Cross-fitting

Chan Park

Wharton
University of Pennsylvania
DEPARTMENT OF STATISTICS AND DATA SCIENCE

(Efficient) Influence Function

- $O^F = (Y^{(0)}, Y^{(1)}, A, L)$ and $\tau^* = E\{\Psi^F(O^F)\} \in \mathbb{R}^p$

- $O = (Y, A, L)$ and $\tau^* = E\{\Psi(O)\} \in \mathbb{R}^p$ under causal assumptions

- $\mathcal{M} = \{P(O) \mid \text{regular law of } O\text{; can be parametric, semiparametric, nonparametric}\}$, $P^* = $ true law $\in \mathcal{M}$
  η^*: nuisance function at P^*

- $\mathcal{M}_t = \{P_t(O) \mid \text{parametric submodel of } \mathcal{M} \text{ indexed by } t\}$, $P_0 = P^*$
  $\tau_t = E_t\{\Psi_t(O)\}$; $\quad \eta_t = $ nuisance function at P_t; $\quad s_O(O;t) = \dfrac{\partial f(O;t)/\partial t}{f(O;t)} = $ score function

- Influence function $\mathtt{IF}(O;\eta,\tau)$ solves $\left.\dfrac{\partial \tau_t}{\partial t}\right|_{t=0} = E\{\mathtt{IF}(O;\eta^*,\tau^*) s_O(O;t=0)\}$

- Example: Average treatment effect with $\mathcal{M} = $ nonparametric model
  $\tau^*_{\mathsf{ATE}} = E\{Y^{(1)} - Y^{(0)}\} = E\{E(Y \mid A=1, X) - E(Y \mid A=0, X)\}$ under consistency, ignorability, positivity
  $\eta^* = (\mu^*, e^*)$ where $\mu^*(a,X) = E(Y \mid A=a, X)$ and $e^*(a,X) = \Pr(A=a \mid X)$

$$\mathtt{IF}(O;\eta^*,\tau^*_{\mathsf{ATE}}) = \frac{A\{Y-\mu^*(1,X)\}}{e^*(1,X)} - \frac{(1-A)\{Y-\mu^*(0,X)\}}{e^*(0,X)} + \{\mu^*(1,X) - \mu^*(0,X)\} - \tau^*_{\mathsf{ATE}}$$

Decomposition

- $\widehat{\tau} = N^{-1} \sum_{i=1}^{N} \widetilde{\texttt{IF}}(O_i ; \widehat{\eta})$

$$
\begin{aligned}
&\sqrt{N}(\widehat{\tau} - \tau^*) \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big[\widetilde{\texttt{IF}}(O_i ; \widehat{\eta}) - E\{\widetilde{\texttt{IF}}(O_i ; \eta^*)\} \Big] \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big[\widetilde{\texttt{IF}}(O_i ; \eta^*) - E\{\widetilde{\texttt{IF}}(O_i ; \eta^*)\} \Big] && \xrightarrow{D} N(0, \sigma^2) \\
&\quad + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big[\widetilde{\texttt{IF}}(O_i ; \widehat{\eta}) - E\{\widetilde{\texttt{IF}}(O_i ; \widehat{\eta})\} - \widetilde{\texttt{IF}}(O_i ; \eta^*) + E\{\widetilde{\texttt{IF}}(O_i ; \eta^*)\} \Big] && \Rightarrow \text{Empirical process} \overset{Want}{=} o_P(1) \\
&\quad + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big[E\{\widetilde{\texttt{IF}}(O_i ; \widehat{\eta})\} - E\{\widetilde{\texttt{IF}}(O_i ; \eta^*)\} \Big] && \Rightarrow \text{Bias} \overset{Want}{=} o_P(1)
\end{aligned}
$$

- $\widehat{\eta}$ and $\sum_{i=1}^{N}$ share the same observation $\Rightarrow$ Unclear to characterize EP and Bias

- Cross-fitting
  $\widehat{\eta}^{(-k)}$ (estimate nuisance ft without using $\mathcal{I}_k$) and $\sum_{i \in \mathcal{I}_k}$ (evaluate nuisance ft over $\mathcal{I}_k$)
  Clearer characterization of EP and Bias (i.e. show $o_P(1)$ easily (?)) for any generic ML learners

