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

Kosuke Imai

Harvard University

Summer Meeting, Japanese Society for Quantitative Political Science
July 4, 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 individual-level 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

- Methodological contributions:
  - functional Successive Projection Algorithm (fSPA) to recover latent basis functions
  - Conformal inference to provide a confidence interval with marginal coverage control
  - Open-source software: MetaHunt `https://cran.r-project.org/package=MetaHunt`

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}

- Examples: Dirichlet regression, 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

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 $|\hat{f}^{(i)}(\boldsymbol{x}) - f^{(i)}(\boldsymbol{x})|$

Solution: conformal prediction interval

- marginal coverage guarantee
- extension to a functional of the target function: e.g., CATE $\rightsquigarrow$ ATE

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)
  - gain vs. loss framing effects (Tversky & Kahneman, 1981)

- 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

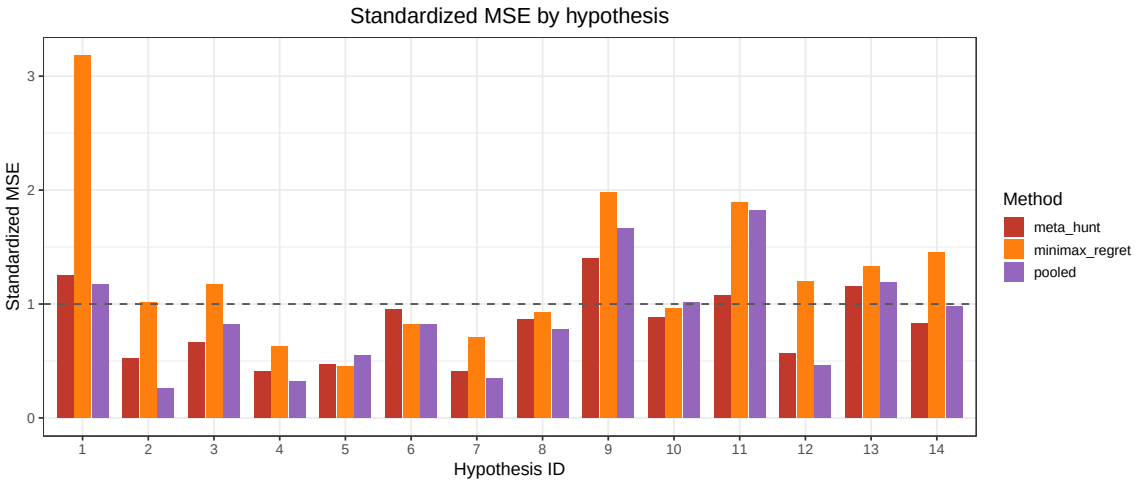
- Site-level covariates $\boldsymbol{W}$:
  - Online vs. laboratory setting
  - US vs. international site
  - Average age
  - Gender ratio
  - Average political ideology
  - Measured covariate shift relative to the target site (varies by target site and hypothesis)
- Individual-level covariates $\boldsymbol{X}$: 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

Standardized MSE by Hypotheses

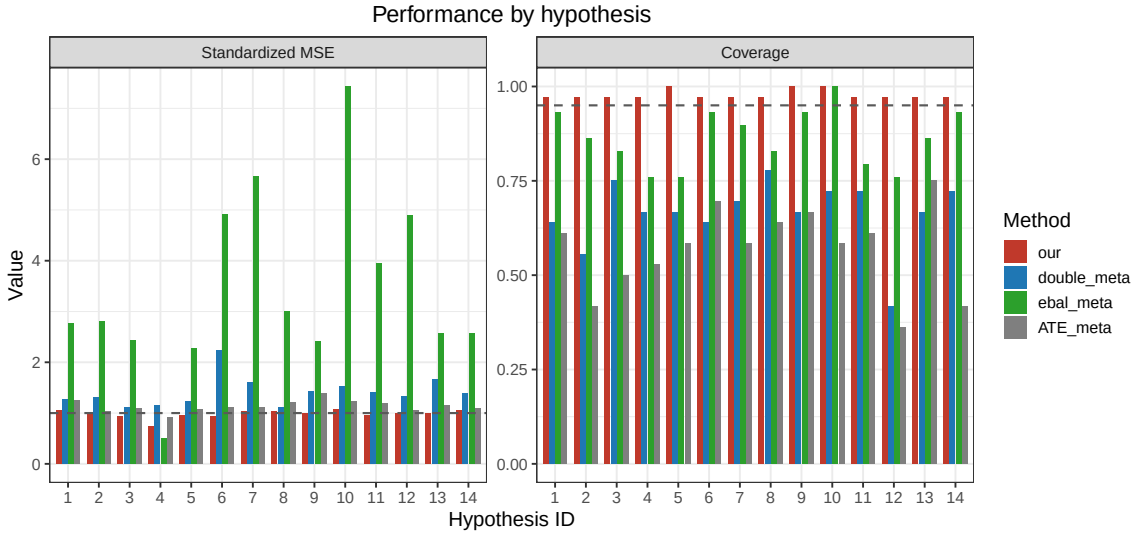


Causal Task Results

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



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 contributions:
  - 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