Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments

Victor Chernozhukov
MIT

Mert Demirer
MIT

Esther Duflo
MIT

Iván Fernández-Val
BU (visiting NYU)

Celebrating Whitney Newey's Contributions to Econometrics

May 17, 2019

Heterogeneous Effects in Randomized Experiments

- Let $Y(1)$ and $Y(0)$ be the potential outcomes in the treatment state 1 and the non-treament state 0. Let Z be a vector of of covariates. The main causal functions are the baseline conditional average:

$$b_0(Z) := \mathrm{E}[Y(0) \mid Z],$$

and the conditional average treatment effect (CATE):

$$s_0(Z) := \mathrm{E}[Y(1) \mid Z] - \mathrm{E}[Y(0) \mid Z].$$

- Suppose the treatment variable D is randomly assigned conditional on Z, with probability of assignment depending only on a subvector of stratifying variables Z_1 in Z, namely $D \perp\!\!\!\perp (Y(1), Y(0)) \mid Z$, and the propensity score is known and is given by

$$p(Z) := \mathrm{P}[D = 1 \mid Z] = \mathrm{P}[D = 1 \mid Z_1].$$

- We assume that the propensity score is bounded away from zero or unity:

$$p(Z) \in [p_0, p_1] \subset (0, 1).$$

- The observed outcome is given by $Y = DY(1) + (1 - D)Y(0)$. Under the stated assumption, the causal functions coincide with the components of the regression function of Y given D, Z:

$$Y = b_0(Z) + Ds_0(Z) + U, \quad \mathrm{E}[U \mid Z, D] = 0,$$

that is,

$$b_0(Z) = \mathrm{E}[Y \mid D = 0, Z]$$

and

$$s_0(Z) = \mathrm{E}[Y \mid D = 1, Z] - \mathrm{E}[Y \mid D = 0, Z].$$

- We observe $\mathrm{Data} = (Y_i, Z_i, D_i)_{i=1}^N$, consisting of i.i.d. copies of random vector (Y, Z, D) having probability law P.

Why Not Use Machine Learning to Estimate $s_0(Z)$?

- A collection of constantly evolving statistical learning methods: Random Forest, Boosted Trees, Neural Networks, Penalized Regression, Ensembles, and Hybrids. Branded "Machine Learning".
- Work well in practice for prediction purposes and pattern recognizition, much better than classical methods in high-dimensional settings.
- We can apply ML methods to try to learn and approximate the CATE function

$$s \mapsto s_0(z)$$

- It is fundamentally difficult to obtain consistency and even harder to get credible inference for CATE using ML.

Fundamental Limitations for ML

- **Impossibility for Consistency** (Stone, 82): If $z \mapsto s_0(z)$ is known to be smooth with p bounded derivatives and $d = dim(z)$ is modest, $d \geqslant \log N$, then there exists no consistent ML estimator of $z \mapsto s_0(z)$.

  By the way, $\log(100,000) = 5$.

  Consistency is possible under *structured forms of sparsity,* but

  Sparsity is untestable, so must be used with caution

- **Valid Adaptive Confidence Sets Do Not Exist** (Low, 97; Genovese and Wasserman, 08) because adaptive estimators have non-negligible biases.

  Can construct partly adaptive confidence sets using self-similarity (Gine and Nickl, 11; Chernozhukov, Chetverikov, Kato, 13), bias-bounding or undersmoothing, but

  It remains unclear how to apply these methods to high-dimensional problems with structured sparsity

A Fundamental Theory-Practice Gap for ML

- Suppose we are willing to make strong assumptions for ML consistency and valid inference
- Many tuning parameters. Real implementations produced by a huge engineering effort. Have to trust the software engineers knowing statistics.
- Justification is very often heuristic and practice based. Theoretical justification is available in some cases, existence type results. There exist tuning parameters that make some of these methods work under assumptions that are hard to verify in practice.
- Even cross-validation remains unjustified in high-dimensional cases (exception:Lasso)
- Very often there are no theoretical guarantees for real implementations with the real tuning parameters (exception: Lasso)

Geometric Illustration of Impossibilities and Existing Gaps

Source: en.kueez.com

Deep Learning?

Our (Agnostic and Generic) Approach

- Motivated by limitations, we proceed agnostically: we will treat ML tools as providing us with predictor proxies for CATE.
- We don't assume they are consistent or unbiased.
- We will post-process the ML proxies to perform inference on key features of CATE.
- Our approach is **generic** with respect to the Machine Learning method being used
- In essence, continuation of work with Whitney on Double/De-biased Machine Learning.

- We shall rely on random data splitting into a main sample, indexed by M, and an auxiliary sample, indexed by A. Here (A, M) form a random partition of $\{1, ..., N\}$.

- From the auxiliary sample A, we obtain **Generic ML** estimates of the baseline and treatment effects, which we call proxy predictors

$$z \mapsto B(z) = B(z; \mathrm{Data}_A)$$

and

$$z \mapsto S(z) = S(z; \mathrm{Data}_A).$$

- We treat $B(Z)$ and $S(Z)$ agnostically as possibly biased and noisy predictors of $b_0(Z)$ and $s_0(Z)$.

- We condition on the auxiliary sample, so we consider these maps as frozen in the main sample.

Target Parameters

- We target and develop valid inference on *key features of* CATE and not the CATE itself:

  (1) Best linear predictor (**BLP**) of CATE $s_0(Z)$ using ML proxy $S(Z)$;

  (2) Group average treatment effects sorted (**GATES**) by the groups induced by ML proxy $S(Z)$;

  (3) Classification Analysis (**CLAN**): Average characteristics of the units in most and least affected groups.

