Evidence aggregation with ignorance in mind: learning what we do not know for archetypes discovery

Emily Breza
Harvard

Arun Chandrasekhar
Stanford

Davide Viviano
Harvard

November 3, 2026

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 evidence aggregation
  - Goal often is to aggregate evidence across all individuals
  - Aggregation relies on model assumptions (e.g. via shrinkage/sparsity)
  - It is often dictated by communication constraints

Problem description

- Growing # of experiments across environments and individuals
  E.g. Male/female enterpreneurs in different countries
- Moved institutions (JPAL) towards evidence aggregation
  - Goal often is to aggregate evidence across all individuals
  - Aggregation relies on model assumptions (e.g. via shrinkage/sparsity)
  - It is often dictated by communication constraints
- 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 can be misleading

Problem description

- Growing # of experiments across environments and individuals
  E.g. Male/female enterpreneurs in different countries
- Moved institutions (JPAL) towards evidence aggregation
  - Goal often is to aggregate evidence across all individuals
  - Aggregation relies on model assumptions (e.g. via shrinkage/sparsity)
  - It is often dictated by communication constraints
- 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 can be misleading
- Goal: learn from the data **when** and **how** to aggregate evidence across contexts/individuals and when instead we need more evidence

Setup and goal

- We conduct experiments across heterogeneous individuals/sites
  - Types $x \in \mathcal{X}$, with large $|\mathcal{X}|$ large (covariates, experiment type, country, ...)
  - Each individual experience an effect $\tau(x)$

Setup and goal

- We conduct experiments across heterogeneous individuals/sites
  - Types $x \in \mathcal{X}$, with large $|\mathcal{X}|$ 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 $\mathbb{V}(\hat{\tau}(x))$

Setup and goal

- We conduct experiments across heterogeneous individuals/sites
  - Types $x \in \mathcal{X}$, with large $|\mathcal{X}|$ 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 $\mathbb{V}(\hat{\tau}(x))$

E.g. $\hat{\tau}(x)$ *diff/ in means for types* x *and* $\hat{\eta}(x)^2$ *its sampling variance*

Setup and goal

- We conduct experiments across heterogeneous individuals/sites
  - Types $x \in \mathcal{X}$, with large $|\mathcal{X}|$ 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 $\mathbb{V}(\hat{\tau}(x))$

  E.g. $\hat{\tau}(x)$ *diff/ in means for types* x *and* $\hat{\eta}(x)^2$ *its sampling variance*

- Goal is two fold:
  - Predict $\tau(x)$ when information is portable across groups
  - Claim for more evidence if cannot reliably aggregate existing evidence

Setup and goal

- We conduct experiments across heterogeneous individuals/sites
  - Types $x \in \mathcal{X}$, with large $|\mathcal{X}|$ 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 $\mathbb{V}(\hat{\tau}(x))$

  E.g. $\hat{\tau}(x)$ *diff/ in means for types* x *and* $\hat{\eta}(x)^2$ *its sampling variance*

- Goal is two fold:
  - Predict $\tau(x)$ when information is portable across groups
  - Claim for more evidence if cannot reliably aggregate existing evidence

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

Setup and goal

- We conduct experiments across heterogeneous individuals/sites
  - Types $x \in \mathcal{X}$, with large $|\mathcal{X}|$ 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 $\mathbb{V}(\hat{\tau}(x))$

  E.g. $\hat{\tau}(x)$ *diff/ in means for types* x *and* $\hat{\eta}(x)^2$ *its sampling variance*

- Goal is two fold:
  - Predict $\tau(x)$ when information is portable across groups
  - Claim for more evidence if cannot reliably aggregate existing evidence

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

- Steps of the analysis
  - Introduce decision problem/objective + estimation
  - Study trades-off through formal guarantees
  - Illustrate implications for anti-poverty programs

Related literature

- Meta-analysis and transfer learning (Borenstein et al., 2021; Meager, 2022; Crosta et al., 2024; Menzel, 2023; Ishihara and Kitagawa, 2021; Deeb and de Chaisemartin, 2019; Adjaho and Christensen, 2022; Andrews et al., 2022, . . .)

  Policy learning and effect heterogeneity (Kitagawa and Tetenov, 2018; Manski, 2004; Athey and Wager, 2021; Chernozhukov et al., 2018; Bonhomme and Manresa, 2015; Viviano, 2024, . . .)

Related literature

- Meta-analysis and transfer learning (Borenstein et al., 2021; Meager, 2022; Crosta et al., 2024; Menzel, 2023; Ishihara and Kitagawa, 2021; Deeb and de Chaisemartin, 2019; Adjaho and Christensen, 2022; Andrews et al., 2022, . . .)

  Policy learning and effect heterogeneity (Kitagawa and Tetenov, 2018; Manski, 2004; Athey and Wager, 2021; Chernozhukov et al., 2018; Bonhomme and Manresa, 2015; Viviano, 2024, . . .)

  $\Rightarrow$ Makes predictions/decisions for all contexts
  $\Rightarrow$ Here detect where to explore and where and how to form predictions

Related literature

- Meta-analysis and transfer learning (Borenstein et al., 2021; Meager, 2022; Crosta et al., 2024; Menzel, 2023; Ishihara and Kitagawa, 2021; Deeb and de Chaisemartin, 2019; Adjaho and Christensen, 2022; Andrews et al., 2022, . . .)

  Policy learning and effect heterogeneity (Kitagawa and Tetenov, 2018; Manski, 2004; Athey and Wager, 2021; Chernozhukov et al., 2018; Bonhomme and Manresa, 2015; Viviano, 2024, . . .)

  $\Rightarrow$ Makes predictions/decisions for all contexts
  $\Rightarrow$ Here detect where to explore and where and how to form predictions

- Site selection/sampling (e.g., Olea et al., 2024; Gechter et al., 2024, . . .)
  $\Rightarrow$ Ex-ante choice of where to run an experiment
  $\Rightarrow$ Here dual goal of evidence aggregation with existing data + exploration

Related literature

- Meta-analysis and transfer learning (Borenstein et al., 2021; Meager, 2022; Crosta et al., 2024; Menzel, 2023; Ishihara and Kitagawa, 2021; Deeb and de Chaisemartin, 2019; Adjaho and Christensen, 2022; Andrews et al., 2022, . . .)

  Policy learning and effect heterogeneity (Kitagawa and Tetenov, 2018; Manski, 2004; Athey and Wager, 2021; Chernozhukov et al., 2018; Bonhomme and Manresa, 2015; Viviano, 2024, . . .)

  $\Rightarrow$ Makes predictions/decisions for all contexts
  $\Rightarrow$ Here detect where to explore and where and how to form predictions

- Site selection/sampling (e.g., Olea et al., 2024; Gechter et al., 2024, . . .)

  $\Rightarrow$ Ex-ante choice of where to run an experiment
  $\Rightarrow$ Here dual goal of evidence aggregation with existing data + exploration

- 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)

  $\Rightarrow$ Here study evidence aggregation w/ future experimentation (for CATEs)

Content

1. Setup and estimator
2. Properties
3. Practical implementation and illustration
4. Empirical application

A framework to learn when to aggregate evidence



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

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

A framework to learn when to aggregate evidence



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

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

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

A framework to learn when to aggregate evidence



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

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

- Making a prediction = predict $\tau(x)$ with $\phi(x), \phi \in \mathcal{F}$
  - $\mathcal{F}$ characterizes model/communication constraints (later: selection of $\mathcal{F}$)
  - But not all $\tau(x)$ may be well approximated by $\phi(x)$
- Admit ignorance = policy function in a set Π (bounded VC-complexity)

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

A framework to learn when to aggregate evidence



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

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

- Making a prediction = predict $\tau(x)$ with $\phi(x), \phi \in \mathcal{F}$
  - $\mathcal{F}$ characterizes model/communication constraints (later: selection of $\mathcal{F}$)
  - But not all $\tau(x)$ may be well approximated by $\phi(x)$
- Admit ignorance = policy function in a set Π (bounded VC-complexity)

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

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

Ignorance-aware predictions: population version

For now, ignore sampling uncertainty and let $c(x) = c$.

For $c = \infty$, the population objective reads as

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

Ignorance-aware predictions: population version

For now, ignore sampling uncertainty and let $c(x) = c$.

For $c \geq 0$, the population objective, denoted by $\boldsymbol{L_c(\phi, \pi)}$, is

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

Ignorance-aware predictions: population version

For now, ignore sampling uncertainty and let $c(x) = c$.

For $c \geq 0$, the population objective, denoted by $\boldsymbol{L_c(\phi, \pi)}$, is

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

Dual characterization

- Minimize error with a budget constraint for ignorance.
- Minimize the cost of ignorance with a constraint on error.

Ignorance-aware predictions: population version

For now, ignore sampling uncertainty and let $c(x) = c$.

For $c \geq 0$, the population objective, denoted by $\boldsymbol{L_c(\phi, \pi)}$, is

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

Dual characterization

- Minimize error with a budget constraint for ignorance.
- Minimize the cost of ignorance with a constraint on error.

$\Rightarrow$ Robustify predictions by detecting "hard to predict" units:

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

Ignorance-aware predictions: population version

For now, ignore sampling uncertainty and let $c(x) = c$.

For $c \geq 0$, the population objective, denoted by $\boldsymbol{L_c(\phi, \pi)}$, is

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

Dual characterization

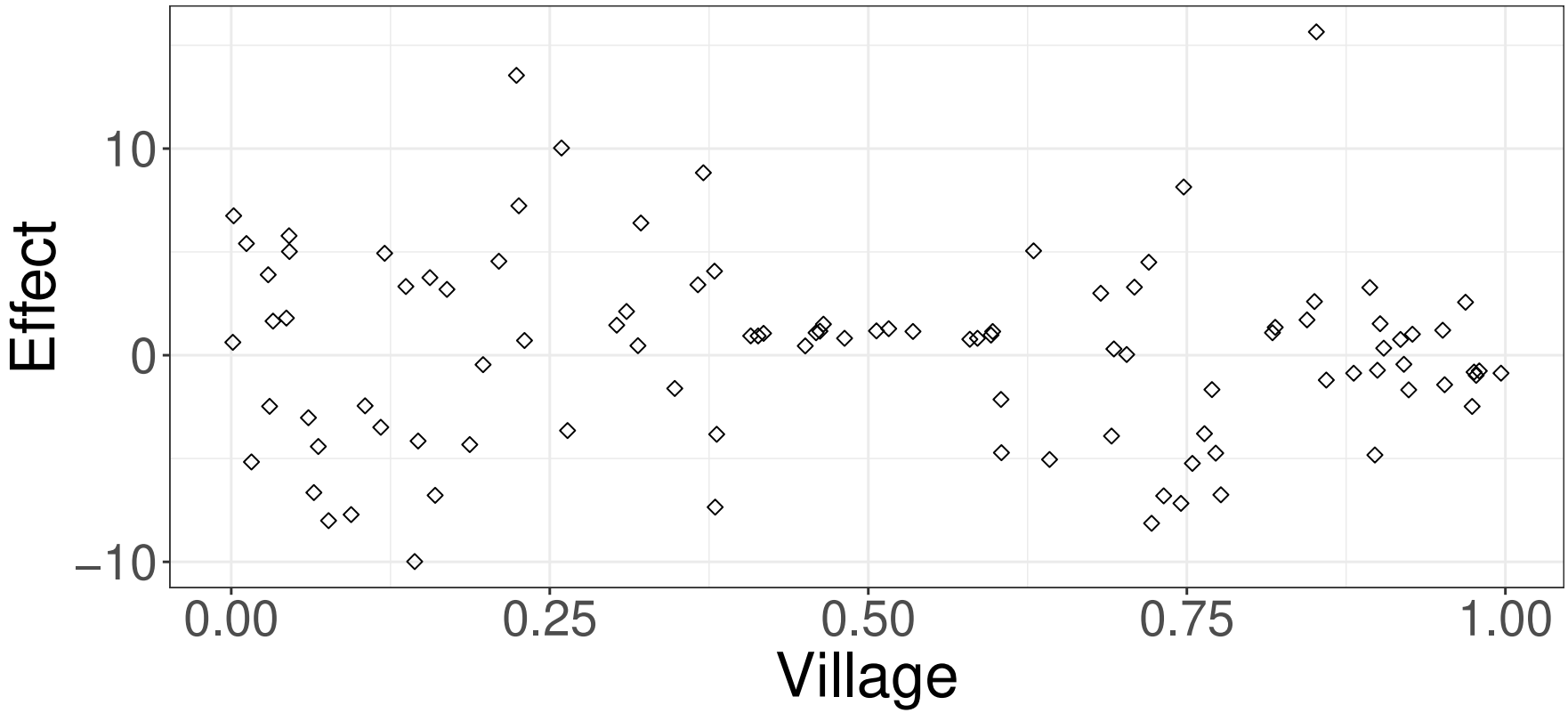
- Minimize error with a budget constraint for ignorance.
- Minimize the cost of ignorance with a constraint on error.

