hat{\lambda}$. Mean zero property of influence adjustment implies the identity

$$E_{F_\tau}[\phi(z_i, \beta, \gamma(F_\tau), \lambda(F_\tau))] \equiv 0.$$

Differentiating left hand side while applying chain rule to $E_{F_\tau}[\cdot]$ and $\gamma(F_\tau), \lambda(F_\tau)$ gives

$$\begin{aligned} 0 &= \int \phi(z, \beta, \gamma_0, \lambda_0) dG + \frac{\partial E[\phi(z_i, \beta, \gamma(F_\tau), \lambda(F_\tau))]}{\partial \tau} \\ &= \frac{\partial E[m(z_i, \beta, \gamma(F_\tau))]}{\partial \tau} + \frac{\partial E[\phi(z_i, \beta, \gamma(F_\tau), \lambda(F_\tau))]}{\partial \tau} \\ &= \frac{\partial E[\psi(z_i, \beta, \gamma(F_\tau), \lambda(F_\tau))]}{\partial \tau}, \end{aligned}$$

where the second equality follows by the definition of ϕ and the third by the definition of ψ.

Thus we see there is no first order effect of γ and λ on $\bar{\psi}(\gamma, \lambda) = E[\psi(z_i, \beta_0, \gamma, \lambda)]$. Under additional conditions get

$$\left|\bar{\psi}(\gamma, \lambda_0)\right| \leq C \left\|\gamma - \gamma_0\right\|^2, \ \bar{\psi}(\gamma_0, \lambda) = E[\phi(z_i, \gamma_0, \lambda)] = 0,$$

where $\|\cdot\|$ is a function norm; directly useful for simple asymptotic theory; double robustness is $\bar{\psi}(\gamma, \lambda_0) = 0$.

MACHINE LEARNING OF DYNAMIC DISCRETE CHOICE

Let x_t be state variables like work experience, number of children, etc, β unknown parameters. Utility of choice j in period t is

$$U_{jt} = v_j(x_t, \beta_0) + \epsilon_{jt}, (j = 1, ..., J; t = 1, 2, ...).$$

Disturbances ϵ_{jt} are i.i.d over time with known distribution with support R^J and x_t is Markov, order one.

Let $y_{jt} \in \{0, 1\}$ be choice indicator, $y_{tj} = 1$ if j chosen in period t.

Let $P_{j0}(x_t) = \mathsf{Pr}(y_{tj} = 1|x_t)$ be choice probabilities.

Rust (1987) shows $P_{j0}(x_t)$ has known functional form depending on expected value function.

Hotz and Miller (1993) shows can invert that relationship to get expected value function differences as function of conditional choice probability and expectation of next period given current period future.

Use Hotz-Miller, CCP identifying moment conditions.

Add influence adjustments for first steps to form LR moment functions.

Let δ be time discount parameter, $\bar{v}(x)$ expected value function,

$$\bar{v}_j(x_t) = u_j(x_t, \beta_0) + \delta E[\bar{v}(x_{t+1})|x_t, j]$$

the expected value function conditional on choice j. CCP is then

$$P_j(\bar{v}_t) = \mathsf{Pr}(\bar{v}_j(x_t) + \epsilon_{jt} \geq \bar{v}_k(x_t) + \epsilon_{kt}; k = 1, ..., J), \ .$$

for $\bar{v}_t = (\bar{v}_1(x_t), ..., \bar{v}_J(x_t))'$.

Suppose there is renewal choice, $j = 1$ without loss of generality, so

$$E[\bar{v}(x_{t+1})|x_t, 1] = C$$

Let $\tilde{v}_j(x_t) = \bar{v}_j(x_t) - \bar{v}_1(x_t)$, so that $\tilde{v}_1(x_t) \equiv 0$. As usual choice probabilities depend only on $\tilde{v}_t = (\tilde{v}_2(x_t), ..., \tilde{v}_J(x_t))$.

