The lasso, persistence, and cross-validation

Darren Homrighausen[†] Daniel J. McDonald[‡]

†Department of Statistics, Colorado State University, Fort Collins
‡Department of Statistics, Indiana University, Bloomington

Suppose we have data

$$\mathcal{D}_n = \{(Y_1, X_1^\top,), \ldots, (Y_n, X_n^\top)\}$$

$X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ are the features

$Y_i \in \mathbb{R}$ are the responses

Use $\mathcal{D}_n$ to choose a function $\widehat{f}$ that can predict Y from X

The regression function is the best predictor

$$m(X) = \mathbb{E}[Y|X] = \operatorname*{argmin}_f \mathbb{E}\left[(Y - f(X))^2\right]$$

Idea: Start with linear approximation of $m(X)$.

Choose $\beta \in \mathbb{R}^{p+1}$, form

$$\widehat{f}(X) = X_1\beta_1 + \ldots + X_p\beta_p = X^\top\beta$$

Important: This does not assume that m is linear in X!

We need to find a good estimator of β.

ℓ_1-regularized regression

Called lasso or basis pursuit

The estimator satisfies

$$\widehat{\beta}_t = \operatorname*{argmin}_\beta ||\mathbb{Y} - \mathbb{X}\beta||_2^2 \text{ subject to } ||\beta||_1 \leq t$$

Alternatively:

$$\widehat{\beta}_\lambda = \operatorname*{argmin}_\beta ||\mathbb{Y} - \mathbb{X}\beta||_2^2 + \lambda||\beta||_1$$

Properties

Suppose $m(X) = X^\top\beta$:

- If $\lambda = o(n)$, then $\widehat{\beta}_\lambda \xrightarrow{\text{a.s.}} \beta$
- If $\frac{\lambda}{n} \to a \in (0, \infty)$, then $\widehat{\beta}_\lambda \not\to \beta$ in general
- If $\frac{\lambda}{n} \to \infty$, then $\widehat{\beta}_\lambda \xrightarrow{\text{a.s.}} 0$

What if $m(X)$ not linear? What if $p \gg n$?

Define $Z^\top = (Y, X^\top)$ to be a new observation (same distribution)

(Predictive) risk

$$R(\beta) = \mathbb{E}_Z\left[(Y - X^\top\beta)^2\right]$$

Oracle estimator

$$\beta_t^* = \operatorname*{argmin}_{\{\beta:||\beta||_1 \leq t\}} R(\beta)$$

Excess risk

$$\mathcal{E}(\widehat{\beta}_t, \beta_t^*) = R(\widehat{\beta}_t) - R(\beta_t^*)$$

A procedure is persistent if

$$\mathcal{E}(\widehat{\beta}_t, \beta_t^*) \xrightarrow{\text{P}} 0$$

The best (oracle) linear model

If $t^4 = o\left(\frac{n}{\log n}\right)$, then $\widehat{\beta}_t$ is persistent relative to β_t^*

$\widehat{\beta}_t$ is not necessarily persistent if $t^4 \notin o\left(\frac{n}{\log n}\right)$

What if choose $t = \widehat{t}$ using $\mathcal{D}_n$?

Create set of validation sets $V_n = \{v_1, \ldots, v_{K_n}\}$

$\widehat{\beta}_t^{(v)}$ lasso estimator ignoring observations in $v \subset \{1, \ldots, n\}$

The cross-validation estimator of the risk is

$$\widehat{R}_{V_n}(t) = \widehat{R}_{V_n}\left(\widehat{\beta}_t^{(v_1)}, \ldots, \widehat{\beta}_t^{(v_{K_n})}\right) := \frac{1}{K_n}\sum_{v \in V_n}\frac{1}{|v|}\sum_{r \in v}\left(Y_r - X_r^\top\widehat{\beta}_t^{(v)}\right)^2$$

