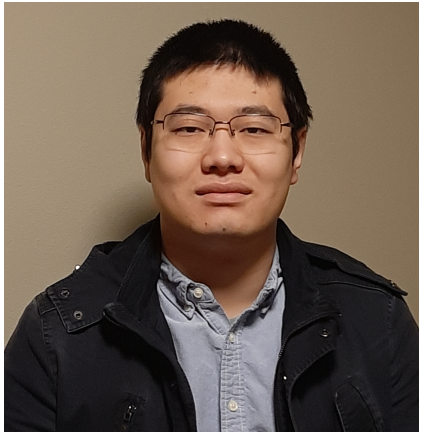
Efficient and Multiply Robust Risk Estimation under General Forms of Dataset Shift

Edgar Dobriban

Department of Statistics and Data Science, the Wharton School, University of Pennsylvania

November 6, 2023

Collaborators



Hongxiang (David) Qiu

Eric Tchetgen Tchetgen

Table of Contents

Motivation

- Machine learning approaches to prediction are increasingly successful.
- A common challenge: limited data available from the target domain/population, despite large related source data sets.[1]

[1] Will use colors to highlight source and target population throughout

Motivation

- Machine learning approaches to prediction are increasingly successful.
- A common challenge: limited data available from the target domain/population, despite large related source data sets.[1]
- It is valid to use target population data alone, but desirable to leverage relevant source data to *increase efficiency/accuracy*.
  - Related areas: transfer learning, domain adaptation, distribution shift

[1] Will use colors to highlight source and target population throughout

Motivation

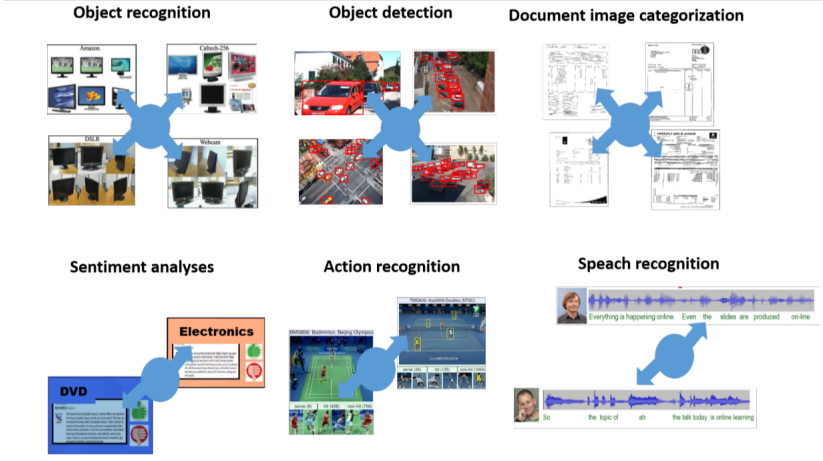


Fig. 1 Example scenarios with domain adaptation needs.

Figure: Csurka (2017)

Motivation

Example: Wish to predict HIV risk in one community with little data, leveraging data from other communities to improve prediction accuracy.

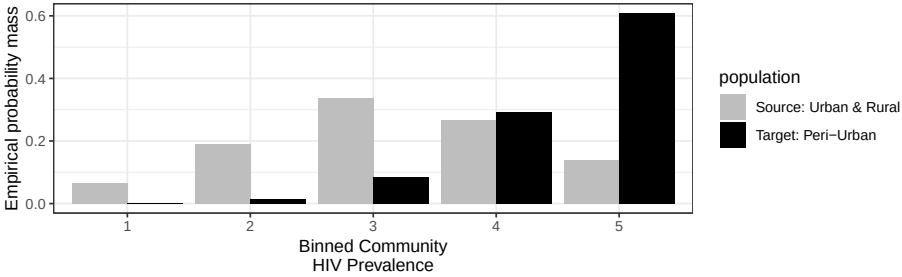
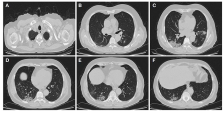


Figure: Qiu et al. (2022)

Motivation

Example: Wish to diagnose lung disease based on CT scans. Have limited labeled CT scans, but might leverage large texture datasets.



Al-Shudifat et al. (2022)

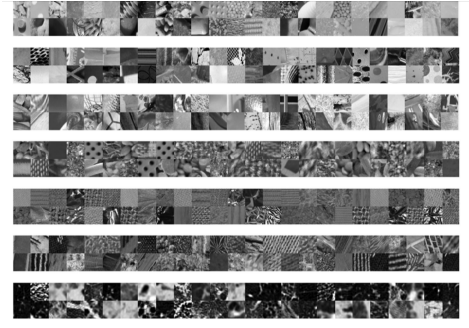


Fig. 1. Typical samples from each dataset. The color databases were converted to gray scale. From top to bottom: ALOT, DTD, FMD, KTB, KTH-TIPS-2b, UIUC, ILD.

Christodoulidis et al. (2017)

Motivation

Most methods for transfer learning/domain adaptation work by fine-tuning/adapting a predictor/estimator fit on the source data.

Motivation

Most methods for transfer learning/domain adaptation work by fine-tuning/adapting a predictor/estimator fit on the source data.

- Considering a supervised setting:
  - First, fit predictor f on—large—source data. Typically by risk minimization.

Motivation

Most methods for transfer learning/domain adaptation work by fine-tuning/adapting a predictor/estimator fit on the source data.

- Considering a supervised setting:
  - First, fit predictor f on—large—source data. Typically by risk minimization.
  - Then, fine-tune f on—small—target data. Also by risk minimization. Requires estimating the risk in the target population.

Motivation

Most methods for transfer learning/domain adaptation work by fine-tuning/adapting a predictor/estimator fit on the source data.

- Considering a supervised setting:
  - First, fit predictor f on—large—source data. Typically by risk minimization.
  - Then, fine-tune f on—small—target data. Also by risk minimization. Requires estimating the risk in the target population.
    - Since target dataset is small, accurate estimation is crucial.

Motivation

We study the estimation of a target population risk; for a generic datapoint Z and loss function ℓ:

$$\mathbb{E}[\ell(Z) \mid \text{target population}]$$

Motivation

We study the estimation of a target population risk; for a generic datapoint Z and loss function ℓ:

$$\mathbb{E}[\ell(Z) \mid \text{target population}]$$

Examples:

- Estimate the MSE: $Z = (X, Y)$, $\ell(Z) = (Y - f(X))^2$ for a given predictor f.

Motivation

We study the estimation of a target population risk; for a generic datapoint Z and loss function ℓ:

$$\mathbb{E}[\ell(Z) \mid \text{target population}]$$

Examples:

- Estimate the MSE: $Z = (X, Y)$, $\ell(Z) = (Y - f(X))^2$ for a given predictor f.
- Construct prediction sets with coverage guarantees and small sizes: estimate the coverage error $P(Y \notin C(X)) = \mathbb{E}(I(Y \notin C(X))$ (Vovk, 2013; Qiu et al., 2022; Yang et al., 2022).

Motivation

We study the estimation of a target population risk; for a generic datapoint Z and loss function ℓ:

$$\mathbb{E}[\ell(Z) \mid \text{target population}]$$

Examples:

- Estimate the MSE: $Z = (X, Y)$, $\ell(Z) = (Y - f(X))^2$ for a given predictor f.
- Construct prediction sets with coverage guarantees and small sizes: estimate the coverage error $P(Y \notin C(X)) = \mathbb{E}(I(Y \notin C(X))$ (Vovk, 2013; Qiu et al., 2022; Yang et al., 2022).
- "Risk" and "loss" can be interpreted broadly: To estimate the target population mean, take "loss" ℓ to be identity

Example dataset shift condition: Semi-supervised learning

- Semi-supervised learning (Concept shift in the features):
  $\{X \mid \text{target}\} \stackrel{d}{=} \{X \mid \text{source}\}$; $Y \mid X$ may differ between source and target

Example dataset shift condition: Semi-supervised learning

- Semi-supervised learning (Concept shift in the features): $\{X \mid \text{target}\} \stackrel{d}{=} \{X \mid \text{source}\}$; $Y \mid X$ may differ between source and target
- e.g., In a sample from the target population, a random subset is labeled (Y observed); the others are unlabeled (Y missing)

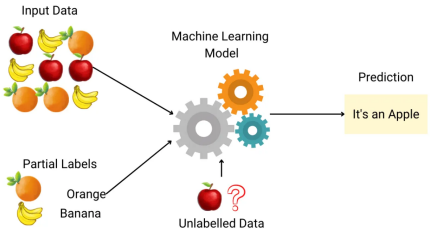


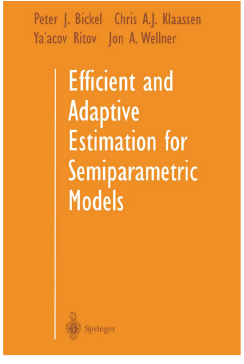
Figure: EnjoyAlgorithms.com

Motivation

- Dataset shift conditions can often be formulated as *restrictions on the observed data generating mechanism*, yielding a semiparametric model.
- Taking the perspective of modern semiparametric efficiency theory (Bickel et al., 1993; Pfanzagl, 1985, 1990; Bolthausen et al., 2002), we can ask: how can we leverage source data under dataset shift *most efficiently*?

