The lasso, persistence, and cross-validation

Daniel J. McDonald

Department of Statistics

Indiana University

http://www.stat.cmu.edu/∼danielmc

Joint work with:

Darren Homrighausen

Colorado State University

The question:

All the results about lasso are for oracle tuning parameter. What happens if you choose it using the data?

The answer: YES!

The question:

All the results about lasso are for oracle tuning parameter. What happens if you choose it using the data?

The answer: YES!

The motivation:

The motivation:

The Setup

Suppose we have data

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

where

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

We use $\mathcal{D}_n$ to find a function $\widehat{f}$ that can predict Y from X.

The regression function is the best possible function

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

Parameterizing this Relationship

A good start is to find the best linear approximation of $m(X)$.

A linear predictor specifies a $\beta \in \mathbb{R}^p$ and forms

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

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

We need to find a good estimator of β.

The lasso

ℓ_1-REGULARIZED REGRESSION

Of course, for large p, small n, we need to regularize

Known as

- 'lasso'
- 'basis pursuit'

The estimator satisfies

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

Alternatively:

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

Some properties

Suppose $m(X)$ IS linear, p is small[1]...

- If $\lambda = o(n)$, then $\widehat{\beta}_\lambda \xrightarrow{\text{a.s.}} \beta$
- If $\frac{\lambda}{n} \to a \in (0, \infty)$, then $\widehat{\beta}_\lambda \nrightarrow \beta$ in general
- If $\frac{\lambda}{n} \to \infty$, then $\widehat{\beta}_\lambda \xrightarrow{\text{a.s.}} 0$
- If $\lambda = o(\sqrt{n})$, then $\sqrt{n}||\widehat{\beta}_\lambda - \beta|| \xrightarrow{\text{d}} A$, A is a random variable.

[1] Knight and Fu (2000), Chatterjee and Lahiri (2011)

Different properties

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

Define $\mathcal{Z}^\top = (\mathcal{Y}, \mathcal{X}^\top)$ to be a new observation [same distribution].

We define the (predictive) risk to be

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

Define the oracle estimator

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

The excess risk is

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

Different properties (cont.)

A procedure is persistent (relative to the oracle) if

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

Then[2]

- If $t^4 = o\left(\frac{n}{\log n}\right)$, $\widehat{\beta}_t$ is persistent relative to β_t^*
- $\widehat{\beta}_t$ is not necessarily persistent if $t^4 \notin o\left(\frac{n}{\log n}\right)$

[2] Greenshtein and Ritov (2004)

You've got data. . .

What t to use?

Methods for choosing t

The tuning parameter can be selected by

- unbiased risk estimation using degrees of freedom
- using an adapted Bayesian information criterion

However…

Many papers recommend cross-validation [3, 4, 7, 8, 9, 10, 11]

[It is also the default method in the `R` package `glmnet`. See Zou, Hastie, and Tibshirani (2010)]

Define

- $V_n = \{v_1, \ldots, v_{K_n}\}$ to be a set of validation sets
- $\widehat{\beta}_t^{(v)}$ lasso estimator computed on observations not in $v \subset \{1, \ldots, n\}$

The cross-validation estimator of the risk is

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

Define

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

Define

- $V_n = \{v_1, \ldots, v_{K_n}\}$ to be a set of validation sets
- $\widehat{\beta}_t^{(v)}$ lasso estimator computed on observations not in $v \subset \{1, \ldots, n\}$

The cross-validation estimator of the risk is

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

Define

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

Choosing T_n

In practice, the optimization set $T_n = [0, t_{\max}]$ needs to be specified

However, if $t_{\max}$ is too small, good solutions might be excluded

What is too small?

Choosing T_n

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

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 $\mathbb{X}$

If $t \ge t_0$, then $\widehat{\beta}_t$ is ‘equal to’ $\widehat{\beta}^0$

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

Does cross-validation work?

Prevailing heuristic:

> *"Regarding the choice of the regularization parameter, we typically use [the tuning parameter chosen by] cross-validation. 'Luckily', empirical and some theoretical indications support [good performance]..."*
>
> — *Peter Bühlmann's comments to Tibshirani (2011).*

What does theory have to say?

Stability (or lack thereoff)

Sparsity inducing algorithms, such as lasso, are not (uniformly) algorithmically stable

Algorithmic stability is sufficient, but not necessary, for persistence[4]

[4] Xu and Mannor (2008) and Bousquet and Elisseeff (2002)

Model selection and cross-validation

There is a close connection between lasso and model selection
[e.g. the LARS algorithm]

For model selection[5]...

- Leave-one-out cross-validation is inconsistent
- If $c_n/n \to 1$ and $n - c_n \to \infty$, then cross-validation is consistent
  [[c_n is the size of the smallest held-out set]]

Very restrictive: asymptotically, all the data is used for validation

[5] Shao (1993)

Results: Cross-validation does work

Conditions

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

C2. Held-out sets contain at least c_n observations, don't overlap.

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

$$\mathbb{E}_{F_n} \max_{0 \leq j,k \leq p} (\mathcal{Z}_j \mathcal{Z}_k - \mathbb{E}_{F_n} \mathcal{Z}_j \mathcal{Z}_k)^2 \leq C$$

Results (cont.)

Theorem

Assume C1–C3 and that $p_n = n^\alpha$ for some $\alpha > 0$.

Then, for any $\delta > 0$,

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

Some remarks

- $c_n \asymp n$ for K-fold cross-validation
- leave-one-out cross-validation has $c_n = 1$
- $\mathcal{E}(\widehat{\beta}_{\widehat{t}}, \beta_{t_n})$ CAN be negative (don't care)

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 grows as fast or faster than $\left(\frac{n}{\log n}\right)^{1/4}$, then $\widehat{\beta}_{t_n}$ is not necessarily persistent

Can $\mathbb{E}\left[||\widehat{\beta}^0||_1^4\right] = o(t_n^4)$ if $t_n = o\left(\left(\frac{n}{\log n}\right)^{1/4}\right)$?

Yes. . .

When it works...

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

Example 1:

- $X_i \in \mathbb{R}^p$ i.i.d sub-Gaussian with independent components

Example 2:

- Fixed design $e_i = i/n$
- $\mathbb{X}_{ij} = h^{-1}\phi(|e_j - e_i|/h)$
- ϕ satisfies $h^{-1}\phi(1/h) \to 0$ as $h \to \infty$

Example 3:

- Orthogonal basis regression

Future work

Show similar results for lasso-type estimators, such as group lasso

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

Theorem

Suppose

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

Then, for any $\delta > 0$,

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

References I

[1] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378, 2000.

[2] A. Chatterjee and SN Lahiri. Strong consistency of lasso estimators. *Sankhya A-Mathematical Statistics and Probability*, 73(1):55–78, 2011.

[3] E. Greenshtein and Y.A. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

[4] H. Zou, T. Hastie, and R. Tibshirani. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.

[5] R.J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *Annals of Statistics*, 40:1198–1232, 2012.

[6] H. Wang and C. Leng. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048, 2007.

[7] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Verlag, 2009.

[9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[10] R. Tibshirani. Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

[11] S. van de Geer and J. Lederer. The lasso, correlated design, and improved oracle inequalities, 2011.

[12] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[13] H. Xu, S. Mannor, and C. Caramanis. Sparse algorithms are not stable: A no-free-lunch theorem. In *46th Annual Allerton Conference on Communication, Control, and Computing*, pages 1299–1303. IEEE, 2008.

[14] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[15] J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88:486–494, 1993.

[16] M. Rudelson and R. Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009.