Uniform consistency of cross-validation estimators for high-dimensional ridge regression

Pratik Patil, Yuting Wei, Alessandro Rinaldo, Ryan J. Tibshirani

Carnegie Mellon University

AISTATS 2021

Main punchline

- Standard regression with n data pairs $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$
- Given a tuning parameter λ, recall that ridge regression solves

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2/n + \lambda \|\beta\|_2^2$$

- Choice λ crucially affects the performance of the fitted estimator

Key question: how to select λ based on observed data in high dimensions

We show: under proportional asymptotics as $n \to \infty$, $p/n \to \gamma \in (0, \infty)$, the leave-one-out and generalized cross-validation almost surely,

1. converge to out-of-sample prediction error uniformly in λ;
2. pick optimal λ for prediction error, including when $\lambda = 0$ or negative

Outline

Problem setup

Main results

Proof intuitions

High-dimensional ridge regression

- Let $X \in \mathbb{R}^{n \times p}$ denote feature matrix, $y \in \mathbb{R}^n$ denote response vector
- Let $\widehat{\beta}_\lambda := \arg\min_{\beta \in \mathbb{R}^p} \; \|y - X\beta\|_2^2/n + \lambda\|\beta\|_2^2$ denote ridge estimate
  - if $\lambda > 0$, problem convex in β and has an explicit solution:

$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{-1} X^T y/n$$

  - for any $\lambda \in \mathbb{R}$, extend using Moore-Penrose inverse:

$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{+} X^T y/n$$

  - when $\lambda = 0$, this reduces to least squares sol with minimum ℓ_2 norm; in particular, when $\text{rank}(X) = n \le p$, the solution interpolates data, i.e. $X\widehat{\beta} = y$, and has minimum ℓ_2 norm among all interpolators

Prediction error and cross validation

- We measure the performance of fitted models $\widehat{\beta}_\lambda$ by their expected squared out-of-sample prediction error defined as

$$\mathrm{err}(\lambda) := \mathbb{E}_{x_0,y_0}\big[(y_0 - x_0^T\widehat{\beta}_\lambda)^2 \mid X, y\big],$$

  where (x_0, y_0) is test pair sampled from same training distribution
  - random (conditional on observed data X and y)
  - unknown (depends on characteristics of data generating distribution)
- Several estimators of prediction error:
  - k-fold cross validation (large bias when $k = 5$ or even when $k = 10$)
  - Generalized cross validation
  - Stein unbiased error estimate (in-sample prediction error)

We study the case when $k = n$ also called leave-one-out cross-validation, and generalized cross-validation

Leave-one-out and generalized cross-validation

- Leave-one-out cross-validation (LOOCV):
  - for every i, train on all data except (x_i, y_i), call the estimate $\widehat{\beta}_\lambda^{-i}$
  - compute test error on the i^{th} point and take average

$$\text{loo}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T\widehat{\beta}_\lambda^{-i}\right)^2$$

$$\overset{\text{(shortcut)}}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1-[L_\lambda]_{ii}}\right)^2$$

where $L_\lambda = X(X^TX/n + \lambda I_p)^+X^T/n$ is the ridge smoothing matrix

- Generalized cross-validation (GCV)
  - same as leave-one-out shortcut but a single re-weighting

$$\text{gcv}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1-\text{tr}[L_\lambda]/n}\right)^2$$

- When $\widehat{\beta}_\lambda$ is an interpolator, i.e. $L_\lambda = I_n$, both estimates are "0/0"; we then define the estimates as their respective limits as $\lambda \to 0$

Goals of the paper

There are two main questions that we answer in this paper:

1. How do $\mathrm{gcv}(\lambda)$ and $\mathrm{loo}(\lambda)$ compare to $\mathrm{err}(\lambda)$ as functions of λ?
2. How do $\mathrm{err}(\widehat{\lambda}_I^{\mathrm{gcv}})$ and $\mathrm{err}(\widehat{\lambda}_I^{\mathrm{loo}})$ compare to $\mathrm{err}(\lambda_I^\star)$ where $\lambda_I^\star$ denotes the optimal oracle ride tuning parameter

$$\lambda_I^\star = \operatorname*{arg\,min}_{\lambda \in I \subseteq \mathbb{R}} \mathrm{err}(\lambda),$$

and $\widehat{\lambda}_I^{\mathrm{gcv}}$ and $\widehat{\lambda}_I^{\mathrm{loo}}$ denote the corresponding tuning parameters that minimize GCV and LOOCV over an interval I?

