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

Hongxiang (David) Qiu

Department of Statistics, the Wharton School, University of Pennsylvania

Department of Statistics, Iowa State University

May 11, 2023

Table of Contents

Motivation

- Statistical machine learning is increasingly popular and successful.
- A common challenge: limited data available from the target domain/ population, despite existing large related source data sets.[1]

[1] I will use these colors to highlight source and target population throughout

Motivation

- Statistical machine learning is increasingly popular and successful.
- A common challenge: limited data available from the target domain/ population, despite existing large related source data sets.[1]
- In principle, it might be valid to use target population data alone, but desirable to leverage relevant source data to *increase efficiency/ accuracy*.

[1] I will use these colors to highlight source and target population throughout

Motivation

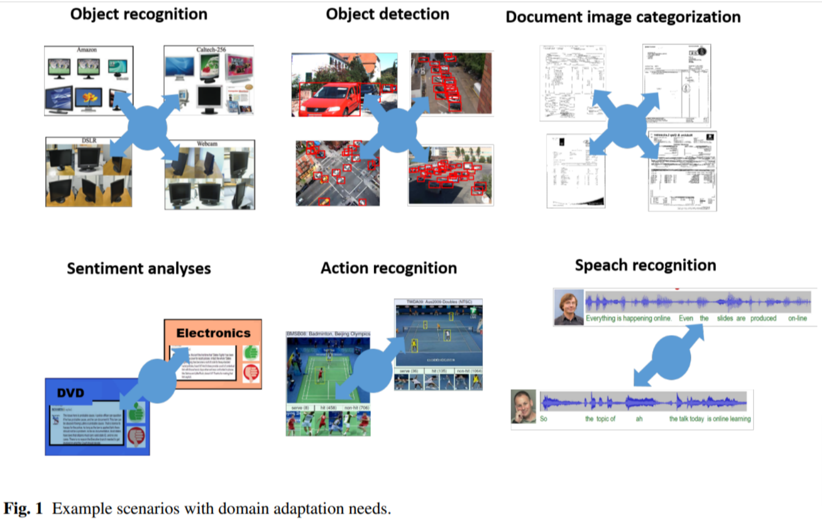


Fig. 1 Example scenarios with domain adaptation needs.

Figure: Csurka (2017)

Example: Large image datasets in certain domains are available, but they are not necessarily representative of the target domain of interest, e.g., forensics.

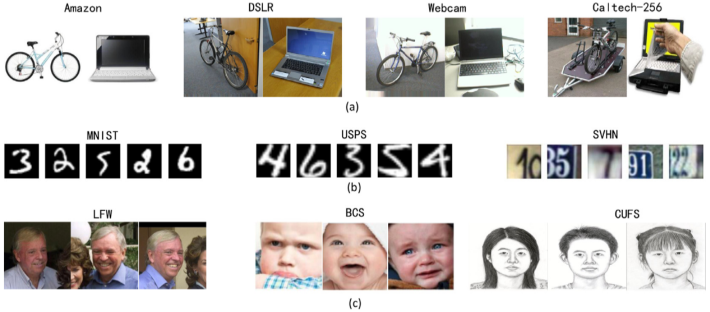


Fig. 1. (a) Some object images from the "Bike" and "Laptop" categories in Amazon, DSLR, Webcam, and Caltech-256 databases. (b) Some digit images from MNIST, USPS, and SVHN databases. (c) Some face images from LFW, BCS and CUFS databases. Realworld computer vision applications, such as face recognition, must learn to adapt to distributions specific to each domain.

Figure: Wang and Deng (2018)

Motivation

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

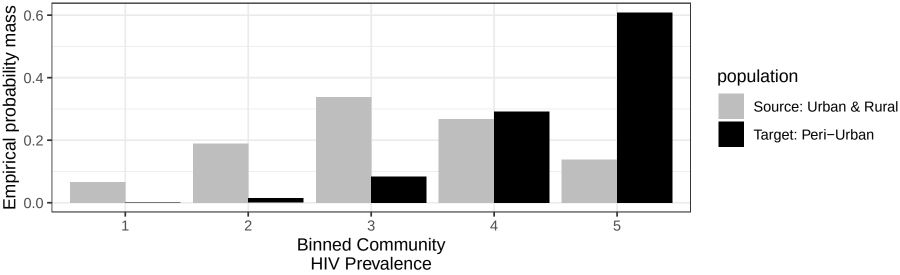
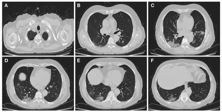


Figure: Qiu et al. (2022c)

Motivation

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



Al-Shudifat et al. (2022)

