

ML-econometrics reading group presentation

Federico Nutarelli¹

¹ Bocconi University

Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India

Chernozhukov V., Demirer M., Duflo E., Fernández-Val I.

Introduction

Main purpose: "propose strategies to estimate and make inference on **key features** of heterogeneous treatment effects (HTE) in randomized experiments".

Notice: not directly on HTE but on **key features** of HTE including:

- ▶ Best Linear Predictor (BLP);
- ▶ Average Sorted Effects by impact groups (GATES);
- ▶ Average *characteristics* of most and least impacted units (CLAN).

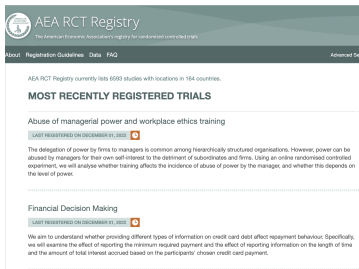
Why is ML useful in randomized experiments?

Accidental imbalances that cannot be foreseen if not with ML tools.

Traditional problems of the literature:

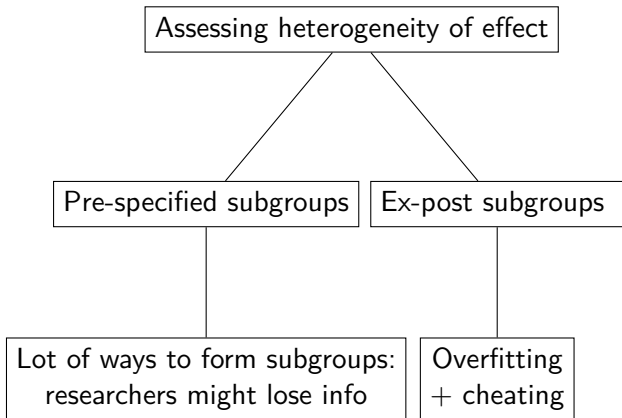
- ▶ Imbalances of the data which requires variable selection (e.g. LASSO). Example: prediction of natural disaster, medical datasets (e.g. more smokers than not), country datasets (e.g. more data on developed countries)...
- ▶ Interested in how does the effect change according to subgroups of population defined on characteristics (see next slide). Example:

Problem: how to define such subgroups in advance?



The screenshot shows the AEA RCT Registry website. At the top, it says "AEA RCT Registry" and "The American Economic Association's registry for randomized controlled trials". Below that, there are navigation links: "About", "Registration Guidelines", "Data", "FAQ", and "Advanced Search". A line of text states: "AEA RCT Registry currently lists 6993 studies with locations in 164 countries." Below this is a section titled "MOST RECENTLY REGISTERED TRIALS". Two trials are listed:

- Abuse of managerial power and workplace ethics training**
LAST REGISTERED ON DECEMBER 01, 2015
The delegation of power by firms to managers is common among hierarchically structured organizations. However, power can be abused by managers for their own self-interest to the detriment of subordinates and firms. Using an online randomized controlled experiment, we will analyse whether training affects the incidence of abuse of power by the managers, and whether this depends on the level of power.
- Financial Decision Making**
LAST REGISTERED ON DECEMBER 01, 2015
We aim to understand whether providing different types of information on credit card debt affect repayment behaviour. Specifically, we will examine the effect of reporting the minimum required payment and the effect of reporting information on the length of time and the amount of total interest accrued based on the participants' chosen credit card payment.



Idea: let the data speak and employ ML to estimate HTE directly!

Problems of estimating HTE directly with ML

While ML tools are successful in prediction empirically, it is much more difficult to obtain uniformly valid inference for CATE (and hence HTE, see Yuang et al. 2008), i.e. inference that remains valid under a large class of data generating processes (DGP) in high dimensional settings.

People: *Dude you can regularize and reduce dimensionality!*

Chernuz.: *No way...I wanna stay agnostic (see [▶ Agnostic property](#))*

Workaround

Not estimating HTE directly but, rather, features of it: BLP of the CATE on the ML proxy predictors, GATES (ATE by heterogeneity groups induced by the ML proxy predictor), CLAN (see Chernuzokov et al. (2019) for the R routine).

Agnostic property

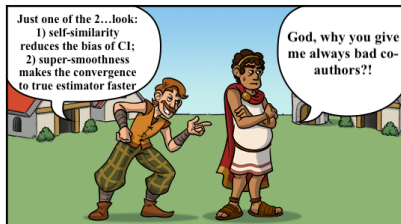
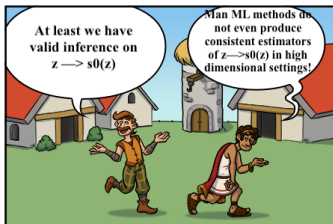
#3 HOLD YOUR HORSES



Sparsity cannot be employed for general ML tools: we must be agnostic if we wanna stay general!

Agnostic property (pt.2)

#5 FUNTIME WITH FRIENDS



Super-smoothness and self-similarity cannot be employed for general ML tools: stay agnostic!

Basic clarifications before math

- ▶ $HTE \neq CATE$ but $CATE \implies HTE$ (often): CATE is employed to assess HTE non-parametrically (see e.g. Athey et al., 2019).
- ▶ Why, in general, people use ML to construct proxies of CATE rather than estimating CATE with usual econometric techniques? Because the essence of the HTE and the CATE is to perform a non-parametric (i.e. without making assumptions on the underlying population) estimation (agnostic). Moreover, (generally) econometric techniques do not catch HTE. Why? HTE analysis often requires highly dimensional data ($n < p$) which invalidates usual asymptotics. Remember: HTE analysis looks for $\theta(X)$ and not θ alone generally. Hence, HTE analysis uses ML and CATE is estimated non-parametrically via ML.

