Some Comments on CV



Trevor Hastie - Stanford University

Breiman Award Lecture

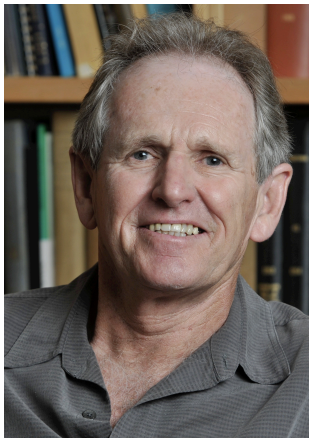
Cross-validation: What does it estimate and how well does it do it? https://arxiv.org/abs/2104.00673

hastie@stanford.edu

The team



Stephen Bates
UC Berkeley Statistics,
EECS



Trevor Hastie
Stanford Statistics,
Biomedical Data Science



Robert Tibshirani
Stanford Statistics,
Biomedical Data Science

Dedicated to Leo Breiman (1928-2007) and Colin Mallows (1930-)





Leo: Random Forests and OOB
Colin: Some Comments on Cp

Estimating prediction accuracy

Foundational statistical topic:

- cross-validation

  Allen (1974), Geisser (1975), Stone (1977)

- bootstrap

  Efron (1983), Efron (1986), Breiman (1996), Efron and Tibshirani (1997)

- Cp, AIC, BIC, Covariance penalties

  Mallows (1973), Akaike (1974), Schwarz (1978), Stein (1981), Efron (2004)

- data splitting

  …

cross-validation: **6,000,000** Google scholar results

Setting

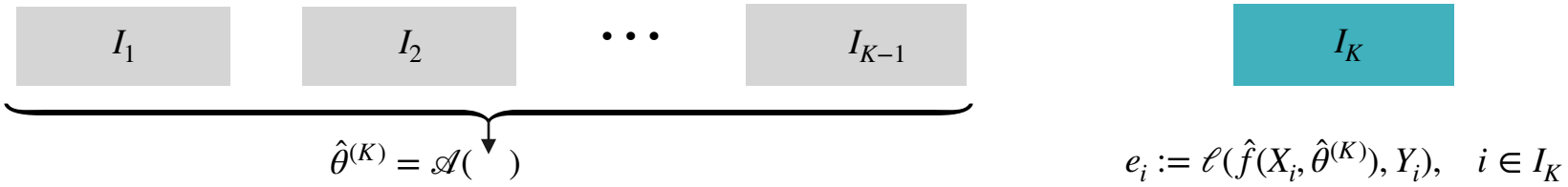
Data: $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} \qquad i = 1, \ldots, n$ (i.i.d.)

Predictive model: $\hat{f}(\,\cdot\,, \theta)$ for $\theta \in \Theta$

Fitting algorithm: $\mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n)) = \hat{\theta}$

Loss function: $\ell(\hat{y}, y) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$

K folds: $I_1 \cup I_2 \cup \ldots \cup I_K = \{1, \ldots, n\}$



Setting

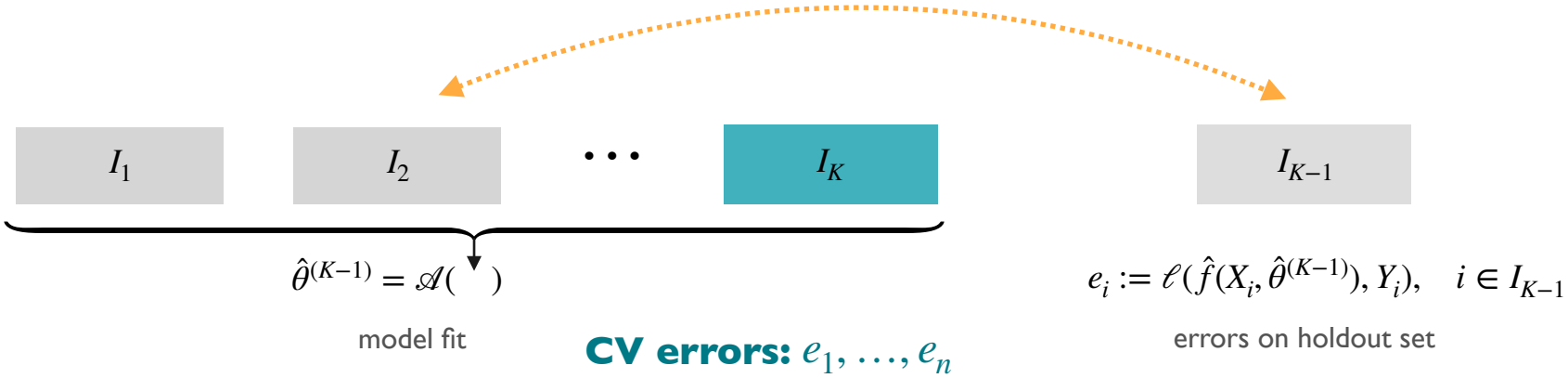
Data: $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} \qquad i = 1, \ldots, n$ (i.i.d.)

