JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–12
https://doi.org/10.1080/01621459.2023.2197686

Taylor & Francis
Taylor & Francis Group

THEORY AND METHODS

Cross-Validation: What Does It Estimate and How Well Does It Do It?

Stephen Bates[a], Trevor Hastie[b], and Robert Tibshirani[c]

[a]Department of Statistics and EECS, University of California, Berkeley, Berkeley, CA; [b]Department of Statistics and Biomedical Data Science, Stanford University, Stanford, CA; [c]Department of Biomedical Data Science and Statistics, Stanford University, Stanford, CA

ABSTRACT

Cross-validation is a widely used technique to estimate prediction error, but its behavior is complex and not fully understood. Ideally, one would like to think that cross-validation estimates the prediction error for the model at hand, fit to the training data. We prove that this is not the case for the linear model fit by ordinary least squares; rather it estimates the average prediction error of models fit on other unseen training sets drawn from the same population. We further show that this phenomenon occurs for most popular estimates of prediction error, including data splitting, bootstrapping, and Mallow's C_p. Next, the standard confidence intervals for prediction error derived from cross-validation may have coverage far below the desired level. Because each data point is used for both training and testing, there are correlations among the measured accuracies for each fold, and so the usual estimate of variance is too small. We introduce a nested cross-validation scheme to estimate this variance more accurately, and show empirically that this modification leads to intervals with approximately correct coverage in many examples where traditional cross-validation intervals fail. Lastly, our analysis also shows that when producing confidence intervals for prediction accuracy with simple data splitting, one should *not* refit the model on the combined data, since this invalidates the confidence intervals. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2021
Accepted February 2023

KEYWORDS
Bootstrap/resampling; Computationally intensive methods; Cross-validation; Goodness-of-fit methods

1. Introduction

When deploying a predictive model, it is important to understand its prediction accuracy on future test points, so both good point estimates and accurate confidence intervals for prediction error are essential. Cross-validation (CV) is a widely-used approach for these two tasks, but in spite of its seeming simplicity, its operating properties remain opaque. Considering first estimation, it is challenging to precisely state the estimand corresponding to the cross-validation point estimate. In this work, we show that the estimand of CV is not the accuracy of the model fit on the data at hand, but is instead the average accuracy over many hypothetical datasets. Specifically, we show that the CV estimate of error has larger mean squared error (MSE) when estimating the prediction error of the final model than when estimating the average prediction error of models across many unseen datasets for the special case of linear regression. Turning to confidence intervals for prediction error, we show that naïve intervals based on CV can fail badly, giving coverage far below the nominal level; we provide a simple example soon in Section 1.1. The source of this behavior is the estimation of the variance used to compute the width of the interval: it does not account for the correlation between the error estimates in different folds, which arises because each data point is used for both training and testing. As a result, the estimate of variance is too small and the intervals are too narrow. To address this issue, we develop *nested cross-validation* (NCV) that achieves coverage near the nominal level.

1.1. A Simple Illustration

As a motivating example where naïve cross-validation confidence intervals fail, we consider a sparse logistic regression model

$$P(Y_i = 1 \mid X_i = x_i) = \frac{1}{1 + \exp\{-x_i^\top \theta\}} \qquad i = 1, \ldots, n, \quad (1)$$

with $n = 90$ observations of $p = 1000$ features, and a coefficient vector $\theta = c \cdot (1, 1, 1, 1, 0, 0, \ldots)^\top \in \mathbb{R}^p$ with four nonzero entries of equal strength. The feature matrix $X \in \mathbb{R}^{n \times p}$ is comprised of iid standard normal variables, and we chose the signal strength c so that the Bayes misclassification rate is 20%. We estimate the parameters using ℓ_1-penalized logistic regression with a fixed penalty level. In this case, naïve confidence intervals for prediction error are far too small: intervals with desired miscoverage of 10% give 31% miscoverage in our simulation. We visualize this in Figure 1. The intervals need to be made larger by a factor of about 1.6 to obtain coverage at the desired level in this case.

1.2. Related Work

Cross-validation is used ubiquitously to estimate the prediction error of a model (Allen 1974; Geisser 1975; Stone 1977). The enduring popularity of CV is due to the fact that it is a conceptually simple improvement over a one-time train-test split (Blum, Kalai, and Langford 1999). CV is part of a broader land-

CONTACT Stephen Bates stephenbates@berkeley.edu Department of Statistics and EECS, University of California, Berkeley, Berkeley, CA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

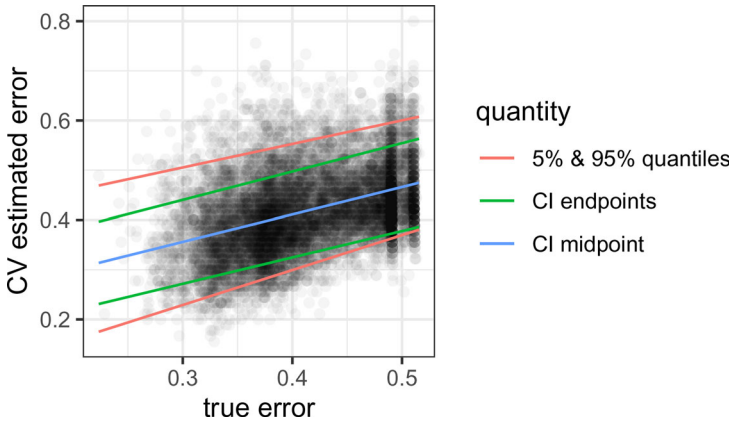


Figure 1. A plot of the true error of a model versus the CV estimates for 1000 replicates of the model from Section 1.1. The blue curve shows the average midpoint of the naïve CV confidence intervals. The green bands show the average 90% confidence interval for prediction error given by naïve CV. The red curves show the 5% and 95% quantiles from a quantile regression fit. To achieve nominal coverage, the green curves should approximate the red curves, but they are too narrow in this case.

scape of resampling techniques to estimate prediction error, with bootstrap-based techniques as the most common alternative (Efron 1983, 1986; Efron and Tibshirani 1997, 1993). The other main category of prediction error estimates are based on analytic adjustments such as Mallow's C_p (Mallows 1973), AIC (Akaike 1974), BIC (Schwarz 1978), and general *covariance penalties* (Stein 1981; Efron 2004). The present work is primarily concerned with CV, but also addresses the properties of bootstrap, data splitting and covariance penalty methods.

In spite of CV's apparent simplicity, the formal properties of this procedure are subtle; the seemingly basic question "what is cross-validation estimating?" has engendered considerable debate. Although the predictive accuracy of the model fit on the observed training data may seem like a natural estimand, it has been observed that the CV estimator tracks this quantity only weakly, suggesting that CV should instead be treated as an estimator of the average prediction error across training sets (Zhang 1995; Hastie, Tibshirani, and Friedman 2009; Yousef 2020). See also Rosset and Tibshirani (2020) and Wager (2020) for a discussion about different potential estimands. In this work, we discuss this phenomenon in detail for the case of the linear model. Our main result uses a conditional independence argument to explain the aforementioned weak relationship between CV and the instance-specific error.

Turning to the question of inference, one important use of CV is to deliver confidence intervals for the prediction error (or, similarly, an estimate of the standard error) to accompany a point estimate. The second primary goal in this work is to provide such confidence intervals, which cannot be reliably created with naïve methods, as shown in our example in Figure 1. A fundamental prior result shows that there is no unbiased estimator of the standard error of the CV point estimate based on one instance of CV (Bengio and Grandvalet 2004). As a result, to obtain standard error estimates, one would either need to modify the CV procedure or make additional assumptions. Pursuing the former, Dietterich (1998) and Nadeau and Bengio (2003) proposes sampling schemes where the data is split in half, and CV is carried out within each half separately. This yields an estimate of standard error, but it will typically be much too conservative since the internal CV model fits each use a samples size that is less than half of the full sample. A related proposal due to Austern and Zhou (2020) involves repeatedly performing leave-one-out CV with datasets of half of the original size, but this proposed estimator is not computationally feasible for most learning algorithms.

In a different direction, Nadeau and Bengio (2003) and Markatou et al. (2005) propose alternative estimates of standard error, but these are based only on the sample size and higher moments of the errors and so do not address the source of the problem: a covariance term that we describe in Section 4.1. For bootstrap estimators, there are proposals to estimate the standard error of the (bootstrap) point estimates of prediction error with methods based on influence functions (Efron 1983; Efron and Tibshirani 1997). The CV proposal of Austern and Zhou (2020) similarly involves leave-one-out resampling, which can be interpreted as an empirical estimate of the influence functions.

