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

Hongxiang (David) Qiu

Department of Epidemiology and Biostatistics, Michigan State University

2026 ICSA Applied Statistics Symposium

Table of Contents

1. Motivation
2. A general dataset shift condition
3. Efficient and multiply robust estimation
4. Data analysis

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 may be valid to use target population data alone, but it is desirable to leverage relevant source data to *increase efficiency/accuracy*.

[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 may be valid to use target population data alone, but it is desirable to leverage relevant source data to *increase efficiency/accuracy*.
- Challenge: Dataset shift, source and target populations differ

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

Motivation: Comparative Effectiveness

Example: Multiphase sampling (Chakrabortty and Cai, 2018)

1. Draw a large sample D_1 from the target population

Motivation: Comparative Effectiveness

Example: Multiphase sampling (Chakrabortty and Cai, 2018)

1. Draw a large sample D_1 from the target population
2. Measure cheap variables Z_1 in electronic health records (EHR) for all individuals in D_1
   E.g., age, sex, diagnosis of rheumatoid arthritis (RA) (a systemic auto-immune disease), medical history

Motivation: Comparative Effectiveness

Example: Multiphase sampling (Chakrabortty and Cai, 2018)

1. Draw a large sample D_1 from the target population
2. Measure cheap variables Z_1 in electronic health records (EHR) for all individuals in D_1
   E.g., age, sex, diagnosis of rheumatoid arthritis (RA) (a systemic auto-immune disease), medical history
3. Measure expensive variable Y for a random subsample $D_2 \subset D_1$
   E.g., biomarker anti-CCP

Motivation: Comparative Effectiveness

Example: Multiphase sampling (Chakrabortty and Cai, 2018)

1. Draw a large sample D_1 from the target population
2. Measure cheap variables Z_1 in electronic health records (EHR) for all individuals in D_1
   E.g., age, sex, diagnosis of rheumatoid arthritis (RA) (a systemic auto-immune disease), medical history
3. Measure expensive variable Y for a random subsample $D_2 \subset D_1$
   E.g., biomarker anti-CCP
4. Wish to study the association between the outcome Y (only observed in D_2) and some clinical variables X of Z_1

Motivation: Comparative Effectiveness

Example: Multiphase sampling (Chakrabortty and Cai, 2018)

1. Draw a large sample D_1 from the target population
2. Measure cheap variables Z_1 in electronic health records (EHR) for all individuals in D_1
   E.g., age, sex, diagnosis of rheumatoid arthritis (RA) (a systemic auto-immune disease), medical history
3. Measure expensive variable Y for a random subsample $D_2 \subset D_1$
   E.g., biomarker anti-CCP
4. Wish to study the association between the outcome Y (only observed in D_2) and some clinical variables X of Z_1

Considering *observed* data, an equivalent formulation in terms of dataset shift:

- Target population data: D_2: both Z_1 and Y are observed
- Source population data: $D_1 \setminus D_2$: Z_1 observed, Y missing

Motivation: HIV Epidemiology

Example: To improve HIV treatment/prevention, wish to predict HIV risk in peri-urban communities with low community antiretroviral therapy (ART) coverage to identify people with high risk. Wish to leverage data from urban & rural communities to improve prediction accuracy.

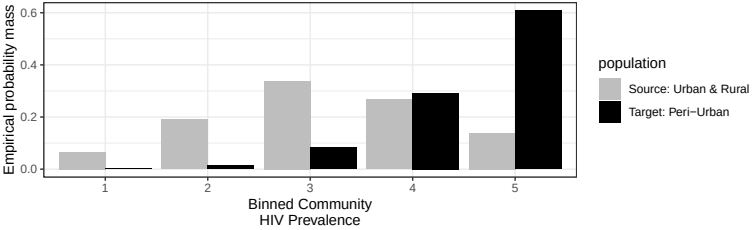


Figure: Qiu et al. (2022)

This is the main example for the rest of this talk.

Motivation: How well does a given predictor perform in the target population?

