Extrapolated Cross-validation for Randomized Ensembles

Jin-Hong Du[1] Pratik Patil[2]
Kathryn Roeder[1] Arun Kumar Kuchibhotla[1]

[1] Department of Statistics and Data Science, Carnegie Mellon University
[2] Department of Statistics, University of California, Berkeley

August 5th, 2023

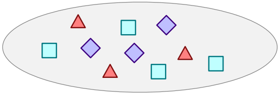
Ensemble learning

- Bagging and its variants combine multiple models, each fitted on different bootstrapped or subsampled datasets, to improve prediction accuracy and stability.

Ensemble learning

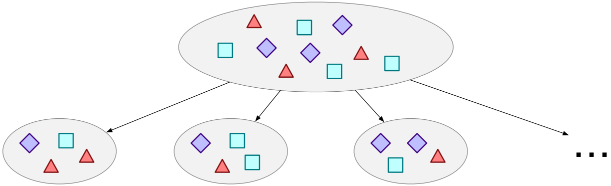
- Bagging and its variants combine multiple models, each fitted on different bootstrapped or subsampled datasets, to improve prediction accuracy and stability.

Full dataset



Ensemble learning

- Bagging and its variants combine multiple models, each fitted on different bootstrapped or subsampled datasets, to improve prediction accuracy and stability.

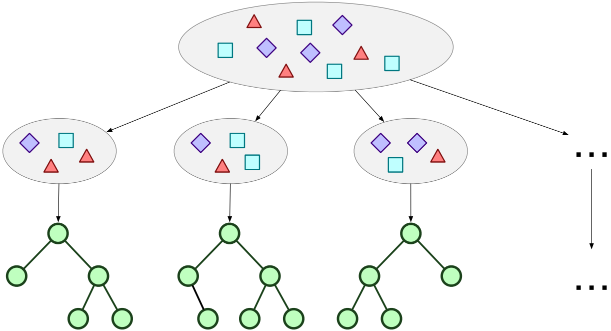


Full dataset

Subsampled/
Bootstrapped
datasets

Ensemble learning

- Bagging and its variants combine multiple models, each fitted on different bootstrapped or subsampled datasets, to improve prediction accuracy and stability.



Full dataset

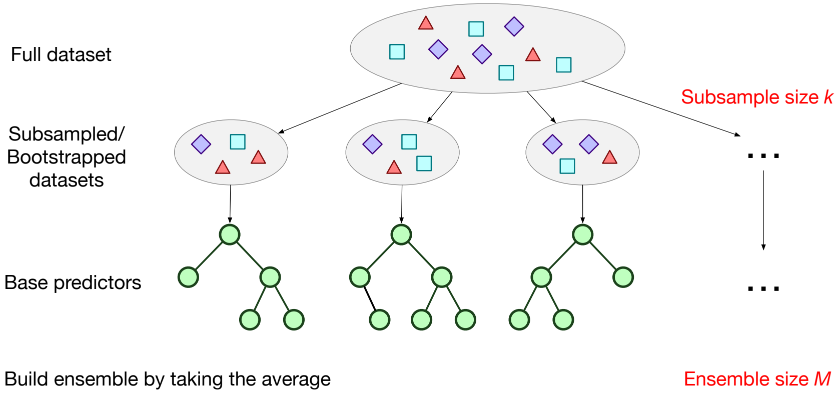
Subsampled/
Bootstrapped
datasets

Base predictors

Build ensemble by taking the average

Ensemble learning

- Bagging and its variants combine multiple models, each fitted on different bootstrapped or subsampled datasets, to improve prediction accuracy and stability.



Full dataset

Subsampled/ Bootstrapped datasets

Subsample size k

Base predictors

Build ensemble by taking the average

Ensemble size M

Ensemble tuning

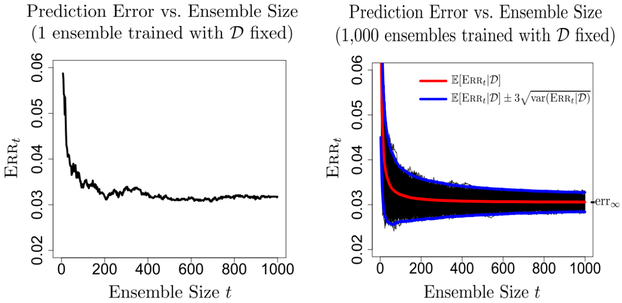
Two key parameters:

- The ensemble size M
  - Role: as $M \to \infty$, the predictive accuracy improves while variance decreases and stabilizes (algorithmic convergence[1,2]).