$\Rightarrow$ Robustify predictions by detecting "hard to predict" units:

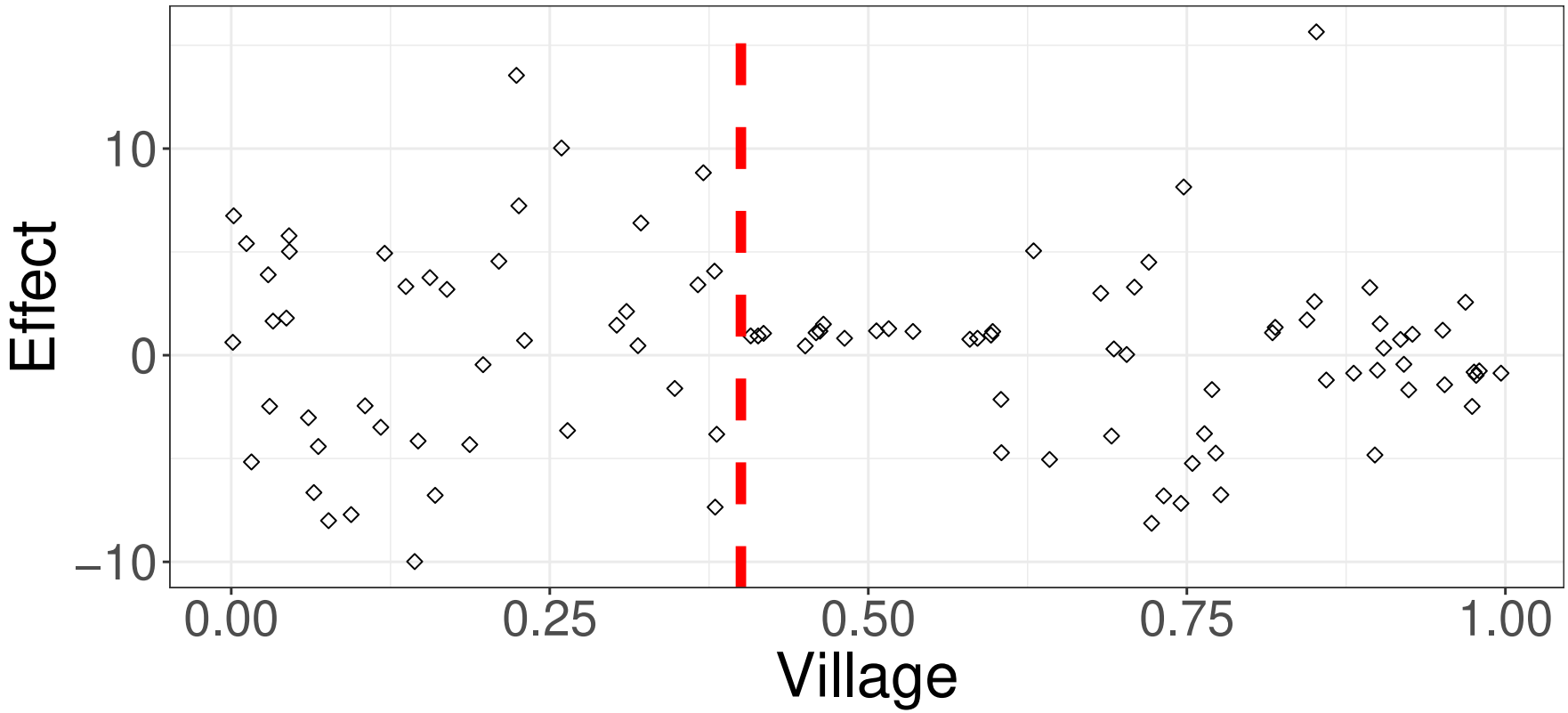
$$\min_{\pi \in \Pi,\, \phi \in \mathcal{F}} \sum_x p(x) \Big(\tau(x) - \phi(x)\Big)^2 \pi(x) \quad \text{s.t.} \quad \sum_x p(x)\pi(x) \geq \lambda$$

$\Rightarrow$ (Later: practical choices of c)

