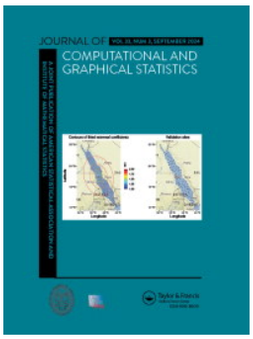

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Extrapolated Cross-Validation for Randomized Ensembles

Jin-Hong Du, Pratik Patil, Kathryn Roeder & Arun Kumar Kuchibhotla

To cite this article: Jin-Hong Du, Pratik Patil, Kathryn Roeder & Arun Kumar Kuchibhotla (2024) Extrapolated Cross-Validation for Randomized Ensembles, Journal of Computational and Graphical Statistics, 33:3, 1061-1072, DOI: 10.1080/10618600.2023.2288194

To link to this article: https://doi.org/10.1080/10618600.2023.2288194

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2024, VOL. 33, NO. 3, 1061–1072
https://doi.org/10.1080/10618600.2023.2288194

Extrapolated Cross-Validation for Randomized Ensembles

Jin-Hong Du[a,b], Pratik Patil[c], Kathryn Roeder[a], and Arun Kumar Kuchibhotla[a]

[a]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA; [b]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA; [c]Department of Statistics, University of California, Berkeley, Berkeley, CA

ABSTRACT

Ensemble methods such as bagging and random forests are ubiquitous in various fields, from finance to genomics. Despite their prevalence, the question of the efficient tuning of ensemble parameters has received relatively little attention. This article introduces a cross-validation method, Extrapolated Cross-Validation (ECV), for tuning the ensemble and subsample sizes in randomized ensembles. Our method builds on two primary ingredients: initial estimators for small ensemble sizes using out-of-bag errors and a novel risk extrapolation technique that leverages the structure of prediction risk decomposition. By establishing uniform consistency of our risk extrapolation technique over ensemble and subsample sizes, we show that ECV yields δ-optimal (with respect to the oracle-tuned risk) ensembles for squared prediction risk. Our theory accommodates general predictors, only requires mild moment assumptions, and allows for high-dimensional regimes where the feature dimension grows with the sample size. As a practical case study, we employ ECV to predict surface protein abundances from gene expressions in single-cell multiomics using random forests under a computational constraint on the maximum ensemble size. Compared to sample-split and K-fold cross-validation, ECV achieves higher accuracy by avoiding sample splitting. Meanwhile, its computational cost is considerably lower owing to the use of the risk extrapolation technique. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2023
Accepted November 2023

KEYWORDS
Bagging; Distributed learning; Ensemble learning; Random forest; Risk extrapolation; Tuning and model selection

1. Introduction

Bagging and its variants are popular randomized ensemble methods in statistics and machine learning. These methods combine multiple models, each fitted on different bootstrapped or subsampled datasets, to improve prediction accuracy and stability (Breiman 1996; Pugh 2002), which is well-suited for large-scale distributed computation. The success of these methods lies in the careful tuning of key parameters: the *ensemble size* M and the *subsample size* k (Hastie, Tibshirani, and Friedman 2009). As M grows, the predictive accuracy improves while prediction variance decreases and stabilizes, a concept known as algorithmic convergence (Lopes 2019; Lopes, Wu, and Lee 2020). However, in the era of big data (Politis 2023), achieving a precise approximation in the infinity ensemble is challenging due to computational costs. This necessitates the selection of a suitable value of M to strike a balance between data-dependent considerations and budget constraints, without the requirement for it to scale proportionally with the sample size. Further, the number of subsampled/bootstrapped observations, k, used for each predictor plays a crucial role in ensemble learning (Martínez-Muñoz and Suárez 2010). In low-dimensional scenarios, a smaller k can yield consistent results (Politis and Romano 1994; Bickel, Götze, and van Zwet 1997); however, in high-dimensional scenarios, the prediction risk may not have a straightforward monotonic relationship with subsample size, exhibiting instead multiple descent behaviors (Chen et al. 2020; Hastie et al. 2022; Patil et al. 2022). By restricting the subsample size, there has been some theoretical work on random forests showing that consistency and/or asymptotic normality can be achieved under certain regularity conditions (Scornet, Biau, and Vert 2015; Mentch and Hooker 2016; Wager and Athey 2018; Peng, Coleman, and Mentch 2022). Selecting the right subsample size is thus also of paramount importance for optimal predictive performance.

Several strategies have been proposed to tune either the ensemble size M or the subsample size k. For instance, to choose an ensemble size M, a variance stabilization strategy is proposed by Lopes (2019) and Lopes, Wu, and Lee (2020). This approach relies on the convergence rate of variance or quantile estimators, using these metrics to gauge the point at which the ensemble's performance stabilizes as M approaches infinity. Such an approach effectively helps reduce computational expenses. On the other hand, determining the optimal subsample size k is a more difficult task and generally tuned by standard cross-validation (CV) methods. As one of the most basic of the CV methods, sample-split CV estimates the predictive risk of every predictor associated with a configuration of parameters using independent hold-out observations (Patil, Du, and Kuchibhotla 2023). Another commonly used CV method, K-fold CV, repeatedly fits each candidate predictor on K different subsets of the data and uses their average to estimate the prediction risk.

CONTACT Jin-Hong Du *jinhongd@andrew.cmu.edu* Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

While these aforementioned methods offer ways to tune M and k, they have several drawbacks. In terms of turning over the ensemble size M, the specialized method proposed by Lopes (2019) and Lopes, Wu, and Lee (2020) involves monitoring the variability of the test errors as a function of M. However, as this approach focuses solely on *the scale of variance of the risk* rather than the *prediction risk* itself, it does not provide any suboptimality guarantee compared to the optimal risk of an infinite ensemble (LeJeune, Javadi, and Baraniuk 2020). Furthermore, the method does not provide any estimators for the prediction risks for any finite ensemble size M. In terms of tuning the subsample size k, the traditional CV methods such as sample-split CV and K-fold CV can be significantly impacted by finite-sample effects due to sample splitting, particularly in high-dimensional scenarios (Wang et al. 2018; Rad and Maleki 2020; Patil, Du, and Kuchibhotla 2023). Furthermore, these generic CV methods must evaluate every possible ensemble and subsample size within an arbitrarily chosen search space, which often requires exploring larger ensemble sizes, thus, demanding more computational resources. Yet, even with these, certifying any optimality outside this predefined search is not generally possible. These drawbacks highlight the need for more efficient tuning methods for ensemble learning.

This article seeks to address both the theoretical and practical challenges associated with ensemble parameter tuning. To this end, we develop a CV method that can efficiently and consistently tune both the ensemble and subsample sizes. We focus primarily on randomized ensembles such as bagging and subagging (Breiman 1996; Bühlmann and Yu 2002) and rely on out-of-bag observations to estimate the conditional prediction risk. Our proposed method, termed ECV (Extrapolated Cross-Validation), enjoys several advantages over the previously mentioned approaches. We highlight two of them below: (a) *Statistical consistency*: ECV is versatile and model-agnostic and is applicable to general ensemble predictors. It provides uniform consistency in estimating the actual prediction risk of ensembles over all ensemble and subsample sizes under mild conditions. It also notably outperforms standard CV methods in finite samples, especially in high-dimensional settings. (b) *Computational efficiency*: ECV operates by estimating the risk of ensembles with small ensemble sizes ($M = 1, 2$) using out-of-bag observations and then extrapolates the risk estimates to arbitrary ensemble sizes. Unlike sample-split and K-fold CV, our method does not require fitting an ensemble for every ensemble size or explicitly estimating prediction risks for every ensemble size, and can serve as a valuable tool for assessing the fitness of ensemble predictors. Though our focus in this article is on ensemble and subsample size tuning, ECV is flexible for tuning other hyperparameters efficiently by risk extrapolation. As a result, ECV significantly reduces the computational burden, making it an efficient method for ensemble parameter tuning.

Before delving into the details of our method, we take the opportunity to demonstrate these key points through a real-world example. We apply ECV on four single-cell datasets and aim to select a δ-optimal random forest in Section 6, so that its prediction risk is no more than $\delta = 0.05$ away from the best random forest consisting of 50 trees. In Figure 1, we compare the performance of the sample-split CV (Patil, Du, and Kuchibhotla 2023), K-fold CV with $K = 3, 5$, and our method ECV applied to four datasets, each corresponding to a different cell type obtained from (Hao et al. 2021). Further details regarding this application can be found in Section 6. Both of the commonly used CV methods require estimating the risks with all possible choices of ensemble size M and subsample size k. To ensure a fair comparison, we use the same search space of (M, k) for all methods. Because sample splitting introduces additional

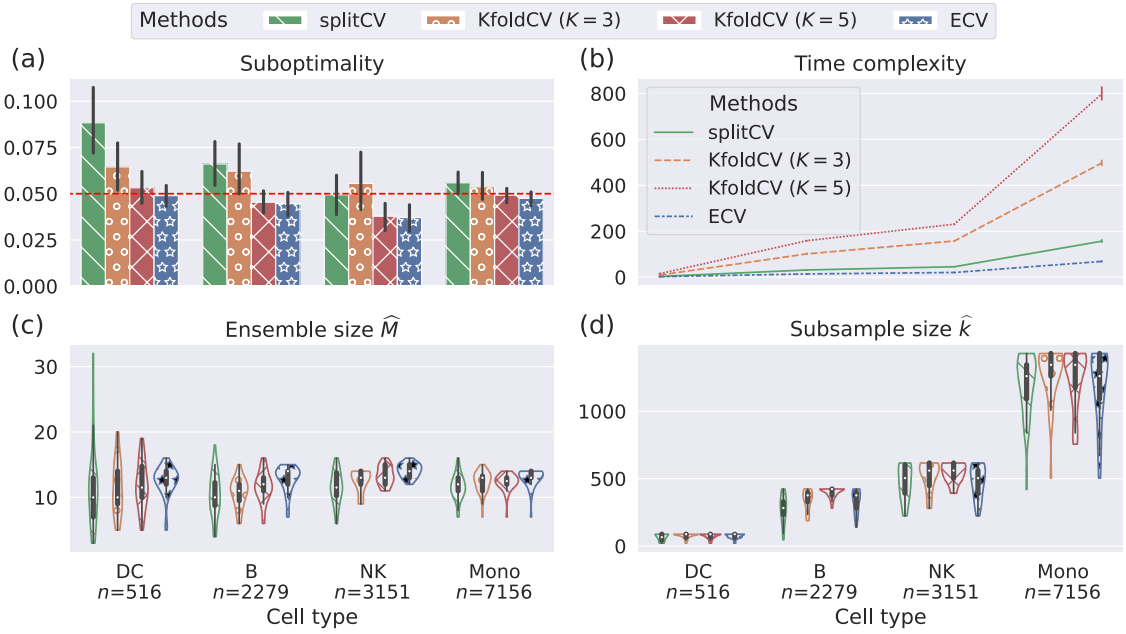

Figure 1. Overview of different cross-validation methods for predicting 50 surface proteins based on 5000 genes in the single-cell sequencing multiomic datasets from Hao et al. (2021) using random forests. The proposed ECV method estimates the prediction risks based on 20 trees. Each panel shows cell types DC, B, NK, and Mono. (a) The out-of-sample suboptimality compared to the optimal random forest with 50 trees. The CV predictors are tuned to be δ-optimal in terms of normalized mean squared errors. The error bars show the standard deviations over 50 proteins, and the red dashed line indicates the optimality threshold $\delta = 0.05$. (b) The time consumption in seconds. (c) The cross-validated ensemble size $\widehat{M}$. (d) The cross-validated subsample size $\widehat{k}$.