Ensemble tuning

Two key parameters:

- The ensemble size *M*
  - Role: as $M \to \infty$, the predictive accuracy improves while variance decreases and stabilizes (algorithmic convergence[1,2]). Figure adapted from [1].



Prediction Error vs. Ensemble Size (1 ensemble trained with $\mathcal{D}$ fixed)

Prediction Error vs. Ensemble Size (1,000 ensembles trained with $\mathcal{D}$ fixed)

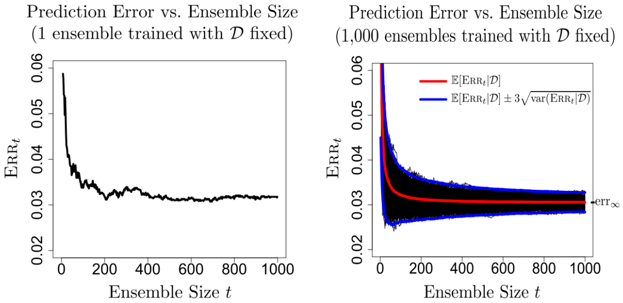
[1] Miles E Lopes. "Estimating the algorithmic variance of randomized ensembles via the bootstrap". In: *The Annals of Statistics* 47.2 (2019), pp. 1088–1112

[2] Miles E Lopes, Suofei Wu, and Thomas CM Lee. "Measuring the algorithmic convergence of randomized ensembles: The regression setting". In: *SIAM Journal on Mathematics of Data Science* 2.4 (2020), pp. 921–943

Ensemble tuning

Two key parameters:

- The ensemble size *M*
  - Role: as $M \to \infty$, the predictive accuracy improves while variance decreases and stabilizes (algorithmic convergence[1,2]). Figure adapted from [1].



Prediction Error vs. Ensemble Size (1 ensemble trained with $\mathcal{D}$ fixed)

Prediction Error vs. Ensemble Size (1,000 ensembles trained with $\mathcal{D}$ fixed)

  - The approach[1,2] relies on the convergence rate of variance or quantile estimators, to gauge the point at which the ensembles performance stabilizes as $M \to \infty$.

Ensemble tuning

Two key parameters:

- The ensemble size M
- The subsample size k

[3] Peter J Bickel, Friedrich Götze, and Willem R van Zwet. “Resampling fewer than n observations: gains, losses, and remedies for losses”. In: *Statistica Sinica* 7.1 (1997), pp. 1–31

[4] Pratik Patil, Jin-Hong Du, and Arun Kumar Kuchibhotla. “Bagging in overparameterized learning: Risk characterization and risk monotonization”. In: *arXiv preprint arXiv:2210.11445* (2022)

Ensemble tuning

Two key parameters:

- The ensemble size M
- The subsample size k
  - In low-dimensional scenarios, only a smaller k yields consistent results for k-of-n bootstrap[3].

[3] Bickel, Götze, and Zwet, “Resampling fewer than n observations: gains, losses, and remedies for losses”

[4] Patil, Du, and Kuchibhotla, “Bagging in overparameterized learning: Risk characterization and risk monotonization”

Ensemble tuning

Two key parameters:

- The ensemble size M
- The subsample size k
  - In low-dimensional scenarios, only a smaller k yields consistent results for k-of-n bootstrap[3].
  - In high-dimensional scenarios, tuning k helps to mitigate the multiple descents of the prediction risk.