Accompanying these algorithmic proposals, there is some theoretical understanding of the asymptotic behavior of CV. Dudoit and van der Laan (2005) proves a central limit theorem (CLT) for a cross-validation estimator, showing asymptotic coverage with a non-CV plug-in estimator for standard error. LeDell, Petersen, and van der Laan (2015) provides a consistent estimator for the standard error in the special case of estimating the AUC, and Benkeser, Petersen, and van der Laan (2020) conducts a higher-order asymptotic analyses for AUC and other metrics, yielding a more efficient estimator for accuracy with a consistent standard error estimate. Further theoretical results establish the asymptotic normality of the CV estimate in more general cases (Austern and Zhou 2020; Bayle et al. 2020). The former considers the average prediction error across training sets (similar to our goal herein), and introduces an asymptotically valid estimate of the standard error; see Supplementary Appendix F.9 for an experiment with this estimator. The latter estimates a different estimand: the average prediction error of the models fit on the subsamples, and introduces a valid estimate of standard error for this quantity. We explain this estimand and the proposed standard error estimator in more detail in Supplementary Appendix I. Both use arguments relying on notions of algorithmic stability (Kale, Kumar, and Vassilvitskii 2011; Kumar et al. 2013; Celisse and Guedj 2016). At present, it is not clear how the large-sample regime considered in these works

relates to the behavior we see in small samples such as in the experiment in Section 1.1. In particular, algorithmic stability may not be satisfied in high-dimensional settings or with small sample sizes; see Supplementary Appendix I and Bayle et al. (2020) for more discussion.

Lastly, we note that CV is often used to compare predictive models, such as when selecting a model or a good value of a learning algorithm's hyperparameters (e.g., Stoica et al. 1986; Shao 1993; Zhang 1993; Dietterich 1998; Xu and Liang 2001). To this end, Yang (2007) and Wager (2020) show that for CV, comparing two models is a statistically easier task than estimating the prediction error, in some sense. While we expect that our proposed estimator would be of use for hyperparameter selection because it yields more accurate confidence intervals for prediction error, we do not pursue this problem further in the present work.

1.3. Our Contribution

This work has two main thrusts. First, we study the choice of estimand for CV, giving results for the special case of the linear model. We prove a finite-sample conditional independence result (Theorem 1) with a supporting asymptotic result (Theorem 2) that together show that CV does not estimate the error of the specific model fit on the observed training set, but is instead estimating the average error over many training sets (Corollaries 2 and 3). We also show that this holds for the other common estimates of prediction error: data splitting (Section 3.4), Mallow's C_p (Section 3.5), and bootstrap (Supplementary Appendix A). Second, we introduce a modified cross-validation scheme to give accurate confidence intervals for prediction error. We prove that our estimate for the MSE of the CV point estimate is unbiased (Theorem 3). Moreover, we validate our method with extensive numerical experiments, confirming that the coverage is consistently better than that of standard cross-validation (Section 5).

2. Setting and Notation

We consider the supervised learning setting where we have features $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ and response $Y = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$, and we assume that the data points (X_i, Y_i) for $i = 1, \ldots, n$ are iid from some distribution P. We wish to understand how well fitted models generalize to unseen data points, which we formalize with a loss function $\ell(\hat{y}, y) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ such that $\ell(y, y) = 0$ for all y. For example, ℓ could be squared error loss, misclassification error, or deviance (cross-entropy). Now consider a class of models parameterized by θ. Let $\hat{f}(x, \theta)$ be the function that predicts y from $x \in \mathbb{R}^p$ using the model with parameters θ, which takes values in some space Θ. Let $\mathcal{A}$ be a model-fitting algorithm that takes any number of data points and returns a parameter vector $\hat{\theta} \in \Theta$. Let $\hat{\theta} = \mathcal{A}(X, Y)$ be the fitted value of the parameter based on the observed data X and Y. We are interested in the *out-of-sample error* with this choice of parameters:

$$\mathrm{Err}_{XY} := \mathbb{E}\left[\ell(\hat{f}(X_{n+1}, \hat{\theta}), Y_{n+1}) \mid (X, Y)\right],$$

where $(X_{n+1}, Y_{n+1}) \sim P$ is an independent test point from the same distribution. Notice Err_{XY} is a random quantity, depending on the training data. We denote the expectation of this quantity across possible training sets as

$$\mathrm{Err} := \mathbb{E}\left[\mathrm{Err}_{XY}\right].$$

We will discuss the relationship between these two quantities further in Section 3. We note that out-of-sample error is materially different from *in-sample-error* which is the focus of methods like the C_p and AIC statistics, and covariance penalties. These are discussed in Section 3.5.

In cross-validation, we partition the observations $\mathcal{I} = \{1, \ldots, n\}$ into K disjoint subsets (*folds*) $\mathcal{I}_1, \ldots, \mathcal{I}_K$ of size $m = n/K$ at random. Throughout this work, we will assume K divides n for convenience, and we will choose $K = 10$ in all of our numerical results. Consider the first fold, and let $\hat{\theta}^{(-1)} = \mathcal{A}((X_j, Y_j)_{j \in \mathcal{I} \setminus \mathcal{I}_1})$ be the model fit to only those points that are not in fold one. Then, let $e_i = \ell(\hat{f}(x_i, \hat{\theta}^{(-1)}), y_i)$ for each $i \in \mathcal{I}_1$. The errors e_i for points in other folds are defined analogously. We let

$$\widehat{\mathrm{Err}}^{(\mathrm{CV})} := \bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{2}$$

be the average error, which is the usual CV estimate of prediction error. If one desires a confidence interval for the prediction error, a straightforward approach is to compute the empirical standard deviation of the e_i divided by $\sqrt{n}$ to get an estimate of the standard error:

$$\widehat{\mathrm{SE}} := \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(e_i - \bar{e})^2}.$$

From here, we can create a confidence interval as

$$(\bar{e} - z_{1-\alpha/2} \cdot \widehat{\mathrm{SE}},\ \bar{e} + z_{1-\alpha/2} \cdot \widehat{\mathrm{SE}}),$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. We call these the *naïve cross-validation intervals* and they serve as our baseline approach. Importantly, we find that these naïve CV intervals are on average too small because the true standard deviation of $\bar{e}$ is larger than the naïve estimate $\widehat{\mathrm{SE}}$ would suggest, so a better estimate of the standard error is needed.

3. What Prediction Error are We Estimating?

We next discuss targets of inference when assessing prediction accuracy. We discuss both Err and Err_{XY}, and also introduce an intermediate quantity Err_X that explains the connection between these two. While cross-validation is our focus, our results hold identically for other estimates of prediction error: covariance penalties (Section 3.5), data splitting (Section 3.4), and bootstrap (Supplementary Appendix A).

3.1. Err_X: A Different Target of Inference

The two most natural estimands of interest to the analyst are Err_{XY}, the error of the model that was fit on our actual training set, and Err, the average error of the fitting algorithm run

on same-sized datasets drawn from the underlying distribution P. The former quantity is of the most interest to a practitioner deploying a specific model, whereas the latter may be of interest to a researcher comparing different fitting algorithms. While it may initially appear that the quantity Err_{XY} is easier to estimate—since it concerns the model at hand—it has been observed that the cross-validation estimate provides little information about Err_{XY} (Zhang 1995; Hastie, Tibshirani, and Friedman 2009; Yousef 2020), a phenomenon sometimes called the *weak correlation* issue.

We now prove that CV has lower MSE for estimating Err than it does for Err_{XY}, for the special case of the linear model. In this sense, CV should be viewed as an estimate of Err rather than of Err_{XY}. In order to state this formally, for this section only, assume the homoscedastic linear model holds:

$$y_i = x_i^\top \theta + \epsilon_i \quad \text{where} \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad i = 1, \ldots, n, \tag{3}$$

with $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ independent of X. In this setting, a key quantity in our analysis is

$$\text{Err}_X := \mathbb{E}[\text{Err}_{XY} \mid X],$$

which falls between Err and Err_{XY}; see Figure 2 for a visualization. This quantity is also considered by Hastie et al. (2019) in a high-dimensional regression setting, but to the best of our knowledge has not been considered in the literature on estimation of prediction error.

While our current focus is on cross-validation, the conclusions hold for a broad class of estimates of prediction error. In particular, we consider estimators of prediction error that satisfy the following property:

Definition 1 (Linearly invariant estimator). We say that an estimator of prediction error $\widehat{\text{Err}}((X_1, Y_1), \ldots, (X_n, Y_n), U)$ is *linearly invariant* if for all x_i, y_i, u we have

$$\begin{aligned}&\widehat{\text{Err}}\big((x_1, y_1), \ldots, (x_n, y_n), u\big) \\ &\quad = \widehat{\text{Err}}\big((x_1, y_1 + x_1^\top \kappa), \ldots, (x_n, y_n + x_n^\top \kappa), u\big).\end{aligned} \tag{4}$$

Here κ is any p-vector and the random variable U is included to allow for randomized procedures like cross-validation, and without loss of generality it is taken to be unif$[0, 1]$ and independent of (X, Y).

With ordinary least square (OLS) fitting, cross-validation satisfies this property:

Lemma 1. When using OLS as the fitting algorithm and squared-error loss, the cross-validation estimate of prediction error, $\widehat{\text{Err}}^{(\text{CV})}$, is linearly invariant.

Note that linear invariance is a deterministic property of an estimator and does not rely on any distributional assumptions.

Recall from classical linear regression theory that when using ordinary least squares (OLS), the estimated coefficient vector is independent of the residual sum of squares. This implies that the sum of squared residuals is independent of the true predictive error. It turns out that even further, the CV estimate of error (and all linearly invariant estimates of error) is independent of the true error, conditional on the feature matrix X.