Cross-fitting

- References: [1, 4] [1]

Data

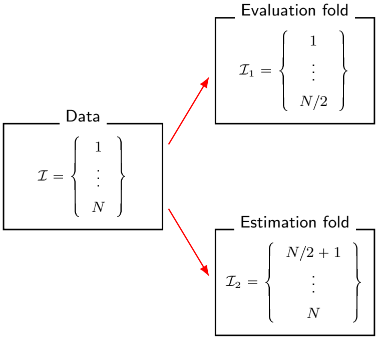
$$\mathcal{I} = \left\{ \begin{array}{c} 1 \\ \vdots \\ N \end{array} \right\}$$

[1] Schick (1986) On asymptotically efficient estimation in semiparametric models. *AoS*

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*

Cross-fitting

- References: [1, 4] [1]

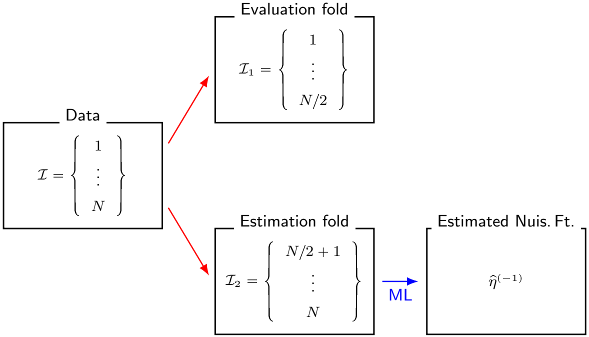


[1] Schick (1986) On asymptotically efficient estimation in semiparametric models. *AoS*

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*

Cross-fitting

- References: [1, 4] [1]

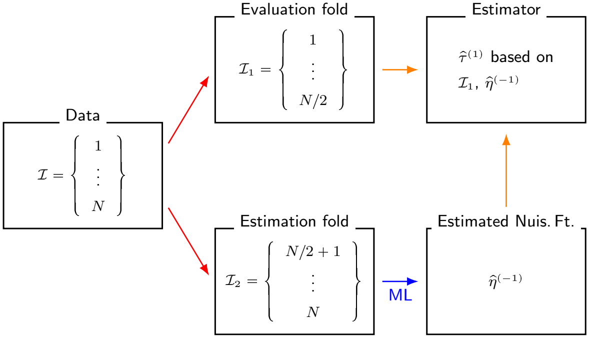


[1] Schick (1986) On asymptotically efficient estimation in semiparametric models. *AoS*

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*

Cross-fitting

- References: [1, 4] [1]

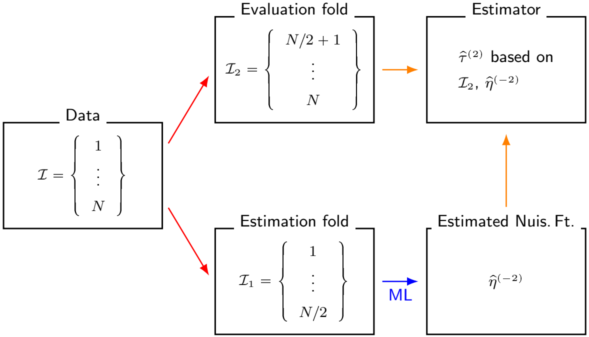


[1] Schick (1986) On asymptotically efficient estimation in semiparametric models. *AoS*

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*

Cross-fitting

- References: [1, 4] [1]

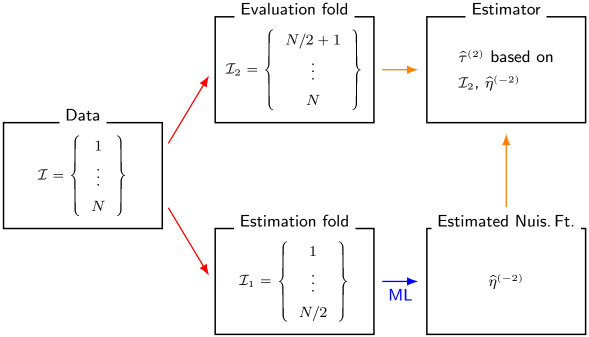


[1] Schick (1986) On asymptotically efficient estimation in semiparametric models. *AoS*