BLP of CATE by ML Proxy

Consider the weighted linear projection:

$$Y = \alpha' X_1 + \beta_1(D - p(Z)) + \beta_2(D - p(Z))(S - \mathrm{E}S) + \epsilon, \quad \mathrm{E}[w(Z)\epsilon X] = 0,$$

where $w(Z) = \{p(Z)(1 - p(Z))\}^{-1}$, $\quad X := (X_1, X_2)$, $X_1 := X_1(Z)$, e.g. $X_1 = (1, B(Z))$, $X_2 := (D - p(Z), (D - p(Z))S(Z))$.

✓ The interaction $(D - p(Z))(S - \mathrm{E}S)$ and the weights $w(Z)$ creates necessary orthogonality with other variables.

Theorem 1: Projection coefficients identify the BLP of CATE:

$$\beta_1 + \beta_2(S(Z) - \mathrm{E}S) = \mathrm{BLP}[s_0(Z) \mid S(Z)],$$

in particular $\beta_1 = \mathrm{E}S_0(Z)$ and $\beta_2 = \mathrm{Cov}(s_0(Z), S(Z))/\mathrm{Var}(S(Z))$.

Special Cases

- If $S(Z)$ is a perfect proxy for $s_0(Z)$, then

$$\beta_2 = 1.$$

- In general, $\beta_2 \neq 1$, correcting for noise in $S(Z)$.
- If $S(Z)$ is complete noise, uncorrelated to $s_0(Z)$, then $\beta_2 = 0$
- If there is no heterogeneity, that is $s_0(Z) = s$, then

$$\beta_2 = 0.$$

- Rejecting the hypothesis

$$\beta_2 = 0$$

means that there is both heterogeneity and $S(Z)$ is its relevant predictor.

BLP Estimation

- Estimation is done through the empirical analog:

$$Y_i = \widehat{\alpha}' X_{1i} + \widehat{\beta}_1 (D_i - p(Z_i)) + \widehat{\beta}_2 (D_i - p(Z_i))(S_i - \mathbb{E}_{N,M} S_i) + \widehat{\epsilon}_i, \quad i \in M,$$

$$\mathbb{E}_{N,M}[w(Z_i)\widehat{\epsilon}_i \widehat{X}_i] = 0,$$

where $\mathbb{E}_{N,M}$ denote the empirical expectation with respect to the main sample.

- What is nice here is that fixed effects and clustered standards are easily accommodated in this stage!

Post Processing ML Proxies with BLP

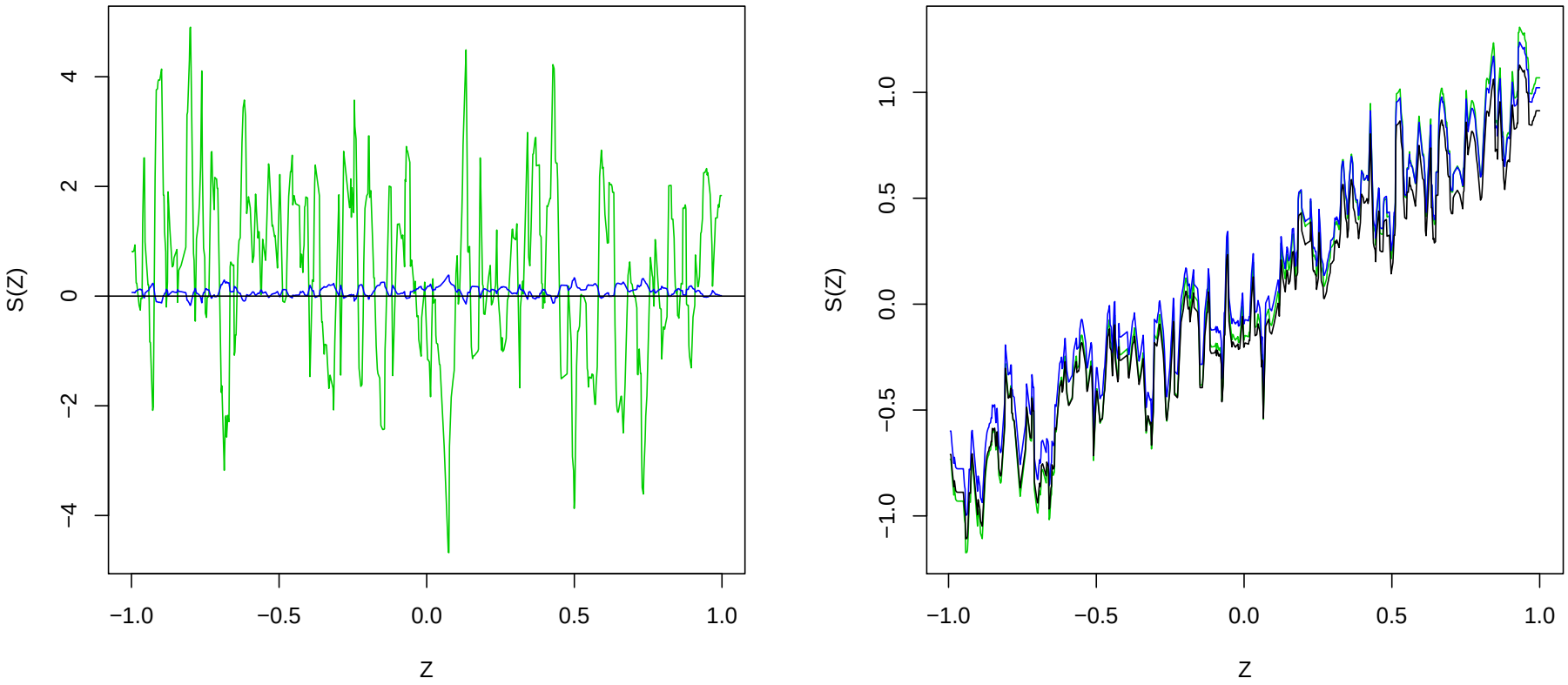


Figure: Left: CATE $s_0(Z) = 0$; Right: CATE $s_0(Z) = Z$; ML proxy $S(Z)$ is produced by Random Forest, shown by green line, BLP is shown by black line, and estimated BLP is shown by blue line.