Basic clarifications part 2

- ▶ Term "*proxies*" might be misleading. Take it as "*estimations using ML*".
- ▶ Why valid inference for general DGP is difficult using ML?
Because in high dimensional settings ML tools might produce inconsistent estimates of CATE (ATE conditional on covariates). Why? Intuitively think about LASSO. LASSO selects the covariates that are more useful for prediction of Y and not necessarily those producing consistent estimates of θ .
- ▶ Authors look for (valid) inference and (valid) estimation for **generic** ML tools. In the literature you find examples of consistent and efficient estimators adopting **specific** ML tools (e.g. RF for Athey et al., 2016...), see again ▶ Agnostic property.

Theory

Framework

- ▶ Baseline Conditional Average (BCA):

$$b_0 := E[Y(0)|Z]$$

- ▶ CATE:

$$s_0 := E[Y(1)|Z] - E[Y(0)|Z]$$

- ▶ Propensity score:

$$p(Z) := P[D = 1|Z]$$

- ▶ Observed outcome (given ANOVA):

$$Y = b_0(Z) + Ds_0(Z) + U, \quad E[U|Z, D] = 0,$$

$$s_0(Z) = E[Y|D = 1, Z] - E[Y|D = 0, Z]$$

This results from $Y = DY(1) + (1-D)Y(0)$

Formalize agnostic approach

- ▶ Split the sample: (A, M) s.t. $Data_A = (Y_i, D_i, Z_i)_{i \in A}$, $Data_M = (Y_i, D_i, Z_i)_{i \in M}$
- ▶ **Stage 1:** From A obtain estimates of $b_0(z)$ and $s_0(z)$ (proxy predictors):

$$z \rightarrow B(z) = B(z; Data_A) \text{ and } z \rightarrow S(z) = S(z; Data_A)$$

Not required consistency for $S(z)$ and $B(z)$

- ▶ **Stage 2:** Post-process the proxies from Stage 1 to estimate and make inference on features (BLP, GATES, CLAN) of the CATE $z \rightarrow s_0(z)$ in the main sample M .

We can now better formalize the features of CATE using the notation above:

1. BLP of the CATE $s_0(Z)$ on the ML proxy predictor $S(Z)$;
2. GATES: average of $s_0(Z)$ (ATE) by heterogeneity groups induced by the ML proxy predictor $S(Z)$;
3. CLAN: average characteristics of the most and least affected units defined in terms of the ML proxy predictor $S(Z)$.

Inference will account for two sources of uncertainty:

1. **Estimation uncertainty** conditional on the auxiliary sample;
2. **Splitting uncertainty** induced by random partitioning of data into A and M .

Results (BLP)

We want to identify and estimate a linear predictor $f(Z)$ (i.e., s.t. $\text{Span}(1, S(Z))$) of $s_0(z)$ using $S(Z)$ (i.e., the linear function includes $S(Z)$):

$$\text{BLP}[s_0(Z) | S(Z)] := \arg \min_{f(Z) \in \text{Span}(1, S(Z))} \mathbb{E}[s_0(Z) - f(Z)]^2,$$

which is identified via 2 strategies:

- ▶ Via manipulation of the observed outcome's equation
(▶ Observed outcome)
- ▶ Using Horvitz-Thompson transformation (skipped here)

BLP (First strategy)

Identify the coefficients of the following linear projection:

$$Y = \alpha' X_1 + \beta_1(D - p(Z)) + \beta_2(D - p(Z))(S(Z) - E[S(Z)]) \quad (1)$$

Where does it come from and how can this be employed to find BLP?

Let's recover it! (my intuition \rightarrow feel free to criticize)

- ▶ Start from $Y = Y(0) + D(Y(1) - Y(0))$, where $Y(1) - Y(0)$ is $s_0(Z)$;
- ▶ Idea: we want an $f(Z)$, linear on $S(Z)$ s.t. $E[s_0(Z) - f(Z)]^2$ is minimized. General form of $f(Z) = \beta_1 + \beta_2 S(Z)$.
- ▶ Substitute the general form of $f(Z)$ in Y and obtain:
$$Y = Y(0) + D(\beta_1 + \beta_2 S(Z)) = Y(0) + D\beta_1 + D\beta_2(S(Z))$$
- ▶ Apply local centering of D and $S(Z)$, i.e.
 $D - E[D] = D - p(Z)$ and $S(Z) - E[S(Z)]$ to obtain the equation.

Fundamental theorem:

Theorem 3.1 (BLP 1). Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. Assume that Y and X have finite second moments, EXX' is full rank, and $\text{Var}(S(Z)) > 0$. Then, (β_1, β_2) defined in (3.1) also solves the best linear predictor/approximation problem for the target $s_0(Z)$:

$$(\beta_1, \beta_2)' = \arg \min_{b_1, b_2} E[s_0(Z) - b_1 - b_2(S(Z) - ES(Z))]^2,$$

in particular $\beta_1 = Es_0(Z)$ and $\beta_2 = \text{Cov}(s_0(Z), S(Z)) / \text{Var}(S(Z))$.

