Privacy-preserving Meta-Analysis through Low-Rank Basis Hunting

Kosuke Imai

Harvard University

Applied Statistics Seminar
Indian Statistical Institute, Kolkata
March 24, 2026

Joint work with Wenqi Shi (Harvard) and Yi Zhang (Netflix)

Motivation

- Meta analysis: statistical analysis to combine the results of multiple independent studies
  - widely used in social and medical sciences for evidence synthesis
  - generalizing results from *sources* to a *target* population
  - common parametric approaches: random effect models

- Key challenges:
  1. unknown heterogeneity across sources and between source and target populations
     - covariate shift $\mathcal{P}(\boldsymbol{X})$
     - conditional shift $\mathcal{P}(Y \mid \boldsymbol{X})$
  2. function-valued quantities of interest, going beyond vector-valued parameters
     - regression functions $\mathbb{E}[Y \mid \boldsymbol{X}] \rightsquigarrow$ means $\mathbb{E}[Y]$
     - conditional average treatment effect (CATE) $\mathbb{E}[Y(1) - Y(0) \mid \boldsymbol{X}] \rightsquigarrow$ ATE $\mathbb{E}[Y(1) - Y(0)]$
  3. Privacy preservation: limited direct access to source data
  4. Use of machine learning (ML) models: flexible nonparametric estimation for each source

Overview of Our Contributions

- MetaHunt: privacy-preserving functional meta-analysis
  - estimates function-valued quantities for a *new target* population
  - requires only aggregate information from sources
  - allows for the use of (possibly different and unknown) ML models in each source
  - provides asymptotically valid (pointwise) statistical inference

- Key idea:
  - source and target populations share a common low-rank structure
  - all study-level functions lie in the convex hull of a small number of latent basis functions

- functional Successive Projection Algorithm (fSPA) to recover latent basis functions
- Conformal inference to provide a confidence interval with marginal coverage control

Low-rank Cross-Study Heterogeneity

- Study-specific functions of interest $f^{(i)}(\boldsymbol{x})$ may differ in complex and unknown ways
- But, we assume that they share a common underlying structure
- All study-specific functions lie within the convex hull of a set of basis functions $\{g_k(\boldsymbol{x})\}_{k=1}^K$
- Enable dimension reduction, nonparametric modeling, and generalization

Assumption 1 (Low-rank cross-study heterogeneity)

There exists a set of basis functions $\{g_k(\boldsymbol{x})\}_{k=1}^K$ *(*$K < m$*) such that for all* $i = 0, 1, \ldots, m$

$$f^{(i)}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_{ik} g_k(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathcal{X},$$

where $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})^\top \in \Delta_{K-1} = \{\boldsymbol{\pi} \in \mathbb{R}^K : \sum_{k=1}^K \pi_k = 1, \pi_k \geq 0\}$

Weight Model

- The weights $\boldsymbol{\pi}_i$ determine how study i combines the shared basis functions $\{g_k\}_{k=1}^K$
- Use study-level information $\boldsymbol{W}_i$ to predict variation in $\boldsymbol{\pi}_i$
- Studies with similar $\boldsymbol{W}_i$ will have similar mixing of the basis functions

Assumption 2 (Weight model)

For all $i \in \{0, 1, \ldots, m\}$, the weight vector $\boldsymbol{\pi}_i$ is drawn independently from a conditional distribution given $\boldsymbol{W}_i \in \mathcal{W}$

$$\boldsymbol{\pi}_i \mid \boldsymbol{W}_i \overset{ind.}{\sim} \mathcal{P}_{\boldsymbol{\pi}|\boldsymbol{W}}(\cdot \mid \boldsymbol{W}_i),$$

where $\mathcal{P}_{\boldsymbol{\pi}|\boldsymbol{W}}$ is an arbitrary distributional map from $\mathcal{W}$ to the simplex Δ_{K-1}

Weight Model Example: Dirichlet Regression

Example 1 (Kernelized Dirichlet regression model)