Motivation

- Dataset shift conditions can often be formulated as *restrictions on the observed data generating mechanism*, yielding a semiparametric model.
- Taking the perspective of modern semiparametric efficiency theory (Bickel et al., 1993; Pfanzagl, 1985, 1990; Bolthausen et al., 2002), we can ask: how can we leverage source data under dataset shift *most efficiently*? .



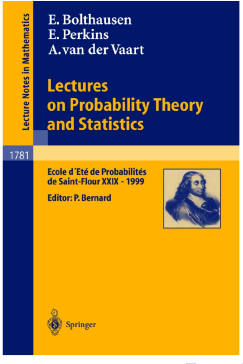


Table of Contents

Problem setup

- Observe i.i.d. copies of $O = (Z, A) \sim P_*$:
  - Actual data $Z \in \mathcal{Z}$: e.g., $Z = (X, Y)$
  - Population index $A \in \mathcal{A}$:

$$A = \begin{cases} 0 & \text{target population} \\ \text{another value, e.g., } 1 & \text{a source population} \end{cases}$$

- Estimand of interest: $r_* := \mathbb{E}_{P_*}[\ell(Z) \mid A = 0]$.

Problem setup

- Observe i.i.d. copies of $O = (Z, A) \sim P_*$:
  - Actual data $Z \in \mathcal{Z}$: e.g., $Z = (X, Y)$
  - Population index $A \in \mathcal{A}$:

$$A = \begin{cases} 0 & \text{target population} \\ \text{another value, e.g., } 1 & \text{a source population} \end{cases}$$

- Estimand of interest: $r_* := \mathbb{E}_{P_*}[\ell(Z) \mid A = 0]$.
- Given data $O_1 = (Z_1, A_1), \ldots, O_n = (Z_n, A_n)$. An "obvious" estimator is the average over the target population data:

$$\hat{r}_{\text{np}} := \frac{\sum_{i=1}^n \mathbb{1}(A_i = 0)\ell(Z_i)}{\sum_{i=1}^n \mathbb{1}(A_i = 0)},$$

but it may be inaccurate when target population data is small (number of i s.t. $A_i = 0$).

A general dataset shift condition

- Let Z be decomposed into K components $(Z_1, \ldots, Z_K)$
- Define $\bar{Z}_0 := \emptyset$, $\bar{Z}_k := (Z_1, \ldots, Z_k)$ for $k = 1, \ldots, K$

A general dataset shift condition

- Let Z be decomposed into K components $(Z_1, \ldots, Z_K)$
- Define $\bar{Z}_0 := \emptyset$, $\bar{Z}_k := (Z_1, \ldots, Z_k)$ for $k = 1, \ldots, K$

Condition (Sequential conditionals, Li & Luedtke '21)

For every k, there is a known *transfer set* $\mathcal{S}_k \subset \mathcal{A} \setminus \{0\}$ such that, for all $a \in \mathcal{S}_k$,

$$\left\{ Z_k \mid \bar{Z}_{k-1} = \bar{z}_{k-1}, A = a \right\} \stackrel{d}{=} \left\{ Z_k \mid \bar{Z}_{k-1} = \bar{z}_{k-1}, A = 0 \right\}$$

for all $\bar{z}_{k-1}$ in the common support of $\bar{Z}_{k-1} \mid A = 0$ and $\bar{Z}_{k-1} \mid A = a$.

Biometrika (2023), **00**, 0, *pp.* 1–14

https://doi.org/10.1093/biomet/asad007
Advance Access Publication 6 February 2023

Efficient estimation under data fusion

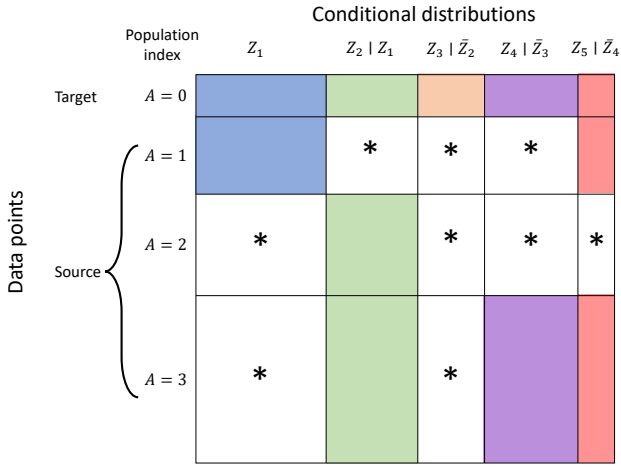
BY SIJIA LI AND ALEX LUEDTKE

SUMMARY

We aim to make inferences about a smooth, finite-dimensional parameter by fusing together data from multiple sources. Previous works have studied the estimation of a variety of parameters in similar data fusion settings, including estimation of the average treatment effect and average reward under a policy, with the majority of them merging one historical data source with covariates, actions and rewards, and one data source of the same covariates. In this article, we consider the general case where one or more data sources align with each part of the distribution of the target population, such as the conditional distribution of the reward given actions and covariates. We describe potential gains in efficiency that can arise from fusing these data sources together in a single analysis, which we characterize by a reduction in the semiparametric efficiency bound.

The sequential conditionals dataset shift condition

For all $a \in \mathcal{S}_k$, $\left\{ Z_k \mid \bar{Z}_{k-1} = \bar{z}_{k-1}, A = a \right\} \overset{d}{=} \left\{ Z_k \mid \bar{Z}_{k-1} = \bar{z}_{k-1}, A = 0 \right\}$.



Popular assumptions are special cases of "sequential conditionals"

"Sequential conditionals" includes the most popular dataset shift assumptions: concept shift in features+labels, covariate shift, label shift as special cases.

One source population ($A \in \{0, 1\}$) and $Z = (X, Y)$.

Popular assumptions are special cases of "sequential conditionals"

"Sequential conditionals" includes the most popular dataset shift assumptions: concept shift in features+labels, covariate shift, label shift as special cases.

One source population ($A \in \{0, 1\}$) and $Z = (X, Y)$.

- Concept shift in the features/semi-supervised learning: $\{X \mid A = 1\} \stackrel{d}{=} \{X \mid A = 0\}$;

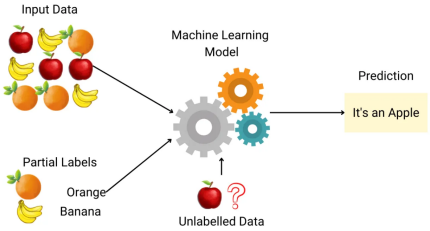


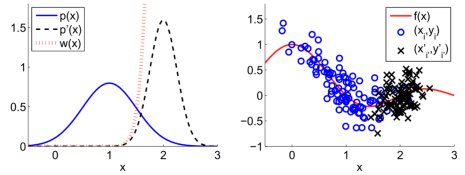
Figure: EnjoyAlgorithms.com

Popular assumptions are special cases of "sequential conditionals"

- Concept shift in the labels: $\{Y \mid A = 1\} \stackrel{d}{=} \{Y \mid A = 0\}$
- E.g., the definition of the features X differs between source and target

Examples of sequential conditionals: covariate shift

- Full-data covariate shift: $\{Y \mid X, A = 1\} \stackrel{d}{=} \{Y \mid X, A = 0\}$; covariate X distribution may differ between source and target populations.



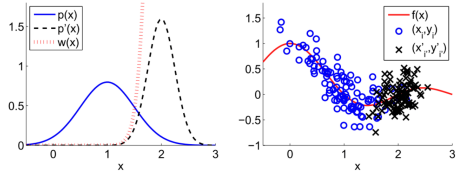
(a) Input densities and importance

(b) Learning target function, training samples, and test samples

Figure: Sugiyama et al., 2013

Examples of sequential conditionals: covariate shift

- Full-data covariate shift: $\{Y \mid X, A = 1\} \stackrel{d}{=} \{Y \mid X, A = 0\}$; covariate X distribution may differ between source and target populations.