Digression: Naive Strategy that is not Quite Right

- Recall that
$$\mathrm{E}[Y \mid Z, D] = b_0(Z) + Ds_0(Z),$$
and
$$\begin{aligned}&\mathrm{BLP}[Y - b_0(Z) \mid D, S(Z)]\\ &\qquad = \mathrm{BLP}\{\mathrm{E}[Y \mid Z, D] - b_0(Z) \mid D, S(Z)\} = D\,\mathrm{BLP}[s_0(Z) \mid S(Z)]\end{aligned}$$
- It might be tempting and "more natural" to estimate
$$Y = \tilde{\alpha}_1 + \tilde{\alpha}_2 B + \tilde{\beta}_1 D + \tilde{\beta}_2 D(S - \mathrm{E}S) + \epsilon,$$
- Good for predicting the conditional expectation of Y given Z and D.
- But, $\tilde{\beta}_2 \neq \beta_2$, and $\tilde{\beta}_1 + \tilde{\beta}_2(S - \mathrm{E}S)$ is not the BLP of $s_0(Z)$.

GATES: Group Average Treatment Effects Sorted by ML Proxy

- The target parameters are

$$\mathrm{E}[s_0(Z) \mid G_k],$$

where G_k is an indicator of a group membership.

- We build the groups to explain as much variation in $s_0(Z)$ as possible

$$G_k = \{S \in I_k\}, \quad k = 1, ..., K,$$

where $I_k = [\ell_{k-1}, \ell_k)$ are non-overlaping intervals that divide the support of proxy S into regions $[\ell_{k-1}, \ell_k)$ with equal masses:

$$-\infty = \ell_0 < \ell_1 < \ldots < \ell_K = +\infty.$$

- Can impose the shape restriction

$$\mathrm{E}[s_0(Z) \mid G_1] \leqslant \ldots \leqslant \mathrm{E}[s_0(Z) \mid G_K]$$

  which holds asymptotically if $S(Z)$ is reasonably close to $s_0(Z)$ and the latter has an absolutely continuous distribution.

- Homogeneous effects, if $s_0(Z) = s$, then

$$\mathrm{E}[s_0(Z) \mid G_1] = \ldots = \mathrm{E}[s_0(Z) \mid G_K]$$

GATES Estimation

Consider the weighted linear projection:

$$Y = \alpha' X_1 + \sum_{k=1}^{K} \gamma_k \cdot (D - p(Z)) \cdot 1(G_k) + \nu, \quad \mathrm{E}[w(Z)\nu W] = 0,$$

$$W = (X_1', \{(D - p(Z))1(G_k)\}_{k=1}^{K})'.$$

✓ $D - p(Z)$ in the interaction $(D - p(Z))1(G_k)$ orthogonalizes this regressor relative to all other regressors that are functions of Z.

✓ X_1, e.g. B, is included to improve precision, but can be omitted.