[3] Bickel, Götze, and Zwet, "Resampling fewer than *n* observations: gains, losses, and remedies for losses"

[4] Patil, Du, and Kuchibhotla, "Bagging in overparameterized learning: Risk characterization and risk monotonization"

Ensemble tuning

Two key parameters:

- The ensemble size M
- The subsample size k
  - In low-dimensional scenarios, only a smaller k yields consistent results for k-of-n bootstrap[3].
  - In high-dimensional scenarios, tuning k helps to mitigate the multiple descents of the prediction risk.
  - Common tuning methods include sample-split CV[4] and K-fold CV, which are computationally and statistically inefficient.

[3] Bickel, Götze, and Zwet, "Resampling fewer than n observations: gains, losses, and remedies for losses"

[4] Patil, Du, and Kuchibhotla, "Bagging in overparameterized learning: Risk characterization and risk monotonization"

Goal

An agnostic procedure to efficiently determine (M, k) of general ensemble predictors for optimal prediction risk.

- Statistical consistency over all $M \in \mathbb{N}$ and a grid of k.
- Computational efficiency while avoiding sample splitting.
- Allow for constraints on the maximum ensemble size (δ-optimal).

Setup

- Let $\mathcal{D}_n = \{(\boldsymbol{x}_j, y_j) \in \mathbb{R}^p \times \mathbb{R} : j \in [n]\}$ denote a dataset and $I_\ell \subseteq [n]$, $\ell = 1, \ldots, M$ be independent indices with $|I_\ell| = k$.
  Given the base predictor $\widehat{f}$, a bagged predictor is defined as

$$\widetilde{f}_{M,k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^{M}) = \frac{1}{M} \sum_{\ell=1}^{M} \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}). \tag{1}$$

  The *conditional prediction risk* for a bagged predictor $\widetilde{f}_{M,k}$:

$$R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^{M}) = \int \left(y_0 - \widetilde{f}_{M,k}(\boldsymbol{x}_0; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^{M})\right)^2 \mathrm{d}P(\boldsymbol{x}_0, y_0). \tag{2}$$

Risk decomposition

- It decomposes into

$$R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) = -\left(1 - \frac{2}{M}\right) a_{1,M} + 2\left(1 - \frac{1}{M}\right) a_{2,M}, \tag{3}$$

where

$$a_{1,M} = \frac{1}{M}\sum_{\ell=1}^{M} R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_\ell\}),$$

$$a_{2,M} = \frac{1}{M(M-1)} \sum_{\substack{\ell,m\in[M]\\ \ell\neq m}} R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_\ell, I_m\}).$$

Risk decomposition

- It decomposes into

$$R(\widetilde{f}_{M,k};\, \mathcal{D}_n, \{I_\ell\}_{\ell=1}^{M}) = -\left(1 - \frac{2}{M}\right) a_{1,M} + 2\left(1 - \frac{1}{M}\right) a_{2,M}, \tag{3}$$

  where

$$a_{1,M} = \frac{1}{M}\sum_{\ell=1}^{M} R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_\ell\}),$$

$$a_{2,M} = \frac{1}{M(M-1)} \sum_{\substack{\ell,m\in[M]\\ \ell\neq m}} R(\widetilde{f}_{2,k};\, \mathcal{D}_n, \{I_\ell, I_m\}).$$

- $a_{1,M}$ and $a_{2,M}$ are $\mathcal{D}_n$-conditional U-statiatics of 1-bagged and 2-bagged risks!

Risk estimation for $M = 1, 2$

Proposition

Let $\widehat{\sigma}_I := \|y_0 - \widehat{f}(\boldsymbol{x}_0; \mathcal{D}_I)\|_{\psi_1 \mid \mathcal{D}_I}$ *be the variance proxy. If* $\widehat{\sigma}_I / \sqrt{|I^c| / \log n} \xrightarrow{\text{p}} 0$*, then*

$$\Big| \underbrace{\widehat{R}(\widehat{f}; \mathcal{D}_{I^c})}_{\textit{OOB estimate}} - \underbrace{R(\widehat{f}; \mathcal{D}_I)}_{\textit{risk}} \Big| \xrightarrow{\text{p}} 0.$$