Let $\kappa : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ be a positive-definite kernel with the associated reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\kappa$ and feature map $\psi : \mathcal{W} \to \mathcal{H}_\kappa$. For all $i \in \{0, 1, \ldots, m\}$, assume

$$\boldsymbol{\pi}_i := (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK}) \overset{ind.}{\sim} \text{Dirichlet}(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK}),$$

with $\alpha_{ik} = \exp\left\{\langle \boldsymbol{\beta}_k, \psi(\boldsymbol{w}_i)\rangle_{\mathcal{H}_\kappa}\right\}$ and $\boldsymbol{\beta}_k \in \mathcal{H}_\kappa$

- Alternative modeling approaches:
  - Log-ratio regression
  - Neural networks with softmax

Exchangeability of Study-level Covariates

Assumption 3 (Study-level covariate exchangeability)

The study-level covariates $\boldsymbol{W}_0, \boldsymbol{W}_1, \ldots, \boldsymbol{W}_m$ are exchangeable

- Exchangeability required for conformal prediction
- No exchangeability assumption at the individual level
- Under Assumptions 1–3, $(\boldsymbol{W}_i, \boldsymbol{\pi}_i, f^{(i)})$ are jointly exchangeable across $i = 0, 1, \ldots, m$

Overview of MetaHunt

Estimation stage

Input: $\{\boldsymbol{W}_i, \hat{f}^{(i)}(\boldsymbol{x})\}_{i=1}^m$

Output: $\{\hat{g}_k\}_{k=1}^K$, $\widehat{\mathcal{M}} : \mathcal{W} \to \Delta_{K-1}$

1. Basis hunting: obtain $\{\hat{g}_k\}_{k=1}^K$
2. Estimate weights $\hat{\boldsymbol{\pi}}_i$ by projection
3. Fit the weight model $\widehat{\mathcal{M}}$

Prediction stage

Input: $\{\hat{g}_k\}_{k=1}^K$, $\widehat{\mathcal{M}}$, $\boldsymbol{W}_0$

Output: $\tilde{f}^{(0)}(\boldsymbol{x})$ and a confidence interval

1. Predict the weights: $\tilde{\boldsymbol{\pi}}_0 = \widehat{\mathcal{M}}(\boldsymbol{W}_0)$
2. Predict the target function:

$$\tilde{f}^{(0)}(\boldsymbol{x}) = \sum_{k=1}^{K} \tilde{\boldsymbol{\pi}}_{0k}\, \hat{g}_k(\boldsymbol{x})$$

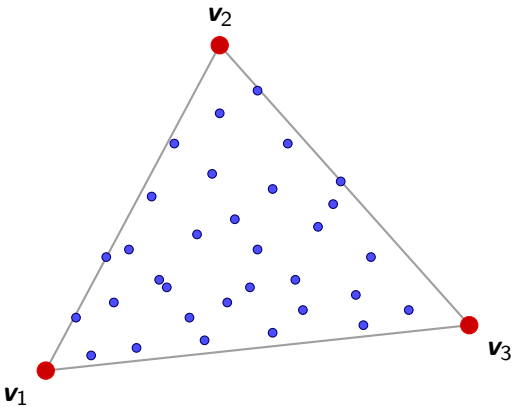
3. Construct a conformal prediction interval

Basis Hunting vs. Vertex Hunting

- Vertex hunting:

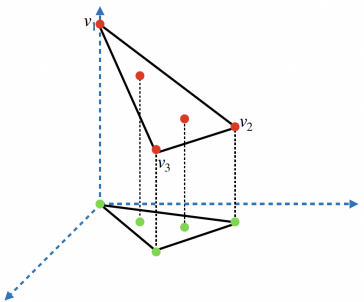
  Find the vertices $\boldsymbol{v}_k$ such that for all $\boldsymbol{x}_i$ we have

$$\boldsymbol{x}_i = \sum_{k=1}^{K} \pi_{ik} \boldsymbol{v}_k, \quad \sum_{k=1}^{K} \pi_{ik} = 1, \ \pi_{ik} \geq 0$$



- Basis hunting:
  Recover the latent basis functions $\{g_k\}_{k=1}^{K}$ from the observed functions $\{\hat{f}^{(i)}\}_{i=1}^{m}$
- Random functions vs random vectors