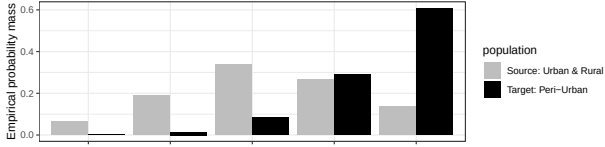


(a) Input densities and importance

(b) Learning target function, training samples, and test samples

Figure: Sugiyama et al., 2013

Example: Predict HIV risk Y with baseline covariates X using data from target and source communities



Examples of sequential conditionals: label shift

- Full-data label shift: $\{X \mid Y, A = 1\} \stackrel{d}{=} \{X \mid Y, A = 0\}$

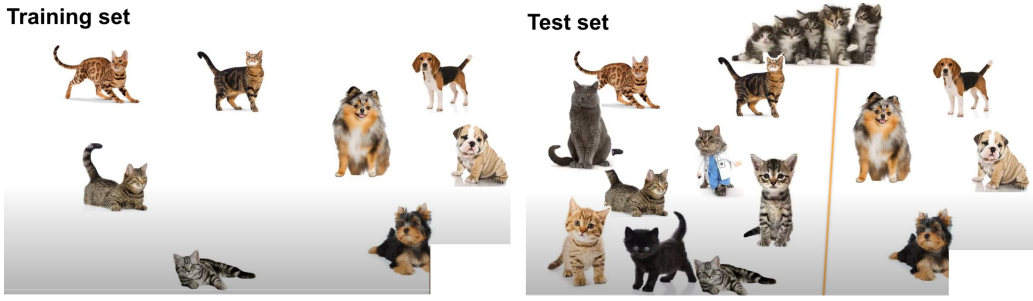


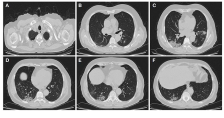
Figure: Alex Smola slides

Example (case-cohort study): Form a cohort from the target population, measure baseline covariates X and HIV risk Y for a random subset and all cases.

More sophisticated example of sequential conditionals

- Improving lung disease diagnosis with CT scans (Christodoulidis et al., 2017):
  - X_1: image
  - X_2: texture
  - Y: diagnosis

In addition to the labeled CT scans+textures, leverage a large image+texture dataset containing (X_1, X_2) and assume $\{X_2 \mid X_1, A = 1\} \overset{d}{=} \{X_2 \mid X_1, A = 0\}$



Al-Shudifat et al. (2022)

Fig. 1. Typical samples from each dataset. The color databases were converted to gray scale. From top to bottom: ALOT, DTD, FMD, KTB, KTH-TIPS-2b, UIUC, ILD.

Christodoulidis et al. (2017)

Details of example

Consider lung disease diagnosis with CT scans: $Z = (X_1, X_2, Y)$, X_1=image, X_2= texture, Y=diagnosis.

Data sets:

- Fully labeled CT scans from target population $(X_1, X_2, Y, A = 0)$
- Unlabeled CT scans from target population, no texture $(X_1, A = 1)$
- Large image+texture dataset $(X_1, X_2, A = 2)$
- Fully labeled CT scans from another population $(X_1, X_2, Y, A = 3)$

Details of example

Consider lung disease diagnosis with CT scans: $Z = (X_1, X_2, Y)$, X_1=image, X_2= texture, Y=diagnosis.

Data sets:

- Fully labeled CT scans from target population $(X_1, X_2, Y, A = 0)$
- Unlabeled CT scans from target population, no texture $(X_1, A = 1)$
- Large image+texture dataset $(X_1, X_2, A = 2)$
- Fully labeled CT scans from another population $(X_1, X_2, Y, A = 3)$

Relevant source data set indices $\mathcal{S}_k$

- $\mathcal{S}_1 = \{1\}$: $\{X_1 \mid A = 1\} \stackrel{d}{=} \{X_1 \mid A = 0\}$
- $\mathcal{S}_2 = \{2, 3\}$: $\{X_2 \mid X_1, A \in \{2, 3\}\} \stackrel{d}{=} \{X_2 \mid X_1, A = 0\}$
- $\mathcal{S}_3 = \{3\}$: $\{Y \mid X_1, X_2, A = 3\} \stackrel{d}{=} \{Y \mid X_1, X_2, A = 0\}$

Efficiency bound: nuisance functions/parameters

- Conditional odds of source vs target:

$$\theta_*^2(X_1, X_2) := \frac{P_*(A \in \mathcal{S}_3 = \{3\} \mid X_1, X_2)}{P_*(A = 0 \mid X_1, X_2)},$$
$$\theta_*^1(X_1) := \frac{P_*(A \in \mathcal{S}_2 = \{2, 3\} \mid X_1)}{P_*(A = 0 \mid X_1)}$$

Efficiency bound: nuisance functions/parameters

- Conditional odds of source vs target:

$$\theta_*^2(X_1, X_2) := \frac{P_*(A \in \mathcal{S}_3 = \{3\} \mid X_1, X_2)}{P_*(A = 0 \mid X_1, X_2)},$$
$$\theta_*^1(X_1) := \frac{P_*(A \in \mathcal{S}_2 = \{2, 3\} \mid X_1)}{P_*(A = 0 \mid X_1)}$$

- Can identify conditional mean loss via the following recursive definition:

$$\begin{aligned}
\ell_*^3 &:= \ell, \\
\ell_*^2(X_1, X_2) &:= \mathbb{E}_{P_*}[\ell_*^3(Z) \mid X_1, X_2, A \in \{0, 3\}] \\
&= \mathbb{E}_{P_*}[\ell(Z) \mid X_1, X_2, A \in \{0, 3\}], \\
\ell_*^1(X_1) &:= \mathbb{E}_{P_*}[\ell_*^2(X_1, X_2) \mid X_1, A \in \{0, 2, 3\}]
\end{aligned}$$

Efficiency bound: nuisance functions/parameters

- We can show that ℓ_*^k is indeed a conditional mean loss in target population:

$$\ell_*^2(X_1, X_2) = \mathbb{E}_{P_*}[\ell(Z) \mid X_1, X_2, A = 0],$$
$$\ell_*^1(X_1) = \mathbb{E}_{P_*}[\ell(Z) \mid X_1, A = 0].$$

Efficiency bound: nuisance functions/parameters

- We can show that ℓ_*^k is indeed a conditional mean loss in target population:

$$\ell_*^2(X_1, X_2) = \mathbb{E}_{P_*}[\ell(Z) \mid X_1, X_2, A = 0],$$
$$\ell_*^1(X_1) = \mathbb{E}_{P_*}[\ell(Z) \mid X_1, A = 0].$$

- Define marginal probabilities of populations: $\pi_*^a := P_*(A = a)$.

Efficiency bound: nuisance functions/parameters

- We can show that ℓ_*^k is indeed a conditional mean loss in target population:

$$\ell_*^2(X_1, X_2) = \mathbb{E}_{P_*}[\ell(Z) \mid X_1, X_2, A = 0],$$
$$\ell_*^1(X_1) = \mathbb{E}_{P_*}[\ell(Z) \mid X_1, A = 0].$$

- Define marginal probabilities of populations: $\pi_*^a := P_*(A = a)$.
- Collect nuisance functions/parameters:

$$\boldsymbol{\theta}_* := (\theta_*^1, \theta_*^2), \quad \boldsymbol{\ell}_* := (\ell_*^1, \ell_*^2), \quad \boldsymbol{\pi}_* := (\pi_*^a)_{a \in \mathcal{A}}.$$

Efficiency bound

Semiparametric theory tells us that, if there is an efficient influence function $D_{\mathrm{SC}}(\ell, \theta, \pi, r)$, then an efficient estimator $\hat{r}$ must satisfy

$$\hat{r} = r_* + \frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{SC}}(\ell_*, \theta_*, \pi_*, r_*)(O_i) + \mathrm{o}_p(n^{-1/2}).$$

Efficiency bound

Semiparametric theory tells us that, if there is an efficient influence function $D_{\mathrm{SC}}(\ell, \theta, \pi, r)$, then an efficient estimator $\hat{r}$ must satisfy

$$\hat{r} = r_* + \frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{SC}}(\ell_*, \theta_*, \pi_*, r_*)(O_i) + \mathrm{o}_p(n^{-1/2}).$$

Results in Li and Luedtke (2021) imply that the efficient influence function under this condition is

$$\begin{aligned} D_{\mathrm{SC}}(\ell, \theta, \pi, r) : o \mapsto & \frac{\mathbb{1}(a \in \{0,3\})}{\pi^0(1+\theta^2(x_1,x_2))}\left\{\ell(z) - \ell^2(x_1,x_2)\right\} \\ & + \frac{\mathbb{1}(a \in \{0,2,3\})}{\pi^0(1+\theta^1(x_1))}\left\{\ell^2(x_1,x_2) - \ell^1(x_1)\right\} \\ & + \frac{\mathbb{1}(a \in \{0,1\})}{\pi^0(1+\theta^0)}\left\{\ell^1(x_1) - r\right\} \end{aligned}$$

Efficient and multiply robust estimation

- When/how can we construct such an efficient estimator $\hat{r}$?