We study the estimation of a target population risk:

$$\mathbb{E}[\underbrace{\ell(Z)}_{\text{user-specified loss}} \mid \text{target population}]$$

Motivation: How well does a given predictor perform in the target population?

We study the estimation of a target population risk:

$$\mathbb{E}[\underbrace{\ell(Z)}_{\text{user-specified loss}} \mid \text{target population}]$$

Example: $Z = (X, Y)$, given predictor f

- Mean squared error: $\ell(Z) = (Y - f(X))^2$
- Cross-entropy loss (negative Bernoulli log-likelihood): $\ell(Z) = -Y \log(f(X)) - (1 - Y) \log(1 - f(X))$
- Classification error: $\ell(Z) = \mathbb{1}(Y \neq f(X))$
- Prediction set coverage error: $\ell(Z) = \mathbb{1}(Y \notin f(X))$

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.

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, we often want to estimate the coverage error (a risk) precisely (e.g., Vovk, 2013; Qiu et al., 2022; Yang et al., 2022).

Motivation

- For source data to be helpful, source and target populations need to be related.

Motivation

- For source data to be helpful, source and target populations need to be related.
- Partly motivated by causal inference and data fusion, we consider a general *dataset shift condition*.

Motivation

- For source data to be helpful, source and target populations need to be related.
- Partly motivated by causal inference and data fusion, we consider a general *dataset shift condition*.
- Multiple valid methods to leverage source data. Which method is *efficient* (asymptotically normal, smallest asymptotic variance)?

Motivation

- For source data to be helpful, source and target populations need to be related.
- Partly motivated by causal inference and data fusion, we consider a general *dataset shift condition*.
- Multiple valid methods to leverage source data. Which method is *efficient* (asymptotically normal, smallest asymptotic variance)?
- More precise risk estimators $\Longrightarrow$ better distinction between predictors.

Motivation

- For source data to be helpful, source and target populations need to be related.
- Partly motivated by causal inference and data fusion, we consider a general *dataset shift condition*.
- Multiple valid methods to leverage source data. Which method is *efficient* (asymptotically normal, smallest asymptotic variance)?
- More precise risk estimators $\Longrightarrow$ better distinction between predictors.
- Can we also achieve *robustness* or *multiple robustness* (against misspecification of some nuisance functions)?

Motivation

- For source data to be helpful, source and target populations need to be related.
- Partly motivated by causal inference and data fusion, we consider a general *dataset shift condition*.
- Multiple valid methods to leverage source data. Which method is *efficient* (asymptotically normal, smallest asymptotic variance)?
- More precise risk estimators $\Longrightarrow$ better distinction between predictors.
- Can we also achieve *robustness* or *multiple robustness* (against misspecification of some nuisance functions)?
- To address these estimation questions, we rely on common tools used in causal inference—semiparametric efficiency theory.

Related works

- Rich 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

- Rich 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 a.k.a. *semi-supervised learning* (Azriel et al. (2021), Gronsbell et al. (2022), Zhang et al. (2021)).

  Particular problems under a particular type of dataset shift.

Related works

- Rich 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 a.k.a. *semi-supervised learning* (Azriel et al. (2021), Gronsbell et al. (2022), 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 (Chakrabortty and Cai, 2018; Chatterjee et al., 2016; Li and Luedtke, 2021; Robins et al., 1995). Target population data might not be fully observed.

Related works

- Rich 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 a.k.a. *semi-supervised learning* (Azriel et al. (2021), Gronsbell et al. (2022), 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 (Chakrabortty and Cai, 2018; Chatterjee et al., 2016; Li and Luedtke, 2021; Robins et al., 1995). Target population data might not be fully observed.
- A general framework for efficient and robust risk estimation under general forms of dataset shift is lacking.

Table of Contents

1. Motivation
2. A general dataset shift condition
3. Efficient and multiply robust estimation
4. Data analysis

Problem setup

- Observe i.i.d. copies of $O = (Z, A)$:
  - 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: $r_* := \mathbb{E}[\ell(Z) \mid A = 0]$.

