Cross-validation: what does it estimate and how well does it do it?

17 V 2021

Alicja Gosiewska

heavily based on
Cross-validation: what does it estimate and how well does it do it?
by
Stephen Bates, Trevor Hastie, Robert Tibshirani

[Submitted on 1 Apr 2021 (v1), last revised 14 Apr 2021 (this version, v2)]

Cross-validation: what does it estimate and how well does it do it?

Stephen Bates, Trevor Hastie, Robert Tibshirani

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 Cp. 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 re-fit the model on the combined data, since this invalidates the confidence intervals.

Subjects: **Methodology (stat.ME)**; Statistics Theory (math.ST); Computation (stat.CO); Machine Learning (stat.ML)

Cite as: arXiv:2104.00673 **[stat.ME]**

(or arXiv:2104.00673v2 **[stat.ME]** for this version)

Submission history

From: Stephen Bates [view email]

[v1] Thu, 1 Apr 2021 17:58:54 UTC (1,695 KB)

[v2] Wed, 14 Apr 2021 16:51:23 UTC (1,734 KB)

Download:

- PDF
- Other formats

(cc) BY

Current browse context:
stat.ME

< prev | next >

new | recent | 2104

Change to browse by:

math
math.ST
stat
stat.CO
stat.ML
stat.TH

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

Export Bibtex Citation

Bookmark

Robert Tibshirani

Robert Tibshirani

OBSERWUJ

Professor of Biomedical Data Sciences, and of Statistics, Stanford University
Zweryfikowany adres z stanford.edu - Strona główna

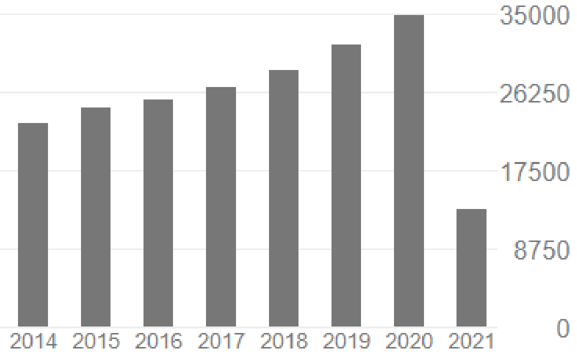
Statistics data science Machine Learning

TYTUŁ	CYTOWANE PRZEZ	ROK
elements of statistical learning hastie	57246 *	
The elements of statistical learning: data mining, inference and prediction T Hastie, R Tibshirani, J Friedman, J Franklin The Mathematical Intelligencer 27 (2), 83-85	~~56540~~ *	2005
An introduction to the bootstrap B Efron, RJ Tibshirani CRC press	46166	1994
Regression shrinkage and selection via the lasso R Tibshirani Journal of the Royal Statistical Society. Series B (Methodological), 267-288	39293	1996
Generalized Additive Models TJ Hastie, RJ Tibshirani CRC Press	19230	1990

UTWÓRZ SWÓJ PROFIL

Cytowane przez WYŚWIETL WSZYSTKO

	Wszystkie	Od 2016
Cytowania	376687	160762
h-indeks	160	110
i10-indeks	486	386



Dostęp publiczny WYŚWIETL WSZYSTKO

Trevor Hastie

Trevor Hastie

Professor of Statistics, Stanford University
Verified email at stanford.edu - Homepage
Statistical learning and mod... data mining machine learning

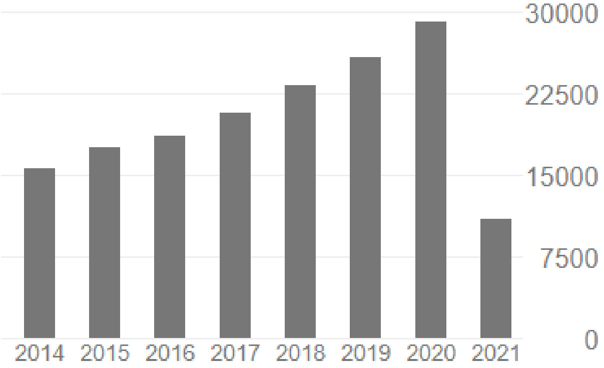
FOLLOW

TITLE	CITED BY	YEAR
The elements of statistical learning: data mining, inference, and prediction T Hastie, R Tibshirani, J Friedman Springer Science & Business Media	57180	2009
Generalized Additive Models THR Tibshirani	19371 *	1990
Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications T Sørlie, CM Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, ... Proceedings of the National Academy of Sciences 98 (19), 10869-10874	14600 *	2001
Regularization and variable selection via the elastic net H Zou, T Hastie Journal of the royal statistical society: series B (statistical methodology ...	13698	2005

GET MY OWN PROFILE

Cited by VIEW ALL

	All	Since 2016
Citations	262943	128607
h-index	138	105
i10-index	375	285



Stephen Bates

Stephen Bates

research
code
teaching
contact

made with
jemdoc

I'm a postdoctoral researcher with Michael I. Jordan in the Statistics and EECS departments at UC Berkeley. I work on developing methods to analyze modern scientific data sets, leveraging sophisticated black box models while providing rigorous statistical guarantees. Specifically, I work on problems in high-dimensional statistics (especially false discovery rate control), statistical machine learning, conformal prediction and causal inference.

Previously, I completed my Ph.D. in the Stanford Department of Statistics advised by Emmanuel Candès. My thesis introduced methods for conditional independence testing and false discovery rate control in genomics, and I was honored to receive the Ric Weiland Graduate Fellowship and the Theodore W. Anderson Theory of Statistics Dissertation Award for this work. Before my Ph.D., I studied statistics and mathematics at Harvard University, and spent a year teaching mathematics at NYU Shanghai. Outside research, I enjoy triathlons, sailing, hiking, and reading speculative fiction novels.