Efficient and multiply robust estimation

- When/how can we construct such an efficient estimator $\hat{r}$?
- How can we make $\hat{r}$ multiply robust against inconsistent estimation of some of the nuisance functions ℓ_* and θ_*?

Efficient and multiply robust estimation

- When/how can we construct such an efficient estimator $\hat{r}$?
- How can we make $\hat{r}$ multiply robust against inconsistent estimation of some of the nuisance functions $\boldsymbol{\ell}_*$ and $\boldsymbol{\theta}_*$?
- Since $\mathbb{E}D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*) = 0$, generalizing usual statement that "score function has mean zero". Motivated by this, we use an estimating equation approach, solving

$$\sum_{i=1}^{n} D_{\mathrm{SC}}(\widehat{\boldsymbol{\ell}}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}}, r)(O_i) = 0 \quad \text{for } r.$$

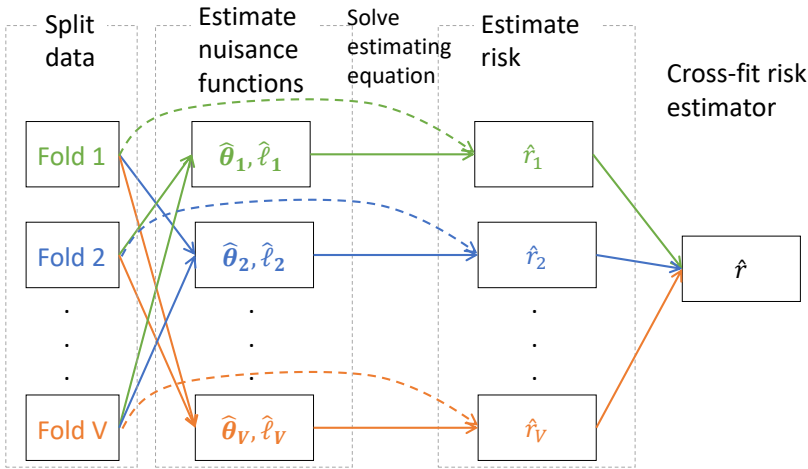
Efficient and multiply robust estimation

- When/how can we construct such an efficient estimator $\hat{r}$?
- How can we make $\hat{r}$ multiply robust against inconsistent estimation of some of the nuisance functions $\boldsymbol{\ell}_*$ and $\boldsymbol{\theta}_*$?
- Since $\mathbb{E}D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*) = 0$, generalizing usual statement that "score function has mean zero". Motivated by this, we use an estimating equation approach, solving

$$\sum_{i=1}^{n} D_{\mathrm{SC}}(\widehat{\boldsymbol{\ell}}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}}, r)(O_i) = 0 \quad \text{for } r.$$

- We also use cross-fitting to relax conditions on nuisance function estimators $(\widehat{\boldsymbol{\ell}}, \widehat{\boldsymbol{\theta}})$.

Cross-fit risk estimator



Cross-fit risk estimator

1: Randomly split data into V folds with index sets I_v $(v = 1, \ldots, V)$.
2: **for** $v = 1, \ldots, V$ **do**
3: For $k \in \{1, 2\}$, estimate θ^k by $\hat{\theta}^k_v$ using data out of fold v
4: Set $\hat{\pi}^a_v := |I_v|^{-1} \sum_{i \in I_v} \mathbb{1}(A_i = a)$ for all $a \in \mathcal{A}$
5: Estimate ℓ^2_* by $\hat{\ell}^2_v$ using data out of fold v: regress $\hat{\ell}^3_v(Z) := \ell(Z)$ on covariates (X_1, X_2) in the subsample with $A \in \{0, 3\}$
6: Estimate ℓ^1_* by $\hat{\ell}^1_v$ using data out of fold v: regress $\hat{\ell}^2_v(X_1, X_2)$ on covariate X_1 in the subsample with $A \in \{0, 2, 3\}$
7: Estimator $\hat{r}_v$ is the solution in r to: ▷ Can be solved explicitly

$$\sum_{i \in I_v} D_{\mathrm{SC}}(\widehat{\ell}_v, \widehat{\theta}_v, \widehat{\pi}_v, r)(O_i) = 0.$$

8: Cross-fit estimator $\hat{r} := \frac{1}{n} \sum_{v=1}^{V} |I_v| \hat{r}_v$ (average of $\hat{r}_v$ over folds).

Efficiency and multiple robustness

Define *oracle conditional mean loss estimator* h_v^{k-1} of ℓ_*^{k-1} based on $\hat{\ell}_v^k$, evaluated under the true distribution P_*:

$$\begin{aligned} h_v^2(X_1, X_2) &:= \mathbb{E}_{P_*}[\hat{\ell}_v^3(Z) \mid X_1, X_2, A \in \{0, 3\}] \\ &= \mathbb{E}_{P_*}[\ell(Z) \mid X_1, X_2, A \in \{0, 3\}], \\ h_v^1(X_1) &:= \mathbb{E}_{P_*}[\hat{\ell}_v^2(X_1, X_2) \mid X_1, A \in \{0, 2, 3\}]. \end{aligned}$$

Efficiency and multiple robustness

Theorem (Qiu et al., 2023)

- *(Efficiency) If, for every fold v and $k = 1, 2$,*

$$\left\| \frac{1}{1+\hat{\theta}_v^k} - \frac{1}{1+\theta_*^k} \right\| \quad \text{and} \quad \left\| \hat{\ell}_v^k - h_v^k \right\|$$

  are both $o_p(1)$ and their product is $o_p(n^{-1/2})$, then $\hat{r}$ is efficient.

- *(2^{K-1}-robustness) If, for every v and $k = 1, 2$,*

$$\left\| \frac{1}{1+\hat{\theta}_v^k} - \frac{1}{1+\theta_*^k} \right\| \quad \text{or} \quad \left\| \hat{\ell}_v^k - h_v^k \right\|$$

  is $o_p(1)$, then $\hat{r}$ is consistent.

Have the same result for general sequential conditionals.

Crucial role of parameterization

Since

$$\ell_*^2(X_1, X_2) = \mathbb{E}_{P_*}[\ell(Z) \mid X_1, X_2, A = 0],$$
$$\ell_*^1(X_1) = \mathbb{E}_{P_*}[\ell(Z) \mid X_1, A = 0],$$

why not obtain $\widehat{\ell}_v$ by directly regressing loss $\ell(Z)$ on covariate (X_1, X_2) or X_1 in the target population data?

Crucial role of parameterization

Since

$$\ell_*^2(X_1, X_2) = \mathbb{E}_{P_*}[\ell(Z) \mid X_1, X_2, A = 0],$$
$$\ell_*^1(X_1) = \mathbb{E}_{P_*}[\ell(Z) \mid X_1, A = 0],$$

why not obtain $\widehat{\ell}_v$ by directly regressing loss $\ell(Z)$ on covariate (X_1, X_2) or X_1 in the target population data?

Heuristically, our sequential regression approach better leverages the "sequential conditionals" condition.

Crucial role of parameterization

Theoretically:

- One term in the second-order bias of $\hat{r}$ takes the form

$$\mathbb{E}_{P_*}\left[\left(\frac{1}{1+\hat{\theta}_v^2(X_1,X_2)}-\frac{1}{1+\theta_*^2(X_1,X_2)}\right)\left(\hat{\ell}_v^2(X_1,X_2)-h_v^2(X_1,X_2)\right) \mid A\in\{0,2,3\}\right]$$
$$+\,\mathbb{E}_{P_*}\left[\left(\frac{1}{1+\hat{\theta}_v^1(X_1)}-\frac{1}{1+\theta_*^1(X_1)}\right)\left(\hat{\ell}_v^1(X_1)-h_v^1(X_1)\right) \mid A\in\{0,1\}\right]$$

- Natural to require $\hat{\ell}_v^k$ to be close to the oracle estimator h_v^k, not necessarily to ℓ_*^k.
- This difference is crucial for achieving 2^{K-1}-robustness.