Risk estimation for $M = 1, 2$

Proposition

Let $\widehat{\sigma}_I := \|y_0 - \widehat{f}(\boldsymbol{x}_0; \mathcal{D}_I)\|_{\psi_1|\mathcal{D}_I}$ *be the variance proxy. If* $\widehat{\sigma}_I / \sqrt{|I^c| / \log n} \xrightarrow{\text{p}} 0$*, then*

$$|\underbrace{\widehat{R}(\widehat{f}; \mathcal{D}_{I^c})}_{\textit{OOB estimate}} - \underbrace{R(\widehat{f}; \mathcal{D}_I)}_{\textit{risk}}| \xrightarrow{\text{p}} 0.$$

- For linear models ($y_0 = \boldsymbol{x}_0^\top \beta_0 + \epsilon$) and linear predictors ($\widehat{f}(\boldsymbol{x}_0; \mathcal{D}_I) = \boldsymbol{x}_0^\top \widehat{\beta}(\mathcal{D}_I)$), $\widehat{\sigma}_I$ is simply $\|\widehat{\beta}(\mathcal{D}_I) - \beta_0\|_{\boldsymbol{\Sigma}}$ (generally bounded, e.g. for ridge predictors).

Risk estimation for $M = 1, 2$

Proposition

Let $\widehat{\sigma}_I := \|y_0 - \widehat{f}(\boldsymbol{x}_0; \mathcal{D}_I)\|_{\psi_1|\mathcal{D}_I}$ *be the variance proxy. If* $\widehat{\sigma}_I/\sqrt{|I^c|/\log n} \xrightarrow{\text{p}} 0$*, then*

$$\Big|\, \underbrace{\widehat{R}(\widehat{f};\mathcal{D}_{I^c})}_{\textit{OOB estimate}} - \underbrace{R(\widehat{f};\mathcal{D}_I)}_{\textit{risk}} \Big| \xrightarrow{\text{p}} 0.$$

- For linear models ($y_0 = \boldsymbol{x}_0^\top \beta_0 + \epsilon$) and linear predictors ($\widehat{f}(\boldsymbol{x}_0;\mathcal{D}_I) = \boldsymbol{x}_0^\top \widehat{\beta}(\mathcal{D}_I)$), $\widehat{\sigma}_I$ is simply $\|\widehat{\beta}(\mathcal{D}_I) - \beta_0\|_{\boldsymbol{\Sigma}}$ (generally bounded, e.g. for ridge predictors).
- Aggregate individual OOB estimates yields more stable risk estimates for $M = 1, 2$:

$$\widehat{R}^{\mathrm{ECV}}_{M,k} = \begin{cases} \frac{1}{M_0}\sum_{\ell=1}^{M_0} \widehat{R}(\widetilde{f}_{1,k}(\cdot;\mathcal{D}_n,\{I_\ell\}),\mathcal{D}_{I_\ell^c}), & M = 1, \\ & \end{cases} \tag{4}$$

Risk estimation for $M = 1, 2$

Proposition

Let $\widehat{\sigma}_I := \|y_0 - \widehat{f}(\boldsymbol{x}_0; \mathcal{D}_I)\|_{\psi_1|\mathcal{D}_I}$ *be the variance proxy. If* $\widehat{\sigma}_I / \sqrt{|I^c| / \log n} \xrightarrow{\text{p}} 0$*, then*

$$\Big| \underbrace{\widehat{R}(\widehat{f}; \mathcal{D}_{I^c})}_{\textit{OOB estimate}} - \underbrace{R(\widehat{f}; \mathcal{D}_I)}_{\textit{risk}} \Big| \xrightarrow{\text{p}} 0.$$

