Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression

Pratik Patil, Alessandro Rinaldo, Ryan J. Tibshirani

Carnegie Mellon University

AISTATS 2022

Motivation and punchline of the paper

- Given $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}, 1 \le i \le n\}$, let $\widehat{\beta}_\lambda$ be ridge estimator:

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

- The out-of-sample error of $\widehat{\beta}_\lambda$ is $y_0 - x_0^\top \widehat{\beta}_\lambda$ for a test point (x_0, y_0)
- Estimating out-of-sample error well is crucial for model assessment
- Prior work shows leave-out-out and generalized cross-validation consistently estimate the expected squared error $\mathbb{E}[(y_0 - x_0^\top \widehat{\beta}_\lambda)^2 \mid \mathcal{D}]$

Key question: can we reliably estimate the entire out-of-sample error distribution and its linear and non-linear functionals in high dimensions?

We show, that under proportional asymptotics, almost surely:

1. the empirical distributions of re-weighted in-sample errors from leave-one-out and generalized cross-validation converge weakly to the out-of-sample error distribution, even when $\lambda = 0$
2. the plug-in estimators of these empirical distributions consistent for a broad class of linear and non-linear functionals of error distribution

Outline

High-dimensional ridge regression

- Let $X \in \mathbb{R}^{n \times p}$ denote feature matrix, $y \in \mathbb{R}^n$ denote response vector
- Let $\widehat{\beta}_\lambda \in \mathbb{R}^p$ denote the ridge estimator at regularization level λ:

$$\widehat{\beta}_\lambda := \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2/n + \lambda\|\beta\|_2^2$$

  - if $\lambda > 0$, the problem is 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 the solution 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 $\text{rank}(X) = n \leq p$, the solution interpolates data, i.e. $X\widehat{\beta} = y$, and has minimum ℓ_2 norm among all interpolators

Out-of-sample error distribution and its functionals

- Let P_λ denote distribution of out-of-sample error of $\widehat{\beta}_\lambda$:

$$P_\lambda = \mathcal{L}\big(y_0 - x_0^\top \widehat{\beta}_\lambda \mid X, y\big),$$

  where (x_0, y_0) is sampled indep from the same training distribution
  - a random distribution (conditional on observed data X and y)
- Let ψ denote a functional such that $P \mapsto \psi(P) \in \mathbb{R}$:
  - Linear functional:

$$\psi(P_\lambda) = \int t(z)\, dP_\lambda(z) = \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big],$$

    where $t : \mathbb{R} \to \mathbb{R}$ is an error function (e.g., squared or absolute error)
  - Nonlinear functional:

$$\psi(P_\lambda) = \mathrm{Quantile}(P_\lambda; \tau) = \inf\{z : F_\lambda(z) \geq \tau\},$$

    where F_λ denotes the cumulative distribution function of P_λ

We construct estimators of P_λ and $\psi(P_\lambda)$ by suitably extending leave-one-out cross-validation and generalized cross-validation procedures.

Standard 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

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

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

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

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

- Standard LOOCV and GCV are consistent for the expected squared out-of-sample prediction error

Proposed estimators

Natural estimators for P_λ and $\psi(P_\lambda)$ building off from GCV and LOOCV.

- Empirical distributions of the GCV, LOO re-weighted errors:

$$\widehat{P}_\lambda^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n}\delta\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n}\delta\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)$$

- 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 $\lambda \to 0$:

$$\widehat{P}_0^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n}\delta\left(\frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}\right) \quad \text{and} \quad \widehat{P}_0^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n}\delta\left(\frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}}\right)$$

- Plug-in GCV and LOO estimators:

$$\widehat{\psi}_\lambda^{\mathrm{gcv}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}}) \quad \text{and} \quad \widehat{\psi}_\lambda^{\mathrm{loo}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}})$$

Outline

Problem setup

Distribution estimation

Functional estimation

Distribution estimation

Under i.i.d. sampling of (x_i, y_i), $i = 1, \ldots, n$ with

1. feature x_i decomposable into $x_i = \Sigma^{1/2} z_i$ where z_i contains i.i.d. entries with mean 0, variance 1 and finite 4+ moment, and max and min eigenvalues of Σ uniformly away from 0 and ∞,
2. response y_i with bounded 4+ moment,

