Generalizability with ignorance in mind: learning what we do (not) know for archetypes discovery

Emily Breza
Harvard

Arun G. Chandrasekhar
Stanford

Davide Viviano
Harvard

June, 2025

Problem description

- Growing # of experiments across environments and individuals
  E.g. Male/female enterpreneurs in different countries

Problem description

- Growing # of experiments across environments and individuals
  E.g. Male/female enterpreneurs in different countries
- Moved institutions (JPAL) towards meta-analysis/heterogeneity
  - Goal often is to aggregate evidence across all individuals
  - Aggregation relies on model assumptions (e.g. via shrinkage/sparsity)

Problem description

- Growing # of experiments across environments and individuals
  E.g. Male/female enterpreneurs in different countries
- Moved institutions (JPAL) towards meta-analysis/heterogeneity
  - Goal often is to aggregate evidence across all individuals
  - Aggregation relies on model assumptions (e.g. via shrinkage/sparsity)
- But goal of such institutions is also to direct research
  - "reasonable" models may be predictive only for some individuals
  - for some units, effects may just be arbitrary different

  $\Rightarrow$ Pooling information across all units often misleading

Problem description

- Growing # of experiments across environments and individuals
  E.g. Male/female enterpreneurs in different countries
- Moved institutions (JPAL) towards meta-analysis/heterogeneity
  - Goal often is to aggregate evidence across all individuals
  - Aggregation relies on model assumptions (e.g. via shrinkage/sparsity)
- But goal of such institutions is also to direct research
  - "reasonable" models may be predictive only for some individuals
  - for some units, effects may just be arbitrary different
  - $\Rightarrow$ Pooling information across all units often misleading
- Goal: learn from the data **when** and **how** evidence is portable across contexts/individuals and when instead we need more evidence

Empirical illustrations

Unconditional Cash Transfers:

A Bayesian Meta-Analysis of Randomized Evaluations

in Low and Middle Income Countries

Tommaso Crosta, Dean Karlan, Finley Ong,

Julius Rüschenpöhler, and Christopher Udry*

March 28, 2024

Abstract

We use Bayesian meta-analysis methods to estimate the impact of unconditional cash transfers (UCTs) on twelve primary outcomes from 114 studies of 73 UCT programs in middle and low income countries. Cash transfers generate strong and positive average treatment effects on nine of twelve outcomes: total consumption, food consumption, food security, income, assets, labor supply, children height-for-age, schooling, and psychological well-being. We draw six conclusions: First, households consume more of streams and invest more of lump sums, however once stream programs end the impacts mirror those of lump sum, indicating some propensity to save a portion of stream transfers. Second, we find long-run treatment effects remain strong, but the effects of lump sum transfers measured more than 18 months after the transfer are substantially smaller. Third, as returns are linear with respect to grant amount, we do not find evidence of either threshold-based poverty traps or diminishing marginal returns (within the observed range of transfers). Fourth, effects on consumption and income are greater for UCTs targeted to women. Fifth, including light-touch framing related to child welfare or food security generates weakly stronger impacts. Sixth, positive impacts on labor supply and income suggest no evidence of "dependency" theories that cash transfers demotivate income-generating activity on average.

Empirical illustrations

Unconditional Cash Transfer

A Bayesian Meta-Analysis of Randomize

in Low and Middle Income Cou

Tommaso Crosta, Dean Karlan, Finley (

Julius Rüschenpöhler, and Christopher U

March 28, 2024

Abstract

We use Bayesian meta-analysis methods to estimate the impact o
transfers (UCTs) on twelve primary outcomes from 114 studies of 7
middle and low income countries. Cash transfers generate strong a
treatment effects on nine of twelve outcomes: total consumption, food
security, income, assets, labor supply, children height-for-age, schoolin
well-being. We draw six conclusions: First, households consume more o
more of lump sums, however once stream programs end the impacts n
sum, indicating some propensity to save a portion of stream transfers. S
run treatment effects remain strong, but the effects of lump sum trans
than 18 months after the transfer are substantially smaller. Third, as re
respect to grant amount, we do not find evidence of either threshold-
or diminishing marginal returns (within the observed range of transfers
consumption and income are greater for UCTs targeted to women. Fi
touch framing related to child welfare or food security generates weakly stronger impacts.
Sixth, positive impacts on labor supply and income suggest no evidence of "dependency"
theories that cash transfers demotivate income-generating activity on average.

Why Do People Stay Poor?

Get access >

Clare Balboni, Oriana Bandiera, Robin Burgess, Maitreesh Ghatak, Anton Heil

The Quarterly Journal of Economics, Volume 137, Issue 2, May 2022, Pages 785–844,
https://doi.org/10.1093/qje/qjab045
Published: 07 December 2021

Cite Permissions Share ▾

Abstract

There are two broad views as to why people stay poor. One emphasizes differences in fundamentals, such as ability, talent, or motivation. The poverty traps view emphasizes differences in opportunities that stem from access to wealth. To test these views, we exploit a large-scale, randomized asset transfer and an 11-year panel of 6,000 households who begin in extreme poverty. The setting is rural Bangladesh, and the assets are cows. The data support the poverty traps view—we identify a threshold level of initial assets above which households accumulate assets, take on better occupations (from casual labor in agriculture or domestic services to running small livestock businesses), and grow out of poverty. The reverse happens for those below the threshold. Structural estimation of an occupational choice model reveals that almost all beneficiaries are misallocated in the work they do at baseline and that the gains arising from eliminating misallocation would far exceed the program costs. Our findings imply that large transfers, which create better jobs for the poor, are an effective means of getting people out of poverty traps and reducing global poverty.

Empirical illustrations

Unconditional Cash Transfer
A Bayesian Meta-Analysis of Randomize
in Low and Middle Income Cou

Tommaso Crosta, Dean Karlan, Finley (
Julius Rüschenpöhler, and Christopher U

March 28, 2024

Abstract

We use Bayesian meta-analysis methods to estimate the impact of
transfers (UCTs) on twelve primary outcomes from 114 studies of 7:
middle and low income countries. Cash transfers generate strong a
treatment effects on nine of twelve outcomes: total consumption, food
security, income, assets, labor supply, children height-for-age, schooling
well-being. We draw six conclusions: First, households consume more o
more of lump sums, however once stream programs end the impacts n
sum, indicating some propensity to save a portion of stream transfers. S
run treatment effects remain strong, but the effects of lump sum trans
than 18 months after the transfer are substantially smaller. Third, as re
respect to grant amount, we do not find evidence of either threshold-
or diminishing marginal returns (within the observed range of transfers
consumption and income are greater for UCTs targeted to women. Fi
touch framing related to child welfare or food security generates weakly
Sixth, positive impacts on labor supply and income suggest no evidence
theories that cash transfers demotivate income-generating activity on av

Why Do People Stay Poor? Get access ›

Clare Balboni, Oriana Bandiera, Robin Burgess, Maitreesh Ghatak, Anton Heil

The Quarterly Journal of Economics, Volume 137, Issue 2, May 2022, Pages 785–844,

Implementation Matters: Generalizing Treatment Effects in Education

Noam Angrist
University of Oxford, Youth Impact

Rachael Meager
London School of Economics

Targeted instruction is one of the most effective educational interventions in low- and middle-income countries, yet reported impacts vary by an order of magnitude. We study this variation by aggregating evidence from prior randomized trials across five contexts, and use the results to inform a new randomized trial. We find two factors explain most of the heterogeneity in effects across contexts: the degree of implementation (intention-to-treat or treatment-on-the-treated) and program delivery model (teachers or volunteers). Accounting for these implementation factors yields high generalizability, with similar effect sizes across studies. Thus, reporting treatment-on-the-treated effects, a practice which remains limited, can enhance external validity. We also introduce a new Bayesian framework to formally incorporate implementation metrics into evidence aggregation. Results show targeted instruction delivers average learning gains of 0.42 SD when taken up and 0.85 SD when implemented with high fidelity. To investigate how implementation can be improved in future settings, we run a new randomized trial of a targeted instruction program in Botswana. Results demonstrate that implementation can be improved in the context of a scaling program with large causal effects on learning. While research on implementation has been limited to date, our findings and framework reveal its importance for impact evaluation and generalizability.