Predictive model: $\hat{f}(\cdot, \theta)$ for $\theta \in \Theta$

Fitting algorithm: $\mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n)) = \hat{\theta}$

Loss function: $\ell(\hat{y}, y) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$

K folds: $I_1 \cup I_2 \cup \ldots \cup I_K = \{1, \ldots, n\}$



Seems simple

$e_1, \ldots, e_n$ are the individual errors from the holdout folds

naive approach: treat as a fresh i.i.d. holdout set

point estimate $\quad \widehat{\text{Err}} = \frac{1}{n} \sum_{i=1}^{n} e_i$  

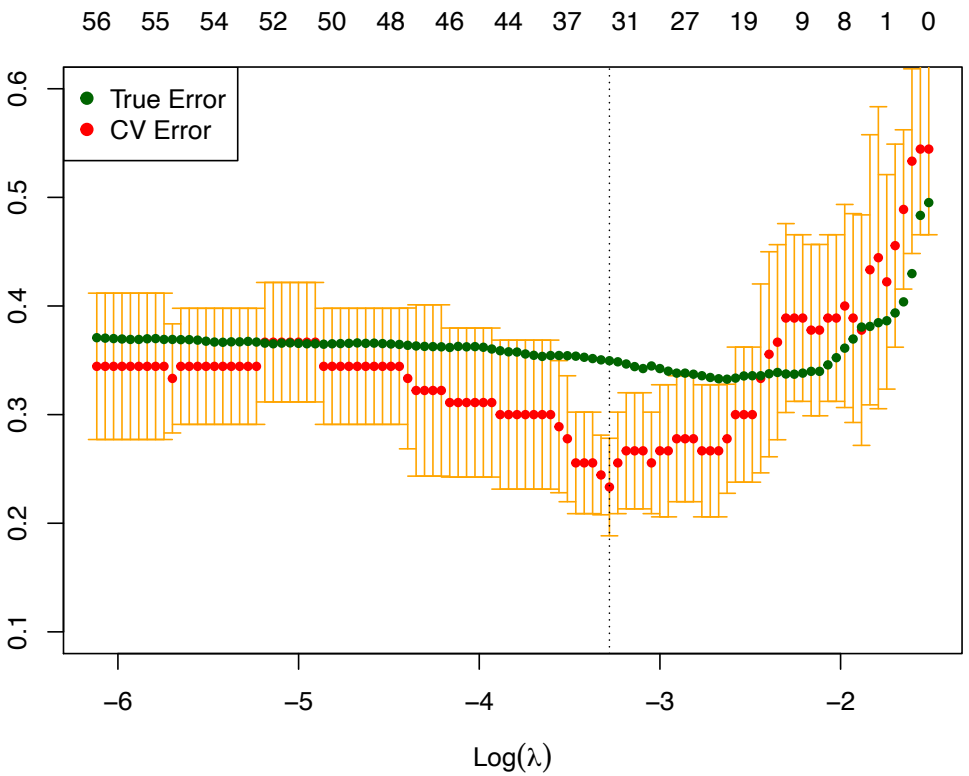
standard error estimate $\quad \widehat{\text{SE}} = \frac{1}{\sqrt{n}} \cdot \text{sd}(e_i)$ 

Can we take these at face value?

Part 1: Estimating the standard error

(to construct valid confidence intervals)

Naive 90% Confidence Interval



- sparse logistic model
- cv.glmnet path
- logistic lasso fit
- n = 90 samples
- p = 1000 features
- Bayes error: 22%

Does this have 90% coverage?

Frank Harrell raised doubts about $p \gg n$ Machine Learning results in a 2019 seminar at Stanford

(example due to Frank Harrell)