Define

$$\widehat{t} := \operatorname*{argmin}_{t \in T_n} \widehat{R}_{V_n}(t)$$

In practice, need to specify $T_n = [0, t_{\max}]$

If $t_{\max}$ is too small, we may exclude good solutions

By definition, $\widehat{\beta}_t \in \{\beta : ||\beta||_1 \leq t\}$

This constraint is only binding if

$$t < \min_{\eta \in \mathcal{K}} ||\widehat{\beta}^0 + \eta||_1 =: t_0,$$

where

$\widehat{\beta}^0 := (\mathbb{X}^\top\mathbb{X})^\dagger\mathbb{X}^\top\mathbb{Y}$ is a least squares solution

$\mathcal{K} := \{a : \mathbb{X}a = 0\}$ is the null space of $\mathbb{X}$

Define $t_{\max} := ||\widehat{\beta}^0||_1$

Conditions

C1. $\mathbb{E}\left[||\widehat{\beta}^0||_1^4\right] = o(t_n^4)$

C2. For any cross-validation procedure V_n, there exists a constant c_n such that for all $v \neq v' \in V_n$

- $|v| \geq c_n$
- $v \cap v' = \emptyset$

C3. Let $Z^\top = (Y, X^\top) \sim F_n$. Then, $(F_n)_{n \geq 1}$ is such that $\exists C < \infty$ for all n where

$$\mathbb{E}_{F_n}\left[\max_{0 \leq j,k \leq p}(Z_jZ_k - \mathbb{E}_{F_n}Z_jZ_k)^2\right] \leq C$$

Results

Theorem: Suppose C1–C3 and that $p_n = n^\alpha$, $\alpha > 0$. Then, for any $\delta > 0$,

$$P(\mathcal{E}(\widehat{\beta}_{\widehat{t}}, \beta_{t_n}^*) > \delta) = o\left(t_n^2\sqrt{\frac{\log n}{c_n}}\right).$$

- $c_n \asymp n$ for K-fold cross-validation
- leave-one-out cross-validation has $c_n = 1$

Properties of t_n

The faster $t_n \to \infty \ldots$

- the less restrictive condition C1 becomes
- $R_n(\beta_{t_n}^*)$ shrinks faster
- But if $t_n^4 = \Omega(n/\log n)$, $\widehat{\beta}_{t_n}$ may not be persistent, let alone $\widehat{\beta}_{\widehat{t}}$

Can $\mathbb{E}\left[||\widehat{\beta}^0||_1^4\right] = o(t_n^4)$ if $t_n^4 = o\left(\frac{n}{\log n}\right)$?

Examples:

Suppose $Y = m(X) + \epsilon$, $m(X)$ bounded, $\mathbb{E}[\epsilon^4] < \infty$

- $X_i \in \mathbb{R}^p$ are i.i.d sub-Gaussian with independent components
- Fixed design, kernel regression satisfying $h^{-1}\phi(1/h) \to 0$ as $h \to \infty$
- Orthogonal basis regression

Future work: Similar results for lasso-type estimators

- G a partition of $\{1, \ldots, p\}$
- $\mathcal{G}_u := \{\beta : \sum_{g \in G}\sqrt{|g|}||\beta_g||_2 \leq u\}$

Theorem: Suppose

- $\mathbb{E}\left[\left(\sum_{g \in G}||\widehat{\beta}_g^0||_2\right)^4\right] = o(u_n^4)$
- $p_n = n^\alpha$ for some $\alpha > 0$
- $\max_{g \in G}|g| = a_n$
- Conditions C2 and C3

Then, for any $\delta > 0$,

$$P_{F_n}\left(\mathcal{E}\left(\widehat{\beta}_{\widehat{u}}, \beta_{u_n}^*\right) > \delta\right) = o\left(a_nu_n^2\sqrt{\frac{\log n}{c_n}}\right).$$

† stat.colostate.edu/~darrenho ‡ mypage.iu.edu/~dajmcdon † darrenho@stat.colostate.edu ‡ dajmcdon@indiana.edu