Machine Learning for the Social Sciences

IIII. ML for Causal Inference: effect heterogeneity, IV, optimal treatment assignments

Andrew Peterson

Université de Genève

Table of contents

Review and Caveat: Double machine learning

Double machine learning: Chernozhukov, et al. (2017)

- Intuition: Combine ideas of double robustness, ML, and orthogonality
- In particular: naïve ML prediction is biased.
- Semi-parametric model: for treatment D, controls Z,
  $Y = D\theta_0 + g_0(Z) + U$
- $\mathbb{E}[U|Z, D] = 0$

Double machine learning: Chernozhukov, et al. (2017)

1. (like double selection) Use ML to predict Y and D from Z to get $\mathbb{E}[\hat{Y}|Z]$, $\mathbb{E}[\hat{D}|Z]$

2. FWL to cleanly estimate the effect (and de-bias):
Residualize Y and D of the control variables:
$\tilde{Y} = Y - \mathbb{E}[\hat{Y}|Z]$;
$\tilde{D} = D - \mathbb{E}[\hat{D}|Z]$

3. Regression of $\tilde{Y}$ on $\tilde{D}$ to get treatment effect θ_0

Caveat

› D'Amour, et al.[1] warn that while additional covariates might allow you to address confounding, it becomes more and more difficult to maintain covariate overlap (analogous to curse of dimensionality)

[1] D'Amour, et al. (2017) "Overlap in Observational Studies with High-Dimensional Covariates"

Effect Heterogeneity

Effect Heterogeneity

- Example: Beyond the general question of whether a drug works (ATE, etc), it's worth understanding the variation in how different individuals respond.

- Some individuals might have no response to a generally effective drug, while others are very sensitive due to reduced kidney function etc. Might need to give different doses, monitor certain patients, etc.

Effect Heterogeneity

- Obvious response is to calculate different ATEs for subgroups defined by covariates
- One way to do this is trees or forests, which partition the data. Then identify the mean difference within each partition.
- Ok great, but what depth, how many leaves? (As usual, we will need to regularize).

Causal trees

› Two estimation procedures,
(1) tree branches based on in-sample MSE (for a given λ), and

(2) Optimize a regularization parameter λ based on grid search with cross-validation OOS, with loss as MSE.

› Actually the details are a bit more complex: They split the sample to get an 'honest' estimate by using one sample to determine the partition and the other to estimate the causal effect.[1]

[1] (See Athey & Imbens (2016) "Recursive partitioning for heterogeneous causal effects"

Causal trees

- The trees approach sacrifices some predictive accuracy for interpretability. So ideal for exploratory work. If you want to max performance, try instead causal forests (below)

- Caveat - while it's easy to understand the treatment effect in a partition, you can't make an inference from what branches exist to what variables are important.

Causal forests

- Review: in random forests, construct multiple trees then average their predictions
- Problem: Lose interpretability - black box approach. See also gradient boosted trees, which in my experience work well for predictive accuracy.
- Practical: *causalTree* package for R (Athey, Kong, and Wager, 2015)

Effect Heterogeneity: Applications

- The foundation of these models is useful for a number of tasks, such as optimal treatment design (assigning treatment dynamically over time) or in generalizing the results from an experiment (extrapolating), etc.

- Some individuals might have no response to a generally effective drug, while others are very sensitive due to reduced kidney function etc. Might need to give different doses, monitor certain patients, etc.

Policy Learning

- There are many applications where we have a finite budget or other constraints but want to get the maximum value by wisely targeting which units (people, towns, etc) to treat.