Successive Projection Algorithm (SPA; Arajúno et al. 2001) for Vertex Hunting



1. Select the data point with the largest length as the first vertex $\boldsymbol{v}_1$
2. Project all data points onto the space orthogonal to the selected vertex
3. Among the projected points, choose the one with the largest length as the next vertex $\boldsymbol{v}_2$
4. Repeat the projection and selection steps until K vertices are chosen

- Denoising step (Jin et al. 2024): For each (projected) point,
  1. If there are fewer than N points in the neighborhood of $\boldsymbol{x}_i$, remove $\boldsymbol{x}_i$
  2. Otherwise, replace $\boldsymbol{x}_i$ with the average of N neighborhood points

d-fSPA: Denoised Basis Hunting Algorithm

Algorithm 1: d-fSPA

Input: Estimated functions $\{\hat{f}^{(i)}(\cdot)\}_{i=1}^{m}$; number of basis K; tuning parameters (N, Δ)

Denoising step:

1. If there are fewer than N functions in $B_\Delta(\hat{f}^{(i)})$, remove this function
2. Otherwise, replace $\hat{f}^{(i)}$ by the average of all functions in $B_\Delta(\hat{f}^{(i)})$

where $B_\Delta(\hat{f}^{(i)}) = \{f : \|f - \hat{f}^{(i)}\| \leq \Delta\}$ denotes the Δ-neighborhood of $\hat{f}^{(i)}$

Initialize $S = \emptyset$, $h_i(\cdot) = \hat{f}^{(i)}(\cdot)$ for $1 \leq i \leq m$

for $k = 1, \ldots, K$ **do**

1. Project $h_i = \hat{f}^{(i)}$ onto the orthogonal space of $\mathrm{span}(S)$: $h_i = h_i - P_{\mathrm{span}(S)} h_i$
2. Find s_k such that h_{s_k} has the largest norm $s_k = \mathrm{argmax}_{1 \leq i \leq m} \|h_i\|$
3. Update the set $S = S \cup \{\hat{f}^{(s_k)}\}$

Output: Estimated basis functions $\hat{g}_k(\cdot) = \hat{f}^{(s_k)}$ for $1 \leq k \leq K$

d-fSPA Implementation Details

- Functional norm:
  - L^2 norm under the target covariate distribution $\mathcal{P}_{0,\boldsymbol{X}}$ (use sampled covariates for evaluation)

$$\|f - \tilde{f}\| = \left\{ \mathbb{E}_{\mathcal{P}_{0,\boldsymbol{X}}} \left[(f(\boldsymbol{X}) - \tilde{f}(\boldsymbol{X}))^2 \right] \right\}^{1/2}$$

  - If the target distribution is unavailable, use proxies (e.g., pooled source covariate distribution)

- Denoising parameters: heuristic choice (Jin et al. 2024)
  $N = 0.5 \log m$ and $\Delta = \max_{ij} \|\hat{f}^{(i)} - \hat{f}^{(j)}\|/10$

- Study-specific functions $\hat{f}^{(i)}$ may be estimated using (possibly different and unknown) black-box ML models

Selecting the Number of Basis Functions K

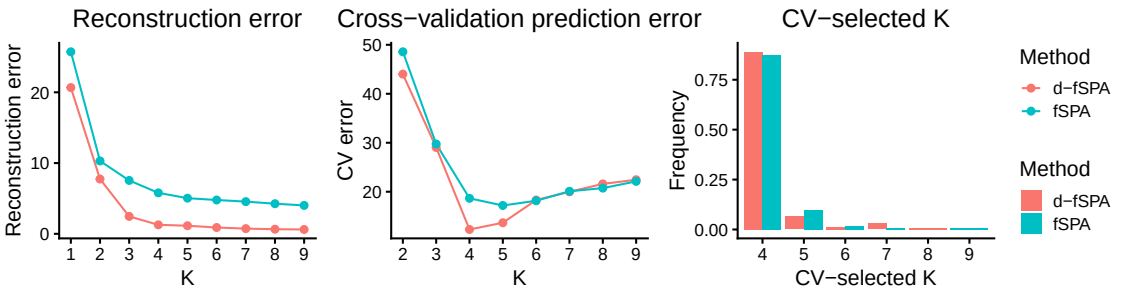
1. Elbow plot:
   - Plot K against the reconstruction error

$$\mathcal{E}(K) := \frac{1}{m} \sum_{i=1}^{m} \min_{\boldsymbol{\pi}_i} \left\| \hat{f}^{(i)} - \sum_{k=1}^{K} \pi_{ik} \, \hat{g}_k \right\|$$