In other words, $\beta_1 + \beta_2(S(Z) - E[S(Z)]) = BLP[s_0(Z)|S(Z)]$. The fact that the latter is the BLP can be shown by proving that it solves the normal equations (not covered here). Can we show that the optimal coefficients are the ones shown in Th. 3.1? (see next slide).

My informal proof ¹

⇒ Notice this: the loss function

$$L(Z) = E[s_0(Z) - f(Z)]^2 = E[s_0(Z) - \beta_1 + \beta_2 S(Z)]^2$$

is a Least square loss and has the same solution as $E[Y - \beta X]^2$ in linear regression.

⇒ Thus, minimizing $L(Z)$ we would obtain the same β s obtained as the solution of the coefficients of the linear regression (1), being X , in our case, represented by $S(Z)$. In other words we expect that:

$$\begin{aligned} \frac{\partial L(Z)}{\partial \beta_1} &= 0, \text{ gives us } E[s_0(Z)]; \\ \frac{\partial L(Z)}{\partial \beta_2} &= 0, \text{ gives us } \frac{\text{Cov}(s_0(Z), S(Z))}{\text{Var}(S(Z))} \quad \text{c.v.d.} \end{aligned}$$

¹Refer to the paper for a formal proof

Results (GATES)

Target parameter:

$$E[s_0(Z)|G]$$

being G an indicator for group membership (groups based upon ML tools applied to the auxiliary data).

Impose **monotonicity** condition:

$$E[s_0(Z)|G_1] \leq E[s_0(Z)|G_2] \leq \dots \leq E[s_0(Z)|G_K]$$

GATES are recovered from the weighted linear projection

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

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

The non-overlapping groups are constructed to explain as much variation in $s_0(Z)$ as possible.

Again: where does this weighted linear projection come from?
Two ways to reach it...(my intuition \rightarrow feel free to criticize)

1) Start from Eq. (1).

Remember that we want

$E[s_0(Z)|G]$. Now "given G "

means that:

- ▶ We are operating within the same group G_k , i.e. we need a dummy $1(G_k)$.
- ▶ By assumption, within the same group the effect $s_0(Z)$ is the same, hence, in Eq. (1), $S(Z) - E[S(Z)] = 0$ within $G_k \quad \forall k \in \{1, K\}$.
- ▶ Without $S(Z) - E[S(Z)] = 0$, β_1 represents $E[s_0(Z)|G] = \gamma_k \cdot 1(G_k) \quad \forall k \in \{1, K\}$.

2) Start from

$Y = Y(0) + D(Y(1) - Y(0)) \simeq b_0(Z) + D \cdot s_0(Z)$. Condition to G : $E[Y|G] \simeq$

$E[b_0(Z)|G] + D \cdot E[s_0(Z)|G]$.

Now:

- ▶ $E[Y_0(Z)|G] \simeq E[b_0(Z)|G] = E[b_0(Z)]$ will only depend on X^* and not on the group, G , hence, putting them all in the same vector X_1 , $E[b_0(Z)|G] = E[b_0(Z)] = \alpha' X_1$.
- ▶ Define $E[s_0(Z)|G] = \gamma_k \cdot 1(G_k) \quad \forall k \in \{1, K\}$.
- ▶ Apply local centering.

* This is so because groups are defined in order to maximize the variability of $s_0(Z)$.

Hence, they are defined on the support of $S(Z)$ (and not on the one of $B(Z)$, the estimator of $b_0(Z)$).

The implicit (plausible) assumption of the authors is that the grouping scheme(s) that maximize the variability of $s_0(Z) \simeq Y(1) - Y(0)$, do not (necessarily) influence the variability of $b_0(Z) \simeq Y(0)$. Is the variability of the difference that counts and not that of $Y(0)$ in the scheme!

ANOVA contributes in guaranteeing so (random assignment in treatment given Z).

GATES (theorems)

Theorem 1

The projection coefficients γ_k are the GATES parameter:

$$\gamma = (\gamma_k)_{k=1}^K = (E[s_0(Z)|G_k])_{k=1}^K$$

Proof: simply FWL theorem!

Theorem 2

Given a bunch of assumptions (see paper), γ is an efficient estimator of GATES, i.e.

$$\gamma_k = E[s_0(Z)|G_k]$$

asymptotically.

Results (CLAN)

When GATES reveals substantial heterogeneity it is interesting to know the properties of most and least affected groups.

Call G_1 the least affected group and G_K the most affected (remember **monotonicity** assumption).

Let $g(\cdot)$ be a vector of characteristics of a unit s.t.

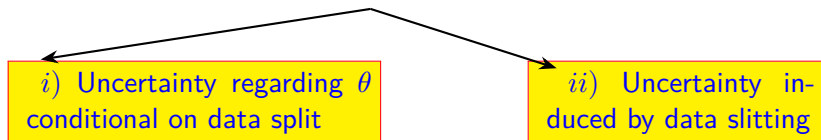
$$\delta_1 = E[g(Y, Z)|G_1] \text{ and } \delta_K = E[g(Y, Z)|G_K]$$

$$CLAN = \delta_K - \delta_1$$

Inference

Let θ be a generic target parameter of interest (e.g.: $\theta = GATES$, $\theta = CLAN...$). Then:

Sources of uncertainty:



i): $A + M$ is a sample and not the entire population \rightarrow uncertainty on $\theta \rightarrow$ we can only have estimations of θ .

ii): Different partitions A and M give different estimations of θ .