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*

Cross-fitting

- References: [1, 4] [1]



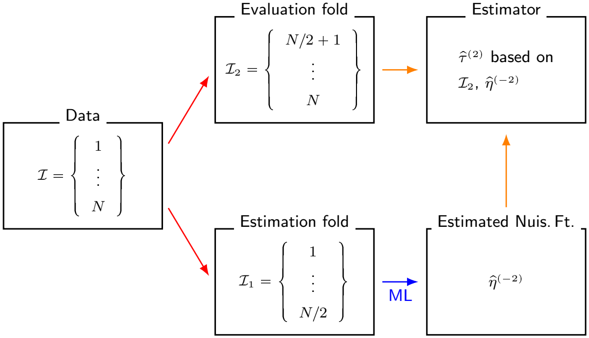
- $\widehat{\tau} = \text{aggregate}(\widehat{\tau}^{(1)}, \widehat{\tau}^{(2)})$ (e.g., $\dfrac{\widehat{\tau}^{(1)} + \widehat{\tau}^{(2)}}{2}$)

[1] Schick (1986) On asymptotically efficient estimation in semiparametric models. *AoS*

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*

Cross-fitting

- References: [1, 4] [1]



- $\widehat{\tau} = \text{aggregate}(\widehat{\tau}^{(1)}, \widehat{\tau}^{(2)})$ (e.g., $\dfrac{\widehat{\tau}^{(1)} + \widehat{\tau}^{(2)}}{2}$)

- ML: superlearner; [5] [2], minimax estimation [2] [3]; Forster-Warmuth Counterfactual Regression [7] [4]

[1] Schick (1986) On asymptotically efficient estimation in semiparametric models. *AoS*

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*

[2] van der Laan, Polley, Hubbard (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*

[3] Ghassami, Ying, Shpitser, Tchetgen Tchetgen (2022). Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. AISTAT

[4] Yang, Kuchibhotla, Tchetgen Tchetgen (2024). Forster-Warmuth counterfactual regression: a unified learning approach. *arXiv*

Superlearner by van der Laan et al. [5]

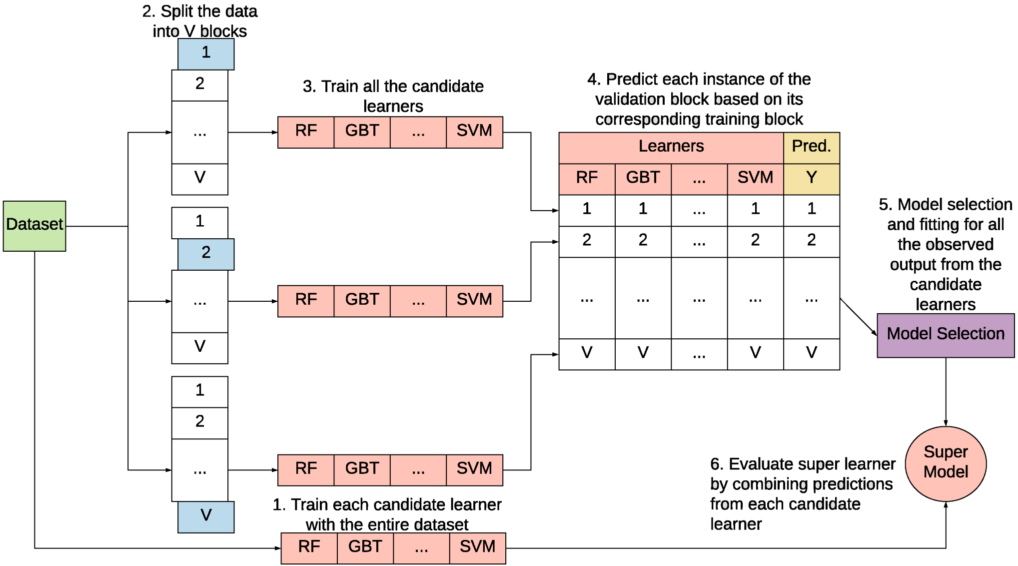


Figure: Reference: [3]