- For linear models ($y_0 = \boldsymbol{x}_0^\top \beta_0 + \epsilon$) and linear predictors ($\widehat{f}(\boldsymbol{x}_0; \mathcal{D}_I) = \boldsymbol{x}_0^\top \widehat{\beta}(\mathcal{D}_I)$), $\widehat{\sigma}_I$ is simply $\|\widehat{\beta}(\mathcal{D}_I) - \beta_0\|_{\boldsymbol{\Sigma}}$ (generally bounded, e.g. for ridge predictors).
- Aggregate individual OOB estimates yields more stable risk estimates for $M = 1, 2$:

$$\widehat{R}^{\mathrm{ECV}}_{M,k} = \begin{cases} \frac{1}{M_0} \sum\limits_{\ell=1}^{M_0} \widehat{R}(\widetilde{f}_{1,k}(\cdot; \mathcal{D}_n, \{I_\ell\}), \mathcal{D}_{I_\ell^c}), & M = 1, \\ \frac{1}{M_0(M_0-1)} \sum\limits_{\substack{\ell, m \in [M_0] \\ \ell \neq m}} \widehat{R}(\widetilde{f}_{2,k}(\cdot; \mathcal{D}_n, \{I_\ell, I_m\}), \mathcal{D}_{(I_\ell \cup I_m)^c}), & M = 2, \end{cases} \tag{4}$$

Extrapolated cross-validation

- Extrapolate the risk estimations $\widehat{R}^{\mathrm{ECV}}_{M,k}$ using

$$\widehat{R}^{\mathrm{ECV}}_{{\color{brown}M},k} = -\left(1-\frac{2}{M}\right)\widehat{R}^{\mathrm{ECV}}_{{\color{red}1},k} + 2\left(1-\frac{1}{M}\right)\widehat{R}^{\mathrm{ECV}}_{{\color{green}2},k}, \quad M > 2.$$

Extrapolated cross-validation

- Extrapolate the risk estimations $\widehat{R}^{\mathrm{ECV}}_{M,k}$ using

$$\widehat{R}^{\mathrm{ECV}}_{M,k} = -\left(1 - \frac{2}{M}\right) \widehat{R}^{\mathrm{ECV}}_{1,k} + 2\left(1 - \frac{1}{M}\right) \widehat{R}^{\mathrm{ECV}}_{2,k}, \quad M > 2.$$

Theorem (Uniform consistency of risk extrapolation)

Under certain conditions, ECV estimates satisfy that

$$\sup_{M \in \mathbb{N}, k \in \mathcal{K}_n} \left| \widehat{R}^{\mathrm{ECV}}_{M,k} - R_{M,k} \right| = \mathcal{O}_p(\zeta_n),$$

Extrapolated cross-validation

- Extrapolate the risk estimations $\widehat{R}^{\mathrm{ECV}}_{M,k}$ using

$$\widehat{R}^{\mathrm{ECV}}_{M,k} = -\left(1 - \frac{2}{M}\right)\widehat{R}^{\mathrm{ECV}}_{1,k} + 2\left(1 - \frac{1}{M}\right)\widehat{R}^{\mathrm{ECV}}_{2,k}, \quad M > 2.$$

Theorem (Uniform consistency of risk extrapolation)

Under certain conditions, ECV estimates satisfy that

$$\sup_{M\in\mathbb{N}, k\in\mathcal{K}_n} \left| \widehat{R}^{\mathrm{ECV}}_{M,k} - R_{M,k} \right| = \mathcal{O}_p(\zeta_n),$$

where

$$\zeta_n = \underbrace{\widehat{\sigma}_n \frac{\log n}{\sqrt{n}}}_{\textit{CV error}} + \underbrace{n^{\epsilon}(\gamma_{1,n} + \gamma_{2,n})}_{\textit{convergence rate for } M = 1,2} .$$

Extrapolated cross-validation

- Extrapolate the risk estimations $\widehat{R}^{\text{ECV}}_{M,k}$ using

$$\widehat{R}^{\text{ECV}}_{M,k} = -\left(1 - \frac{2}{M}\right)\widehat{R}^{\text{ECV}}_{1,k} + 2\left(1 - \frac{1}{M}\right)\widehat{R}^{\text{ECV}}_{2,k}, \quad M > 2.$$

- Tuning: Select a subsample size $\widehat{k} \in \mathcal{K}_n$ and a *smallest* ensemble size $\widehat{M} \in \mathbb{N}$ such that $\widehat{R}^{\text{ECV}}_{\widehat{M},\widehat{k}}$ is δ-close to the oracle.