News

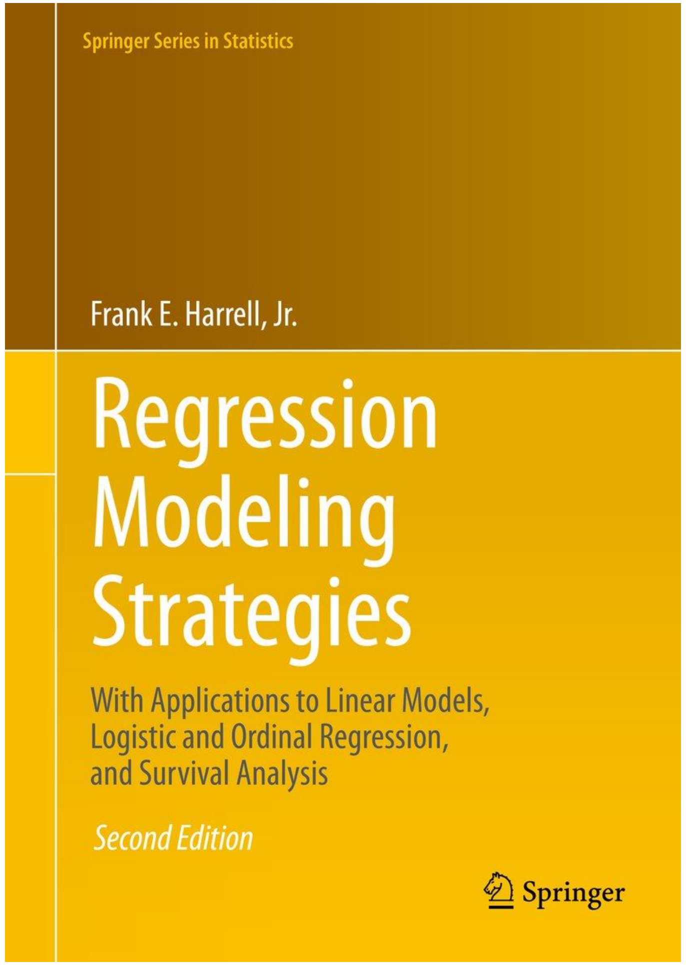
I'm co-organizing the **2021 ICML Workshop on Distribution-free Uncertainty Quantification**, which will take place on Saturday, July 24, 2021.

Select recent papers

"Cross-validation: what does it estimate and how well does it do it?"
S. Bates, T. Hastie, and R. Tibshirani. *arXiv preprint*, 2021.
[arXiv] [code] [bibtex]

https://github.com/stephenbates19/nestedcv_experiments

But not only them ☺



rob tibshirani @robtibshirani · Apr 1

With postdoc Stephen Bates and Trevor Hastie, I have just completed a new paper "Cross-validation: what does it estimate and how well does it do it?" statweb.stanford.edu/~tibs/ftp/NCV.... ✅

18 390 1.2K

rob tibshirani @robtibshirani · Apr 1

We do two things (a) we establish, for the first time, what exactly CV is estimating and (b) we show that the SEs from CV are often WAY too small, and show how to fix them. We also show similar properties to those in (a) for bootstrap, Cp, AIC and data splitting. I'm excited!

1 10 85

rob tibshirani
@robtibshirani

Replying to @robtibshirani

This work was directly motivated by a great talk given in 2010 at Stanford by @f2harrell

2:51 AM · Apr 1, 2021 · Twitter Web App

2 Retweets **57** Likes

rob tibshirani @robtibshirani · Apr 2

Replying to @robtibshirani

Not to downplay the contribution of my long-term co-author and friend Trevor Hastie, but our former Stanford Statistics student @stats_stephen deserves much of the credit for this work. He is amazing.