Empirical illustrations

WHEN LESS IS MORE:
EXPERIMENTAL EVIDENCE ON INFORMATION DELIVERY
DURING INDIA'S DEMONETIZATION

ABHIJIT BANERJEE*, EMILY BREZA§, ARUN G. CHANDRASEKHAR‡, AND BENJAMIN GOLUB†

ABSTRACT. How should information be disseminated to large populations? The options include broadcasting (e.g., via mass media) and informing a small number of "seeds" who then spread the message. While it may seem natural to try to reach the maximum number of people from the beginning, we show, theoretically and experimentally, that when incentives to seek information are endogenous, this is not necessarily true. In a field experiment during the 2016 Indian demonetization, we varied how information about the policy was delivered to villages along three dimensions: how many people were initially informed (i.e., broadcasting versus seeding); whether the identities of the initially informed were made common knowledge; and number of facts delivered (2 versus 24). The quality of information aggregation was measured in three ways: the volume of conversations about demonetization, the level of knowledge about demonetization rules, and the likelihood of making the correct choice in a strongly incentivized decision where understanding the rules is key. Under common knowledge, seeding dominates broadcasting. Moreover, common knowledge makes seeding more effective but broadcast *less so*. These comparisons hold for all three outcomes and underscore the importance of the incentive to engage in social learning. Using data on differential behavior across different ability categories, we interpret our results via a model of image concerns, and also consider several alternative explanations.

ing

mics

and middle-income
tion by aggregating evidence
ew randomized trial. We find
of implementation
hers or volunteers).
lar effect sizes across studies.
, can enhance external
plementation metrics into
gains of 0.42 SD when taken
entation can be improved in
future settings, we run a new randomized trial of a targeted instruction program in Botswana. Results demonstrate that implementation can be improved in the context of a scaling program with large causal effects on learning. While research on implementation has been limited to date, our findings and framework reveal its importance for impact evaluation and generalizability.

theories that cash tra

of 6.4 percentage points over a baseli
of the students using our preferred p

This paper in one slide

- We conduct experiments (pilots) with heterogeneous individuals
  - Types $x \in \mathcal{X}$ + $|\mathcal{X}|$ large (covariates, experiment type, country, ...)
  - Each individual experience an effect $\tau(x)$
  - Observe noisy unbiased estimates $\hat{\tau}(x)$ and $\hat{\eta}(x)^2$ of $\tau(x)$ and $\mathbb{V}(\hat{\tau}(x))$

This paper in one slide

- We conduct experiments (pilots) with heterogeneous individuals
  - Types $x \in \mathcal{X}$ + $|\mathcal{X}|$ large (covariates, experiment type, country, ...)
  - Each individual experience an effect $\tau(x)$
  - Observe noisy unbiased estimates $\hat{\tau}(x)$ and $\hat{\eta}(x)^2$ of $\tau(x)$ and $\mathbb{V}(\hat{\tau}(x))$
- Goal is two fold:
  - Predict $\tau(x)$ when information is portable across groups
  - Admit ignorance and claim for more evidence if generalizability fails

This paper in one slide

- We conduct experiments (pilots) with heterogeneous individuals
  - Types $x \in \mathcal{X}$ + $|\mathcal{X}|$ large (covariates, experiment type, country, ...)
  - Each individual experience an effect $\tau(x)$
  - Observe noisy unbiased estimates $\hat{\tau}(x)$ and $\hat{\eta}(x)^2$ of $\tau(x)$ and $\mathbb{V}(\hat{\tau}(x))$

- Goal is two fold:
  - Predict $\tau(x)$ when information is portable across groups
  - Admit ignorance and claim for more evidence if generalizability fails

  $\Rightarrow$ Learn from the data what we do (not) know to inform future research

This paper in one slide

- We conduct experiments (pilots) with heterogeneous individuals
  - Types $x \in \mathcal{X}$ + $|\mathcal{X}|$ large (covariates, experiment type, country, ...)
  - Each individual experience an effect $\tau(x)$
  - Observe noisy unbiased estimates $\hat{\tau}(x)$ and $\hat{\eta}(x)^2$ of $\tau(x)$ and $\mathbb{V}(\hat{\tau}(x))$
- Goal is two fold:
  - Predict $\tau(x)$ when information is portable across groups
  - Admit ignorance and claim for more evidence if generalizability fails

  $\Rightarrow$ Learn from the data what we do (not) know to inform future research
- Steps of the analysis
  - Introduce decision problem for generalizability
  - Construct (robust) predictions for units where generalizability occurs
  - Theoretical guarantees and implications for anti-poverty programs

Related literature

- Meta-analysis and transfer learning [Borenstein et al., 2021; Meager, 2022; Crosta et al., 2024; Menzel, 2023; Ishiara and Kitagawa, 2023; Deeb and de Chaisemartin, 2019; Adjaho and Chistensen, 2022; Andrews et al., 2022; ...]
- Policy learning and effect heterogeneity [Athey and Wager, 2019; Kitagawa and Tetenov, 2018; Manski, 2004; Murphy, 2003; Athey and Wager, 2021; Kennedy, 2023; Chernozhukov et al., 2018; Bonhomme and Manresa, 2015; Viviano and Bradic, 2024; ...]

Related literature

- Meta-analysis and transfer learning [Borenstein et al., 2021; Meager, 2022; Crosta et al., 2024; Menzel, 2023; Ishiara and Kitagawa, 2023; Deeb and de Chaisemartin, 2019; Adjaho and Chistensen, 2022; Andrews et al., 2022; ...]
- Policy learning and effect heterogeneity [Athey and Wager, 2019; Kitagawa and Tetenov, 2018; Manski, 2004; Murphy, 2003; Athey and Wager, 2021; Kennedy, 2023; Chernozhukov et al., 2018; Bonhomme and Manresa, 2015; Viviano and Bradic, 2024; ...]
  - $\Rightarrow$ Literature forces predictions across all units (no ignorance component)

Related literature

- Meta-analysis and transfer learning [Borenstein et al., 2021; Meager, 2022; Crosta et al., 2024; Menzel, 2023; Ishiara and Kitagawa, 2023; Deeb and de Chaisemartin, 2019; Adjaho and Chistensen, 2022; Andrews et al., 2022; ...]
- Policy learning and effect heterogeneity [Athey and Wager, 2019; Kitagawa and Tetenov, 2018; Manski, 2004; Murphy, 2003; Athey and Wager, 2021; Kennedy, 2023; Chernozhukov et al., 2018; Bonhomme and Manresa, 2015; Viviano and Bradic, 2024; ...]
  - ⇒ Literature forces predictions across all units (no ignorance component)
- Site selection/sampling [e.g., Olea et al., 2024; Gechter et al., 2024;...]
  - ⇒ Uses a model to choose a site among set of sites for experiment
  - ⇒ Here evidence aggregation: pilot to study where to rely on current data for predictions and where run an experiment

Related literature

- Meta-analysis and transfer learning [Borenstein et al., 2021; Meager, 2022; Crosta et al., 2024; Menzel, 2023; Ishiara and Kitagawa, 2023; Deeb and de Chaisemartin, 2019; Adjaho and Chistensen, 2022; Andrews et al., 2022; ...]
- Policy learning and effect heterogeneity [Athey and Wager, 2019; Kitagawa and Tetenov, 2018; Manski, 2004; Murphy, 2003; Athey and Wager, 2021; Kennedy, 2023; Chernozhukov et al., 2018; Bonhomme and Manresa, 2015; Viviano and Bradic, 2024; ...]
  - $\Rightarrow$ Literature forces predictions across all units (no ignorance component)
- Site selection/sampling [e.g., Olea et al., 2024; Gechter et al., 2024;...]
  - $\Rightarrow$ Uses a model to choose a site among set of sites for experiment
  - $\Rightarrow$ Here evidence aggregation: pilot to study where to rely on current data for predictions and where run an experiment
- Rejection options in ML [Chow, 1970; Cortes et al., 2016; Franc et al., 2023; ...] , Robust statistics [e.g., Huber and Ronchetti, 2011; Broderick et al., 2020]
  - $\Rightarrow$ Reduce observations influence/provide robustness metric
  - $\Rightarrow$ Here model misspecification + future experimentation (for CATEs)