Quantifying i)

1. Parameters depend on A , specifically $\theta = \theta_A$;
2. An estimator $\hat{\theta}_A$ is admitted such that:

$$\hat{\theta}_A | \text{Data}_A \sim \mathcal{N}(\theta_A, \hat{\sigma}_A^2).$$

3. As a consequence, the confidence intervals (CI) are:

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

being $\Phi(\cdot)$ the usual cumulative of the normal distribution.

Quantifying i)

1. Parameters depend on A , specifically $\theta = \theta_A$;
2. An estimator $\hat{\theta}_A$ is admitted such that:

$$\hat{\theta}_A | \text{Data}_A \sim \mathcal{N}(\theta_A, \hat{\sigma}_A^2).$$

3. As a consequence, the confidence intervals (CI) are:

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

being $\Phi(\cdot)$ the usual cumulative of the normal distribution.

Problem: $[L_A, U_A]$ are still random (depend on the partition (A, M))!

In order to obtain an estimator and a confidence set which are **non-random** conditional on data, we should turn to quantifying ii)

Quantifying ii), strategy 1

How to account for the uncertainty given by the fact that different partitions give different estimates?

Adjusted Point and Interval Estimators. Our proposal is as follows. As a point estimator, we shall report the median of $\hat{\theta}_A$ as (A, M) vary (as random partitions):

$$\hat{\theta} := \text{Med}[\hat{\theta}_A \mid \text{Data}].$$

This estimator is more robust than the estimator based on a single split. To account for partition uncertainty, we propose to report the following confidence interval (CI) with the nominal confidence level $1 - 2\alpha$:

$$[l, u] := [\overline{\text{Med}}[L_A \mid \text{Data}], \underline{\text{Med}}[U_A \mid \text{Data}]].$$

Note that the price of splitting uncertainty is reflected in the discounting of the confidence level from $1 - \alpha$ to $1 - 2\alpha$. Alternatively, we can report the confidence interval based on inversion of a test based upon p-values, constructed below.

Quantifying ii) continue...

- ▶ The above CI are overall fine (*strategy 1*);
- ▶ However more precise CI can be found using p-values. This is explained in the next slides. Let's call it *strategy 2*.

The latter exercise is useful since by constructing the more precise CI we compute also the p-value and hence can see exactly how hypothesis testing is done!

Quantifying ii), strategy 2

Define:

$$\underline{\text{Med}}(X) := \inf\{x \in \mathbb{R} : P_X(X \leq x) \geq 1/2\}, \quad \overline{\text{Med}}(X) := \sup\{x \in \mathbb{R} : P_X(X \geq x) \geq 1/2\},$$

$$\text{Med}(X) := (\underline{\text{Med}}(X) + \overline{\text{Med}}(X))/2.$$

and also...

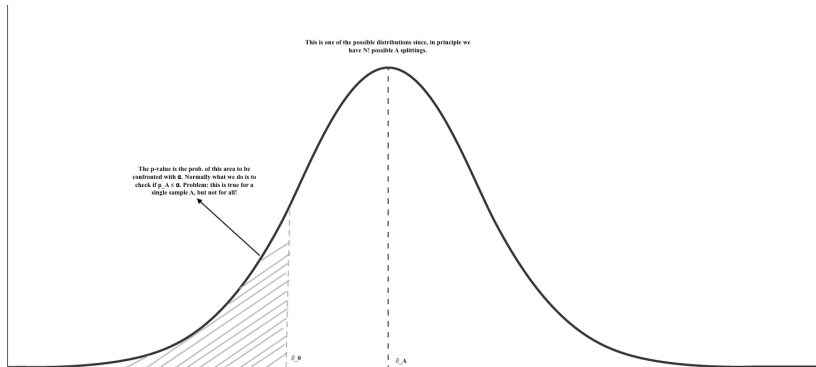
Suppose we are testing $H_0 : \theta_A = \theta_0$ against $H_1 : \theta_A < \theta_0$, conditional on the auxiliary data, then the p-value is given by

$$p_A = \Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_0)).$$

The p-value for testing $H_0 : \theta_A = \theta_0$ against $H_1 : \theta_A > \theta_0$, is given by $p_A = 1 - \Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_0))$.

Quantifying ii), strategy 2

In *strategy 2* we still have to adjust for the fact that different partitions give different estimates, i.e. we still have to define the **adjusted** p-values.



In other words, we cannot state that a *would be normal* test such as $p_A \leq \alpha$ defines a test with significance level α because p_A depends A . Thus we should somehow aggregate the different p-values p_A for each choice of A to reach a unique p :

Adjusted P-values. We say that testing the null hypothesis, based on the p-values p_A , that are random conditional on data, has significance level α if

$$\mathbb{P}(p_A \leq \alpha/2 \mid \text{Data}) \geq 1/2 \quad \text{or} \quad p_{.5} = \text{Med}(p_A \mid \text{Data}) \leq \alpha/2.$$

That is, for at least 50% of the random data splits, the realized p-value p_A falls below the level $\alpha/2$. Hence we can call $p = 2p_{.5}$ the *sample splitting-adjusted p-value*, and consider its small values as providing evidence against the null hypothesis.

Clarify: So if we want to make a test at significance α , we should take as p the median of all the p_A ("most frequent" p_A) computed for different choices of A . Then we check for how many splits A the p_A falls below $\alpha/2$. If this happens for more than 50% of the splits, then we have a valid test.