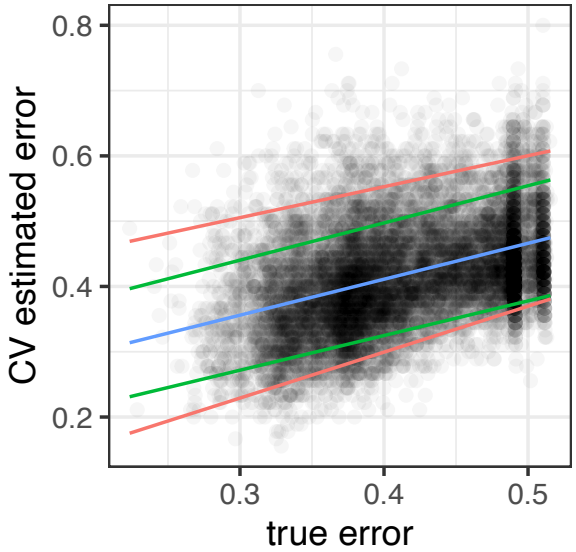
Naive confidence intervals can fail

$e_1, \ldots, e_n$ error when in holdout fold

$$\widehat{\text{Err}} = \frac{1}{n}\sum_{i=1}^{n} e_i \qquad \widehat{\text{SE}} = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2}$$

naive 90% confidence interval

$$\left(\widehat{\text{Err}} - 1.65 \cdot \widehat{\text{SE}}\,,\ \widehat{\text{Err}} + 1.65 \cdot \widehat{\text{SE}}\right)$$



quantity

- 5% & 95% quantiles
- CI endpoints
- CI midpoint

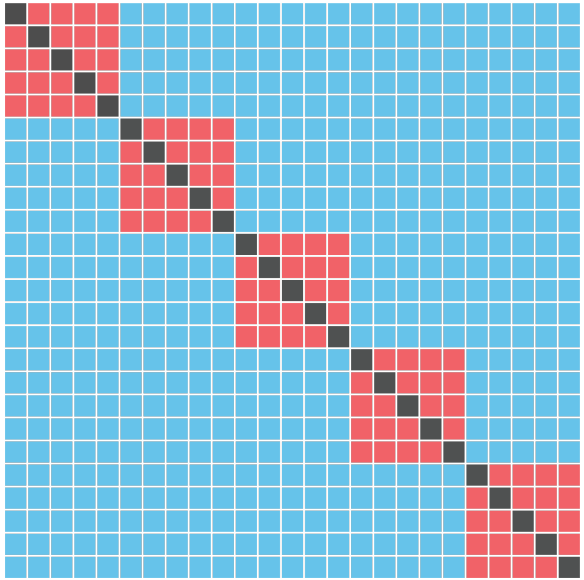
- sparse logistic model
- logistic lasso fit
- n = 90 samples
- p = 1000 features
- Bayes error: 20%

want 10% miscoverage, get 30% miscoverage

CIs need to be 70% bigger to be valid

CV errors have positive dependence (Bengio and Grandvalet '04)

covariance matrix of $e_1, \ldots, e_n$



$a_1 := \mathrm{Var}(e_1)$

$a_2 := \mathrm{Cov}(e_1, e_2)$ (same fold)

$a_3 := \mathrm{Cov}(e_1, e_n)$ (different folds)

Why dependence? **sample re-use**

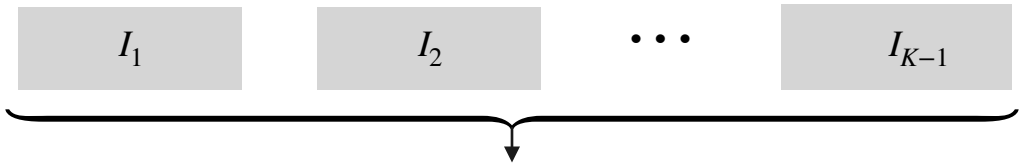
positive dependence → variance estimate too small!

$$\mathrm{Var}(\widehat{\mathrm{Err}}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} e_i\right) = \frac{1}{n}a_1 + \frac{n/K - 1}{n}a_2 + \frac{n - n/K}{n}a_3$$

sometimes $\gg \frac{1}{n}a_1$ ← (variance under independence)

but $\mathbb{E}\left[\widehat{\mathrm{SE}}^2\right] < \frac{1}{n}a_1$

Estimating the variance without assuming independence



$\widehat{\mathrm{Err}}^{(\mathrm{in})}$

estimate of prediction error

$e_i^{(\mathrm{out})}$

error on fresh holdout set

Lemma. (holdout MSE identity)

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

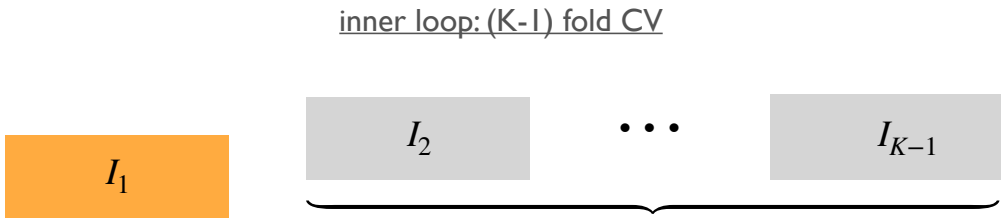
have unbiased plug-in estimate

have unbiased variance estimate: $\frac{1}{n/K-1}\sum\left(e_i^{(\mathrm{out})} - \bar{e}^{(\mathrm{out})}\right)^2$

repeat over many splits to estimate (a) and (b)