Theorem 2: Projection coefficients identify GATES

$$\gamma_k = \mathrm{E}[s_0(Z) \mid G_k].$$

Examples of γ_k with $s_0(Z) = 0$ and $s_0(Z) = Z$

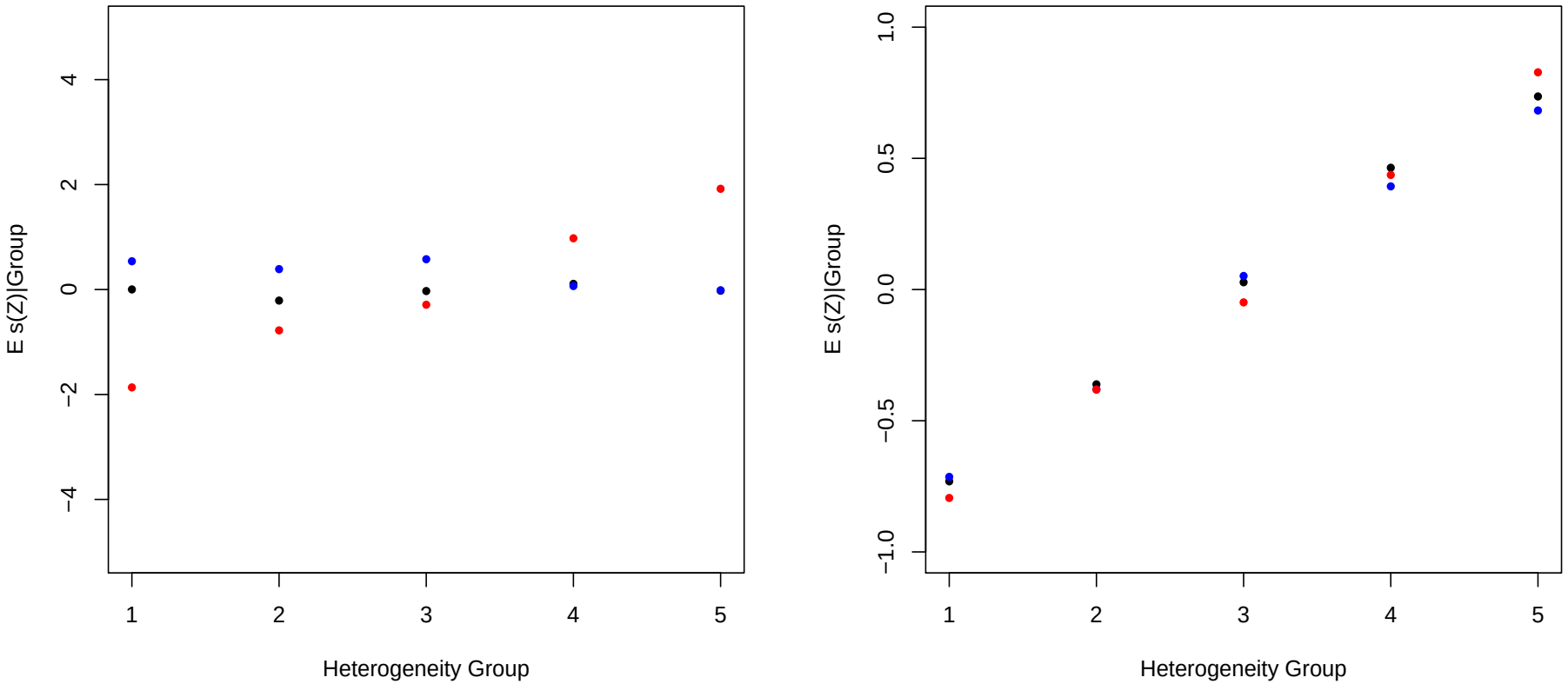


Figure: Left: $s_0(Z) = 0$; Right: $s_0(Z) = Z$; $S(Z)$ is produced by random forest, whose averages over groups are shown in red, the true averages by groups are shown by black dots, and estimated averages are shown by blue dots.

Classification Analysis (CLAN)

- Focus on the "least affected group" G_1 and "most affect group" G_K.
- Let $g(Y, Z)$ be a vector of characteristics of a unit, typically

$$g(Y, Z) = (Y, Z')'$$

- The parameters of interest are the average characteristics of the most and least affected groups:

$$\delta_1 = \mathrm{E}[g(Y, Z) \mid G_1] \quad \text{and} \quad \delta_K = \mathrm{E}[g(Y, Z) \mid G_K].$$

- Compare δ_K and δ_1 to quantify differences between the most and least affected groups.
- δ_K and δ_1 are identified because they are directly observed.

Inference: Target

Let θ denote a generic target parameter or functional, e.g.,

- $\theta = \beta_2$ is the heterogeneity loading parameter;
- $\theta = \beta_1 + \beta_2(S(z) - \mathrm{E}S)$ is the personalized BLP of CATE;
- $\theta = \gamma_k$ is GATE for the group $\{S \in I_k\}$;
- $\theta = \gamma_K - \gamma_1$ is the difference in GATEs between the most and least affected groups;
- $\theta = \delta_K - \delta_1$ is the difference CLAN parameters

Quantification of Uncertainty: Two Sources

- Two sources:
  - (I) Estimation uncertainty regarding the parameter θ, conditional on the data split;
  - (II) Uncertainty induced by data splitting.
- Conditional on the data split, (I) is standard.
- To account for (II), will do many splits and aggregate (how?).

Inference Conditional on a Data Split: Trivial

- Parameters implicitly depend on A, the auxiliary sample, used to create the ML proxies $B = B_A$ and $S = S_A$.
- Make dependence explicit: $\theta = \theta_A$. Unconditionally, this is a random variable.
- All of the examples admit an estimator $\widehat{\theta}_A$ such that

$$\widehat{\theta}_A \mid \mathrm{Data}_A \sim_a N(\theta_A, \widehat{\sigma}_A^2),$$

- Conditional on the split, the confidence interval (CI)

$$[L_A, U_A] = [\widehat{\theta}_A \pm \Phi^{-1}(1 - \alpha/2)\widehat{\sigma}_A]$$

covers θ_A with approximate probability $1 - \alpha$:

$$\mathbb{P}[\theta_A \in [L_A, U_A] \mid \mathrm{Data}_A] = 1 - \alpha - o(1).$$

Unconditional Inference

- Different partitions (A, M) of $\{1, ..., N\}$ yield different targets θ_A and estimators $\widehat{\theta}_A$ with different distributions.
- Splitting risks:
  - (malevolent) researchers may choose to use a favorable split/seed
  - (benevolent) researchers who have honestly fixed the seed in advance are subjected to the risk that the split may be non-typical
  - (arbitrariness) two researchers can draw different conclusions from the same data
- To avoid various risks arising by taking a single split we will rely on multiple splits and take medians over the splits.
- Quantify the uncertainty induced by the random splitting

Accounting for Splitting Uncertainty

- Report the median of $\widehat{\theta}_A$ across random partitions:

$$\widehat{\theta} := \mathrm{Median}[\widehat{\theta}_A \mid \mathrm{Data}].$$

- Report the median CI

$$[l, u] := [\mathrm{Median}[L_A \mid \mathrm{Data}], \mathrm{Median}[U_A \mid \mathrm{Data}]]$$

and discount the confidence level from $1 - \alpha$ to $1 - 2\alpha$.

✓ Unlike $\widehat{\theta}_A$ and $[L_A, U_A]$, $\widehat{\theta}$ and $[l, u]$ are non-random conditional on Data

Main Inference Result

We assume that $\widehat{\theta}_A$ and $[L_A, U_A]$ behave regularly for most realizations of the data and random data splits.

Theorem 3: *Under a mild regularity condition,*

$$\mathbb{P}(\theta_A \in [l, u]) \geqslant 1 - 2\alpha - o(1),$$

where $\mathbb{P}$ is probability measure over data and random partitions.

Splitting uncertainty is reflected in discounting the nominal level of the confidence interval from $1 - \alpha$ to $1 - 2\alpha$.

✓ Similar logic extends to simultaneous confidence bands and p-values (e.g., $\mathrm{Median}[2p_A \mid \mathrm{Data}]$ is *sample-splitting adjusted p-value*)

✓ Key inferential result, could be of interest in other ML applications.