   - The "elbow" point indicates where adding more bases yields small improvement
   - Requires only the estimated functions $\{\hat{f}^{(i)}(\boldsymbol{x})\}$

2. Cross-validation:
   - Select K by minimizing cross-validation prediction error
   - Requires both the estimation and prediction stages
   - Requires site-level covariates $\boldsymbol{W}_i$

Selection of K: Simulation Illustration



Exact Recovery of Basis Functions by fSPA under No Estimation Error

Proposition 1 (Exact recovery by fSPA under no estimation error)

Suppose the basis functions are non-degenerate, Assumption 1 holds, and the following conditions are satisfied:

1. *No estimation error:* $\|f^{(i)} - \hat{f}^{(i)}\| = 0$ *for all* $i = 1, \ldots, m$
2. *Pure node condition: For each* $k = 1, \ldots, K$, *there exists a study* i *such that* $\pi_{ik} = 1$

Then, up to permutation of $\{1, \ldots, K\}$,

$$\max_{1 \le k \le K} \|\hat{g}_k - g_k\| = 0$$

Dealing with Estimation Error and Degree of Purity

- Estimation error:

$$\epsilon^{(i)}(\boldsymbol{x}) := \hat{f}^{(i)}(\boldsymbol{x}) - f^{(i)}(\boldsymbol{x})$$

- Degree of purity:

$$\delta_m := \max_{1 \le k \le K} \min_{1 \le i \le m} (1 - \pi_{ik})$$

Assumption 4 (Error control)

1. $\mathbb{E}[\epsilon^{(i)}(\boldsymbol{x})^2] = O(n_i^{-r})$ *for some* $r > 0$, *for all* i *and* $\boldsymbol{x}$
2. *There exists* $0 < a < r$ *such that* $m = o(\inf_i n_i^a)$
3. $\delta_m \max_{1 \le k \le K} \|g_k\| = o(1)$

1. Parametric models $r = 1$; ML models $r < 1$
2. The number of studies m remains moderate relative to study sample sizes
3. As m increases, we expect some studies to approach pure nodes $\delta_m \to 0$

Asymptotic Recovery of Basis Functions by d-SPA

Theorem 2 (Asymptotic recovery by d-fSPA)

Under Assumptions 1 and 4, if the basis functions are non-degenerate, then, with an appropriate choice of the denoising parameters (N, Δ), we have, up to permutation,

$$\max_{1 \le k \le K} \|\hat{g}_k - g_k\| = o_P(1)$$

- Exact conditions for the denoising parameters are currently being finalized...

Fitting the Weight Model

1. Estimate study-level weights:

$$\hat{\boldsymbol{\pi}}_i = \underset{\boldsymbol{\pi}_i \in \Delta_{K-1}}{\operatorname{argmin}} \left\| \hat{f}^{(i)} - \sum_{k=1}^{K} \pi_{ik}\, \hat{g}_k \right\|$$

2. Fit the weight model: Use $\{(\boldsymbol{W}_i, \hat{\boldsymbol{\pi}}_i)\}_{i=1}^{m}$ to estimate

$$\widehat{\mathcal{M}} : \mathcal{W} \to \Delta_{K-1}$$

Predicting for a Target Site

1. Predict mixing weights:

$$\tilde{\boldsymbol{\pi}}_0 = \widehat{\mathcal{M}}(\boldsymbol{W}_0).$$

2. Predict the target function:

$$\tilde{f}^{(0)}(\boldsymbol{x}) = \sum_{k=1}^{K} \tilde{\pi}_{0k}\, \hat{g}_k(\boldsymbol{x}).$$