Content

A framework to learn what is generalizable

Make a prediction about $\tau(x)$ using (existing) data

Admit ignorance about $\tau(x)$ and elicit more evidence

A framework to learn what is generalizable

$\boxed{\text{Make a prediction about } \tau(x) \text{ using (existing) data}}$

$\boxed{\text{Admit ignorance about } \tau(x) \text{ and elicit more evidence}}$

- Making a prediction $=$ predict $\tau(x)$ with $\phi(x), \phi \in \mathcal{F}$
  - $\mathcal{F}$ characterizes prior/communication constraints/data feasibility
  - But not all $\tau(x)$ may be well approximated by $\phi(x)$

A framework to learn what is generalizable

$\boxed{\text{Make a prediction about } \tau(x) \text{ using (existing) data}}$

$\boxed{\text{Admit ignorance about } \tau(x) \text{ and elicit more evidence}}$

- Making a prediction = predict $\tau(x)$ with $\phi(x), \phi \in \mathcal{F}$
  - $\mathcal{F}$ characterizes prior/communication constraints/data feasibility
  - But not all $\tau(x)$ may be well approximated by $\phi(x)$
- Admit ignorance = policy function in a set Π

$$\pi(x) = \begin{cases} 1 & \text{if make a prediction about } \tau(x) \\ 0 & \text{if admit ignorance at cost } \sigma^2 \end{cases}, \quad \pi \in \Pi$$

A framework to learn what is generalizable

$\boxed{\text{Make a prediction about } \tau(x) \text{ using (existing) data}}$

$\boxed{\text{Admit ignorance about } \tau(x) \text{ and elicit more evidence}}$

- Making a prediction = predict $\tau(x)$ with $\phi(x), \phi \in \mathcal{F}$
  - $\mathcal{F}$ characterizes prior/communication constraints/data feasibility
  - But not all $\tau(x)$ may be well approximated by $\phi(x)$
- Admit ignorance = policy function in a set Π

$$\pi(x) = \begin{cases} 1 & \text{if make a prediction about } \tau(x) \\ 0 & \text{if admit ignorance at cost } \sigma^2 \end{cases}, \quad \pi \in \Pi$$

  - Π has bounded complexity (constraints/feasibility)

A framework to learn what is generalizable

$\boxed{\text{Make a prediction about } \tau(x) \text{ using (existing) data}}$

$\boxed{\text{Admit ignorance about } \tau(x) \text{ and elicit more evidence}}$

- Making a prediction = predict $\tau(x)$ with $\phi(x), \phi \in \mathcal{F}$
  - $\mathcal{F}$ characterizes prior/communication constraints/data feasibility
  - But not all $\tau(x)$ may be well approximated by $\phi(x)$
- Admit ignorance = policy function in a set Π

$$\pi(x) = \begin{cases} 1 & \text{if make a prediction about } \tau(x) \\ 0 & \text{if admit ignorance at cost } \sigma^2 \end{cases}, \quad \pi \in \Pi$$

  - Π has bounded complexity (constraints/feasibility)

$\Rightarrow$ "Detect for which units cannot predict well $\tau(x)$"

Generalizability-aware predictions: population version

- For now, ignore sampling uncertainty
- For $\sigma^2 = \infty$ the population objective reads as

$$\min_{\phi \in \mathcal{F}} \sum_x \underbrace{\Big(\tau(x) - \phi(x)\Big)^2}_{\text{loss from prediction}}$$

Generalizability-aware predictions: population version

- For now, ignore sampling uncertainty
- For $\sigma^2 \geq 0$ the population objective defined as $\boldsymbol{L_\sigma(\phi, \pi)}$ is

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x \underbrace{\Big(\tau(x) - \phi(x)\Big)^2 \pi(x)}_{\text{loss from prediction}}$$

Generalizability-aware predictions: population version

- For now, ignore sampling uncertainty
- For $\sigma^2 \geq 0$ the population objective defined as $\boldsymbol{L_\sigma(\phi, \pi)}$ is

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x \underbrace{\Big(\tau(x) - \phi(x)\Big)^2 \pi(x)}_{\text{loss from prediction}} + \underbrace{\sigma^2(1 - \pi(x))}_{\text{cost of ignorance}}$$

Generalizability-aware predictions: population version

- For now, ignore sampling uncertainty
- For $\sigma^2 \geq 0$ the population objective defined as $\boldsymbol{L_\sigma(\phi, \pi)}$ is

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x \underbrace{\Big(\tau(x) - \phi(x)\Big)^2 \pi(x)}_{\text{loss from prediction}} + \underbrace{\sigma^2(1 - \pi(x))}_{\text{cost of ignorance}}$$

Equivalent formulations

- Minimize error with minimum number of units not in ignorance
- Maximizing number of units not in ignorance with constraints on error

Generalizability-aware predictions: population version

- For now, ignore sampling uncertainty
- For $\sigma^2 \geq 0$ the population objective defined as $\boldsymbol{L_\sigma(\phi, \pi)}$ is

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x \underbrace{\Big(\tau(x) - \phi(x)\Big)^2 \pi(x)}_{\text{loss from prediction}} + \underbrace{\sigma^2(1 - \pi(x))}_{\text{cost of ignorance}}$$

Equivalent formulations

- Minimize error with minimum number of units not in ignorance
- Maximizing number of units not in ignorance with constraints on error

$\Rightarrow$ Robustify predictions by making predictions only over subset

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x \Big(\tau(x) - \phi(x)\Big)^2 \pi(x),$$

Generalizability-aware predictions: population version

- For now, ignore sampling uncertainty
- For $\sigma^2 \geq 0$ the population objective defined as $\boldsymbol{L_\sigma(\phi, \pi)}$ is

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x \underbrace{\Big(\tau(x) - \phi(x)\Big)^2 \pi(x)}_{\text{loss from prediction}} + \underbrace{\sigma^2(1 - \pi(x))}_{\text{cost of ignorance}}$$

Equivalent formulations

- Minimize error with minimum number of units not in ignorance
- Maximizing number of units not in ignorance with constraints on error

$\Rightarrow$ Robustify predictions by making predictions only over subset

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x \Big(\tau(x) - \phi(x)\Big)^2 \pi(x), \quad s.t. \sum_x \pi(x) \geq \lambda$$

Generalizability-aware predictions: population version

- For now, ignore sampling uncertainty
- For $\sigma^2 \geq 0$ the population objective defined as $\boldsymbol{L_\sigma(\phi, \pi)}$ is

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x \underbrace{\Big(\tau(x) - \phi(x)\Big)^2 \pi(x)}_{\text{loss from prediction}} + \underbrace{\sigma^2 (1 - \pi(x))}_{\text{cost of ignorance}}$$

Equivalent formulations

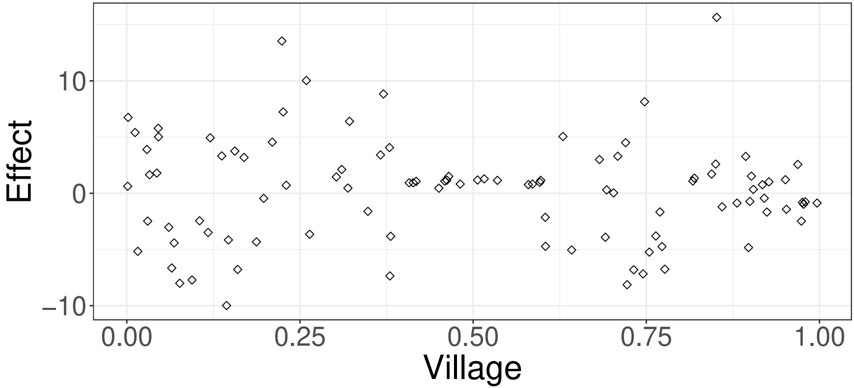
- Minimize error with minimum number of units not in ignorance
- Maximizing number of units not in ignorance with constraints on error