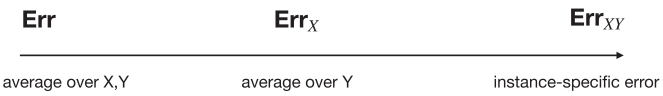


Figure 2. Possible targets of inference for cross-validation. Here, (X, Y) is the training data and Err_{XY} is the average error of the model fit on (X, Y) on a test dataset of infinite size. From left to right, the random variables above are a constant, a function of X only, and a function of (X, Y).

Theorem 1. Assume the homoscedastic Gaussian linear model (3) holds and that we use squared-error loss. Let $\widehat{\text{Err}}$ be a linearly invariant estimate of prediction error (such as $\widehat{\text{Err}}^{(\text{CV})}$ using OLS as the fitting algorithm). Then,

$$\widehat{\text{Err}} \perp\!\!\!\perp \text{Err}_{XY} \mid X. \tag{5}$$

The proof of this theorem rests primarily on the fact that the OLS residuals are independent of the fitted coefficient vector in the linear model, together with the observation that linearly invariant estimators are a function *only* of the residuals of an OLS model fit. Due to its simplicity, we give the proof explicitly here; all other proofs are given in Supplementary Appendix D.

Proof of Theorem 1. The true predictive error (Err_{XY}) is a function only of $\hat{\theta}$, the OLS estimate of θ based on the full sample $(x_1, y_1), \ldots, (x_n, y_n)$. On the other hand, any linearly invariant $\widehat{\text{Err}}$ is a function only of the residuals $Y - X\hat{\theta} = (I - X(X^\top X)^{-1}X^\top)Y$, by the invariance property (see Supplementary Appendix Lemma 6). Since $\hat{\theta} \perp\!\!\!\perp (Y - X\hat{\theta}) \mid X$, from classical linear model results, the proof is complete. □

As a result, any linearly invariant estimator (such as cross-validation) has lower MSE as an estimate of Err_X than as an estimate of Err_{XY}:

Corollary 1. Under the conditions of Theorem 1,

$$\mathbb{E}\left[\left(\widehat{\text{Err}} - \text{Err}_{XY}\right)^2\right] = \mathbb{E}\left[\left(\widehat{\text{Err}} - \text{Err}_X\right)^2\right] + \underbrace{\mathbb{E}\left[\text{var}(\text{Err}_{XY} \mid X)\right]}_{\geq 0}.$$

We demonstrate this in an experiment in a simple linear model with $n = 100$ observations and $p = 20$ features, where the features are iid standard normal variables; see Figure 3. As predicted by Corollary 1, we see that the CV point estimate has lower MSE for Err_X than for Err_{XY}. Similarly, the naïve CV intervals cover Err_X more often than they cover Err_{XY}.

3.2. Relationship with Average Error

The results of the previous section suggest that Err_X is a more natural target of inference than Err_{XY}. Next, we examine the relationship between Err and Err_X, showing that Err_X is close to Err, in that the variance of Err_X (which has mean Err) is small compared with the variance of Err_{XY} (which also has mean Err). Combined with the results of the previous section, this gives a formal statement that cross-validation is a better estimator for Err than for Err_{XY}.

To make this precise, consider the conditional variance decomposition of the variance of Err_{XY},

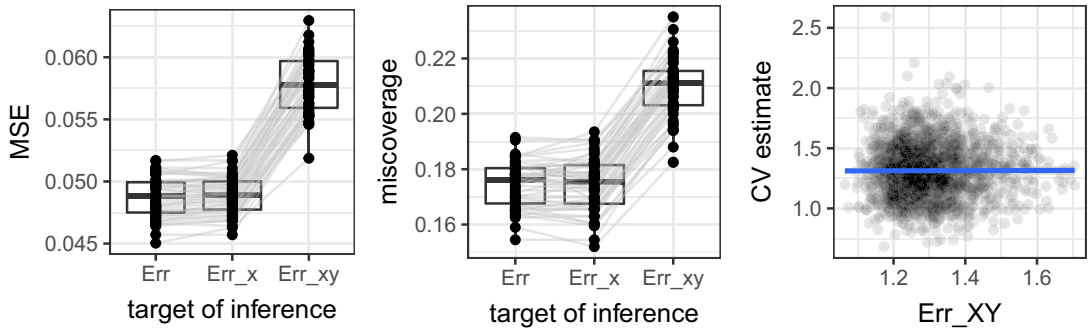


Figure 3. Left: mean squared error of the CV point estimate of prediction error relative to three different estimands: Err, Err_X, and Err_{XY}. Center: coverage of Err, Err_X, and Err_{XY} by the naïve cross-validation intervals in a homoscedastic Gaussian linear model. The nominal miscoverage rate is 10%. Each pair of points connected by a line represents 2000 replicates with the same feature matrix X. Right: 2000 replicates with the same feature matrix and the line of best fit (blue).

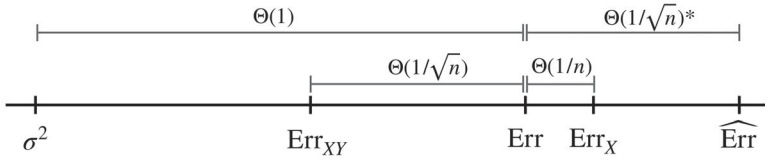


Figure 4. The relationship among various notions of prediction error in the proportional asymptotic limit (7). Recall that σ^2 is the Bayes error: the error rate of the best possible model. See Figure 5 for a simulation experiment demonstrating these rates. *The variance of $\widehat{\mathrm{Err}}$ scales as $1/\sqrt{n}$; see Section 3.3 for details about the bias.

$$\mathrm{var}(\mathrm{Err}_{XY}) = \underbrace{\mathbb{E}_X\left[\mathrm{var}(\mathrm{Err}_{XY} \mid X)\right]}_{\text{var due to } Y \mid X} + \underbrace{\mathrm{var}(\mathrm{Err}_X)}_{\text{var due to } X}. \tag{6}$$

To quantify the relative contribution of the two terms in the right-hand side of (6), we will use a *proportional asymptotic limit*, where

$$n > p, \qquad n, p \to \infty, \qquad n/p \to \lambda > 1. \tag{7}$$

We use the proportional asymptotic limit rather than traditional p fixed, $n \to \infty$ asymptotics, because in the latter asymptotic regime, the difference between Err, Err_X, and Err_{XY} is asymptotic order lower than $1/\sqrt{n}$, so one always estimates these three targets with equal precision, and the analysis is less informative. See supplementary Appendix H for a complementary analysis in the traditional p fixed, $n \to \infty$ asymptotic regime and Yang (2007) and Wager (2020) for a related discussion. By contrast, in the proportional asymptotic limit we will see that $\widehat{\mathrm{Err}}^{(\mathrm{CV})}$ is closer to Err and Err_X than to Err_{XY}.

Theorem 2. Suppose the homoscedastic Gaussian linear model in (3) holds and that we use squared-error loss. In addition, assume that feature vectors $X_i \sim \mathcal{N}(0, \Sigma_p)$ for any full-rank Σ_p. Then, in the proportional asymptotic limit in (7), we have $\mathbb{E}_X\left[\mathrm{var}(\mathrm{Err}_{XY} \mid X)\right] = \Theta(1/n)$ and $\mathrm{var}(\mathrm{Err}_X) = \mathbb{E}(\mathrm{Err}_X - \mathrm{Err})^2 = \Theta(1/n^2)$, as $n, p \to \infty$.

We summarize the asymptotic relationship among the various estimands in Figure 4. We see that the randomness caused by Y given X is of a larger order than that due to the randomness in X. This explains why in Figure 3, the coverage and MSE of cross-validation is similar when estimating either Err or Err_X, but is significantly different when estimating Err_{XY}. As a result, Err_X and Err_{XY} are asymptotically uncorrelated, and moreover, combining this with Theorem 1 shows that $\widehat{\mathrm{Err}}^{(\mathrm{CV})}$ is asymptotically uncorrelated with Err_{XY}, as stated next.