Define $P_t = P(\tilde{v}_t) = (P_1(\bar{v}_t), ...P_J(\bar{v}_t))'$ to be vector of choice probabilities. As in Hotz and Miller (1993), there is a function $P^{-1}(P)$ such that $\tilde{v}_t = P^{-1}(P_t)$. Let $H(P)$ denote the function such that

$$H_t = H(P_t) = E[\max_{1 \le j \le J} \{\mathcal{P}^{-1}(P_t)_j + \epsilon_{jt}\} | x_t] = E[\max_{1 \le j \le J} \{\tilde{v}_{jt} + \epsilon_{jt}\} | x_t].$$

For example, for multinomial logit $H_t = .5772 - \ln(P_{1t})$.

Note that by the definition of H_t,

$$\begin{aligned} \bar{v}(x_t) &= \bar{v}_1(x_t) + H_t = u_1(x_t, \beta_0) + \delta E[\bar{v}(x_{t+1}) | x_t, 1] + H_t \\ &= u_1(x_t, \beta_0) + \delta C + H_t. \end{aligned}$$

Then the choice specific value functions are

$$\begin{aligned} \bar{v}_j(x_t) &= u_j(x_t, \beta_0) + \delta E[\bar{v}(x_{t+1}) | x_t, j] \\ &= u_j(x_t, \beta_0) + \delta E[u_1(x_{t+1}, \beta_0) + H_{t+1} | x_t, j] + \delta^2 C. \end{aligned}$$

Differencing and using the renewal property again then gives

$$\begin{aligned} \tilde{v}_j(x_t) &= \bar{v}_j(x_t) - \bar{v}_1(x_t) \\ &= u_j(x_t, \beta_0) - u_1(x_t, \beta_0) \\ &\quad + \delta \{E[u_1(x_{t+1}, \beta_0) + H_{t+1} | x_t, j] - E[u_1(x_{t+1}, \beta_0) + H_{t+1} | 1]\}. \end{aligned}$$

Consider special case where

$$u_j(x_t, \beta) = w'_{jt}\beta, \; w_{jt} = w_j(x_t).$$

Define $\bar{w}^j_{1t} = E[w_{1,t+1}|x_t, j]$. Then previous formula for $\tilde{v}_j(x_t)$ becomes

$$\begin{aligned} \tilde{v}_j(x_t) &= (w_{jt} - w_{1t})'\beta_0 + \delta\{\bar{w}^j_{1t} - \bar{w}^1_{1t}\}'\beta_0 \\ &\quad + \delta\{E[H_{t+1}|x_t, j] - E[H_{t+1}|1]\} \end{aligned}$$

Hotz-Miller CCP discrete choice estimator substitutes these in choice probabilities replacing β_0 by β, $\bar{w}^j_{1t}$ by estimators $\hat{w}^j_{1t}$, $(j = 1, ..., J)$, H_t by $\hat{H}_{t+1} = H(\hat{P}_{t+1})$, and $E[H_{t+1}|x_t, j]$ by an estimator $\hat{E}[\hat{H}_{t+1}|x_t, j]$.

For binary choice where $e_{t2} - e_{t1}$ has CDF Λ this leads to estimated choice probability for choice 2 given by

$$\begin{aligned} & \hat{\pi}(x_t, \beta) \\ = \; & \Lambda(\{w_{2t} - w_{1t} + \delta[w^j_{1t} - \hat{w}^1_{1t}]\}'\beta + \delta\{\hat{E}[\hat{H}_{t+1}|x_t, 2] - \hat{E}[\hat{H}_{t+1}|1]\}) \end{aligned}$$

To make things especially simple suppose that $w_{1t} = (-1, 0')'$, $w_{2t} = (0, x_{2t}')'$ and let $X_t = (1, x_{2t}')'$. Then

$$\hat{\pi}(x_t, \beta) = \Lambda(X_t'\beta + \delta\{\hat{E}[\hat{H}_{t+1}|x_t, 2] - \hat{E}[\hat{H}_{t+1}|1]\}).$$

Let $y_t = y_{2t}$ and $A_t = A(x_t)$ be a vector of functions of the state variables.

Then semiparametric moment conditions are

$$m(z_i, \beta, \hat{\gamma}) = A_t\{y_t - \Lambda(X_t'\beta + \delta\{\hat{E}[\hat{H}_{t+1}|x_t, 2] - \hat{E}[\hat{H}_{t+1}|1]\})\}.$$

Here the first step $\hat{\gamma}$ consists of the $\hat{H}_{t+1}$ as well as the $\hat{E}[\cdot|x_t, 2]$ and $\hat{E}[\cdot|1]$.

