Is Cross-validation the Gold Standard to Estimate Out-of-sample Model Performance?

Garud Iyengar, Henry Lam, **Tianyu Wang**

{gi10, khl2114, **tw2837**}@columbia.edu

Problem Setup

Data D_n: $\{(X_i, Y_i)\}_{i\in[n]} \sim P$

Predictor: $\hat{z}(x) = A(D_n; x)$

Model A: Parametric + **Nonparametric**

Evaluation of $c(\hat{z})$ $= E_{P_{(X,Y)}}[\ell(\hat{z}(X); Y)]$

Goal: Valid point estimator $\hat{A}$, interval $I(\alpha)$

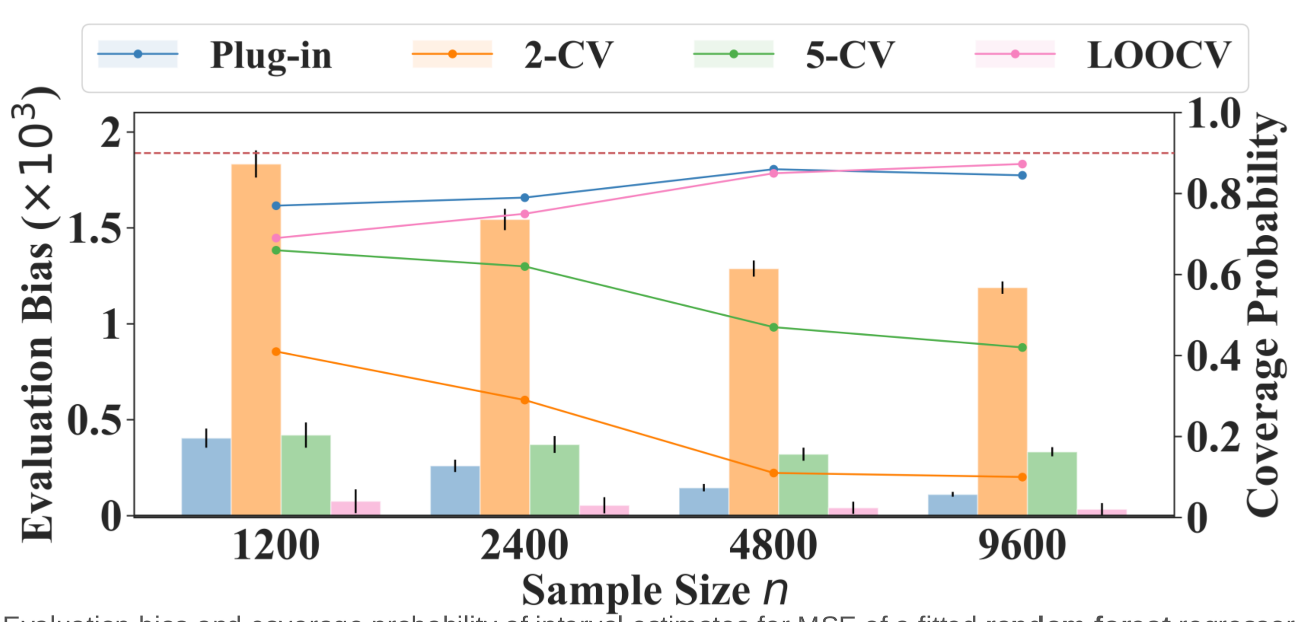
Small bias size: $E[c(\hat{z}) - \hat{A}]$

Validity coverage: $\lim_{n\to\infty} P(c(\hat{z}) \in I(\alpha)) = 1 - \alpha$

Out-of-sample Model Performance

Motivation

- Cross-validation is the default choice of estimating model performance.
- Despite wide utility, their statistical benefits remain unknown.



Evaluation bias and coverage probability of interval estimates for MSE of a fitted **random forest** regressor

Research Question:

- ***Are LOOCV and K-fold CV a "must use" in estimating out-of-sample model performance in general?***
- ***If not, when are these evaluation procedures worthwhile?***

What has been understood

Cross-validation (CV)

- Limiting theorems centered at the average-across-fold $\hat{z}^{(-k)}$ [1].
- Standard procedure suffers under high-dimensional (linear) models [2].