Challenges in Constructing Prediction Interval

1. Our target $f^{(0)}$ is a function
   - Building uniform confidence bands is hard, i.e., $\Pr\left(l(\boldsymbol{x}) \leq f^{(0)}(\boldsymbol{x}) \leq u(\boldsymbol{x})\right) \geq 1-\alpha$ for all $\boldsymbol{x}$
   - We focus on a point-wise interval for $f^{(0)}(\boldsymbol{x})$ given $\boldsymbol{x}$

2. Multiple sources of error
   - Possibly correlated errors in basis hunting, weight estimation, weight model
   - We focus on conformal prediction with valid marginal coverage

3. Estimation error in $\hat{f}^{(i)}$
   - We aim to predict $f^{(0)}(\boldsymbol{x})$ but only observe $\hat{f}^{(i)}(\boldsymbol{x})$ as a proxy
   - Assumptions are needed to control the estimation error $|\hat{f}^{(i)}(\boldsymbol{x}) - f^{(i)}(\boldsymbol{x})|$

Conformal Prediction Interval

Algorithm 2: Split conformal prediction for $f^{(0)}(\boldsymbol{x})$

Input: Estimated functions $\{\hat{f}^{(i)}\}_{i=1}^{m}$; target-study covariates $\boldsymbol{W}_0$; evaluation point $\boldsymbol{x}$; miscoverage level $\alpha \in (0,1)$; estimation pipeline

1. **Data splitting:** randomly split the index set $\{1,\ldots,m\}$ into $\mathcal{I}_{\text{tr}}$ and $\mathcal{I}_{\text{cal}}$
2. **Training:** using only training data, run the estimation pipeline to obtain a prediction rule
3. **Calibration:** for each $i \in \mathcal{I}_{\text{cal}}$, calculate the conformity score $r_i := \left|\hat{f}^{(i)}(\boldsymbol{x}) - \tilde{f}^{(i)}(\boldsymbol{x})\right|$
4. **Conformal quantile:** set $q_{1-\alpha}(\boldsymbol{x}) := r_{(\lceil (1-\alpha)(m_{\text{cal}}+1) \rceil)}$
5. **Prediction:** compute the point prediction for the new study, $\tilde{f}^{(0)}(\boldsymbol{x})$

Output: Split conformal prediction interval

$$C_\alpha(\boldsymbol{x}) = [\tilde{f}^{(0)}(\boldsymbol{x}) - q_{1-\alpha}(\boldsymbol{x}), \tilde{f}^{(0)}(\boldsymbol{x}) + q_{1-\alpha}(\boldsymbol{x})]$$

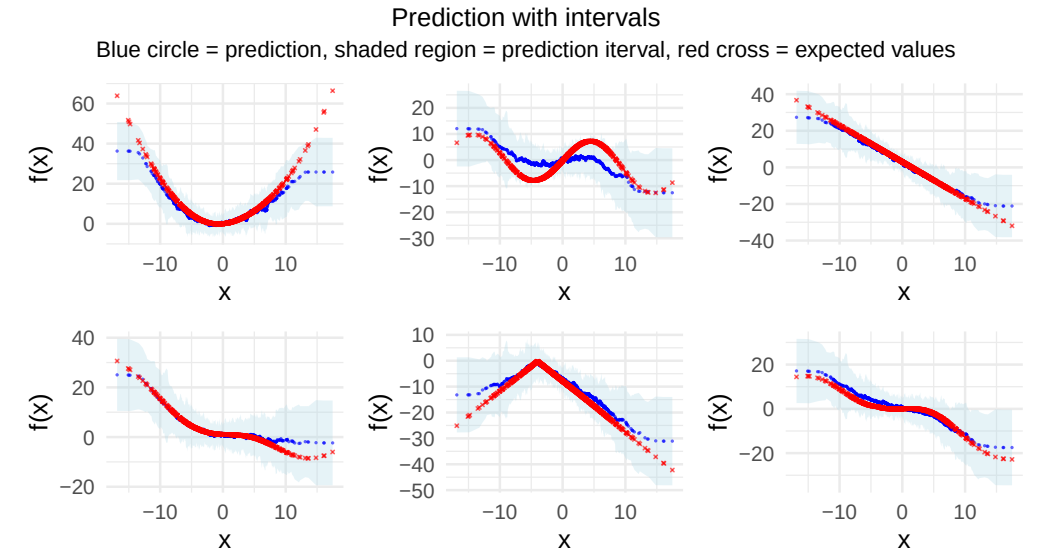
Asymptotic Marginal Coverage

Theorem 3 (Asymptotic marginal coverage guarantee)