Crucial role of parameterization

$$\mathbb{E}_{P_*}\left[\left(\frac{1}{1+\hat{\theta}_v^2(X_1,X_2)}-\frac{1}{1+\theta_*^2(X_1,X_2)}\right)\left(\hat{\ell}_v^2(X_1,X_2)-h_v^2(X_1,X_2)\right)\mid A\in\{0,2\}\right]$$
$$+\,\mathbb{E}_{P_*}\left[\left(\frac{1}{1+\hat{\theta}_v^1(X_1)}-\frac{1}{1+\theta_*^1(X_1)}\right)\left(\hat{\ell}_v^1(X_1)-h_v^1(X_1)\right)\mid A\in\{0,1\}\right] \qquad (1)$$

If we obtain conditional mean loss estimators $\widehat{\ell}_v$ by direct regression:

- Suppose that $\hat{\ell}_v^2$ is inconsistent; $\hat{\ell}_v^3=\ell$ and $\hat{\ell}_v^1$ are consistent.
- To make (1) small, we would need both $1/(1+\hat{\theta}_v^2)$ and $1/(1+\hat{\theta}_v^1)$ to be consistent. The reason is that the inconsistency of $\hat{\ell}_v^2$ propagates to h_v^1.
- This approach does not achieve 2^{K-1}-robustness: the estimator may still be inconsistent, if, for every $k\in\{1,2\}$, only one of $\hat{\ell}_v^k$ and $1/(1+\hat{\theta}_v^k)$ is inconsistent.

Table of Contents

Notations

- Set $Z = (X, Y)$ and $A \in \{0, 1\}$.
- Concept shift in the features/semi-supervised learning: $\{X \mid A = 1\} \overset{d}{=} \{X \mid A = 0\}$. Equivalently, observe $(X, Y, A = 0)$ and $(X, A = 1)$
- Define conditional mean loss

$$\mathcal{E}_* : x \mapsto \mathbb{E}_{P_*}[\ell(X, Y) \mid X = x, A = 0]$$

and probability of target population $\rho_* := P_*(A = 0)$.

Efficiency bound and gain

According to the results for "sequential conditionals", the efficient influence function is

$$D_{\mathrm{Xcon}}(\rho, \mathcal{E}, r) : o \mapsto \frac{\mathbb{1}(a = 0)}{\rho}\{\ell(x, y) - \mathcal{E}(x)\} + \mathcal{E}(x) - r.$$

Efficiency bound and gain

According to the results for "sequential conditionals", the efficient influence function is

$$D_{\mathrm{Xcon}}(\rho, \mathcal{E}, r) : o \mapsto \frac{\mathbb{1}(a = 0)}{\rho}\{\ell(x, y) - \mathcal{E}(x)\} + \mathcal{E}(x) - r.$$

The relative efficiency gain from using an efficient estimator vs. $\hat{r}_{\mathrm{np}}$ is, with $\mathcal{E}_* : x \mapsto \mathbb{E}_{P_*}[\ell(X, Y) \mid X = x, A = 0]$

$$\begin{aligned} RE =& 1 - \frac{\text{efficient asymptotic variance}}{\text{asymptotic variance of } \hat{r}_{\mathrm{np}}} \\ =& \frac{(1 - \rho_*)\mathbb{E}_{P_*}\left[(\mathcal{E}_*(X) - r_*)^2\right]}{\mathbb{E}_{P_*}\left[\mathbb{E}_{P_*}\left[\{\ell(X, Y) - \mathcal{E}_*(X)\}^2 \mid A = 0, X\right]\right] + \mathbb{E}_{P_*}\left[\{\mathcal{E}_*(X) - r_*\}^2\right]} \end{aligned}$$

- Variability of $\ell(X, Y)$ due to X
- Variability of $\ell(X, Y)$ not due to X

Efficiency bound and gain

$$RE = \frac{(1-\rho_*)\mathbb{E}_{P_*}\left[(\mathcal{E}_*(X)-r_*)^2\right]}{\mathbb{E}_{P_*}\left[\mathbb{E}_{P_*}\left[\{\ell(X,Y)-\mathcal{E}_*(X)\}^2 \mid A=0,X\right]\right]+\mathbb{E}_{P_*}\left[\{\mathcal{E}_*(X)-r_*\}^2\right]}$$

To gain large efficiency, P_* should satisfy:

1. ρ_* is small, i.e., limited target population data
2. In the target population, variability of $\ell(X,Y)$ due to X is large compared to variability of $\ell(X,Y)$ not due to X

Efficiency bound and gain

$$RE = \frac{(1-\rho_*)\mathbb{E}_{P_*}\left[(\mathcal{E}_*(X) - r_*)^2\right]}{\mathbb{E}_{P_*}\left[\mathbb{E}_{P_*}\left[\{\ell(X,Y) - \mathcal{E}_*(X)\}^2 \mid A = 0, X\right]\right] + \mathbb{E}_{P_*}\left[\{\mathcal{E}_*(X) - r_*\}^2\right]}$$

To gain large efficiency, P_* should satisfy:

1. ρ_* is small, i.e., limited target population data
2. In the target population, variability of $\ell(X, Y)$ due to X is large compared to variability of $\ell(X, Y)$ not due to X

MSE estimation example:

- $\ell(x, y) = (y - f(x))^2$ for a given predictor f; while $Y = \mu_*(X) + \epsilon$ where $\epsilon \perp\!\!\!\perp X$
- Variability of $\ell(X, Y)$ due to X is determined by the bias $f - \mu_*$
- Variability of $\ell(X, Y)$ not due to X is determined by ϵ
- We gain efficiency for f far from the truth μ_*. Extends linear regression results from Azriel et al. (2021) to general risk estimation

Efficiency & fully robust regularity and asymptotic linearity

- The cross-fit estimator $\hat{r}_{\mathrm{Xcon}}$ is a special case of the general one for "sequential conditionals"
- Rely on out-of-fold estimator $\hat{\mathcal{E}}^{-v}$ of $\mathcal{E}_*$

Efficiency & fully robust regularity and asymptotic linearity

- The cross-fit estimator $\hat{r}_{\mathrm{Xcon}}$ is a special case of the general one for "sequential conditionals"
- Rely on out-of-fold estimator $\hat{\mathcal{E}}^{-v}$ of $\mathcal{E}_*$

Theorem

If $\|\hat{\mathcal{E}}^{-v} - \mathcal{E}_\infty\| = \mathrm{o}_p(1)$ for some function $\mathcal{E}_\infty$, then the cross-fit estimator $\hat{r}_{\mathrm{Xcon}}$ is regular and asymptotically linear:

$$\hat{r}_{\mathrm{Xcon}} - r_* = \frac{1}{n}\sum_{i=1}^{n}\left\{D_{\mathrm{Xcon}}(\rho_*, \mathcal{E}_\infty, r_*)(O_i) + \frac{\mathbb{E}_{P_*}\left[\mathcal{E}_\infty(X)\right] - r_*}{\rho_*}(1 - A_i - \rho_*)\right\} + \mathrm{o}_p(n^{-1/2}).$$

If $\mathcal{E}_\infty = \mathcal{E}_*$, then $\hat{r}_{\mathrm{Xcon}}$ is efficient.

Efficiency & fully robust regularity and asymptotic linearity

In this special case, we have some additional desirable properties:

- Efficiency: *no convergence rate requirement* on $\hat{\mathcal{E}}^{-v}$
- Fully robust regularity and asymptotic linearity: even if the nuisance function estimator $\hat{\mathcal{E}}^{-v}$ is inconsistent,
  - $\hat{r}_{\mathrm{Xcon}}$ is still consistent and asymptotically normal
  - we have valid inference about r_*; crucial e.g., for constructing prediction sets with training-set conditional coverage (Bates et al., 2021; Qiu et al., 2022)

Simulation

Estimate MSE in five scenarios ($\rho_* = 0.1$):

(A) No efficiency gain: $f = \mu_*$

(B) Little efficiency gain: $f \approx \mu_*$

(C) Large efficiency gain: f far from μ_*

(D) Very large efficiency gain: f far from μ_* and no noise ($\epsilon = 0$)

(E) Concept shift does not hold: $\{X \mid A = 1\} \overset{d}{\neq} \{X \mid A = 0\}$

Three estimators:

- `np`: straightforward but imprecise nonparametric estimator $\hat{r}_{\text{np}}$
- `Xconshift`: $\hat{r}_{\text{Xcon}}$ with consistent $\hat{\mathcal{E}}^{-v}$. Super Learner (GLM/GAM, GLM+Lasso, gradient boosting)
- `Xconshift,mis.E`: $\hat{r}_{\text{Xcon}}$ with inconsistent $\hat{\mathcal{E}}^{-v}$ (fixed function for A, D; Super Learner w/o gradient boosting for B, C)

Simulation details

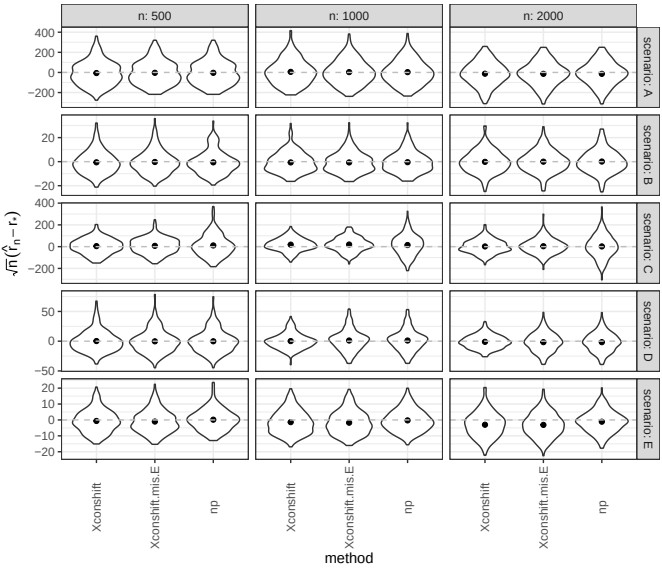
- Generate $X = (X_1, X_2, X_3) \sim N(0, I_3)$.
- A-D: Generate $A \sim \text{Bernoulli}(0.9)$ independent of X.
- Generate label Y in the target population: (A) $Y \mid X = x, A = 0 \sim \text{N}(\mu_*(x), 5^2)$; (B) & (C): $Y \mid X = x, A = 0 \sim \text{N}(\mu_*(x), 1)$; (D): $Y = \mu_*(X)$, where

$$\mu_*(x) = x_1 + x_2 + x_3 + 0.4x_1x_3 - 0.5x_2x_3 + \sin(x_1 + x_3).$$

  We set different predictors f for these scenarios:
  (A) f is the truth μ_*;
  (B) f is a linear function close to the best linear approximation to μ_* in $L^2(P_*)$-sense: $f(x) = 1.4x_1 + x_2 + 1.4x_3$;
  (C) f substantially differs from μ_*: $f(x) = -1 - 3x_1 + 0.5x_3$;
  (D) f substantially differs from μ_*: $f(x) = x_1$.
- For E, f as in B. Include dependence of A on X:

$$A \mid X = x \sim \text{Bernoulli}(\text{expit}\{\cos(x_1 + x_2x_3) + 2x_1^2x_2^2 + 3|x_1x_3| + |x_2|(0.5 - x_3)\}).$$

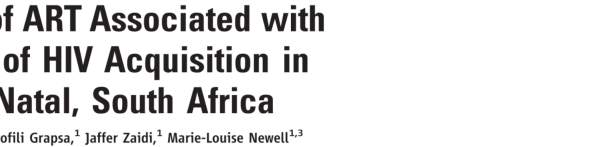
Simulation results



Data analysis example: HIV risk prediction

Data from a large population-based prospective cohort study in KwaZulu-Natal, South Africa (Tanser et al., 2013).

- Y: HIV seroconversion (Y/N); X: baseline covariates including age, sex, marital status, wealth quintile etc.
- Target population: peri-urban communities with ART coverage below 15%
- Source population: urban and rural communities
- Train a classifier f using half of the source population data (6192)
- Use 50 target population datapoints and the other half of the source population data to estimate inaccuracy $\mathbb{E}_{P_*}[\mathbb{1}(Y \neq f(X)) \mid A = 0]$
- Use the rest of the target population data for validation



High Coverage of ART Associated with Decline in Risk of HIV Acquisition in Rural KwaZulu-Natal, South Africa

Frank Tanser,[1*] Till Bärnighausen,[1,2] Erofili Grapsa,[1] Jaffer Zaidi,[1] Marie-Louise Newell[1,3]

Data analysis: HIV risk prediction under the four common dataset shift conditions

- Concept shift assumptions not plausible: require X or Y to have identical distrib.
- Label shift not plausible either: requires $X|Y$ unchanged (most reasonable when Y causes X)

Data analysis: HIV risk prediction under the four common dataset shift conditions

- Concept shift assumptions not plausible: require X or Y to have identical distrib.
- Label shift not plausible either: requires $X|Y$ unchanged (most reasonable when Y causes X)

Table: Risk estimates from HIV risk prediction data. The risk estimate from the validation dataset, $n_{\text{val}} \approx 1300$ is 0.24 (95% CI: 0.22–0.26).

Dataset Shift Condition	Estimate	S.E.	95% CI
None	0.24	0.060	(0.12, 0.36)
Concept shift in the features	0.26	0.057	(0.15, 0.38)
Concept shift in the labels	0.10	0.010	(0.08, 0.12)
Full-data covariate shift	0.19	0.026	(0.14, 0.25)
Full-data label shift	0.23	0.059	(0.11, 0.34)

Low power: not rejecting dataset shift conditions is weak evidence for plausibility.

Summary and Outlook

- We introduced a general framework for estimating the risk of a fixed predictor in distribution shift scenarios
  - One target population, multiple source populations
  - "Sequential conditionals": conditional distributions of data components are shared with the target (includes covariate/label/concept shift)
  - Developed methods for efficient and multiply robust risk estimation
  - Illustrated when this can lead to efficiency gains
- Key next step: How to use this for transfer learning? (simultaneously learn predictor)

References

A. E. Al-Shudifat, A. Al-Radaideh, S. Hammad, N. Hijjawi, S. Abu-Baker, M. Azab, and R. Tayyem. Association of Lung CT Findings in Coronavirus Disease 2019 (COVID-19) With Patients' Age, Body Weight, Vital Signs, and Medical Regimen. *Frontiers in Medicine*, 9:1925, 2022. ISSN 2296858X. doi: 10.3389/fmed.2022.912752.

D. Azriel, L. D. Brown, M. Sklar, R. Berk, A. Buja, and L. Zhao. Semi-Supervised Linear Regression. *Journal of the American Statistical Association*, 117(540):2238–2251, 2021. ISSN 1537274X. doi: 10.1080/01621459.2021.1915320.

S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan. Distribution-free, risk-controlling prediction sets. *arXiv preprint arXiv:2101.02703*, 2021.

P. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, 1993. ISBN 9780387984735.

E. Bolthausen, E. Perkins, and A. van der Vaart. *Lectures on Probability Theory and Statistics: Ecole D'Eté de Probabilités de Saint-Flour XXIX-1999*, volume 1781 of *Lecture Notes in Mathematics*. Springer Science & Business Media, Berlin, Heidelberg, 2002. ISBN 978-3-540-43736-9. doi: 10.1007/B93152.

References

N. Chatterjee, Y. H. Chen, P. Maas, and R. J. Carroll. Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-Level Information From External Big Data Sources. *Journal of the American Statistical Association*, 111(513):107–117, 2016. ISSN 1537274X. doi: 10.1080/01621459.2015.1123157.

S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, and S. Mougiakakou. Multisource Transfer Learning with Convolutional Neural Networks for Lung Pattern Analysis. *IEEE Journal of Biomedical and Health Informatics*, 21(1):76–84, 2017. ISSN 21682208. doi: 10.1109/JBHI.2016.2636929.

G. Csurka. A comprehensive survey on domain adaptation for visual applications. *Advances in Computer Vision and Pattern Recognition*, (9783319583464):1–35, 2017. ISSN 21916594. doi: 10.1007/978-3-319-58347-1_1. URL `www.xrce.xerox.com`.

J. Gronsbell, M. Liu, L. Tian, and T. Cai. Efficient evaluation of prediction rules in semi-supervised settings under stratified sampling. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 84(4):1353–1391, 2022.

S. Li and A. Luedtke. Efficient Estimation Under Data Fusion. *Biometrika*, 2021. ISSN 0006-3444. doi: 10.1093/BIOMET/ASAD007.

References

J. Pfanzagl. *Contributions to a general asymptotic statistical theory*, volume 3 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 1985. ISBN 978-0-387-90776-5. doi: 10.1524/strm.1985.3.34.379.

J. Pfanzagl. *Estimation in semiparametric models*, volume 63 of *Lecture Notes in Statistics*. Springer, New York, NY, 1990. doi: 10.1007/978-1-4612-3396-1_5.