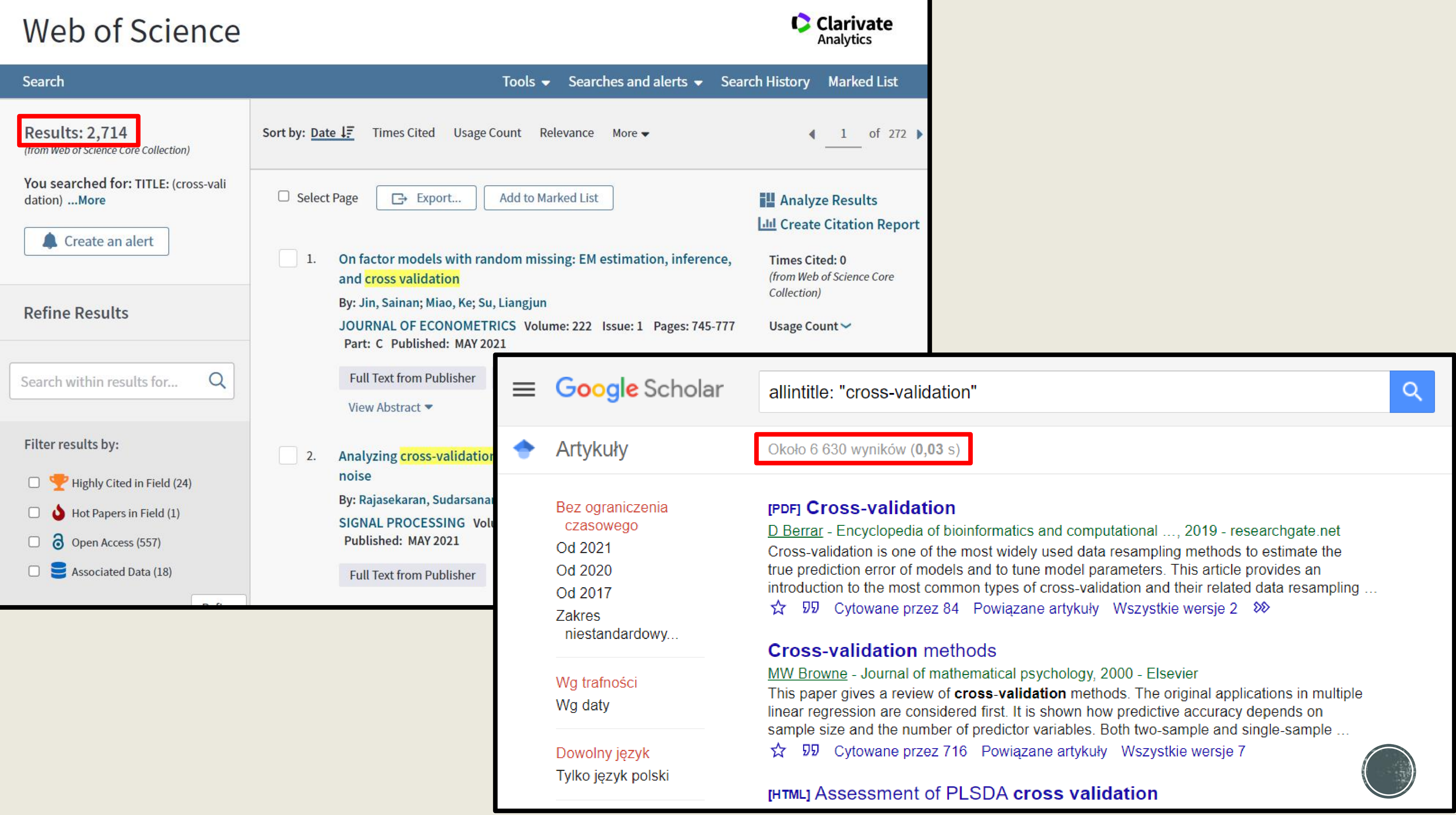
1 19

Frank Harrell @f2harrell · Apr 1

Replying to @robtibshirani

Rob you are so generous to say that.   Can't wait to read the paper.   Minor correction: talk was in Oct. 2019.

6


Web of Science
Clarivate Analytics
Search
Tools
Searches and alerts
Search History
Marked List
Results: 2,714
(from Web of Science Core Collection)
You searched for: TITLE: (cross-validation) ...More
Create an alert
Refine Results
Search within results for...
Filter results by:
Highly Cited in Field (24)
Hot Papers in Field (1)
Open Access (557)
Associated Data (18)
Sort by: Date
Times Cited
Usage Count
Relevance
More
1 of 272
Select Page
Export...
Add to Marked List
Analyze Results
Create Citation Report
1. On factor models with random missing: EM estimation, inference, and cross validation
By: Jin, Sainan; Miao, Ke; Su, Liangjun
JOURNAL OF ECONOMETRICS Volume: 222 Issue: 1 Pages: 745-777 Part: C Published: MAY 2021
Times Cited: 0
(from Web of Science Core Collection)
Usage Count
Full Text from Publisher
View Abstract
2. Analyzing cross-validation
noise
By: Rajasekaran, Sudarsana
SIGNAL PROCESSING Vol
Published: MAY 2021
Full Text from Publisher
Google Scholar
allintitle: "cross-validation"
Artykuły
Około 6 630 wyników (0,03 s)
Bez ograniczenia czasowego
Od 2021
Od 2020
Od 2017
Zakres niestandardowy...
Wg trafności
Wg daty
Dowolny język
Tylko język polski
[PDF] Cross-validation
D Berrar - Encyclopedia of bioinformatics and computational ..., 2019 - researchgate.net
Cross-validation is one of the most widely used data resampling methods to estimate the true prediction error of models and to tune model parameters. This article provides an introduction to the most common types of cross-validation and their related data resampling ...
Cytowane przez 84 Powiązane artykuły Wszystkie wersje 2
Cross-validation methods
MW Browne - Journal of mathematical psychology, 2000 - Elsevier
This paper gives a review of cross-validation methods. The original applications in multiple linear regression are considered first. It is shown how predictive accuracy depends on sample size and the number of predictor variables. Both two-sample and single-sample ...
Cytowane przez 716 Powiązane artykuły Wszystkie wersje 7
[HTML] Assessment of PLSDA cross validation

CROSS-VALIDATION BEHAVIOR IS COMPLEX AND NOT FULLY UNDERSTOOD


Training set
Training folds
Test fold
1st iteration
2nd iteration
3rd iteration
10th iteration
E_1
E_2
E_3
E_{10}
$E = \frac{1}{10}\sum_{i=1}^{10} E_i$