Superlearner by van der Laan et al. [5]: with an R example

```
SL.Fit ← MySL(       Data = data,
                     locY = response variable columns,
                     locX = explanatory variable columns,
                     Ydist = gaussian() or binomial(),
                     SL.list = SL.hpara$SLL,              # ML learners
                     MTRY = SL.hpara$MTRY,                # hyperparameters for random forest
                     MLPL = SL.hpara$MLPL,                # hyperparameters for multi-layer perceptron
                     MLPdecay=SL.hpara$MLPdecay,          # hyperparameters for multi-layer perceptron
                     NMN=SL.hpara$NMN)                    # hyperparameters for GBM

SL.Eval ← predict(   SL.Fit
                     newdata = newdata[,locX],
                     onlySL = TRUE)$pred
```

Minimax Estimation by Ghassami et al. [2]

- $\tau_1^* = E\{Y^{(1)}\}$

- $\texttt{IF}(O;\nu^*,\pi^*,\tau_1^*) = A\pi^*(X)\{Y-\nu^*(X)\} + \nu^* - \tau_1^*$ where $\pi^*(X) = 1/\Pr(A=1 \mid X)$ and $\nu^*(X) = \mu^*(1,X)$

- From the DR property:

$$E\{\texttt{IF}(O;\nu^*,\pi^*,\tau_1^*)\} = E\{\texttt{IF}(O;\nu^*,\pi^\dagger,\tau_1^*)\} = E\{\texttt{IF}(O;\nu^\dagger,\pi^*,\tau_1^*)\} = 0$$

$$\Rightarrow \quad 0 = E\big\{\texttt{IF}(O;\nu^*,\pi^\dagger,\tau_1^*) - \texttt{IF}(O;\nu^*,\pi^*,\tau_1^*)\big\} = E\big[A\{\pi^\dagger(X)-\pi^*(X)\}\{Y-\nu^*(X)\}\big] = E\big[p(X)\{AY - A\nu^*(X)\}\big]$$

- One can show

$$\nu^* = \arg\min_{\nu}\max_{p} E\big[p(X)\{AY - A\nu(X)\} - d^2(X)\big]$$

- We estimate ν^*

$$\widehat{\nu}^{(-k)} = \arg\min_{\nu\in\mathcal{H}_X}\left[\max_{p\in\mathcal{H}_X}\left[\mathbb{P}^{(-k)}\big[p(X)\{AY - A\nu(X)\} - p^2(X)\big] - \lambda_p\|p\|^2_{\mathcal{H}_X}\right] + \lambda_\nu\|\nu\|^2_{\mathcal{H}_X}\right]$$

$\mathcal{H}_X$: RKHS $\quad\Rightarrow\quad \widehat{\nu}^{(-k)}(x) = \sum \alpha_i K(x,X_i);$ see [2] for a closed-form representation of α_i

Minimax Estimation by Ghassami et al. [2]: with an R example

$$\widehat{\nu}^{(-k)} = \underset{\nu \in \mathcal{H}_X}{\arg\min} \left[\max_{p \in \mathcal{H}_X} \left[\mathbb{P}^{(-k)} [p(X)\{AY - A\nu(X)\} - p^2(X)] - \lambda_p \|p\|^2_{\mathcal{H}_X} \right] + \lambda_\nu \|\nu\|^2_{\mathcal{H}_X} \right.$$

Coef $= -A$

Intercept $= AY$

Perturb $= p(X)$

Target $= \nu(X)$

$$\widehat{\pi}^{(-k)} = \underset{\pi \in \mathcal{H}_X}{\arg\min} \left[\max_{q \in \mathcal{H}_X} \left[\mathbb{P}^{(-k)} [q(X)\{A\pi(X) - 1\} - q^2(X)] - \lambda_q \|q\|^2_{\mathcal{H}_X} \right] + \lambda_\pi \|\pi\|^2_{\mathcal{H}_X} \right.$$

Coef $= A$

Intercept $= -1$

Perturb $= q(X)$

Target $= \pi(X)$

Multiplier Bootstrap

- Reference: [6][1]