randomness and the reduced sample size has significant finite sample effects, we observe from Figure 1(a) that sample-split CV does not control the out-of-sample error within the specified tolerance of $\delta = 0.05$ away from the best possible error. On the other hand, even though K-fold CV gives the valid error control as ECV, it costs extra computational time, which significantly increases as the sample size increases; see Figure 1(b). Overall, the distributions of tuned ensemble parameters $(\widehat{M}, \widehat{k})$ are similar between K-fold CV and ECV; see Figure 1(c)–(d). These results demonstrate the practical effectiveness of ECV in addressing the above drawbacks for ensemble parameter tuning.

We next provide an overview of the main results and delineate the structure for the remainder of the article. In Section 2, we outline the setup of the tuning problem in the context of randomized ensemble learning. We also provide a comprehensive review of prior work on cross-validation and tuning and contrast our method to earlier work. In Section 3, we lay the foundation for our proposed method by constructing the theoretical ingredients essential to our main algorithm, in particular, the extrapolation risk estimation strategy. Through a nonasymptotic analysis, we further demonstrate the convergence rate and uniform consistency of the proposed risk estimator for general predictors and data distributions under a mild moment condition (see Theorem 1).

In Section 4, we introduce our main algorithm and discuss its theoretical properties and various practical considerations. We prove that the resulting tuned ensemble obtains the best possible ensemble risk over all ensemble and subsample sizes up to a specified tolerance of δ (see Theorem 2). In Section 5, we examine ECV's generality and effectiveness with various types of predictors. In Section 6, comparisons with sample-split and K-fold CV in the protein prediction problem highlight the statistical and computational benefits of ECV on low- and high-dimensional datasets, under a computational budget for the maximum ensemble size. In Section 7, we conclude the article with a brief discussion that acknowledges some limitations of the method and provides avenues for future work. The code to replicate all our experiments can be obtained from *https://jaydu1.github.io/overparameterized-ensembling/ecv* and the Python package implementing the ECV method can be found on the GitHub repository *https://github.com/jaydu1/ensemble-cross-validation*.

2. Randomized Ensembles

We consider a supervised regression setup. Suppose $\mathcal{D}_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ represents a dataset with independent and identically distributed random vectors from $\mathbb{R}^p \times \mathbb{R}$. We will not assume any specific data model, only that the second moment of the response is finite, that is, $\mathbb{E}(y_1^2) < \infty$. A prediction procedure $\widehat{f}(\cdot;\cdot)$ is defined as a map from $\mathbb{R}^p \times \mathscr{P}(\mathcal{D}_n) \to \mathbb{R}$, where $\mathscr{P}(A)$, for any set A, represents the power set of A.

Bagging (as in bootstrap-aggregating) traditionally refers to computing predictors multiple times based on bootstrapped data (Breiman 1996), which can involve repeated observations. There is another version of bagging called subagging (as in subsample-aggregating) in Bühlmann and Yu (2002, sec. 3.2) where we sample observations without replacement. Our method and analysis apply to both of these sampling strategies. Formally, these can be understood as simple random samples from a finite set, commonly used in survey sampling. Fix any $k \in [n]$, we define the indices $\{I_\ell\}_{\ell=1}^M$ to be M independent samples with replacement from $\mathcal{I}_k$ (denoted by $\{I_\ell\}_{\ell=1}^M \overset{\text{SRS}}{\sim} \mathcal{I}_k$). Here $\mathcal{I}_k$ is defined for bagging and subagging as

$$\mathcal{I}_k = \begin{cases} \{\{i_1, i_2, \ldots, i_k\} : 1 \le i_1 \le i_2 \le \cdots \le i_k \le n\}, & \text{(bagging)} \\ \{\{i_1, i_2, \ldots, i_k\} : 1 \le i_1 < i_2 < \cdots < i_k \le n\}. & \text{(subagging)} \end{cases} \tag{1}$$

For bagging, $\mathcal{I}_k$ represents the set of all possible independent draws from $[n]$ with replacement, and there are n^k many of them. For subagging, $\mathcal{I}_k$ represents the set of all k subset choices from $[n]$, and there are $n!/(k!(n-k)!)$ many of them. Throughout the paper, we mainly focus on subagging but the results apply equally well to bagging. For this reason, we do not distinguish different choices of $\mathcal{I}_k$. For any $I \in \mathcal{I}_k$, let $\mathcal{D}_I$ and the corresponding subsampled predictor be defined as $\mathcal{D}_I = \{(\boldsymbol{x}_j, y_j) : j \in I\}$ and $\widehat{f}(\boldsymbol{x}; \mathcal{D}_I) = \widehat{f}(\boldsymbol{x}; \{(\boldsymbol{x}_j, y_j) : j \in I\})$. Then the randomized ensemble using either bootstrap or subsampling is defined as follows:

$$\begin{aligned} &\widetilde{f}_{M,k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) \\ &\quad = \frac{1}{M}\sum_{\ell=1}^M \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}) \quad \text{with} \quad I_1, \ldots, I_M \overset{\text{SRS}}{\sim} \mathcal{I}_k. \end{aligned} \tag{2}$$

In the context where we want to highlight the size of bootstrap/subsample, k, we write $I_{k,1}, \ldots, I_{k,M}$ instead of $I_1, \ldots, I_M$.

We are interested in the performance of our predictors (computed on $\mathcal{D}_n$) on future data from the same distribution P. We consider the behavior of the predictors conditional on $\mathcal{D}_n$ and $\{I_\ell\}_{\ell=1}^M$. More specifically, for a predictor $\widehat{f}$ fitted on $\mathcal{D}_n$ and its bagged predictor $\widetilde{f}_{M,k}$ fitted on $\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M$, with $\{I_\ell\}_{\ell=1}^M \overset{\text{SRS}}{\sim} \mathcal{I}_k$, the data and subsample conditioned risks are defined as:

$$\begin{aligned} &R(\widehat{f}; \mathcal{D}_n) = \int (y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_n))^2 \, \mathrm{d}P(\boldsymbol{x}, y), \\ &R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) \\ &\quad = \int \left(y - \widetilde{f}_{M,k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\right)^2 \, \mathrm{d}P(\boldsymbol{x}, y). \end{aligned} \tag{3}$$

The two critical quantities for ensemble learning are the ensemble size M and the subsample size k. Different values of M tradeoff model stability and computational burden. As M increases, the bagged predictors get more stabilized while requiring more time to fit them. On the other hand, the subsample size trades off the bias and variance of the bagged predictors. A smaller subsample size has a considerable bias but may reduce the variance. For example, in the context of subagging minimum norm least square predictors with $M = \infty$, a properly chosen subsample size k strictly less than n can have a higher variance reduction compared to the inflation of bias. This raises the question: how to efficiently choose both the ensemble size (M) and the subsample size (k) to minimize prediction risk (3) for general predictors. We address this question in the next section. Before that, we review some related work on cross-validation and situate our work in the context of other related work.

There is extensive literature on cross-validation (CV) approaches; see Appendix S1 for a detailed survey. In the context of bagging and subagging, Liu, Chin, and Tran (2019) study parameters selection for bagging in sparse regression based on the derived error bound. For subsample size tuning, a sample-split CV method is proposed in (Patil et al. 2022; Patil, Du, and Kuchibhotla 2023). To estimate the prediction risk without sample splitting, the other line of research uses the out-of-bag (OOB) observations (Breiman 2001). For example, the algorithmic variance of ensemble regression functions at a fixed test point is studied in Oshiro, Perez, and Baranauskas (2012), Wager, Hastie, and Efron (2014); in Lopes (2019) and Lopes, Wu, and Lee (2020), the authors extrapolate the algorithmic variance and quantile of random forests for classification and regression problems, respectively. Their extrapolated estimators based on the heuristic scaling improve computation empirically, but theoretically, the statistical property is still unclear. Politis (2023) discuss scalable subbagging estimator when the subsample size k and the ensemble size M scale with the sample size n.

This article differs from the previously mentioned works in two significant aspects. First, the consistency of sample-split CV is shown in Patil, Du, and Kuchibhotla (2023) for a fixed ensemble size M for subagging, which suffers from the finite-sample effects because of sample splitting. Additionally, their results rely on stringent assumptions that require the asymptotic risk to satisfy certain analytic properties and do not provide any convergence rates. In contrast, our work establishes uniform consistency over both the ensemble size M and the subsample size k for bagging as well as subagging. More importantly, we also characterize the proposed estimators' convergence rate and require much milder assumptions on the risk, in the form of certain moment conditions. Second, Lopes (2019) and Lopes, Wu, and Lee (2020) rely on the convergence rate of variance to extrapolate the fluctuations of the estimates and require the ensemble size M to approach infinity. In contrast, ECV directly estimates the extrapolated risk (not just the scale of variances or quantiles) for an arbitrary range of ensemble sizes in a consistent manner. Additionally, while their papers only focus on tuning the ensemble size M, we also tune the subsample size k, which is crucial to minimizing the predictive risks, especially in high-dimensional scenarios, as alluded to in the introduction.