Problem setup

- Observe i.i.d. copies of $O = (Z, A)$:
  - 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: $r_* := \mathbb{E}[\ell(Z) \mid A = 0]$.
- Without any additional assumption, a sensible estimator is the empirical mean over 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 given limited target population data, particularly with relevant source 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 adapted from Li and Luedtke (2021):

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 $a \in \mathcal{S}_k$,

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

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

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 adapted from Li and Luedtke (2021):

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 $a \in \mathcal{S}_k$,

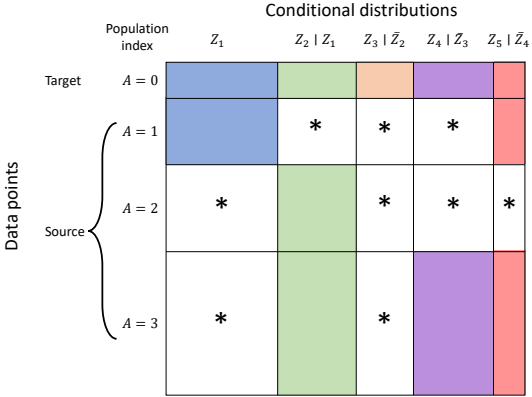
$$\left\{ Z_k \mid \bar{Z}_{k-1} = \bar{z}_{k-1}, A = a \right\} \stackrel{d}{=} \left\{ Z_k \mid \bar{Z}_{k-1} = \bar{z}_{k-1}, A = 0 \right\}$$

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

$\mathcal{S}_k$ can be selected based on prior knowledge (e.g., study design, causal mechanism). We can also test "sequential conditionals" condition (details in paper).

A general dataset shift condition

Figure: Blocks with same colors are shared conditional distributions. Blocks with * are not assumed to share same conditional distributions.



Common dataset shift conditions are special cases of "sequential conditionals"

- Full-data covariate shift: $\{Y \mid X, A = 1\} \stackrel{d}{=} \{Y \mid X, A = 0\}$ (similar to unconfoundedness/ignorability; covariate-dependent sampling)

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

Common dataset shift conditions are special cases of "sequential conditionals"

- Full-data covariate shift: $\{Y \mid X, A = 1\} \stackrel{d}{=} \{Y \mid X, A = 0\}$ (similar to unconfoundedness/ignorability; covariate-dependent sampling)

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

- Full-data label shift: $\{X \mid Y, A = 1\} \stackrel{d}{=} \{X \mid Y, A = 0\}$ (anti-causal; outcome-dependent sampling)

  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.

Common dataset shift conditions are special cases of "sequential conditionals"

- Full-data covariate shift: $\{Y \mid X, A = 1\} \stackrel{d}{=} \{Y \mid X, A = 0\}$ (similar to unconfoundedness/ignorability; covariate-dependent sampling)

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

- Full-data label shift: $\{X \mid Y, A = 1\} \stackrel{d}{=} \{X \mid Y, A = 0\}$ (anti-causal; outcome-dependent sampling)

  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.

- Concept shift in the features: $\{X \mid A = 1\} \stackrel{d}{=} \{X \mid A = 0\}$ (semi-supervised learning; multiphase sampling)
- Concept shift in the labels: $\{Y \mid A = 1\} \stackrel{d}{=} \{Y \mid A = 0\}$

Table of Contents

1 Motivation

2 A general dataset shift condition

3 Efficient and multiply robust estimation

4 Data analysis

How to leverage source data under sequential conditionals?

Example: Estimate prediction error of a given predictor f:

$$\Pr(Y \neq f(X) \mid A = 0) = \mathbb{E}[\underbrace{\mathbb{1}(Y \neq f(X))}_{\ell(X,Y)} \mid A = 0]$$

X = covariate (age, sex, etc.), Y = binary outcome (HIV seroconversion)

How to leverage source data under sequential conditionals?

