Cross-fitting in causal machine learning with correlated units

Salvador Balkus
sbalkus@g.harvard.edu
Harvard Biostatistics

Hasan Laith
hasanhussein@college.harvard.edu
Harvard Statistics

Nima Hejazi
nhejazi@hsph.harvard.edu
Harvard Biostatistics

August 3, 2026

Cross-fitting background

A growing number of statistical estimators take the form

$$\mathsf{P}_n f_{\hat{\eta}} = \frac{1}{n} \sum_{i=1}^{n} f_{\hat{\eta}}(X_i)$$

where…

- we observe data $X_1, \dots, X_n \sim \mathsf{P}$
- $\hat{\eta}$ contains nuisance parameters fit using machine learning

Cross-fitting background

A growing number of statistical estimators take the form

$$\mathsf{P}_n f_{\hat{\eta}} = \frac{1}{n} \sum_{i=1}^{n} f_{\hat{\eta}}(X_i)$$

e.g. One-step estimator of counterfactual mean $\mathbb{E}[\mathbb{E}(Y \mid A = 1, L)]$*:*

$$\frac{1}{n} \sum_{i=1}^{n} \underbrace{\hat{\mathbb{E}}(Y_i \mid A_i = 1, L_i)}_{\text{conditional mean}} + \frac{\mathbb{I}(A_i = 1)}{\underbrace{\hat{\mathbb{P}}(A_i = 1 \mid L_i)}_{\text{inverse probability}}} \underbrace{[Y_i - \hat{\mathbb{E}}(Y_i \mid A_i, L_i)]}_{\text{residuals}}$$

Cross-fitting background

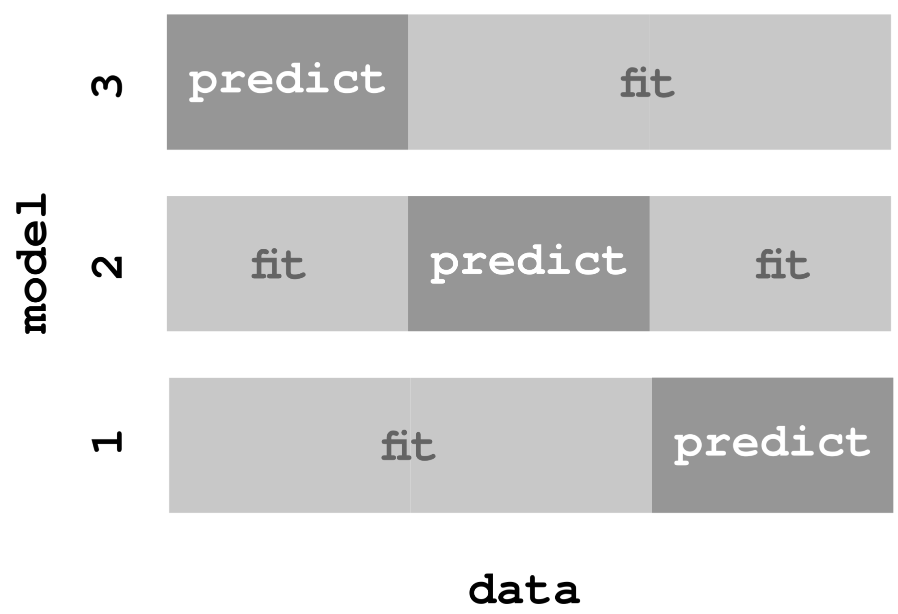
A growing number of statistical estimators take the form

