

Cross-fitting

Chan Park



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
- $\mathcal{M} = \{P(O) \mid \text{regular law of } O; \text{ can be parametric, semiparametric, nonparametric}\}$, $P^* = \text{true law} \in \mathcal{M}$
 η^* : nuisance function at P^*
- $\mathcal{M}_t = \{P_t(O) \mid \text{parametric submodel of } \mathcal{M} \text{ indexed by } t\}$, $P_0 = P^*$
 $\tau_t = E_t\{\Psi_t(O)\}$; $\eta_t = \text{nuisance function at } P_t$; $s_O(O; t) = \frac{\partial f(O; t)/\partial t}{f(O; t)} = \text{score function}$
- Influence function $\text{IF}(O; \eta, \tau)$ solves $\left. \frac{\partial \tau_t}{\partial t} \right|_{t=0} = E\{\text{IF}(O; \eta^*, \tau^*) s_O(O; t=0)\}$
- Example: Average treatment effect with $\mathcal{M} = \text{nonparametric model}$
 $\tau_{\text{ATE}}^* = E\{Y^{(1)} - Y^{(0)}\} = E\{E(Y \mid A = 1, X) - E(Y \mid A = 0, X)\}$ under consistency, ignorability, positivity
 $\eta^* = (\mu^*, e^*)$ where $\mu^*(a, X) = E(Y \mid A = a, X)$ and $e^*(a, X) = \Pr(A = a \mid X)$

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

Decomposition

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

$$\sqrt{N}(\hat{\tau} - \tau^*)$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[\widetilde{\mathbf{IF}}(O_i; \hat{\eta}) - E\{\widetilde{\mathbf{IF}}(O_i; \eta^*)\} \right]$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[\widetilde{\mathbf{IF}}(O_i; \eta^*) - E\{\widetilde{\mathbf{IF}}(O_i; \eta^*)\} \right] \xrightarrow{D} N(0, \sigma^2)$$

$$+ \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[\widetilde{\mathbf{IF}}(O_i; \hat{\eta}) - E\{\widetilde{\mathbf{IF}}(O_i; \hat{\eta})\} - \widetilde{\mathbf{IF}}(O_i; \eta^*) + E\{\widetilde{\mathbf{IF}}(O_i; \eta^*)\} \right] \Rightarrow \text{Empirical process } \overset{W_{ant}}{=} o_P(1)$$

$$+ \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[E\{\widetilde{\mathbf{IF}}(O_i; \hat{\eta})\} - E\{\widetilde{\mathbf{IF}}(O_i; \eta^*)\} \right] \Rightarrow \text{Bias } \overset{W_{ant}}{=} o_P(1)$$

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

- Cross-fitting

$\hat{\eta}^{(-k)}$ (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

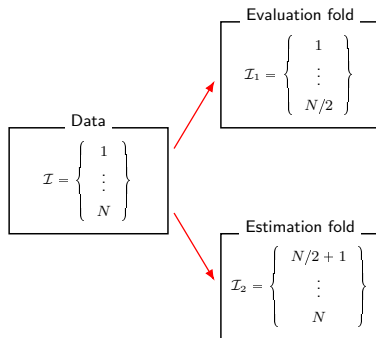
$$\begin{array}{c} \text{Data} \\ \boxed{\mathcal{I} = \left\{ \begin{array}{c} 1 \\ \vdots \\ N \end{array} \right\}} \end{array}$$

¹Schick (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

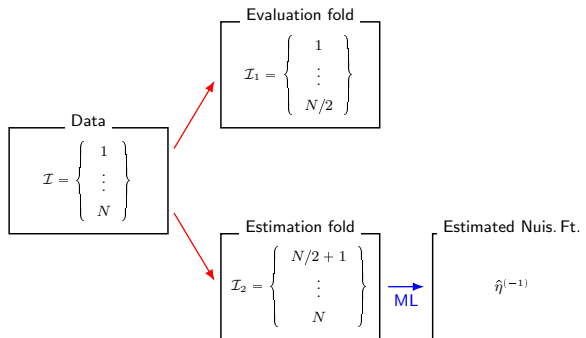


¹Schick (1986) On asymptotically efficient estimation in semiparametric models. *AoS*

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

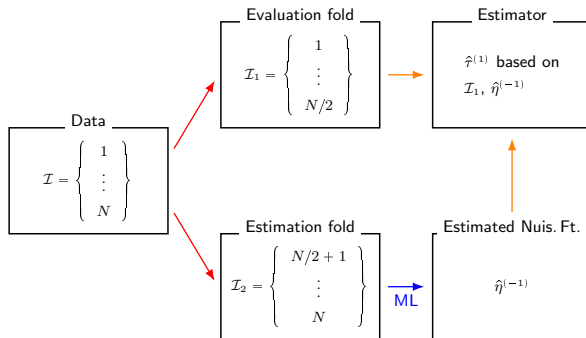


¹Schick (1986) On asymptotically efficient estimation in semiparametric models. *AoS*