3. Method Motivation

In this section, we derive preliminary results that serve as the foundation for our extrapolated cross-validation method in Section 4. Our approach relies on utilizing out-of-bag (OOB) observations to tune both the ensemble size M and the subsample size k. Let us now describe the main components behind our methodology.

3.1. Decomposition and Risk Estimation

To begin with, we will fix k and focus on tuning over $M \in \mathbb{N}$. Recall the conditional risk $R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M)$ associated with an M-bagged predictor $\widetilde{f}_{M,k}$, as defined in (3). The subsequent proposition demonstrates that $R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M)$ can be expressed as a linear combination of the conditional risks for $M = 1$ and $M = 2$.

Proposition 1 (Squared conditional risk decomposition). The conditional prediction risk defined in (3) for a bagged predictor $\widetilde{f}_{M,k}$ decomposes into

$$R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) = -\left(1 - \frac{2}{M}\right) a_{1,M} + 2\left(1 - \frac{1}{M}\right) a_{2,M}, \tag{4}$$

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

The proof of Proposition 1 can be found in Appendix S2.1. The statement follows due to a special decomposition that governs the squared risk of the M-bagged predictor. The components $a_{1,M}$ and $a_{2,M}$ in the decomposition (4) are the averages of the 1-bagged and 2-bagged conditional risks, respectively. Conditional on $\mathcal{D}_n$, note that $a_{1,M}$ and $a_{2,M}$ are U-statistic based on iid elements $\{I_\ell\}_{\ell=1}^M$ sampled from $\mathcal{I}_k$.

Toward motivating our ECV method, let us assume that there exist constants $\mathfrak{R}_{1,k}$ and $\mathfrak{R}_{2,k}$ such that as $n \to \infty$, both $a_{1,M} - \mathfrak{R}_{1,k}$ and $a_{2,M} - \mathfrak{R}_{2,k}$ converge almost surely to 0. Then Proposition 1 implies that the conditional prediction risk of $\widetilde{f}_{M,k}$ can be approximated asymptotically by $-(1-2/M)\mathfrak{R}_{1,k} + 2(1-1/M)\mathfrak{R}_{2,k}$. This serves as the basis for the concept of ECV, where consistent estimation of $\mathfrak{R}_{1,k}$ and $\mathfrak{R}_{2,k}$ allows for a consistent estimator of the risk of $\widetilde{f}_{M,k}$ for every $M \in \mathbb{N}$, hence, justifying the name "extrapolated" cross-validation.

To consistently estimate the basic components $\mathfrak{R}_{1,k}$ and $\mathfrak{R}_{2,k}$ in (4), we first leverage the OOB risk estimator for an arbitrary predictor $\widehat{f}$ fitted on $\mathcal{D}_I$ based on the OOB test dataset $\mathcal{D}_{I^c} = \mathcal{D}_n \setminus \mathcal{D}_I$. One can simply consider the average of the squared loss (MEAN) on OOB observations in $\mathcal{D}_{I^c}$ as a risk estimator. This choice, however, is not suitable for heavy-tailed data and hence we also consider a median-of-means (MOM) estimator:

$$\widehat{R}(\widehat{f}, \mathcal{D}_{I^c}) = \begin{cases} \frac{1}{|I^c|}\sum_{i\in I^c}(y_i - \widehat{f}(\boldsymbol{x}_i))^2, & \text{if EST = MEAN,} \\ \text{median}\left\{\frac{1}{|I^{(b)}|}\sum_{i\in I^{(b)}}(y_i - \widehat{f}(\boldsymbol{x}_i))^2,\ b \in [B]\right\}, & \text{if EST = MOM,} \end{cases} \tag{5}$$

with $B = \lceil 8\log(1/\eta)\rceil$ and $I^{(1)}, \ldots, I^{(B)}$ being $B = \lceil 8\log(1/\eta)\rceil$ random splits of I^c for some $\eta > 0$. The median-of-means estimator was developed for heavy-tailed mean estimation and is commonly used in robust statistics (Lugosi and Mendelson 2019).

We will provide a condition to certify the pointwise consistency of risk estimates $\widehat{R}$ under certain assumptions on the data distribution. Toward that end, for any nonnegative loss function $\mathcal{L}$ and a given test observation $(\boldsymbol{x}_0, y_0)$ from P, define the conditional ψ_1-Orlicz norm of $\mathcal{L}(y_0, \widehat{f}(\boldsymbol{x}_0))$ given $\mathcal{D}_I$ as $\|\mathcal{L}(y_0, \widehat{f}(\boldsymbol{x}_0))\|_{\psi_1|\mathcal{D}_I} = \inf\{C > 0 : \mathbb{E}[\exp(\mathcal{L}(y_0, \widehat{f}(\boldsymbol{x}_0))C^{-1}) \mid \mathcal{D}_I] \le 2\}$. Similarly, for $r \ge 1$, define the conditional L_r-norm as $\|\mathcal{L}(y_0, \widehat{f}(\boldsymbol{x}_0))\|_{L_r|\mathcal{D}_I} = (\mathbb{E}[\mathcal{L}(y_0, \widehat{f}(\boldsymbol{x}_0))^r \mid \mathcal{D}_I])^{1/r}$. See Vershynin (2018, chap. 2) for more details. The following proposition provides the condition for consistency.

Proposition 2 (Consistent risk estimators). Let $\widehat{f}(\cdot; \mathcal{D}_I)$ be any predictor trained on $\mathcal{D}_I \subset \mathcal{D}_n$, and EST = MEAN or EST = MOM with $\eta = n^{-A}$, where $A \in (0, \infty)$ is a fixed constant. Define $\widehat{\sigma}_I = \|(y_0 - \widehat{f}(\boldsymbol{x}_0; \mathcal{D}_I))^2\|_{\psi_1|\mathcal{D}_I}$. If $\widehat{\sigma}_I / \sqrt{|I^c|/\log n} \to 0$ in probability, then $|\widehat{R}(\widehat{f}, \mathcal{D}_{I^c}) - R(\widehat{f}; \mathcal{D}_I)| \to 0$ in probability. The conclusion remains true for EST = MOM even when the conditional ψ_1-Orlicz norm in the definition of $\widehat{\sigma}_I$ is replaced with $\|\cdot\|_{L_2|\mathcal{D}_I}$.

The proof of Proposition 2 can be found in Appendix S2.2. It follows from the results in Patil, Du, and Kuchibhotla (2023, Lemmas 2.9 and 2.10), which are used to demonstrate the strong consistency of sample-split CV in (Patil et al. 2022; Patil, Du, and Kuchibhotla 2023) for subsample tuning. However, our analysis will utilize the finite-sample tail bounds that underlie Proposition 2 to obtain convergence rates.

Recall from (2) that $\widetilde{f}_{1,k}$ is computed from one dataset $\mathcal{D}_{I_1}$ and $\widetilde{f}_{2,k}$ is computed from two datasets $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$ or equivalently that $\widetilde{f}_{2,k}$ is computed on $\mathcal{D}_{I_1 \cup I_2}$. Hence, Proposition 2 can be applied to $\widetilde{f}_{1,k}$ with $I = I_1$ and to $\widetilde{f}_{2,k}$ with $I = I_1 \cup I_2$ to consistently estimate the conditional prediction risks of $\widetilde{f}_{1,k}$ and $\widetilde{f}_{2,k}$. To obtain consistency, we need to ensure that the assumption on $\widehat{\sigma}_I$ becomes reasonable if $|I^c|/\log n \to \infty$ as $n \to \infty$. For subagging predictors $\widetilde{f}_{1,k}$ and $\widetilde{f}_{2,k}$, we have $|I_1^c| = n(1 - k/n)$ and by Lemma S3, $|(I_1 \cup I_2)^c| \approx n(1 - k/n)^2$, asymptotically. On the other hand, for bagging predictors $\widetilde{f}_{1,k}$ and $\widetilde{f}_{2,k}$, we have $|I_1^c| \approx n\exp(-k/n)$ and $|(I_1 \cup I_2)^c| \approx n\exp(-2k/n)$, asymptotically. Hence, collectively, assuming $k \le n(1 - 1/\log n)$ implies that $|I^c| \gtrsim n/\log^2 n$, asymptotically.

Under the assumption $k \le n(1 - 1/\log n)$, the risks of $\widetilde{f}_{1,k}$ and $\widetilde{f}_{2,k}$ computed on $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_1 \cup I_2}$, respectively, can be estimated consistently. Further, if we assume that the limiting conditional risks ($\mathfrak{R}_{1,k}$ and $\mathfrak{R}_{2,k}$) of $\widetilde{f}_{1,k}$ and $\widetilde{f}_{2,k}$ do not depend on the particular subsets I_1, I_2, then $\widehat{R}(\widetilde{f}_{1,k}, \mathcal{D}_{I_1^c})$ and $\widehat{R}(\widetilde{f}_{2,k}, \mathcal{D}_{(I_1 \cup I_2)^c})$ consistently estimate $\mathfrak{R}_{1,k}$ and $\mathfrak{R}_{2,k}$, as $n \to \infty$. Note, however, that because the limiting conditional risks do not depend on specific subsets I_1, I_2, we can reduce the variance in our estimates of $\mathfrak{R}_{1,k}$ and $\mathfrak{R}_{2,k}$ by averaging the estimated risks over several subsets I_ℓ's. This observation suggests the out-of-bag risk estimates for $M = 1, 2$ as

$$\widehat{R}^{\mathrm{ECV}}_{M,k} = \begin{cases} \frac{1}{M_0} \sum_{\ell=1}^{M_0} \widehat{R}(\widetilde{f}_{1,k}(\cdot; \mathcal{D}_n, \{I_\ell\}), \mathcal{D}_{I_\ell^c}), & M = 1, \\ \frac{1}{M_0(M_0-1)} \sum_{\substack{\ell, m \in [M_0] \\ \ell \ne m}} \widehat{R}(\widetilde{f}_{2,k}(\cdot; \mathcal{D}_n, \{I_\ell, I_m\}), \mathcal{D}_{(I_\ell \cup I_m)^c}), & M = 2, \end{cases} \tag{6}$$

where $I_1, \ldots, I_{M_0}$ are iid samples from $\mathcal{I}_k$ and $M_0 \ge 2$ is a prespecified natural number. Increasing M_0 improves estimates but also increases computation.