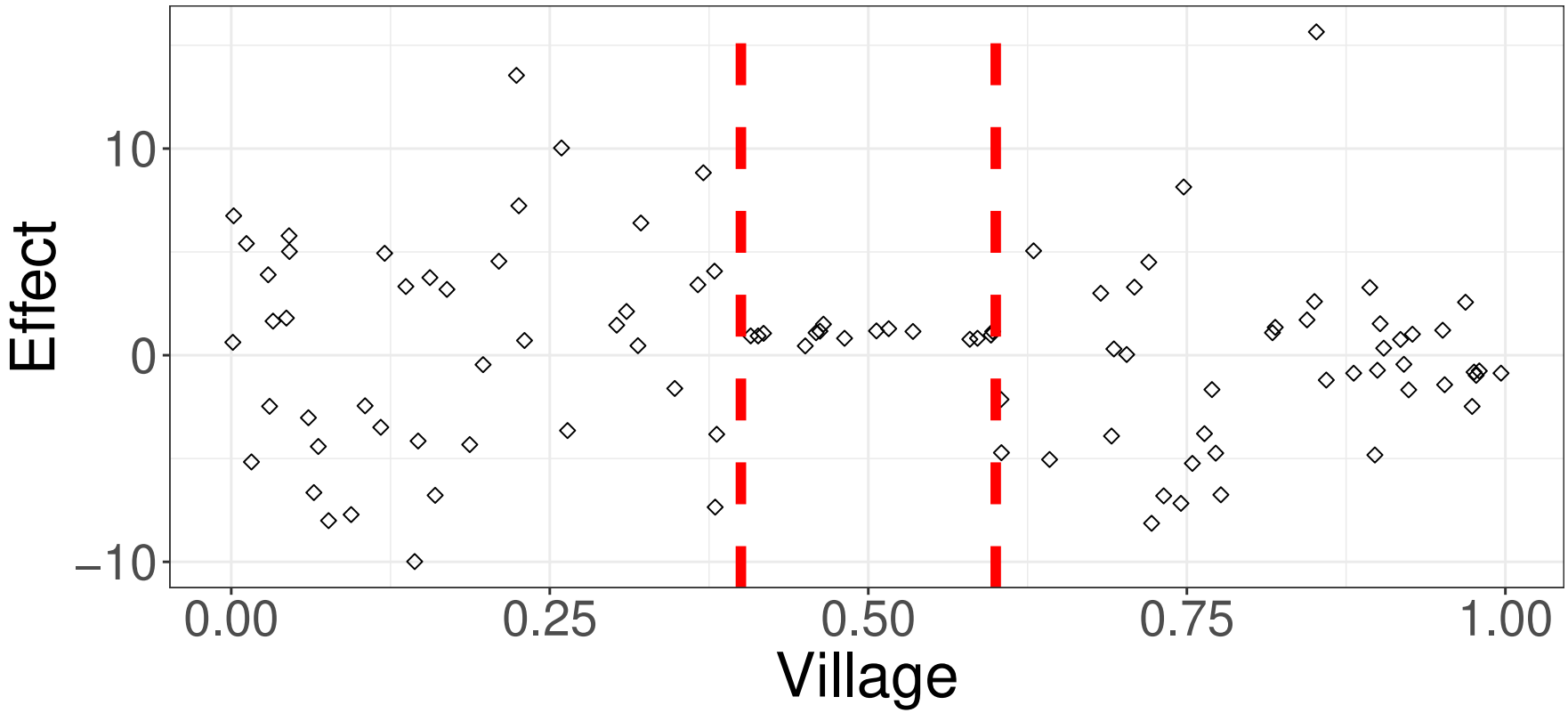
Example: small c



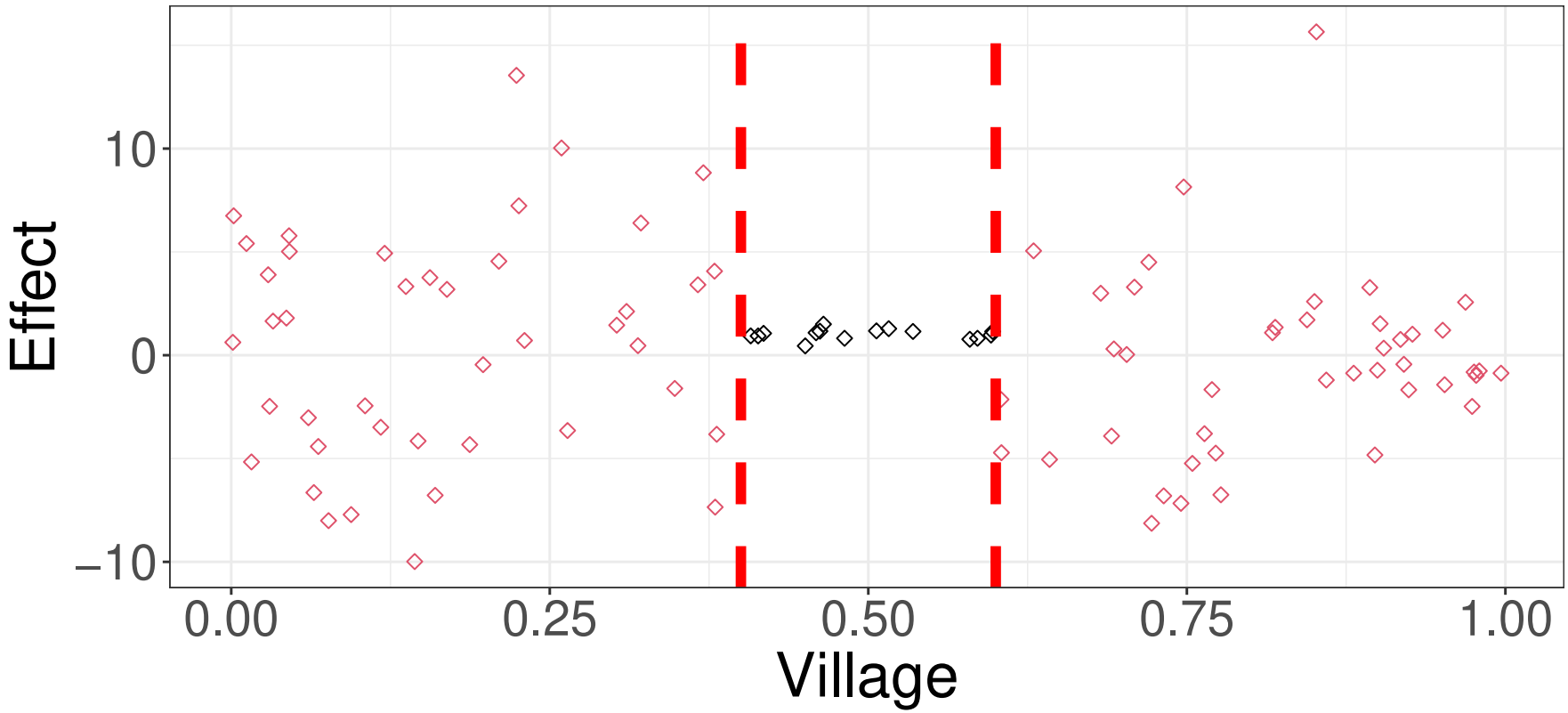
Example: small c



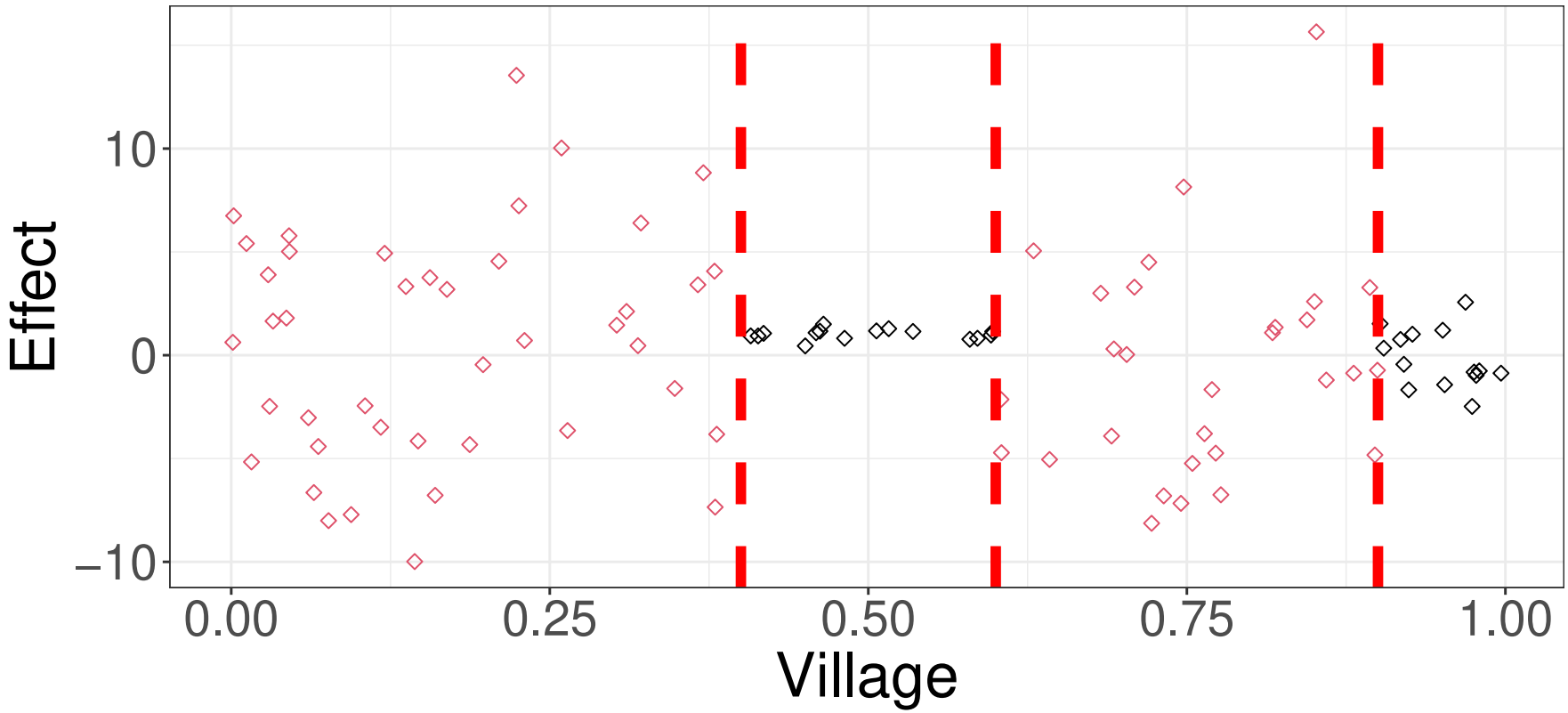
Example: small c



Example: small c



Example: increase c

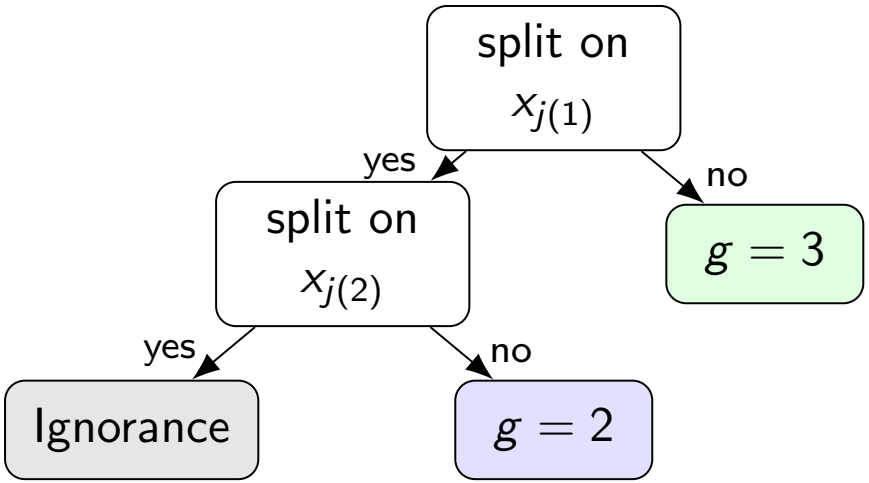


Practical methods

- Study anti-poverty program [Banerjee et al., 2015]
- Each type x is (small/noisy) group of size $n(x)$, large number of types $|\mathcal{X}|$
- x: age, food security, baseline assets/consumption, ...

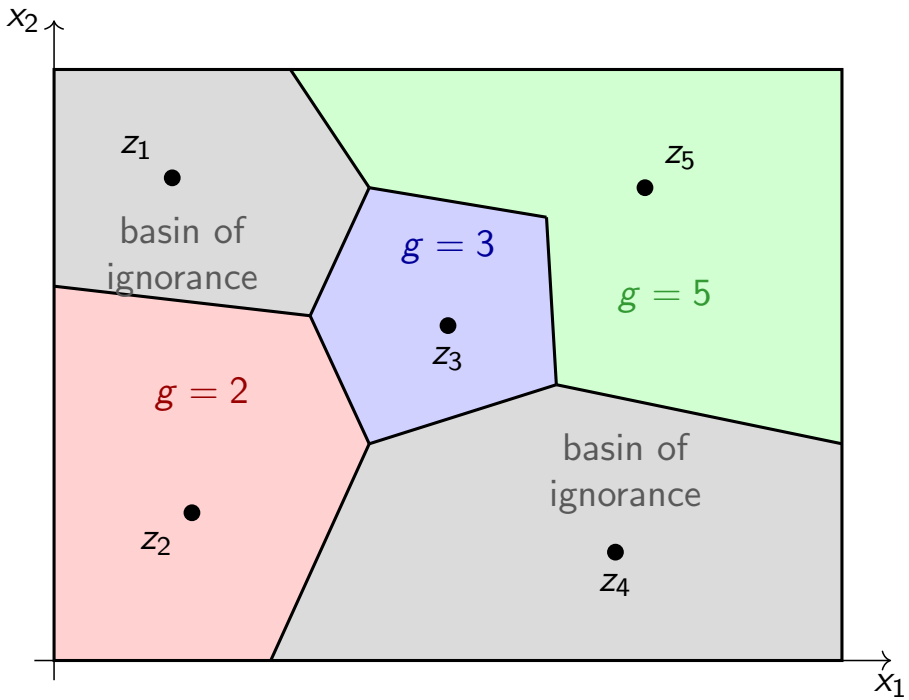
Practical methods

- Study anti-poverty program [Banerjee et al., 2015]
- Each type x is (small/noisy) group of size $n(x)$, large number of types $|\mathcal{X}|$
- x: age, food security, baseline assets/consumption, ...

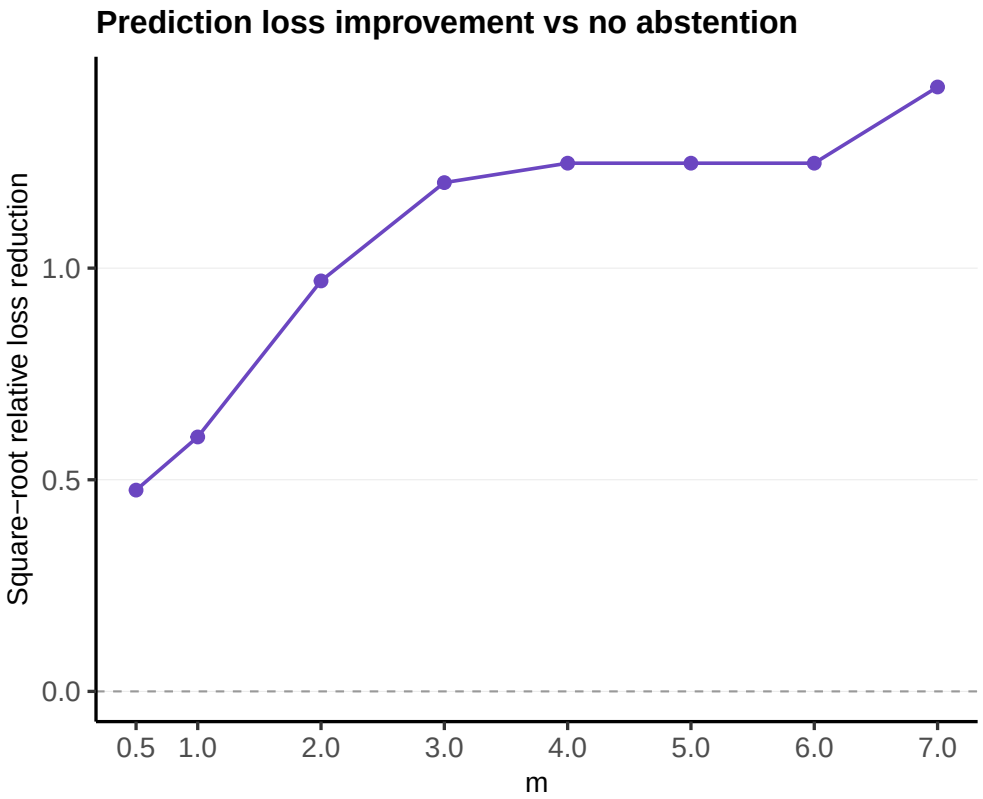


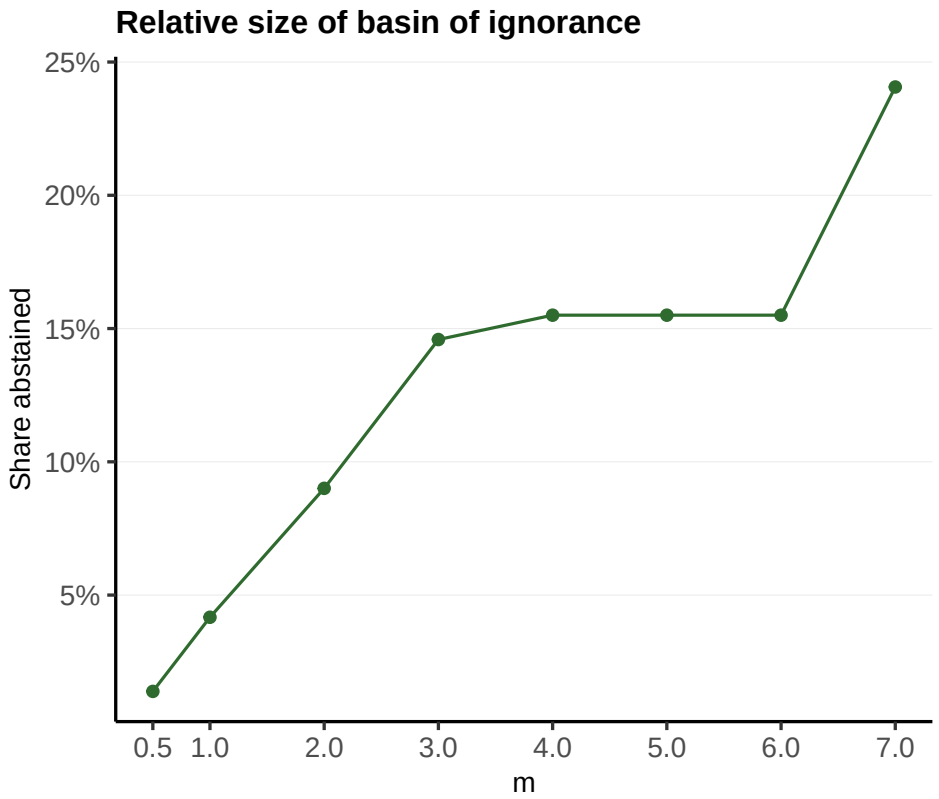
Practical methods

- Study anti-poverty program [Banerjee et al., 2015]
- Each type x is (small/noisy) group of size $n(x)$, large number of types $|\mathcal{X}|$
- x: age, food security, baseline assets/consumption, ...