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

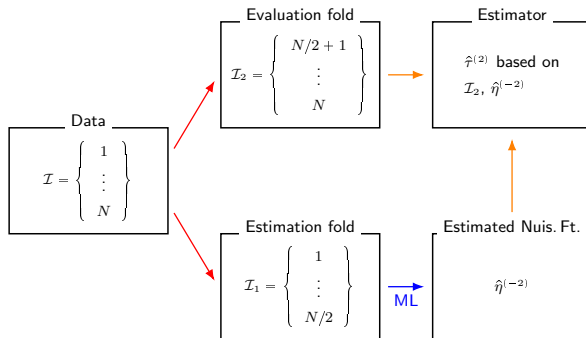


¹Schick (1986) On asymptotically efficient estimation in semiparametric models. *AoS*

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

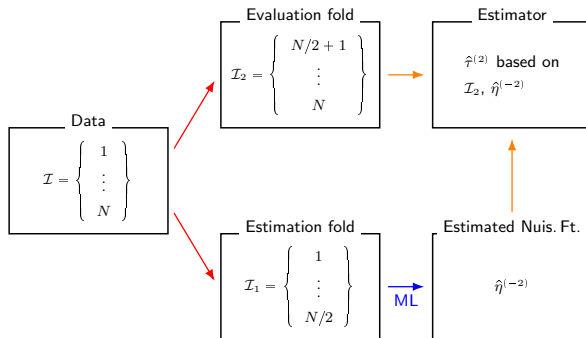


¹Schick (1986) On asymptotically efficient estimation in semiparametric models. *AoS*

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



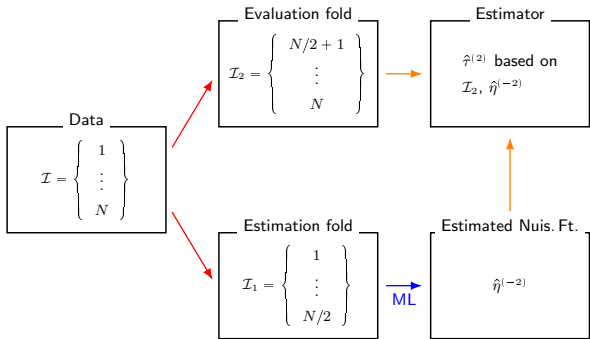
- $\hat{\tau} = \text{aggregate}(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$ (e.g., $\frac{\hat{\tau}^{(1)} + \hat{\tau}^{(2)}}{2}$)

¹Schick (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



- $\hat{\tau} = \text{aggregate}(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$ (e.g., $\frac{\hat{\tau}^{(1)} + \hat{\tau}^{(2)}}{2}$)
- ML**: superlearner; [5]², minimax estimation [2]³; Forster-Warmuth Counterfactual Regression [7]⁴

¹Schick (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*

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

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

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

Superlearner by van der Laan et al. [5]

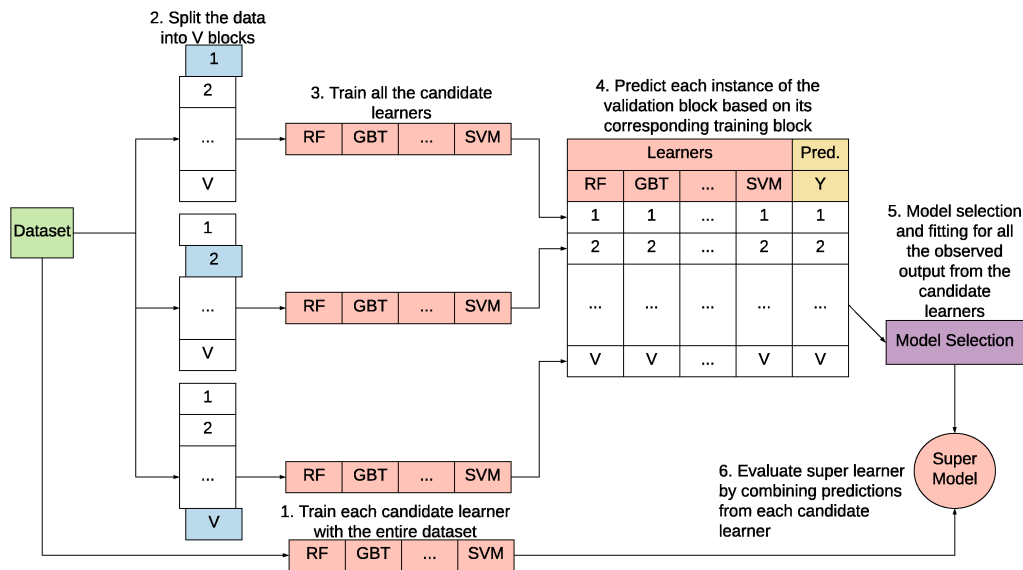


Figure: Reference: [3]