- One way would be to (1) estimate the (heterogenous) treatment effects, then use this decide who to treat (say by having a cutoff and treating everyone who has an effect greater than the cutoff)

Policy Learning

- But this two-step process is sub-optimal since it cobbles together two loss functions: MSE for learning the treatment effects, and some other policy loss function (call it 'π') which relates to the cost of treating different individuals, etc.
- Might also need to think about how to incorporate covariates that we observe only after treatment, or fairness issues, etc.
- Simple case: binary treatment, but can also consider continuous treatment

Policy Learning

› Can apply the doubly-robust estimator, causal trees, etc

Instrumental Variables

Instrumental Variables: Setup

› Recall the basic 2SLS setup. While in practice we estimate 2SLS in one-step (for variances), can run separately:

(1) $D_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 X_i' + \epsilon_i$ (first-stage)

(2) $Y_i = \beta_0 + \beta_1 D_i + \beta_2' X_i + \epsilon_i$

ML & Instrumental Variables

- Chernozhukov, et al. (2017) - for partially linear IV: Use post-LASSO to estimate (1).
- Then use predicted values of the treatment (D_i) in the second stage to estimate the treatment effect β_1

Double debiased IV, applied

- Reanaysis of Acemoglu et al. (2001)- mortality rates as instrument for institutions
- ML allows relaxation of assumptions on geography as confounder

Table 4. Estimated Effect of Institutions on Output

	Lasso	Reg. Tree	Forest	Boosting	Neural Net.	Ensemble	Best
2 fold	0.85	0.81	0.84	0.77	0.94	0.8	0.83
	[0.28]	[0.42]	[0.38]	[0.33]	[0.32]	[0.35]	[0.34]
	(0.22)	(0.29)	(0.3)	(0.27)	(0.28)	(0.3)	(0.29)
5 fold	0.77	0.95	0.9	0.73	1.00	0.83	0.88
	[0.24]	[0.46]	[0.41]	[0.33]	[0.33]	[0.37]	[0.41]
	(0.17)	(0.45)	(0.4)	(0.27)	(0.3)	(0.34)	(0.39)

Note: Estimated coefficient from a linear instrumental variables model based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Results are based on 100 splits with point estimates calculated the median method. The median standard error across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses. Further details about the methods are provided in the main text.

[Compared to orig coef estim, s.e. of 1.10 (0.46)]

Deep IV

- Unlike above or 2SLS, do not assume a linear model.
- Could use non-parametric approaches (basis functions, kernel methods, e.g. Darolles 2011), but doesn't scale well

Deep IV

- Turn IV into 2 ML tasks
- Want to identify the effect of a treatment variable p, with covars x on outcome y:

$$y = g(p, x) + \epsilon$$

where g() is some possibly non-linear continuous fn

Deep IV

› Counterfactual prediction fn:

$$h(p, x) \equiv g(p, x) + \mathbb{E}[e|x]$$

"conditional expectation of y given the observables p and x, holding the distribution of e constant as p is changed."

Deep IV

› Under relevance (strong 1st stage), exclusion and unconfoundedness:

$$\mathbb{E}[y|x,z] = \int h(p,x)dF(p|x,z)$$

Deep IV: 1st stage

- Use deep neural network to learn $\hat{F}(p|x, z)$
- If treatment p discrete, categorical prediction with softmax, if p, continuous, use mix of Gaussian distributions