A simple illustration
A sparse logistic model

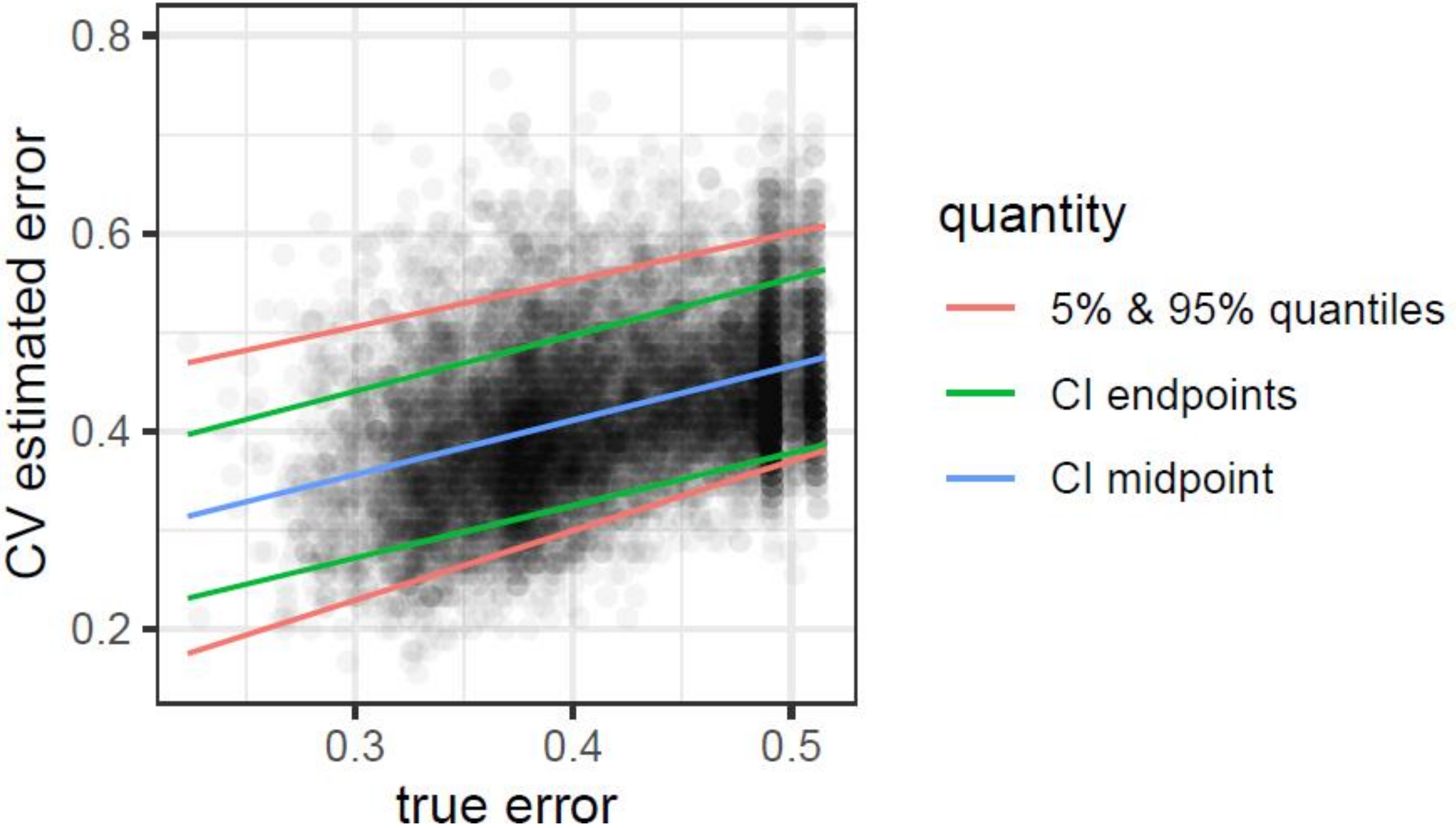
$$P(Y_i = 1 \mid X_i = x_i) = \frac{1}{1 + \exp\{-x_i^\top \theta\}} \qquad i = 1, \ldots, n,$$

$n = 90$ observations of $p = 1000$ features

coefficient vector $\theta = c \cdot (1, 1, 1, 1, 0, 0, \ldots)^\top \in \mathbb{R}^p$

Bayes misclassification rate is 20%.

L1-penalized logistic regression



Two main contributions:

- They 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.
- They introduce a modfied cross-validation scheme to give accurate confidence intervals for prediction error.

What prediction error are we estimating?

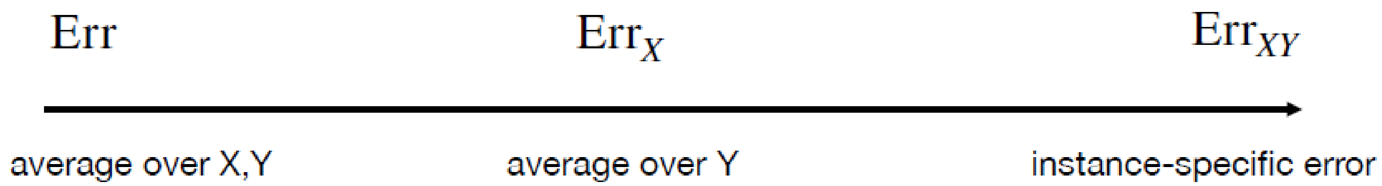
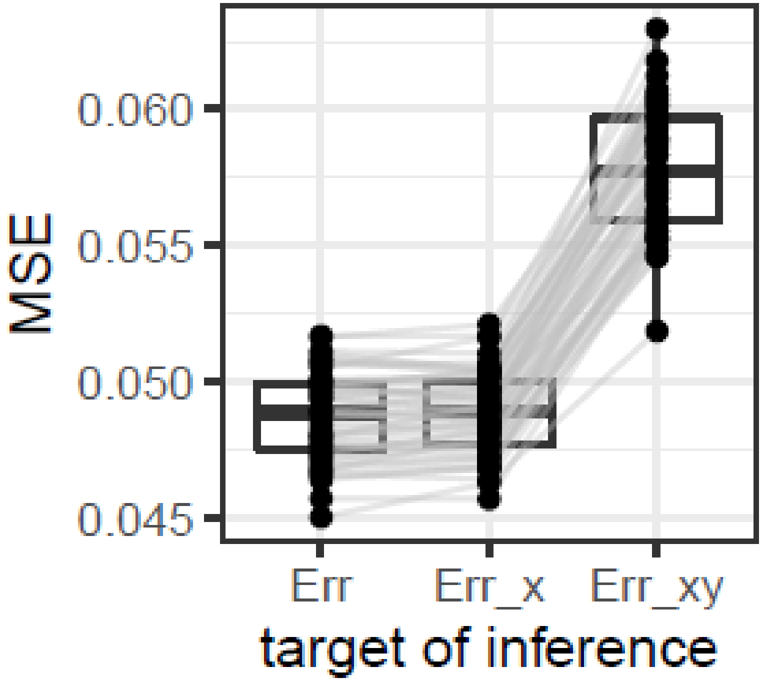