As hinted above, the risk decomposition (4), along with the component risk estimation (6), suggests a natural estimator for the M-bagged risk:

$$\widehat{R}^{\mathrm{ECV}}_{M,k} = -\left(1 - \frac{2}{M}\right)\widehat{R}^{\mathrm{ECV}}_{1,k} + 2\left(1 - \frac{1}{M}\right)\widehat{R}^{\mathrm{ECV}}_{2,k}, \quad M > 2. \tag{7}$$

We call this "extrapolated" risk estimation according to (4) because the M-bagged risk is *extrapolated* from the 1- and 2-bagged risks. If the prediction risk of $M = 1, 2$ can be consistently estimated, the extrapolated estimates (7) are also pointwise consistent over $M \in \mathbb{N}$ for k fixed.

3.2. Uniform Risk Consistency

The next step is then to tune both M and k. To tune k, we define $\mathcal{K}_n \subset [n]$ to be a grid of subsample sizes. In practice, we would like $\mathcal{K}_n$ to cover the full range of n asymptotically (in the sense that $\mathcal{K}_n/n \approx [0, 1]$), and one simple choice is to set $\mathcal{K}_n = \{0, k_0, 2k_0, \ldots, \lfloor n(1 - (\log n)^{-1})/k_0 \rfloor k_0\}$ where the minimum subsample size is $k_0 = \lfloor n^\nu \rfloor$ for some $\nu \in (0, 1)$. Here we adopt the convention that when $k = 0$, the ensemble predictor reduces to the *null predictor* that always outputs zero. To facilitate our theoretical results for general predictors, we make the following two assumptions. The results are stated asymptotically as n tends to infinity, where we view both k and p as sequences $\{k_n\}$ and $\{p_n\}$ indexed by n, and assume k diverges with the sample size n (except when $k \equiv 0$), but the feature dimension p may or may not diverge.

Assumption 1 (Variance proxy). For $k \in \mathcal{K}_n$, assume for all $\{I_1, I_2\} \overset{\mathrm{SRS}}{\sim} \mathcal{I}_k$, as $n \to \infty$,

$$\frac{\log n}{\sqrt{n}}\widehat{\sigma}_{I_1} \to 0, \quad \text{and} \quad \frac{(\log n)^{3/2}}{\sqrt{n}}\widehat{\sigma}_{I_1 \cup I_2} \to 0,$$

in probability, where the variance proxies $\widehat{\sigma}_{I_1}$ and $\widehat{\sigma}_{I_1 \cup I_2}$ are defined in Proposition 2.

Assumption 2 (Convergence of asymptotic risks). For $k \in \mathcal{K}_n$, assume for all $\{I_1, I_2\} \overset{\mathrm{SRS}}{\sim} \mathcal{I}_k$ and for $M = 1, 2$, there exist

constants $\epsilon \in (0,1)$, $C_0 > 0$, $\eta_0 \geq 1$, $\gamma_{M,n} = o(n^{-\epsilon})$, and $\mathfrak{R}_{M,k} \geq 0$, such that the following holds:

$$\limsup_{n\to\infty} \sup_{\eta\geq\eta_0} \eta^{1/\epsilon} \mathbb{P}(\gamma_{M,n}^{-1}|R_{M,k}(\widetilde{f}_{M,k};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) - \mathfrak{R}_{M,k}| \leq \eta) \leq C_0. \quad (8)$$

Assumption 1 is used to show consistent risk estimation with $M = 1, 2$ in Appendix S2.2. Assumption 2 formalizes the assumption that the limiting values of the conditional risks $R_{M,k}(\widetilde{f}_{M,k};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M)$ do not depend on $\{I_\ell\}_{\ell=1}^M$. This assumption also requires certain rate and tail assumptions, which are used to provide the rate of consistency of our ECV procedure. In Assumption 2, $\gamma_{M,n}$ for $M = 1, 2$ represent the lower bounds of the rates of convergence over $k \in \mathcal{K}_n$, which are typically on a scale of $n^{-\alpha}$ for some constant $\alpha > 0$. Condition (8) is also known as the weak moment norm condition (Rio 2017; Guo and Peterson 2019), which ensures that the expected differences between the risks and the limits also converge to zero in certain rates. Under classical linear models with fixed-X design and Gaussian noises, the risk of linear predictor concentrates around its mean and satisfies (8); see, for example, Bellec (2018, Lemma 3.1). Another sufficient condition for (8) is the strong moment condition that $\mathbb{E}(\gamma_{M,n}^{-1/\epsilon}|R_{M,k}(\widetilde{f}_{M,k};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) - \mathfrak{R}_{M,k}|^{1/\epsilon})$ is bounded.

From now on, we shall write $R_{M,k} = R(\widetilde{f}_{M,k};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M)$ to indicate the dependency only on M and k and to simplify the notations. In Appendix S4.3, we present an example of ridge regression where the convergence is under the proportional asymptotics (i.e., both the *data aspect ratio* p/n and the *subsample aspect ratio* p/k converge to fixed constants), and Assumptions 1–2 are satisfied. The following theorem guarantees uniform consistency over both M and k.

Theorem 1 (Uniform consistency of risk extrapolation). Suppose Assumptions 1 and 2 hold for all $k \in \mathcal{K}_n$, then ECV estimates defined in (7) satisfy that

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

where $\zeta_n = \widehat{\sigma}_n \log n/n^{1/2} + n^{\epsilon}(\gamma_{1,n} + \gamma_{2,n})$, and $\widehat{\sigma}_n = \max_{m,\ell\in[M_0],k\in\mathcal{K}_n} \widehat{\sigma}_{I_{k,\ell}\cup I_{k,m}}$.

The result in Theorem 1 is of paramount importance to establish the convergence rate of the CV-tuned estimator returned by our algorithm. In words, the theorem says that the extrapolated error depends on three factors: the cross-validation error and the rates for $M = 1, 2$. While the convergence rates of asymptotic risks of $M = 1, 2$ usually depend on the chosen predictor, the non-asymptotic analysis in Theorem 1 allows us to derive convergence rates even for general predictors. This is particularly advantageous when compared to the consistency results established in Patil, Du, and Kuchibhotla (2023) for subagged ridge predictors.

The proof Theorem 1 is rather involved and can be found in Appendix S3. For the convenience of the readers, we provide a schematic of the whole proof in Figure S1. We will now explain the key ideas involved in the proof. The proof strategy relies on deriving concentration results of varied random quantities to their limits in a specific order. First, we establish the uniform consistency of the risk estimates $\widehat{R}_{M,k}^{\mathrm{ECV}}$ over $k \in \mathcal{K}_n$ to the risks $R_{M,k}$ for $M = 1, 2$ (see Proposition S3). Building upon this result and the risk decomposition presented in Proposition 1, we then derive the uniform consistency of the risk estimates $\widehat{R}_{M,k}^{\mathrm{ECV}}$ over $(M,k) \in \mathbb{N}\times\mathcal{K}_n$ to the deterministic limits $\mathfrak{R}_{M,k}$ (Proposition S4). On the other hand, to establish the concentration for subsample conditional risks over $M \in \mathbb{N}$ and $k \in \mathcal{K}_n$, we first establish the concentration for the expected conditional risk $\mathbb{E}[R_{M,k} \mid \mathcal{D}_n]$ over $k \in \mathcal{K}_n$ (Lemma S1). We then apply the reverse martingale concentration bound (Lemma S2).

4. Main Proposal: Extrapolated Cross-Validation

Based on the previous discussion in Section 3, we present the proposed cross-validation algorithm for tuning the ensemble parameters without sample splitting in Algorithm 1. The procedure requires a dataset $\mathcal{D}_n$ of n observations, a base prediction procedure $\widehat{f}$, a natural number M_0 for risk estimation, and some other parameters. It first constructs the grid of subsample sizes $\mathcal{K}_n$ and fits only M_0 base predictors accordingly. Then, the prediction risk for M-bagged predictors can be estimated based on the OOB observations through (6) and (7). Observe that the optimal risk of (7) for any k is obtained at $\widehat{R}_{\infty,k}^{\mathrm{ECV}} = 2\widehat{R}_{2,k}^{\mathrm{ECV}} - \widehat{R}_{1,k}^{\mathrm{ECV}}$ when $M = \infty$. Thus, to tune k, it suffices to perform a grid search to minimize $\widehat{R}_{\infty,k}^{\mathrm{ECV}}$ over $k \in \mathcal{K}_n$, because the optimal ensemble size happens to be infinity from the previous results (Lopes 2019; Patil, Du, and Kuchibhotla 2023). However, calculating it is prohibitive in practice. Thus, we pick the smallest $\widehat{M}$ such that $\widehat{R}_{\widehat{M},\widehat{k}}^{\mathrm{ECV}}$ is close to $\widehat{R}_{\infty,\widehat{k}}^{\mathrm{ECV}}$ within δ error, where δ is the suboptimality parameter. Finally, Algorithm 1 returns a $\widehat{M}$-bagged predictor using subsample size $\widehat{k}$. Note that Algorithm 1 naturally applies to other randomized ensemble methods, such as random forests, when fixing other hyperparameters.

As a byproduct, Algorithm 1 also gives the ECV risk estimates $\widehat{R}_{M,k}^{\mathrm{ECV}}$ for all $M \in \mathbb{N}$ and $k \in \mathcal{K}_n$. This risk profile in (M,k) is helpful for users to investigate whether the given base predictor $\widehat{f}$ well fits the dataset $\mathcal{D}_n$ or not. For instance, one can tune the ensemble predictors under a computational budget on the maximum ensemble size $M_{\max}$. This gives rise to practical considerations presented later in Section 4.2.

4.1. Theoretical Guarantees

By combining the ingredients in Section 3, our main theorem states that Algorithm 1 yields an ensemble predictor whose risk is at most δ away from the best ensemble predictor. Further, it provides an estimator of the risk of the selected ensemble predictor.