Example: Estimate prediction error of a given predictor f:

$$\Pr(Y \neq f(X) \mid A = 0) = \mathbb{E}[\underbrace{\mathbb{1}(Y \neq f(X))}_{\ell(X,Y)} \mid A = 0]$$

X = covariate (age, sex, etc.), Y = binary outcome (HIV seroconversion)

Data sets:

- Fully observed data (X, Y) from peri-urban communities with low ART coverage $(A = 0)$
- Covariate data X with missing outcome Y from peri-urban communities with low ART coverage $(A = 1)$
- Fully observed data (X, Y) from urban & rural communities $(A = 2)$

How to leverage source data under sequential conditionals?

With $Z = (Z_1 = X, Z_2 = Y)$, relevant source data set indices $\mathcal{S}_k$

- $\mathcal{S}_1 = \{1\}$: $\{X \mid A = 1\} \stackrel{d}{=} \{X \mid A = 0\}$
  Shared covariate distribution between the fully observed data from peri-urban communities and covariate data from peri-urban communities
- $\mathcal{S}_2 = \{2\}$: $\{Y \mid X, A = 2\} \stackrel{d}{=} \{Y \mid X, A = 0\}$
  Shared distribution of HIV seroconversion given covariate between peri-urban communities and urban & rural communities

How to leverage source data under sequential conditionals?

By Law of Iterated Expectation, the risk of interest can be written as

$$r_* = \mathbb{E}[\ell(X, Y) \mid A = 0] = \mathbb{E}[\underbrace{\mathbb{E}[\ell(X, Y) \mid X, A = 0]}_{\mathcal{E}_*(X)} \mid A = 0]$$

How to leverage source data under sequential conditionals?

By Law of Iterated Expectation, the risk of interest can be written as

$$r_* = \mathbb{E}[\ell(X, Y) \mid A = 0] = \mathbb{E}[\underbrace{\mathbb{E}[\ell(X, Y) \mid X, A = 0]}_{\mathcal{E}_*(X)} \mid A = 0]$$

HIV risk prediction example: $\mathcal{E}_*(X) = \Pr(Y \neq f(X) \mid X, A = 0)$ is the prediction error of the given predictor f conditional on covariate X

How to leverage source data under sequential conditionals?

By Law of Iterated Expectation, the risk of interest can be written as

$$r_* = \mathbb{E}[\ell(X, Y) \mid A = 0] = \mathbb{E}[\underbrace{\mathbb{E}[\ell(X, Y) \mid X, A = 0]}_{\mathcal{E}_*(X)} \mid A = 0]$$

HIV risk prediction example: $\mathcal{E}_*(X) = \Pr(Y \neq f(X) \mid X, A = 0)$ is the prediction error of the given predictor f conditional on covariate X

- Inner expectation $\mathbb{E}[\ell(X, Y) \mid X, A = 0]$ concerns the conditional distribution $Y \mid X$

How to leverage source data under sequential conditionals?

By Law of Iterated Expectation, the risk of interest can be written as

$$r_* = \mathbb{E}[\ell(X, Y) \mid A = 0] = \mathbb{E}[\underbrace{\mathbb{E}[\ell(X, Y) \mid X, A = 0]}_{\mathcal{E}_*(X)} \mid A = 0]$$

HIV risk prediction example: $\mathcal{E}_*(X) = \Pr(Y \neq f(X) \mid X, A = 0)$ is the prediction error of the given predictor f conditional on covariate X

- Inner expectation $\mathbb{E}[\ell(X, Y) \mid X, A = 0]$ concerns the conditional distribution $Y \mid X$

  We can leverage data from urban & rural communities ($A = 2$) to estimate this expectation $\mathcal{E}_*$

How to leverage source data under sequential conditionals?