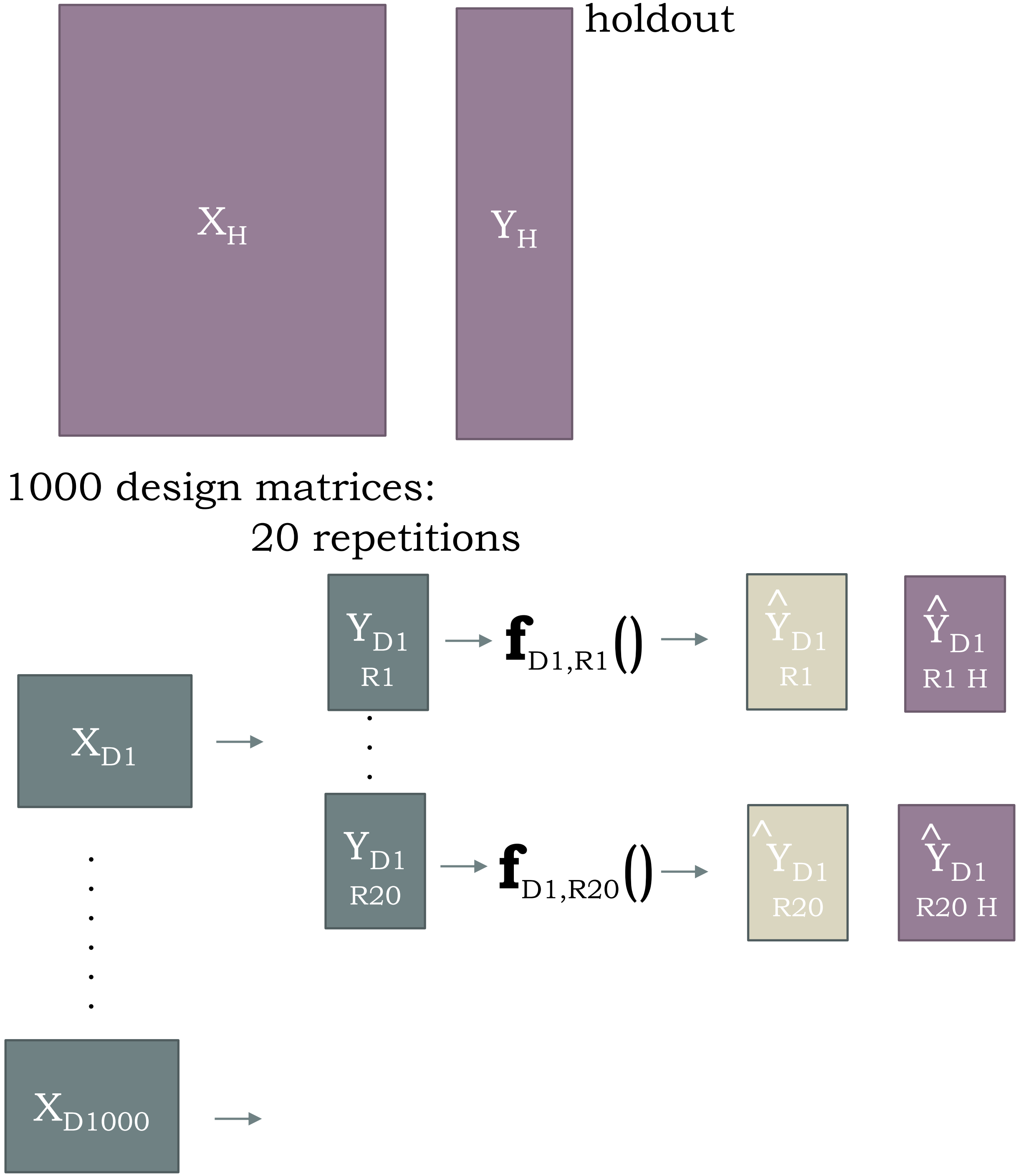
Figure 2: Possible targets of inference for cross-validation. Here, (X, Y) *is the training data and* Err_{XY} *is the average error of the model fit on* (X, Y) *on a test data set of infinite size. From left to right, the random variables above are a constant, a function of* X *only, and a function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

Weak correlation issue

A simple linear model with n=100 observations and p=20 features, where the features are i.i.d. standard normal variables.



Mean squared error of the CV point estimate of prediction error relative to three different estimands: Err, Err_X, and Err_{XY}.



$$\hat{Err}_{CV} = \text{mean}\left(\left(\hat{Y}_{D1,\,R1} - Y_{D1,\,R1}\right)^2\right)$$

$$Err_{XY}(R1, D1) = \text{mean}\left(\left(\hat{Y}_{D1,\,R1\ H} - \hat{Y}_{H}\right)^2\right)$$

$$Err_X(R1) = \text{mean}\left(Err_{XY}(D1, R1), \ldots, Err_{XY}(D1, R20)\right)$$

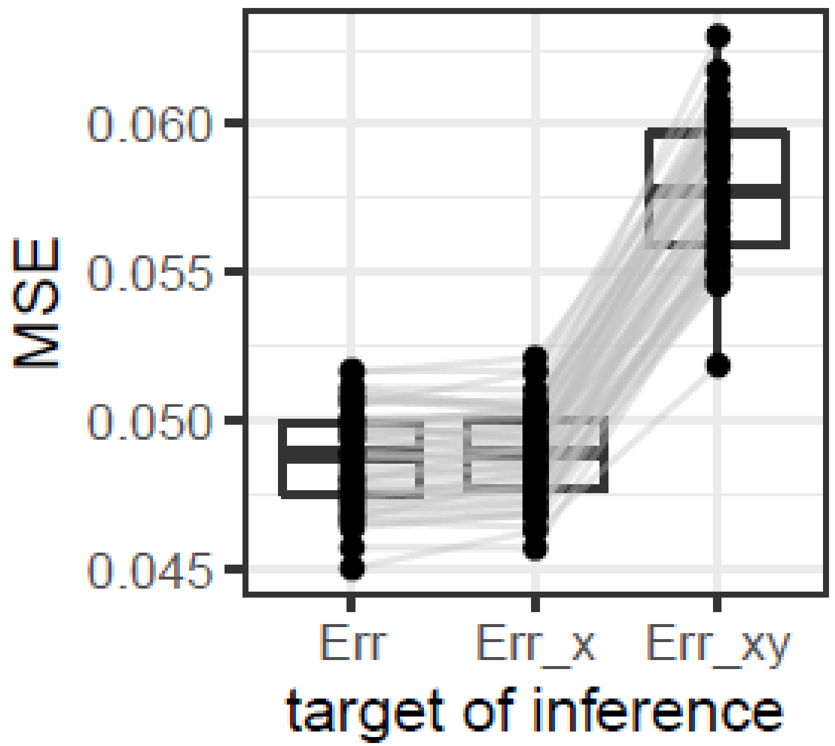
$$Err = \text{mean}\left(Err_X(D1), \ldots, Err_X(D1000)\right)$$