$\Rightarrow$ Robustify predictions by making predictions only over subset

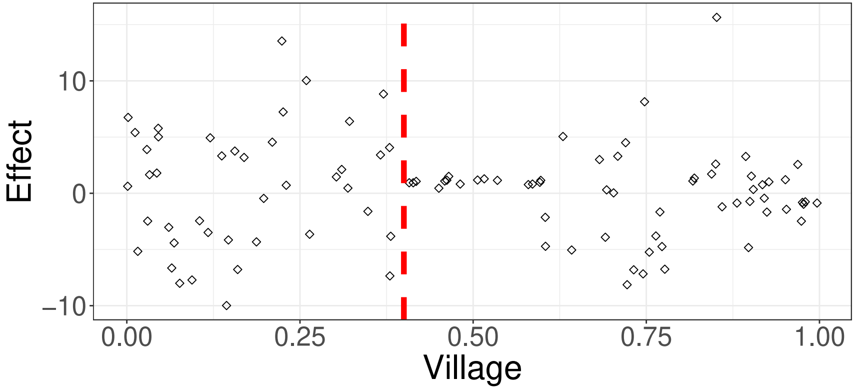
$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x \Big(\tau(x) - \phi(x)\Big)^2 \pi(x), \quad s.t. \sum_x \pi(x) \geq \lambda$$

$\Rightarrow$ Existing estimators for effect heterogeneity always pick $\sigma^2 = \infty$

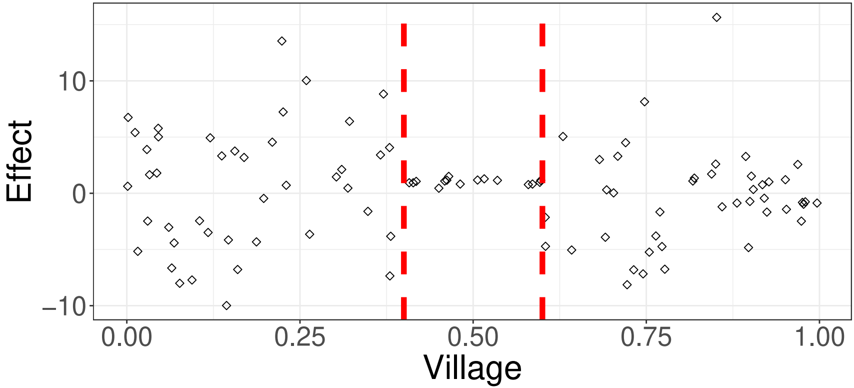
Example: small σ^2



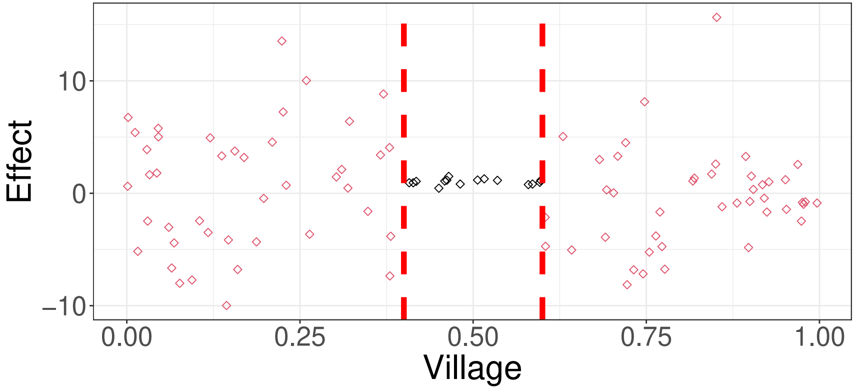
Example: small σ^2



Example: small σ^2



Example: small σ^2



Example: increase σ^2

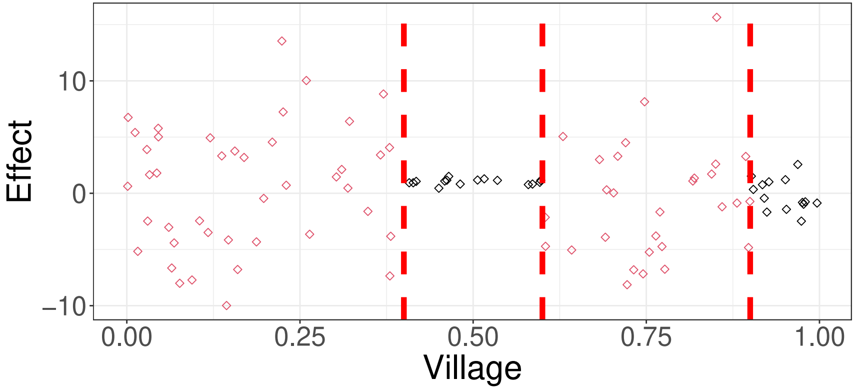


Illustration [Calibrated to Banerjee et al., 2015]

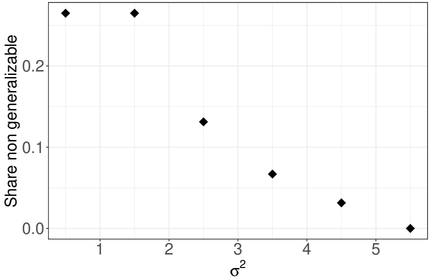
Trade-off between

- how general your model/predictions are
- vs. prediction accuracy

Illustration [Calibrated to Banerjee et al., 2015]

Trade-off between

- how general your model/predictions are
- vs. prediction accuracy



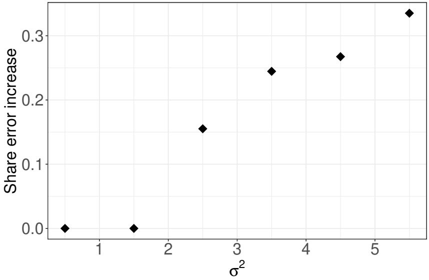
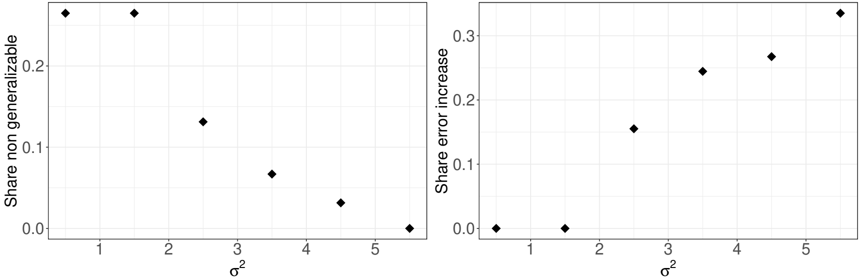


Illustration [Calibrated to Banerjee et al., 2015]

Trade-off between

- how general your model/predictions are
- vs. prediction accuracy



Interpretation of σ^2

- Cost from collecting more data about x in a follow up experiment
- Tolerable inaccuracy of the model given data

Content

Motivation: from models to experiments

- Researchers observe/can report $\phi \in \mathcal{F}$ ($\mathcal{F}$ "simple" $\Rightarrow$ negligible estimation error)
- Researchers are also given the option to sample (e.g., from new study)

$$\tau^{new}(x)|\tau(x) \sim \mathcal{N}(\tau(x), \sigma^2)$$

Motivation: from models to experiments

- Researchers observe/can report $\phi \in \mathcal{F}$ ($\mathcal{F}$ "simple" $\Rightarrow$ negligible estimation error)
- Researchers are also given the option to sample (e.g., from new study)

$$\tau^{new}(x)|\tau(x) \sim \mathcal{N}(\tau(x), \sigma^2)$$

- Audience forms posterior $\mathbb{E}_\eta[\tau(x)|\phi, \tau^{new}]$ under prior $\tau|\mathcal{F} \sim \rho_\eta$. Define risk

$$R_\eta(\phi, \tau) = \mathbb{E}\left[\left(\tau(x) - \mathbb{E}_\eta[\tau(x)|\phi, \tau^{new}]\right)^2 \middle| \tau, \mathcal{F}\right]$$