Outline

Summary of main results

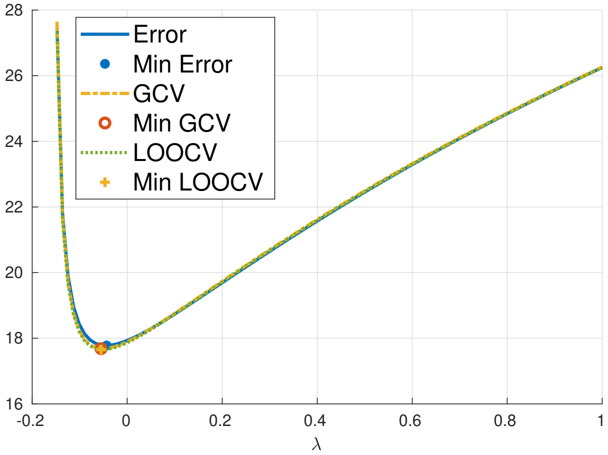
Under i.i.d. sampling with

- a well-specified model $y = x^T \beta_0 + \varepsilon$ where ε is independent of x
- decomposable features $x = \Sigma^{1/2} z$ where z contains i.i.d. entries
- certain bnd moment and norm cond. on ε and z, and β_0 and Σ, resp.

as $n \to \infty$ and $p/n \to \gamma \in (0, \infty)$, we show

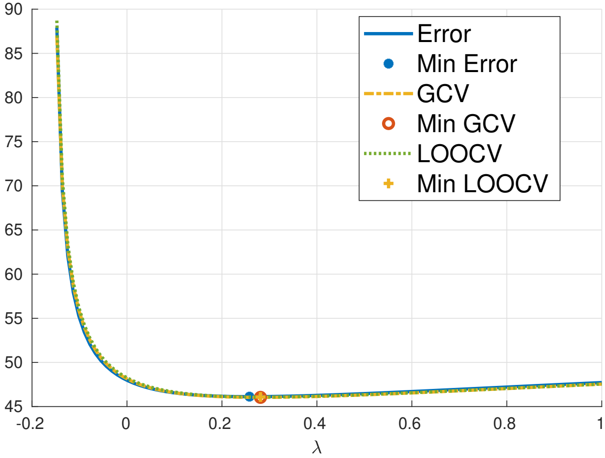
1. GCV pointwise convergence
   - $\mathrm{gcv}(\lambda)$ converges to $\mathrm{err}(\lambda)$ pointwise in λ
2. GCV uniform convergences
   - convergence holds uniformly over compact intervals of λ including 0
3. LOOCV convergences
   - the analogous results hold for $\mathrm{loo}(\lambda)$ by relating it to $\mathrm{gcv}(\lambda)$
4. Optimal tuned prediction errors
   - both $\mathrm{err}(\widehat{\lambda}_I^{\mathrm{gcv}})$ and $\mathrm{err}(\widehat{\lambda}_I^{\mathrm{loo}})$ converge to $\mathrm{err}(\lambda_I^\star)$

Numerical illustration (negative optimal regularization)



- Overparametrized regime ($p = 12000$, $n = 6000$)
- Autoregressive Σ
- β_0 aligned with the top eigendirection of Σ

Numerical illustration (positive optimal regularization)



- Overparametrized regime ($p = 12000$, $n = 6000$)
- Autoregressive Σ
- β_0 aligned with the bottom eigendirection of Σ

Outline

GCV versus prediction error: two key proof steps

Step 1: bias and variance decompositions of prediction error and GCV

Let $\widehat{\Sigma} := X^T X/n$ denote the sample covariance matrix.