Extrapolated cross-validation

- Extrapolate the risk estimations $\widehat{R}^{\mathrm{ECV}}_{M,k}$ using

$$\widehat{R}^{\mathrm{ECV}}_{M,k} = -\left(1-\frac{2}{M}\right)\widehat{R}^{\mathrm{ECV}}_{1,k} + 2\left(1-\frac{1}{M}\right)\widehat{R}^{\mathrm{ECV}}_{2,k}, \quad M > 2.$$

- Tuning: Select a subsample size $\widehat{k} \in \mathcal{K}_n$ and a *smallest* ensemble size $\widehat{M} \in \mathbb{N}$ such that $\widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}}$ is δ-close to the oracle.

Theorem (Sub-optimality of the tuned risk (w.r.t. the infinite-ensemble))

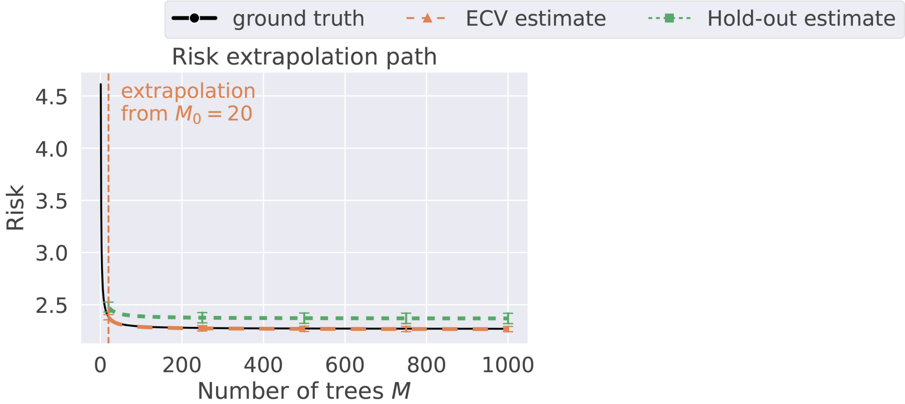
$$\left| R_{\widehat{M},\widehat{k}} - \inf_{M\in\mathbb{N},k\in\mathcal{K}_n} R_{M,k} \right| = \delta + \mathcal{O}_p(\zeta_n).$$

Experiment

- Tuning ensemble sizes of random forests ($n = 1,000$):

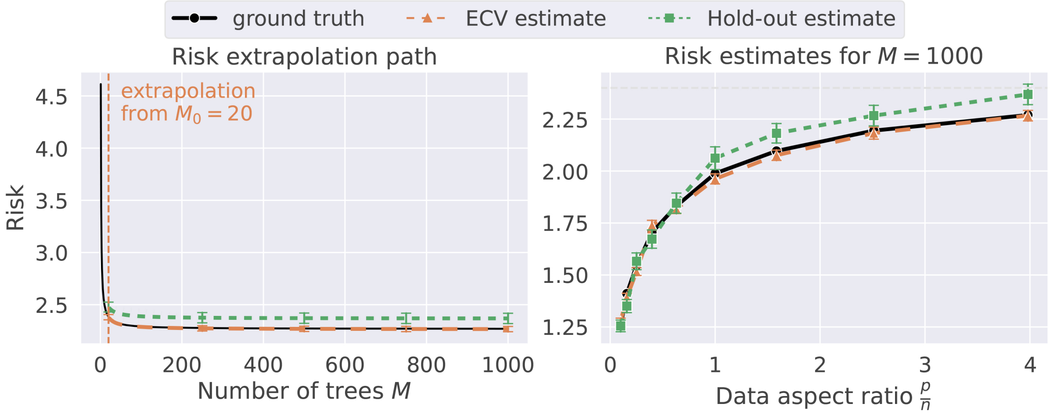
Experiment

- Tuning ensemble sizes of random forests ($n = 1,000$):