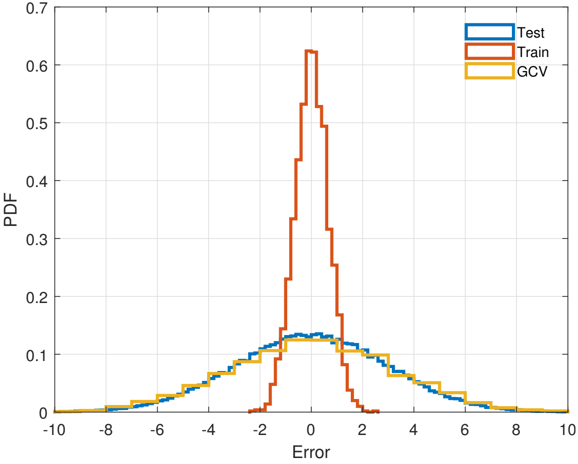
as $n, p \to \infty$ such that $p/n \to \gamma \in (0, \infty)$, almost surely

$$\widehat{P}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{d}} P_\lambda \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{d}} P_\lambda.$$

Remarks:

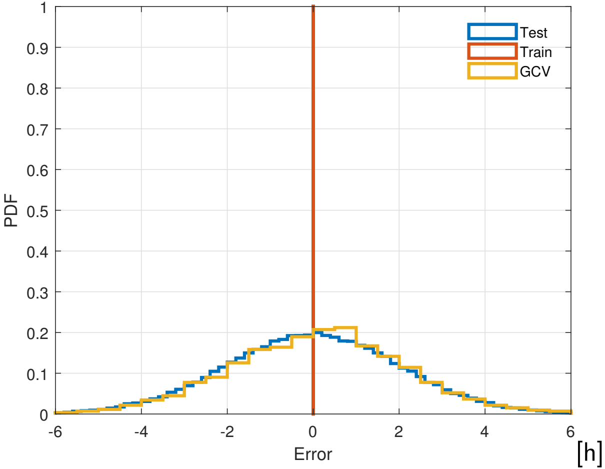
- Almost sure convergence with respet to the training data
- The regression function does not need to be linear in x
- Amazingly, this results also holds when $\lambda = 0$ (min-norm estimator)

Distribution estimation: illustration ($p < n$)



- $n = 2500$, $p = 2000$, $p/n = 0.8$
- $\lambda = 0$, i.e., least squares

Distribution estimation: illustration ($p > n$)



- $n = 2500$, $p = 5000$, $p/n = 2$
- $\lambda = 0$, i.e., the min-norm estimator, zero in-sample errors

Outline

Problem setup

Distribution estimation

Functional estimation

Linear functional estimation (pointwise)

- Let T_λ be a linear functional of the out-of-sample error distribution:

$$T_\lambda = \mathbb{E}\left[t(y_0 - x_0^T \widehat{\beta}_\lambda) \mid X, y\right]$$

- Let $\widehat{T}_\lambda^{\mathrm{gcv}}$ and $\widehat{T}_\lambda^{\mathrm{loo}}$ be plug-in estimators from GCV and LOOCV:

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

For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are continuous,
2. have quadratic growth, i.e., there exist constats $a, b, c > 0$ such that $|t(z)| \leq az^2 + b|z| + c$ for any $z \in \mathbb{R}$,

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, almost surely

$$\widehat{T}_\lambda^{\mathrm{gcv}} \to T_\lambda \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} \to T_\lambda.$$

Linar functional estimation (uniform)

For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are differentiable,
2. have derivative with linear growth rate, i.e., there exist constants $g, h > 0$ such that $|t'(z)| \le g|z| + h$ for any $z \in \mathbb{R}$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$ for any compact set Λ,

$$\sup_{\lambda \in \Lambda} |\widehat{T}_\lambda^{\mathrm{gcv}} - T_\lambda| \to 0 \quad \text{and} \quad \sup_{\lambda \in \Lambda} |\widehat{T}_\lambda^{\mathrm{loo}} - T_\lambda| \to 0.$$

Remarks:

- Special case of $t(r) = r^2$ exploits bias-variance decomposition
- No bias-variance decomposition for general error functions and result requires a different proof technique via leave-one-out arguments
- Using uniformity arguments, the result can be extended for non-linear variational functionals (see paper for more details)

Discussion and future directions

Take-away from this work: empirical distributions of GCV and LOOCV track out-of-sample error distribution and a wide class of its functionals for ridge regression under proportional asymptotics framework

Key relation that we exploit:

$$y_i - x_i^\top \widehat{\beta}_{-i,\lambda} = \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} \approx \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}$$

$$y_i - x_i^\top \widehat{\beta}_{-i,0} = \frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}} \approx \frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}$$

Going beyond ...

- Equivalences for ridge variants and other smoothers
- Finite sample analysis and rates of convergence

⋮

Thanks for listening!

Questions/comments/thoughts?