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

Pratik Patil Yuting Wei Alessandro Rinaldo Ryan J. Tibshirani

Carnegie Mellon University

High-dimensional ridge regression

- Consider standard regression with feature matrix $X \in \mathbb{R}^{n \times p}$ and response vector $y \in \mathbb{R}^n$
- Given a tuning parameter λ, recall that ridge estimator $\widehat{\beta}_\lambda$ solves the optimization problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \|y - X\beta\|_2^2/n + \lambda\|\beta\|_2^2$$

  - for any $\lambda > 0$, the problem is convex in β and has an explicit closed-form solution given by
$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{-1} X^T y/n$$
  - for any $\lambda \in \mathbb{R}$, we can extend the solution using the Moore-Penrose inverse as
$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{+} X^T y/n$$
  - when $\lambda = 0$, this reduces to least squares solution with minimum ℓ_2 norm; in particular, when $\text{rank}(X) = n \le p$, the solution also interpolates data, i.e. $X\widehat{\beta} = y$, and has minimum ℓ_2 norm among all interpolators
- In general, the choice of λ crucially affects the performance of the fitted model

Key question: how to select λ based on observed data in high dimensions (p much larger than n)

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

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

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

We study the case when $k = n$ also called leave-one-out cross-validation and its approximation generalized cross-validation and provide theoretical guarantees for tuning λ

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} data point and take average

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

  where $L_\lambda = X(X^T X/n + \lambda I_p)^{+} X^T/n$ 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 in $0/0$ form; in this case, we 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 $\text{gcv}(\lambda)$ and $\text{loo}(\lambda)$ compare to $\text{err}(\lambda)$ as functions of λ?
2. How do $\text{err}(\widehat{\lambda}_I^{\text{gcv}})$ and $\text{err}(\widehat{\lambda}_I^{\text{loo}})$ compare to $\text{err}(\lambda_I^\star)$ where $\lambda_I^\star$ denotes the optimal oracle ride tuning parameter
$$\lambda_I^\star = \underset{\lambda \in I \subseteq \mathbb{R}}{\arg\min}\ \text{err}(\lambda),$$
and $\widehat{\lambda}_I^{\text{gcv}}$ and $\widehat{\lambda}_I^{\text{loo}}$ denote the corresponding tuning parameters that minimize GCV and LOOCV over an interval I?

Summary of contributions

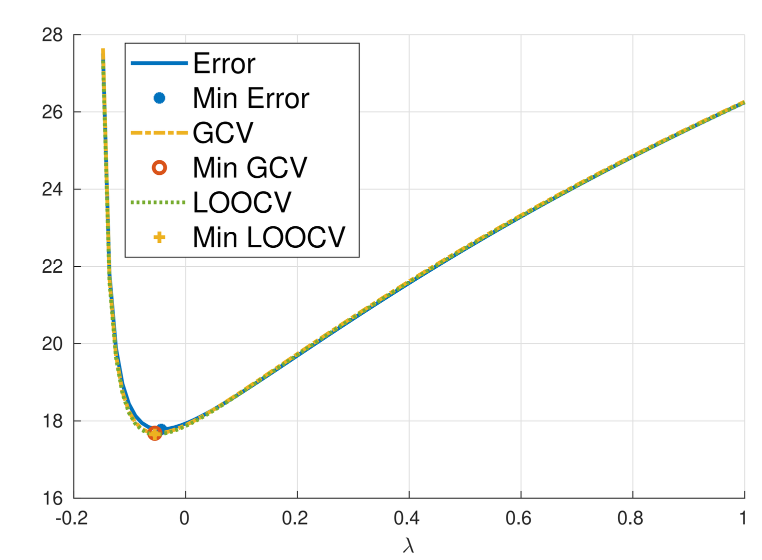
Under i.i.d. sampling with

- 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;
- bounded moments of order $(4 + \eta)$ of ε and z for some $\eta > 0$;
- bounded norm and eigenvalue conditions on β_0 and Σ, respectively