Clarify: So if we want to make a test at significance α , we should take as p the median of all the p_A ("most frequent" p_A) computed for different choices of A . Then we check for how many splits A the p_A falls below $\alpha/2$. If this happens for more than 50% of the splits, then we have a valid test.

Problem: How can we be sure that the threshold of 50% is a good one for defining the significance level at α ? we will see it in 2 ways...

Quantifying ii), strategy 2

First:

We want that the average number of times in which U_j is lower or equal than $\alpha/2$ is greater than 50% so that we are sure that $U_j \leq \alpha/2$ is a median of all the random variables U_j (i.e. it happens in more than 50% of the uniform random variables). In other words if $M \leq \alpha/2$ then, the majority of the U_j is below $\alpha/2$.

Lemma 4.1 (A Property of Uniform Variables). Consider M , the (usual, lower) median of a sequence $\{U_j\}_{j \in J}$ of uniformly distributed variables, $U_j \sim U(0, 1)$ for each $j \in J$, where variables are not necessarily independent. Then, **(the p_A for us)**

$$\mathbb{P}(M \leq \alpha/2) \leq \alpha.$$

Proof. Let M denote the median of $\{U_j\}_{j \in J}$. Then $M \leq \alpha/2$ is equivalent to $|J|^{-1} \sum_{j \in J} [1(U_j \leq \alpha/2)] - 1/2 \geq 0$. So

$$\mathbb{P}[M \leq \alpha/2] = \mathbb{E}1\left\{|J|^{-1} \sum_{j \in J} [1(U_j \leq \alpha/2)] \geq 1/2\right\}. \quad (4.0)$$

By Markov inequality this is bounded by **Call it: G**

$$2\mathbb{E}|J|^{-1} \sum_{j \in J} [1(U_j \leq \alpha/2)] \leq 2\mathbb{E}[1(U_j \leq \alpha/2)] \leq 2\alpha/2 = \alpha.$$

where the last inequality holds by the marginal uniformity.¹⁰

* Markov Inequality states that: $\mathbb{P}(X \geq \alpha) \leq \mathbb{E}[X]/\alpha$. In our case by 4.0, $\alpha=1/2$, and $X = G$. So by Markov we say that $\mathbb{P}(G \geq 1/2) \leq \mathbb{E}[G]/(1/2) = 2 * \mathbb{E}[G] \leq 2 * \mathbb{E}[1(U_j \leq \alpha/2)] \leq 2 * \alpha/2$ where the penultimate inequality stands because there is no more the division by J . The last is complicated. Intuitively say that $\mathbb{E}[1(U_j \leq \alpha/2)]$ is dichotomous and hence is $= p_1 * 1 + (1-p_1) * 0 = p_1 = \mathbb{P}(U_j \leq \alpha/2) = U_j$ if $0 \leq U_j \leq \alpha/2$. So $\alpha/2$ is a max for the latter probability, meaning that we are sure that $\mathbb{E}[1(U_j \leq \alpha/2)] = p_1 \leq \alpha/2$.

Second which also defines the more precise CI:

Notice that this and Lemma 4.1 show that $p_{.5}$ is a valid test at level α as $p_{.A}$ is a median lower or equal than $\alpha/2$ and we showed that its probability is lower or equal than α and, thanks to th. 4.1 that it converges to α .

Theorem 4.1 (Uniform Validity of Variational P-Value). *Under condition PV and the null hypothesis holding,*

$$\mathbb{P}_P(p_{.5} \leq \alpha/2) \leq \alpha + 2(\delta + \gamma) = \alpha + o(1),$$

uniformly in $P \in \mathcal{P}$.

In order to establish the properties of the confidence interval $[l, u]$, we first consider the properties of the related confidence interval, which is based on the inversion of the p-value based tests:

$$\text{CI} := \{\theta \in \mathbb{R} : p_u(\theta) > \alpha/2, p_l(\theta) > \alpha/2\}, \quad (4.1)$$

for $\alpha < .25$, where, for $\hat{\sigma}_A > 0$,

$$p_l(\theta) := \underline{\text{Med}}(1 - \Phi[\hat{\sigma}_A^{-1}(\hat{\theta}_A - \theta)] \mid \text{Data}), \quad (4.2)$$

$$p_u(\theta) := \underline{\text{Med}}(\Phi[\hat{\sigma}_A^{-1}(\hat{\theta}_A - \theta)] \mid \text{Data}). \quad (4.3)$$

The confidence interval CI has the following representation in terms of the medians of t-statistics implied by the proof Theorem 4.2 stated below:

$$\text{CI} = \left\{ \theta \in \mathbb{R} : \begin{array}{l} \overline{\text{Med}} \left[\frac{\theta - \hat{\theta}_A}{\hat{\sigma}_A} - \Phi^{-1}(1 - \alpha/2) \mid \text{Data} \right] < 0 \\ \underline{\text{Med}} \left[\frac{\theta - \hat{\theta}_A}{\hat{\sigma}_A} + \Phi^{-1}(1 - \alpha/2) \mid \text{Data} \right] > 0 \end{array} \right\}. \quad (4.4)$$

This CI can be (slightly) tighter than $[l, u]$, while the latter is much simpler to construct.

The following theorem establishes that both confidence sets maintain the approximate confidence level $1 - 2\alpha$.

Application

Disclaimer

If you are willing to apply the described methodology, I would recommend going through Chernuzokov sorted effects theoretical paper ([4]) plus to go and check out the following link:
<https://cran.r-project.org/web/packages/SortedEffects/vignettes/SortedEffects.html>.