Application to Immunization Incentives in India

- RCT in the state of Haryana for improving immunization coverage
- 3 treatments cross randomized in 903 villages from Dec-16 to Apr-18
  - Incentives: parents/caregivers received mobile phone credit upon bringing children for vaccinations
  - Seeds: information about immunization camps was diffused through key members of a social network
  - Reminders: a fraction of parents/caregivers received phone reminders for pending vaccinations for the children

Experimental Setting

- Most effective program is a 'package' of treatments combining incentives, seeds and reminders
- Look for heterogeneity in the 'package' treatment effect on immunization
  - Y is the number of immunizations per village per month
  - D is an indicator of the 'package' treatment
  - Z are 36 baseline village level characteristics, including religion, caste, financial status, marriage and family status, education, and baseline immunization
- Propensity score is constant, 25 treated villages and 78 control villages

Descriptive Statistics for Selected Variables

	All	Treated	Control
Number of immunizations per village per month	4.48	5.84	4.04
Financial status of households 1-10 scale	3.45	3.23	3.52
Scheduled Caste-Scheduled Tribes households	0.22	0.22	0.22
Other Backward Caste households	0.24	0.17	0.26
Hinduism households	0.92	0.88	0.93
Islam households	0.05	0.06	0.05
Highest level of education nursery	0.15	0.15	0.15
Highest level of education class 4	0.08	0.08	0.08
Highest level of education class 9	0.16	0.16	0.16
Highest level of education class 12	0.25	0.23	0.25
Highest level of education graduate or other diploma	0.09	0.08	0.09
Num vaccines administered to pregnant mother	2.34	2.34	2.34
Num vaccines administered to kids since birth	4.40	4.57	4.34
Num of polio drops	2.99	2.99	2.98
Kids receiving immunization card YN	0.92	0.93	0.91
Number of Observations	1,321	320	1,001

Implementation Details

- Four ML methods: Elastic Net, Boosted Trees, Neural Networks with Feature Extraction, and Random Forest
- Elastic net uses regression specification that includes all two-way interactions between the variables
- Model parameters are tuned by 2-fold cross validation for all ML methods except Random Forest.
- District-time fixed effects and standard errors clustered at the village level
- 250 random splits by half
- Unconditional Inference with $\alpha = 0.05 \rightarrow 10\%$ significance level

Choosing Best ML producing best BLP of CATE

- We propose the measure:

$$\Lambda := |\beta_2|^2 \mathrm{Var}(S(Z)) \propto \mathrm{Corr}^2(s_0(Z), S(Z))$$

  which is proportional to the correlation of CATE and ML proxy.

- Maximizing Λ gives us the best ML proxy.

- Can construct similar measure based on GATES

	Elastic Net	Boosting	Neural Network	Random Forest
Best BLP (Λ)	55.83	24.86	35.67	15.83
Best GATES ($\bar{\Lambda}$)	7.16	4.63	5.28	3.77

Notes: Medians over 250 splits in half.