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

https://github.com/qkrcks0218/CCI_Crossfitting

```
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]

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

- From the DR property:

$$\begin{aligned} E\{\text{IF}(O; \nu^*, \pi^*, \tau_1^*)\} &= E\{\text{IF}(O; \nu^*, \pi^\dagger, \tau_1^*)\} = E\{\text{IF}(O; \nu^\dagger, \pi^*, \tau_1^*)\} = 0 \\ \Rightarrow 0 &= E\{\text{IF}(O; \nu^*, \pi^\dagger, \tau_1^*) - \text{IF}(O; \nu^*, \pi^*, \tau_1^*)\} = E[A\{\pi^\dagger(X) - \pi^*(X)\}\{Y - \nu^*(X)\}] = E[p(X)\{AY - A\nu^*(X)\}] \end{aligned}$$

- One can show

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

- We estimate ν^*

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

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

- Minimax estimator can be used to solve integral equations

$$g(v) = E\{f^*(W) \mid V = v\}$$

$$\Rightarrow 0 = E[p(V)\{f^*(W) - g(V)\}] , \quad \forall p$$

$$\Rightarrow \hat{f}^{(-k)} = \arg \min_{f \in \mathcal{H}_W} \left[\max_{p \in \mathcal{H}_V} \left[\mathbb{P}^{(-k)}[p(V)\{f(W) - g(V)\} - p^2(V)] - \lambda_p \|p\|_{\mathcal{H}_V}^2 \right] + \lambda_f \|f\|_{\mathcal{H}_W}^2 \right]$$

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

https://github.com/qkrcks0218/CCI_Crossfitting

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

Coef = $-A$; Intercept = AY ; Perturb = $p(X)$; Target = $\nu(X)$

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

Coef = A ; Intercept = -1 ; Perturb = $q(X)$; Target = $\pi(X)$

Multiplier Bootstrap

- Reference: [6]¹

- $\hat{\tau}$ solves

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

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

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

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

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

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

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

Median Adjustment

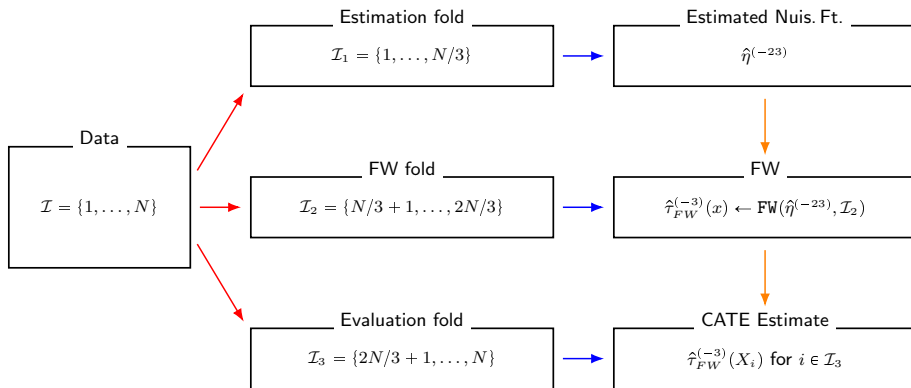
- Cross-fitting estimates depend on particular split samples
- Repeat cross-fitting S times, say $S = 100$
- Obtain estimates $\hat{\tau}^{[1]}, \dots, \hat{\tau}^{[S]}$ with the corresponding SE $\hat{\sigma}^{2,[1]}, \dots, \hat{\sigma}^{2,[S]}$
- Median adjustment [1]¹

$$\begin{aligned}\hat{\tau}^{(\text{report})} &= \text{median}_{s=1, \dots, S} \hat{\tau}^{[s]} \\ \hat{\sigma}^{2, (\text{report})} &= \text{median}_{s=1, \dots, S} \left[\hat{\sigma}^{2, [s]} + \left\{ \hat{\tau}^{[s]} - \hat{\tau}^{(\text{report})} \right\}^2 \right]\end{aligned}$$

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 $\tau^*(x)$



¹Yang, 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 $E\{f(O; \eta^*) \mid X\} = \tau^*(X)$ (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 $\hat{\eta}^{(-23)}$
- Using \mathcal{I}_2 , define

$$\hat{\tau}_{FW}^{(-3)}(x) = \frac{\phi^\top(x) \left[\Phi_{\mathcal{I}_2}^\top \Phi_{\mathcal{I}_2} + \phi(x) \phi^\top(x) \right]^{-1} \left[\Phi_{\mathcal{I}_2}^\top \mathbf{f}_{\mathcal{I}_2}^{(-23)} \right]}{1 - \phi^\top(x) \left[\Phi_{\mathcal{I}_2}^\top \Phi_{\mathcal{I}_2} + \phi(x) \phi^\top(x) \right]^{-1} \phi(x)}$$

where

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

- ϕ : splines, polynomials, sin/cos, etc; J is chosen from cross-validation

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

- https://github.com/qkrcks0218/CCI_Crossfitting

- Also check Yachong Yang's Github

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*.