- limiting bias-like components:
  - prediction error

$$\mathrm{err}_b(\lambda) := \lambda^2 \beta_0^T (\widehat{\Sigma} + \lambda I)^+ \Sigma (\widehat{\Sigma} + \lambda I)^+ \beta_0$$

  - gcv

$$\mathrm{gcv}_b(\lambda) := \frac{\lambda^2 \beta_0^T (\widehat{\Sigma} + \lambda I)^+ \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^+ \beta_0}{\left(1 - \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n\right)^2}$$

- limiting variance-like components:
  - prediction error

$$\mathrm{err}_v(\lambda) := \sigma^2 \left[1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n\right] - \sigma^2\, \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma (\widehat{\Sigma} + \lambda I_p)^+\right]/n$$

  - gcv

$$\mathrm{gcv}_v(\lambda) := \sigma^2 \left[\frac{1}{1 - \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n}\right] - \frac{\sigma^2\, \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma} (\widehat{\Sigma} + \lambda I_p)^+\right]/n}{\left(1 - \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n\right)^2}$$

GCV versus prediction error: two key proof steps

Step 2: bias and variance equivalences for prediction error and GCV

- bias component equivalence:

$$\lambda^2 \beta_0^T (\widehat{\Sigma} + \lambda I)^+ \Sigma (\widehat{\Sigma} + \lambda I)^+ \beta_0 - \frac{\lambda^2 \beta_0^T (\widehat{\Sigma} + \lambda I)^+ \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^+ \beta_0}{\left(1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n\right)^2} \xrightarrow{\text{a.s.}} 0$$

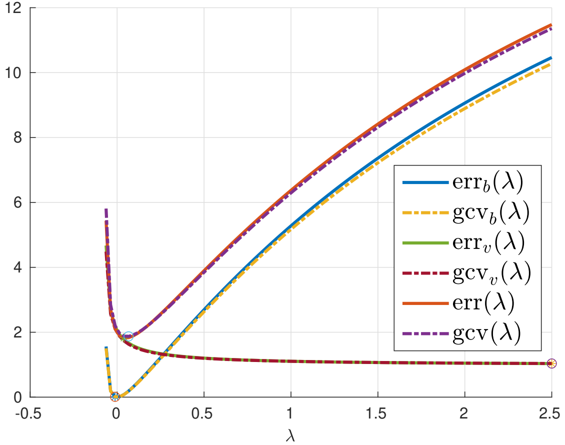
- variance component equivalences:

$$\sigma^2 \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma (\widehat{\Sigma} + \lambda I_p)^+\right]/n - \frac{\sigma^2 \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma} (\widehat{\Sigma} + \lambda I_p)^+\right]/n}{\left(1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n\right)^2} \xrightarrow{\text{a.s.}} 0$$

$$\sigma^2 \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n - \frac{\sigma^2 \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n}{1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n} \xrightarrow{\text{a.s.}} 0$$

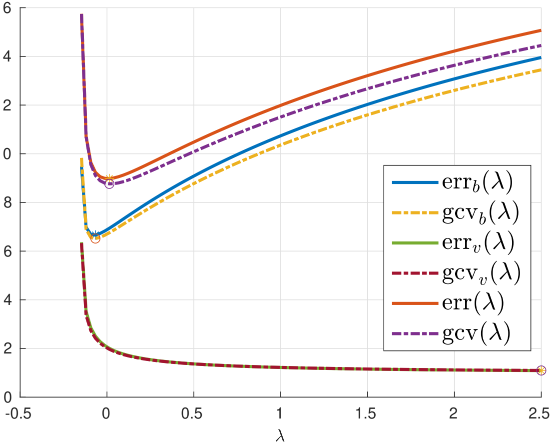
Main message: the GCV denominator proves to be the right correction for for the excess optimism in the biased GCV numerator of training error

Bias and variance equivalence numerical illustration



- Underparametrized ($n = 6000$, $p = 3000$)
- Bias minimized at $\lambda = 0$ and variance decreases as λ increases
- Optimal λ always positive is underparametrized regime!

Bias and variance equivalence numerical illustration



- Overparametrized ($p = 12000$, $n = 6000$)
- Bias no longer minimized $\lambda = 0$ and variance still decreasing in λ
- Optimal λ may be negative in overparametrized regime!

Discussion and future directions

This work shows GCV and LOOCV uniformly track squared out-of-sample prediction error for ridge regression under proportional asymptotics.

Main tool:

$$(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma \asymp \frac{(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}}{1 - \mathrm{tr}[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}]/n}$$

where for any two sequences of matrices A_p and B_p, $A_p \asymp B_p$ is used to mean $\mathrm{tr}[C_p(A_p - B_p)] \xrightarrow{\text{a.s.}} 0$ for any deterministic seq of matrices C_p of bnd trace norm

Going beyond . . .

- Equivalences for general functionals of out-of-sample distributions
- Equivalences for general estimators
- Finite sample analysis and rates of convergence

⋮

Thanks for listening!

Questions/comments/thoughts?