Under Assumptions 1–4, we have asymptotic marginal coverage for any $\boldsymbol{x}$*,*

$$\lim \Pr(f^{(0)}(\boldsymbol{x}) \in C_\alpha(\boldsymbol{x})) \geq 1 - \alpha$$

Conformal Prediction: Simulation Illustration



Extension to a Functional of the Target Function

- $\mathcal{H} : \mathcal{F} \to \mathbb{R}$: a functional mapping a function to scaler, i.e., $C^{(0)} = \mathcal{H}\big(f^{(0)}\big)$
- Example: ATE $\tau := \mathbb{E}[Y(1) - Y(0)]$
  - CATE may be smoother than the conditional outcome function
  - shared low-rank structure of the CATE $f(\boldsymbol{x}) = \mathbb{E}[Y(1) - Y(0) \mid \boldsymbol{X} = \boldsymbol{x}]$

$$\tau^{(0)} = \mathcal{H}\big(f^{(0)}\big) = \mathbb{E}_{\mathcal{P}_{0,\boldsymbol{x}}}\Big[f^{(0)}(\boldsymbol{X})\Big]$$

  where $\mathcal{P}_{0,X}$ is the covariate distribution of the target population

- Plug-in point prediction:

$$\tilde{\tau}^{(0)} = \mathcal{H}\big(\tilde{f}^{(0)}\big)$$

- Conformal interval: For each calibration study i, define the conformity score

$$r_i = \Big|\mathcal{H}\big(\hat{f}^{(i)}\big) - \mathcal{H}\big(\tilde{f}^{(i)}\big)\Big|$$

Empirical Application: Many Labs 1 (Klein et al. 2014)

- Large-scale multisite replication project in psychology
- Evaluates the replicability of 13 classic experimental findings, including:
  - anchoring (Jacowitz & Kahneman, 1995)
    whether exposure to a high or low numerical anchor influences participants' estimates of the population of Chicago
  - gain vs. loss framing effects (Tversky & Kahneman, 1981)
    whether participants make different choices when outcomes are framed in terms of lives saved (gains) versus lives lost (losses)

- Experiments across 36 independent data collection sites ($m = 36$)
- Each hypothesis is tested under a common experimental protocol across studies
- Total sample size: approximately 6,000 participants per hypothesis on average

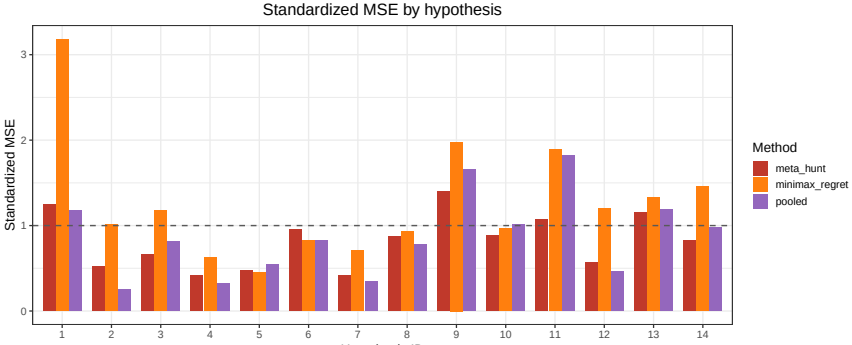
Setup

- Site-level covariates $\boldsymbol{W}$:
  - Online vs. laboratory setting
  - US vs. international site
  - Average age
  - Gender ratio
  - Average political ideology
  - Measured covariate shift relative to the target site (varies by target site and hypothesis)
- Individual-level covariates $\boldsymbol{X}$: gender, age, race, political ideology, American identity
- Prediction task: regression function estimated using random forest
- Causal task: site-level CATE functions estimated using Causal Forest
- Selection of K based on cross-validation for each hypothesis
- Leave-one-site-out prediction for validation:
  - for each hypothesis, treat one site as the target and use the remaining sites for training
  - repeat so that every site serves as the target
- Empirical benchmark: Target-site ATE estimated directly from the experiment

