Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments

Victor Chernozhukov
MIT

Mert Demirer
MIT

Esther Duflo
MIT

Iván Fernández-Val
BU

Available via www.ArXiv.org (Econometrics!)

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, much better than classical methods in the 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

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

  By the way, $Log(100,000) = 5$**.**

  Adaptive Possibilities Under Structured Sparsity: Consistency is possible under *structured forms of linear and nonlinear sparsity.*

  - The assumption is untestable, so must be used with caution.

- **Valid Adaptive Confidence Sets Do Not Exist**(Lou 90, Genovese and Wasserman 90; Annals). This has to do with bias dominating the behavior of adaptive estimators.

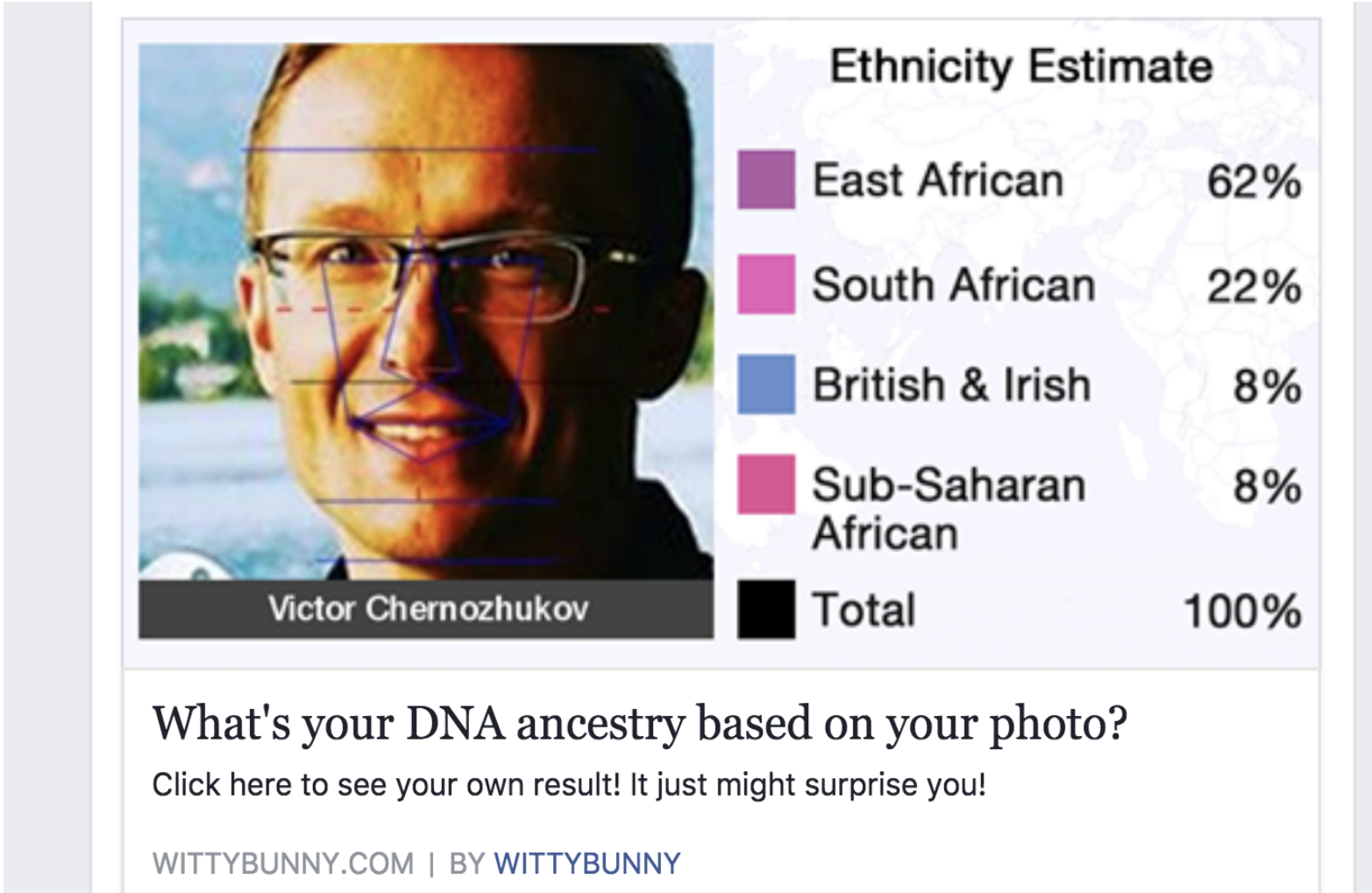
  Can do partly adaptive confidence sets using the assumptions of self-similarity (Gine and Nickl 2011, C. Chetverikov, Kato 2013; Annals), allowing to bound bias or undersmooth.