Corollary 2. In the setting of Theorem 2, $\mathrm{cor}(\mathrm{Err}_{XY}, \mathrm{Err}_X) \to 0$ as $n, p \to \infty$. Moreover, for any linearly invariant estimator $\widehat{\mathrm{Err}}$ (such as $\widehat{\mathrm{Err}}^{(\mathrm{CV})}$ using OLS as the fitting algorithm),

$$\mathrm{cor}(\mathrm{Err}_{XY}, \widehat{\mathrm{Err}}) \to 0 \quad \text{as } n, p \to \infty.$$

Notice that this is a marginal result, whereas the similar Theorem 1 is conditional on X. With respect to Figure 4, this result means that the fluctuations of Err_{XY} around Err are asymptotically uncorrelated with the fluctuations of $\widehat{\mathrm{Err}}$ around Err. Combining Theorem 2 with Theorem 1, we conclude that CV has larger error for estimating Err_{XY} than for Err or Err_X:

Corollary 3. In the setting of Theorem 2, let $\widehat{\mathrm{Err}}$ be any linearly invariant estimator (such as $\widehat{\mathrm{Err}}^{(\mathrm{CV})}$ using OLS as the fitting algorithm). Suppose in addition that $\mathrm{var}(\widehat{\mathrm{Err}}) \to 0$ (an extremely weak condition satisfied by any reasonable estimator). Then,

$$\begin{aligned}
&\mathbb{E}\left[\left(\widehat{\mathrm{Err}} - \mathrm{Err}_{XY}\right)^2\right] - \mathbb{E}\left[\left(\widehat{\mathrm{Err}} - \mathrm{Err}_X\right)^2\right] = \Omega(1/n), \\
&\mathbb{E}\left[\left(\widehat{\mathrm{Err}} - \mathrm{Err}_{XY}\right)^2\right] - \mathbb{E}\left[\left(\widehat{\mathrm{Err}} - \mathrm{Err}\right)^2\right] = \Omega(1/n), \quad \text{and} \\
&\left|\mathbb{E}\left[\left(\widehat{\mathrm{Err}} - \mathrm{Err}\right)^2\right] - \mathbb{E}\left[\left(\widehat{\mathrm{Err}} - \mathrm{Err}_X\right)^2\right]\right| = o(1/n).
\end{aligned}$$

The asymptotic theory perfectly predicts the experimental results presented in Figure 5; we see that even for moderate sample size, the scalings is exactly as anticipated. The main conclusion is that for a linearly invariant estimate of prediction error that has precision $1/\sqrt{n}$, our results show that asymptotically one has lower estimation error when estimating Err compared to Err_{XY}. Similarly, the correlation between a linearly invariant estimate and Err_{XY} goes to zero. These theoretical predictions are also corroborated by the experimental results presented in supplementary Appendix Figure 6. Thus, cross-validation is estimating the average error Err more so than the specific error Err_{XY}.

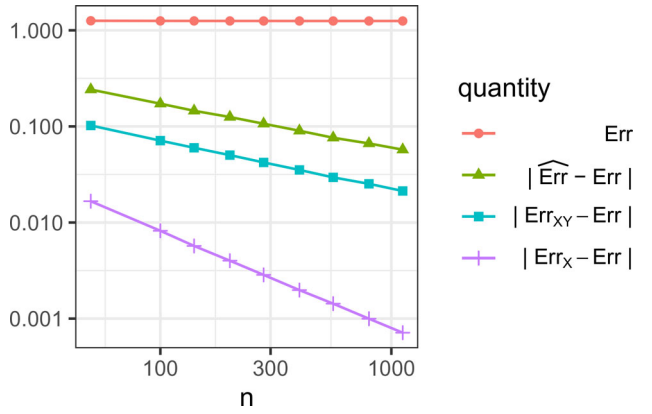


Figure 5. Simulation results demonstrating the asympotic scaling presented in Figure 4. The fitted slopes of the lines (after log-transforming both axes) are 0.00, −0.46, −0.50, −1.01, from top to bottom. See Section 3.3 for details about the rate of $\widehat{\text{Err}}$.

Remark 1. Note that the results in this section apply both to K-fold cross-validation with fixed K, and leave-one-out cross-validation where $K = n$. Formally, the results require only that one is using some sequence of linearly invariant estimators.

3.3. The Bias of Cross-Validation

Up until this point, we have not explicitly mentioned the bias in the CV point estimate $\widehat{\text{Err}}$ that comes from the difference in sample size. That is, $\widehat{\text{Err}}$ uses models of size $n(K-1)/K$, whereas Err and Err_{XY} are defined for models fit on data of size n, so $\mathbb{E}[\widehat{\text{Err}}]$ is typically smaller than $\text{Err} = \mathbb{E}[\text{Err}_{XY}]$. We now pause for a few remarks about this bias. First, notice that Corollary 3 sidesteps the bias issue by considering differences between two mean squared error quantities. The bias is important, however, if we wish to understand absolute quantities such as $\mathbb{E}[(\widehat{\text{Err}} - \text{Err})^2]$. To this end, the bias exhibits different behavior in different regimes[1]:

- *The parametric regime.* Suppose p is fixed, $n \to \infty$, and the model class has fixed dimension. Here, the bias will typically be of order $1/n$, which means that it is negligible compared to the variance. (In fact, the dimension of the model class can grow, provided the rate is slow enough; see Wager (2020) for discussion.)
- *The proportional, dense regime.* Consider the setting above in (7), fitting a dense model. If the number of folds is fixed, the bias of $\widehat{\text{Err}}$ will converge to a nonzero constant as n and p grow (e.g., Liu and Dobriban 2020). What this means is that in Figure 5, the $|\widehat{\text{Err}} - \text{Err}|$ curve will eventually cease to decay at a $1/\sqrt{n}$ rate, bottoming out due to the constant bias. We do not see this behavior in the plot, because the bias is still much smaller than the variance at the sample sizes we consider; supplementary Appendix Figure F.1 reports the bias and variance in this setting.
- *The proportional, sparse regime.* The setting is the most delicate. Here, the behavior of sparse regression algorithms may have very different behavior on samples of size $n(K-1)/K$ versus samples of size n (e.g., Reeves, Xu, and Zadik 2019). Thus, the bias here may be appreciable.

In all cases, the bias can be mitigated by taking a larger number of folds as n grows.

Incorporating bias alongside our results from Section 3.2 leads to an interesting bias-variance-variance decomposition of $\mathbb{E}[(\widehat{\text{Err}} - \text{Err}_{XY})^2]$. Since both $\widehat{\text{Err}}$ and Err_{XY} are random quantities, we cannot use the usual bias-variance decomposition. However, by Corollary 2, these two quantities are asymptotically uncorrelated, yielding the following:

$$\mathbb{E}\left[(\widehat{\text{Err}} - \text{Err}_{XY})^2\right] \approx \underbrace{\left(\mathbb{E}\left[\widehat{\text{Err}}\right] - \text{Err}\right)^2}_{\text{bias}} + \underbrace{\mathbb{E}\left[(\widehat{\text{Err}} - \mathbb{E}[\widehat{\text{Err}}])^2\right]}_{\text{variance of } \widehat{\text{Err}}} + \underbrace{\mathbb{E}\left[(\text{Err}_{XY} - \text{Err})^2\right]}_{\text{variance of } \text{Err}_{XY}}.$$

The first two terms on the right hand side are the bias-variance decomposition for $\widehat{\text{Err}}$ as an estimate of Err. Thus, because the additional third term is positive, we again see that $\widehat{\text{Err}}$ is a more precise estimate of Err than of Err_{XY}.

3.4. Data Splitting

Perhaps the simplest way to estimate prediction error is to split the data into two disjoint sets, one for training and one for estimating the prediction accuracy. The previous results also shed light on the properties of data splitting. In particular, we will show that when estimating prediction error with data splitting, refitting the model on the full data incurs additional variance that can make the confidence intervals slightly too small, even asymptotically. This is not a cause for practical concern, but it is another manifestation of the fact that linearly invariant estimators are estimating average prediction error, and Err_{XY} contains additional, independent variation. We report on the details in supplementary Appendix B.

3.5. Connection with Covariance Penalties

For parametric models, there is an alternative theory of the estimation of prediction accuracy based on *covariance penalties*; see Stein (1981), Efron (2004), Rosset and Tibshirani (2020) for overviews of this approach. For the linear model with OLS and squared error loss, this approach specializes to the well-known Mallows C_p (Mallows 1973; Akaike 1974) estimate of prediction error:

$$\widehat{\text{Err}}^{(C_p)} := \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i, \hat{\theta}))^2 + \frac{2p\hat{\sigma}^2}{n}.$$

We first consider the estimation of prediction error with Mallows C_p, showing that (like cross-validation) it is a worse estimator for Err_{XY} than for Err. The results from Sections 3.1 and 3.2 continue to hold for $\widehat{\text{Err}}^{(C_p)}$ and $\widehat{\text{Err}}^{(RC_p)}$, since they are linearly invariant:

[1] We thank an anonymous reviewer for feedback on this topic.

Lemma 2. The estimators $\widehat{\text{Err}}^{(C_p)}$ and $\widehat{\text{Err}}^{(RC_p)}$ are linearly invariant.

This result is immediate from the fact the $\widehat{\text{Err}}^{(C_p)}$ and $\widehat{\text{Err}}^{(RC_p)}$ are functions only of the residuals of the OLS fit. Thus, the conclusions of Theorem 1, Corollaries 1–3 hold for $\widehat{\text{Err}}^{(C_p)}$. In particular, $\widehat{\text{Err}}^{(C_p)}$ has lower error for estimating Err and Err_X than for estimating Err_{XY}, and $\widehat{\text{Err}}^{(C_p)}$ is asymptotically uncorrelated with Err_{XY}. In summary, as before with cross-validation, Mallow's C_p is not able to estimate Err_{XY}, but is rather an estimate of $\text{Err}_{(\text{in})}$, Err or Err_X (the latter two are close for large samples).