Motivation: from models to experiments

- Researchers observe/can report $\phi \in \mathcal{F}$ ($\mathcal{F}$ "simple" $\Rightarrow$ negligible estimation error)
- Researchers are also given the option to sample (e.g., from new study)

$$\tau^{new}(x)|\tau(x) \sim \mathcal{N}(\tau(x), \sigma^2)$$

- Audience forms posterior $\mathbb{E}_\eta[\tau(x)|\phi, \tau^{new}]$ under prior $\tau|\mathcal{F} \sim \rho_\eta$. Define risk

$$R_\eta(\phi, \tau) = \mathbb{E}\left[\left(\tau(x) - \mathbb{E}_\eta[\tau(x)|\phi, \tau^{new}]\right)^2 \middle| \tau, \mathcal{F}\right]$$

$\phi \in \mathcal{F}$	$\tau^{new}\|\tau \sim \mathcal{N}(\tau, \sigma^2)$	$E_\eta[\tau\|\mathcal{F}, \tau^{new}]$	$R_\eta(\phi, \tau)$
Model prediction from pilot	New data	Audience	Risk

Motivation: from models to experiments

- Misspecification in prior ρ_η: For *some* $\phi \in \mathcal{F}, \pi \in \Pi$:

$$\tau(x)|\mathcal{F} \begin{cases} = \phi(x) & \text{if } \pi(x) = 1 \\ \sim \mathcal{N}(b(x), \eta^2) & \text{otherwise} \end{cases}$$

b arbitrary $+ \ \eta^2$ "radius" of heterogeneity

Motivation: from models to experiments

- Misspecification in prior ρ_η: For *some* $\phi \in \mathcal{F}, \pi \in \Pi$:

$$\tau(x)|\mathcal{F} \begin{cases} = \phi(x) & \text{if } \pi(x) = 1 \\ \sim \mathcal{N}(b(x), \eta^2) & \text{otherwise} \end{cases}$$

  b arbitrary + η^2 "radius" of heterogeneity

- Prior: ϕ is only correct **locally**

Motivation: from models to experiments

- Misspecification in prior ρ_η: For *some* $\phi \in \mathcal{F}, \pi \in \Pi$:

$$\tau(x)|\mathcal{F} \begin{cases} = \phi(x) & \text{if } \pi(x) = 1 \\ \sim \mathcal{N}(b(x), \eta^2) & \text{otherwise} \end{cases}$$

  b arbitrary $+\ \eta^2$ "radius" of heterogeneity

- Prior: ϕ is only correct **locally**

Thm (Informal) For some ϕ, π,

$$L_\sigma(\phi, \pi) = \lim_{\eta \to \infty} \mathbb{E}\left[\left(\tau(x) - \mathbb{E}_\eta[\tau(x)|\mathcal{F}, \tau^{new}]\right)^2 \Big| \tau, \mathcal{F}\right]$$

$\Rightarrow$ We balance local misspecification with exploration

Content

Estimation with existing (pilot) study

- How to construct optimal $\phi(x)$ from previous studies?
- Observe noisy unbiased estimates $\hat{\tau}(x)$ and $\hat{\eta}(x)^2$ of $\tau(x)$ and $\mathbb{V}(\hat{\tau}(x))$

Estimation with existing (pilot) study

- How to construct optimal $\phi(x)$ from previous studies?
- Observe noisy unbiased estimates $\hat{\tau}(x)$ and $\hat{\eta}(x)^2$ of $\tau(x)$ and $\mathbb{V}(\hat{\tau}(x))$
- Estimate $\hat{\pi}, \hat{\phi}$ by minimizing empirical loss $\hat{L}_\sigma(\cdot)$ defined as

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x \Big\{ \underbrace{\Big(\hat{\tau}(x) - \phi(x)\Big)^2}_{\text{est prediction err}} - \underbrace{\hat{\eta}(x)^2}_{\text{est variance}} \Big\} \pi(x) - \underbrace{\sigma^2 \pi(x)}_{\text{cost of ignorance}}$$

Estimation with existing (pilot) study

- How to construct optimal $\phi(x)$ from previous studies?
- Observe noisy unbiased estimates $\hat{\tau}(x)$ and $\hat{\eta}(x)^2$ of $\tau(x)$ and $\mathbb{V}(\hat{\tau}(x))$
- Estimate $\hat{\pi}, \hat{\phi}$ by minimizing empirical loss $\hat{L}_\sigma(\cdot)$ defined as

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x \Big\{ \underbrace{\Big(\hat{\tau}(x) - \phi(x)\Big)^2}_{\text{est prediction err}} - \underbrace{\hat{\eta}(x)^2}_{\text{est variance}} \Big\} \pi(x) - \underbrace{\sigma^2 \pi(x)}_{\text{cost of ignorance}}$$

- Basic intuition
  - here $|\mathcal{X}|$ is large/ characterizes effective sample size
  - we need to pool some x to reduce noise, $\mathcal{F}$ posit how to pool obs/

Estimation with existing (pilot) study

- How to construct optimal $\phi(x)$ from previous studies?
- Observe noisy unbiased estimates $\hat{\tau}(x)$ and $\hat{\eta}(x)^2$ of $\tau(x)$ and $\mathbb{V}(\hat{\tau}(x))$
- Estimate $\hat{\pi}, \hat{\phi}$ by minimizing empirical loss $\hat{L}_\sigma(\cdot)$ defined as

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x \Big\{ \underbrace{\Big(\hat{\tau}(x) - \phi(x)\Big)^2}_{\text{est prediction err}} - \underbrace{\hat{\eta}(x)^2}_{\text{est variance}} \Big\} \pi(x) - \underbrace{\sigma^2 \pi(x)}_{\text{cost of ignorance}}$$

- Basic intuition
  - here $|\mathcal{X}|$ is large/ characterizes effective sample size
  - we need to pool some x to reduce noise, $\mathcal{F}$ posit how to pool obs/

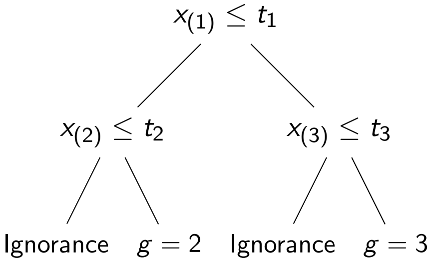
  - To decide when to pool, compare between to within variation
  - If between variation much larger than within, do not pool

Example with simple regression tree

- Each "unit" is a (small) group of obs/ with same x
- At each leaf node, either predict (with sample mean) or abstain

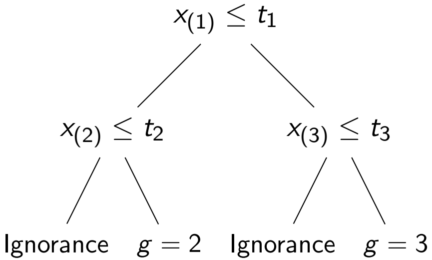
Example with simple regression tree

- Each "unit" is a (small) group of obs/ with same x
- At each leaf node, either predict (with sample mean) or abstain



Example with simple regression tree

- Each "unit" is a (small) group of obs/ with same x
- At each leaf node, either predict (with sample mean) or abstain

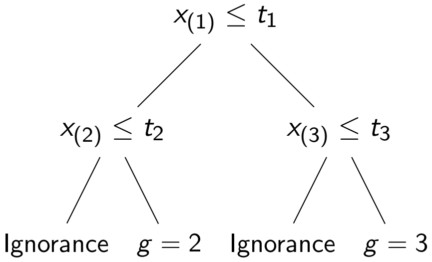


- For each leaf node (and given splits $(x \leq t)$), assign to ignorance if:

  Between variation of $\hat{\tau}(x)$ exceeds by σ^2 sum of within variations

Example with simple regression tree

- Each "unit" is a (small) group of obs/ with same x
- At each leaf node, either predict (with sample mean) or abstain