- $\hat{\tau}$ solves

$$0 = \frac{1}{N} \sum_{i \in \mathcal{I}_k} \texttt{IF}(O_i; \hat{\eta}^{(-k)}, \hat{\tau})$$

$$\hat{\tau}_{\mathsf{ATE}} = \frac{1}{N} \sum_{i \in \mathcal{I}_k} \left[\frac{A_i \{Y_i - \hat{\mu}^{(-k)}(1, X_i)\}}{\hat{e}^{(-k)}(1, X_i)} - \frac{(1 - A_i)\{Y_i - \hat{\mu}^{(-k)}(0, X_i)\}}{\hat{e}^{(-k)}(0, X_i)} + \{\hat{\mu}^{(-k)}(1, X_i) - \hat{\mu}^{(-k)}(0, X_i)\} \right]$$

- $\sqrt{N}(\hat{\tau} - \tau^*) \xrightarrow{D} N(0, \sigma^2)$

- A consistent estimator of σ^2 and SE of $\hat{\tau}$:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i \in \mathcal{I}_k} \left[\texttt{IF}(O_i; \hat{\eta}^{(-k)}, \hat{\tau}) \right]^2 \quad \Rightarrow \quad \mathsf{SE}(\hat{\tau}) = \frac{\hat{\sigma}}{\sqrt{N}}$$

- A multiplier bootstrap standard error of $\hat{\tau}$:

$$\mathsf{BSE}(\hat{\tau}) = \texttt{sd}(\hat{e}^{[1]}, \cdots, \hat{e}^{[B]}) \text{ where } \hat{e}^{[b]} = \frac{1}{N} \sum_{i \in \mathcal{I}_k} \epsilon_i^{[b]} \texttt{IF}(O_i; \hat{\eta}^{(-k)}, \hat{\tau}) \quad \Leftarrow \quad \epsilon_i^{[b]} \overset{iid}{\sim} N(0, 1)$$

Median Adjustment

- Cross-fitting estimates depend on particular split samples
- Repeat cross-fitting S times, say $S = 100$
- Obtain estimates $\widehat{\tau}^{[1]}, \ldots, \widehat{\tau}^{[S]}$ with the corresponding SE $\widehat{\sigma}^{2,[1]}, \ldots, \widehat{\sigma}^{2,[S]}$
- Median adjustment [1][1]

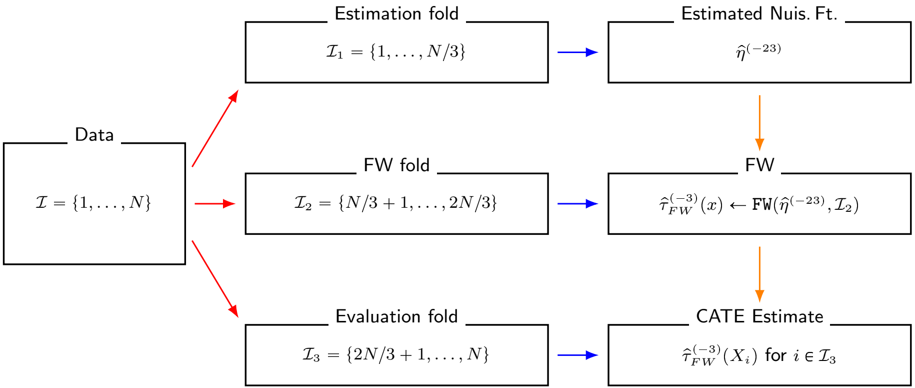
$$\widehat{\tau}^{(\text{report})} = \operatorname*{median}_{s=1,\ldots,S} \widehat{\tau}^{[s]}$$

$$\widehat{\sigma}^{2,(\text{report})} = \operatorname*{median}_{s=1,\ldots,S} \left[\widehat{\sigma}^{2,[s]} + \left\{ \widehat{\tau}^{[s]} - \widehat{\tau}^{(\text{report})} \right\}^2 \right]$$

[1] Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*

Forster-Warmuth Counterfactual Regression by Yang et al. [7]