Plug-in (In-sample Loss)

- Valid asymptotic normality when model is stable [3].
- Overfit under complex models.

No results for their performance difference, especially under nonparametric models with a slow convergence rate.

Main Results

***Key takeaway:** In terms of estimating out-of-sample model performance, for various parametric and nonparametric estimators, cross-validation does not statistically outperform plug-in, both asymptotic bias and coverage accuracy of the associate interval.*

Model Assumptions

- [Smoothness] $\forall x$, $E_{P_{Y|x}}[\ell(z;Y)]$ is twice differentiable and bounded. Optimality condition holds.
- [Model Convergence Rate] $E_{D_n}\left[\|\hat{z}(x) - z^*(x)\|_2\right] = \Theta(n^{-\gamma}), \gamma \in (0, ½]$.

  $\gamma = \min\{\gamma_b, \gamma_v\} \in (0, ½]$ denotes bias and variability convergence.

  $\gamma = \frac{1}{2}$: parametric model $\gamma < \frac{1}{2}$: kNN, NW kernel, random forest [4][5]
- [LOO Stability] $E_{P,D_n}[|\hat{z}(X) - \hat{z}^{(-i)}(X)|^2] = o(n^{-1})$.

Point Estimators and Interval Procedures

$$I_m(\alpha) = [\hat{A}_m - z_{1-\frac{\alpha}{2}}\hat{\sigma}_m/\sqrt{n},\ \hat{A}_m + z_{1-\frac{\alpha}{2}}\hat{\sigma}_m/\sqrt{n}],\ m \in \{p, cv\}.$$

Plug-in: $\hat{A}_p = \frac{1}{n}\sum_{i\in[n]} \ell(\hat{z}(X_i); Y_i)$, $\sigma_p^2 = \frac{1}{n}\sum_{i\in[n]}\left(\ell(\hat{z}(X_i); Y_i) - \hat{A}_p\right)^2$,

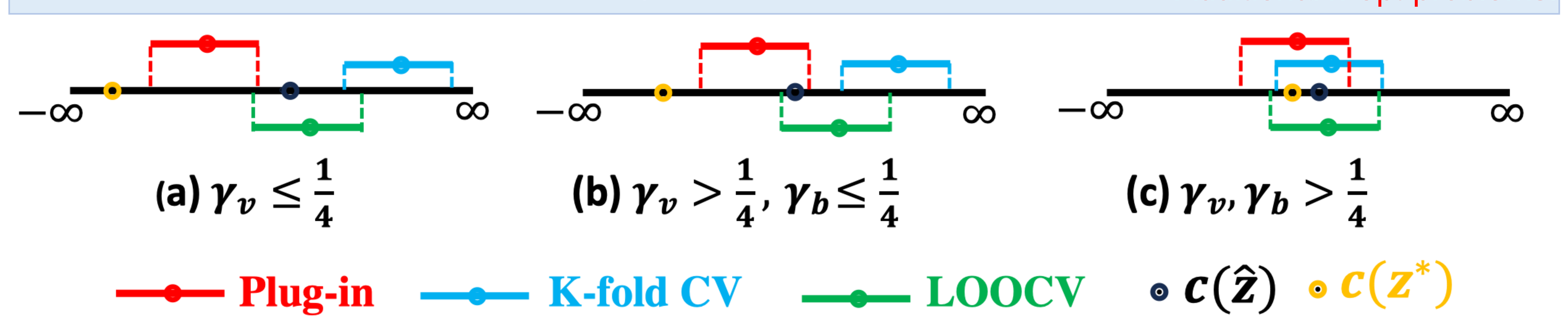
Cross-validation: $\hat{A}_{cv} = \frac{1}{n}\sum_{k\in[K]}\sum_{i\in N_k} \ell\left(\hat{z}^{(-N_k)}(X_i); Y_i\right)$, $\sigma_p^2 = \frac{1}{n}\sum_{k\in[K]}\sum_{i\in N_k}\left(\ell\left(\hat{z}^{(-N_k)}(X_i); Y_i\right) - \hat{A}_{cv}\right)^2$

Main Theorem

	K-Fold CV	Plug-in	LOOCV
Bias size	$\Theta(n^{-2\gamma})$, < 0	$\Theta(n^{-2\gamma_v})$, > 0	$o(n^{-1})$, < 0
Condition of the validity of interval coverage	$\gamma > \frac{1}{4}$	$\gamma_v > \frac{1}{4}$	Any γ_b, γ_v

Never outperform plug-in: $\gamma \le \gamma_v$

Computationally worse: Additional n opt problems.