The winners are Elastic Net and Neural Network

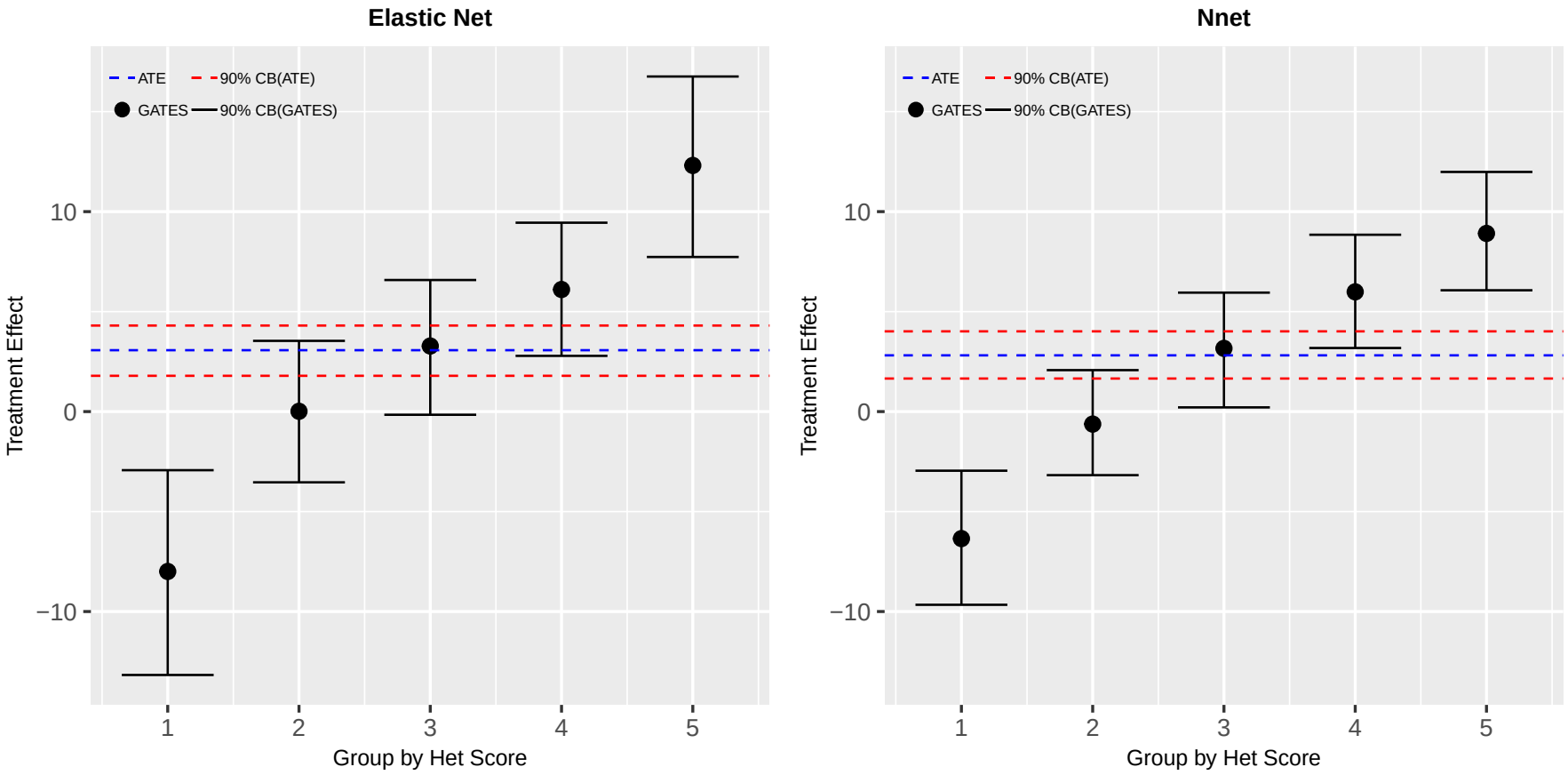
BLP of package treatment CATE

Elastic Net		Nnet	
ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
3.07	**1.09**	**1.90**	**0.92**
(1.79, 4.30)	(0.87, 1.29)	(0.75, 3.02)	(0.73, 1.11)
[0.000]	[0.000]	[0.003]	[0.000]

Notes: Medians over 250 splits. 90% confidence interval in parenthesis. P-values for the hypothesis that the parameter is equal to zero in brackets.

- Package treatment increases number of immunizations by 2-3 on average
- There is detectable strong heterogeneity CATE

GATES by Quintiles of ML Proxies



CIs are simultaneous across groups

GATES of 20% Most and Least Affected Groups

Most Affected (γ_5)	Least Affected (γ_1)	Difference ($\gamma_5 - \gamma_1$)
	Elastic Net	
12.31	-7.96	**20.32**
(8.43, 16.00)	(-12.03, -3.76)	(14.08, 26.35)
[0.000]	[0.000]	[0.000]
	Nnet	
8.72	-6.34	**15.04**
(6.38, 11.15)	(-9.07, -3.56)	(11.18, 18.73)
[0.000]	[0.000]	[0.000]

Notes: Medians over 250 splits. 90% confidence interval in parenthesis. P-values for the hypothesis that the parameter is equal to zero in brackets.

- GATEs are significantly different for Most and Least Affected Groups.

Classification Analysis: Baseline Immunization Variables

	Elastic Net			Nnet		
	20% Most	20% Least	Difference	20% Most	20% Least	Difference
	(δ_5)	(δ_1)	($\delta_5 - \delta_1$)	(δ_5)	(δ_1)	($\delta_5 - \delta_1$)
Num vaccines to pregnant mother	2.12	2.31	-0.10	2.20	2.29	-0.09
	(2.15, 2.25)	(2.26, 2.35)	(-0.17, -0.04)	(2.15, 2.24)	(2.25, 2.33)	(-0.15,-0.04)
	-	-	[0.003]	-	-	[0.003]
Num vaccines to kids since birth	4.11	4.65	-0.51	4.33	4.70	-0.37
	(3.97, 4.25)	(4.52, 4.78)	(-0.70, -0.32)	(4.22, 4.44)	(4.58, 4.81)	(-0.53, -0.22)
	-	-	[0.000]	-	-	[0.000]
Num of polio drops	2.95	2.99	-0.05	2.96	3.00	-0.04
	(2.93, 2.96)	(2.98, 3.01)	(-0.07, -0.03)	(2.95, 2.97)	(2.99, 3.01)	(-0.06, -0.03)
	-	-	[0.000]	-	-	[0.000]
Kids with immunization card YN	0.81	0.93	-0.12	0.91	0.93	-0.03
	(0.78, 0.84)	(0.90, 0.95)	(-0.16, -0.08)	(0.89, 0.92)	(0.91, 0.95)	(-0.05,-0.01)
	-	-	[0.000]	-	-	[0.019]

- Villages with low levels of pretreatment immunization are most affected by the incentives

Classification Analysis: Socioeconomic Variables

	Elastic Net			Nnet		
	20% Most	20% Least	Difference	20% Most	20% Least	Difference
	(δ_5)	(δ_1)	($\delta_5 - \delta_1$)	(δ_5)	(δ_1)	($\delta_5 - \delta_1$)
Financial status of households 1-10 scale	3.345	3.683	-0.341	3.459	3.750	-0.255
	(3.167, 3.513)	(3.513, 3.854)	(-0.565, -0.107)	(3.304, 3.616)	(3.598, 3.903)	(-0.479, -0.038)
	-	-	[0.009]	-	-	[0.043]
Scheduled Caste-Scheduled Tribes households	0.174	0.126	0.050	0.189	0.139	0.047
	(0.148, 0.201)	(0.100, 0.151)	(0.013, 0.087)	(0.162, 0.217)	(0.114, 0.164)	(0.011, 0.086)
	-	-	[0.014]	-	-	[0.023]
Other Backward Caste households	0.276	0.154	0.124	0.335	0.168	0.169
	(0.243, 0.309)	(0.123, 0.185)	(0.078, 0.170)	(0.305, 0.367)	(0.139, 0.196)	(0.126, 0.212)
	-	-	[0.000]	-	-	[0.000]
Islam households	0.165	0.026	0.135	0.020	0.020	0.005
	(0.119, 0.210)	(-0.014, 0.066)	(0.071, 0.198)	(0.009, 0.037)	(0.003, 0.046)	(-0.010, 0.020)
	-	-	[0.000]	-	-	[1.000]
Highest level of education nursery	0.152	0.170	-0.017	0.133	0.168	-0.033
	(0.144, 0.162)	(0.162, 0.178)	(-0.029, -0.005)	(0.127, 0.140)	(0.162, 0.175)	(-0.043, -0.024)
	-	-	[0.014]	-	-	[0.000]
Highest level of education class 4	0.079	0.090	-0.011	0.079	0.090	-0.011
	(0.074, 0.084)	(0.085, 0.095)	(-0.018, -0.005)	(0.074, 0.084)	(0.085, 0.094)	(-0.018, -0.004)
	-	-	[0.002]	-	-	[0.004]
Highest level of education class 9	0.171	0.160	0.010	0.161	0.155	0.006
	(0.164, 0.179)	(0.153, 0.167)	(0.000, 0.021)	(0.155, 0.168)	(0.148, 0.162)	(-0.004, 0.015)
	-	-	[0.095]	-	-	[0.451]
Highest level of education class 12	0.208	0.224	-0.013	0.246	0.225	0.020
	(0.195, 0.220)	(0.213, 0.235)	(-0.029, 0.003)	(0.235, 0.257)	(0.215, 0.235)	(0.006, 0.034)
	-	-	[0.219]	-	-	[0.009]