  - It remains unclear how to define self-similarity and do bias-bounding/undersmoothing for high-dimensional problems with structured sparsity.

A Fundamental Theory-Practice Gap for ML

- 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



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

- We shall rely on the random data splitting into the 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 Machine Learning** 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 about *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:

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

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.[1]

Theorem 1: Projection coefficients identify the BLP of CATE:

$$\beta_1 + \beta_2 (S(Z) - \mathrm{E}S) = \mathrm{BLP}[s_0(Z) \mid 1, S(Z)],$$

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

[1] Like our DML paper, but does not require consistency/allows misspecification!

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.

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 with BLP: Examples of with $s_0(Z) = 0$ and $s_0(Z) = Z$

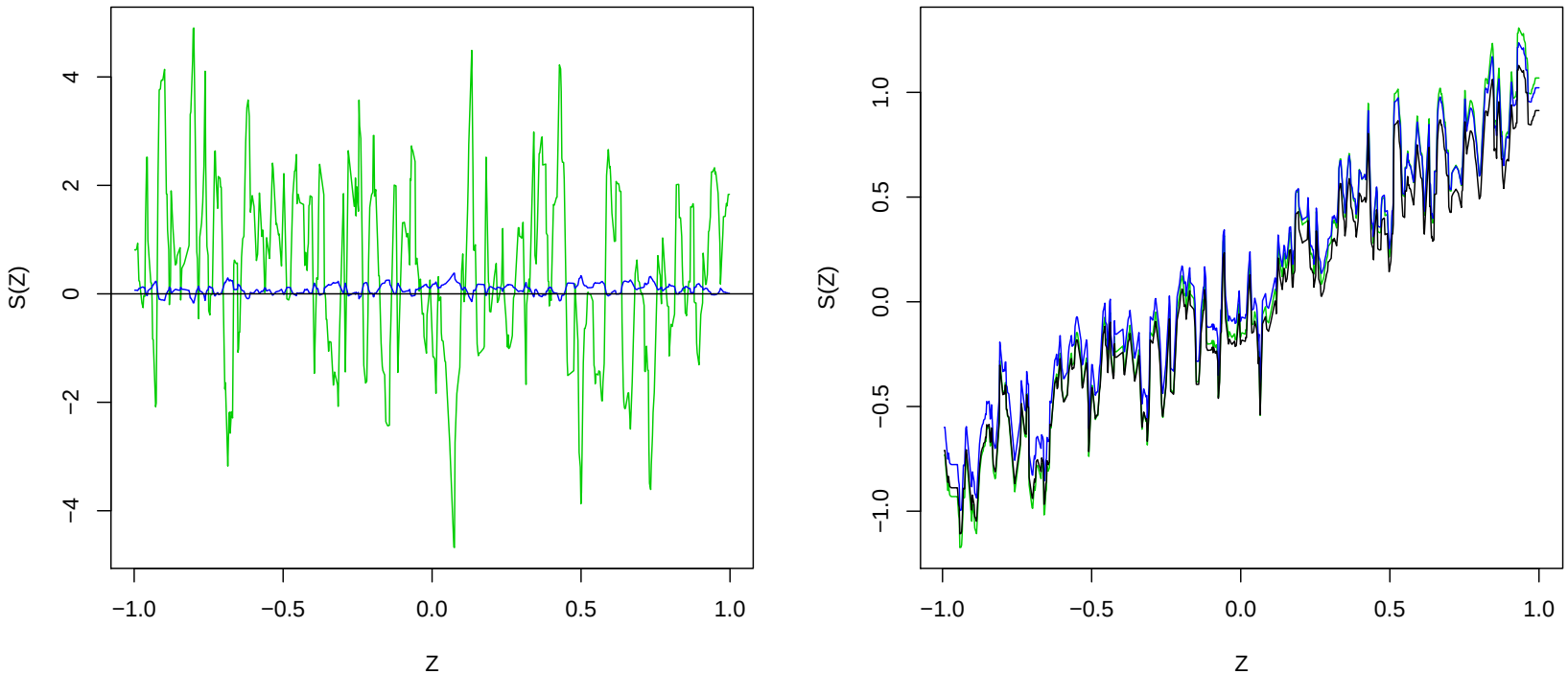


Figure: Left: CATE $s_0(Z) = 0$; Right: CATE $s_0(Z) = Z$; ML proxy $S(Z)$ is produced by Random Forest, shown by green line, BLP is shown by black line, and estimated BLP is shown by blue line.

Digression: Naive Strategy that is not Quite Right

- It is 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 Heterogeneity Proxies

- 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

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

Average $s_0(Z)$ by Groups

- Consider the weighted linear projection:

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

$$W = (X_1', W_2')' = (X_1', \{(D - p(Z))1(G_k)\}_{k=1}^{K})'.$$

- $D - p(Z)$ 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 $s_0(Z) = Z$

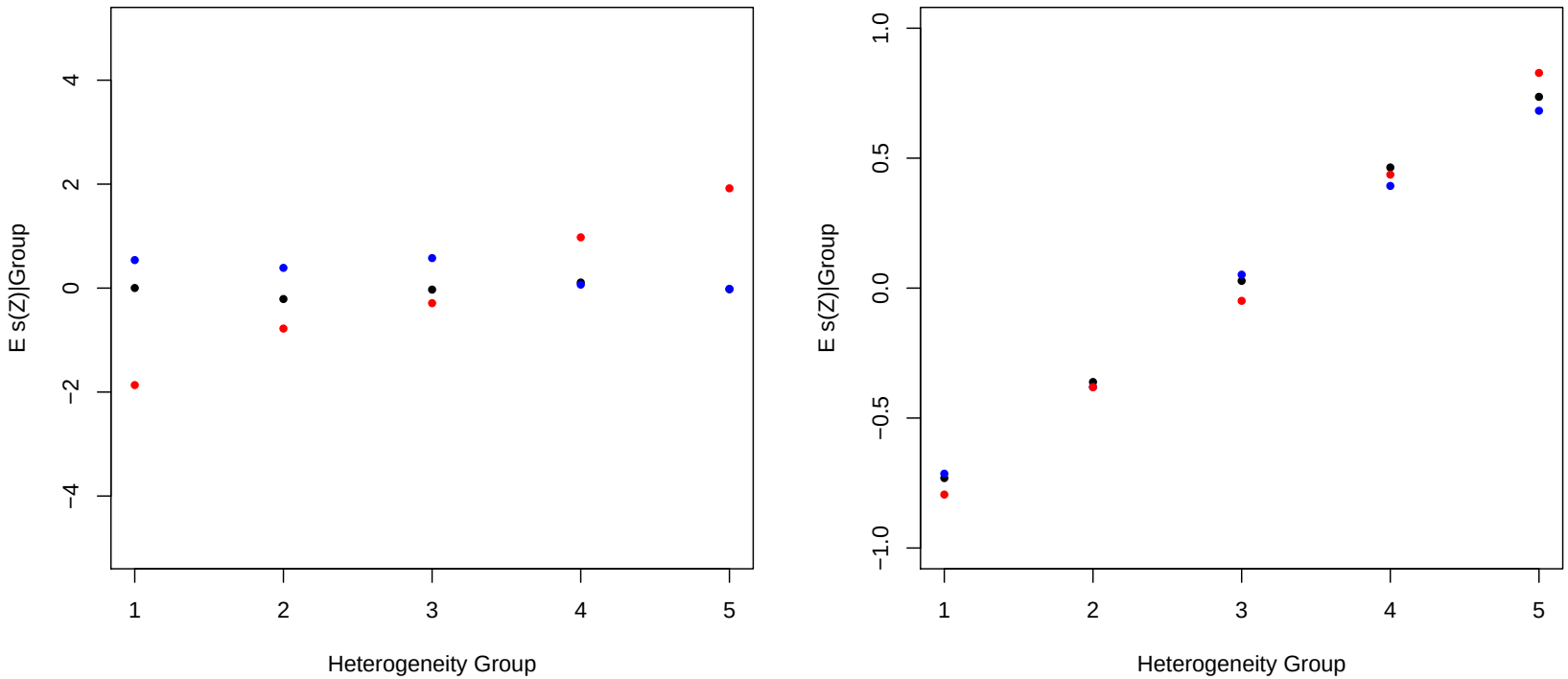


Figure: Left: $s_0(Z) = 0$; Right: $s_0(Z) = Z$; $S(Z)$ is produced by random forest, whose averages over groups are shown in red, the true averages by groups are shown by black dots, and estimated averages are shown by blue dots.

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.
- 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 the 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: Accounting for Splitting Uncertainty

- Different partitions (A, M) of $\{1, ..., N\}$ yield different targets θ_A and estimators $\widehat{\theta}_A$ with different distributions.
- To avoid various risks arising by taking a sinlge split we will rely on multiple splits and take medians over the splits.
- Quantify the uncertainty induced by the random splitting.

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