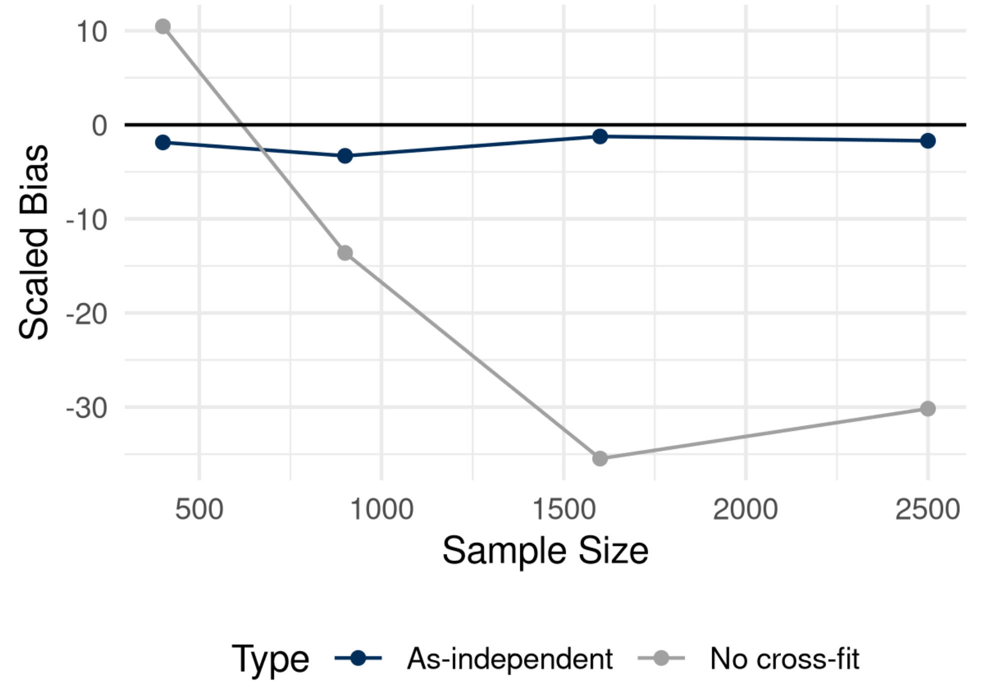
$$\mathsf{P}_n f_{\hat{\eta}} = \frac{1}{n}\sum_{i=1}^{n} f_{\hat{\eta}}(X_i)$$

Problem: Estimating $\hat{\eta}$ and $\mathsf{P}_n f_{\hat{\eta}}$ on the *same data* = **bias!**