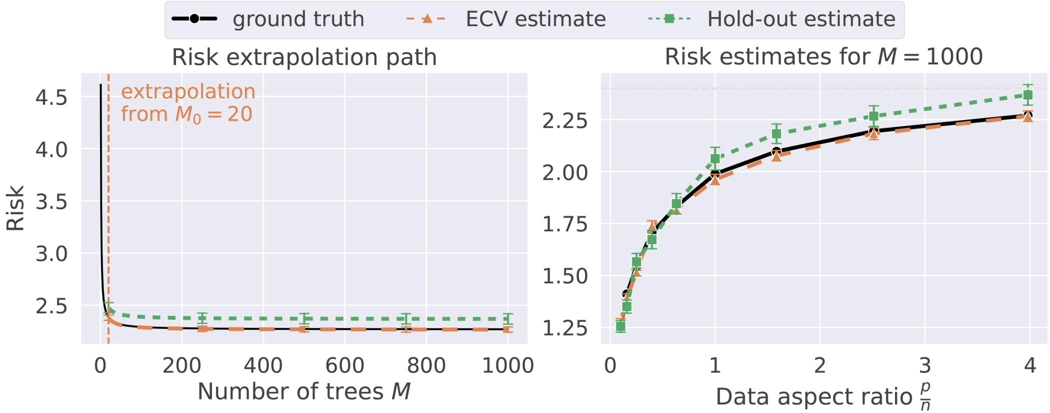
Experiment

- Tuning ensemble sizes of random forests ($n = 1,000$):



Experiment

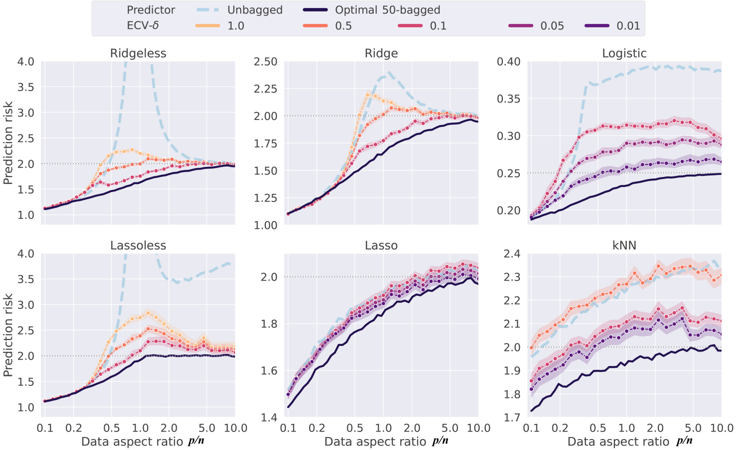
- Tuning ensemble sizes of random forests ($n = 1,000$):



ECV estimates provide valid extrapolation paths in both low- and high-dimensional scenarios.

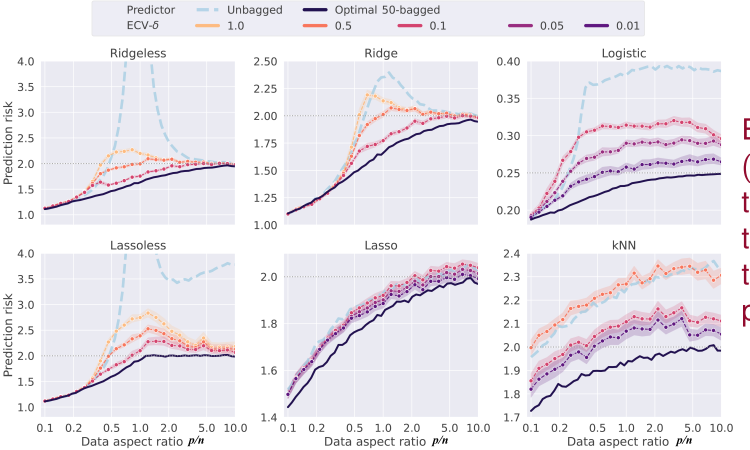
Experiment

- Tuning ensemble and subsample sizes with $M_{\max} = 50$:



Experiment

- Tuning ensemble and subsample sizes with $M_{\max} = 50$:



ECV-tuned parameters $(\widehat{M}, \widehat{k})$ give risks close to the oracle choices within the desired optimality threshold δ in finite samples.