3.6. Bootstrap Estimates of Prediction Error

Bootstrap estimates of prediction error are also linearly invariant, and so they are also estimates of the average prediction error. For brevity, we present these results in supplementary Appendix A.

4. Confidence Intervals with Nested Cross-Validation

In this section, we develop an estimator for the MSE of the cross-validation point estimate. Our ultimate goal is then to use the estimated MSE to give confidence intervals for prediction error with approximately valid coverage.

4.1. Dependence Structure of CV Errors

Before developing our estimator for the cross-validation MSE, we pause here to build up intuition for why the naïve CV confidence intervals for prediction error can fail, as seen previously in our example in Section 1.1. The naïve CV intervals are too small, on average, because the true sampling variance of $\widehat{\text{Err}}^{(\text{CV})}$ is larger than the naïve estimate $\widehat{\text{SE}}$ would suggest. In particular, this estimate of the variance of the CV point estimate assumes that the observed errors $e_1, \ldots, e_n$ are independent. This is not true—the observed errors have less information than an independent sample since each point is used for both training and testing, which induces dependence among these terms. In particular, the covariance matrix of the errors $e_1, \ldots, e_n$ has the block structure shown in Figure 6; see (Bengio and Grandvalet 2004). Thus, the usual estimate of the variance of $\widehat{\text{Err}}^{(\text{CV})}$ is too small, resulting in poor coverage. To remedy this issue, we now develop an estimator that empirically estimates the variance of $\widehat{\text{Err}}^{(\text{CV})}$ across many subsamples. Avoiding the faulty independence approximation leads to intervals with superior coverage.

4.2. Our Target of Inference

In this section, our primary goal will be to give confidence intervals for test accuracy by estimating the mean-squared error (MSE) of cross-validation:

Definition 2. For a sample of size n split into K folds, the *cross-validation MSE* is

$$\text{MSE}_{K,n} := \mathbb{E}\left[\left(\widehat{\text{Err}}^{(\text{CV})} - \text{Err}_{XY}\right)^2\right]. \tag{8}$$

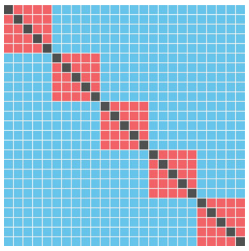


Figure 6. Covariance structure of the CV errors. Red corresponds to the covariance between points in the same fold, and blue corresponds to the covariance between points in different folds.

In particular, we define MSE with respect to Err_{XY} and thus will calibrate our test intervals to cover the quantity Err_{XY}. At this point, the reader should wonder why we define the MSE with respect to Err_{XY} in view of the results from Section 3 that show that $\widehat{\text{Err}}^{(\text{CV})}$ is a more precise estimate of Err than of Err_{XY}, and we will next discuss this issue carefully.

To be clear, we choose to pursue confidence intervals for Err_{XY} because we are able to do so; the MSE quantity above can be estimated in a convenient way due to an upcoming decomposition (Lemma 3). At present, we do not know how to obtain a similar MSE estimate with respect to Err. Second, we emphasize that our results from Section 3 do *not* mean that giving confidence intervals for Err_{XY} is impossible. Rather, our results say that in the linear model, *confidence intervals for* Err_{XY} *will be larger than confidence intervals for Err*. Still, confidence intervals for either Err or Err_{XY} would be of interest to the analyst. We are able to derive an estimator for the MSE with respect to Err_{XY}, and we will turn to the details next.

The MSE in (8) contains both a bias term (due to the reduced sample size used by $\widehat{\text{Err}}^{(\text{CV})}$) and variance term. See Section 3.3 for a discussion of the bias. Thus, we can view the MSE as a slightly conservative version of the variance of the cross-validation estimator. In any case, the MSE is the relevant quantity for creating confidence intervals around a possibly biased point estimate, since it accounts for both bias and variance. With this in mind, we will use an estimate of the MSE to construct confidence intervals for Err_{XY}. Previewing the remainder of this section, we will produce confidence intervals for Err_{XY} as follows:

$$\left(\widehat{\text{Err}}^{(\text{NCV})} - \widehat{\text{bias}} - z_{1-\alpha/2} \cdot \sqrt{\widehat{\text{MSE}}},\right.$$
$$\left.\widehat{\text{Err}}^{(\text{NCV})} - \widehat{\text{bias}} + z_{1-\alpha/2} \cdot \sqrt{\widehat{\text{MSE}}}\right). \tag{9}$$

Above, $\widehat{\text{Err}}^{(\text{NCV})}$ is similar to the CV estimate of error except across many random splits, and $\widehat{\text{MSE}}$ is our estimator for MSE—the heart of this section. In addition, we allow for the possibility of correcting for the sample size bias with an estimator $\widehat{\text{bias}}$; we use one that arises naturally from the computations already carried out to estimate the MSE (Section 4.3.3).

4.3. A Nested CV Estimate of MSE

4.3.1. A Holdout MSE Identity

We now give a generic decomposition of the mean-squared error of an estimate of prediction error, which will enable use

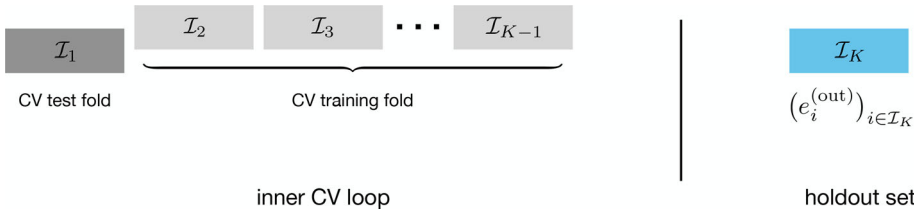


Figure 7. Visualization of nested CV. Using only the folds on left of the vertical line, we perform the usual cross-validation by holding out one fold at a time (the dark grey fold) and then fitting on the remaining folds (the light grey folds). The fresh holdout points (the blue fold) are never used in the inner CV step.

to estimate $\text{MSE}_{K,n}$. Consider a single split of the data into a training set and holdout set, that is, we partition $\mathcal{I} = \{1, \ldots, n\}$ into $\mathcal{I}_{(\text{train})}$ and $\mathcal{I}_{(\text{out})}$ calling the training set $(\widetilde{X}, \widetilde{Y})$. Using only $(\widetilde{X}, \widetilde{Y})$, we use our fitting procedure to obtain estimated parameters $\hat{\theta}^{(\text{train})} = \mathcal{A}(\widetilde{X}, \widetilde{Y})$, and further assume we have some estimate of $\widehat{\text{Err}}_{\widetilde{X}\widetilde{Y}}$ of the prediction error $\text{Err}_{\widetilde{X}\widetilde{Y}}$ defined in Supplementary Appendix (12). Here, $\widehat{\text{Err}}_{\widetilde{X}\widetilde{Y}}$ is any estimator of $\text{Err}_{\widetilde{X}\widetilde{Y}}$ based only on $(\widetilde{X}, \widetilde{Y})$, such as cross-validation using only $(\widetilde{X}, \widetilde{Y})$. Let $\{e_i^{(\text{out})}\}_{i\in\mathcal{I}_{(\text{out})}}$ be the losses of the fitted model $\hat{f}(\cdot, \hat{\theta}^{(\text{train})})$ on the holdout set, and let $\bar{e}^{(\text{out})}$ be their average. The MSE of $\widehat{\text{Err}}_{\widetilde{X}\widetilde{Y}}$ can be written as follows:

Lemma 3 (Holdout MSE identity). In the setting above

$$\underbrace{\mathbb{E}\left[\left(\widehat{\text{Err}}_{\widetilde{X}\widetilde{Y}} - \text{Err}_{\widetilde{X}\widetilde{Y}}\right)^2\right]}_{\text{MSE}} = \underbrace{\mathbb{E}\left[\left(\widehat{\text{Err}}_{\widetilde{X}\widetilde{Y}} - \bar{e}^{(\text{out})}\right)^2\right]}_{\text{(a)}} - \underbrace{\mathbb{E}\left[\left(\bar{e}^{(\text{out})} - \text{Err}_{\widetilde{X}\widetilde{Y}}\right)^2\right]}_{\text{(b)}}. \quad (10)$$

The expectations above are over the complete data (X, Y). The lemma follows from adding and subtracting $\text{Err}_{\widetilde{X}\widetilde{Y}}$ within term (a) then showing the cross-term is zero with a nested conditional expectation argument.

This identity is of interest, since both (a) and (b) can be estimated from the data, which leads to an estimate of the MSE term. Specifically, we propose the following estimation strategy:

1. Repeatedly split the data into $\mathcal{I}_{(\text{train})}$ and $\mathcal{I}_{(\text{out})}$, and for each split, do the following:
   (i) Apply cross-validation to $\mathcal{I}_{(\text{train})}$ to obtain $\widehat{\text{Err}}_{\widetilde{X}\widetilde{Y}}$ and use $\mathcal{I}_{(\text{out})}$ to obtain $\bar{e}^{(\text{out})}$, and then estimate (a) with $(\widehat{\text{Err}}_{\widetilde{X}\widetilde{Y}} - \bar{e}^{(\text{out})})^2$.
   (ii) Estimate (b) with empirical variance of $\{e_i\}_{i\in\mathcal{I}_{(\text{out})}}$ divided by the size of $\mathcal{I}_{(\text{out})}$.
2. Average together estimates of (a) and (b) across all random splits and take their difference as in (10) to obtain an estimate of MSE.