- The most affected group is more likely to live in "poorer" villages with higher proportion of muslim and backward cast households

Literature

I. **Orthogonalized/Double ML**:

- Chernozhukov, Chetverikov, Demirer, Duflo, Newey, Robins (17, Econometrics Journal)
- Belloni, Chernozhukov, Hansen (14, ReStud): double selection
- Belloni, Chernozhukov, Wang (14, Annals): partialling out

II. **Heterogenous Effects:**

- **Using Trees**: Athey and Imbens (15, PNAS) – like ours, assumption free, but limited to ATE for tree leaves; no accounting for splitting uncertainty; Wager and Athey (16) on forests, restricted only to low-d cases.
- **Using Sparsity**: Hansen Kozbur, Misra (17); Belloni, Chernozhukov, Kato (14, Biometrika, high-dimensional treatments); restrictive assumptions;
- **Partial Sparsity**: D. Small et al. (17); Chernozhukov, Goldman, Semenova, Taddy (17); somewhat less restrictive assumptions.

Concluding Remarks

- Propose generic, assumption-free strategies to make inference on key features of heterogeneous effects in randomized experiments.
- Key features include BLP, GATEs, and CLAN.
- Estimation and inference relies on repeated data splitting to avoid overfitting.
- Valid inference quantifies uncertainty coming from parameter estimation and data splitting.
- We are talking to the government about the immunization experiment. They are considering using the results to implement a program

Application to Micro-Credit in Morocco

- Randomized experiment in Morocco to measure the impact of microfinance on outcomes (Crépon et al (2015))
- 162 villages with $N \approx 5000$ households in rural areas are divided into 81 pairs.
- One treatment and one control village were randomly assigned within each pair. In treated villages a microfinance institution opened branches
- Introduced in 2006, outcomes from follow-up surveys in 2009.
- Y is profit; D is indicator of offering access to microfinance services; Z are 22 household characteristics including the number of household members, number of adults, head age and 81 village pair fixed effects.
- Standard errors are clustered at the village level.

Choosing Best ML producing best BLP of CATE

We propose the measure:

$$\Lambda := |\beta_2|^2 \mathrm{Var}(S(Z)) \propto \mathrm{Corr}^2(s_0(Z), S(Z))$$

which is proportional to the correlation of CATE and ML proxy.
Maximizing Λ gives us the best ML proxy.

	Elastic Net	Boosting	Nnet	Random Forest
Profit (Λ)	32,462,874	16,674,642	13,411,383	43,184,732

Notes: Medians over 1,000 splits.

The winners are the Elastic Net and Random Forest.

BLP of the Effect of Microfinance on Profits

Table: BLP of the Effect of Microfinance on Profits

	Elastic Net		Random Forest	
	ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
Profit	1,554	0.243	1,625	0.275
	(-1,344, 4,388)	(0.079, 0.416)	(-1,332, 4,576)	(0.036,0.510)
	[0.584]	[0.008]	[0.577]	[0.045]

Median estimates, CIs, and p-values computed over 1000 splits.

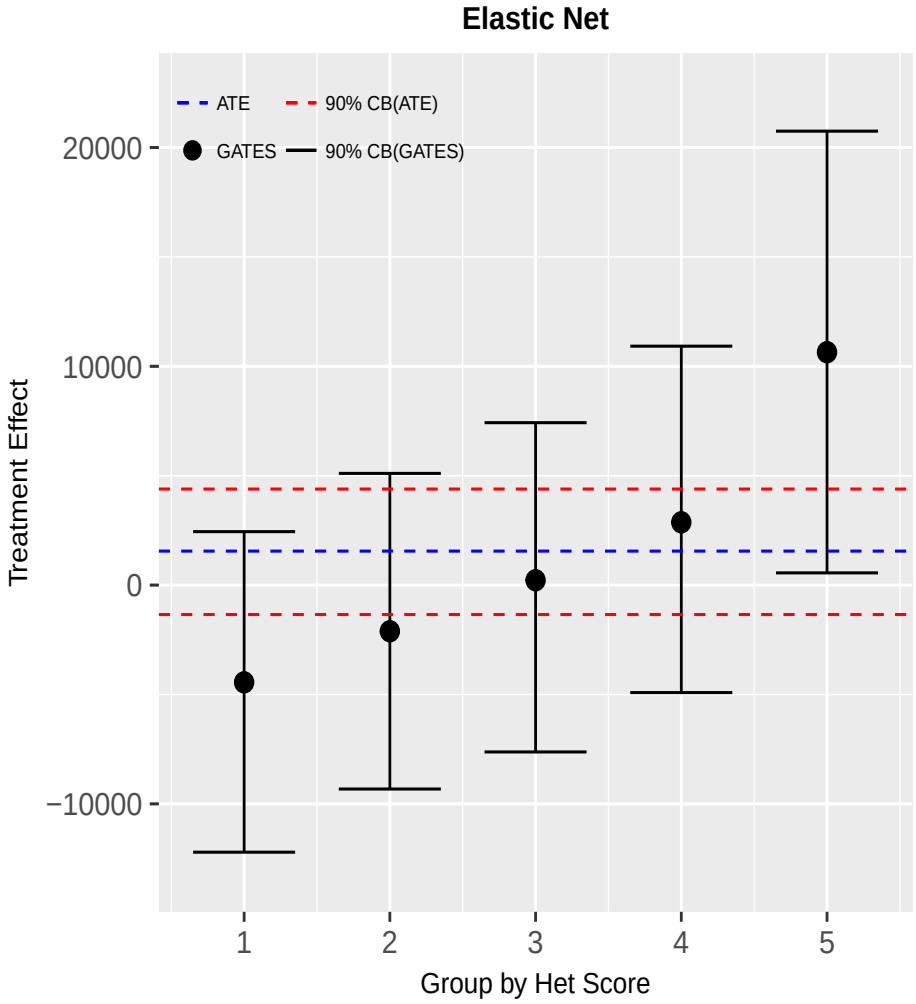
- There is detectable heterogeneity in Profits.