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

- Report the median CI

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

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

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 p-values and simultaneous confidence bands.

This is a **key inferential result**, which could be of independent interest in numerous ML applications.

Application to Morocco Micro-Credit Data (Crépon et al(2015))

- Randomized experiment in Marocco to measure the impact of microfinance on outcomes.
- 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:

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

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)	32307828	17105855	20404000	39286050

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	1553	0.244	1603	0.279
	(-1344,4389)	(0.079,0.416)	(-1276.,4536)	(0.046,0.518)
	[0.584]	[0.008]	[0.521]	[0.039]

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

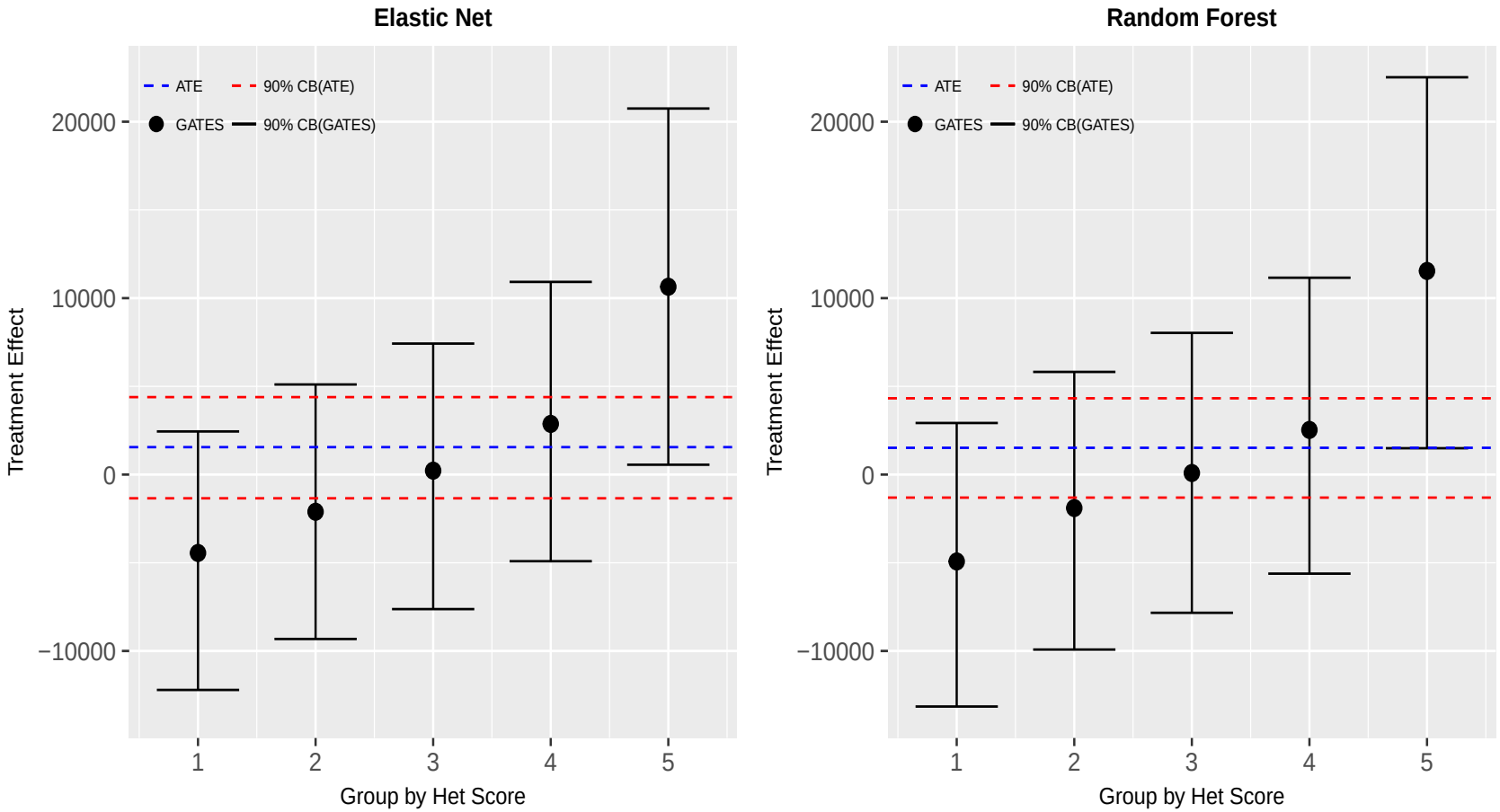
- There is detectable heterogeneity in Profits.