By Law of Iterated Expectation, the risk of interest can be written as

$$r_* = \mathbb{E}[\ell(X, Y) \mid A = 0] = \mathbb{E}[\underbrace{\mathbb{E}[\ell(X, Y) \mid X, A = 0]}_{\mathcal{E}_*(X)} \mid A = 0]$$

HIV risk prediction example: $\mathcal{E}_*(X) = \Pr(Y \neq f(X) \mid X, A = 0)$ is the prediction error of the given predictor f conditional on covariate X

- Inner expectation $\mathbb{E}[\ell(X, Y) \mid X, A = 0]$ concerns the conditional distribution $Y \mid X$

  We can leverage data from urban & rural communities ($A = 2$) to estimate this expectation $\mathcal{E}_*$

- Outer expectation $\mathbb{E}[\mathcal{E}_*(X) \mid A = 0]$ concerns the marginal distribution of covariate X

How to leverage source data under sequential conditionals?

By Law of Iterated Expectation, the risk of interest can be written as

$$r_* = \mathbb{E}[\ell(X, Y) \mid A = 0] = \mathbb{E}[\underbrace{\mathbb{E}[\ell(X, Y) \mid X, A = 0]}_{\mathcal{E}_*(X)} \mid A = 0]$$

HIV risk prediction example: $\mathcal{E}_*(X) = \Pr(Y \neq f(X) \mid X, A = 0)$ is the prediction error of the given predictor f conditional on covariate X

- Inner expectation $\mathbb{E}[\ell(X, Y) \mid X, A = 0]$ concerns the conditional distribution $Y \mid X$

  We can leverage data from urban & rural communities ($A = 2$) to estimate this expectation $\mathcal{E}_*$

- Outer expectation $\mathbb{E}[\mathcal{E}_*(X) \mid A = 0]$ concerns the marginal distribution of covariate X

  We can leverage covariate data from peri-urban communities ($A = 1$) to estimate this expectation

How to leverage source data under sequential conditionals?

$$r_* = \mathbb{E}[\underbrace{\mathbb{E}[\ell(X, Y) \mid X, A \in \{0, 2\}]}_{\mathcal{E}_*(X)} \mid A \in \{0, 1\}]$$

How to leverage source data under sequential conditionals?

$$r_* = \mathbb{E}[\underbrace{\mathbb{E}[\ell(X,Y) \mid X, A \in \{0,2\}]}_{\mathcal{E}_*(X)} \mid A \in \{0,1\}]$$

One intuitive plug-in approach:

1. Estimate $\mathcal{E}_*(X)$ with $\hat{\mathcal{E}}$ by regressing $\ell(X,Y)$ on X in the subsample with $A \in \{0,2\}$ (e.g., linear regression, neural networks, etc.)
2. Estimate risk r_* by the empirical mean of $\hat{\mathcal{E}}(X)$ in the subsample with $A \in \{0,1\}$:

$$\frac{\sum_{i=1}^n \mathbb{1}(A_i \in \{0,1\})\hat{\mathcal{E}}(X_i)}{\sum_{i=1}^n \mathbb{1}(A_i \in \{0,1\})}$$

Efficient and multiply robust estimation

- We wish to use flexible machine learning (ML) methods to estimate $\mathcal{E}_*$ (lower bias, but slower convergence rate)

Efficient and multiply robust estimation

- We wish to use flexible machine learning (ML) methods to estimate $\mathcal{E}_*$ (lower bias, but slower convergence rate)
- However, in general, such a plug-in estimator would be dominated by the slow convergence of the ML estimator and would be inefficient (McGrath and Mukherjee, 2022).

Efficient and multiply robust estimation

- We wish to use flexible machine learning (ML) methods to estimate $\mathcal{E}_*$ (lower bias, but slower convergence rate)
- However, in general, such a plug-in estimator would be dominated by the slow convergence of the ML estimator and would be inefficient (McGrath and Mukherjee, 2022).
- We have developed an estimator $\hat{r}$ based on the *efficient influence function* to address this issue