Concept plots of interval coverage of $c(\hat{z})$ and $c(z^*) := E_P[c(z^*(X);Y)]$ for evaluation procedures across model rates

Proof Sketch

Step 1: Variability term fixed as $O_p\left(n^{-\frac{1}{2}}\right)$. (stability)

Step 2.1: Plug-in (Optimistic) Bias (controlled by γ_v)

- M-estimator asymptotics → local samples → bias decomposition via a novel Taylor expansion.

Step 2.2: CV Bias (controlled by γ)

- Fewer samples in each evaluation that affects the convergence.

Examples & Experiments

Nonparametric examples

$\hat{z}(x) \in \mathrm{argmin}_{z\in Z}\sum_{i\in[n]} w_{n,i}(x)\ell(z; Y_i)$

kNN: $w_{n,i}(x) = 1_{\{X_i \text{ is a kNN of } x\}}$;

Random Forest (RF): $w_{n,i}(x) = \sum_{j=1}^{T} 1_{\{\tau_j(x_i)=\tau_j(x)\}}/T$, $F = \{\tau_1, \dots, \tau_T\}$

	Bias			Coverage Validity		
	Plug-in	K-CV	LOOCV	Plug-in	K-CV	LOOCV
LERM	$\mathbf{o(n^{-1/2})}$	$\mathbf{o(n^{-1/2})}$	$\boldsymbol{o(n^{-1/2})}$	√	√	√
kNN (large k)	$\mathbf{o(n^{-1/2})}$	$\Omega(n^{-1/2})$		√	×	√
kNN (small k)	$\Omega(n^{-1/2})$	$\Omega(n^{-1/2})$		×	×	√
RF	$\mathbf{o(n^{-1/2})}$	$\Omega(n^{-1/2})$		√	×	√

Numerical simulation

Regression problem: $\ell(z; Y) = (z - Y)^2$

CVaR portfolio optimization: $\ell(z; Y) = z_v + \frac{1}{\eta}\left(-z_p^\top Y - z_v\right)^+$

Coverage probability of different methods (target **90%** interval), where boldfaced values mean "almost" **valid coverage** for $c(\hat{z})$ (i.e., within [0.85, 0.95]).

	n	Plug-in	5-CV	LOOCV
kNN $k = n^{1/4}$	2400	0.00	0.00	**0.92**
	4800	0.00	0.00	**0.88**
	9600	0.00	0.00	**0.89**
RF	2400	0.77	0.66	0.72
	4800	**0.86**	0.47	**0.85**
	9600	**0.85**	0.42	**0.90**
Ridge	1200	0.78	0.55	0.78
	2400	**0.85**	**0.84**	**0.86**
	4800	**0.88**	**0.89**	**0.89**

Extensions

When $\gamma > 1/4$, all intervals provide valid coverages for:

$E_P[c(z^*(X); Y)]$) limiting decisions).

$E_Q[c(\hat{z}(X); Y)]$ with Q under covariate shift.

Other plug-in, cross validation variants:

Bias corrected CV / plug-in; Nested CV.

More efficient than LOOCV while retaining statistical guarantees.

References

[1] Pierre Bayle, Alexandre Bayle, Lucas Janson, and Lester Mackey. Cross-validation confidence intervals for test error. *NeurIPS* 2020.

[2] Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it? *JASA* 2023.

[3] Qizhao Chen, Vasilis Syrgkanis, and Morgane Austern. Debiased machine learning without sample-splitting for stable estimators. *NeurIPS* 2022.

[4] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. A distribution-free theory of nonparametric regression. 2006.

[5] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics* 2019.