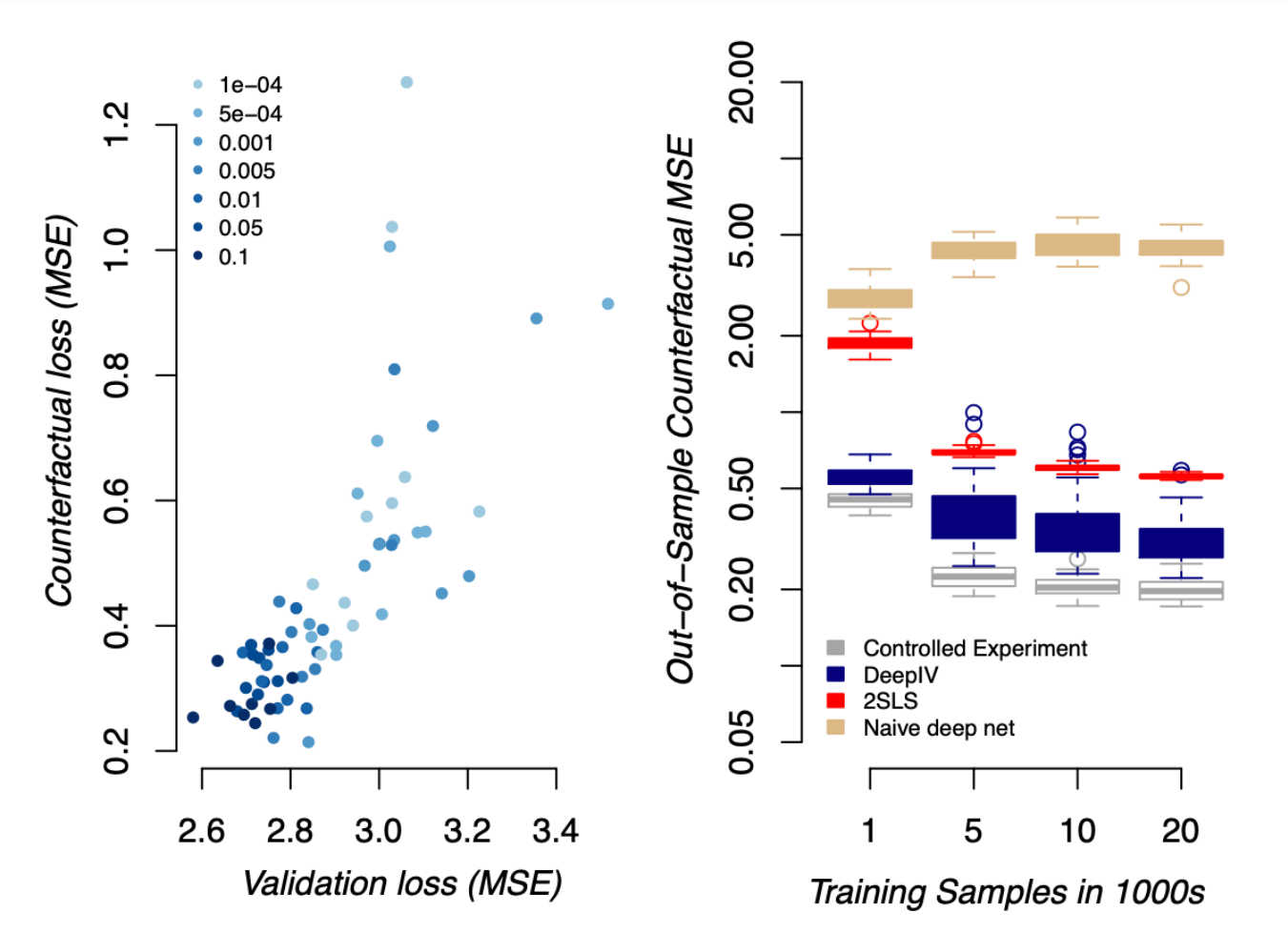
Deep IV: 2nd stage

› Minimize mean squared error via SGD for DNN parameters θ:

$$y_i = \int h_\theta(p, x_i) d\hat{F}_\theta(p|x_i, z_i)$$

› See additional details about unbiasedness by ensuring independent treatment draws; validation; discrete outcomes, inference, etc.

Deep IV: Results, code



Code available on Github: jhartford/DeepIV

Optimal treatment assignments & bandits

Imai & Strauss 2011

› Want to use info from previous study to optimize a new experiment, such as a GOTV campaign

› Incorporate heterogeneous treatment effect estimates into a Bayesian decision theoretic framework