Cross-fitting background

Solution: Partition data, train on subset, predict on independent held-out fold

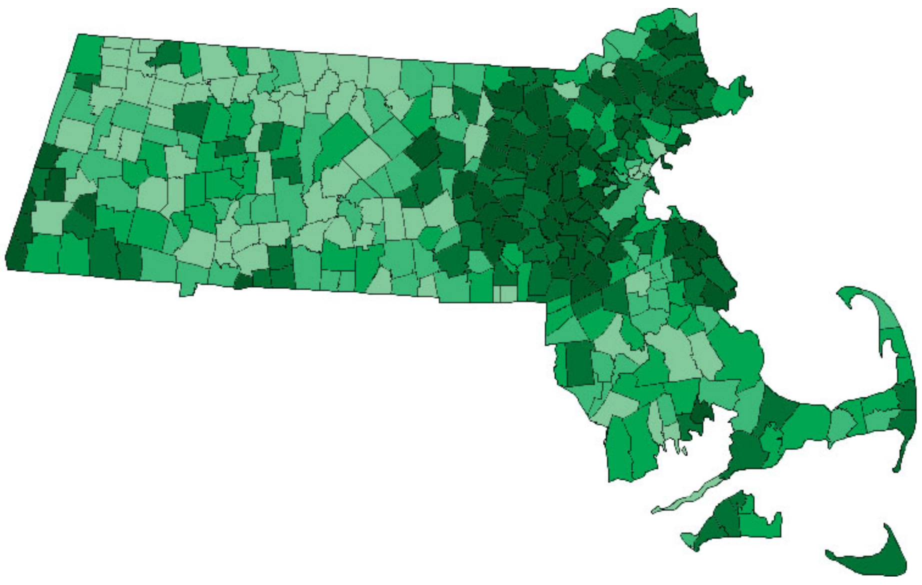


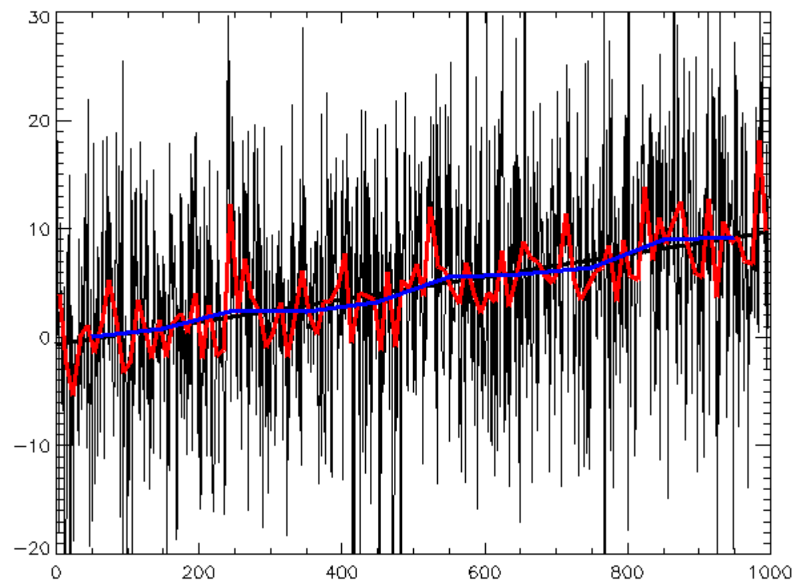


What if $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are correlated?

What if $X_1, \dots, X_n$ are correlated?

Examples: Census tracts, remote sensing, cluster trials, time series





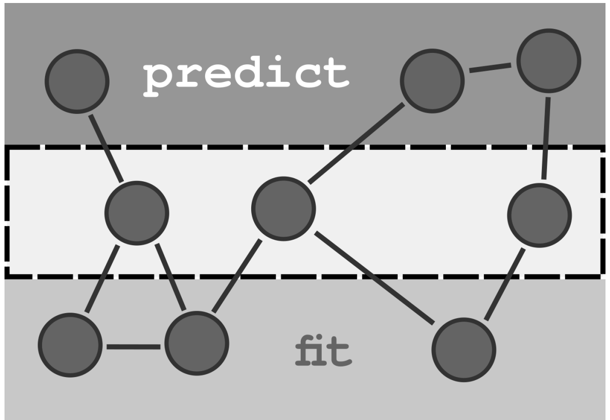
Wikimedia Commons

Research question: How to deal with correlated folds?

Existing approaches

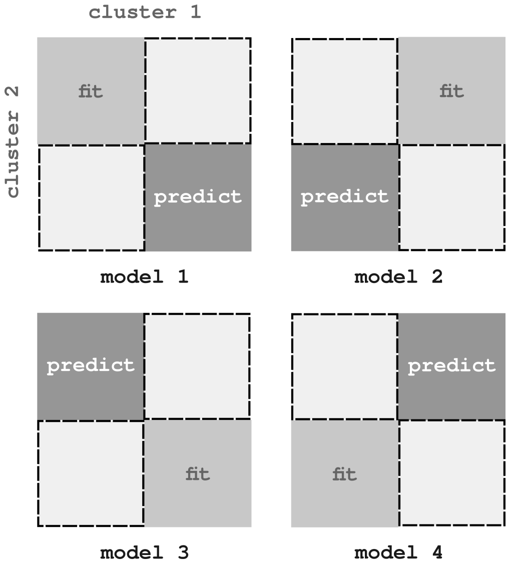
Block / drop units to create independent folds

Network



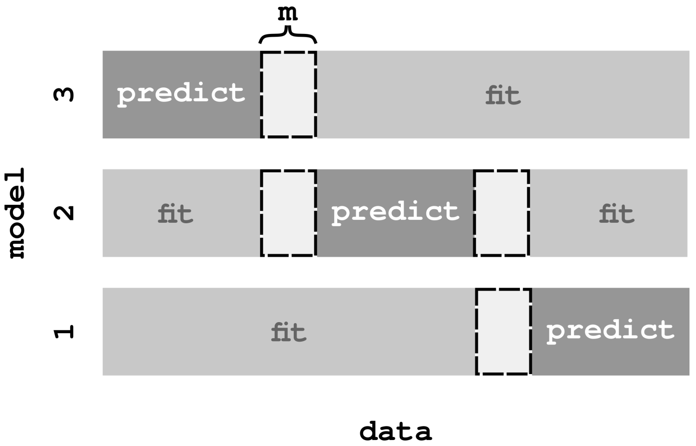
Emmenegger *et al.* (2025)

Cluster



Chiang *et al.* (2021)

Time series



Semenova *et al.* (2023)

Our approach

Our approach

Do nothing.