Key fact: by using expectations, immune to correlation issue

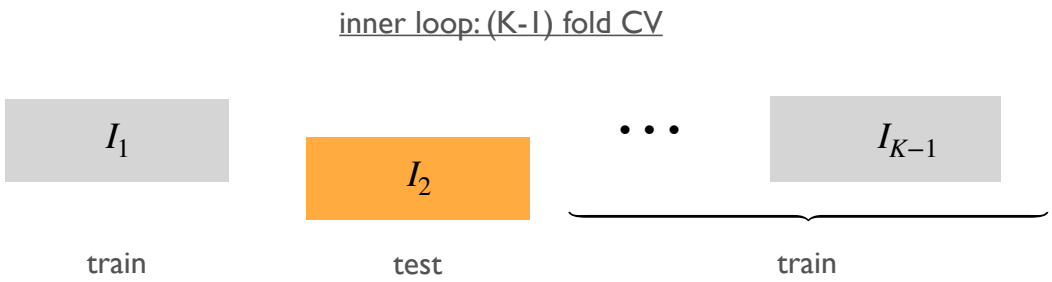
Nested cross-validation



$e_i^{(\text{out})}$

error on fresh holdout set

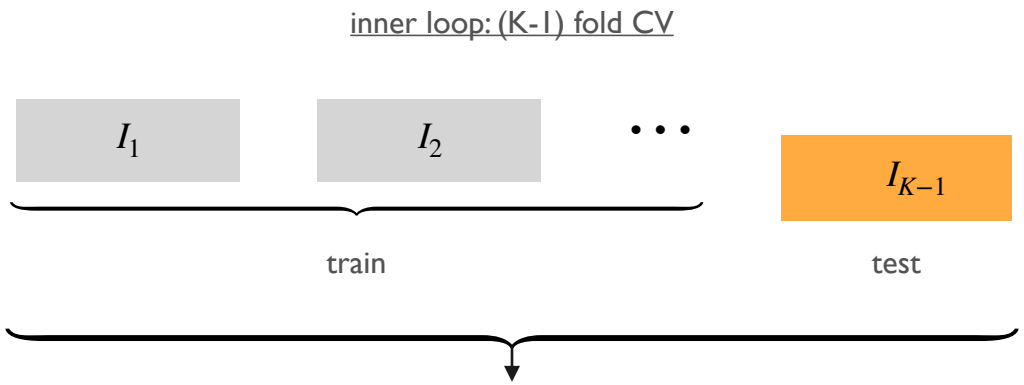
Nested cross-validation



$e_i^{(\text{out})}$

error on fresh holdout set

Nested cross-validation



$e_i^{(\text{out})}$

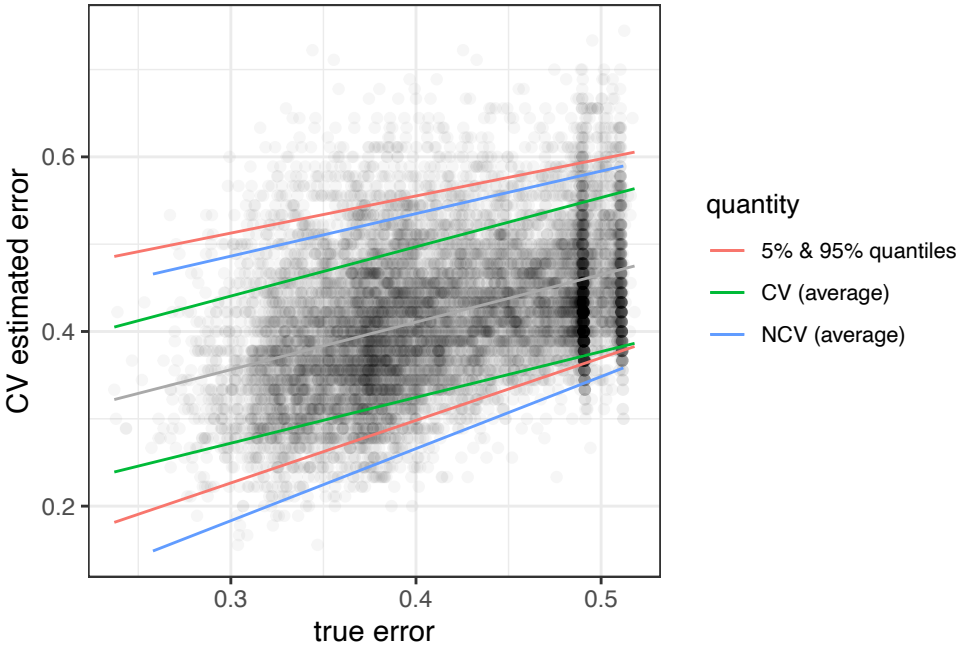
error on fresh holdout set

$$\widehat{\text{MSE}} := \frac{1}{\#\text{ splits}} \sum_{\text{random splits}} \left[\underbrace{\left(\widehat{\text{Err}}^{(\text{in})} - \bar{e}^{(\text{out})}\right)^2}_{\text{(a)}} - \underbrace{\frac{1}{n/K-1} \sum \left(e_i^{(\text{out})} - \bar{e}^{(\text{out})}\right)^2}_{\text{(b)}} \right]$$

unbiased for the true MSE (by last slide)

Example: high-dimensional logistic regression

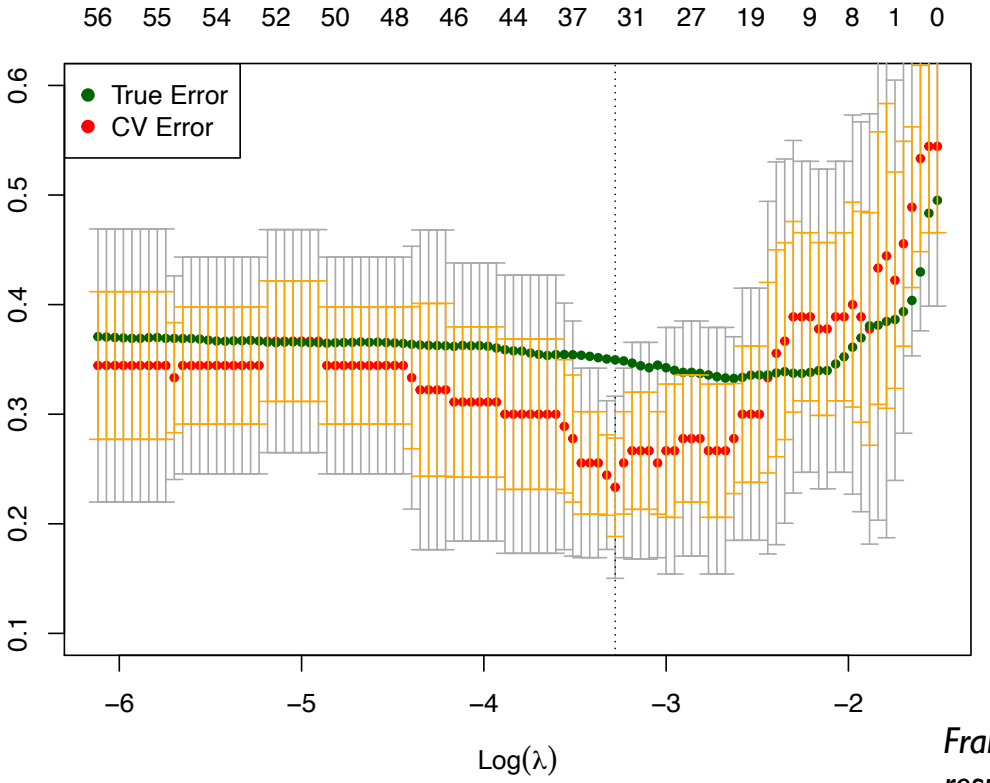
- sparse logistic model
- logistic lasso fit
- n = 90 samples
- p = 1000 features
- Bayes error: 20%



miscoverage (want 5%, 5%)

CV:	16% above, 12% below
nested CV:	6% above, 7% below

Corrected 90% Confidence Interval



- sparse logistic model
- cv.glmnet path
- logistic lasso fit
- n = 90 samples
- p = 1000 features
- Bayes error: 22%

Does this have 90% coverage?

Frank Harrell raised doubts about $p \gg n$ Machine Learning results in a 2019 seminar at Stanford

Recap

- CV confidence intervals can be much too small

sample re-use → positive dependence → standard error estimate too small

- Nested CV: estimated standard error without assuming independence
- Takes 100x more computation (1 minute in the high-d logistic regression)
- When does normal CV fail? Worse with n/p small and with less regularization

Software online!

Part 11: What is cross-validation estimating?