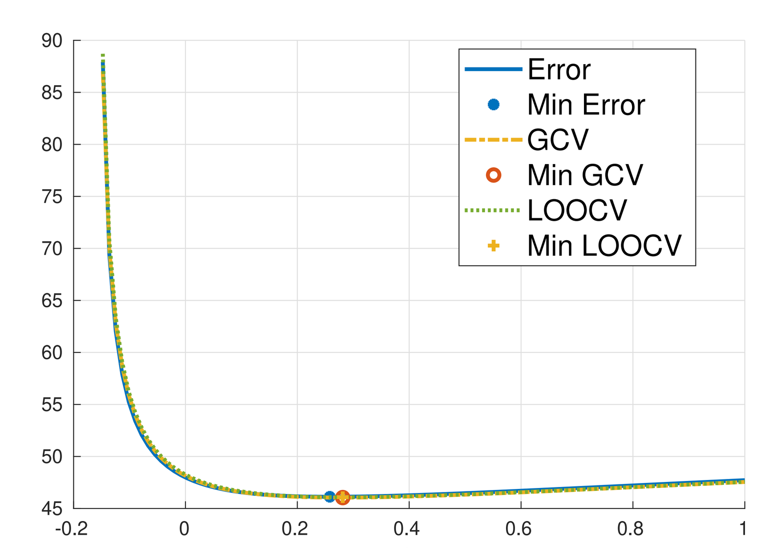
as $n \to \infty$ and $p/n \to \gamma \in (0, \infty)$, we show the following:

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

Numerical illustration



- Overparametrized regime
- Autoregressive Σ
- β_0 aligned with top eigendirection of Σ



- Overparametrized regime
- Autoregressive Σ
- β_0 aligned with bottom eigendirection of Σ

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
$$\text{err}_b(\lambda) := \lambda^2 \beta_0^T (\widehat{\Sigma} + \lambda I)^{+} \Sigma (\widehat{\Sigma} + \lambda I)^{+} \beta_0$$
  - GCV
$$\text{gcv}_b(\lambda) := \frac{\lambda^2 \beta_0^T (\widehat{\Sigma} + \lambda I)^{+} \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{+} \beta_0}{\big(1 - \text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}]/n\big)^2}$$
- limiting variance-like components:
  - prediction error
$$\text{err}_v(\lambda) := \sigma^2\Big[1 + \text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma]/n\Big] - \sigma^2\,\text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma(\widehat{\Sigma} + \lambda I_p)^{+}]/n$$
  - GCV
$$\text{gcv}_v(\lambda) := \sigma^2\left[\frac{1}{1 - \text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}]/n}\right] - \frac{\sigma^2\,\text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}]/n}{\big(1 - \text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}]/n\big)^2}$$

GCV versus prediction error: two key proof steps

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

- bias components equivalence:
$$\lambda^2 \beta_0^T (\widehat{\Sigma} + \lambda I)^{+} \Sigma (\widehat{\Sigma} + \lambda I)^{+} \beta_0 - \frac{\lambda^2 \beta_0^T (\widehat{\Sigma} + \lambda I)^{+} \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{+} \beta_0}{\big(1 - \text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}]/n\big)^2} \xrightarrow{\text{a.s.}} 0$$
- variance components equivalences:
$$\sigma^2\,\text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma(\widehat{\Sigma} + \lambda I_p)^{+}]/n - \frac{\sigma^2\,\text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}]/n}{\big(1 - \text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}]/n\big)^2} \xrightarrow{\text{a.s.}} 0$$
$$\sigma^2\,\text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma]/n - \frac{\sigma^2\,\text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}]/n}{1 - \text{tr}\,[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}]/n} \xrightarrow{\text{a.s.}} 0$$

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

Discussion and future work

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

Main tool:

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

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

Going beyond ...

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

⋮