Theorem 2 (Optimality of OOB estimate and ECV-tuned risk). Under the assumed conditions in Theorem 1, for any $\delta > 0$, the OOB estimate and the ECV-tuned risk output by Algorithm 1 satisfy the following properties respectively:

$$\left|\widehat{R}_{\widehat{M},\widehat{k}}^{\mathrm{ECV}} - R_{\widehat{M},\widehat{k}}\right| = \mathcal{O}_p(\zeta_n),$$
$$\left|R_{\widehat{M},\widehat{k}} - \inf_{M\in\mathbb{N},k\in\mathcal{K}_n} R_{M,k}\right| = \delta + \mathcal{O}_p(\zeta_n), \quad (9)$$

where ζ_n is the quantity defined in Theorem 1.

Algorithm 1 Tuning of ensemble and subsample sizes without sample splitting

Input: a dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R} : 1 \le i \le n\}$, a base prediction procedure $\widehat{f}$, a real number $\nu \in (0,1)$ (subsample size unit parameter), a ensemble size $M_0 \ge 2$ for risk estimation, a centering procedure EST $\in$ {MEAN, MOM}, a real number A used to compute η when EST = MOM, and optimality tolerance parameter δ.

1: Construct a grid $\mathcal{K}_n = \{0, k_0, 2k_0, \ldots, \lfloor n(1-1/\log n)/k_0 \rfloor k_0\}$ where $k_0 = \lfloor n^\nu \rfloor$.

2: Build ensembles $\widetilde{f}_{M_0,k}(\cdot) = \widetilde{f}_{M_0}(\cdot; \{\mathcal{D}_{I_{k,\ell}}\}_{\ell=1}^{M_0})$ with M_0 base predictors, where $I_{k,1}, \ldots, I_{k,M_0} \overset{\text{SRS}}{\sim} \mathcal{I}_k$ for each $k \in \mathcal{K}_n$.

3: Estimate the conditional prediction risk on OOB observations of $\widetilde{f}_{M_0,k}$ with $\widehat{R}^{\text{ECV}}_{M,k}$ defined in (6) for $k \in \mathcal{K}_n$ and $M = 1, 2$.

4: Extrapolate the risk estimations $\widehat{R}^{\text{ECV}}_{M,k}$ using (7).

5: Select a subsample size $\widehat{k} \in \mathcal{K}_n$ that minimizes the extrapolated estimates using

$$\widehat{k} \in \operatorname*{argmin}_{k \in \mathcal{K}_n} 2\widehat{R}^{\text{ECV}}_{2,k} - \widehat{R}^{\text{ECV}}_{1,k}.$$

6: Select an ensemble size $\widehat{M} \in \mathbb{N}$ for the δ-optimal risk with a plug-in estimator:

$$\widehat{M} = \left\lceil \frac{2}{\max\{\delta, n^{-1/2}\}} (\widehat{R}^{\text{ECV}}_{1,\widehat{k}} - \widehat{R}^{\text{ECV}}_{2,\widehat{k}}) \right\rceil.$$

7: If $\widehat{M} > M_0$, fit a $\widehat{M}$-bagged predictor $\widetilde{f}_{\widehat{M},\widehat{k}} = \widetilde{f}_{\widehat{M},\widehat{k}}(\cdot; \{\mathcal{D}_{I_{\widehat{k},\ell}}\}_{\ell=1}^{\widehat{M}})$.

Output: Return the ECV-tuned predictor $\widetilde{f}_{\widehat{M},\widehat{k}}$, and the risk estimators $\widehat{R}^{\text{ECV}}_{M,k}$ for all M, k.

Theorem 2 implies that the OOB estimate is close to the true risk because of the uniform consistency from Theorem 1. Furthermore, the ECV-tuned ensemble parameters $\widehat{M}$ and $\widehat{k}$ produce a bagged predictor with a risk δ-close to the optimal predictor. The optimality is model-agnostic because it does not directly depend on the feature and response models. On the other hand, in real-world applications, the infinite ensemble may not be of interest because of the computational cost. Because of the uniform consistency established in Theorem 1, one can naturally extend Theorem 2 to tuning with restriction on maximum ensemble sizes. In Appendix S4.3, we also provide a concrete example of the application of Theorem 2 to ridge predictors, which verifies Assumptions 1 and 2 under mild moment assumptions.

Remark 1 (From additive to multiplicative optimality). Theorem 2 provides a guarantee of additive optimality for tuned predictor returned by Algorithm 1, while it is also useful to consider the multiplicative optimality. Toward that end, with the choice of $\widehat{k}$ in Step 5 of Algorithm 1 and change the choice of $\widehat{M}$ in Step 6 to

$$\widehat{M} = \left\lceil \frac{2}{\max\{\delta, n^{-1/2}\}} \frac{\widehat{R}^{\text{ECV}}_{1,\widehat{k}} - \widehat{R}^{\text{ECV}}_{2,\widehat{k}}}{2\widehat{R}^{\text{ECV}}_{2,\widehat{k}} - \widehat{R}^{\text{ECV}}_{1,\widehat{k}}} \right\rceil,$$

then, under the assumption that the irreducible risk $\int \big(y - \mathbb{E}(y \mid \boldsymbol{x})\big)^2 \, \mathrm{d}P(\boldsymbol{x}, y)$ is strictly positive, Proposition S5 guarantees

$$R_{\widehat{M},\widehat{k}} = (1+\delta) \inf_{M \in \mathbb{N}, k \in \mathcal{K}_n} R_{M,k}(1 + \mathcal{O}_p(\zeta_n)). \tag{10}$$

Compared to (9), the optimality upper bound on the right-hand side of (10) depends on the scale of the optimal prediction risk.

4.2. Computational Considerations

Algorithm 1 estimates the ensemble parameters $\widehat{k}$ and $\widehat{M}$ to derive a δ-optimal bagged predictor. Here, we compare the computational complexity of Algorithm 1 with other common CV methods. For each $k \in \mathcal{K}_n$, suppose the computational complexity of fitting one base predictor on k subsampled observations and obtain their predicted values on $n-k$ OOB observations is $\mathcal{O}(C_n)$ (ignoring k). Then the computational complexity of estimating $\widehat{M}$ and $\widehat{k}$ for all three validation methods are given as below.

- ECV: For each $k \in \mathcal{K}_n$, we need to fit M_0 base predictors. Then we estimate $R_{1,k}$ and $R_{2,k}$ in $\mathcal{O}(M_0(n-k))$ and $\mathcal{O}(M_0^2(n-2k+i_0))$, respectively, where $i_0 = k^2/n$ is the intersect observations between two indices of a simple random sample. The computational complexity of risk extrapolation is negligible compared to the above time consumption. All in all, it takes $\mathcal{O}(C_n(|\mathcal{K}_n|M_0 + M_{\max}))$ to obtain tuned bagging parameter by ECV.
- sample-split CV: Suppose the ratio of training data is $\alpha \in (0,1)$. Similar to ECV, each base predictor is fitted and evaluated on $\lceil n\alpha \rceil$ observations and we need to fit $M_{\max}$ base predictors. We then compute the moving average of the predicted values for M varying from 1 to $M_{\max}$, which gives the predicted values of the M-bagged predictors, which takes $\mathcal{O}(M_{\max})$ operations. We note that one can alternatively fit one bagged predictor for each k and each M; however, this will cause much more time consumption compared to the simple matrix computation operations we used above. All in all, it takes $\mathcal{O}(|\mathcal{K}_n|M_{\max}(C_{n\alpha} + n))$ to obtain the tuned parameter.
- K-fold CV: We follow the same strategy for fitting base predictors so that K-fold CV has roughly K times of complexity as sample-split CV. Specifically, it takes $\mathcal{O}(K|\mathcal{K}_n|M_{\max}(C_{n/K} + n))$ to obtain the tuned parameter.

In general, we expect C_n to grow much faster than n, because fitting one base predictor may involve matrix multiplication operation, which takes $\mathcal{O}(n^2)$. Therefore, the computational complexity of the three methods has the ordering: ECV $\le$ sample-split CV $\le K$-fold CV, provided that M_0 is much smaller than $M_{\max}$.

Besides the computational efficiency gained by ECV, we also discuss some considerations when the proposed method is used in practice.

1. Maximum ensemble size: Algorithm 1 determines $\widehat{k}$ and $\widehat{M}$ by minimizing the estimated risk (7) with the infinite ensemble. However, it may still be computationally infeasible if $\widehat{M}$ is too large. In such cases, based on the extrapolated risk estimation in Algorithm 1, we can also derive the δ-optimal bagged predictor whose ensemble size is no more than a pre-specified number $M_{\max}$, which we call the restricted oracle. That is, we choose subsample and ensemble size to restrict the computational cost: $\widehat{k} \in \operatorname{argmin}_{k \in \mathcal{K}_n} \widehat{R}^{\text{ECV}}_{M_{\max},k}$ and $\widehat{M} = \left\lceil 2(\delta + \widehat{R}^{\text{ECV}}_{M_{\max},\widehat{k}} - \widehat{R}^{\text{ECV}}_{\infty,\widehat{k}})^{-1}(\widehat{R}^{\text{ECV}}_{1,\widehat{k}} - \widehat{R}^{\text{ECV}}_{2,\widehat{k}}) \right\rceil$. On the

other hand, it also controls the suboptimality to the oracle:

$$\underbrace{R_{\widehat{M},\widehat{k}} - \min_{k\in\mathcal{K}_n} R_{\infty,k}}_{\text{suboptimality to the oracle}} = \underbrace{R_{\widehat{M},\widehat{k}} - \min_{k\in\mathcal{K}_n} R_{M_{\max},k}}_{\text{suboptimality to the restricted oracle}} + \underbrace{\min_{k\in\mathcal{K}_n} R_{M_{\max},k} - \min_{k\in\mathcal{K}_n} R_{\infty,k}}_{\text{unavoidable budget error}}.$$

When $\delta = 0$, the suboptimality to the restricted oracle vanishes, and the tuned ensemble simply tracks the optimal $M_{\max}$-ensemble.