Example from multifaceted anti-poverty program





Estimation

- We want to obtain unbiased estimator of L_c
- We collect *noisy* unbiased $\hat{\tau}(x)$, $\hat{c}(x)$, and $\hat{\eta}(x)^2$ of $\mathbb{V}(\hat{\tau}(x))$

Estimation

- We want to obtain unbiased estimator of L_c
- We collect *noisy* unbiased $\hat{\tau}(x)$, $\hat{c}(x)$, and $\hat{\eta}(x)^2$ of $\mathbb{V}(\hat{\tau}(x))$
  - In experiment, mean difference & sample variance for small group x

Estimation

- We want to obtain unbiased estimator of L_c
- We collect *noisy* unbiased $\hat{\tau}(x)$, $\hat{c}(x)$, and $\hat{\eta}(x)^2$ of $\mathbb{V}(\hat{\tau}(x))$
  - In experiment, mean difference & sample variance for small group x

    Ex For outcomes Y_i, binary randomized treatments $D_i \sim_{iid} \mathrm{Bern}(0.5)$

$$\hat{\tau}(x) = \frac{1}{n(x)} \sum_{i:X_i=x} \tilde{Y}_i, \qquad \tilde{Y}_i = \frac{Y_i D_i}{0.5} - \frac{Y_i(1-D_i)}{0.5}$$

Estimation

- We want to obtain unbiased estimator of L_c
- We collect *noisy* unbiased $\hat{\tau}(x)$, $\hat{c}(x)$, and $\hat{\eta}(x)^2$ of $\mathbb{V}(\hat{\tau}(x))$
  - In experiment, mean difference & sample variance for small group x

    Ex For outcomes Y_i, binary randomized treatments $D_i \sim_{iid} \mathrm{Bern}(0.5)$

$$\hat{\tau}(x) = \frac{1}{n(x)} \sum_{i:X_i=x} \tilde{Y}_i, \qquad \tilde{Y}_i = \frac{Y_i D_i}{0.5} - \frac{Y_i(1-D_i)}{0.5}$$

- Estimate $\hat{\pi}, \hat{\phi}$ by minimizing empirical loss $\hat{L}_c(\cdot)$ proportional to

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x p(x) \Big\{ \underbrace{\Big(\hat{\tau}(x) - \phi(x)\Big)^2}_{\text{estimated error}}$$

Estimation

- We want to obtain unbiased estimator of L_c
- We collect *noisy* unbiased $\hat{\tau}(x)$, $\hat{c}(x)$, and $\hat{\eta}(x)^2$ of $\mathbb{V}(\hat{\tau}(x))$
  - In experiment, mean difference & sample variance for small group x

    Ex For outcomes Y_i, binary randomized treatments $D_i \sim_{iid} \mathrm{Bern}(0.5)$

$$\hat{\tau}(x) = \frac{1}{n(x)} \sum_{i:X_i=x} \tilde{Y}_i, \qquad \tilde{Y}_i = \frac{Y_i D_i}{0.5} - \frac{Y_i(1-D_i)}{0.5}$$

- Estimate $\hat{\pi}, \hat{\phi}$ by minimizing empirical loss $\hat{L}_c(\cdot)$ proportional to

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x p(x) \Big\{ \underbrace{\Big(\hat{\tau}(x) - \phi(x)\Big)^2}_{\text{estimated error}} - \underbrace{\hat{\eta}(x)^2}_{\text{bias correction}}$$

Estimation

- We want to obtain unbiased estimator of L_c
- We collect *noisy* unbiased $\hat{\tau}(x)$, $\hat{c}(x)$, and $\hat{\eta}(x)^2$ of $\mathbb{V}(\hat{\tau}(x))$
  - In experiment, mean difference & sample variance for small group x

    Ex For outcomes Y_i, binary randomized treatments $D_i \sim_{iid} \mathrm{Bern}(0.5)$

$$\hat{\tau}(x) = \frac{1}{n(x)} \sum_{i:X_i=x} \tilde{Y}_i, \qquad \tilde{Y}_i = \frac{Y_i D_i}{0.5} - \frac{Y_i(1-D_i)}{0.5}$$

- Estimate $\hat{\pi}, \hat{\phi}$ by minimizing empirical loss $\hat{L}_c(\cdot)$ proportional to

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x p(x) \Big\{ \underbrace{\Big(\hat{\tau}(x) - \phi(x)\Big)^2}_{\text{estimated error}} - \underbrace{\hat{\eta}(x)^2}_{\text{bias correction}} - \underbrace{\hat{c}(x)}_{\text{cost of ignorance}} \Big\} \pi(x)$$

Estimation

- We want to obtain unbiased estimator of L_c
- We collect *noisy* unbiased $\hat{\tau}(x)$, $\hat{c}(x)$, and $\hat{\eta}(x)^2$ of $\mathbb{V}(\hat{\tau}(x))$
  - In experiment, mean difference & sample variance for small group x

    Ex For outcomes Y_i, binary randomized treatments $D_i \sim_{iid} \mathrm{Bern}(0.5)$

$$\hat{\tau}(x) = \frac{1}{n(x)} \sum_{i:X_i=x} \tilde{Y}_i, \qquad \tilde{Y}_i = \frac{Y_i D_i}{0.5} - \frac{Y_i(1-D_i)}{0.5}$$

- Estimate $\hat{\pi}, \hat{\phi}$ by minimizing empirical loss $\hat{L}_c(\cdot)$ proportional to

$$\min_{\pi \in \Pi, \phi \in \mathcal{F}} \sum_x p(x) \Big\{ \underbrace{\big(\hat{\tau}(x) - \phi(x)\big)^2}_{\text{estimated error}} - \underbrace{\hat{\eta}(x)^2}_{\text{bias correction}} - \underbrace{\hat{c}(x)}_{\text{cost of ignorance}} \Big\} \pi(x)$$

- Target: $\mathbb{E}[(\tau(x) - \phi(x))^2] = \mathbb{E}[(\hat{\tau}(x) - \phi(x))^2] - \mathbb{V}(\hat{\tau}(x))$
  - To decide when to pool, compare between to within variation

Content

1. Setup and estimator
2. Properties
3. Practical implementation and illustration
4. Empirical application

Regret and conditions

- Complexity restrictions: Π has VC-complexity ν_Π and $\mathcal{F}$ Dudley's entropy integral $\sqrt{C(\mathcal{F})} \Rightarrow$ Large set of classes (parametric and high-dim)

Regret and conditions

- Complexity restrictions: Π has VC-complexity v_Π and $\mathcal{F}$ Dudley's entropy integral $\sqrt{C(\mathcal{F})} \Rightarrow$ Large set of classes (parametric and high-dim)

Thm Assuming estimators are unbiased with bounded third moment,

$$\mathbb{E}\left[L_c(\hat{\phi}, \hat{\pi}) - \inf_{\pi \in \Pi, f \in \mathcal{F}} L_c(f, \pi)\right] \lesssim \sqrt{\frac{C(\mathcal{F}) + v_\Pi}{|\mathcal{X}|}}$$

Regret and conditions

- Complexity restrictions: Π has VC-complexity v_Π and $\mathcal{F}$ Dudley's entropy integral $\sqrt{C(\mathcal{F})} \Rightarrow$ Large set of classes (parametric and high-dim)

Thm Assuming estimators are unbiased with bounded third moment,

$$\mathbb{E}\left[L_c(\hat{\phi}, \hat{\pi}) - \inf_{\pi \in \Pi, f \in \mathcal{F}} L_c(f, \pi)\right] \lesssim \sqrt{\frac{C(\mathcal{F}) + v_\Pi}{|\mathcal{X}|}}$$

$\Rightarrow$ Leading case is high-dim x, rate is $|\mathcal{X}|^{-1/2}$

$\Rightarrow$ A large number of (noisy) types x improves error on average

$\Rightarrow$ Bound include trees, clustering on covariates, RF, ridge, ...

Regret and conditions

- Complexity restrictions: Π has VC-complexity v_Π and $\mathcal{F}$ Dudley's entropy integral $\sqrt{C(\mathcal{F})} \Rightarrow$ Large set of classes (parametric and high-dim)

Thm Assuming estimators are unbiased with bounded third moment,

$$\mathbb{E}\left[L_c(\hat{\phi}, \hat{\pi}) - \inf_{\pi \in \Pi, f \in \mathcal{F}} L_c(f, \pi)\right] \lesssim \sqrt{\frac{C(\mathcal{F}) + v_\Pi}{|\mathcal{X}|}}$$

$\Rightarrow$ Leading case is high-dim x, rate is $|\mathcal{X}|^{-1/2}$

$\Rightarrow$ A large number of (noisy) types x improves error on average

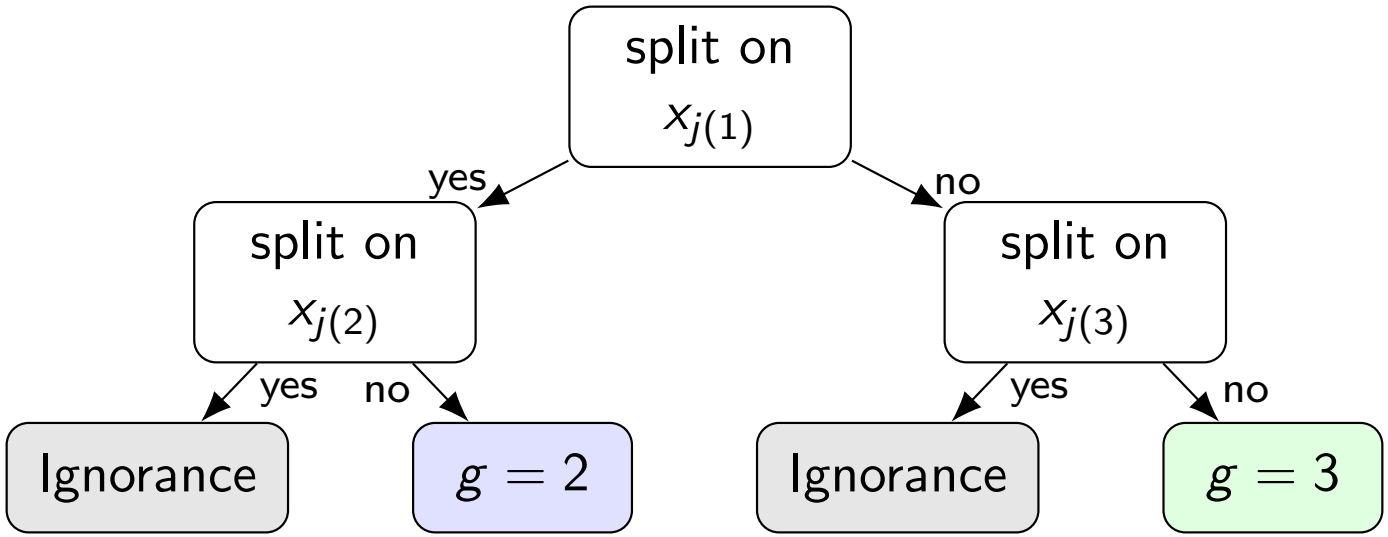
$\Rightarrow$ Bound include trees, clustering on covariates, RF, ridge, ...

$\Rightarrow$ In the paper:

- with growing sample size within each x gives us faster rate
- minimax matching lower bound for clustering algorithms

Examples with G groups

- Trees with $x \in \mathbb{R}^d$, complexity $\lesssim dG\log(G)$

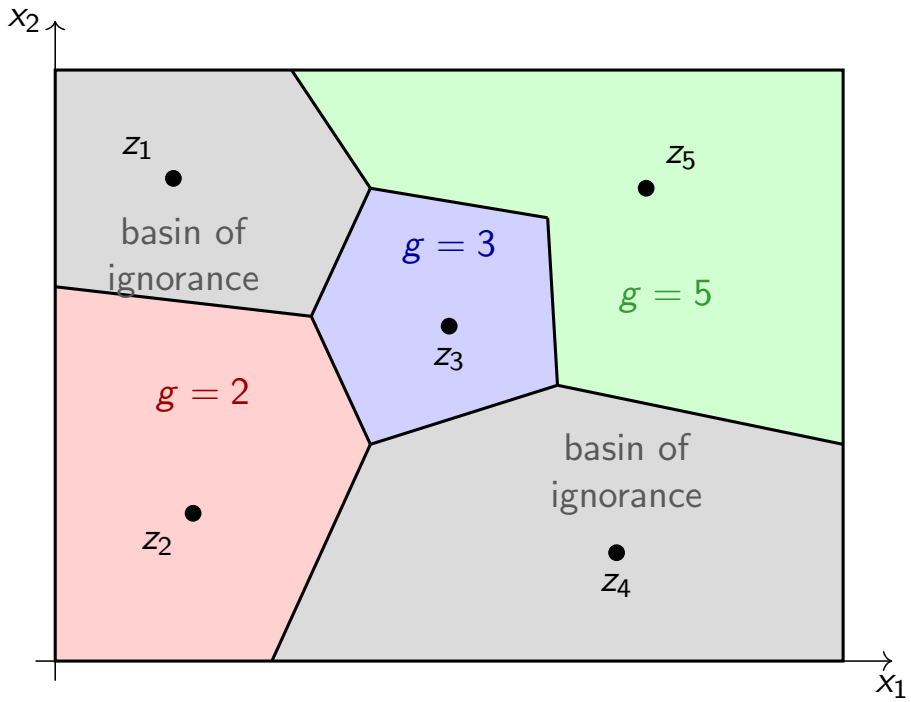


Examples with G groups

- G-clustering in $x \in \mathbb{R}^d$, complexity $\lesssim dG^2 \log(G)^2$

Examples with G groups

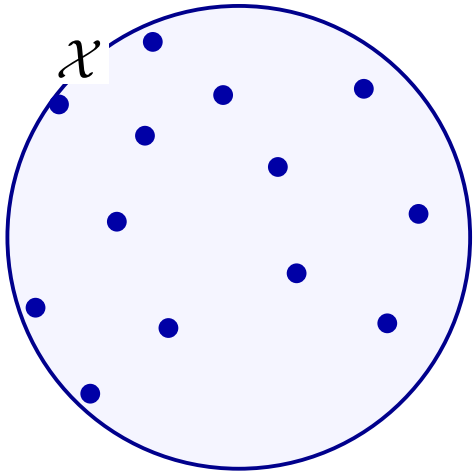
- G-clustering in $x \in \mathbb{R}^d$, complexity $\lesssim dG^2 \log(G)^2$