- Reference: [7][1]
- Estimand is now infinite-dimensional function $\tau^*(x)$



[1] Yang, Kuchibhotla, Tchetgen Tchetgen (2024). Forster-Warmuth counterfactual regression: a unified learning approach. *arXiv*

Forster-Warmuth Counterfactual Regression by Yang et al. [7]

- $\tau^*(X)$: target estimand (CATE)
- Find a good f satisfying $E\{f(O;\eta^*) \,|\, X\} = \tau^*(X)$ (uncentered EIF in general)

$$\tau^*(X) = \mu^*(1,X) - \mu^*(0,X) \quad \Rightarrow \quad f(O;\eta^*) = \frac{A\{Y - \mu^*(1,X)\}}{e^*(1,X)} - \frac{(1-A)\{Y - \mu^*(0,X)\}}{e^*(0,X)} + \{\mu^*(1,X) - \mu^*(0,X)\}$$

- Using $\mathcal{I}_1$, get $\widehat{\eta}^{(-23)}$
- Using $\mathcal{I}_2$, define

$$\widehat{\tau}_{FW}^{(-3)}(x) = \frac{\boldsymbol{\phi}^\intercal(x)\Big[\boldsymbol{\Phi}_{\mathcal{I}_2}^\intercal \boldsymbol{\Phi}_{\mathcal{I}_2} + \boldsymbol{\phi}(x)\boldsymbol{\phi}^\intercal(x)\Big]^{-1}\Big[\boldsymbol{\Phi}_{\mathcal{I}_2}^\intercal \boldsymbol{f}_{\mathcal{I}_2}^{(-23)}\Big]}{1 - \boldsymbol{\phi}^\intercal(x)\Big[\boldsymbol{\Phi}_{\mathcal{I}_2}^\intercal \boldsymbol{\Phi}_{\mathcal{I}_2} + \boldsymbol{\phi}(x)\boldsymbol{\phi}^\intercal(x)\Big]^{-1}\boldsymbol{\phi}(x)}$$

where

$$\boldsymbol{\phi}(x) = \begin{bmatrix} \phi_1(x) \equiv 1 \\ \phi_2(x) \\ \vdots \\ \phi_J(x) \end{bmatrix} \in \mathbb{R}^J \quad \boldsymbol{\Phi}_{\mathcal{I}_2} = \Big[\boldsymbol{\phi}^\intercal(X_i)\Big]_{i\in\mathcal{I}_2} = \begin{bmatrix} \boldsymbol{\phi}^\intercal(X_{1+N/3}) \\ \vdots \\ \boldsymbol{\phi}^\intercal(X_{2N/3}) \end{bmatrix} \in \mathbb{R}^{(N/3)\times J} \quad \boldsymbol{f}_{\mathcal{I}_2}^{(-23)} = \Big[f(O_i;\widehat{\eta}^{(-23)})\Big]_{i\in\mathcal{I}_2} = \begin{bmatrix} f(O_{1+N/3};\widehat{\eta}^{(-23)}) \\ \vdots \\ f(O_{2N/3};\widehat{\eta}^{(-23)}) \end{bmatrix} \in \mathbb{R}^{N/3}$$

- $\boldsymbol{\phi}$: splines, polynomials, sin/cos, etc; J is chosen from cross-validation

Forster-Warmuth Counterfactual Regression by Yang et al. [7]: with an R example

- Check Yachong Yang's Github Code
- https://github.com/Elsa-Yang98/Forster_Warmuth_counterfactual_regression

References

[1] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

[2] Ghassami, A., Ying, A., Shpitser, I., and Tchetgen Tchetgen, E. (2022). Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7210–7239. PMLR.

[3] Neto, H. N. C., Lopez, M. A., Fernandes, N. C., and Mattos, D. M. (2020). Minecap: super incremental learning for detecting and blocking cryptocurrency mining on software-defined networking. *Annals of Telecommunications*, 75(3):121–131.

[4] Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139–1151.

[5] van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

[6] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

[7] Yang, Y., Kuchibhotla, A. K., and Tchetgen Tchetgen, E. (2024). Forster-warmuth counterfactual regression: A unified learning approach. *Preprint arXiv:2307.16798*.