Further suggestions: if you decide to write your own code starting from the theoretical paper, be as patient as you can and prepare to suffer for a week...the result will amaze you anyway, see for instance our paper (under review) here https://www.dropbox.com/s/dvmlf564cbth3yb/Machine_Learning_Trade_all_2020_sorted_CADiff_8_12_22_%20%281%29.pdf?dl=0. Appendix D will delight you with the difficulties in constructing joint p-values.

A useful algorithm to apply generalized ML (is the one applied for India's immunization effectiveness):

Step 0 Fix number of splits R

Step 1 Compute propensity score $p(Z_i)$

Step 2 Divide each split in half into A and M . For each split $r = 1, \dots, R$:

▶ Learn $B(\cdot)$ and $S(\cdot)$ on A and predict them on M to obtain predicted baseline treatment effect $B(\cdot)$ and predicted treatment effect $S(\cdot)$

▶ Estimate BLP parameters via weighted OLS in M , i.e.:

$$Y_i = \hat{\alpha}' X_{1i} + \hat{\beta}_1 (D_i - p(Z_i)) + \hat{\beta}_2 (D_i - p(Z_i))(S_i - E_{N,M}[S_i]) + \hat{\epsilon}_i, \quad i \in M$$

▶ Estimate GATES parameters via weighted OLS in M , i.e.:

$$Y_i = \hat{\alpha}' X_{1i} + \sum_{k=1}^K \hat{\gamma}_k \cdot (D_i - p(Z_i)) \cdot 1(S_i \in I_k) + \hat{\nu}_i, \quad i \in M$$

▶ Estimate CLAN parameters via weighted OLS in M , i.e.:

$$\hat{\delta}_1 = E_{N,M}[g(Y_i, Z_i) | S_i \in I_1] \text{ and } \hat{\delta}_K = E_{N,M}[g(Y_i, Z_i) | S_i \in I_K]$$

How are they obtained?
(B,S) are minimizer of $\{Y_i - B(Z_i) - S(Z_i)(D_i - p(Z_i))\}^2$

As such they can be solved using debiased-ML.

India application (quick)

Problem: (i) Poor Indian population demand for vaccine; (ii) dis-information lead to delays in vaccine injections in children; (iii) parents lose steam over the course of the immunization process; (iv) less children are vaccinated;

Proposed solution: Introduce nudges and evaluate their effectiveness. Three types of nudges:

- ▶ Small incentives: phone credit upon bringing children to vaccination;
- ▶ Information diffused through key members of the community;
- ▶ Reminders via SMS;
- ▶ Mixes of the above selected with post-LASSO (they selected the relevant interactions among interventions).

Set-up

- ▶ **Treatment group:** 25 villages where such interventions were applied; **Control group:** 78 villages;
- ▶ Y : of children of 15 months or lower in a month in a village that received the measles shot \rightarrow why? Measles shot is the last in the sequence and should be done at the 10th month of living. However it is usually done with delay within the 15th. Hence it marks people
 1. who **demand vaccines** (because they vaccinated children until measles shot);
 2. who **lost trust** in the vaccination process as they probably did the measles (and hence all the preceding vaccines) with delay

Therefore children of 15 months or lower represent the eligible population.

- ▶ D : treatment: 1 if the village received the treatment;
- ▶ Z : village-level characteristics.

Results (BLP)

Classical ATE: the number of immunized children increased of almost 3 more in treated villages

TABLE 3. BLP of Immunization Incentives

This indicates that there is heterogeneity. $\beta_2 = (S-ES)$
-> if no het. $S=ES$ and $\beta_2=0$

Elastic Net		Neural Network	
ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
2.812	0.876	2.530	1.059
(0.867,4.774)	(0.656,1.105)	(0.984,4.079)	(0.724,1.401)
[0.008]	[0.000]	[0.003]	[0.000]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.

P-values for the hypothesis that the parameter is equal to zero in brackets.

Results (GATES)

Notice: GATES is done on the treatment effect $s_0(Z)$!

TABLE 4. GATES of 20% Most and Least Affected Groups

	Elastic Net			Nnet		
	20% Most (G_5)	20% Least (G_1)	Difference	20% Most (G_5)	20% Least (G_1)	Difference
GATE $\gamma_k := \hat{E}[s_0(Z) G_k]$	10.36 (7.42,13.52) [0.00]	-6.12 (-9.83,-2.23) [0.00]	16.34 (11.21,21.62) [0.00]	10.39 (6.22,14.60) [0.00]	-6.20 (-11.43,-0.73) [0.05]	16.80 (9.50,23.85) [0.00]
Control Mean $:= \hat{E}[b_0(Z) G_k]$	2.19 (1.36,2.98) [0.00]	12.24 (11.45,13.10) [0.00]	-9.87 (-11.16,-8.73) [0.00]	1.18 (0.44,1.87) [0.00]	10.32 (9.65,11.02) [0.00]	-9.17 (-10.17,-8.14) [0.00]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.

P-values for the hypothesis that the parameter is equal to zero in brackets.

Results (CLAN)

Notice: while GATES is done on the effects, CLAN is done on covariates!