Targets of estimation

- Test error of my model with my training data

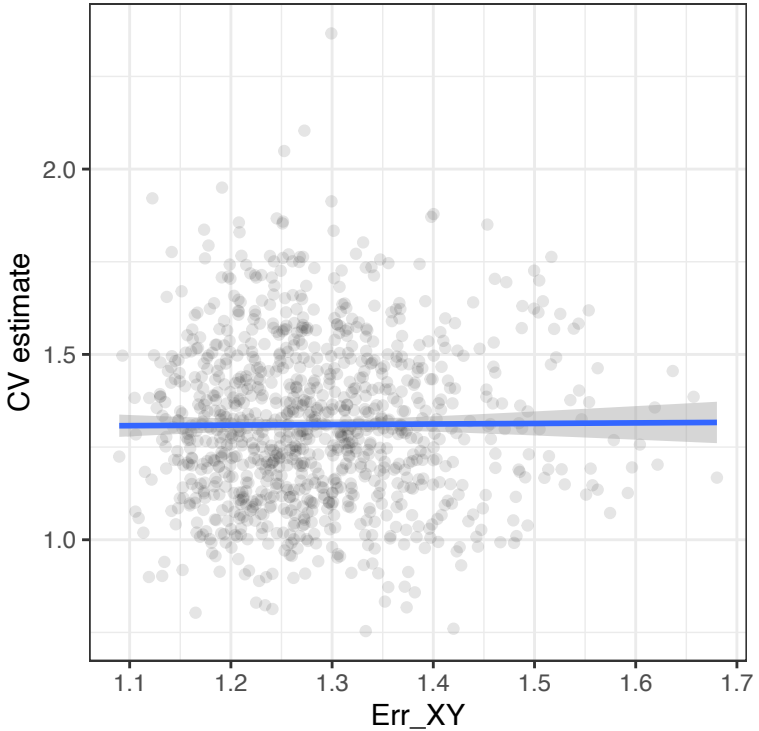
$$\mathrm{Err}_{XY} := \mathbb{E}\left[\ell(\hat{f}(X_{n+1}, \hat{\theta}), Y_{n+1}) \mid (X, Y)\right]$$

model fit on training data — test point — training data

- Average test error (across unseen training sets)

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

Can't estimate Err_{XY}??



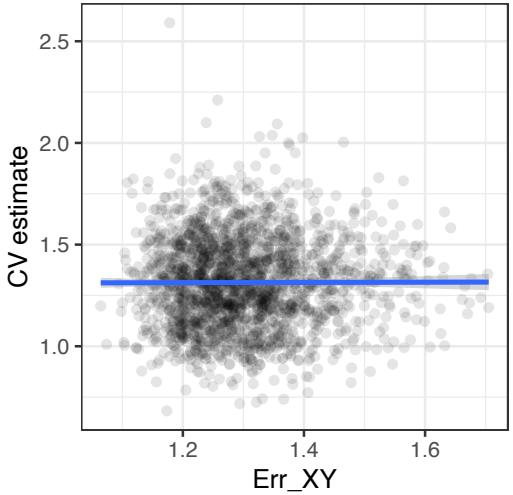
linear model, n = 100 samples, p = 20 features

correlation 0.01!

What happens in the linear model?

$$Y_i = X_i^\top \theta + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2) \qquad X_i \sim N(0, \Sigma)$$

(less important)



- linear model
- n = 100 samples
- p = 20 features
- **fixed design matrix**

correlation 0.00 ✓

Theorem. In the linear model, $\widehat{\mathrm{Err}} \perp \mathrm{Err}_{XY} \mid X$



Stephen's data: $(X_1, Y_1), \ldots, (X_n, Y_n)$ → $\mathrm{Err}_{XY}^{(\mathrm{Stephen})}$ and $\widehat{\mathrm{Err}}(\mathrm{Stephen})$



Rob's data: $(X_1, Y_1'), \ldots, (X_n, Y_n')$ → $\widehat{\mathrm{Err}}(\mathrm{Rob})$

(same X, fresh Y)

$$\left(\widehat{\mathrm{Err}}(\mathrm{Rob}) - \mathrm{Err}_{XY}^{(\mathrm{Stephen})}\right)^2 \overset{d}{=} \left(\widehat{\mathrm{Err}}(\mathrm{Stephen}) - \mathrm{Err}_{XY}^{(\mathrm{Stephen})}\right)^2$$

Rob's estimate is just as good for Stephen's test error!

What happens in the linear model?