2. To bag or not to bag: The benefit of ensemble learning may be slight in some cases. For instance, when the number of samples is much larger than the feature dimensions and the signal-noise ratio is large, ensemble learning can only provide minor improvements over the non-ensemble predictor. Suppose that $\widehat{R}_0^{\mathrm{ECV}}$ is the estimated risk of the null predictor, $\min_{k\in\mathcal{K}_n} \widehat{R}_{1,k}^{\mathrm{ECV}}$ is the optimal estimated risk due to subsampling when $M = 1$, and $\min_{k\in\mathcal{K}_n} \widehat{R}_{M_{\max},k}^{\mathrm{ECV}}$ is the optimal ECV estimate with maximum ensemble size $M_{\max}$. Let $\zeta > 0$ be a user-specified improvement factor that encodes the desired excess risk improvement in a multiplicative sense. Then we can decide to bag if either the null risk is smaller in the sense that $\widehat{R}_0^{\mathrm{ECV}} < \min_{k\in\mathcal{K}_n} \widehat{R}_{1,k}^{\mathrm{ECV}}$, or the improvement due to ensemble exceeds ζ times the improvement due to subsampling:

$$\underbrace{\min_{k\in\mathcal{K}_n} \widehat{R}_{1,k}^{\mathrm{ECV}} - \min_{k\in\mathcal{K}_n} \widehat{R}_{M_{\max},k}^{\mathrm{ECV}}}_{\text{improvement due to ensemble}} > \underbrace{\zeta}_{\text{improvement factor}} \times \underbrace{\widehat{R}_0^{\mathrm{ECV}} - \min_{k\in\mathcal{K}_n} \widehat{R}_{1,k}^{\mathrm{ECV}}}_{\text{improvement due to subsampling}}. \quad (11)$$

3. Absolute versus normalized tolerances: The choice of the tolerance threshold δ is for controlling the absolute suboptimality, but the scale of the prediction risk may be different for different predictors and datasets. One can normalize the estimated risk by the null predictor's estimated risk to make the tolerance threshold comparable across different predictors and datasets, or tune based on the multiplicative guarantee (10).

5. Numerical Illustrations

In this section, we evaluate Algorithm 1 on synthetic data. In Section 5.1, we inspect whether the extrapolated risk estimates $\widehat{R}_{M,k}^{\mathrm{ECV}}$ serve as reasonable proxies for the actual out-of-sample prediction errors for various base predictors on uncorrelated features. In Section 5.2, we further evaluate the risk minimization performance with tuned ensemble parameters $(\widehat{M},\widehat{k})$. Finally, in Section 5.3, we consider tuning M for random forests on correlated features.

5.1. Validating Extrapolated Risk Estimates

In this simulation, we examine whether our risk extrapolation strategy provides reasonable risk estimates for specific values of ensemble size M and subsample size k based on only the risk estimates for $M = 1$ and $M = 2$. We evaluate six base predictors (in Figure 2) on data models:

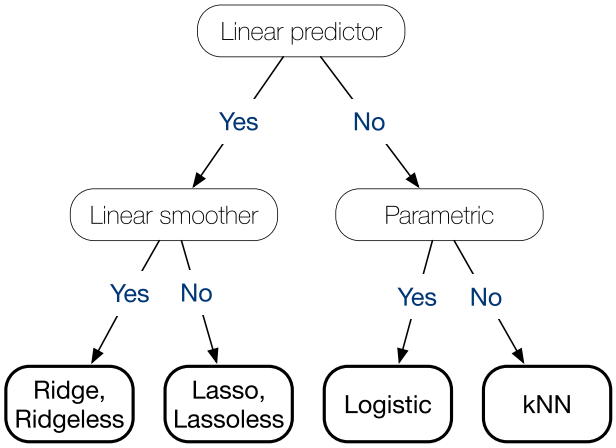

Figure 2. Predictors for regression tasks evaluated in Sections 5.1 and 5.2.

(M1) Linear: A linear model: $y = \boldsymbol{x}^\top\boldsymbol{\beta} + \epsilon$,
(M2) Quad: A polynomial regression model: $y = \boldsymbol{x}^\top\boldsymbol{\beta} + ((\boldsymbol{x}^\top\boldsymbol{\beta})^2 - \mathrm{tr}(\boldsymbol{\Sigma}_{\rho_{\mathrm{ar1}}})/p) + \epsilon$,
(M3) Tanh: A single-index regression model: $y = \tanh(\boldsymbol{x}^\top\boldsymbol{\beta}) + \epsilon$,

where the features and the coefficients are generated from $\boldsymbol{x} \sim \mathcal{N}_p(0, \boldsymbol{\Sigma}_{\rho_{\mathrm{ar1}}})$ and $\boldsymbol{\beta}$ is the average of $\boldsymbol{\Sigma}_{\rho_{\mathrm{ar1}}}$'s eigenvectors associated with the top-5 eigenvalues. Here $\boldsymbol{\Sigma}_{\rho_{\mathrm{ar1}}} = (\rho_{\mathrm{ar1}}^{|i-j|})_{1\le i,j\le p}$ is the covariance matrix of an auto-regressive process of order 1 (AR(1)) with $\rho_{\mathrm{ar1}} = 0.5$. The additive noise is sampled from $\epsilon \sim \mathcal{N}(0,\sigma^2)$ with $\sigma = 0.5$. In this setup, 5.1 has a signal-noise ratio of 2.4. Here, the data aspect ratio $\phi = p/n$ varies from 0.1 (low-dimensional regime) to 10 (high-dimensional regime), and the subsample aspect ratio $\phi_s = p/k$ varies from 0.1 to 10 and from 10 to 100, respectively. The *null risk*, the risk of the null predictor that always outputs zero, can also be estimated at each ϕ. For ridgeless and lassoless predictors, we use rule (11) in Section 4.2 to exclude k with exploding risks more than 5 times the estimated null risk.

The ECV estimates and the corresponding prediction errors for different ensemble sizes M and subsample aspect ratios ϕ_s are then summarized in Appendix S6 for bagged and subagged predictors. As k decreases, the ensemble with subsample size k behaves more like the null predictor. As a result, the risk curves approach a particular value as the subsample aspect ratio p/k increases. From Figure S2 to S7, we observe a good match between the ECV estimates and the out-of-sample prediction errors. This suggests that Theorem 2 potentially applies to various types of predictors. Comparing the results of bagging and subagging, the risk estimates are very similar, especially in the overparameterized regime when $p > n$. Thus, we will only present the results using bagging for illustration purposes.

5.2. Tuning Ensemble and Subsample Sizes

Next, we examine the performance of ECV on predictive risk minimization. More specifically, we apply Algorithm 1 using the rule (11) with $\zeta = 5$ to tune an ensemble that is close to

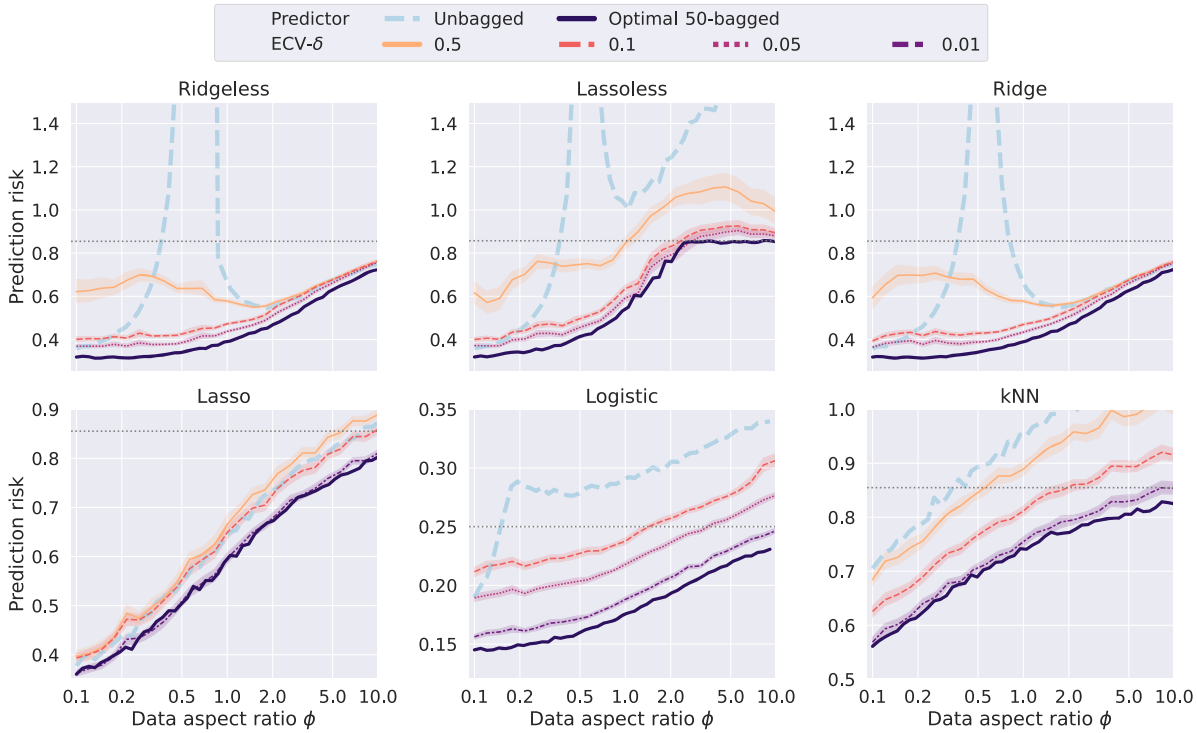

Figure 3. Prediction risk for different bagged predictors by ECV, under model 5.1 with $\sigma = \rho_{\text{ar1}} = 0.5$, $M_0 = 10$, and $M_{\max} = 50$, for varying ϕ and tolerance threshold δ. An ensemble is fitted when (11) is satisfied with $\zeta = 5$. The null risks and the risks for the non-ensemble predictors are marked as gray dotted lines and blue thick dashed lines, respectively. The points denote finite-sample risks averaged over 100 dataset repetitions, and the shaded regions denote the values within one standard deviation, with $n = 1000$ and $p = \lfloor n\phi \rfloor$.

the optimal $M_{\max}$-ensemble up to additive error δ, where the maximum ensemble size $M_{\max}$ is 50 and optimality threshold δ ranges from 0.01 to 1. With the same predictors and data used in Section 5.1, their out-of-sample mean squared errors are evaluated on the same test set.