- For each leaf node (and given splits $(x \leq t)$), assign to ignorance if:

  Between variation of $\hat{\tau}(x)$ exceeds by σ^2 sum of within variations

- Repeat and search for combinations of splits that minimize loss

More general class of predictions and policies

- $\alpha \in \mathcal{G}$ groups individuals into G groups with complexity $\mathrm{VC}(\mathcal{G})$
- Predictions are the same in each group (call them $\phi \in \mathcal{F}_\alpha$)

$$\mathcal{F}_\alpha = \left\{ \phi : \phi(x) = \phi(x') \text{ if } \alpha(x) = \alpha(x') \right\}$$

More general class of predictions and policies

- $\alpha \in \mathcal{G}$ groups individuals into G groups with complexity $\mathrm{VC}(\mathcal{G})$
- Predictions are the same in each group (call them $\phi \in \mathcal{F}_\alpha$)

$$\mathcal{F}_\alpha = \Big\{\phi : \phi(x) = \phi(x') \text{ if } \alpha(x) = \alpha(x')\Big\}$$

- Each group $g > 1$ has either zero individuals, or a few of them $(\underline{\kappa}|\mathcal{X}|)$:

$$\pi_\alpha(x) = \begin{cases} 1 & \text{if } \alpha(x) > 1 \quad \text{(generalizable)} \\ 0 & \text{if } \alpha(x) = 1 \quad \text{(ignorance)} \end{cases}.$$

More general class of predictions and policies

- $\alpha \in \mathcal{G}$ groups individuals into G groups with complexity $\mathrm{VC}(\mathcal{G})$
- Predictions are the same in each group (call them $\phi \in \mathcal{F}_\alpha$)

$$\mathcal{F}_\alpha = \Big\{\phi : \phi(x) = \phi(x') \text{ if } \alpha(x) = \alpha(x')\Big\}$$

- Each group $g > 1$ has either zero individuals, or a few of them $(\underline{\kappa}|\mathcal{X}|)$:

$$\pi_\alpha(x) = \begin{cases} 1 & \text{if } \alpha(x) > 1 \quad \text{(generalizable)} \\ 0 & \text{if } \alpha(x) = 1 \quad \text{(ignorance)} \end{cases}.$$

E.g. Reg trees and group fixed effects with bounded complexity

Main guarantee: regret

Thm Let $\hat{\tau}(x), \hat{\eta}(x)^2$ have bounded third moment. Study the regret

$$\mathbb{E}\left[L_\sigma(\hat{\phi}, \hat{\pi}) - \min_{\alpha \in \mathcal{G}, \phi \in \mathcal{F}_\alpha} L_\sigma(\phi, \pi_\alpha)\right]$$

Main guarantee: regret

Thm Let $\hat{\tau}(x), \hat{\eta}(x)^2$ have bounded third moment. Study the regret

$$\mathbb{E}\left[L_\sigma(\hat{\phi}, \hat{\pi}) - \min_{\alpha \in \mathcal{G}, \phi \in \mathcal{F}_\alpha} L_\sigma(\phi, \pi_\alpha)\right] \leq c_0 G \sqrt{\frac{\mathrm{VC}(\mathcal{G})}{|\mathcal{X}|}}$$

$\Rightarrow$ proof combines chaining argument with group-fixed effects

$\Rightarrow$ bounds only depend on $G\sqrt{\mathrm{VC}(\mathcal{G})}$ and distribution free

Main guarantee: regret

Thm Let $\hat{\tau}(x), \hat{\eta}(x)^2$ have bounded third moment. Study the regret

$$\mathbb{E}\left[L_\sigma(\hat{\phi}, \hat{\pi}) - \min_{\alpha \in \mathcal{G}, \phi \in \mathcal{F}_\alpha} L_\sigma(\phi, \pi_\alpha)\right] \leq c_0 G \sqrt{\frac{\mathrm{VC}(\mathcal{G})}{|\mathcal{X}|}}$$

$\Rightarrow$ proof combines chaining argument with group-fixed effects
$\Rightarrow$ bounds only depend on $G\sqrt{\mathrm{VC}(\mathcal{G})}$ and distribution free

Additional results in the paper:

- minimax rate ($n^{-1/2}$) as function of total n
- guarantees with weights for observations x
- computational algorithms for regression trees (more)

Main guarantee: regret

Thm Let $\hat{\tau}(x), \hat{\eta}(x)^2$ have bounded third moment. Study the regret

$$\mathbb{E}\left[L_\sigma(\hat{\phi}, \hat{\pi}) - \min_{\alpha \in \mathcal{G}, \phi \in \mathcal{F}_\alpha} L_\sigma(\phi, \pi_\alpha)\right] \leq c_0 G \sqrt{\frac{\mathrm{VC}(\mathcal{G})}{|\mathcal{X}|}}$$

$\Rightarrow$ proof combines chaining argument with group-fixed effects
$\Rightarrow$ bounds only depend on $G\sqrt{\mathrm{VC}(\mathcal{G})}$ and distribution free

Additional results in the paper:

- minimax rate ($n^{-1/2}$) as function of total n
- guarantees with weights for observations x
- computational algorithms for regression trees (more)
- asymptotic inference on the set of optimal partitions: (inference guarantees)

$$H_0 : \mathcal{G}' \in \mathcal{G}^\star \quad \mathcal{G}^\star := \left\{\alpha \in \mathcal{G} : \min_{\alpha' \in \mathcal{G}, \phi \in \mathcal{F}_{\alpha'}} L_\sigma(\phi, \pi_{\alpha'}) = \min_{\phi \in \mathcal{F}_\alpha} L_\sigma(\phi, \pi_\alpha)\right\}$$

Content

Empirical illustration

- Heterogeneity in anti-poverty programs often depends on baseline poverty level. Can we find predictable heterogeneity?
  - $\Rightarrow$ Study multifacet program in six countries [Banerjee et al., 2015]
  - $\Rightarrow$ Implement Generalized Aware trees with several covariates + country.

Empirical illustration

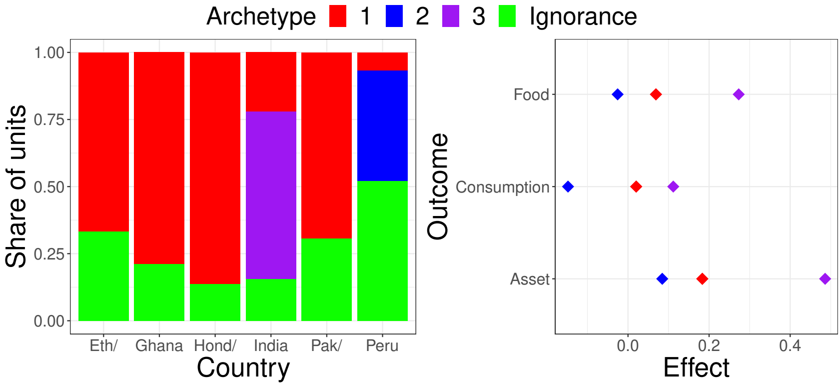
- Heterogeneity in anti-poverty programs often depends on baseline poverty level. Can we find predictable heterogeneity?
  - $\Rightarrow$ Study multifacet program in six countries [Banerjee et al., 2015]
  - $\Rightarrow$ Implement Generalized Aware trees with several covariates + country.