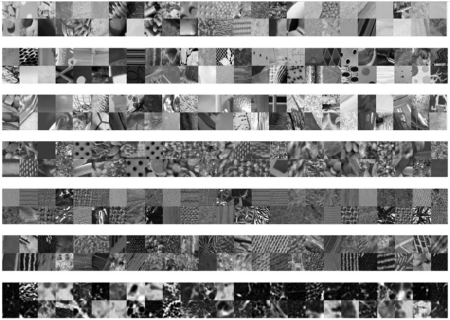


Fig. 1. Typical samples from each dataset. The color databases were converted to gray scale. From top to bottom: ALOT, DTD, FMD, KTB, KTH-TIPS-2b, UIUC, ILD.

Christodoulidis et al. (2017)

Motivation

We study the estimation of a target population risk:

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

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

Motivation

We study the estimation of a target population risk:

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

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

Risk has a central role in training prediction/classification models.

Motivation

We study the estimation of a target population risk:

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

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

Risk has a central role in training prediction/classification models.

- We often minimize the risk when training a model and evaluate the performance of a model by its risk.

Motivation

We study the estimation of a target population risk:

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

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

Risk has a central role in training prediction/classification models.

- We often minimize the risk when training a model and evaluate the performance of a model by its risk.
- To construct prediction sets with coverage guarantees and small sizes, we often need to estimate the coverage error (a risk) precisely (Vovk, 2013; Qiu et al., 2022c; Yang et al., 2022).

Motivation

We study the estimation of a target population risk:

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

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

Risk has a central role in training prediction/classification models.

- We often minimize the risk when training a model and evaluate the performance of a model by its risk.
- To construct prediction sets with coverage guarantees and small sizes, we often need to estimate the coverage error (a risk) precisely (Vovk, 2013; Qiu et al., 2022c; Yang et al., 2022).
- "Risk" and "loss" can be interpreted broadly:
    - To estimate the target population mean, take "loss" ℓ to be identity

Motivation

- Multiple valid methods to leverage source data under a dataset shift condition

Motivation

- Multiple valid methods to leverage source data under a dataset shift condition
- Which method is *efficient*?

Motivation

- Multiple valid methods to leverage source data under a dataset shift condition
- Which method is *efficient*?
- Can we also achieve *robustness* or *multiple robustness*?

Motivation

- Multiple valid methods to leverage source data under a dataset shift condition
- Which method is *efficient*?
- Can we also achieve *robustness* or *multiple robustness*?

We take the perspective of modern semiparametric efficiency theory (Bickel and Doksum, 2015; Pfanzagl, 1985, 1990; van der Vaart, 1998).

Dataset shift conditions can often be formulated as *restrictions on the observed data generating mechanism*, yielding a semiparametric model.

Related works

- Vast amount of literature on [transfer learning]/[domain adaptation]/[dataset shift], but most papers study the case where target population data is *not fully observed*, a different scenario.

Related works

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

  Particular problems under a particular type of dataset shift.

Related works

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

  Particular problems under a particular type of dataset shift.
- Another related area is *data fusion* with an emphasis on causal inference applications (Chatterjee et al. (2016) JASA, Li and Luedtke (2021) Biometrika, Robins et al. (1995) JRSS-B). Target population data might not be fully observed.

Related works

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

  Particular problems under a particular type of dataset shift.
- Another related area is *data fusion* with an emphasis on causal inference applications (Chatterjee et al. (2016) JASA, Li and Luedtke (2021) Biometrika, Robins et al. (1995) JRSS-B). Target population data might not be fully observed.
- A general framework for efficient risk estimation under general forms of dataset shift is lacking.

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: $r_* := \mathbb{E}_{P_*}[\ell(Z) \mid A = 0]$.
- An "obvious" estimator is the average over the target population data:

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

but it may be inaccurate with limited target population data.

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 $k = 1, \ldots, K$

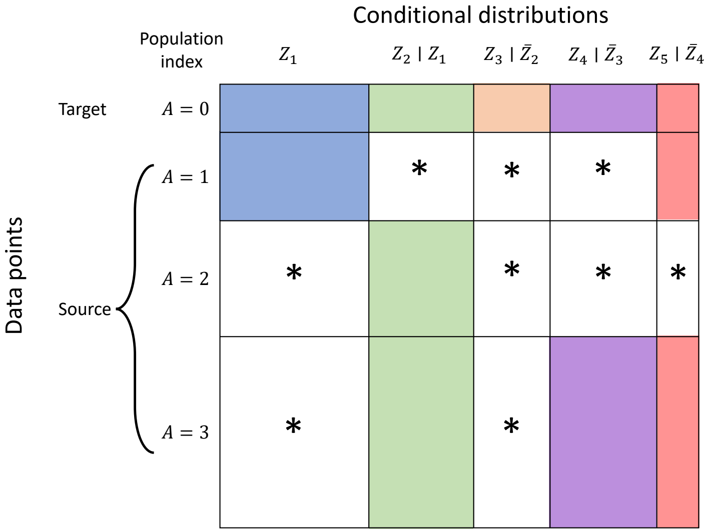
Condition (Sequential conditionals)