Note that the estimates for both (a) and (b) are unbiased, so the resulting MSE estimate is unbiased for the MSE term in (10). In the next section, we will pursue this strategy for the particular case where $\widehat{\text{Err}}_{\widetilde{X}\widetilde{Y}}$ is itself a cross-validation estimate based only on $(\widetilde{X}, \widetilde{Y})$.

4.3.2. The MSE Estimator

Building from Lemma 3, we now turn to our proposed estimate of MSE, the heart of this section. We follow the estimation strategy described above, using $(K-1)$-fold CV as the estimator $\widehat{\text{Err}}_{\widetilde{X}\widetilde{Y}}$. This gives an estimate of (a) and (b), and hence an estimate for the MSE, as described above. We also get a point estimate of error by taking the empirical mean of $\widehat{\text{Err}}_{\widetilde{X}\widetilde{Y}}$ across the many splits. See Figure 7 for a visualization of the nested CV sample splitting, and see Algorithm 1 in the Supplementary Appendix for a detailed description. We denote the resulting estimate of mean squared error by $\widehat{\text{MSE}}^{(\text{NCV})}$ and the point estimate for prediction error $\widehat{\text{Err}}^{(\text{NCV})}$.

In view of Lemma 3, we see that the estimator $\widehat{\text{MSE}}^{(\text{NCV})}$ is targeting the MSE of $\widehat{\text{Err}}^{(\text{CV})}$ as an estimate of Err_{XY}, as we record formally next.

Theorem 3 (Estimand of nested CV). For a nested CV with a sample of size n, $\mathbb{E}[\widehat{\text{MSE}}^{(\text{NCV})}] = \text{MSE}_{K-1,n'}$, where $n' = n(K-1)/K$.

This result shows that $\widehat{\text{MSE}}^{(\text{NCV})}$ obtained by nested CV is estimating the MSE of $(K-1)$-fold cross-validation on a sample of size $n(K-1)/K$. Since nested CV uses an inner loop with samples of size $n(K-1)/K$, we recommend rescaling to obtain an estimate for a sample of size n by instead taking $\widehat{\text{MSE}} = (K-1)/K \cdot \widehat{\text{MSE}}^{(\text{NCV})}$ (although this rescaled version is not guaranteed to be exactly unbiased for $\text{MSE}_{K,n}$). As a minor detail, in practice we also restrict $\sqrt{\widehat{\text{MSE}}}$ to fall between $\widehat{\text{SE}}$ (the estimated standard error if one had n independent points) and $\sqrt{K} \cdot \widehat{\text{SE}}$ (the estimated standard error if one had only n/K independent points). This is a minor implementation detail prevents implausible values of $\widehat{\text{MSE}}$ from arising. After adjusting the point estimate $\widehat{\text{Err}}^{(\text{NCV})}$ with a bias correction discussed next, we form our final confidence intervals as in (9).

Remark 2 (The sample size difference and the target of inference.). Note that the estimator $\widehat{\text{Err}}^{(\text{NCV})}$ uses models fit on with $n(K-2)/K$ data points, whereas the target of inference in Theorem 3, $\text{MSE}_{K-1,n'}$, is defined with respect to the prediction accuracy of a model fit with $n(K-1)/K$ points. How can the former be used to estimate the latter? The answer is that the nested CV procedure relies also on fits of size $n(K-1)/K$, see the definition of $\hat{\theta}$ in the "nested_crossval" subroutine of Algorithm 1. Nested CV compares the predicted accuracy on the models from $n(K-2)/K$ data points to the estimated accuracy of the

Table 1. Performance of cross-validation (CV), nested cross-validation (NCV), and data splitting with refitting (DS) in the low-dimensional logistic regression model from Section 5.1.1.

Setting		Width		Point estimates				Miscoverage					
								CV		NCV		DS	
Bayes Error	Target	NCV	DS	Err	CV	NCV	DS	Hi	Lo	Hi	Lo	Hi	Lo
33.2%	Err_{XY}	1.23	2.23	39.1%	39.6%	39.0%	40.1%	10%	8%	3%	5%	7%	6%
″	Err	″	″	″	″	″	″	9%	8%	3%	4%	6%	5%
22.5%	Err_{XY}	1.47	2.25	28.7%	30.4%	28.1%	33.3%	11%	3%	4%	1%	16%	4%
″	Err	″	″	″	″	″	″	10%	2%	5%	0%	15%	3%

NOTE: Each row is a setting with a different signal strength, indexed by the Bayes error: the error of the true model. The nominal total error rate is 10%, that is, 5% above and below. A "Hi" miscoverage is one where the confidence interval is too large and the point estimate falls below the interval; conversely for a "Lo" miscoverage. The standard error in each coverage estimate reported is about 0.5%. The "Target" column indicates the target of coverage—the intervals are always generated identically, but we report the coverage of both Err and Err_{XY}.

models with $n(K-1)/K$ data points, using the extra holdout data to asses this accuracy.

4.3.3. Estimation of Bias

The nested CV computations also yield a convenient estimate of the bias of the NCV point estimate of error, $\widehat{\text{Err}}^{(\text{NCV})}$. They key idea is that nested CV considers both models fit with $n(K-2)/K$ data points and with $n(K-1)/K$ data points, and comparing their these models gives an estimate of bias; see Supplementary Appendix C. This aspect of nested CV is not critical—the MSE estimation above is the core of our proposal.

5. Simulation Experiments

We now explore the coverage of nested CV in a variety of settings. In each case, we will report the coverage of naïve CV (CV), nested CV (NCV), and data splitting with refitting (DS), where the nominal miscoverage rate is 10% (5% miscoverage in each tail). We also report on the width of the intervals, expressed relative to the width of the standard CV intervals. (We wish to produce intervals that are as narrow as possible while maintaining correct coverage.) We use 10-fold CV (the number of folds has little impact; see Supplementary Appendix F.4) and NCV, with 200 random splits for the latter; see Supplementary Appendix F.2 for the runtime of each experiment. For classification examples we use binary loss, and form confidence intervals for CV, NCV, and data splitting after taking the binomial variance-stabilizing transformation, described in detail in Supplementary Appendix G. For regression examples, we use squared error loss.

For data splitting, we use 80% of the samples for training and 20% for estimating prediction error. Note that the data splitting without refitting intervals are the same as the data splitting with refitting intervals; the difference is that they are intended to cover different quantities. To make this comparable to CV and nested CV, we report on the coverage of Err and Err_{XY} here, which corresponds to data splitting with refitting. Data splitting without refitting (which seeks to cover the quantity in Supplementary Appendix (12)) will typically have better coverage; we observed relatively accurate coverage in the classification examples and worse coverage in the regression examples, but do not explicitly report these results herein.

Scripts reproducing these experiments are available at *https://github.com/stephenbates19/nestedcv_experiments*.

5.1. Classification

5.1.1. Low-Dimensional Logistic Regression

We consider the logistic regression data generating model 1 with $n = 100$ observations and $p = 20$ features, sampled as i.i.d. standard Gaussian variables. Due to the rotational symmetry of the features, the only parameter that affects behavior is the signal strength, and we explore models with Bayes error of either 33% or 23%. Here, we use (un-regularized) logistic regression as our fitting algorithm. We report the results in Table 1, finding that nested CV gives coverage much closer to the nominal level. Moreover, the point estimates have slightly less bias. We report the size of the NCV intervals relative to their CV counterparts per instance in Supplementary Appendix Figure F.2.

Next, we return to the question of estimands, as in Section 3. Since we do not have analytical results for the logistic regression model, we explore this in simulation. Here, we consider problems where the Bayes error rate is 22.5%, and vary n and p. We investigate two quantities. First, we investigate the correlation of $\widehat{\text{Err}}^{(\text{CV})}$ and Err_{XY}, and we find that it is small but larger than in the OLS case. See Figure 8. Next, we check whether $\widehat{\text{Err}}$ has higher precision for Err than for Err_{XY}. To this end, we compute the expected value of $|\widehat{\text{Err}}^{(\text{CV})} - \text{Err}|$ and of $|\widehat{\text{Err}}^{(\text{CV})} - \text{Err}_{XY}|$, and plot their relative difference in the right panel of Figure 8. We find that the CV point estimate is again slightly more precise as an estimate of Err than of Err_{XY} in this setting.

5.1.2. High-Dimensional Sparse Logistic Regression

We return to the high-dimensional logistic regression model introduced in Section 1.1, generalizing slightly. We consider $n \in \{90, 200\}$ with $p = 1000$ features. The feature matrix has standard normal entries with an autoregressive covariance pattern such that adjacent columns have covariance ρ. In each case, we take $k = 4$ nonzero entries of the covariance matrix and use sparse logistic regression. We report on the results in Table 2 and give the width[2] in Figure 9. Again, NCV gives intervals with coverage much closer to the nominal level.