Global and local approximation error

$f \in \mathcal{F}_j$: candidate model $\quad$ $f^\star$: oracle target

Predict everywhere



Global approximation error

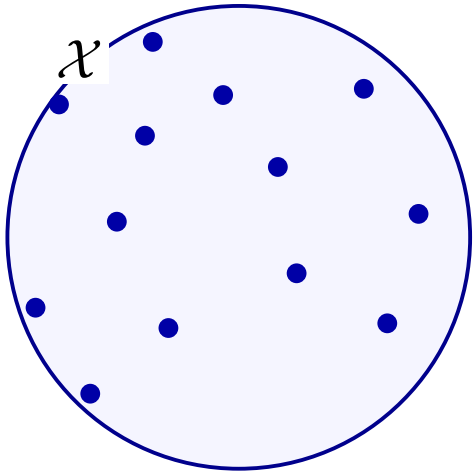
$f \longleftrightarrow f^\star$

Compare over all of $\mathcal{X}$

Global and local approximation error

$f \in \mathcal{F}_j$: candidate model $\quad$ $f^\star$: oracle target

Predict everywhere



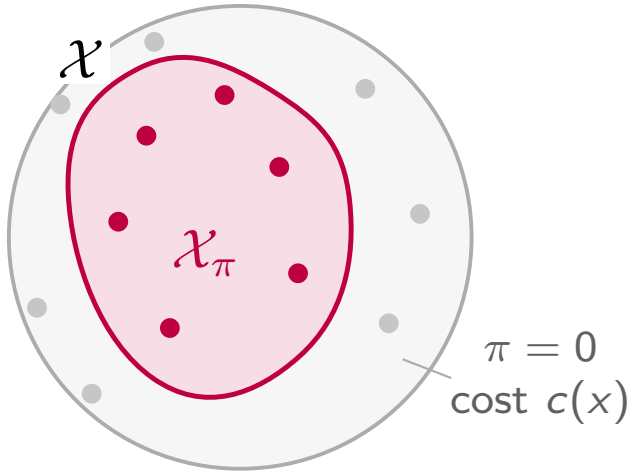
Global approximation error

$f \longleftrightarrow f^\star$

Compare over all of $\mathcal{X}$

choose π

Predict selectively



Local approximation error

$f \longleftrightarrow f^\star$

Compare on $\mathcal{X}_\pi = \{x : \pi(x) = 1\}$

A simpler model can approximate $f^\star$ well on a selected set.

Using abstention to improve model selection

Choice between models

- set of models $\mathcal{F}_j$ of increasing complexity $j \in \{1, \cdots, J\}$
- allow for abstention
- compare to oracle $f^\star \in \mathcal{F}_J$ with most complex model and no noise

Using abstention to improve model selection

Choice between models

- set of models $\mathcal{F}_j$ of increasing complexity $j \in \{1, \cdots, J\}$
- allow for abstention
- compare to oracle $f^\star \in \mathcal{F}_J$ with most complex model and no noise

How we choose

- fit each on training sample
- evaluate and select out-of-sample

$\Rightarrow$ select best prediction and prediction/ignorance sets

Using abstention to improve model selection

Choice between models

- set of models $\mathcal{F}_j$ of increasing complexity $j \in \{1, \cdots, J\}$
- allow for abstention
- compare to oracle $f^\star \in \mathcal{F}_J$ with most complex model and no noise

How we choose

- fit each on training sample
- evaluate and select out-of-sample

$\Rightarrow$ select best prediction and prediction/ignorance sets

$$\textbf{Gap to oracle} \lesssim \min_j \Big\{ \quad \textbf{precision}_j \quad + \textbf{local approx. error}_j + \textbf{abstention cost}_j \Big\}$$

Using abstention to improve model selection

Choice between models

- set of models $\mathcal{F}_j$ of increasing complexity $j \in \{1, \cdots, J\}$
- allow for abstention
- compare to oracle $f^\star \in \mathcal{F}_J$ with most complex model and no noise

How we choose

- fit each on training sample
- evaluate and select out-of-sample

$\Rightarrow$ select best prediction and prediction/ignorance sets

Gap to oracle $\lesssim \min_j \Big\{$ **precision**$_j$ $+$**local approx. error**$_j$$+$**abstention cost**$_j \Big\}$

Loss above the oracle

$\mathbb{E}\big[L_c(\hat{\phi}_{\hat{j}}, \hat{\pi}_{\hat{j}}) - L_c^\star\big]$

Using abstention to improve model selection

Choice between models

- set of models $\mathcal{F}_j$ of increasing complexity $j \in \{1, \cdots, J\}$
- allow for abstention
- compare to oracle $f^\star \in \mathcal{F}_J$ with most complex model and no noise

How we choose

- fit each on training sample
- evaluate and select out-of-sample

$\Rightarrow$ select best prediction and prediction/ignorance sets

$$\textbf{Gap to oracle} \lesssim \min_j \Big\{ \quad \textbf{precision}_j \quad + \textbf{local approx. error}_j + \textbf{abstention cost}_j \Big\}$$

Loss above the oracle

$$\mathbb{E}\big[L_c(\hat{\phi}_{\hat{j}}, \hat{\pi}_{\hat{j}}) - L_c^\star\big]$$

Estimation error rising with complexity

$$\sim \sqrt{\frac{C(\mathcal{F}_j) + v_{\Pi_j}}{|\mathcal{X}|}}$$

Using abstention to improve model selection

Choice between models

- set of models $\mathcal{F}_j$ of increasing complexity $j \in \{1, \cdots, J\}$
- allow for abstention
- compare to oracle $f^\star \in \mathcal{F}_J$ with most complex model and no noise

How we choose

- fit each on training sample
- evaluate and select out-of-sample

$\Rightarrow$ select best prediction and prediction/ignorance sets

$$\textbf{Gap to oracle} \lesssim \min_j \Big\{ \quad \textbf{precision}_j \quad + \textbf{local approx. error}_j + \textbf{abstention cost}_j \Big\}$$

Loss above the oracle

$$\mathbb{E}\big[L_c(\hat{\phi}_{\hat{j}}, \hat{\pi}_{\hat{j}}) - L_c^\star\big]$$

Estimation error rising with complexity

$$\sim \sqrt{\frac{C(\mathcal{F}_j) + v_{\Pi_j}}{|\mathcal{X}|}}$$

Approximation error within the prediction set

$$L_{|1}(f_j^\star, \pi_j^\star) - L_{|1}(f^\star, \pi_j^\star)$$

$$L_{|1}(f, \pi) = \mathbb{E}_x\big[(f(x) - \tau(x))^2 \mid \pi(x) = 1\big].$$

Using abstention to improve model selection

Choice between models

- set of models $\mathcal{F}_j$ of increasing complexity $j \in \{1, \cdots, J\}$
- allow for abstention
- compare to oracle $f^\star \in \mathcal{F}_J$ with most complex model and no noise

How we choose

- fit each on training sample
- evaluate and select out-of-sample

$\Rightarrow$ select best prediction and prediction/ignorance sets

$$\textbf{Gap to oracle} \lesssim \min_j \Big\{ \quad \textbf{precision}_j \quad + \textbf{local approx. error}_j + \textbf{abstention cost}_j \Big\}$$

Loss above the oracle

$$\mathbb{E}\big[L_c(\hat{\phi}_{\hat{j}}, \hat{\pi}_{\hat{j}}) - L_c^\star\big]$$

Estimation error rising with complexity

$$\sim \sqrt{\frac{C(\mathcal{F}_j) + v_{\Pi_j}}{|\mathcal{X}|}}$$

Approximation error within the prediction set

$$L_{|1}(f_j^\star, \pi_j^\star) - L_{|1}(f^\star, \pi_j^\star)$$

$$\mathbb{E}_x\big[c(x) \mid \pi(x) = 0\big]$$

$$L_{|1}(f, \pi) = \mathbb{E}_x\big[(f(x) - \tau(x))^2 \mid \pi(x) = 1\big].$$

Content

1. Setup and estimator
2. Properties
3. Practical implementation and illustration
4. Empirical application

Optimization with trees

- consider trees with depth D ($G \leq 2^D$), p covariates, and n observations

Prop. Computational complexity is $\mathcal{O}(n^D p^D)$

Optimization with trees

- consider trees with depth D ($G \leq 2^D$), p covariates, and n observations

Prop. Computational complexity is $\mathcal{O}(n^D p^D)$

Intuition

- For $D = 1$:
  1. run over all covariates and splits;
  2. ignorance is at each leaf;
  3. the loss adds across leaves.

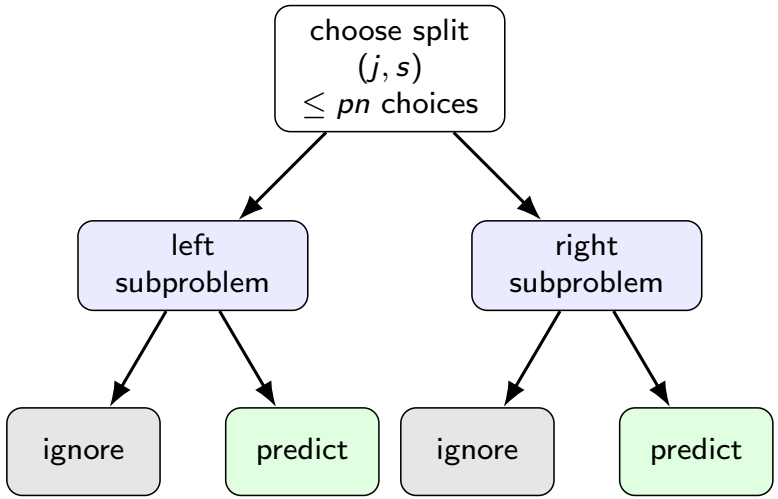
Optimization with trees

- consider trees with depth D ($G \leq 2^D$), p covariates, and n observations

Prop. Computational complexity is $\mathcal{O}(n^D p^D)$

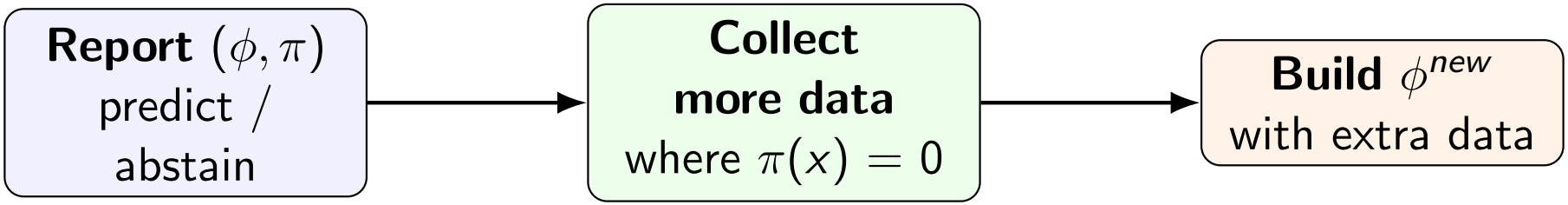
Intuition

- For $D = 1$:
  1. run over all covariates and splits;
  2. ignorance is at each leaf;
  3. the loss adds across leaves.
- For $D \geq 1$:
  1. run over all covariates and splits;
  2. solve the two child subproblems;
  3. add the two subproblem losses.
- (If $G < 2^D$, perform a final greedy step)
- For arbitrary clustering we propose a smoothed gradient descent

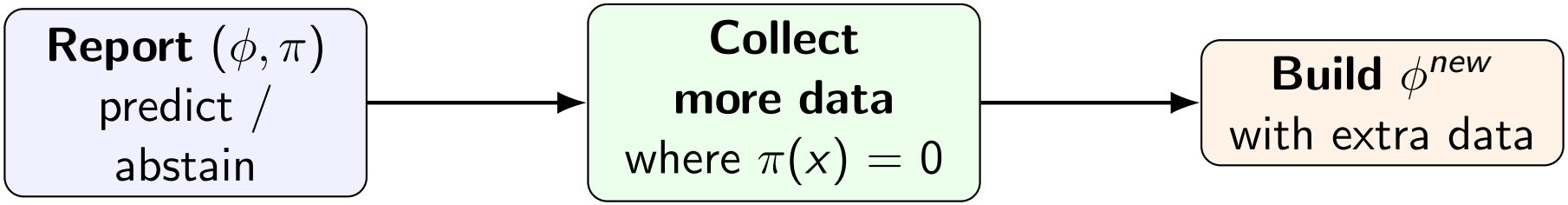


$\text{Loss} = \text{Loss}_L + \text{Loss}_R$

Interpretation of the cost of ignorance



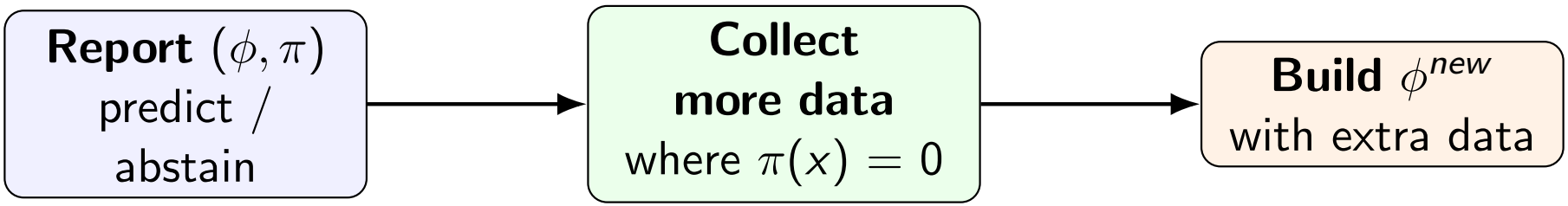
Interpretation of the cost of ignorance



If researchers admit ignorance

- Cost = expected prediction error after we defer and collect more data.

