STAT 154: Penalized methods in regression

Noureddine El Karoui

Please do not redistribute without consent of instructor

Slides have not been very carefully proof-read

Department of Statistics
UC, Berkeley

September 24, 2013

General outlook

We are moving from

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2$$

to

$$\widehat{\beta}_\lambda = \text{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda P(\beta)$$

Examples:

1. $P(\beta) = \|\beta\|_0$, i.e number of non-zero entries in β
2. $P(\beta) = \|\beta\|_2^2 = \sum_{i=1}^{p} \beta_i^2$. Leads to ridge regression
3. $P(\beta) = \|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|$. LASSO.
4. $P(\beta) = \|\beta\|_q^2 = \sum_{i=1}^{p} |\beta_i|^q$, $q \geq 1$.

Role of Penalty

How to think about the penalty? What is its role:

- Makes the problem have a unique solution
- Forces a certain structure on $\widehat{\beta}$: sparsity. Bayesian point of view
- Stabilizes/regularizes the solution: ridge
- Automatically picks a subset for us: lasso

Potential downsides:

- Numerical cost: how easy is it to solve the penalized problem?
- Stability of solution? Bias in solution? (Elastic net)
- Interpretation

Interpretation for LASSO

$$\widehat{\beta}_\lambda = \text{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \text{ . or}$$

$$\widehat{\beta}_\lambda = \begin{cases} \text{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 \\ \text{subject to } \|\beta\|_1 \leq t(\lambda) \end{cases}$$

What is LASSO doing? Hard to quantify though see picture.

1. A simple example: the case of orthogonal predictors.
2. Utility: cyclical coordinate descent. Very fast numerically
3. Bias in predictors.

Variants

1. Elastic net: compromise between ridge and lasso
2. Adaptive lasso: $P(\beta) = \sum w_i |\beta_i|$. Example $w_i = |\widehat{\beta}_{LS,i}|^{-\nu}$, $\nu > 0$. Attempt at solving the non-convex problems that arise with ℓ_q norm penalization with $q < 1$.
3. Non-negative garrote: requires $\beta_i \geq 0$ for all i.
4. Why not look at $\|Y - X\beta_2\| + \lambda \|\beta\|_1$?

Historically, preceded LASSO. One motivation: improve prediction.
Recall:

$$\widehat{\beta}_{\text{ridge}} = \text{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

Find $\widehat{\beta}$:

$$\widehat{\beta}_{\text{ridge}} = (X'X + \lambda \mathrm{Id}_p)^{-1} X'Y \; .$$

So if $Y = X\beta_0 + \epsilon$,

$$\widehat{\beta}_{\text{ridge}} = (X'X + \lambda \mathrm{Id}_p)^{-1} X'X\beta_0 + (X'X + \lambda \mathrm{Id}_p)^{-1} X'\epsilon \; .$$

Bias-variance tradeoff?

AIC, BIC and all that

Recall that our "old problem" was that

$$\|Y - X\widehat{\beta}_{LS}\|_2^2 \text{ IS NOT A GOOD MEASURE OF EPE}$$

Add penalties to sequentially built models to get a good measure of EPE.

- AIC: Optimize $\frac{1}{n}\|Y - X\widehat{\beta}_{LS,d}\|_2^2 + 2\frac{d}{n}\sigma_\epsilon^2$
- BIC: Optimize $\frac{1}{n}\|Y - X\widehat{\beta}_{LS,d}\|_2^2 + d\frac{\log n}{n}\sigma_\epsilon^2$

Link with above methods?

Other ideas

Why not change the representation of the data?

Idea: why not change the basis in which the predictors are given?
In particular, we saw that the LS estimator had covariance essentially $(X'X)^{-1}$. Can we improve that?
Idea of **Principal Components Regression** See sections 3.5.1 in the book.