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	10644.939	-1152.242	**11768**	11540.	-2031	**14037**
	(2146,19096)	(-7250,4952)	(1077,22422)	(2965,20955.576)	(-8721,4796)	(2459,25833)
	[0.028]	[1.000]	[0.061]	[0.014]	[1.000]	[0.037]

- 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		
	10 % Most	10 % Least	Difference	10 % Most	10 % Least	Difference
Profit						
Head Age	**34.1**	**40.4**	**-6.5**	29.2	33.7	**-5.8**
	(31.2,37.0)	(37.5,43.4)	(-10.7,-2.5)	(25.7,32.6)	(30.390,37.108)	(-10.566,-1.217)
	-	-	[0.003]	-	-	[0.029]
Non-agricultural self-emp.	0.181	0.108	0.082	0.153	0.099	0.051
	(0.140,0.222)	(0.068,0.149)	(0.022,0.138)	(0.113,0.192)	(0.058,0.139)	(-0.003,0.105)
	-	-	[0.014]	-	-	[0.129]
Borrowed from Any Source	**0.180**	**0.257**	**-0.091**	0.144	0.162	-0.032
	(0.130,0.230)	(0.207,0.307)	(-0.160,-0.022)	(0.098,0.190)	(0.122,0.206)	(-0.095,0.029)
	-	-	[0.020]	-	-	[0.578]

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

Literature

I. **Orthogonalized/Double ML**. A very nuanced continuation of our work on DML where orthogonalization is key:
   1. Chernozhukov, Chetverikov, Demirer, Duflo, Newey, Robins (2016, Econometrics Journal 2017))
   2. Belloni, Chernozhukov, Hansen (2011, ReStud, 2014): double selection
   3. Belloni, Chernozhukov, Wang (2012, Annals, 2014): partialling out

II. **Heterogenous Effects:**
   1. **Using Trees**: Athey and Imbens (2015, PNAS) – like ours, assumption free, but limited to ATE for tree leaves; no accounting for splitting uncertainty; Wager and Athey (2016) on forests, restricted only to low-d cases.
   2. **Using Sparsity**: Hansen Kozbur, Misra (2017); Belloni, Chernozhukov, Kato (Biometrika, 2014, high-dimensional treatments); restrictive assumptions;
   3. **Partial Sparsity**: D. Small et al paper (2017); Semenova, Goldman, Taddy, C. (2017); 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.