Interpretation of the cost of ignorance



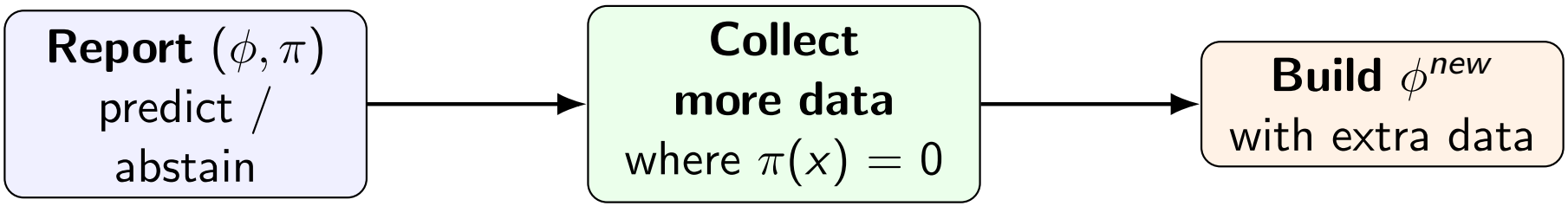
If researchers admit ignorance

- Cost = expected prediction error after we defer and collect more data.

Calibration

- Start from variance $\eta^2(x)$.
- m: extra sample size.
- Rule of thumb: $c(x) \approx \frac{\eta^2(x)}{m+1}$ i.e., precision of the mean with addition m data (e.g., $m = 1$)

Interpretation of the cost of ignorance



If researchers admit ignorance

- Cost = expected prediction error after we defer and collect more data.

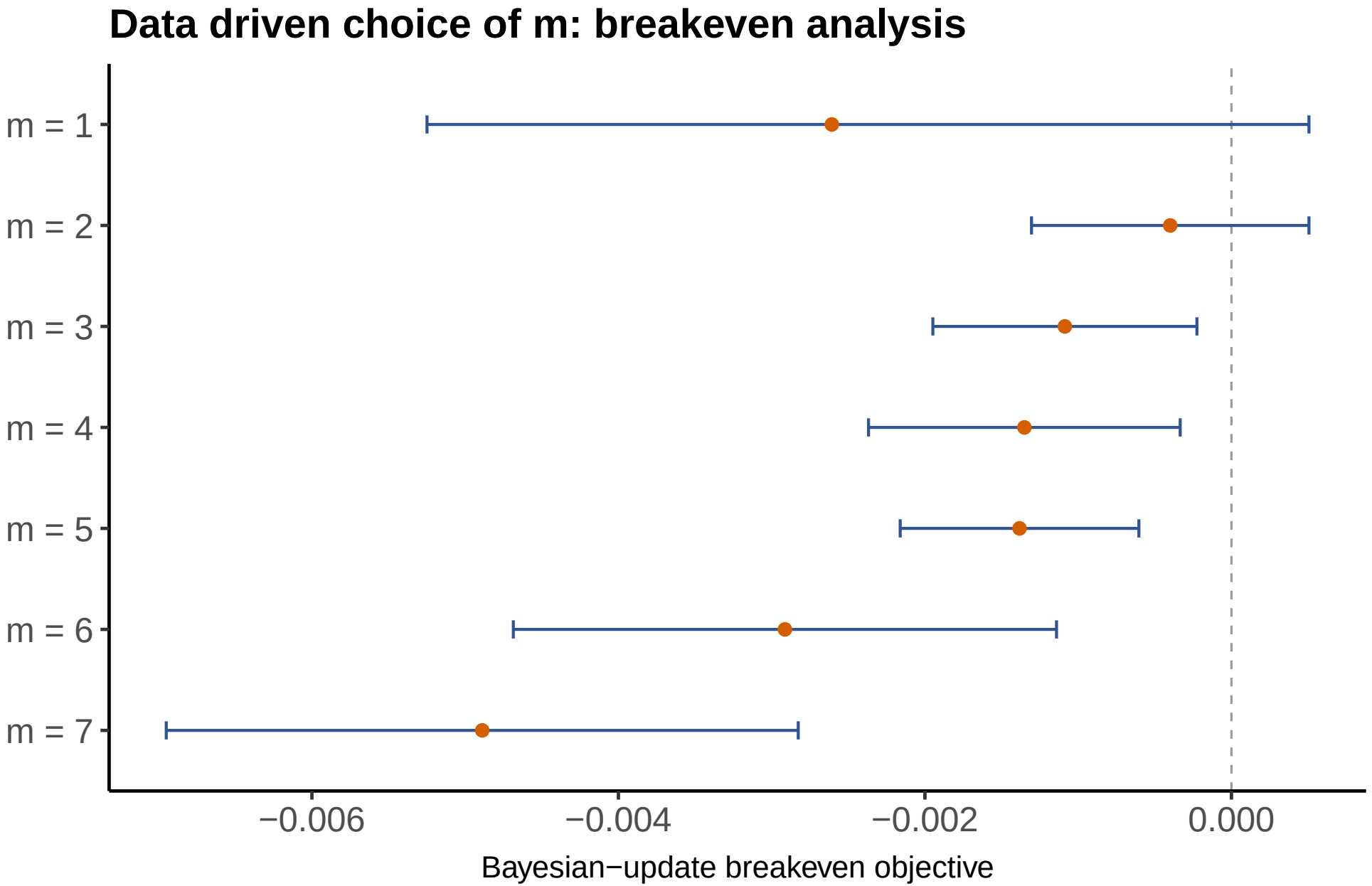
Calibration

- Start from variance $\eta^2(x)$.
- m: extra sample size.
- Rule of thumb: $c(x) \approx \frac{\eta^2(x)}{m+1}$ i.e., precision of the mean with addition m data (e.g., $m = 1$)

- In the paper data-driven choice of m: inference on break-even m so that abstention improves loss upon no abstention with high-prob