Efficient and multiply robust estimation

- We wish to use flexible machine learning (ML) methods to estimate $\mathcal{E}_*$ (lower bias, but slower convergence rate)
- However, in general, such a plug-in estimator would be dominated by the slow convergence of the ML estimator and would be inefficient (McGrath and Mukherjee, 2022).
- We have developed an estimator $\hat{r}$ based on the *efficient influence function* to address this issue
- Generally need to estimate two sets of nuisance functions flexibly:
  1. the mean loss conditional on variables $\bar{Z}_k$ (e.g., $\mathcal{E}_*$)
  2. the odds of target vs. relevant source population conditional on variables $\bar{Z}_k$ (similar to weighting)

Efficient and multiply robust estimation

Theorem (Informal)

- *(Efficiency) If conditional mean loss and conditional odds are all estimated reasonably well, then our proposed estimator $\hat{r}$ is asymptotically normal and efficient (smallest asymptotic variance)*
- *(Multiple robustness) If conditional mean loss or conditional odds is estimated consistently for every pair, then our proposed estimator $\hat{r}$ is consistent*

Formal statement: Theorem 1 in paper

Table of Contents

1. Motivation
2. A general dataset shift condition
3. Efficient and multiply robust estimation
4. Data analysis

Data analysis: HIV risk prediction under full-data covariate shift

Table: Risk estimates from HIV risk prediction data. "Gold standard": Risk estimate from large held-out validation dataset is 0.24 (95% CI: 0.22–0.26).

Dataset Shift Condition	Estimate	S.E.	95% CI	P-value
None	0.24	0.060	(0.12, 0.36)	—
Full-data covariate shift	0.19	0.026	(0.14, 0.25)	0.41

Acknowledgment

Edgar Dobriban

Eric Tchetgen Tchetgen

Paper with same title published in AoS and available on arXiv
`https://arxiv.org/abs/2306.16406`

This presentation is supported by NSF Travel Award.

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.

M. Carone, A. R. Luedtke, and M. J. van der Laan. Toward Computerized Efficient Estimation in Infinite-Dimensional Models. *Journal of the American Statistical Association*, 114(527):1174–1190, 2019. ISSN 1537274X. doi: 10.1080/01621459.2018.1482752.

A. Chakrabortty and T. Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.

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.

References

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.

S. McGrath and R. Mukherjee. On Undersmoothing and Sample Splitting for Estimating a Doubly Robust Functional. *arXiv preprint arXiv:2212.14857v1*, 2022.

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.

References

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.

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

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.

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.

“Risk” and “loss” can be interpreted broadly

- To estimate the target population mean (e.g., HIV prevalence), take “loss” $\ell(z) = \mathbb{1}(\text{HIV}+)$

"Risk" and "loss" can be interpreted broadly

- To estimate the target population mean (e.g., HIV prevalence), take "loss" $\ell(z) = \mathbb{1}(\text{HIV}+)$
- To estimate a target population quantile, consider ℓ ranging over $\{z \mapsto \mathbb{1}(z \leq t) : t \in \mathbb{R}\}$.

"Risk" and "loss" can be interpreted broadly

- To estimate the target population mean (e.g., HIV prevalence), take "loss" $\ell(z) = \mathbb{1}(\text{HIV}+)$
- To estimate a target population quantile, consider ℓ ranging over $\{z \mapsto \mathbb{1}(z \leq t) : t \in \mathbb{R}\}$.
- Any functional related to expectation may fit into our framework.

A general dataset shift condition: more sophisticated examples

- 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

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

A general dataset shift condition: more sophisticated examples

- 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

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

- HIV risk prediction example in next slides

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):