Single-cell sequencing multiomic datasets

- Gene expressions ($X \in \mathbb{R}^{5,000}$) and protein abundances ($Y \in \mathbb{R}^{50}$) in each cell are measured.

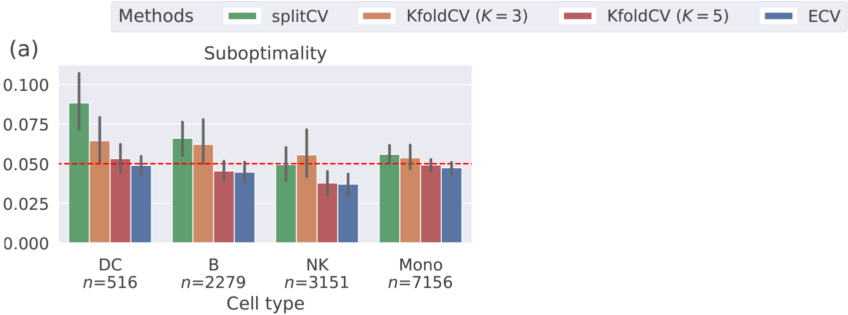
Single-cell sequencing multiomic datasets

- Gene expressions ($X \in \mathbb{R}^{5,000}$) and protein abundances ($Y \in \mathbb{R}^{50}$) in each cell are measured.
- We use all the gene expressions to predict the abundance of each protein.

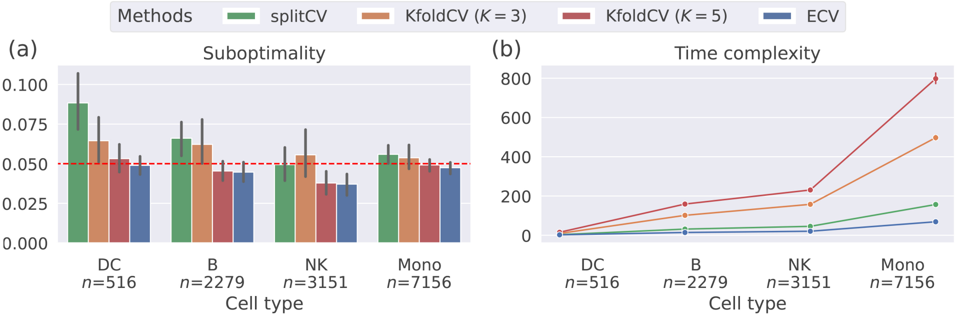
Single-cell sequencing multiomic datasets

- Gene expressions ($X \in \mathbb{R}^{5,000}$) and protein abundances ($Y \in \mathbb{R}^{50}$) in each cell are measured.
- We use all the gene expressions to predict the abundance of each protein.
- Our target is to select a δ-optimal random forest so that its prediction risk is no more than $\delta = 0.05$ away from the best random forest with 50 trees.

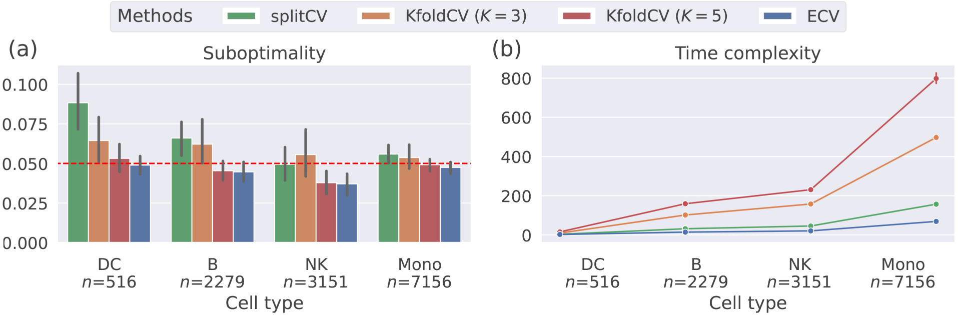
Single-cell sequencing multiomic datasets



Single-cell sequencing multiomic datasets



Single-cell sequencing multiomic datasets



Better out-of-sample errors and time complexity!

Thanks for your attention!
Any questions?