TABLE 5. CLAN of Immunization Incentives

	Elastic Net			Nnet		
	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)
For instance this suggests that the most affected group by the treatment are those villages having less vaccinated pregnant mothers. Logical: these are in fact those villages with less educated people and for which treatment is more effective!						
Number of vaccines to pregnant mother	2.161 (2.110,2.212)	2.288 (2.237,2.337)	-0.128 (-0.200,-0.055) [0.001]	2.164 (2.107,2.221)	2.328 (2.273,2.385)	-0.160 (-0.245,-0.082) [0.000]
Number of vaccines to child since birth	4.230 (4.100,4.369)	4.714 (4.573,4.860)	-0.513 (-0.710,-0.311) [0.000]	3.995 (3.816,4.165)	4.670 (4.507,4.835)	-0.690 (-0.937,-0.454) [0.000]
Fraction of children received polio drops	1.000 (1.000,1.000)	1.000 (1.000,1.000)	0.000 (0.000,0.000) [0.000]	0.998 (0.996,1.000)	1.000 (0.998,1.002)	-0.002 (-0.005,0.001) [0.485]
Number of polio drops to child	2.964 (2.954,2.975)	2.998 (2.987,3.007)	-0.033 (-0.047,-0.019) [0.000]	2.956 (2.940,2.971)	2.994 (2.980,3.008)	-0.038 (-0.059,-0.016) [0.001]
Fraction of children received immunization card	0.899 (0.878,0.922)	0.932 (0.908,0.956)	-0.036 (-0.065,-0.004) [0.000]	0.804 (0.765,0.842)	0.930 (0.895,0.966)	-0.125 (-0.178,-0.070) [0.006]
Fraction of children received Measles vaccine by 15 months of age	0.127 (0.100,0.155)	0.255 (0.230,0.282)	-0.131 (-0.167,-0.094) [0.052]	0.125 (0.098,0.152)	0.254 (0.229,0.279)	-0.134 (-0.169,-0.098) [0.000]
Fraction of children received Measles at credible locations	0.290 (0.252,0.327)	0.435 (0.400,0.470)	-0.152 (-0.198,-0.097) [0.000]	0.275 (0.236,0.315)	0.426 (0.391,0.461)	-0.151 (-0.203,-0.100) [0.000]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.

Notes: P-values for the hypothesis that the parameter is equal to zero in brackets.

Matrix Completion of World Trade: An Analysis of Interpretability through Shapley Values

Gnecco G., Nutarelli F., Riccaboni M.

Introduction

Disclaimer: this part is just to provide an application of Shapley values discussed by Luigi. We will focus on the application skipping the technical details of the Shapley algorithm. This is part of a paper of mine with Massimo and Giorgio.

Idea: We have constructed a complexity index through Matrix Completion (a ML tool) ranking countries. We would like to know **how much the information of the Relative Comparative Advantage of each country contributed to the final value of the index.** So countries are features for us here.

Notation (simplified)

Define $RCA_{c,p}$ the relative comparative advantage of a country c for a product p (in essence the capability of a country to produce and export a product). Typically, if $RCA_{c,p} \geq 1$ c has a comparative advantage in exporting p .

Define an incidence matrix s.t.

$$M_{c,p} = \begin{cases} 1, & \text{if } RCA_{c,p} \geq 1 \\ 0, & \text{otherwise} \end{cases} .$$

The proposed complexity index is based, among others, on the degree of predictability through MC of the incidence matrix associated with each country.

Shapley

The Shapley value is used to measure the role of each country (part of training and test set in MC) in predicting the elements of $M_{c,p}$ for selected countries in EU in 2018 ².

²Applications for further years in the paper.

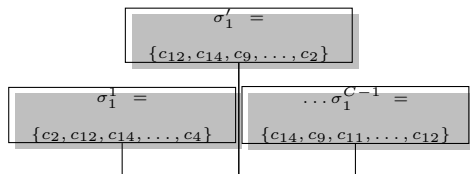
Shapley (reminder)

- ▶ Consider a Transferable Utility (TU) cooperative game, i.e. a pair (N, ν) being $n = |N|$ the number of players (**countries**) and ν a characteristic function (**probability of correct classification** based on that subset of countries S) associating a utility $\nu(S)$ to each subset S of players;
- ▶ The Shapley value divides in a fair way the utility $\nu(N)$ of the coalition $S = N$ among all its players. In other words, it represents a measure of the importance of each player $i \in N$ for a specific TU game:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)] .$$

Algorithm (quick overview)

1. Generate Q' country permutations as $\sigma'_1, \sigma'_2, \dots, \sigma'_Q$
2. For each permutation σ_q generate other $C - 1$ permutations (σ_q^r) so that each country appears in first position exactly once.



The test set will be the last country in such a set. The training will progressively grow and will be the remaining countries.