Our approach

Do nothing.*

** Just partition data into cross-fitting folds as if units were independent*

What error does cross-fitting eliminate?

$$\underbrace{\mathsf{P}_n\hat{f} - \mathsf{P}f}_{\text{Total bias}} = \underbrace{(\mathsf{P}_n - \mathsf{P})f}_{\text{CLT term}} + \underbrace{\mathsf{P}(\hat{f} - f)}_{\text{nuisance bias}} + \underbrace{(\mathsf{P}_n - \mathsf{P})(\hat{f} - f)}_{\text{empirical process bias } Z_n}$$

What error does cross-fitting eliminate?

$$\underbrace{\mathsf{P}_n\hat{f} - \mathsf{P}f}_{\text{Total bias}} = \underbrace{(\mathsf{P}_n - \mathsf{P})f}_{\text{CLT term}} + \underbrace{\mathsf{P}(\hat{f} - f)}_{\text{nuisance bias}} + \underbrace{(\mathsf{P}_n - \mathsf{P})(\hat{f} - f)}_{\text{empirical process bias } Z_n}$$

In independent units, cross-fitting ensures $\sqrt{n}Z_n = o_{\mathsf{P}}(1)$

(*rest addressed by regularity conditions*)

What error does cross-fitting eliminate?

$$\underbrace{\mathsf{P}_n\hat{f} - \mathsf{P}f}_{\text{Total bias}} = \underbrace{(\mathsf{P}_n - \mathsf{P})f}_{\text{CLT term}} + \underbrace{\mathsf{P}(\hat{f} - f)}_{\text{nuisance bias}} + \underbrace{(\mathsf{P}_n - \mathsf{P})(\hat{f} - f)}_{\text{empirical process bias } Z_n}$$

Theorem 1. If, over correlated units, an r_n-CLT exists, and nuisance bias is $o_{\mathsf{P}}(1/r_n)$, then under "as-independent" cross-fitting

$$r_n Z_n = r_n(\mathsf{P}_n - \mathsf{P})(\hat{f} - f) = o_{\mathsf{P}}(1)$$

(*meaning it is asymptotically negligible*)

Proof idea

An important lemma

Lemma 1. Suppose $g(X) \sim \mathsf{Q}$ is **sub-exponential** (MGF is well-controlled):

$$\mathbb{E}_{\mathsf{P}}\left(e^{\lambda(g(X)-E_{\mathsf{P}}(g(X)))}\right) \leq e^{\nu^2\lambda^2} \text{ for } -1/\nu \leq \lambda \leq 1/\nu$$

Let P denote an arbitrary probability distribution absolutely continuous with respect to Q. Then,

$$|\mathbb{E}_{\mathsf{P}}(g(X)) - \mathbb{E}_{\mathsf{Q}}(g(X))| \leq 2\nu\sqrt{D_{KL}(\mathsf{P} \parallel \mathsf{Q})}$$

- Adapted from Xu and Raginsky (**2017**)

An important lemma

Lemma 1. Suppose $g(X) \sim \mathsf{Q}$ is **sub-exponential** (MGF is well-controlled):

$$\mathbb{E}_{\mathsf{P}}\left(e^{\lambda(g(X)-E_{\mathsf{P}}(g(X)))}\right) \leq e^{\nu^2\lambda^2} \text{ for } -1/\nu \leq \lambda \leq 1/\nu$$

Let P denote an arbitrary probability distribution absolutely continuous with respect to Q. Then,

$$|\mathbb{E}_{\mathsf{P}}(g(X)) - \mathbb{E}_{\mathsf{Q}}(g(X))| \leq 2\nu\sqrt{D_{KL}(\mathsf{P} \parallel \mathsf{Q})}$$