H. Qiu, E. Dobriban, and E. Tchetgen Tchetgen. Prediction Sets Adaptive to Unknown Covariate Shift. *arXiv preprint arXiv:2203.06126v5*, 2022. doi: 10.48550/arxiv.2203.06126.

J. M. Robins, F. Hsieh, and W. Newey. Semiparametric Efficient Estimation of a Conditional Density with Missing or Mismeasured Covariates. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(2):409–424, 1995. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1995.tb02036.x.

A. Rotnitzky, D. Faraggi, and E. Schisterman. Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association*, 101(475):1276–1288, 2006. ISSN 01621459. doi: 10.1198/016214505000001339.

References

F. Tanser, T. Bärnighausen, E. Grapsa, J. Zaidi, and M. L. Newell. High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. *Science*, 339(6122): 966–971, 2013. ISSN 10959203. doi: 10.1126/science.1228160.

V. Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92(2-3):349–376, 2013. ISSN 08856125. doi: 10.1007/s10994-013-5355-6.

Y. Yang, A. K. Kuchibhotla, and E. T. Tchetgen. Doubly Robust Calibration of Prediction Sets under Covariate Shift. *arXiv preprint arXiv:2203.01761*, 2022. doi: 10.48550/arxiv.2203.01761.

Y. Zhang, A. Chakrabortty, and J. Bradic. Double robust semi-supervised inference for the mean: Selection bias under mar labeling with decaying overlap. *arXiv preprint arXiv:2104.06667*, 2021.

Related works

- Vast literature on [transfer learning]/[domain adaptation]/[dataset shift], but most for a *not fully observed* data from target population, a different scenario.

Related works

- Vast literature on [transfer learning]/[domain adaptation]/[dataset shift], but most for a *not fully observed* data from target population, a different scenario.
- Some works study estimation of means or generalized linear models with both labeled and unlabeled target population data (semi-supervised learning) (Azriel et al. (2021), Gronsbell et al. (2022), Zhang et al. (2021)).

  Particular cases of our framework.

Related works

- Vast literature on [transfer learning]/[domain adaptation]/[dataset shift], but most for a *not fully observed* data from target population, a different scenario.
- Some works study estimation of means or generalized linear models with both labeled and unlabeled target population data (semi-supervised learning) (Azriel et al. (2021), Gronsbell et al. (2022), Zhang et al. (2021)).

  Particular cases of our framework.
- Another related area is *data fusion* with an emphasis on causal inference applications (Chatterjee et al. (2016), Li and Luedtke (2021), Robins et al. (1995)). Data from target population might not be fully observed.

Related works

- Vast literature on [transfer learning]/[domain adaptation]/[dataset shift], but most for a *not fully observed* data from target population, a different scenario.
- Some works study estimation of means or generalized linear models with both labeled and unlabeled target population data (semi-supervised learning) (Azriel et al. (2021), Gronsbell et al. (2022), Zhang et al. (2021)).

  Particular cases of our framework.
- Another related area is *data fusion* with an emphasis on causal inference applications (Chatterjee et al. (2016), Li and Luedtke (2021), Robins et al. (1995)). Data from target population might not be fully observed.
- A general framework for efficient risk estimation under general forms of dataset shift is lacking.

Efficiency bound

- Conditional odds of source vs target:

$$\theta_*^{k-1} : \bar{z}_{k-1} \mapsto \frac{P_*(A \in \mathcal{S}_k \mid \bar{Z}_{k-1} = \bar{z}_{k-1})}{P_*(A = 0 \mid \bar{Z}_{k-1} = \bar{z}_{k-1})},$$

- Conditional mean loss (recursive definition): $\ell_*^K := \ell$,

$$\ell_*^k : \bar{z}_k \mapsto \mathbb{E}_{P_*}[\ell_*^{k+1}(\bar{Z}_{k+1}) \mid \bar{Z}_k = \bar{z}_k, A \in \mathcal{S}'_{k+1}],$$

  We can show that $\ell_*^k(\bar{z}_k) = \mathbb{E}_{P_*}[\ell(Z) \mid \bar{Z}_k = \bar{z}_k, A = 0]$ for $\bar{z}_k$ in the support of $\bar{Z}_{k-1} \mid A = 0$.

- Marginal probabilities of populations: $\pi_*^a := P_*(A = a)$.
- Collections of nuisance functions: $\boldsymbol{\theta}_* := (\theta_*^k)_{k=1}^{K-1}$, $\boldsymbol{\ell}_* := (\ell_*^k)_{k=1}^{K-1}$, $\boldsymbol{\pi}_* := (\pi_*^a)_{a \in \mathcal{A}}$.

Efficiency bound

- Pseudo-loss/unbiased transformation (Rotnitzky et al. (2006) JASA):

$$\mathcal{T}(\ell, \theta, \pi) : o \mapsto \sum_{k=2}^{K} \frac{\mathbb{1}(a \in \mathcal{S}'_k)}{\pi^0(1 + \theta^{k-1}(\bar{z}_{k-1}))} \left\{ \ell^k(\bar{z}_k) - \ell^{k-1}(\bar{z}_{k-1}) \right\} + \frac{\mathbb{1}(a \in \mathcal{S}'_1)}{\pi^0(1 + \theta^0)} \ell^1(z_1).$$

- Li and Luedtke (2021) showed that the efficient influence function is

$$D_{\mathrm{SC}}(\ell, \theta, \pi, r) : o \mapsto \mathcal{T}(\ell, \theta, \pi)(o) - \frac{\mathbb{1}(a \in \mathcal{S}'_1)}{\pi^0(1 + \theta^0)} r.$$

In other words, an efficient estimator $\hat{r}$ must satisfy

$$\hat{r} = r_* + \frac{1}{n} \sum_{i=1}^{n} D_{\mathrm{SC}}(\ell_*, \theta_*, \pi_*, r_*)(O_i) + \mathrm{o}_p(n^{-1/2}).$$

Cross-fit risk estimator

1: Randomly split data into V folds with index sets I_v $(v = 1, \ldots, V)$.
2: **for** $v = 1, \ldots, V$ **do**
3: For all $k = 1, \ldots, K-1$, estimate θ^k by $\hat{\theta}_v^k$ using data out of fold v
4: Set $\hat{\pi}_v^a := |I_v|^{-1} \sum_{i \in I_v} \mathbb{1}(A_i = a)$ for all $a \in \mathcal{A}$
5: **for** $k = K-1, \ldots, 1$ **do** ▷ Sequential regression
6: Estimate ℓ_*^k by $\hat{\ell}_v^k$ using data out of fold v by regressing $\hat{\ell}_v^{k+1}(\bar{Z}_{k+1})$ on covariate $\bar{Z}_k$ in the subsample $A \in \mathcal{S}'_{k+1}$.
7: Estimator of r_* for fold v:

$$\hat{r}_v := \frac{1}{|I_v|} \sum_{i \in I_v} \mathcal{T}(\widehat{\ell}_v, \widehat{\theta}_v, \widehat{\pi}_v)(O_i)$$

8: Cross-fit estimator $\hat{r} := \frac{1}{n} \sum_{v=1}^{V} |I_v| \hat{r}_v$.

Efficiency and multiple robustness of cross-fit estimator

Define oracle estimator h^{k-1} of ℓ_*^{k-1} based on $\hat{\ell}_v^k$, evaluated under the true distribution P_*:

$$h_v^{k-1} : \bar{z}_{k-1} \mapsto \mathbb{E}_{P_*}[\hat{\ell}_v^k(\bar{Z}_k) \mid \bar{Z}_{k-1} = \bar{z}_{k-1}, A \in \mathcal{S}_k'].$$

Theorem (Informal)

- *(Efficiency) If, for all v and all k, $\|\frac{1}{1+\hat{\theta}_v^k} - \frac{1}{1+\theta_*^k}\|$ and $\|\hat{\ell}_v^k - h_v^k\|$ are both $o_p(1)$ and their product is $o_p(n^{-1/2})$, then $\hat{r}$ is efficient.*
- *(2^{K-1}-robustness) If, for all v and all k, $\|\frac{1}{1+\hat{\theta}_v^k} - \frac{1}{1+\theta_*^k}\|$ or $\|\hat{\ell}_v^k - h_v^k\|$ is $o_p(1)$, then $\hat{r}$ is consistent.*

What if "sequential conditionals" does not hold?