$$\text{MSE } Err_{XY} = \text{mean}\left(\left(Err_{CV} - Err_{XY}(Ri, Di)\right)^2\right)$$

$$\text{MSE } Err_X = \text{mean}\left(\left(Err_{CV} - Err_X(Ri)\right)^2\right)$$

$$\text{MSE } Err = \text{mean}\left(\left(Err_{CV} - Err\right)^2\right)$$

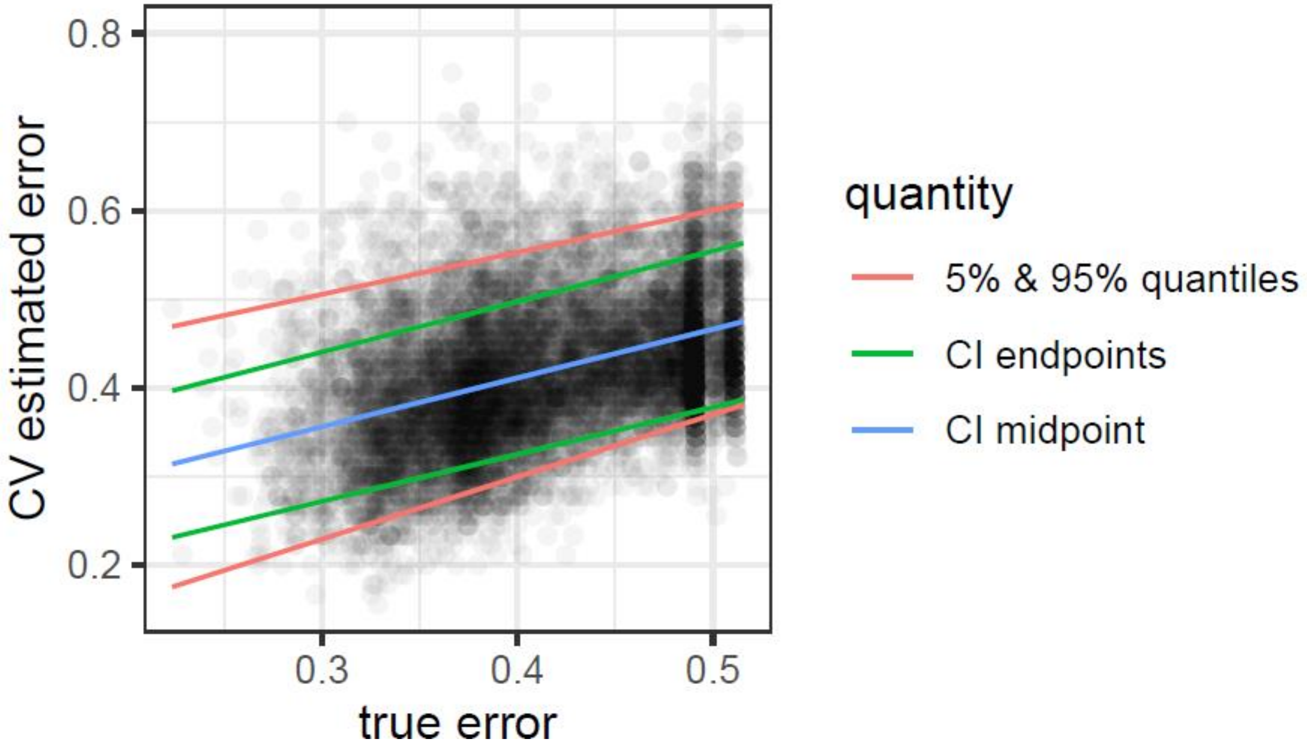
Weak correlation issue



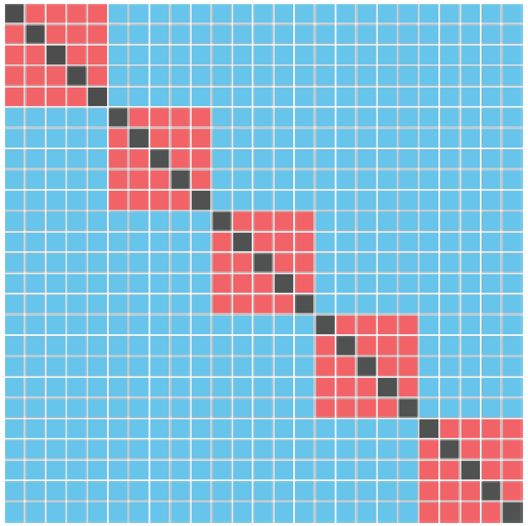
Also a formal proof!