- Setup
  - We consider depth three tree, with $G \leq 4$
  - Obtain $\hat{\tau}(x)$ through IPW and $\hat{\eta}(x)$ with lasso (possible also to use non-parametric estimators)
  - Consider three outcomes with same group structure α
  - Look at σ^2 so that $\leq 15\%$ are non-generalizable

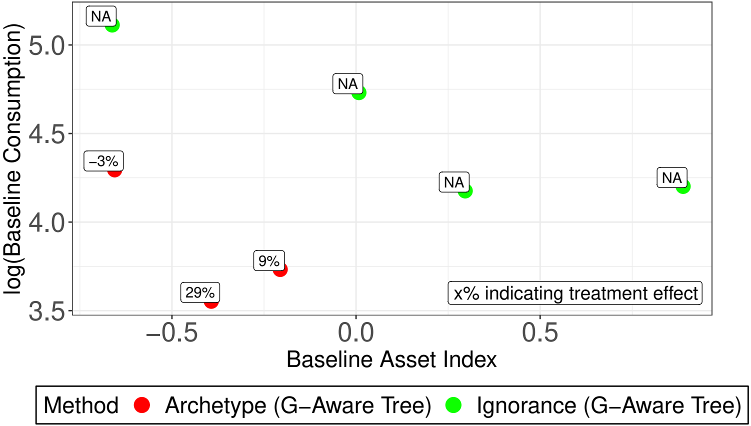
Empirical illustration

- Heterogeneity in anti-poverty programs often depends on baseline poverty level. Can we find predictable heterogeneity?
  - $\Rightarrow$ Study multifacet program in six countries [Banerjee et al., 2015]
  - $\Rightarrow$ Implement Generalized Aware trees with several covariates + country.
- Setup
  - We consider depth three tree, with $G \leq 4$
  - Obtain $\hat{\tau}(x)$ through IPW and $\hat{\eta}(x)$ with lasso (possible also to use non-parametric estimators)
  - Consider three outcomes with same group structure α
  - Look at σ^2 so that $\leq 15\%$ are non-generalizable
- Findings
  - Large effects for ultra-poor individuals
  - Effects are arbitrary heterogeneous for richer individuals (within poor)
  - Comparable existing regressions report unstable estimates
  - Policy interventions should consider gather more evidence on richer

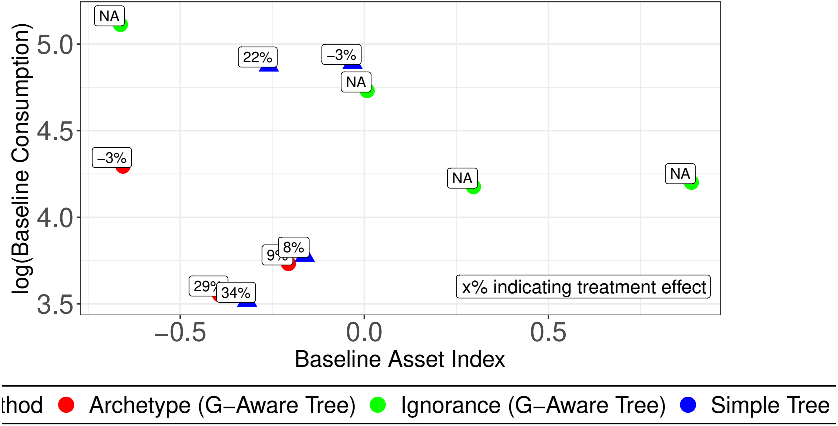
Compositions of archetypes by country



Predicted treatment effects with two-depth tre3



Predicted treatment effects with two-depth tree

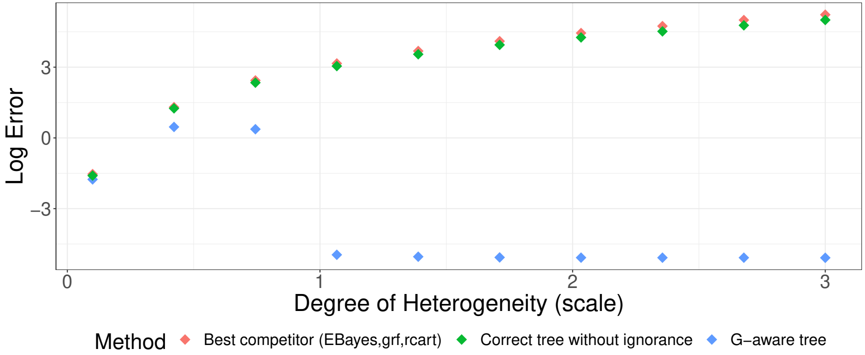


Calibrated simulations

- Calibrate simulations to estimated DGP with estimated tree
- For 4% of observations in ignorance, generate treatment effects from Cauchy
- Compute error conditional on treatment effects

Calibrated simulations

- Calibrate simulations to estimated DGP with estimated tree
- For 4% of observations in ignorance, generate treatment effects from Cauchy
- Compute error conditional on treatment effects



Conclusions

- Goal should be produce predictions but also direct research

$\Rightarrow$ Learn from the data what we know and what we do not know

Conclusions

- Goal should be produce predictions but also direct research

$\Rightarrow$ Learn from the data what we know and what we do not know

- Propose abstaining from predictions at cost of experimentation
- We propose estimators to optimize over predictions and ignorance
- We study theoretical properties and provide an application

Conclusions

- Goal should be produce predictions but also direct research

$\Rightarrow$ Learn from the data what we know and what we do not know

- Propose abstaining from predictions at cost of experimentation
- We propose estimators to optimize over predictions and ignorance
- We study theoretical properties and provide an application

What is next?

- Implications for experimental design
- Application to ensamble methods
- Large scale empirical implementation

Thanks very much, questions?

Computational complexity

$\mathcal{G}$ is class of trees with G leaf nodes, p covariates and n observaions:

Thm Computational complexity is $\mathcal{O}(n^G p^G)$.

Computational complexity

$\mathcal{G}$ is class of trees with G leaf nodes, p covariates and n observaions:

Thm Computational complexity is $\mathcal{O}(n^G p^G)$.

Intuition

- Let $G = 2$
  1. Run over all covariates and splits (at most pn)
  2. Ignorance decision ($\pi(x) = 0$) is independent at each leaf node
  3. Loss function equal some of losses betw/ leaf nodes

Computational complexity

$\mathcal{G}$ is class of trees with G leaf nodes, p covariates and n observaions:

Thm Computational complexity is $\mathcal{O}(n^G p^G)$.

Intuition

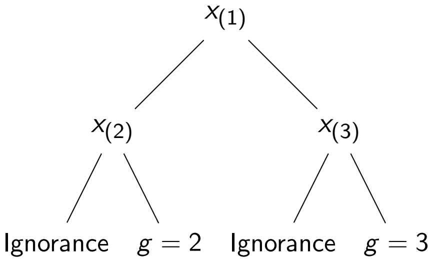
- Let $G = 2$
  1. Run over all covariates and splits (at most pn)
  2. Ignorance decision ($\pi(x) = 0$) is independent at each leaf node
  3. Loss function equal some of losses betw/ leaf nodes

- For $G > 2$ we can repeat recursively
  1. Run over all covariates and splits (at most pn)
  2. Solve each subproblem independently within each split
  3. Loss function equal some of losses from each subproblem

Computational complexity

$\mathcal{G}$ is class of trees with G leaf nodes, p covariates and n observaions:

Thm Computational complexity is $\mathcal{O}(n^G p^G)$. (back)



Inference on optimal partitions $\mathcal{G}^\star$

- Estimate $\hat{\alpha}$ out-of-sample and construct test stat and quantile $q_{\alpha,1-\gamma}$

$$\hat{T}_\alpha(\hat{\alpha}^o) = \min_{\phi \in \mathcal{F}_\alpha} \hat{L}_\sigma(\pi^\alpha, \phi_\alpha) - \min_{\phi \in \mathcal{F}_{\hat{\alpha}}} \hat{L}_\sigma(\pi^{\hat{\alpha}}, \phi_\alpha)$$

Inference on optimal partitions $\mathcal{G}^\star$

- Estimate $\hat{\alpha}$ out-of-sample and construct test stat and quantile $q_{\alpha,1-\gamma}$

$$\hat{T}_\alpha(\hat{\alpha}^o) = \min_{\phi \in \mathcal{F}_\alpha} \hat{L}_\sigma(\pi^\alpha, \phi_\alpha) - \min_{\phi \in \mathcal{F}_{\hat{\alpha}}} \hat{L}_\sigma(\pi^{\hat{\alpha}}, \phi_\alpha)$$

such that $\lim_{|\mathcal{X}| \to \infty} P\left(\sqrt{|\mathcal{X}|}\,\hat{T}_\alpha(\hat{\alpha}^o) \leq q_{\alpha,1-\gamma}(\hat{\alpha}) | \hat{\alpha}\right) \geq 1 - \gamma$ (back)