As summarized in Figure 3, the thick dashed lines represent the non-ensemble prediction risk, and the thick solid lines represent the prediction risk of optimal 50-bagged predictors using a finer grid. Note that the former may be non-monotonic in the data aspect ratio ϕ, but the latter is increasing in ϕ. The finite-sample prediction errors of the ECV-tuned predictor are shown as thin lines. As we can see, when δ decreases, the prediction errors of ECV get closer to those of the optimal 50-bagged predictor. The slight discrepancy between the ECV-tuned risks with the least δ and the oracle risks comes from the fact that a coarser grid $\mathcal{K}_n$ is used for ECV tuning. Overall, the results suggest that the ECV-tuned ensemble parameters $(\widehat{M}, \widehat{k})$ give risks close to the oracle choices for various predictors within the desired optimality threshold δ in finite samples. Lastly, though ECV is proposed for regression tasks, the numerical results in Appendix S6.3 support its superiority over K-fold CV in imbalanced binary classification scenarios.

5.3. Tuning Ensemble Sizes of Random Forests

When the data apsect ratio p/n is too small, tuning both the subsample size k and the ensemble size M may be unnecessary. In such cases, tuning the ensemble size M (in the sense that how large M is sufficient to have good performance) is a more substantial and practical consideration. In this experiment, we apply Algorithm 1 to tune only the ensemble size of random forests.

Since the most crucial advantage of the random forest model is its flexibility to incorporate highly correlated variables while avoiding multi-collinearity issues, we consider the nonlinear model 5.1 with a non-isotropic AR(1) covariance matrix. For a given dataset, we examine two strategies to estimate the conditional prediction risks. The first uses the OOB observations according to Algorithm 1, while the other uses a hold-out subset to estimate the risks. Similar to Lopes (2019), $\lfloor n/6 \rfloor$ observations are randomly selected as the evaluation set for the hold-out estimates. As suggested by Hastie, Tibshirani, and Friedman (2009), each decision tree uses $\lfloor p/3 \rfloor$ randomly selected features with a minimum node size of 5 as the default without pruning. To build each tree, we fix the subsample size $k = n(1 - 1/\log n)$ observations for subagging. The results are shown in Figure 4, where the standard deviation of the estimates are also visualized as error bars. In the underparameterized regime when $n > p$, we observe that ECV and hold-out estimates have similar performance. Both of them are close to the out-of-sample errors in this case. However, in the overparameterized regime when $n < p$, the hold-out estimates suffer from biases due to sample splitting. On the contrary, the ECV estimates are still accurate and have smaller variability compared to the hold-out estimates. In the right panel of Figure 4, we see that ECV estimates provide a valid extrapolation path from $M = 20$ to $M = 500$ in the high-dimensional scenarios.

Finally, we conduct a sensitivity analysis of the hyperparameter M_0, the correlative strength ρ_{ar1} of the feature covariance matrix, and the covariance structures in Appendices S6.4.1 to

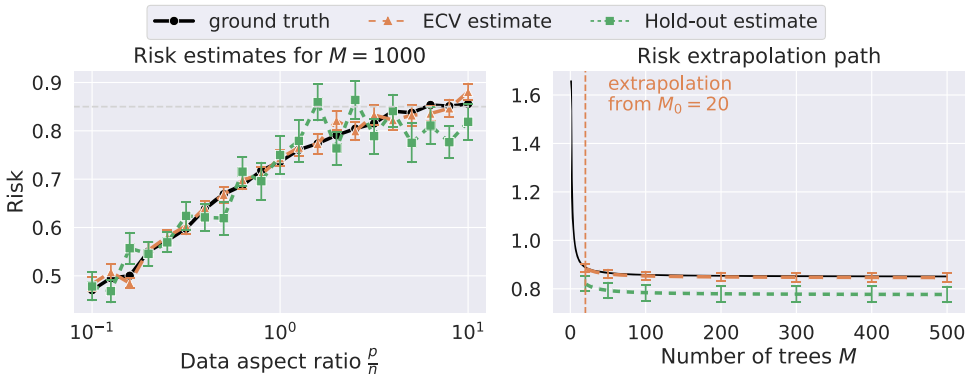

Figure 4. Risk extrapolation for random forest based on the first $M_0 = 20$ trees, under model 5.1 with $\sigma = \rho_{\text{ar1}} = 0.5$. The left panel shows the risk estimates for random forests with $M = 500$ trees and varying data aspect ratios, and the gray dash line $y = 0.85$ denotes the risk of the null predictor; the right panel shows the full risk extrapolation path in M when $p/n = 8$. The error bar denotes one standard deviation across 100 simulations, and the ground truth is estimated from 2000 test observations, with $n = 500$.

S6.4.3. The results suggest that ECV is relatively robust under various scenarios and various choices of hyperparameters. In Appendix S6.4.4, we illustrate the utility of ECV for tuning the number of features drawn when splitting each node of random forests.

6. Applications to Single-Cell Multiomics

In genomics, cell surface proteins act as primary targets for therapeutic intervention and universal indicators of particular cellular processes. More importantly, immunophenotyping of cell surface proteins has become an essential tool in hematopoiesis, immunology, and cancer research over the past 30 years (Hao et al. 2021). However, most single-cell investigations only quantify the transcriptome without cell-matched measures of related surface proteins due to technical limitations and financial constraints (Zhou et al. 2020; Du, Cai, and Roeder 2022). The specific cell types and differentially abundant surface proteins are determined after thoroughly analyzing the transcriptome. This has led researchers to investigate how to reliably predict protein abundances in individual cells using their gene expressions. Specifically, the effectiveness of ensemble methods has been illustrated by (Heckmann et al. 2018; Li et al. 2019; Xu et al. 2021) on the protein prediction problem. Yet, in practice, because of the lack of theoretical results and pragmatic guidelines, the ensemble and subsample sizes are generally determined by ad hoc criteria.

In this section, we apply the proposed method to real datasets in single-cell multi-omics (Hao et al. 2021). See Appendix S6.5 for details on the datasets and the preprocessing steps. Based on these real-world datasets, we compare three different cross-validation methods for tuning both the ensemble size and the subsample size of random forests: (a) K-fold CV: the K-fold CV ($K = 5$); (b) sample-split CV: sample-split or holdout CV (the ratio of training to validation observations is 5:1); and (c) ECV: the proposed extrapolated CV. The grid of subsample sizes $\mathcal{K}_n$ is generated according to Algorithm 1. To ensure all the CV methods are comparable and fairly evaluated, we evaluate the three CV methods on the same grid $[M_{\max}] \times \mathcal{K}_n$ for $M_{\max} = 50$. Decision trees are used as the base predictors to predict the abundance of each protein based on the gene expressions of subsampled cells. After the tuning parameters $(\widehat{M}, \widehat{k})$ are obtained, we refit the ensemble on the entire training set and evaluate it on the test set. For each method, the $M_{\max}$ base predictors are fitted once so that the training costs are almost the same for all three. The computational complexity of the three methods is discussed in Section 4.2. Because different proteins may have different variances, we measure the overall protein prediction accuracy by the normalized mean squared error (NMSE), which is the ratio of the mean squared error to the empirical variance on the test set. Our target is to select a δ-optimal random forest so that its NMSE is no more than $\delta = 0.05$ away from the best random forest with 50 trees.

To illustrate our proposed method, we visualize the prediction risk estimate and out-of-sample error for surface protein CD103 in Figure S16 and S17. Because the response is centered, the empirical variance of the response serves as an estimate of the null risk, that is, the risk of the null predictor that always outputs zeros. A value of NMSE less than one indicates that the predictor performs better than the null risk. From Figure S16, we see that the ECV extrapolated estimates in the left panel are largely consistent with the actual prediction errors in the right panel. The out-of-sample error can still be considerable when $M = 10$ as shown in Figure S16. As the ensemble size M increases, both become more stable for various subsample sizes. Further, we observe that the tuned ensemble and subsample sizes are close to the optimal ones on the test dataset in finite samples. The tune subsample size k is much smaller than the total sample size n. This indicates that our proposed method tracks the out-of-sample optimal parameter well.

We compare the performance of different CV methods on various cell types. As shown in Figure 5(a), K-fold CV is very sensitive to the choice of the number of fold K. sample-split CV and K-fold CV with $K = 3$ have the worst out-of-sample performance among all methods. In the low-dimensional dataset of the Mono cell type, all methods have similar performance. However, ECV significantly improves upon sample-split CV with insufficient sample sizes, as in the DC, B, and NK cell types. In all cases, the out-of-sample NMSEs of ECV are comparable to K-fold CV with $K = 5$. Yet, ECV is more flexible to tuning large ensemble sizes, such as $M_{\max} = 100$ and 250. This suggests that the extrapolated risk estimates of ECV based on only $M_0 = 20$ trees are accurate for tuning ensemble parameters on these datasets.

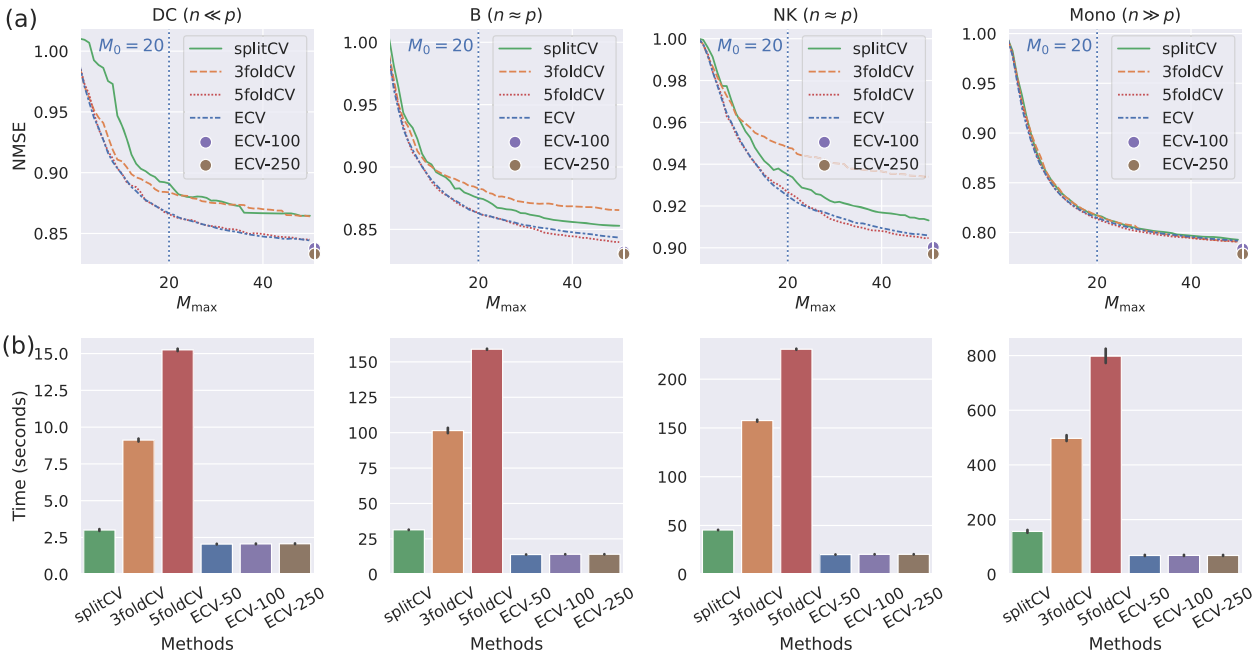

Figure 5. Performance of CV methods on predicting the protein abundances in different cell types. (a) The average NMSEs of the cross-validated predictors for different methods. ECV uses $M_0 = 20$ trees to extrapolate risk estimates, and the points correspond to $M_{\max} \in \{100, 250\}$. (b) The average CV time consumption in seconds.