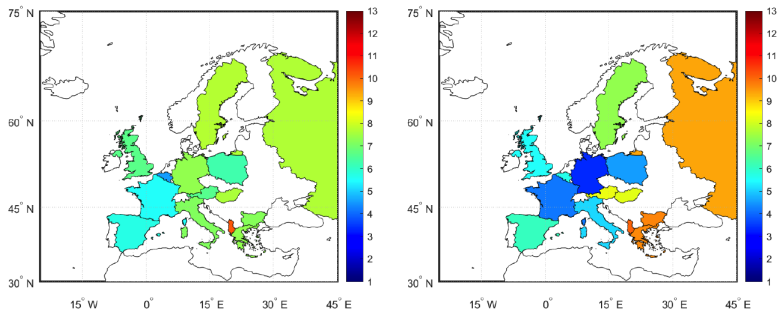
3. For each permutation σ_q^r generate $C - 1$ coalitions $S_{\sigma_q^r}^1, \dots, S_{\sigma_q^r}^{C-1}$ with cardinalities $|S_{\sigma_q^r}^1| = 1, \dots, |S_{\sigma_q^r}^{C-1}| = C - 1$.
4. Solve the MC optimization problem for each subset $S_{\sigma_q^r}^u$ by choosing a training and a test set within $S_{\sigma_q^r}^u$.
5. The resulting optimal matrix obtained with the optimal values of regularization parameters is $Z_{S_{\sigma_q^r}^u}$.
6. For each $Z_{S_{\sigma_q^r}^u}$ construct a binary classifier
7. Approximate ν_j with the accuracy of the MC classifier on the test set
8. Shapley value is obtained by applying the formula above (more or less)

Results (ranking)

Ranking	Country j													
	ALB	AUT	BEL	BGR	DEU	ESP	FRA	GBR	GRC	HUN	ITA	POL	RUS	SWE
1 st	BGR	BEL	ESP	ALB	GBR	ITA	ITA	ESP	HUN	POL	ESP	FRA	SWE	DEU
2 nd	POL	DEU	GRC	BEL	POL	GBR	BEL	BEL	AUT	BEL	POL	BEL	ALB	FRA
3 rd	FRA	GRC	GBR	FRA	ITA	DEU	DEU	FRA	ESP	SWE	BEL	AUT	AUT	GRC
4 th	AUT	GBR	FRA	ESP	HUN	FRA	BGR	BGR	RUS	ESP	HUN	BGR	BEL	ESP
5 th	GRC	BGR	HUN	GRC	SWE	RUS	POL	SWE	BEL	DEU	GBR	GRC	FRA	AUT
6 th	RUS	FRA	POL	POL	ESP	AUT	HUN	POL	ITA	RUS	AUT	RUS	ITA	BEL
7 th	BEL	SWE	RUS	ITA	FRA	HUN	GBR	AUT	FRA	GRC	SWE	GBR	DEU	ITA
8 th	HUN	ESP	SWE	AUT	AUT	BEL	ESP	ITA	DEU	GBR	RUS	SWE	BGR	POL
9 th	ITA	RUS	BGR	SWE	GRC	SWE	GRC	RUS	GBR	BGR	DEU	DEU	ESP	GBR
10 th	SWE	POL	AUT	DEU	RUS	BGR	AUT	DEU	POL	FRA	BGR	ITA	GRC	BGR
11 th	ESP	HUN	DEU	GBR	BGR	GRC	RUS	GRC	SWE	ITA	FRA	ESP	GBR	HUN
12 th	GBR	ITA	ALB	HUN	BEL	POL	SWE	HUN	ALB	AUT	ALB	HUN	HUN	RUS
13 th	DEU	ALB	ITA	RUS	ALB	ALB	ALB	ALB	BGR	ALB	GRC	ALB	POL	ALB

Table 2. For each column j : ranking of countries $i \neq j$ induced by the approximate Shapley values $\hat{\phi}_i(v_j)$ reported in Table 1.

Results (map)



(a) Average rankings induced by the approximate Shapley values.

(b) Average rankings induced by the cosine similarities.

Figure 1. Visual representations of the average rankings (from 1st to 13th) induced by the approximate Shapley values and by the cosine similarities for the selected European countries.

Interpretation (map)

Maps are obtained averaging the Shapley values (Tab.2 above) and another similarity index (cosine similarity).

Both Figures, therefore represent how much each country is similar on average –according to the Shapley value (left panel) and cosine similarity (right panel) – to the other 13 countries considered in the analysis.

As expected, the average rankings obtained in the two cases differ substantially. In particular, for the dataset analyzed, it turns out that **the average ranking obtained by using approximate Shapley values has a narrower distribution with respect to the one obtained by using cosine similarities.**





Please find further technical details in our paper.

We did not apply SHAP but rather constructed our own algorithm in MATLAB according to Mitchell et al. (2022).




Maybe we will check how results change if SHAP is applied instead. Difference? Check Luigi's presentation (basically in SHAP contributions are part of a linear model).

THANK YOU FOR THE ATTENTION

References I

-  Athey, S., Tibshirani, J. and Wager, S., 2019. Generalized random forests. *The Annals of Statistics*, 47(2), pp.1148-1178.
-  Athey, S., Tibshirani, J. and Wager, S., 2016. Solving heterogeneous estimating equations with gradient forests (No. 3475).
-  Chernozhukov, V., Demirer, M., Duflo, E. and Fernandez-Val, I., 2018. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India (No. w24678). National Bureau of Economic Research.
-  Chernozhukov, V., Fernández-Val, I. and Luo, Y., 2018. The sorted effects method: discovering heterogeneous effects beyond their averages. *Econometrica*, 86(6), pp.1911-1938.

References II

-  Gnecco, G., Nutarelli, F. and Riccaboni, M., 2022. A machine learning approach to economic complexity based on matrix completion. *Scientific Reports*, 12(1), pp.1-10.
-  Huang, M.Y. and Yang, S., 2020. Robust inference of conditional average treatment effects using dimension reduction. *arXiv preprint arXiv:2008.13137*.
-  Jacob, D., 2021. CATE meets ML—The Conditional Average Treatment Effect and Machine Learning. *arXiv preprint arXiv:2104.09935*.