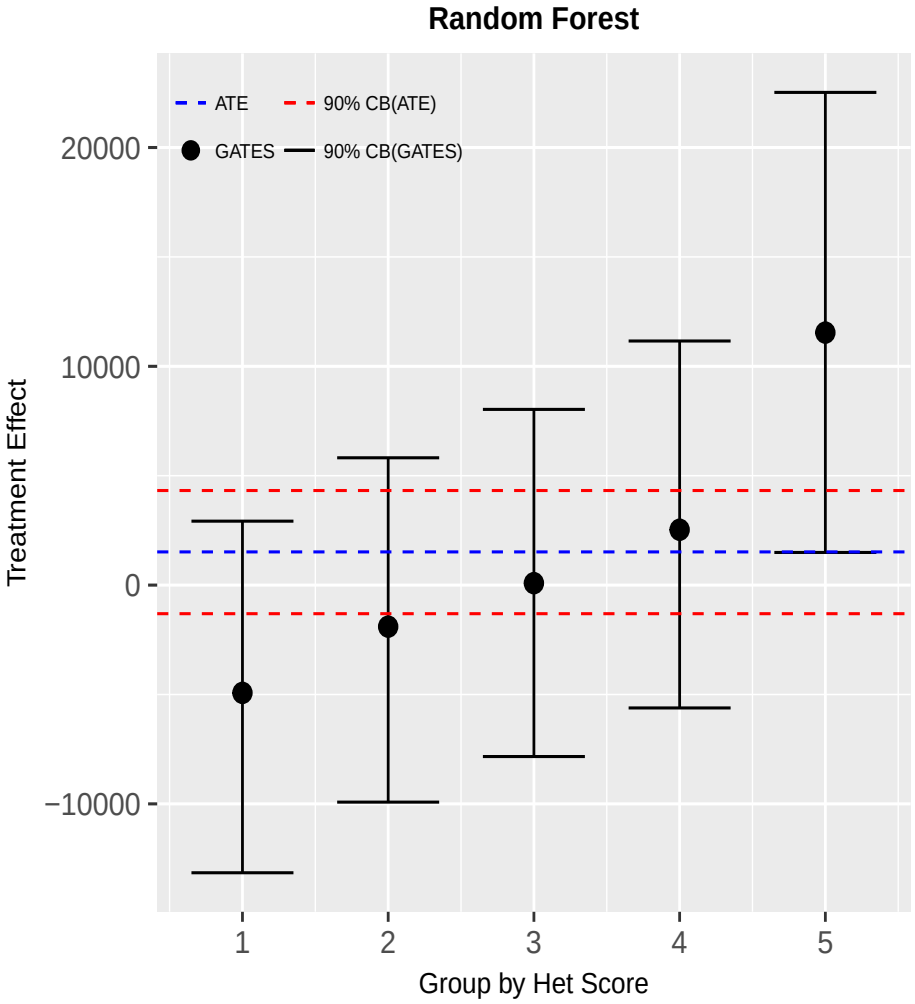
GATES of Microfinance on Profits

	Elastic Net			Random Forest		
	Most Affected (γ_5)	Least Affected (γ_1)	Difference ($\gamma_5 - \gamma_1$)	Most Affected (γ_5)	Least Affected (γ_1)	Difference ($\gamma_5 - \gamma_1$)
Profit	10,707	-1,227	**11,768**	12,000	-2,130	**14,056**
	(1,628, 19,032)	(-7,273, 5,003)	(1,186, 22,485)	(2,911, 20,638)	(-9,135, 4,853)	(2,292, 25,698)
	[0.028]	[1.000]	[0.059]	[0.018]	[1.000]	[0.035]

- GATEs are Dramatically Different for Most and Least Affected Groups.

ATE and GATES of Microfinance on Profits





CIs simultaneous across groups.

Classification Analysis for Microfinance Effects

	Elastic Net			Random Forest		
	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)
Profit						
Non-agricultural self-emp.	0.198	0.103	0.086	0.186	0.108	0.074
	(0.169, 0.227)	(0.073, 0.132)	(0.046, 0.127)	(0.156, 0.215)	(0.079, 0.138)	(0.033, 0.115)
	-	-	[0.000]	-	-	[0.001]
Animal Husbandry self-emp.	0.321	0.570	-0.243	0.378	0.483	-0.113
	(0.280, 0.361)	(0.529, 0.610)	(-0.300, -0.186)	(0.336, 0.419)	(0.442, 0.525)	(-0.171, -0.054)
	-	-	[0.000]	-	-	[0.000]
Head Age	34.11	39.99	-6.08	31.83	35.77	-4.20
	(32.06, 36.18)	(37.90, 42.06)	(-9.05, -3.10)	(29.56, 34.14)	(33.52, 37.99)	(-7.29, -1.10)
	-	-	[0.000]	-	-	[0.017]

- The Most Affected Group tends to be Younger Households with Less Borrowing Experience.