[2]The width in Figure 9 is reported relative to the version of cross-validation that holds out two folds at a time, since this is what is computed internally during NCV. In table Table 2 and elsewhere, we instead report widths relative to the usual K-fold CV.

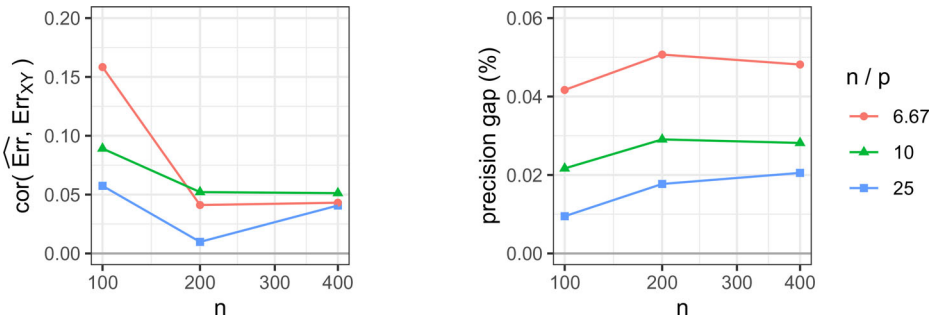
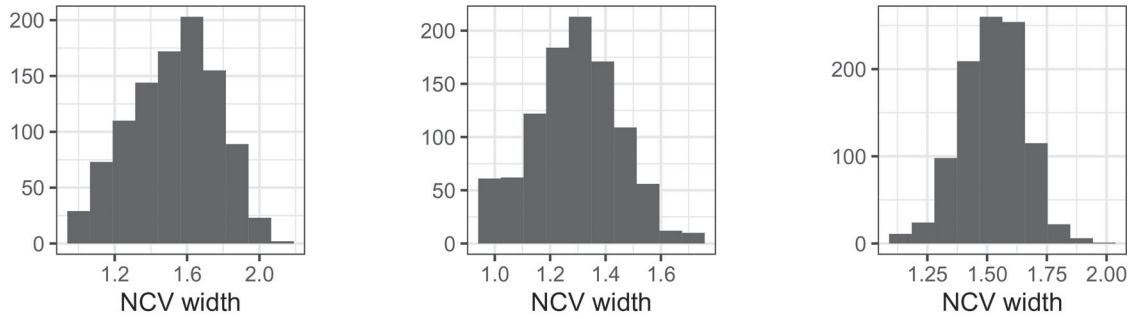


Figure 8. Behavior of cross-validation with a logistic regression model. Left: the correlation between the point estimate and instance-specific error. Right: fraction change in mean absolute deviation of the point estimate with respect to Err_{XY} versus Err.

Table 2. Performance of cross-validation (CV), nested cross-validation (NCV), and data splitting (DS) in the high-d logistic regression model from Section 5.1.2.

Setting			Width		Point estimates				Miscoverage					
									CV		NCV		DS	
n	ρ	Target	NCV	DS	Bayes error	Err	CV	NCV	Hi	Lo	Hi	Lo	Hi	Lo
90	0	Err_{XY}	1.53	2.24	22%	41.3%	41.8%	41.1%	16%	12%	6%	7%	9%	7%
〃	〃	Err	〃	〃	〃	〃	〃	〃	17%	13%	6%	8%	11%	9%
200	0	Err_{XY}	1.66	2.26	22%	25.6%	26.7%	25.6%	14%	7%	3%	5%	9%	4%
〃	〃	Err	〃	〃	〃	〃	〃	〃	15%	7%	4%	6%	8%	5%
90	0.5	Err_{XY}	1.80	2.25	13%	25.6%	27.5%	28.6%	20%	10%	5%	8%	15%	4%
〃	〃	Err	〃	〃	〃	〃	〃	〃	20%	11%	7%	9%	14%	3%

NOTE: The nominal (target) error rate is 10%, that is, 5% above and below. Other details as in Table 1.



(a) $n = 90, \rho = 0$ (b) $n = 200, \rho = 0$ (c) $n = 90, \rho = 0.5$

Figure 9. Size of the nested CV intervals relative to the size of the naïve CV intervals in the high-dimensional sparse logistic regression experiment from Section 5.1.2.

5.2. Regression

We next consider an OLS example. We take $X \in \mathbb{R}^{n \times p}$ with $p = 20$ comprised of iid $\mathcal{N}(0, 1)$. Further, we generate a response from the standard linear model: $Y = X\theta + \epsilon$, where ϵ is likewise iid $\mathcal{N}(0, 1)$. We use OLS to estimate θ. Note that by Lemma 1, the choice of θ does not affect the coverage rate of CV. The same argument shows that the choice of θ will not affect the coverage rate of nested CV, so we can take θ to be 0 without loss of generality. Similarly, both CV and NCV are unchanged when X is transformed by an full-rank linear operator, so the results in this section would remain unchanged for Gaussian features with any full-rank correlation structure. We report the coverage of nested CV in Figure 10. We find that this scheme works well and has good coverage for any n, overcovering somewhat for very small n. By contrast, naïve CV has poor coverage until n is 400. In Supplementary Appendix Figure F.3 we report on the width of the NCV intervals relative to their CV counterparts—the usual ratio is not that large for samples sizes of $n = 100$ or greater. See Supplementary Appendix Section F.11 for an experiment in a high-dimensional regression setting.

6. Real Data Examples

Lastly, we evaluate the nested CV procedure on real datasets from the UCI repository (Dua and Graff 2017). In each case, we repeatedly subsample a small number of observations, perform nested CV on the subsample, and then use the many remaining observations to determine the accuracy of the fitted model. We consider the following datasets:

Communities and crimes (CC). This dataset is comprised of measurements of 1994 communities in the United States. We predict the crime rate of each community, a real number normalized to be between 0 and 1, based on 99 demographic features of the community.

Crop mapping (crp). This dataset is comprised of optical radar measurements of cropland in Manitoba in 2012. We filter the dataset to contain two classes, corn and oats, and then do

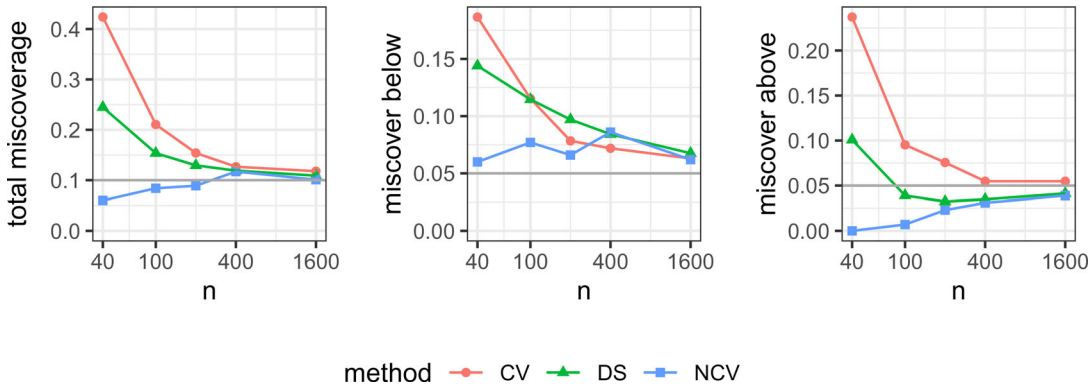


Figure 10. Coverage of CV, data splitting, and nested CV in the OLS case.

Table 3. Performance of cross-validation (CV), nested cross-validation (NCV), and data splitting (DS) with the real datasets.

Setting			Width		Point estimates				Miscoverage					
									CV		NCV		DS	
data	n	Target	NCV	DS	Err	CV	NCV	DS	Hi	Lo	Hi	Lo	Hi	Lo
CC	50	Err_{XY}	2.82	1.77	0.029	0.031	0.029	.034	4%	20%	1%	13%	1%	33%
〃	〃	Err	〃	〃	〃	〃	〃	〃	2%	22%	0%	12%	1%	37%
CC	100	Err_{XY}	1.46	2.05	0.023	0.024	0.023	.025	4%	13%	2%	7%	1%	24%
〃	〃	Err	〃	〃	〃	〃	〃	〃	2%	12%	1%	4%	1%	24%
crp	50	Err_{XY}	1.21	1.89	10.6%	10.7%	10.6%	10.8%	6%	8%	2%	6%	4%	31%
〃	〃	Err	〃	〃	〃	〃	〃	〃	7%	12%	3%	8%	2%	31%
crp	100	Err_{XY}	1.52	2.00	9.5%	9.7%	9.5%	9.4%	6%	6%	4%	5%	4%	15%
〃	〃	Err	〃	〃	〃	〃	〃	〃	8%	9%	5%	7%	4%	15%

NOTE: The nominal (target) error rate is 10%, that is, 5% above and below. Other details as in Table 1.

binary classification based on 174 features. Here, we add a small amount of label noise so that the best possible classifier has a misclassification rate of about 5%.

We again use sparse linear or logistic regression as our fitting algorithm. The results are reported in Table 3. We find that nested CV generally has coverage that is much closer to the nominal rate than naïve CV. Data splitting has poor coverage in this case due to the small sample size, but is significantly better with $= 100$ samples than with $n = 50$ samples.