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

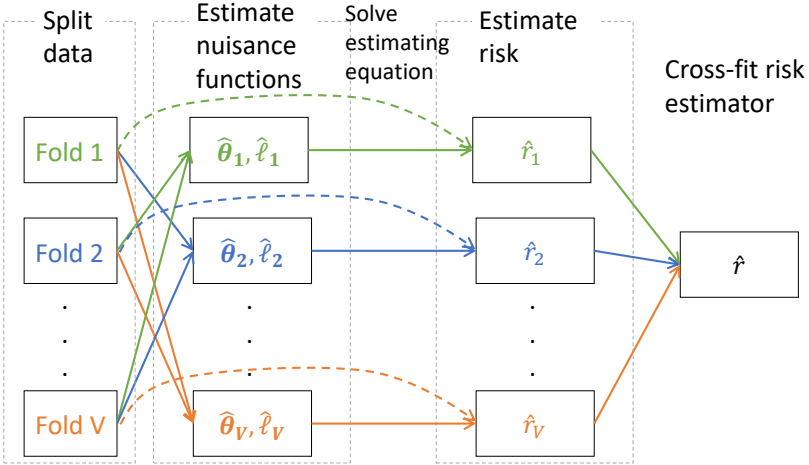
- Li and Luedtke (2021) showed that the efficient influence function is

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

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

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

Cross-fit risk estimator



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 $k \in \{1, 2\}$, 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 $k = 2, 1$ **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 $A \in \{0\} \cup \mathcal{S}_{k+1}$.
    Estimator $\hat{r}_v$ is the solution in r to: ▷ Can be solved explicitly

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

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 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

- *(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.*

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

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

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.
- This approach might 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.

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.

Concept shift: notations

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

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

  and probability of target population $\rho_* := P_*(A = 0)$.
- According to the results for "sequential conditionals", the efficient influence function is

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

Can we check whether "sequential conditionals" holds?

- The nonparametric estimator $\hat{r}_{\mathrm{np}}$ of r_* is always valid regardless of whether "sequential conditionals" holds

Can we check whether "sequential conditionals" holds?

- The nonparametric estimator $\hat{r}_{\mathrm{np}}$ of r_* is always valid regardless of whether "sequential conditionals" holds
- We can use $\hat{r}_{\mathrm{np}}$ as an anchor to test whether $\hat{r}$ is consistent for r_* or whether "sequential conditionals" holds.

Can we check whether "sequential conditionals" holds?

- The nonparametric estimator $\hat{r}_{\mathrm{np}}$ of r_* is always valid regardless of whether "sequential conditionals" holds
- We can use $\hat{r}_{\mathrm{np}}$ as an anchor to test whether $\hat{r}$ is consistent for r_* or whether "sequential conditionals" holds.
- If "sequential conditionals" holds, then

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

Can we check whether "sequential conditionals" holds?

- The nonparametric estimator $\hat{r}_{\mathrm{np}}$ of r_* is always valid regardless of whether "sequential conditionals" holds
- We can use $\hat{r}_{\mathrm{np}}$ as an anchor to test whether $\hat{r}$ is consistent for r_* or whether "sequential conditionals" holds.
- If "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 "sequential conditionals" holds.

Concept shift: efficiency bound and gain

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

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

Concept shift: 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

Concept shift: 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

Concept shift: 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

Concept shift: 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 $\mathcal{E}_*$

Concept shift: 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:

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

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

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

Simulation: concept shift

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_*

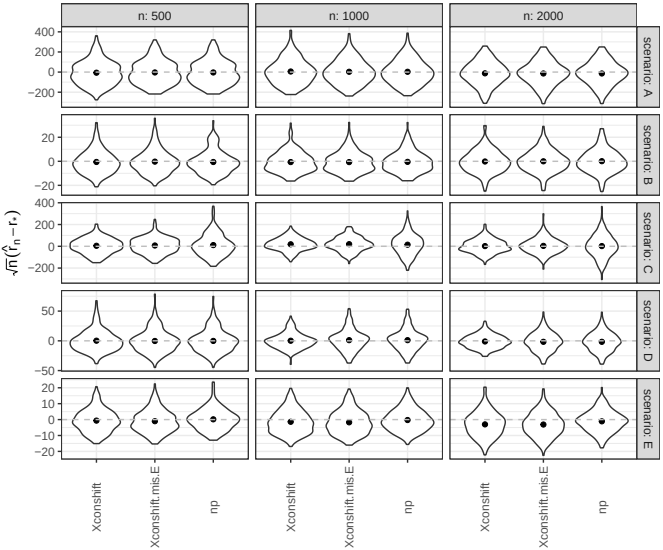
(D) Very large efficiency gain: f far from μ_* and no noise ($\epsilon = 0$)

(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: Violin plot of sampling distributions



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 $(P_A, P_{X|A}, P_{Y|X})$.