Inference on optimal partitions $\mathcal{G}^\star$

- Estimate $\hat{\alpha}$ out-of-sample and construct test stat and quantile $q_{\alpha,1-\gamma}$

$$\hat{T}_\alpha(\hat{\alpha}^o) = \min_{\phi \in \mathcal{F}_\alpha} \hat{L}_\sigma(\pi^\alpha, \phi_\alpha) - \min_{\phi \in \mathcal{F}_{\hat{\alpha}}} \hat{L}_\sigma(\pi^{\hat{\alpha}}, \phi_\alpha)$$

such that $\lim_{|\mathcal{X}|\to\infty} P\left(\sqrt{|\mathcal{X}|}\,\hat{T}_\alpha(\hat{\alpha}^o) \leq q_{\alpha,1-\gamma}(\hat{\alpha})|\hat{\alpha}\right) \geq 1-\gamma$ (back)

$\Rightarrow$ Expression for $q_{\alpha,1-\gamma}$ in the paper provides exact coverage for non-degenerate distribution of $\hat{T}_\alpha$ and conservative coverage otherwise

Inference on optimal partitions $\mathcal{G}^\star$

- Estimate $\hat{\alpha}$ out-of-sample and construct test stat and quantile $q_{\alpha,1-\gamma}$

$$\hat{T}_\alpha(\hat{\alpha}^o) = \min_{\phi \in \mathcal{F}_\alpha} \hat{L}_\sigma(\pi^\alpha, \phi_\alpha) - \min_{\phi \in \mathcal{F}_{\hat{\alpha}}} \hat{L}_\sigma(\pi^{\hat{\alpha}}, \phi_\alpha)$$

such that $\lim_{|\mathcal{X}| \to \infty} P\left(\sqrt{|\mathcal{X}|}\,\hat{T}_\alpha(\hat{\alpha}^o) \leq q_{\alpha,1-\gamma}(\hat{\alpha})|\hat{\alpha}\right) \geq 1 - \gamma$ (back)

$\Rightarrow$ Expression for $q_{\alpha,1-\gamma}$ in the paper provides exact coverage for non-degenerate distribution of $\hat{T}_\alpha$ and conservative coverage otherwise

- For a given subset $\mathcal{G}'$ of interest estimate the set of optimal α

$$\hat{\mathcal{G}}_\gamma = \left\{\alpha \in \mathcal{G}' : \sqrt{|\mathcal{X}|}\,\hat{T}_\alpha(\hat{\alpha}) \leq q_{\alpha,1-\gamma}(\hat{\alpha})\right\},$$

Inference on optimal partitions $\mathcal{G}^\star$

- Estimate $\hat{\alpha}$ out-of-sample and construct test stat and quantile $q_{\alpha,1-\gamma}$

$$\hat{T}_\alpha(\hat{\alpha}^o) = \min_{\phi \in \mathcal{F}_\alpha} \hat{L}_\sigma(\pi^\alpha, \phi_\alpha) - \min_{\phi \in \mathcal{F}_{\hat{\alpha}}} \hat{L}_\sigma(\pi^{\hat{\alpha}}, \phi_\alpha)$$

such that $\lim_{|\mathcal{X}| \to \infty} P\left(\sqrt{|\mathcal{X}|}\,\hat{T}_\alpha(\hat{\alpha}^o) \leq q_{\alpha,1-\gamma}(\hat{\alpha})|\hat{\alpha}\right) \geq 1 - \gamma$ (back)

$\Rightarrow$ Expression for $q_{\alpha,1-\gamma}$ in the paper provides exact coverage for non-degenerate distribution of $\hat{T}_\alpha$ and conservative coverage otherwise

- For a given subset $\mathcal{G}'$ of interest estimate the set of optimal α

$$\hat{\mathcal{G}}_\gamma = \left\{\alpha \in \mathcal{G}' : \sqrt{|\mathcal{X}|}\,\hat{T}_\alpha(\hat{\alpha}) \leq q_{\alpha,1-\gamma}(\hat{\alpha})\right\}, \quad \gamma = \gamma^\star/|\mathcal{G}'|$$

Inference on optimal partitions $\mathcal{G}^\star$

- Estimate $\hat{\alpha}$ out-of-sample and construct test stat and quantile $q_{\alpha,1-\gamma}$

$$\hat{T}_\alpha(\hat{\alpha}^o) = \min_{\phi \in \mathcal{F}_\alpha} \hat{L}_\sigma(\pi^\alpha, \phi_\alpha) - \min_{\phi \in \mathcal{F}_{\hat{\alpha}}} \hat{L}_\sigma(\pi^{\hat{\alpha}}, \phi_\alpha)$$

such that $\lim_{|\mathcal{X}| \to \infty} P\left(\sqrt{|\mathcal{X}|}\,\hat{T}_\alpha(\hat{\alpha}^o) \leq q_{\alpha,1-\gamma}(\hat{\alpha}) | \hat{\alpha}\right) \geq 1 - \gamma$ (back)

$\Rightarrow$ Expression for $q_{\alpha,1-\gamma}$ in the paper provides exact coverage for non-degenerate distribution of $\hat{T}_\alpha$ and conservative coverage otherwise

- For a given subset $\mathcal{G}'$ of interest estimate the set of optimal α

$$\hat{\mathcal{G}}_\gamma = \left\{\alpha \in \mathcal{G}' : \sqrt{|\mathcal{X}|}\,\hat{T}_\alpha(\hat{\alpha}) \leq q_{\alpha,1-\gamma}(\hat{\alpha})\right\}, \quad \gamma = \gamma^\star / |\mathcal{G}'|$$

Thm Under regularities, $\lim_{|\mathcal{X}| \to \infty} P(\mathcal{G}^\star \cap \mathcal{G}' \subseteq \hat{\mathcal{G}}_\gamma) \geq 1 - \gamma^\star$.

Inference on optimal partitions $\mathcal{G}^\star$

- Estimate $\hat{\alpha}$ out-of-sample and construct test stat and quantile $q_{\alpha,1-\gamma}$

$$\hat{T}_\alpha(\hat{\alpha}^o) = \min_{\phi \in \mathcal{F}_\alpha} \hat{L}_\sigma(\pi^\alpha, \phi_\alpha) - \min_{\phi \in \mathcal{F}_{\hat{\alpha}}} \hat{L}_\sigma(\pi^{\hat{\alpha}}, \phi_\alpha)$$

such that $\lim_{|\mathcal{X}| \to \infty} P\left(\sqrt{|\mathcal{X}|}\,\hat{T}_\alpha(\hat{\alpha}^o) \leq q_{\alpha,1-\gamma}(\hat{\alpha})|\hat{\alpha}\right) \geq 1-\gamma$ (back)

$\Rightarrow$ Expression for $q_{\alpha,1-\gamma}$ in the paper provides exact coverage for non-degenerate distribution of $\hat{T}_\alpha$ and conservative coverage otherwise

- For a given subset $\mathcal{G}'$ of interest estimate the set of optimal α

$$\hat{\mathcal{G}}_\gamma = \left\{\alpha \in \mathcal{G}' : \sqrt{|\mathcal{X}|}\,\hat{T}_\alpha(\hat{\alpha}) \leq q_{\alpha,1-\gamma}(\hat{\alpha})\right\}, \quad \gamma = \gamma^\star/|\mathcal{G}'|$$

Thm Under regularities, $\lim_{|\mathcal{X}| \to \infty} P(\mathcal{G}^\star \cap \mathcal{G}' \subseteq \hat{\mathcal{G}}_\gamma) \geq 1-\gamma^\star$. $\forall \alpha \in \mathcal{G}'$ with $L_\sigma(\alpha) - \min_{\alpha'} L_\sigma(\alpha')$ bounded from below, $\lim_{|\mathcal{X}| \to \infty} P(\alpha \in \hat{\mathcal{G}}_\gamma) = 0$