For every k, there exists a known (possibly empty) set $\mathcal{S}_k \subset \mathcal{A} \setminus \{0\}$ such that, for all $a \in \mathcal{S}_k$,

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

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

A general dataset shift condition



Common conditions are special cases of "sequential conditionals"

This "sequential conditionals" condition includes the four most common dataset shift conditions (Moreno-Torres et al., 2012) as special cases.

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

Common conditions are special cases of "sequential conditionals"

This "sequential conditionals" condition includes the four most common dataset shift conditions (Moreno-Torres et al., 2012) as special cases.

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

- Concept shift in the features: $\{X \mid A = 1\} \overset{d}{=} \{X \mid A = 0\}$; $Y \mid X$ may differ between source and target populations.

  Example (two-phase sampling/semi-supervised learning): In a sample from the target population, a random subset is labeled (Y observed); the others are unlabeled (Y missing)

Common conditions are special cases of "sequential conditionals"

This "sequential conditionals" condition includes the four most common dataset shift conditions (Moreno-Torres et al., 2012) as special cases.

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

- Concept shift in the features: $\{X \mid A = 1\} \overset{d}{=} \{X \mid A = 0\}$; $Y \mid X$ may differ between source and target populations.

  Example (two-phase sampling/semi-supervised learning): In a sample from the target population, a random subset is labeled (Y observed); the others are unlabeled (Y missing)

- Concept shift in the labels: $\{Y \mid A = 1\} \overset{d}{=} \{Y \mid A = 0\}$

Common conditions are special cases of "sequential conditionals"

- 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.

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

Common conditions are special cases of "sequential conditionals"

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

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

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

  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.
  Other outcome-dependent sampling schemes might satisfy label shift.

A general dataset shift condition

More sophisticated examples:

- Covariate & concept shift: Three available data sets:
  - labeled target population data ($A = 0$)
  - unlabeled target population data ($A = 1$)
  - labeled data from another population satisfying covariate shift ($A = 2$)

A general dataset shift condition

More sophisticated examples:

- Covariate & concept shift: Three available data sets:
  - labeled target population data ($A = 0$)
  - unlabeled target population data ($A = 1$)
  - labeled data from another population satisfying covariate shift ($A = 2$)
- 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, might wish to leverage a large texture dataset containing (X_1, X_2) and assume

$$\{X_2 \mid X_1, A = 1\} \overset{d}{=} \{X_2 \mid X_1, A = 0\}$$

Efficiency bound: 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 ($A = 0$)
- Unlabeled CT scans from target population ($A = 1$)
- Large texture dataset ($A = 2$)
- Fully labeled CT scans from another population ($A = 3$)

Efficiency bound: 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 ($A = 0$)
- Unlabeled CT scans from target population ($A = 1$)
- Large texture dataset ($A = 2$)
- Fully labeled CT scans from another population ($A = 3$)

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:

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

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

- Conditional mean loss (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:

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

- 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:

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

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

Results in Li and Luedtke (2021) imply the efficient influence function

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

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

$$
\hat{r} = r_* + \frac{1}{n} \sum_{i=1}^{n} D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*)(O_i) + \mathrm{o}_p(n^{-1/2}).
$$

Efficient and multiply robust estimation

- Can we construct such an efficient estimator $\hat{r}$?

Efficient and multiply robust estimation

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

- Can we construct such an efficient estimator $\hat{r}$?
- Can we make $\hat{r}$ multiply robust against inconsistent estimation of some of the nuisance functions ℓ_* and $\boldsymbol{\theta}_*$?