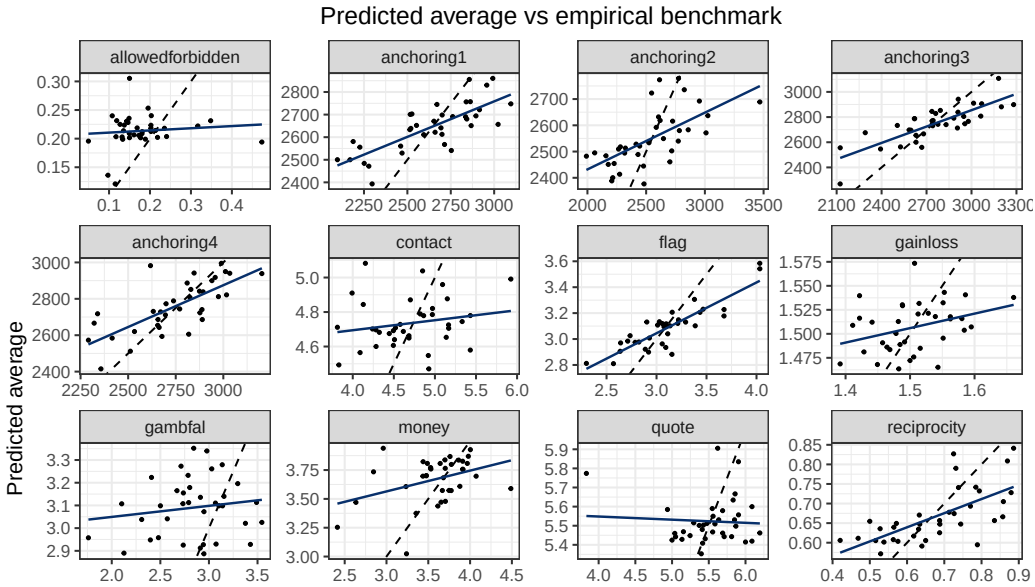
Prediction Task Results

Method	Std. MSE	Coverage	Interval Length (relative)
MetaHunt	0.817	0.988	0.969
Minimax regret (Zhang et al.)	1.266	0.982	0.964
Pooled ML	0.871	0.974	0.806

- Standardized relative to the empirical benchmark
- Pooled ML: ML with conformal inference, using both individual and site-level covariates