For linear model and OLS

Confidence intervals with nested cross-validation



The estimate of the variance of the CV assumes that the observed errors $e_1, \ldots, e_n$ are independent.



n x n

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

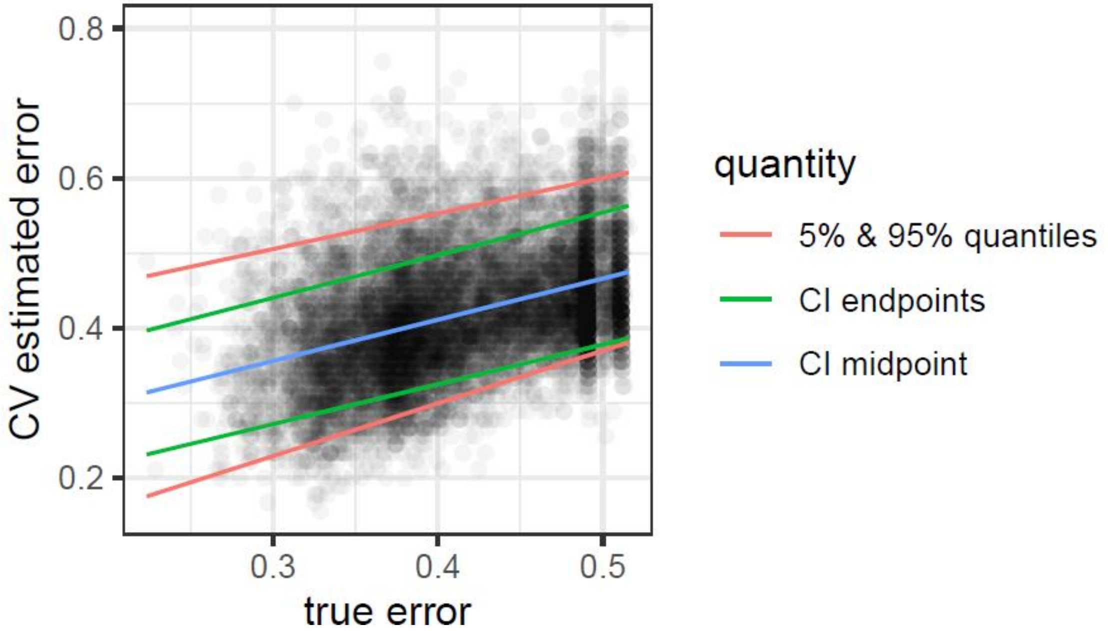
The covariance matrix is parameterized by only three numbers:
$a_1 = \mathrm{var}(e_1)$,
$a_2 := \mathrm{cov}(e_i, e_j)$ for i, j in the same fold,
$a_3 := \mathrm{cov}(e_i, e_j)$ for i; j in different folds

$$\mathrm{var}(\bar{e}) = \frac{1}{n}a_1 + \frac{n/K - 1}{n}a_2 + \frac{n - n/K}{n}a_3$$

The covariance matrix is parameterized by only three numbers:
$a_1 = \mathrm{var}(e_1)$,
$a_2 := \mathrm{cov}(e_i, e_j)$ for i, j in the same fold,
$a_3 := \mathrm{cov}(e_i, e_j)$ for i; j in different folds

$$\mathrm{var}(\bar{e}) = \frac{1}{n} a_1 + \frac{n/K - 1}{n} \underbrace{a_2}_{>0} + \frac{n - n/K}{n} \underbrace{a_3}_{>0}; \longrightarrow \mathrm{var}(\bar{e}) > \frac{1}{n} a_1$$

typically



The estimated variance is approximately a factor of 2.65 too small, so the naive confidence intervals are too small by a factor of $x = \sqrt{2.65} \sim\ 1.6$.

Beyond the usual cross-validation

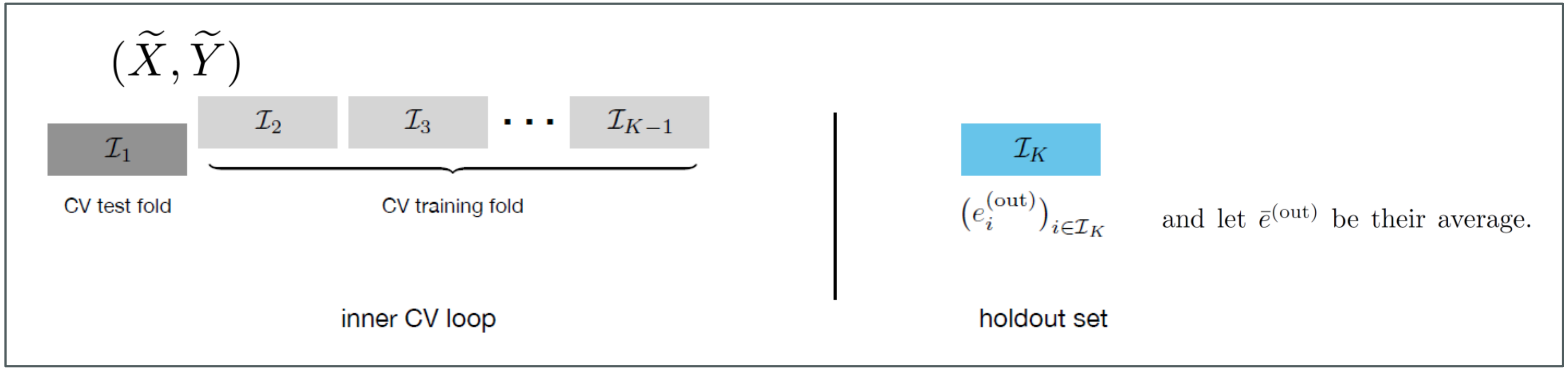
The covariance matrix is parameterized by only three numbers:
a_1= var(e_1),
a_2 := cov(e_i, e_j) for i, j in the same fold,
a_3 := cov(e_i, e_j) for i; j in different folds

Bengio and Grandvalet (2004) proves surprising fact that there is no unbiased estimator of var(e) based on a single run of cross-validation. Thus, estimating a1, a2, and a3 cannot be done from a single run of cross-validation.

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.