Regarding time consumption, recall that the base predictors are only fitted once (Section 4.2) so that the comparison is fair for all methods. From Figure 5(b), we see that tuning by ECV is significantly faster than K-fold CV because we are extrapolating the risks from 20 trees rather than estimating them for every $M > 20$. Though ECV has similar time complexity as sample-split CV when the sample sizes are small, it uses less than 50% of the time used by sample-split CV as the sample size increases. Thus, ECV achieves better computational efficiency than the alternative CV methods in a variety of settings.

7. Discussion

This article addresses the challenge of tuning the ensemble and subsample sizes for randomized ensemble learning. While previous work has extensively focused on the statistical properties of bagging and random forests, ensemble tuning remains an area that has received comparatively less attention. To bridge this gap, this article introduces a method that efficiently provides δ-optimal ensemble parameters (for the tuned risk) without requiring an exhaustive search over all feasible parameter combinations, a common limitation of conventional cross-validation methods.

Our method hinges on the extrapolation of risk estimates, which is afforded due to a special decomposition of the squared risk and the consistent risk estimators of each component. Our method is sample-efficient and can be naturally extended to tuning hyperparameters other than subsample sizes, as shown in Appendix S6.4.4. Furthermore, our algorithm guarantees the δ-optimality of the tuned ensemble risk in relation to the oracle ensemble risk.

To demonstrate the practical utility of ECV, we apply it to the task of predicting proteins in single-cell data. In scenarios with small sample sizes, ECV achieves smaller out-of-sample errors without sample splitting compared to traditional sample-split CV. On the other hand, it drastically reduces the time complexity compared to K-fold CV and maintains comparable accuracy without repeatedly fitting numerous predictors. In summary, ECV exhibits both statistical and computational efficiency in protein prediction tasks.

We next point out some limitations of this work and propose possible future directions. Future directions for this work include extending the theoretical framework to other types of loss functions, such as smooth loss functions, by applying the "sandwich" sub-optimality gap approach (Patil, Du, and Kuchibhotla 2023, Proposition 3.6). Additionally, exploring variance estimation and extrapolation for the proposed risk estimators could provide a more comprehensive understanding of their uncertainty. Integrating the concept of performing CV with confidence (Lei 2020) could address the important issue of quantifying confidence in a tuned model for model selection.

Supplementary Materials

Supplementary notes include the proof for all the theorems and extra experiment results.

Acknowledgments

This work used the Bridges-2 system at the Pittsburgh Supercomputing Center (PSC) through allocations BIO220140 and MTH230020 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This project was partially funded by the National Institute of Mental Health (NIMH) grant R01MH123184.

ORCID

Jin-Hong Du https://orcid.org/0000-0001-9683-4146
Kathryn Roeder https://orcid.org/0000-0002-8869-6254

References

Bellec, P. C. (2018), "Optimal Bounds for Aggregation of Affine Estimators," *The Annals of Statistics*, 46, 30–59. [1066]

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997), "Resampling Fewer than *n* Observations: Gains, Losses, and Remedies for Losses," *Statistica Sinica*, 7, 1–31. [1061]

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140. [1061,1062,1063]

——— (2001), "Random Forests," *Machine Learning*, 45, 5–32. [1064]

Bühlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927–961. [1062,1063]

Chen, L., Min, Y., Belkin, M., and Karbasi, A. (2020), "Multiple Descent: Design Your Own Generalization Curve," arXiv preprint arXiv:2008.01036. [1061]

Du, J.-H., Cai, Z., and Roeder, K. (2022), "Robust Probabilistic Modeling for Single-Cell Multimodal Mosaic Integration and Imputation via Scvaeit," *Proceedings of the National Academy of Sciences*, 119, e2214414119. [1070]

Guo, X., and Peterson, J. (2019), "Berry–Esseen Estimates for Regenerative Processes Under Weak Moment Assumptions," *Stochastic Processes and their Applications*, 129, 1379–1412. [1066]

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. (2021), "Integrated Analysis of Multimodal Single-Cell Data," *Cell*, 184, 3573–3587. [1062,1070]

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022), "Surprises in High-Dimensional Ridgeless Least Squares Interpolation," *The Annals of Statistics*, 50, 949–986. [1061]

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), Springer. [1061,1069]

Heckmann, D., Lloyd, C. J., Mih, N., Ha, Y., Zielinski, D. C., Haiman, Z. B., Desouki, A. A., Lercher, M. J., and Palsson, B. O. (2018), "Machine Learning Applied to Enzyme Turnover Numbers Reveals Protein Structural Correlates and Improves Metabolic Models," *Nature Communications*, 9, 1–10. [1070]

Lei, J. (2020), "Cross-Validation with Confidence," *Journal of the American Statistical Association*, 115, 1978–1997. [1071]

LeJeune, D., Javadi, H., and Baraniuk, R. (2020), "The Implicit Regularization of Ordinary Least Squares Ensembles," in *International Conference on Artificial Intelligence and Statistics*. [1062]

Li, H., Siddiqui, O., Zhang, H., and Guan, Y. (2019), "Joint Learning Improves Protein Abundance Prediction in Cancers," *BMC Biology*, 17, 1–14. [1070]

Liu, L., Chin, S. P., and Tran, T. D. (2019), "Reducing Sampling Ratios and Increasing Number of Estimates Improve Bagging in Sparse Regression," in *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–5, IEEE. [1064]

Lopes, M. E. (2019), "Estimating the Algorithmic Variance of Randomized Ensembles via the Bootstrap," *The Annals of Statistics*, 47, 1088–1112. [1061,1062,1064,1066,1069]

Lopes, M. E., Wu, S., and Lee, T. C. (2020), "Measuring the Algorithmic Convergence of Randomized Ensembles: The Regression Setting," *SIAM Journal on Mathematics of Data Science*, 2, 921–943. [1061,1062,1064]

Lugosi, G., and Mendelson, S. (2019), "Mean Estimation and Regression under Heavy-Tailed Distributions: A Survey," *Foundations of Computational Mathematics*, 19, 1145–1190. [1065]

Martínez-Muñoz, G., and Suárez, A. (2010), "Out-of-Bag Estimation of the Optimal Sample Size in Bagging," *Pattern Recognition*, 43, 143–152. [1061]

Mentch, L., and Hooker, G. (2016), "Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests," *The Journal of Machine Learning Research*, 17, 841–881. [1061]

Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. (2012), "How Many Trees in a Random Forest?" in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. [1064]

Patil, P., Du, J.-H., and Kuchibhotla, A. K. (2023), "Bagging in Overparameterized Learning: Risk Characterization and Risk Monotonization," *Journal of Machine Learning Research*, 24, 1–113. [1061,1062,1064,1065,1066,1071]

Patil, P., Kuchibhotla, A. K., Wei, Y., and Rinaldo, A. (2022), "Mitigating Multiple Descents: A Model-Agnostic Framework for Risk Monotonization," arXiv preprint arXiv:2205.12937. [1061,1064,1065]

Peng, W., Coleman, T., and Mentch, L. (2022), "Rates of Convergence for Random Forests via Generalized U-Statistics," *Electronic Journal of Statistics*, 16, 232–292. [1061]

Politis, D. N. (2023), "Scalable Subsampling: Computation, Aggregation and Inference," *Biometrika*, asad021. [1061,1064]

Politis, D. N., and Romano, J. P. (1994), "Large Sample Confidence Regions based on Subsamples under Minimal Assumptions," *The Annals of Statistics*, 22, 2031–2050. [1061]

Pugh, C. C. (2002), *Real Mathematical Analysis*, Cham: Springer. [1061]

Rad, K. R., and Maleki, A. (2020), "A Scalable Estimate of the Out-of-Sample Prediction Error via Approximate Leave-One-Out Cross-Validation," *Journal of the Royal Statistical Society*, Series B, 82, 965–996. [1062]

Rio, E. (2017), "About the Constants in the Fuk-Nagaev Inequalities," *Electronic Communications in Probability*, 22, 1–12. [1066]

Scornet, E., Biau, G., and Vert, J.-P. (2015), "Consistency of Random Forests," *The Annals of Statistics*, 43, 1716–1741. [1061]

Vershynin, R. (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge: Cambridge University Press. [1065]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [1061]

Wager, S., Hastie, T., and Efron, B. (2014), "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife," *The Journal of Machine Learning Research*, 15, 1625–1651. [1064]

Wang, S., Zhou, W., Lu, H., Maleki, A., and Mirrokni, V. (2018), "Approximate Leave-One-Out for Fast Parameter Tuning in High Dimensions," in *International Conference on Machine Learning*. [1062]

Xu, F., Wang, S., Dai, X., Mundra, P. A., and Zheng, J. (2021),"Ensemble Learning Models that Predict Surface Protein Abundance from Single-Cell Multimodal Omics Data," *Methods*, 189, 65–73. [1070]

Zhou, Z., Ye, C., Wang, J., and Zhang, N. R. (2020), "Surface Protein Imputation from Single Cell Transcriptomes by Deep Neural Networks," *Nature Communications*, 11, 651. [1070]