7. Discussion

Our investigation had two main components. First, we discussed point estimates of prediction error via subsampling techniques. Our primary result is that common estimates of prediction error—cross-validation, bootstrap, data splitting, and covariance penalties—should be viewed as estimates of the *average* prediction error, averaged across other hypothetical datasets from the same distribution. The formal results here were all for the special case of the linear model using unregularized OLS for model-fitting, although we also saw similar behavior in simulation for logistic regression; see Figure 8. A further important question is how regularization affects this behavior. In an additional experiment, we find that $\widehat{\text{Err}}$ does track Err_{XY}, albeit weakly, when there is regularization; see Supplementary Appendix F.10. We look forward to future work explaining the behavior of cross-validation and other estimates of prediction error in these settings.

Second, we discussed inference for cross-validation, deriving an estimator for the MSE of the CV point estimate, nested CV. The nested CV scheme has consistently superior coverage compared to naïve cross-validation confidence intervals, which makes it an appealing choice for providing confidence intervals for prediction error. Nonetheless, we wish to be clear that nested CV is more computationally intensive than standard CV—we use about 1000 times more model fits per example because of the repeated splitting. For example, in the logistic regression example from Section 1.1, nested CV takes about 10 sec on a personal computer.

A fundamental open question is to understand under what conditions the standard CV intervals will be badly behaved, making the nested CV computations necessary. Roughly speaking, we expect the standard CV intervals to perform better when n/p is larger and when more regularization is used. In our experiments, we saw that even in the mundane linear model with $n/p = 10$, the miscoverage rate of standard CV was about 50% larger than the nominal rate. As n increases, however, the violation decreases. Moreover, the asymptotic results in Austern and Zhou (2020) and Bayle et al. (2020) show that the coverage is correct in the p fixed, $n \to \infty$ limiting regime. The stability conditions therein may also be able to shed light on settings with small samples or high-dimensions. We look forward to future work in this direction.

Supplementary Materials

Results for bootstrap estimates of prediction error, additional results for data splitting, details of bias estimation, proofs, additional technical results, additional simulation results, variance stabilization details, low-dimensional asymptotic analyses, and connection with k-fold test error.

Acknowledgments

The authors would like to acknowledge Frank Harrell for a seminar and personal correspondence alerting them to the miscoverage of cross-validation in the high-dimensional logistic regression model. We would like to thank Alexandre Bayle, Michael Celentano, Bradley Efron, Lester Mackey, Adam Smoulder, Ryan Tibshirani, Larry Wasserman, and three anonymous reviewers/editors for helpful comments on earlier versions of this manuscript.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

S. B. was partially supported by a Ric Weiland Graduate Fellowship. T.H. was partially supported by grants DMS-2013736 and IIS 1837931 from the National Science Foundation, and grants 5R01 EB 001988-21 and 5R01 EB001988-16 from the National Institutes of Health. R.T. was supported by the National Institutes of Health (5R01 EB001988-16) and the National Science Foundation (19 DMS1208164).

ORCID

Stephen Bates http://orcid.org/0000-0002-3273-8179
Trevor Hastie http://orcid.org/0000-0002-0164-3142

References

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723. [2,6]

Allen, D. (1974), "The Relationship between Variable Selection and Data Augmentation and a Method of Prediction," *Technometrics*, 16, 125–127. [1]

Austern, M., and Zhou, W. (2020), "Asymptotics of Cross-Validation," arXiv preprint. arXiv:2001.11111. [2,11]

Bayle, P., Bayle, A., Mackey, L., and Janson, L. (2020), "Cross-Validation Confidence Intervals for Test Error," in *Conference on Neural Information Processing Systems*. [2,3,11]

Bengio, Y., and Grandvalet, Y. (2004), "No Unbiased Estimator of the Variance of k-fold Cross-Validation," *Journal of Machine Learning Research*, 5, 1089–1105. [2,7]

Benkeser, D., Petersen, M., and van der Laan, M. J. (2020), "Improved Small-Sample Estimation of Nonlinear Cross-Validated Prediction Metrics," *Journal of the American Statistical Association*, 115, 1917–1932. [2]

Blum, A., Kalai, A. T., and Langford, J. (1999), "Beating the Hold-Out: Bounds for k-fold and Progressive Cross-Validation," in *Proceedings of the Twelfth Annual Conference on Learning Theory*. [1]

Celisse, A., and Guedj, B. (2016), "Stability Revisited: New Generalisation Bounds for the Leave-One-Out," arXiv preprint. arXiv:1608.06412. [2]

Dietterich, T. G. (1998), "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, 10, 1895–1923. [2,3]

Dua, D., and Graff, C. (2017), "UCI Machine Learning Repository." [10]

Dudoit, S., and van der Laan, M. J. (2005), "Asymptotics of Cross-Validated Risk Estimation in Estimator Selection and Performance Assessment," *Statistical Methodology*, 2, 131–154. [2]

Efron, B. (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331. [2]

——— (1986), "How Biased is the Apparent Error Rate of a Prediction Rule?" *Journal of the American Statistical Association*, 81, 461–470. [2]

——— (2004), "The Estimation of Prediction Error," *Journal of the American Statistical Association*, 99, 619–632. [2,6]

Efron, B., and Tibshirani, R. (1997), "Improvements on Cross-Validation: The .632+ Bootstrap Method," *Journal of the American Statistical Association*, 92, 548–560. [2]

——— (1993), *An Introduction to the Bootstrap*, Boca Raton, FL: Chapman & Hall/CRC. [2]

Geisser, S. (1975), "The Predictive Sample Reuse Method with Applications," *Journal of the American Statistical Association*, 70, 320–328. [1]

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019), "Surprises in High-Dimensional Ridgeless Least Squares Interpolation," arXiv preprint. arXiv:1903.08560. [4]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning* (2nd ed.), New York: Springer. [2,4]

Kale, S., Kumar, R., and Vassilvitskii, S. (2011), "Cross-Validation and Mean-Square Stability," in *Proceedings of Innovations in Computer Science*. [2]

Kumar, R., Lokshtanov, D., Vassilvitskii, S., and Vattani, A. (2013), "Near-Optimal Bounds for Cross-Validation via Loss Stability," in *Proceedings of the 30th International Conference on Machine Learning*. [2]

LeDell, E., Petersen, M., and van der Laan, M. (2015), "Computationally Efficient Confidence Intervals for Cross-Validated Area under the ROC Curve Estimates," *Electronic Journal of Statistics*, 9, 1583–1607. [2]

Liu, S., and Dobriban, E. (2020), "Ridge Regression: Structure, Cross-Validation, and Sketching," in *International Conference on Learning Representations*. [6]

Mallows, C. L. (1973), "Some Comments on Cp," *Technometrics*, 15, 661–675. [2,6]

Markatou, M., Tian, H., Biswas, S., and Hripcsak, G. (2005), "Analysis of Variance of Cross-Validation Estimators of the Generalization Error," *Journal of Machine Learning Research*, 6, 1127–1168. [2]

Nadeau, C., and Bengio, Y. (2003), "Inference for the Generalization Error," *Machine Learning*, 52, 239–281. [2]

Reeves, G., Xu, J., and Zadik, I. (2019), "The All-or-Nothing Phenomenon in Sparse Linear Regression," in *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of Proceedings of Machine Learning Research, eds. A. Beygelzimer and D. Hsu, pp. 2652–2663. [6]

Rosset, S., and Tibshirani, R. J. (2020), "From Fixed-x to Random-x Regression: Bias-Variance Decompositions, Covariance Penalties, and Prediction Error Estimation," *Journal of the American Statistical Association*, 115, 138–151. [2,6]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [2]

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486–494. [3]

Stein, C. M. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 9, 1135–1151. [2,6]

Stoica, P., Eykhoff, P., Janssen, P., and Soderstrom, T. (1986), "Model-Structure Selection by Cross-Validation," *International Journal of Control*, 43, 1841–1878. [3]

Stone, M. (1977), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Series B, 36, 111–147. [1]

Wager, S. (2020), "Cross-Validation, Risk Estimation, and Model Selection: Comment on a Paper by Rosset and Tibshirani," *Journal of the American Statistical Association*, 115, 157–160. [2,3,5,6]

Xu, Q.-S., and Liang, Y.-Z. (2001), "Monte Carlo Cross Validation," *Chemometrics and Intelligent Laboratory Systems*, 56, 1–11. [3]

Yang, Y. (2007), "Consistency of Cross Validation for Comparing Regression Procedures," *The Annals of Statistics*, 35, 2450–2473. [3,5]

Yousef, W. A. (2020), "A Leisurely Look at Versions and Variants of the Cross Validation Estimator," arXiv preprint. arXiv:1907.13413. [2,4]

Zhang, P. (1993), "Model Selection Via Multifold Cross Validation," *The Annals of Statistics*, 21, 299–313. [3]

——— (1995), "Assessing Prediction Error in Non-parametric Regression," *Scandinavian Journal of Statistics*, 22, 83–94. [2,4]