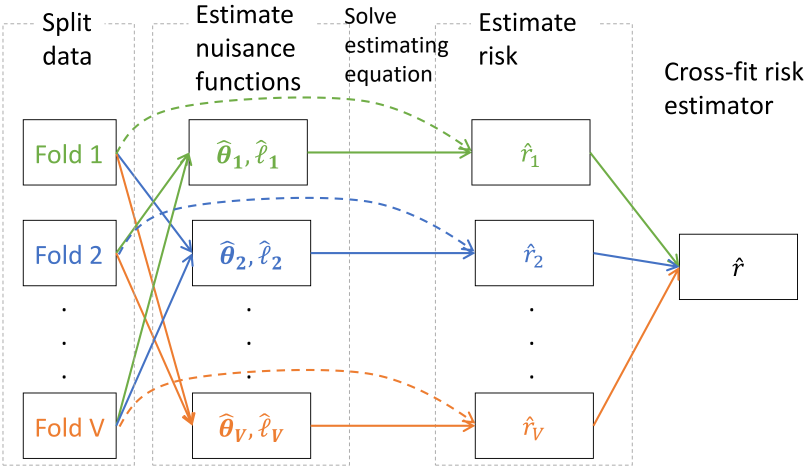
Yes!

We use the estimating equation approach (Bolthausen et al., 2002), essentially solving

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

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

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

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

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

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

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

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

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

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

  is $\mathrm{o}_p(1)$, then $\hat{r}$ is consistent.

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 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

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

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.
- 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

- From now on, $Z = (X, Y)$ and $A \in \{0, 1\}$.
- Concept shift in the features: $\{X \mid A = 1\} \overset{d}{=} \{X \mid A = 0\}$
- 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

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

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

Efficiency bound and gain

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

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

More on item 2 in MSE estimation example:

- $\ell(x, y) = (y - f(x))^2$ for a given predictor f
- $Y = \mu_*(X) + \epsilon$ 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 large efficiency for f far from the truth μ_* (heterogeneously)
- An extension of results in Azriel et al. (2021) (linear regression under concept shift) to general risk estimation problem

Efficiency & fully robust regularity and asymptotic linearity

- The cross-fit estimator $\hat{r}_{\mathrm{Xcon}}$ follows from "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 $\hat{r}_{\mathrm{Xcon}}$ follows from "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:

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

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

Efficiency & fully robust regularity and asymptotic linearity

More desirable properties than under "sequential conditionals":

- 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_*
  - inference about r_* is crucial for constructing prediction sets with training-set conditional coverage (Bates et al., 2021; Qiu et al., 2022c)

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_*

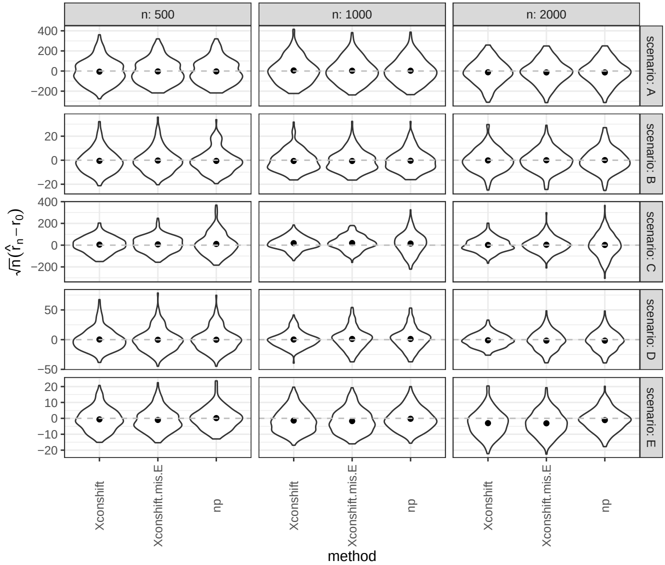
(D) Very large efficiency gain: f far from μ_* and no noise ($\epsilon = 0$)

(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 $\hat{r}_{\mathrm{np}}$
- `Xconshift`: $\hat{r}_{\mathrm{Xcon}}$ with consistent $\hat{\mathcal{E}}^{-v}$
- `Xconshift,mis.E`: $\hat{r}_{\mathrm{Xcon}}$ with inconsistent $\hat{\mathcal{E}}^{-v}$

Simulation



n: 500
n: 1000
n: 2000
scenario: A
scenario: B
scenario: C
scenario: D
scenario: E
400
200
0
−200
20
0
−20
400
200
0
−200
50
0
−50
20
10
0
−10
−20
$\sqrt{n}(\hat{r}_n - r_0)$
Xconshift
Xconshift.mis.E
np
method

Other common dataset shift conditions

We studied the other three most common dataset shift conditions:

- Concept shift in the labels ($\approx$ concept shift in the features)
- Full-data covariate/label shift ($\approx$ "sequential conditionals")

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

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, etc.
- Target population: peri-urban communities with community 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

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

Table: Risk estimates from HIV risk prediction data. The risk estimate from the validation dataset 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)

Discussion

- We also characterized efficiency bounds for several other widely applicable dataset shift conditions (Scott, 2018; Tasche, 2017; Zhang et al., 2013)
  - These efficient influence functions are hardly tractable
  - Challenging to construct efficient estimators
- Possible to construct efficient and multiply robust plug-in estimators, using targeted minimum-loss based estimation (TMLE) (Van der Laan and Rose, 2018), so that the estimator always satisfies known bounds on the true risk r_*.

