SureMap: Simultaneous Mean Estimation for Single-Task and Multi-Task Disaggregated Evaluation

Misha Khodak, Lester Mackey, Alexandra Chouldechova, Miro Dudík

mkhodak@princeton.edu, {lmackey, alexandrac, mdudik}@microsoft.com

What is disaggregated evaluation?

- reporting **performance** (e.g. accuracy, error, AUC) on **subpopulations** of data (e.g. race, sex, age)
- applications:
  - detecting fairness-related harms
  - detecting underperforming slices of data

Why is it hard?

- **lack of samples** from subpopulations, especially intersectional ones, makes naïve performance estimates inaccurate
- low-sample subpopulations **most** likely to be affected by model underperformance

Our contributions

- new disaggregated evaluation method (**SureMap**) inspired by classical mean estimation
- first **multi-task** method (MT SureMap) for making use of multiple model client distributions
- three new disaggregated evaluation **benchmarks**
- https://github.com/mkhodak/SureMap

Methodology

setup / assumptions:

- goal: estimate $\mu_g = \mathbb{E}_{z\sim\mathcal{D}|\mathcal{Z}_g}[f(z)]$ for model & loss $f\colon \mathcal{Z} \mapsto \mathbb{R}$, distribution $\mathcal{D}$ over $\mathcal{Z}$, and partition $\mathcal{Z}_1, \dots, \mathcal{Z}_d$ of the sample space
- data: a sample $S \subset \mathcal{Z}$ with n_g points in subpopulation $g \in [d]$
- assumption: there exists $\sigma^2 > 0$ such that for all $g \in [d]$

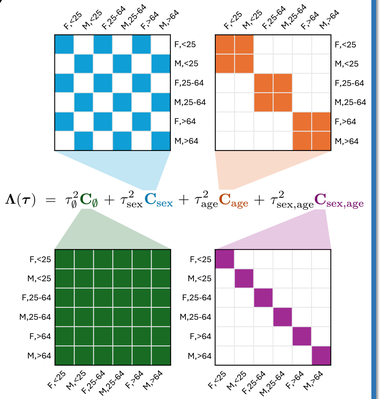
$$y_g = \frac{1}{n_g} \sum_{z \in S \cap \mathcal{Z}_g} f(z) \sim \mathcal{N}\left(\mu_g, \frac{\sigma^2}{n_g}\right)$$

MAP: for Gaussian data $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and prior $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Lambda})$

$$\widehat{\boldsymbol{\mu}}(\mathbf{y}) = (\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Sigma}^{-1}\mathbf{y} + \boldsymbol{\Lambda}^{-1}\boldsymbol{\theta})$$

single-task: use structured prior $\mathcal{N}(\mathbf{0}_d, \boldsymbol{\Lambda}(\boldsymbol{\tau}))$

- equivalent to a linear intersectional effects model, e.g. if subopulation g is an intersection of sex s and age a then $\mu_g = \tau_\emptyset\zeta + \tau_{\text{sex}}\zeta_s^{\text{sex}} + \tau_{\text{age}}\zeta_a^{\text{age}} + \tau_{\text{sex,age}}\zeta_g$, where $\zeta, \zeta_s^{\text{sex}}, \zeta_a^{\text{age}}, \zeta_g \sim \mathcal{N}(0,1)$
- generalizes naïve means, pooling, and Bock-Stein shrinkage, and is a flexible variant of structured regression
  [Herlihy-Truong-Chouldechova-Dudík, FAccT 2024]



multi-task: use prior $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Lambda}(\boldsymbol{\tau}))$ with hyperprior $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{\Lambda}(\boldsymbol{v}))$

SURE: tune $\boldsymbol{\tau}$ and $\boldsymbol{v}$ by minimizing Stein's unbiased risk estimate

$$\frac{\sigma^2}{d}\left(\|\widehat{\boldsymbol{\mu}}(\mathbf{y}) - \mathbf{y}\|_{\boldsymbol{\Sigma}^{-1}}^2 - d + 2\boldsymbol{\nabla}_{\mathbf{y}} \cdot \widehat{\boldsymbol{\mu}}(\mathbf{y})\right)$$

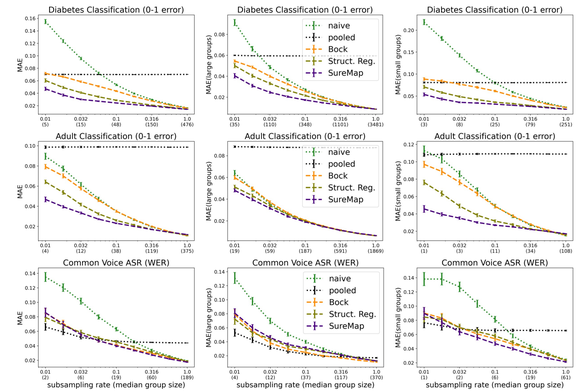
Single-task results

metrics: MAE $(\|\widehat{\boldsymbol{\mu}}(\mathbf{y}) - \boldsymbol{\mu}\|_1)$ of model evaluations when disaggregating by sex, age, and race (where available)

logistic regression

in-context llama3-70b

ASR with whisper-tiny

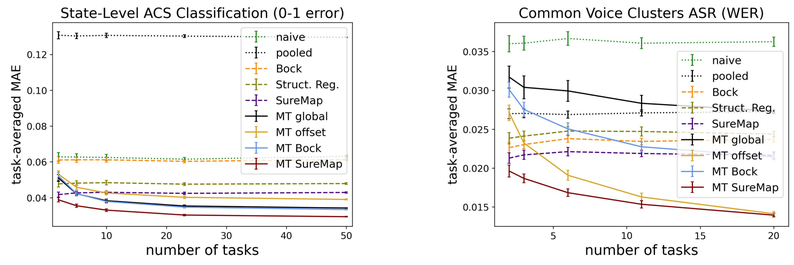


results:

- SureMap matches or outperforms competitors (left)
- most improvement on small groups (right)

Multi-task results

metrics: average MAE across tasks, which are defined by state census data (left) and clustered utterances (right)



results: MT SureMap outperforms single-task and multi-task approaches, even with just one additional task