Example from multifaceted anti-poverty program

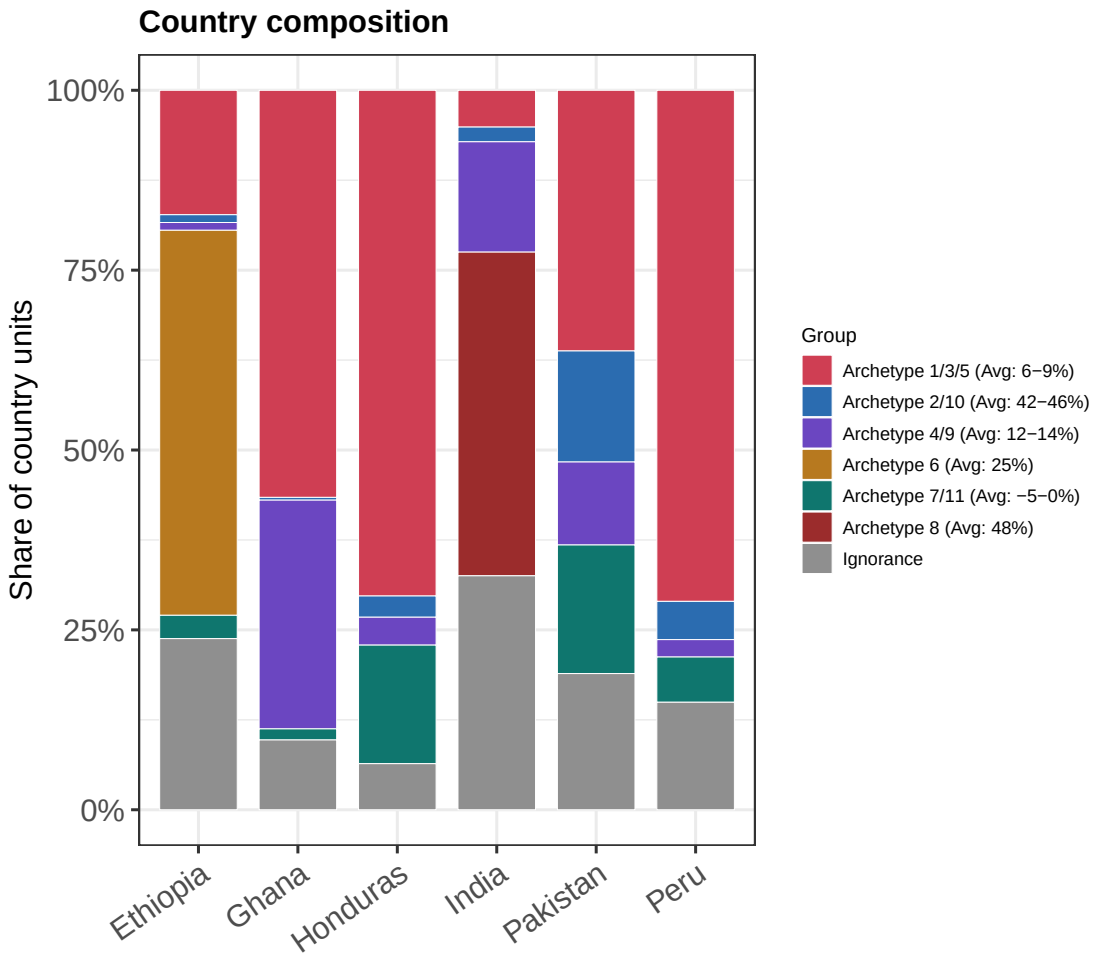
Choosing m

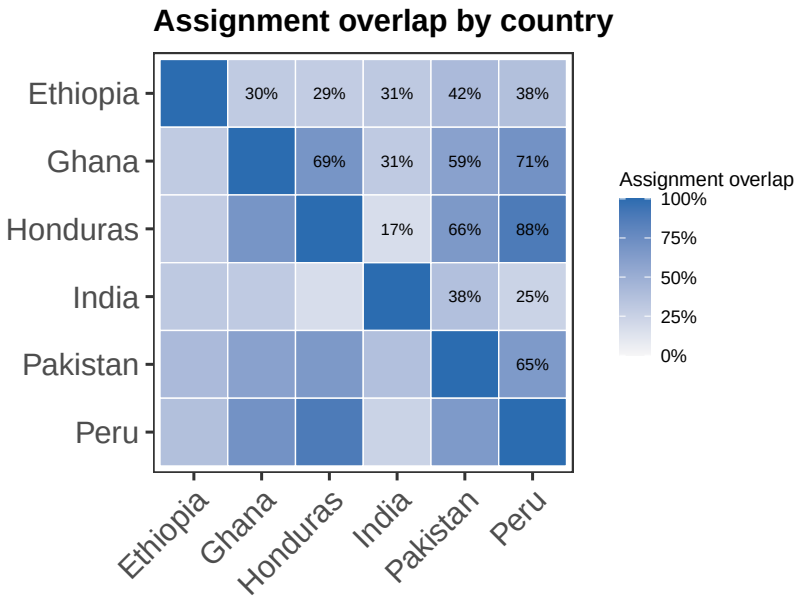


Content

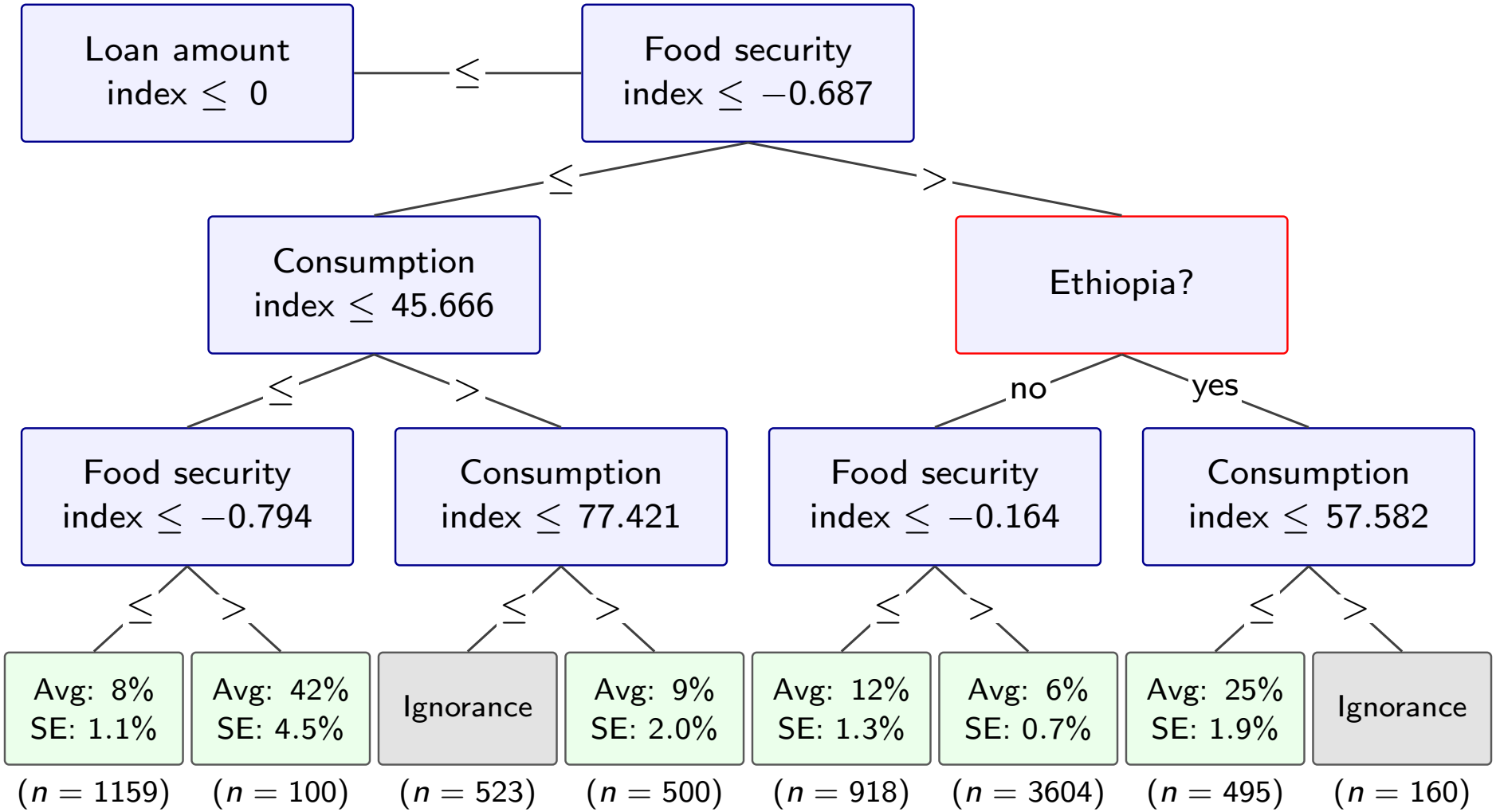
1. Setup and estimator
2. Properties
3. Practical implementation and illustration
4. Empirical application