CV independent of this gap! (previous slide)

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

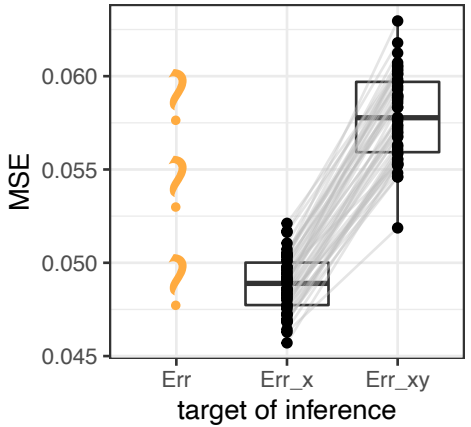
average error — error of model with my training data

Corollary. In the linear model,

$$\mathbb{E}\left(\widehat{\text{Err}} - \text{Err}_X\right)^2 < \mathbb{E}\left(\widehat{\text{Err}} - \text{Err}_{XY}\right)^2$$

(Rao-Blackwell)

better at estimating Err_X !



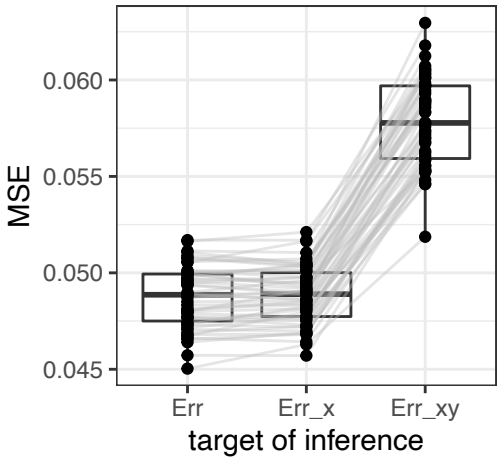
Without assuming Normality, one can prove:

Corollary. $\operatorname{cor}(\widehat{\mathrm{Err}}, \mathrm{Err}_{XY}) \to 0$

Corollary.

$$\mathbb{E}\left(\widehat{\mathrm{Err}} - \mathrm{Err}\right)^2 \approx \mathbb{E}\left(\widehat{\mathrm{Err}} - \mathrm{Err}_X\right)^2 < \mathbb{E}\left(\widehat{\mathrm{Err}} - \mathrm{Err}_{XY}\right)^2$$

CV better at estimating average error!



Other estimators for prediction accuracy

Analysis relies on invariance property:

Definition. An estimator of prediction error $\widehat{\mathrm{Err}} : (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{R}$ is *linearly invariant* if

$$\widehat{\mathrm{Err}}\left((x_1, y_1), \ldots, (x_n, y_n)\right) = \widehat{\mathrm{Err}}\left((x_1, y_1 + \kappa^\top x_1), \ldots, (x_n, y_n + \kappa^\top x_n)\right)$$

for any $\kappa \in \mathbb{R}^p$.

- cross-validation
- bootstrap (out-of-bag, .632, …)
- Mallow's Cp / AIC
- RCp
- data splitting with refitting

(with OLS fitting)

same conclusion: all are better at estimating average error

Part III: Random forests and OOB inference

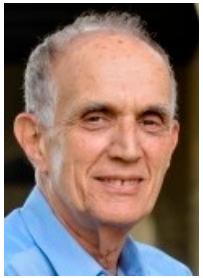
$\widehat{\text{Err}}_{\text{OOB}} \approx \widehat{\text{Err}}_{\text{LOO}}$

- Random forests fit randomized trees to B bootstrapped versions of the n training points, and average their predictions.
- For each observation i in the training set, OOB averages the predictions from all trees grown to bootstrap samples that did *not* include observation i (about two thirds of the trees), to create a leave-one-out or LOO prediction, and a LOO error Q_i for that observation.
- These errors are averaged for all n observations, and this gives the $\widehat{\text{Err}}_{\text{OOB}} = \text{Ave}(Q_i)$ error estimate.
- If the number B of trees is large, then modulo a small n versus $n-1$ correction, this is exactly the leave-one-out cross-validation estimate of error.

- Can we efficiently produce confidence intervals for Err or Err_X based on $\widehat{\text{Err}}_{\text{OOB}}$?
- Can we efficiently produce standard error estimates for random forest predictions $\hat{\mu}(x)$?