- Adapted from Xu and Raginsky (2017)

Proof idea

Let P denote the *true* distribution, and Q the "ideal" where folds are independent. Then, we can bound $\mathbb{P}_{\mathsf{P}}(|r_n Z_n| > \varepsilon)$

$$\leq \underbrace{|\mathbb{P}_{\mathsf{Q}}(|r_n Z_n| > \varepsilon)|}_{\text{Ideal folds}} + \underbrace{|\mathbb{P}_{\mathsf{P}}(|r_n Z_n| > \varepsilon) - \mathbb{P}_{\mathsf{Q}}(|r_n Z_n| > \varepsilon)|}_{\leq 2\nu\sqrt{D_{KL}(\mathsf{P}\|\mathsf{Q})}\ \text{(Lemma 1)}}$$

KL-divergence of two datasets is $O(n)$, and (by classical assumption)

$$\nu = O(r_n \|\hat{f} - f\|_{L_2(Q)}/\sqrt{n}) = o_{\mathsf{P}}(1/\sqrt{n})$$

Hence, scaled empirical process error is $o_{\mathsf{P}}(1/\sqrt{n}) \cdot O(\sqrt{n}) = o_{\mathsf{P}}(1)$.

Proof idea

Let P denote the *true* distribution, and Q the "ideal" where folds are independent. Then, we can bound $\mathbb{P}_{\mathsf{P}}(|r_n Z_n| > \varepsilon)$

$$\leq \underbrace{|\mathbb{P}_{\mathsf{Q}}(|r_n Z_n| > \varepsilon)|}_{\textbf{Ideal folds}} + \underbrace{|\mathbb{P}_{\mathsf{P}}(|r_n Z_n| > \varepsilon) - \mathbb{P}_{\mathsf{Q}}(|r_n Z_n| > \varepsilon)|}_{\leq 2\nu\sqrt{D_{KL}(\mathsf{P}\|\mathsf{Q})}\ \textbf{(Lemma 1)}}$$

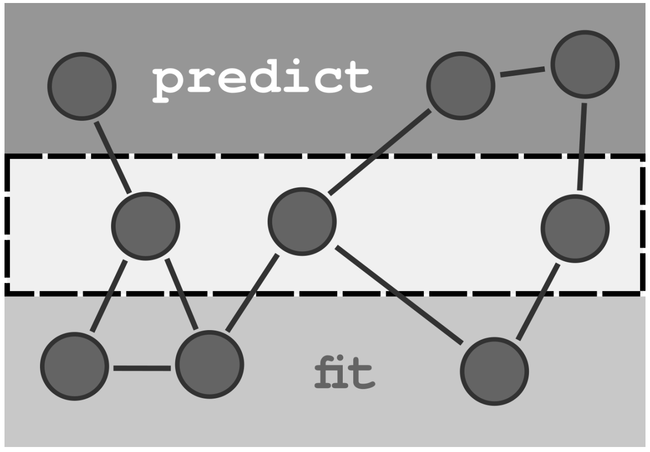
KL-divergence of two datasets is $O(n)$, and (by classical assumption)