Nested cross-validation

(X, Y)



Lemma 4 (Holdout MSE identity). *In the setting above*

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

1. Repeatedly split the data into $\mathcal{I}_{(\mathrm{train})}$ and $\mathcal{I}_{(\mathrm{out})}$, and for each split, do the following:
   (i) Compute $\widehat{\mathrm{Err}}_{\tilde{X}\tilde{Y}}$ and $\bar{e}^{(\mathrm{out})}$, and estimate (a) with $(\widehat{\mathrm{Err}}_{\tilde{X}\tilde{Y}}$ - $\bar{e}^{(\mathrm{out})})^2$.
   (ii) Estimate (b) with empirical variance of $\{e_i\}_{i \in \mathcal{I}_{(\mathrm{out})}}$.
2. Average together estimates of (a) and (b) across all random splits and take their difference

Nested CV Algorithm

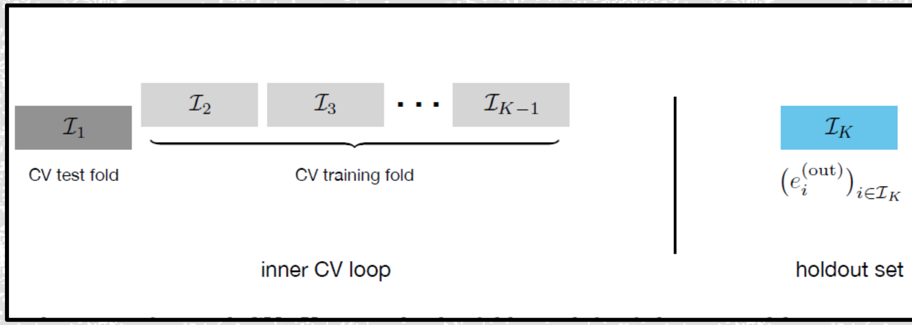
Algorithm 1 Nested Cross-validation

Input: data (X, Y), fitting algorithm $\mathcal{A}$, loss ℓ, number of folds, K, number of repetitions R

procedure NESTED_CROSSVAL(X, Y) ▷ primary algorithm
  es $\leftarrow [\,]$ ▷ initialize empty vectors
  a_list $\leftarrow [\,]$ ▷ (a) terms
  b_list $\leftarrow [\,]$ ▷ (b) terms
  for $r \in \{1, \ldots, R\}$ **do**
    Randomly assign points to folds $\mathcal{I}_1, \ldots, \mathcal{I}_K$
    for $k \in \{1, \ldots, K\}$ **do** ▷ outer CV loop
      $e^{(\mathrm{in})} \leftarrow$ INNER_CROSSVAL$(X, Y, \{\mathcal{I}_1, \ldots, \mathcal{I}_K\} \setminus \mathcal{I}_k)$ ▷ inner CV loop
      $\hat{\theta} \leftarrow \mathcal{A}\left((X_i, Y_i)_{i \in \mathcal{I} \setminus \mathcal{I}_k}\right)$
      $e^{(\mathrm{out})} \leftarrow \left(\ell(\hat{f}(X_i, \hat{\theta}), Y_i)\right)_{i \in \mathcal{I}_k}$
      a_list $\leftarrow$ append $\left(\text{a\_list}, \left(\mathrm{mean}(e^{(\mathrm{in})}) - \mathrm{mean}(e^{(\mathrm{out})})\right)^2\right)$
      b_list $\leftarrow$ append $\left(\text{b\_list}, \mathrm{var}(e^{(\mathrm{out})})\right)$
      es $\leftarrow$ append(es, $e^{(\mathrm{in})}$)
  $\widehat{\mathrm{MSE}} \leftarrow$ mean(a_list) $-$ mean(b_list) ▷ plug-in estimator based on (15)
  $\widehat{\mathrm{Err}}^{(\mathrm{NCV})} \leftarrow$ mean(es)
  return: $(\widehat{\mathrm{Err}}^{(\mathrm{NCV})}, \widehat{\mathrm{MSE}})$ ▷ prediction error estimate and MSE estimate

procedure INNER_CROSSVAL$(X, Y, \{\mathcal{I}_1, \ldots, \mathcal{I}_{K-1}\})$ ▷ inner cross-validation subroutine
  $e^{(\mathrm{in})} \leftarrow [\,]$
  for $k \in \{1, \ldots, K-1\}$ **do**
    $\hat{\theta} \leftarrow \mathcal{A}\left((X_i, Y_i)_{i \in \mathcal{I}_1 \cup \cdots \cup \mathcal{I}_{K-1} \setminus \mathcal{I}_k}\right)$
    $e^{(\mathrm{temp})} \leftarrow \left(\ell(\hat{f}(X_i, \hat{\theta}), Y_i)\right)_{i \in \mathcal{I}_k}$
    $e^{(\mathrm{in})} \leftarrow$ append$(e^{(\mathrm{in})}, e^{(\mathrm{temp})})$
  return: $e^{(\mathrm{in})}$

Output: NESTED_CROSSVAL(X, Y)



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

Simulation experiments
Low-dimensional logistic regression

Setting		Width	Point estimates			Miscoverage			
						CV		NCV	
Bayes Error	Target	NCV	Err	CV	NCV	Hi	Lo	Hi	Lo
33.2%	Err_{XY}	1.23	39.1%	39.6%	40.1%	10%	8%	3%	5%

Takeouts

- Common estimate of prediction error - cross-validation cannot be viewed as estimates of the prediction error of the final model fit on the whole data. Rather, the estimate of prediction error is an estimate of the average prediction error of the final model across other hypothetical data sets from the same distribution.
- The nested CV scheme has consistently superior coverage compared to naive crossvalidation confidence intervals.

Discussion

- Note that the formal results here were all for the special case of the linear model using unregularized OLS for model fitting.
  What about regularization?
- The nested CV is computationally intensive.
- A fundamental open question is to understand under what conditions the standard CV intervals will be badly behaved, making the nested CV computations necessary.