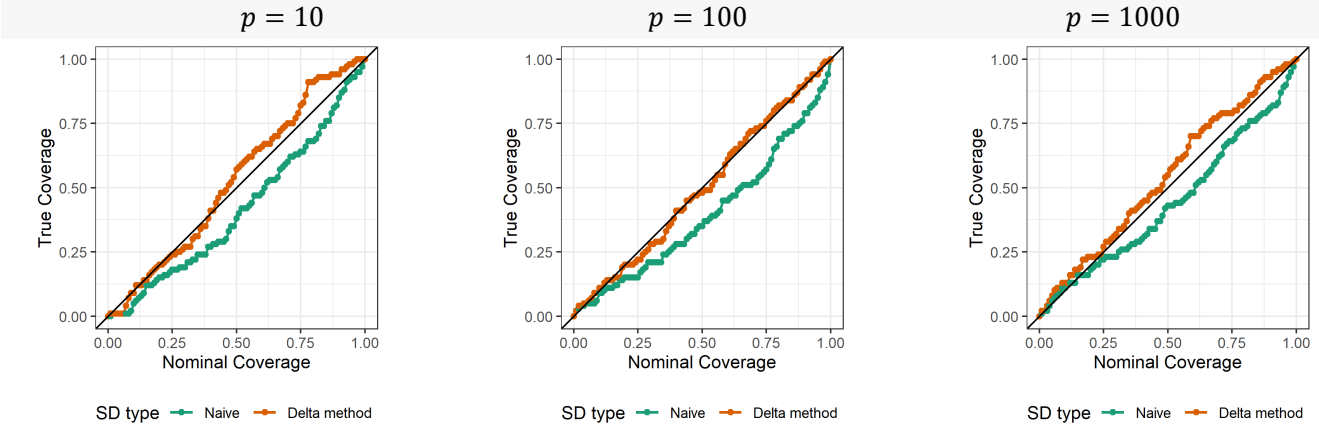
Standard error estimate for $\widehat{\mathrm{Err}}_{\mathrm{OOB}}$





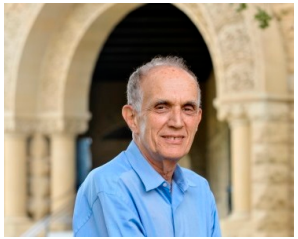
Work with us three and Ph.D. student Samyak Rajanala, using clever tricks from *Jackknife-after-bootstrap standard errors and influence functions* (Efron 1992, JRSSB)

- Naive standard error estimate of $\mathrm{Ave}(Q_i)$ ignores the correlation in the Q_i
- $\widehat{\mathrm{Err}}_{\mathrm{OOB}} = S(\widehat{F})$, a smooth function of the EDF. The idea is to use the infinitesimal jackknife to estimate the variance of $\widehat{\mathrm{Err}}_{\mathrm{OOB}}$
- *This can be done without any additional bootstrap samples!*



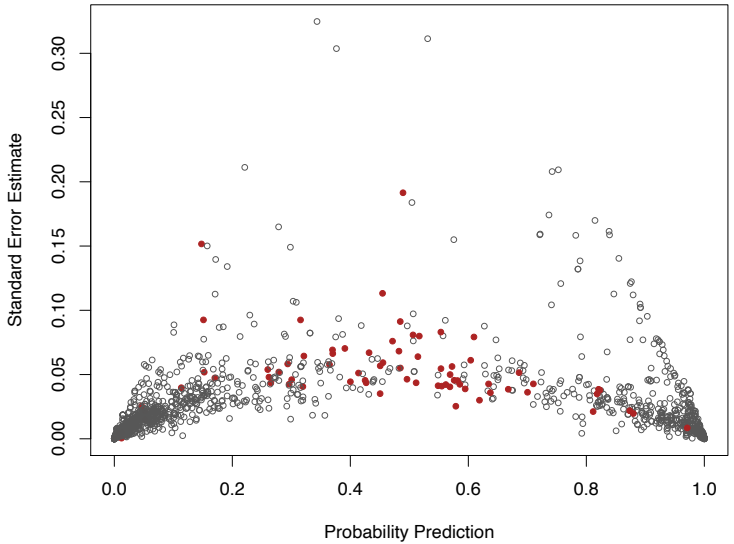
Standard Errors for random forest predictions $\hat{\mu}(x)$





Stefan Wager, Brad Efron, TH: *Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife (JMLR 2014)*

- Similar tricks can produce infinitesimal jackknife standard error estimates for each prediction $\hat{\mu}(x)$ made by random forest, without additional computations.



- Email SPAM data
- points are SE estimates
- red points are incorrect predictions

Conclusion

Part I: confidence intervals

- CV confidence intervals can be much too small

  sample re-use → positive dependence → standard error estimate too small

- worse with n/p small and with less regularization
- nested CV: estimate standard error without assuming independence

Part II: point estimates (linear model with no regularization)

- CV is estimating average prediction error (not prediction error of my fitted model)
- same conclusion for bootstrap, Cp, …

Part III: Random Forests and OOB inference

- Jackknife estimate of $\mathrm{SE}(\widehat{\mathrm{Err}}_{\mathrm{OOB}})$
- Jackknife estimate of $\mathrm{SE}(\hat{\mu}(x))$

} without additional resampling!

Special thanks to Frank Harrell and Brad Efron

Thank you!