Define

$$\Delta_v := \frac{\sum_{a \in \mathcal{S}_1'} \pi_*^a}{\sum_{a \in \mathcal{S}_1'} \hat{\pi}_v^a} \sum_{k=1}^{K} \mathbb{E}_{P_*} \left[h_v^{k-1}(\bar{Z}_{k-1}) - \hat{\ell}_v^k(\bar{Z}_k) \mid A = 0 \right]$$

and $\Delta := n^{-1} \sum_{v=1}^{V} |I_v| \Delta_v$ (average of Δ_v over folds).

- Both Δ_v and Δ are zero under "sequential conditionals".
- Δ is the bias of $\hat{r}$ due to failure of "sequential conditionals".
- If $\hat{\ell}_v^k$ or $1/(1 + \hat{\theta}_v^k)$ is consistent, $\hat{r} - \Delta$ is consistent for r_*

Sanity check: test of consistency

Since we have a straightforward but imprecise estimator $\hat{r}_{\mathrm{np}}$ of r_*, we can use $\hat{r}_{\mathrm{np}}$ as an anchor to test whether $\hat{r}$ is consistent for r_*.

Sanity check: test of consistency

Since we have a straightforward but imprecise estimator $\hat{r}_{\mathrm{np}}$ of r_*, we can use $\hat{r}_{\mathrm{np}}$ as an anchor to test whether $\hat{r}$ is consistent for r_*.

If the nuisance function estimators converge sufficiently fast (product rate $\mathrm{o}_p(n^{-1/2})$) and "sequential conditionals" holds, then

$$\sqrt{n}(\hat{r} - \hat{r}_{\mathrm{np}}) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2_{*,\mathrm{np}} - \sigma^2_{*,\mathrm{SC}}\right).$$

Sanity check: test of consistency

Since we have a straightforward but imprecise estimator $\hat{r}_{\mathrm{np}}$ of r_*, we can use $\hat{r}_{\mathrm{np}}$ as an anchor to test whether $\hat{r}$ is consistent for r_*.

If the nuisance function estimators converge sufficiently fast (product rate $\mathrm{o}_p(n^{-1/2})$) and "sequential conditionals" holds, then

$$\sqrt{n}(\hat{r} - \hat{r}_{\mathrm{np}}) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2_{*,\mathrm{np}} - \sigma^2_{*,\mathrm{SC}}\right).$$

After computing the estimators $\hat{r}_{\mathrm{np}}$ and $\hat{r}$ with respective standard errors SE_1 and SE_2, we can immediately compute the test statistic

$$\frac{\hat{r} - \hat{r}_{\mathrm{np}}}{(\mathrm{SE}_1^2 - \mathrm{SE}_2^2)^{1/2}},$$

which is approximately $\mathrm{N}(0, 1)$ if $\hat{r}$ is consistent for r_*.

Full-data covariate shift: notations

- Full-data covariate shift: $Y \perp\!\!\!\perp A \mid X$.
- Define conditional mean loss

$$\mathcal{L}_* : x \mapsto \mathbb{E}_{P_*}[\ell(X, Y) \mid X = x]$$

and propensity score for target population

$$g_* : x \mapsto P_*(A = 0 \mid X = x).$$

Full-data covariate shift: efficiency bound and gain

The efficient influence function is

$$D_{\mathrm{cov}}(\rho, g, \mathcal{L}, r) : o \mapsto \frac{g(x)}{\rho}\{\ell(x, y) - \mathcal{L}(x)\} + \frac{\mathbb{1}(a = 0)}{\rho}\{\mathcal{L}(x) - r\}.$$

The relative efficiency gain from using an efficient estimator vs $\hat{r}_{\mathrm{np}}$ is

$$1 - \frac{\text{efficient asymptotic variance}}{\text{asymptotic variance of } \hat{r}_{\mathrm{np}}}$$
$$= \frac{\mathbb{E}\left[g_*(X)(1 - g_*(X))\mathbb{E}_{P_*}\left[\{\ell(X, Y) - \mathcal{L}_*(X)\}^2 \mid X\right]\right]}{\mathbb{E}_{P_*}\left[g_*(X)\mathbb{E}_{P_*}[\{\ell(X, Y) - \mathcal{L}_*(X)\}^2 \mid X]\right] + \mathbb{E}_{P_*}\left[g_*(X)\{\mathcal{L}_*(X) - r_*\}^2\right]}$$

- Variability of $\ell(X, Y)$ due to X
- Variability of $\ell(X, Y)$ not due to X

Full-data covariate shift: efficiency bound and gain

To gain large efficiency, P_* should satisfy:

1. g_* is small, i.e., limited data from target population
2. Variability of $\ell(X, Y)$ not due to X is large compared to variability of $\ell(X, Y)$ due to X

Item 2 is the opposite of the case under concept shift in the features.

Full-data covariate shift: cross-fit estimator

- We use a similar cross-fit estimator $\hat{r}_{\mathrm{cov}}$ involving out-of-fold estimators $\hat{\mathcal{L}}^{-v}$ of $\mathcal{L}_*$ and $\hat{g}^{-v}$ of g_*.
- Asymptotic results similar to the general "sequential conditionals", in contrast to concept shift:
  - $\hat{r}_{\mathrm{cov}}$ is efficient if both $\hat{\mathcal{L}}^{-v}$ and $\hat{g}^{-v}$ are consistent with product rate $\mathrm{o}_p(n^{-1/2})$
  - $\hat{r}_{\mathrm{cov}}$ is consistent if $\hat{\mathcal{L}}^{-v}$ or $\hat{g}^{-v}$ is consistent (double robustness)

Full-data covariate shift: impossibility of efficiency & fully robust RAL

Lemma

Under the parameterization $(P_X, P_{A|X}, P_{Y|X})$ *of a distribution* P*, suppose that* $\mathrm{IF}(P_{*,X}, P_{*,A|X}, P_{*,Y|X}, r_*)$ *is an influence function for estimating* r_* *at* P_**, and so is* $\mathrm{IF}(P_{*,X}, P_{A|X}, P_{Y|X}, r_*)$*, for arbitrary* $(P_{A|X}, P_{Y|X})$*. Then,* $\mathrm{IF}(P_{*,X}, P_{A|X}, P_{Y|X}, r_*)$ *equals the influence function of* $\hat{r}_{\mathrm{np}}$*.*

Interpretation: if an estimator $\hat{r}'$ of r_* is regular and asymptotically linear even if both $P_{A|X}$ and $P_{Y|X}$ are misspecified, then $\hat{r}'$ must be asymptotically equivalent to $\hat{r}_{\mathrm{np}}$ and thus achieve no efficiency gain.

Full-data covariate shift: impossibility of efficiency & fully robust RAL

Lemma

Under the parameterization $(P_X, P_{A|X}, P_{Y|X})$ *of a distribution* P*, suppose that* $\mathrm{IF}(P_{*,X}, P_{*,A|X}, P_{*,Y|X}, r_*)$ *is an influence function for estimating* r_* *at* P_**, and so is* $\mathrm{IF}(P_{*,X}, P_{A|X}, P_{Y|X}, r_*)$*, for arbitrary* $(P_{A|X}, P_{Y|X})$*. Then,* $\mathrm{IF}(P_{*,X}, P_{A|X}, P_{Y|X}, r_*)$ *equals the influence function of* $\hat{r}_{\mathrm{np}}$*.*

Interpretation: if an estimator $\hat{r}'$ of r_* is regular and asymptotically linear even if both $P_{A|X}$ and $P_{Y|X}$ are misspecified, then $\hat{r}'$ must be asymptotically equivalent to $\hat{r}_{\mathrm{np}}$ and thus achieve no efficiency gain.

The same holds under the parameterization $(P_A, P_{X|A}, P_{Y|X})$.

Full-data covariate shift: simulation

Estimate MSE in five scenarios ($\rho_* = 0.1$):

(A) Very large efficiency gain: $f = \mu_*$ and large $\mathrm{Var}(\epsilon)$

(B) Large efficiency gain: $f \approx \mu_*$

(C) Little efficiency gain: f far from μ_*

(D) No efficiency gain: f far from μ_* and no noise ($\epsilon = 0$)

(E) Covariate shift does not hold: $Y \not\perp\!\!\!\perp A \mid X$

Four estimators:

- `np`: nonparametric estimator $\hat{r}_{\mathrm{np}}$
- `covshift`: $\hat{r}_{\mathrm{cov}}$ with consistent nuisance function estimators
- `covshift.mis.L`: $\hat{r}_{\mathrm{Xcon}}$ with inconsistent $\hat{\mathcal{L}}^{-v}$
- `covshift.mis.g`: $\hat{r}_{\mathrm{Xcon}}$ with inconsistent $\hat{g}^{-v}$

Full-data covariate shift: simulation