For high dimensional x_t (and low dimensional X_t) can estimate $\hat{P}_t$ by ML.

Just plugging this in is bad, e.g. plugging in and doing binary choice; Plug-in estimator is biased due to regularization and/or model selection.

Add adjustment term to get LR moment conditions.

Two approaches:

1. Derive adjustment term and directly estimate nonparametric components, here.

2. Use parametric delta method and Lasso minimum distance, like Chernozhukov, Newey, Singh (2018).

Adjustment term is always the sum of components, with one component for each first step.

Here there are 3 first steps, so adjustment term is sum of 3 components

The "Locally Robust Semiparametric Estimation" paper derives results; straightfoard but tedious.

$$P_{\tilde{v}}(\tilde{v}) = \partial P(\tilde{v})/\partial \tilde{v},\ \pi_1 = \mathsf{Pr}(y_{t1} = 1),\ \lambda_{10}(x) = E[y_{1t}|x_{t+1} = x],$$
$$\lambda_{20}(x) = E[A_t P_{2\tilde{v}}(\tilde{v}_t)\frac{y_{t2}}{P_2(\tilde{v}_t)}|x_{t+1} = x].$$

Then for $w_t = x_{t+1}$ and $z = (y, x, w)$ let

$$\phi_1(z, \beta, \gamma, \lambda) = -\delta\{\lambda_2(x) - E[A(x_t)P_{\tilde{v}2}(\tilde{v}_t)]\pi_1^{-1}\lambda_1(x)\}[\partial H(P(x))/\partial P]'\{y - P(x)$$
$$\phi_2(z, \beta, \gamma, \lambda) = -\delta A(x)P_{\tilde{v}j}(\tilde{v}(x, \beta, \gamma))\frac{y_2}{P_2(\tilde{v})}\{H(P(w)) - E[H_{t+1}|x_t = x]\},$$

Design of Monte Carlo experiment is loosely like bus replacement example of Rust (1987), with bus replacement happening every 8 periods, more often than in Rust (1987) data.

Here y_t is binary replacement choice. State transition is

$$x_{t+1} = \begin{cases} x_t + N(.25, 1)^2, y_t = 1, \\ x_t = 1 + N(.25, 1)^2, y_t = 0. \end{cases}$$

Agent chooses y_t contingent on state to maximize

$$\sum_{t=1}^{\infty} \delta^{t-1}[y_t(\alpha\sqrt{x_t} + \varepsilon_t) + (1 - y_t)RC], \alpha = -.3, RC = -4.$$

Estimate CCP using kernel, series, Lasso, random forest.

Estimate adjustment term using series regression throughout.

Report some results for 1000 observations:

CCP Estimators, Dynamic Discrete Choice

	Bias		Std Err		95% Cov	
	α	RC	α	RC	α	RC
Two step kernel	-.24	.08	.08	.32	.01	.86
LR kernel	-.05	.02	.06	.32	.95	.92
Two step quad	-.00	.14	.049	.33	.91	.89
LR quad	-.00	.01	.085	.39	.95	.92
Logit Lasso	-.12	.25	.06	.28	.74	.84
LR Logit Lasso	-.09	.01	.08	.36	.93	.95
Random Forest	-.15	-.44	.09	.50	.91	.98
LR Ran. For.	.00	.00	.06	.44	1.0	.98
Boosted Trees	-.10	-.28	.08	.50	.99	.99
LR Boost Tr.	.03	.09	.07	.47	.99	.97

Example shows.

–Bias reduction.

–Smaller MSE in many cases.

–Confidence interval coverage probabilities are closer to nominal.

We expect these good properties to hold more generally, as they do in previous graphs.

SUMMARY

–Plugging in machine learners to parameter formulas is biased.

–Regularization bias destroys root-n consistency.

–Model selection bias leads to bad confidence intervals.

–Debiased/locally robust/orthogonal moment conditions removes these biases.

–Use for machine learning of dynamic discrete choice.

–Find smaller bias and MSE and more accurate confidence intervals.