Simulation: full-data covariate shift

Five scenarios ($\Pr(A = 0) = 0.1$):

(A) f is the true optimal predictor; noisy outcome Y (very large efficiency gain)

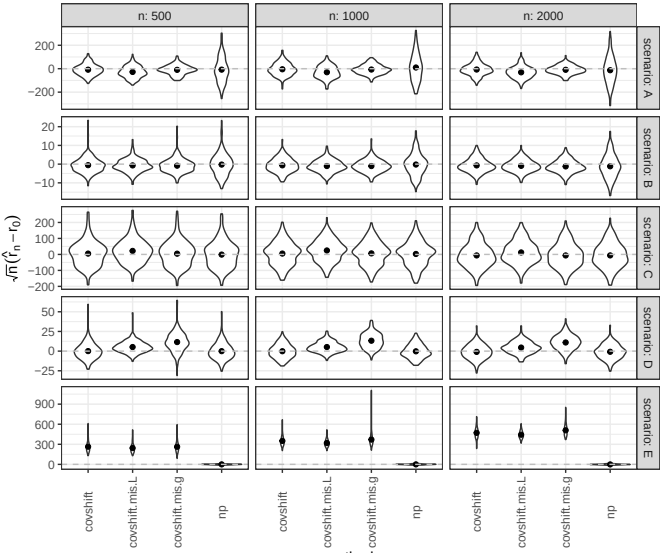
(B) f is a good predictor (large efficiency gain)

(C) f is a poor predictor (little efficiency gain)

(D) f is a poor predictor; deterministic outcome Y given covariate X (no efficiency gain)

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

Simulation: Violin plot of sampling distributions



Simulation findings

When covariate shift holds,

- both our proposed estimator $\hat{r}$ and the nonparametric estimator $\hat{r}_{\mathrm{np}}$ appear approximately normal
- when one nuisance function estimator is inconsistent, our proposed estimator $\hat{r}$ appears consistent, though it might not be asymptotically normal
- we expect a large efficiency gain for a good predictor f and noisy Y

When we assume covariate shift but it fails to hold,

- the nonparametric estimator $\hat{r}_{\mathrm{np}}$ is still consistent, because it does not leverage covariate shift
- our proposed estimator $\hat{r}$ can be severely biased
- there is a trade-off between efficiency and robustness against misassuming dataset shift condition

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 antiretroviral therapy (ART) coverage below 15% ($n = 1,418$)
- Source population: urban and rural communities ($n = 12,385$)

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

- Train a classifier f using half of the source population data ($n = 6192$)
- Use $n = 50$ target population datapoints and the other half of the source population data to estimate prediction error

$$r_* = \Pr(Y \neq f(X) \mid A = 0) = \mathbb{E}[\mathbb{1}(Y \neq f(X)) \mid A = 0]$$

- 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	P-value
None	0.24	0.060	(0.12, 0.36)	—
Concept shift in the features	0.26	0.057	(0.15, 0.38)	0.29
Concept shift in the labels	0.10	0.010	(0.08, 0.12)	0.02
Full-data covariate shift	0.19	0.026	(0.14, 0.25)	0.41
Full-data label shift	0.23	0.059	(0.11, 0.34)	0.42

Data analysis findings

- For the most plausible condition *a prior* (covariate shift), we do not reject this condition and obtain a large efficiency gain

  $>$ 50% smaller S.E. and shorter confidence interval compared to the nonparametric estimator
- Under a plausible dataset shift condition, using our proposed estimator can lead to substantial efficiency gain
- Our test rejected concept shit in the labels but did not reject the others
- Our test might be underpowered. We recommend using prior knowledge to judge what dataset shift condition is plausible