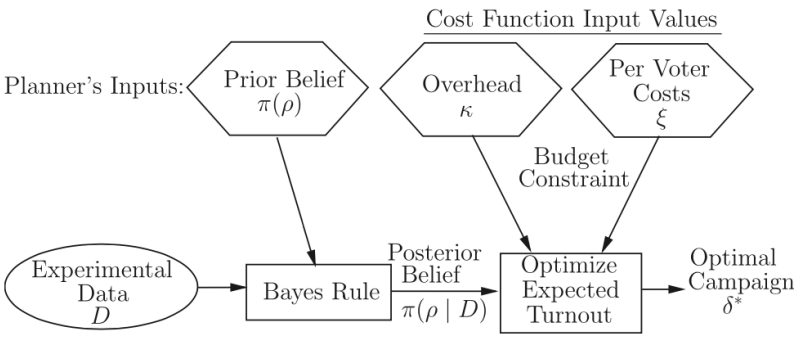


Fig. 1 An overview of the Bayesian optimal campaign planning process. Inputs over which the planner has direct control are represented by hexagons and are (1) the prior belief about the effects of various mobilization strategies on different voters, $\pi(\rho)$, (2) the overhead costs of each mobilization method, κ, and (3) the cost per voter for each strategy, ξ. Data from randomized field experiments, D, are represented by the oval. These data and the planner's prior distribution are combined via Bayes's rule to produce a posterior belief about the effects of mobilization strategies $\pi(\rho|D)$. Finally, our proposed optimization method uses this posterior belief and the exogenous costs ξ to find the optimal campaign strategy δ^* for the planner.

Imai & Strauss 2011

- They use a 4 step process to select vars, order vars by importance, fit the model, then derive the optimal strategy
- Classification trees with k-fold validation (steps 1,3), but can substitute other alternatives (Causal trees, forests, etc)

Bandits

- Multi-armed bandits (Robbins, 1952) - optimal exploration-exploitation tradeoff. Of interest when high-dim: many possible treatments, heterogeneous effects, etc.

- Markov-decision process - Policy fn π to chose a sequence of actions $a \in 1, ...K$ to maximize unknown reward function r

- Contextual bandits - also an exogenous variable x in the policy and reward fns

Upper confidence bound approach

- Measure of uncertainty on estimates based on previous observations (bound on potential outcome for an individual under a given treatment)

- (informally) Assign treatment to where we are most uncertain, that is, to the arm with the highest UCB

Thompson sampling

- Given likelihood fn, prior distribution on parameters, past observations, posterior distribution;
- sample parameters from the posterior, then chose the action that maximizes expected reward

Parametric approach

Dimakopoulou et al 2017:

- conditional expectation fn for reward given context x: $\mu_a(x) = \mathbb{E}[r(a)|x]$
- linear model: $\mu_a(x) = x^\intercal \theta_a$ or GLM
- LASSO/ridge regression. Compute the closed form mean estimate $\hat{\mu}_a(x)$ and $Var[\hat{\mu}_a(x)]$ or by bootstrap estimate.

Non-parametric approach

Dimakopoulou et al 2017:

- Generalized random forest (as above)

Balancing to reduce bias

- IPW - estimate propensity scores (multi-class logistic regr for UCB, or MC for Thompson sampling, then weighted versions of LASSO/ridge
- Approximate residual balancing: high dimensionality exact matching impossible; approx balance sufficient for eliminating bias with Lasso (R 'balanceHD')

Approx residual balancing: results

(Misspecified model: simulations roughly modeled on Lalonde (1986) study)

n			400					1000		
p	100	200	400	800	1600	100	200	400	800	1600
Naive	1.734	1.738	1.734	1.736	1.747	1.724	1.679	1.706	1.698	1.720
Elastic Net	0.446	0.468	0.492	0.517	0.540	0.376	0.380	0.389	0.401	0.413
Approximate Balance	0.523	0.582	0.609	0.656	0.700	0.297	0.327	0.379	0.395	0.464
Approx. Residual Balance	**0.249**	**0.276**	**0.270**	**0.295**	**0.310**	**0.168**	**0.175**	**0.176**	**0.179**	**0.194**
Inv. Propensity Weight	1.060	1.081	1.111	1.154	1.189	0.831	0.831	0.874	0.875	0.940
Augmented IPW	0.340	0.359	0.377	0.406	0.425	0.249	0.254	0.261	0.266	0.285
Weighted Elastic Net	0.313	0.338	0.355	0.385	0.412	0.204	0.209	0.220	0.221	0.249
TMLE Elastic Net	0.347	0.365	0.381	0.407	0.428	0.273	0.275	0.282	0.286	0.301
Double-Select + OLS	0.285	0.292	0.301	0.320	0.339	0.250	0.250	0.246	0.244	0.246

Table 5: Root-mean-squared error $\sqrt{\mathbb{E}\left[(\hat{\tau}-\tau)^2\right]}$ in the misspecified setting. All numbers are averaged over 400 simulation replications.

R package - 'balanceHD'

Questions?