Collaborators

Edgar Dobriban

Eric Tchetgen Tchetgen

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. J. Bickel and K. A. Doksum. *Mathematical statistics: Basic ideas and selected topics, second edition*, volume 1. Chapman and Hall/CRC, 2015. ISBN 9781498723817. doi: 10.1201/b18312.

References

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.

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.

References

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.

J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

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 and A. Luedtke. Adversarial Meta-Learning of Gamma-Minimax Estimators That Leverage Prior Knowledge. *arXiv preprint arXiv:2012.05465v4*, pages 1–45, 2023.

References

H. Qiu, M. Carone, E. Sadikova, M. Petukhova, R. C. Kessler, and A. Luedtke. Optimal Individualized Decision Rules Using Instrumental Variable Methods. *Journal of the American Statistical Association*, 116(533):174–191, 2021a. ISSN 1537274X. doi: 10.1080/01621459.2020.1745814.

H. Qiu, A. Luedtke, and M. Carone. Universal sieve-based strategies for efficient estimation using machine learning tools. *Bernoulli*, 27(4):2300–2336, 2021b. ISSN 13507265. doi: 10.3150/20-BEJ1309. URL https://arxiv.org/abs/2003.01856.

H. Qiu, M. Carone, and A. Luedtke. Individualized treatment rules under stochastic treatment cost constraints. *Journal of Causal Inference*, 10(1):480–493, dec 2022a. ISSN 2193-3685. doi: 10.1515/jci-2022-0005.

H. Qiu, A. J. Cook, and J. F. Bobb. Evaluating tests for cluster-randomized trials with few clusters under generalized linear mixed models with covariate adjustment: a simulation study. *arXiv preprint arXiv:2209.04364v1*, 2022b. doi: 10.48550/arxiv.2209.04364.

References

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

H. Qiu, X. Shi, W. Miao, E. Dobriban, and E. Tchetgen Tchetgen. Doubly Robust Proximal Synthetic Controls. *arXiv preprint arXiv:2210.02014*, 2022d. doi: 10.48550/arxiv.2210.02014.

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.

C. Scott. A Generalized Neyman-Pearson Criterion for Optimal Domain Adaptation. *Proceedings of Machine Learning Research*, 98:1–24, 2018.

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.

D. Tasche. Fisher consistency for prior probability shift. *Journal of Machine Learning Research*, 18:1–32, 2017. ISSN 15337928.

M. J. Van der Laan and S. Rose. *Targeted learning in data science: causal inference for complex longitudinal studies*. Springer, 2018.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. doi: 10.1017/cbo9780511802256.

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.

M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. ISSN 18728286. doi: 10.1016/j.neucom.2018.05.083.

References

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.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. PMLR, 2013.

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.

Overview of my research

I am interested in a variety of areas.

- Prediction sets: Qiu et al. (2022c)
- Causal inference & individualized treatment rules: Qiu et al. (2022d), Qiu et al. (2021a) JASA, Qiu et al. (2022a)
- Estimation under semi-/non-parametric models: Qiu et al. (2021b); Qiu and Luedtke (2023)
- Cluster-randomized trials: Qiu et al. (2022b)

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):

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

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

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

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

$$\hat{r} = r_* + \frac{1}{n} \sum_{i=1}^{n} D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*)(O_i) + \mathrm{o}_p(n^{-1/2}).$$

Cross-fit risk estimator

Randomly split data into V folds with index sets I_v $(v = 1, \ldots, V)$.

for $v = 1, \ldots, V$ **do**

  For all $k = 1, \ldots, K-1$, estimate θ^k by $\hat{\theta}_v^k$ using data out of fold v

  Set $\hat{\pi}_v^a := |I_v|^{-1} \sum_{i \in I_v} \mathbb{1}(A_i = a)$ for all $a \in \mathcal{A}$

  for $k = K-1, \ldots, 1$ **do** ▷ Sequential regression

    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 $A \in \mathcal{S}'_{k+1}$.

  Estimator of r_* for fold v:

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

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 $\mathrm{o}_p(1)$ and their product is $\mathrm{o}_p(n^{-1/2})$, 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 $\mathrm{o}_p(1)$, then $\hat{r}$ is consistent.*

What if "sequential conditionals" condition fails?

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_*
- A trade-off between efficiency and robustness.

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