Empirical Performance by Hypothesis

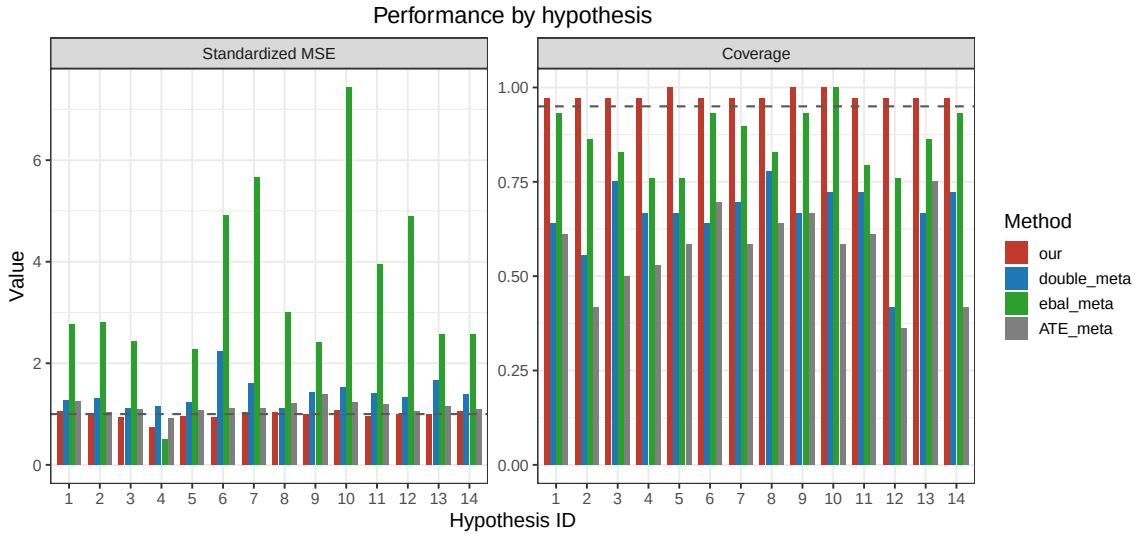


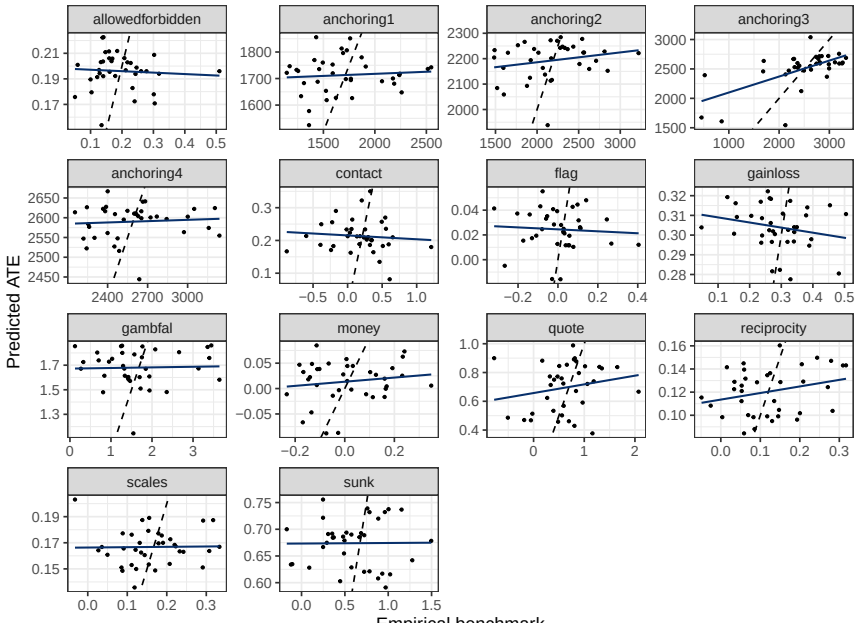
Causal Task Results

Method	Std. MSE	Coverage	Interval Length (relative)
MetaHunt	0.962	0.976	0.780
DR	1.305	0.700	0.405
ebal	2.532	0.855	0.965
DiffMeans	1.117	0.645	0.316
Pooled ML	0.977	0.958	0.714

- DR: doubly robust estimator at each source site with covariate density ratios
- ebal: entropy balancing estimator (reweighted toward the target) at each source site
- DiffMeans: average of difference-in-means ATE at each source site
- All followed by inverse-variance weighted meta-regression adjusting for site-level covariates

Empirical Performance by Hypotheses




Predicted ATE vs empirical benchmark
allowedforbidden
anchoring1
anchoring2
anchoring3
anchoring4
contact
flag
gainloss
gambfal
money
quote
reciprocity
scales
sunk
Predicted ATE
Empirical benchmark

Concluding Remarks

- How to generalize function-valued quantities across heterogeneous studies using only aggregate-level information?
- Cross-study heterogeneity can be captured by a low-rank structure
- Methodological contribution:
  - functional basis recovery via denoised functional SPA (d-fSPA)
  - flexible weight modeling linking study-level covariates to mixing proportions
  - distribution-free uncertainty quantification via conformal prediction
- Theoretical guarantees:
  - consistent recovery of latent basis functions
  - asymptotically valid marginal coverage for target predictions
- MetaHunt enables privacy-preserving, ML-compatible functional meta-analysis that accommodates both covariate shift and conditional shift