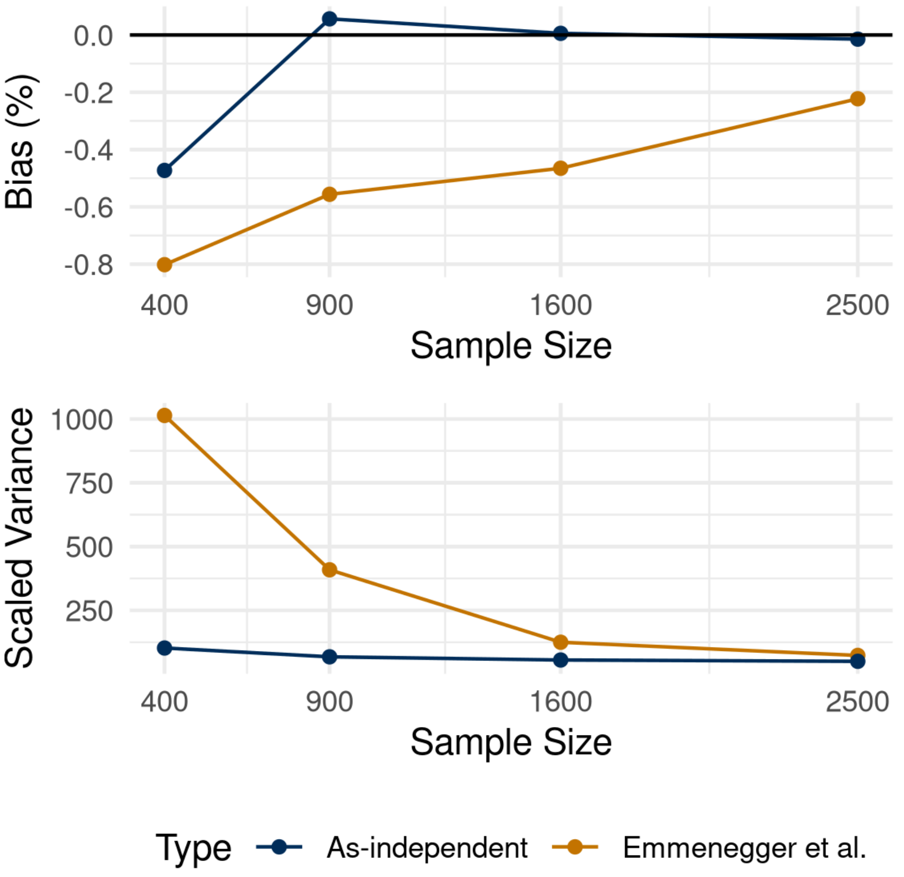
$$\nu = O(r_n \|\hat{f} - f\|_{L_2(Q)}/\sqrt{n}) = o_{\mathsf{P}}(1/\sqrt{n})$$

Hence, scaled empirical process error is $o_{\mathsf{P}}(1/\sqrt{n}) \cdot O(\sqrt{n}) = o_{\mathsf{P}}(1)$.

Why is this so useful?

Comparing "as-independent" splits to throwing out correlated observations:





Conclusions

- We show that, in spite of conventional wisdom, “doing nothing” to account for correlation still allows cross-fitting to control bias

- Of course, other challenges remain (uncertainty quantification, nuisance convergence, etc.)

- Avoid loss of effective sample size, computational inefficiency, implementation complexity

Conclusions

- We show that, in spite of conventional wisdom, "doing nothing" to account for correlation still allows cross-fitting to control bias
- Of course, other challenges remain (uncertainty quantification, nuisance convergence, etc.)
- Avoid loss of effective sample size, computational inefficiency, implementation complexity

Conclusions

- We show that, in spite of conventional wisdom, “doing nothing” to account for correlation still allows cross-fitting to control bias

- Of course, other challenges remain (uncertainty quantification, nuisance convergence, etc.)

- Avoid loss of effective sample size, computational inefficiency, implementation complexity

Thank you! Questions?

Preprint

Personal Website

HDSI

Harvard Data Science Initiative

All opinions are my own and do not necessarily reflect the views of the National Science Foundation or Harvard Data Science Initiative.

References

Chiang, H. D., Kato, K., Ma, Y., et al. (2021) Multiway cluster robust double/debiased machine learning. *Journal of Business & Economic Statistics*, **40**, 1046–1056. Informa UK Limited. DOI: 10.1080/07350015.2021.1895815.

Emmenegger, C., Spohn, M.-L., Elmer, T., et al. (2025) Treatment effect estimation with observational network data using machine learning. *Journal of Causal Inference*, **13**, 20230082. Walter de Gruyter GmbH. DOI: 10.1515/jci-2023-0082.

Semenova, V., Goldman, M., Chernozhukov, V., et al. (2023) Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics*, **14**, 471–510. The Econometric Society. DOI: 10.3982/qe1670.

Xu, A. and Raginsky, M. (2017) Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, **30**.