Composition of archetypes by country

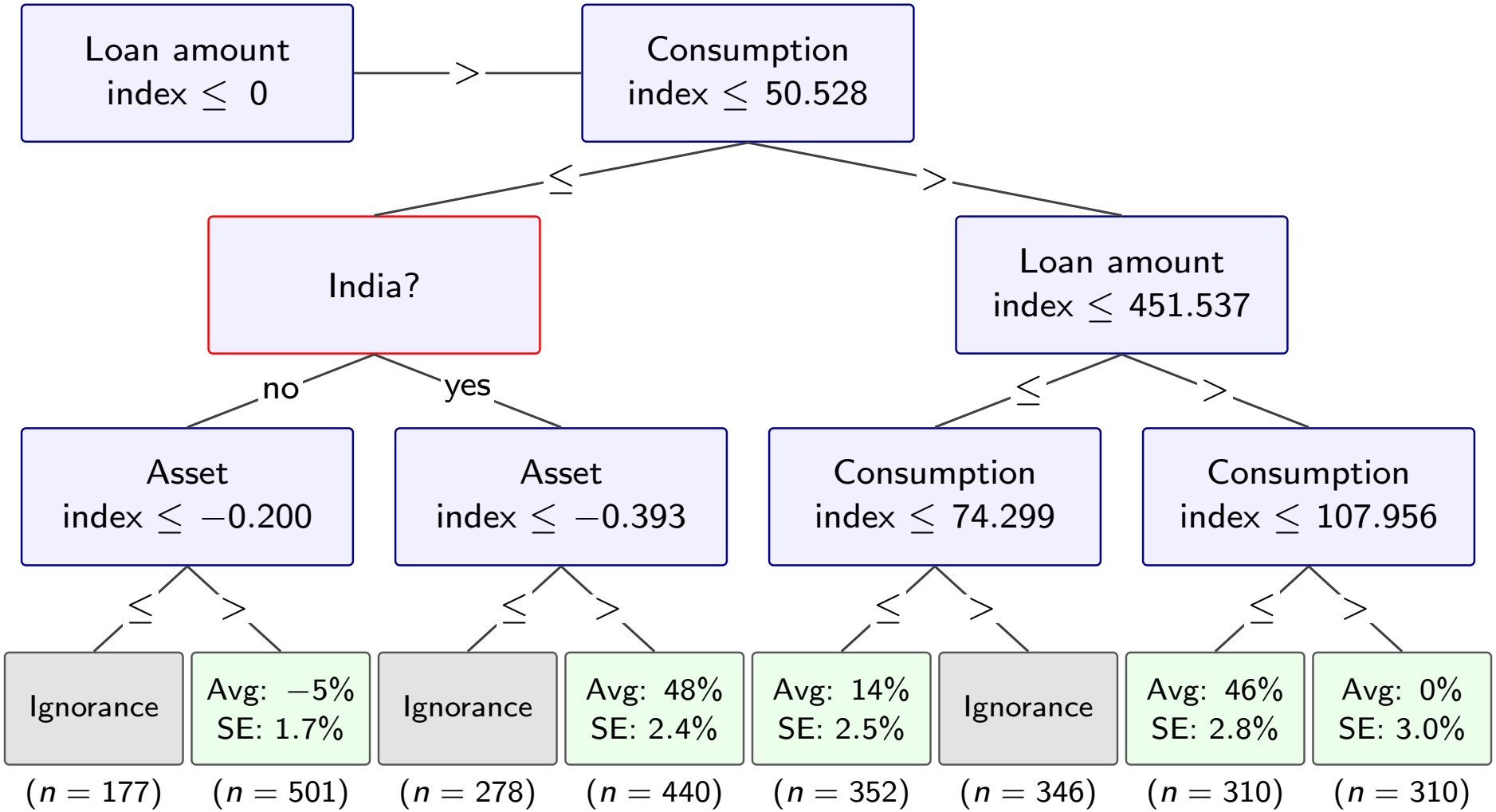




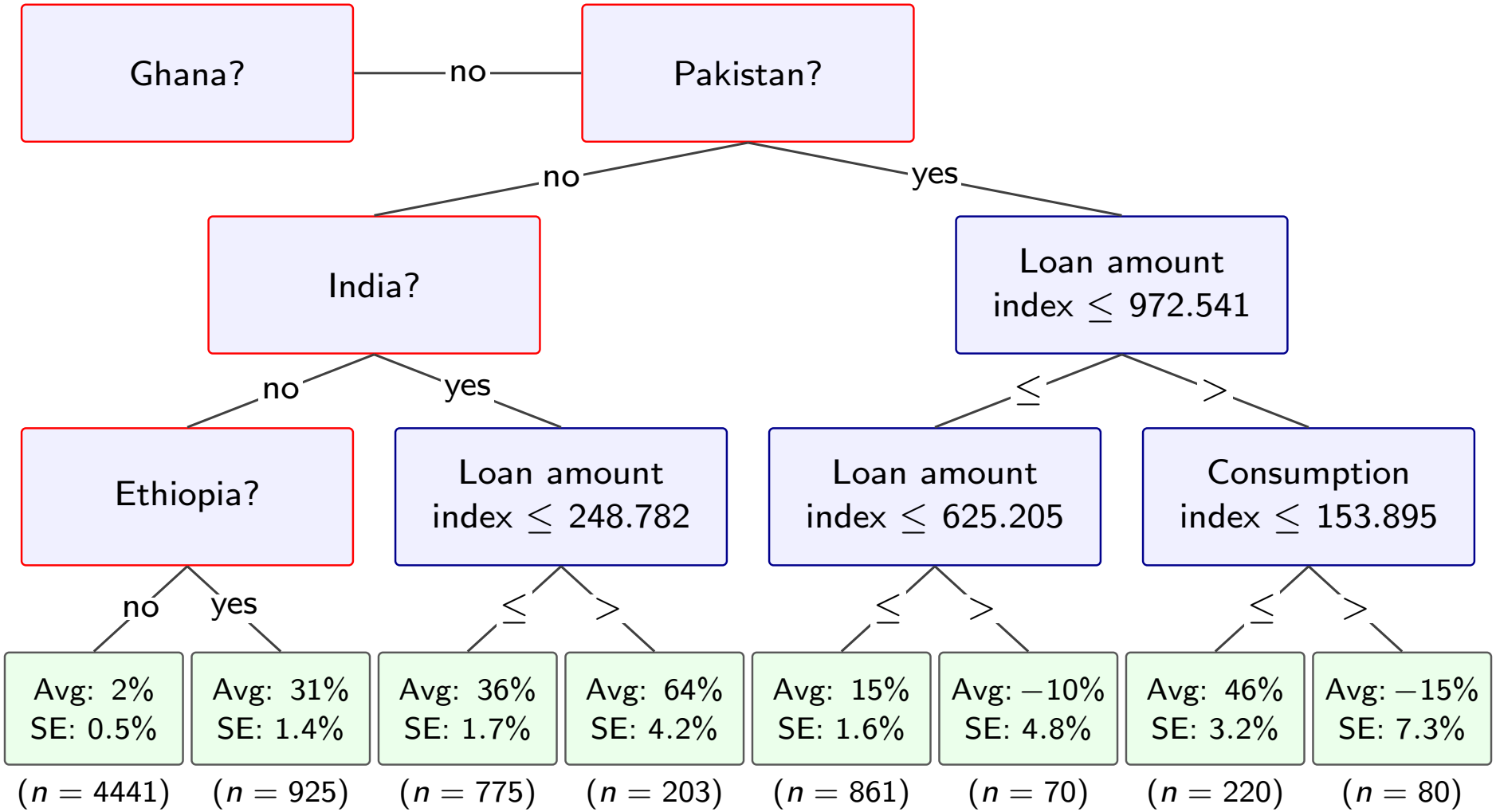
Ignorance-aware tree: loan amount ≤ 0



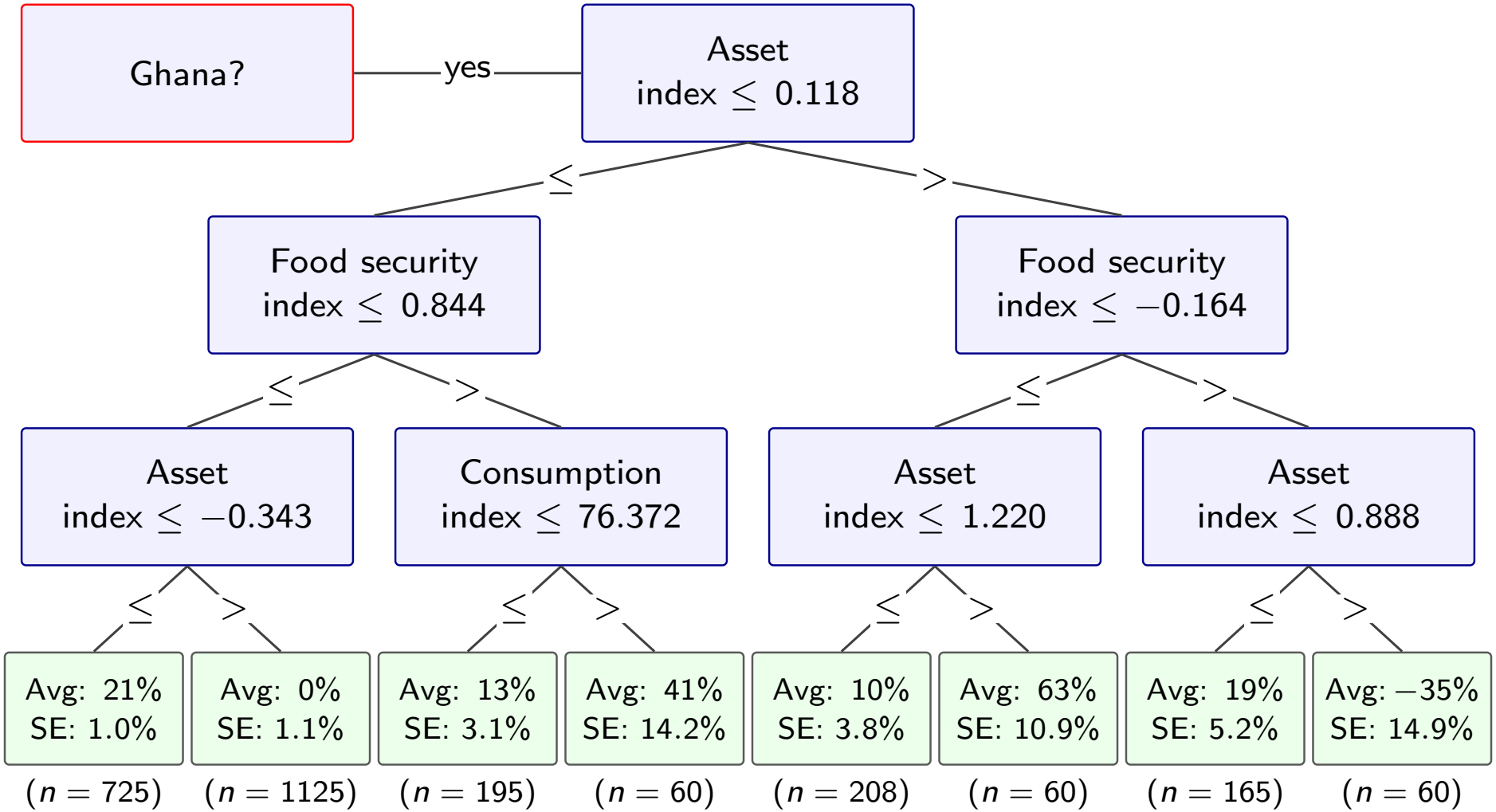
Ignorance-aware tree: loan amount > 0



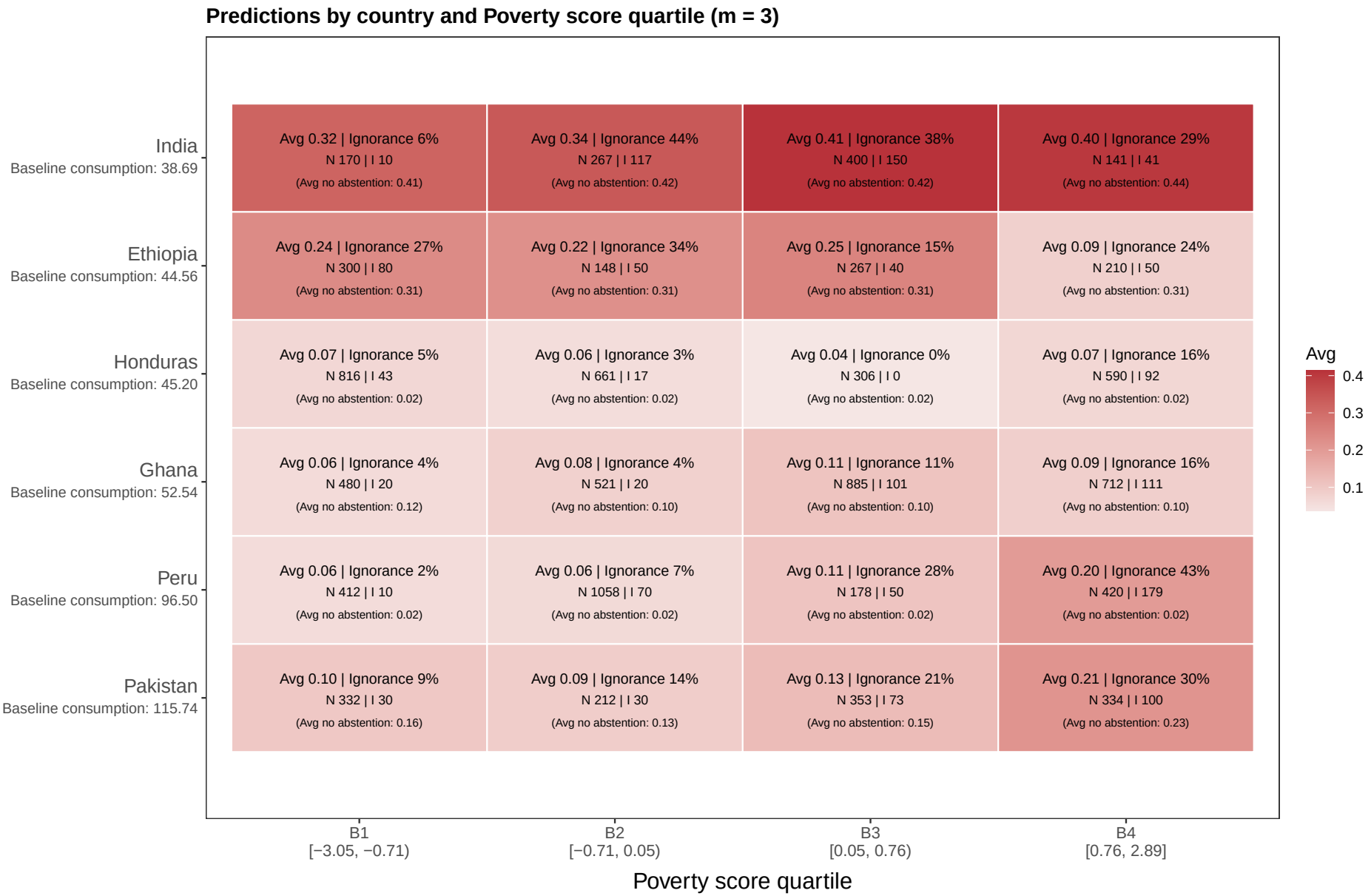
Tree without ignorance: outside Ghana



Tree without ignorance: Ghana



Predictions by country and poverty score

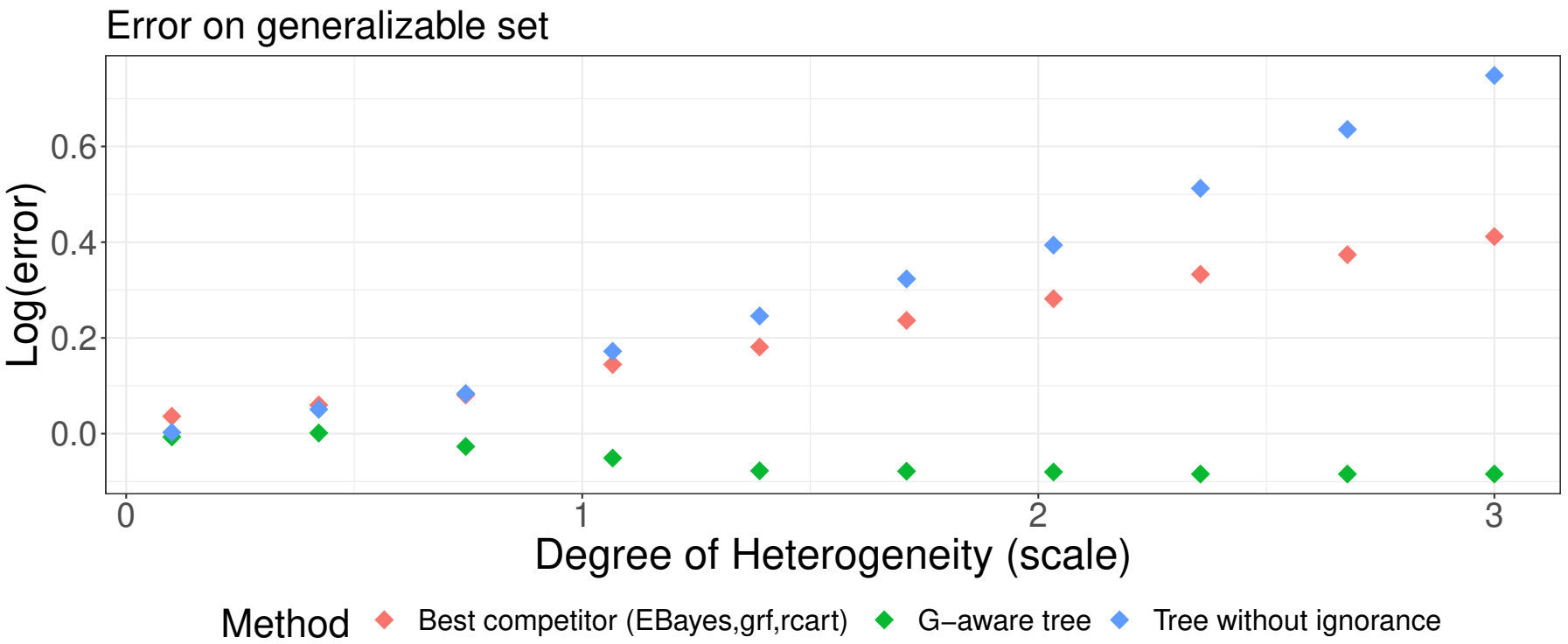


Calibrated simulations

- Calibrate simulations to estimated DGP from empirical application
- For 4% of observations in ignorance, generate treatment effects from Cauchy
- Compute error conditional on treatment effects outside this 4% units

Calibrated simulations

- Calibrate simulations to estimated DGP from empirical application
- For 4% of observations in ignorance, generate treatment effects from Cauchy
- Compute error conditional on treatment effects outside this 4% units



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

- We propose abstaining from predictions at cost of experimentation
- Estimator search for when evidence must be aggregated
- Key output is robust regression + where more data collection needed

$\Rightarrow$ Part of agenda betw/ econometrics and development

References I

Christopher Adjaho and Timothy Christensen. Externally valid treatment choice. *arXiv preprint arXiv:2205.05561*, 1, 2022.

Isaiah Andrews, Drew Fudenberg, Lihua Lei, Annie Liang, and Chaofeng Wu. The transfer performance of economic models. *arXiv preprint arXiv:2202.04796*, 2022.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184, 2015.

Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis*. John Wiley & Sons, 2021.

Tamara Broderick, Ryan Giordano, and Rachael Meager. An automatic finite-sample robustness metric: when can dropping a little data make a big difference? *arXiv preprint arXiv:2011.14999*, 2020.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

References II

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International conference on algorithmic learning theory*, pages 67–82. Springer, 2016.

Tommaso Crosta, Dean Karlan, Finley Ong, Julius Rüschenpöhler, and Christopher R Udry. Unconditional cash transfers: A bayesian meta-analysis of randomized evaluations in low and middle income countries. Technical report, National Bureau of Economic Research, 2024.

Antoine Deeb and Clément de Chaisemartin. Clustering and external validity in randomized controlled trials. *arXiv preprint arXiv:1912.01052*, 2019.

Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.

Michael Gechter, Keisuke Hirano, Jean Lee, Mahreen Mahmud, Orville Mondal, Jonathan Morduch, Saravana Ravindran, and Abu S Shonchoy. Selecting experimental sites for external validity. *arXiv preprint arXiv:2405.13241*, 2024.

Peter J Huber and Elvezio M Ronchetti. *Robust statistics*. John Wiley & Sons, 2011.

Takuya Ishihara and Toru Kitagawa. Evidence aggregation for treatment choice. *arXiv preprint arXiv:2108.06473*, 2021.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Charles F Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72 (4):1221–1246, 2004.

References III

Rachael Meager. Aggregating distributional treatment effects: A bayesian hierarchical analysis of the microcredit literature. *American Economic Review*, 112(6):1818–1847, 2022.

Konrad Menzel. Transfer estimates for causal effects across heterogeneous sites. *arXiv preprint arXiv:2305.01435*, 2023.

José Luis Montiel Olea, Brenda Prallon, Chen Qiu, Jörg Stoye, and Yiwei Sun. Externally valid selection of experimental sites via the k-median problem. *arXiv preprint arXiv:2408.09187*, 2024.

Davide Viviano. Policy targeting under network interference. *Review of Economic Studies*, page rdae041, 2024.