Statistical Machine Learning — 7 — Not Low Assumption Lasso — Sept. 30, 2014

Lecturer: Prof. Homrighausen — Scribe: Miranda Fix

1 Low Assumption Lasso

Recall that when we do **not** assume a linear model, an additive stochastic component, or almost uncorrelated design, we get that with probability at least $1 - C\delta^{-1}t^2\sqrt{\frac{log(p)}{n}}$, $R(\hat{\beta}_t) \leq R(\beta_t^*) + \delta$.

2 Not Low Assumption Lasso

There are several theoretical results under stronger assumptions. Let us assume that the truth is linear: $Y = \mathbb{X}\beta^* + \epsilon$. The basis for derivations related to prediction error is the so-called Basic Inequality:

Lemma 2.1. ***The Basic Inequality***

$$||\mathbb{X}(\hat{\beta} - \beta^*||_2^2/n + \lambda||\hat{\beta}||_1 \leq 2\epsilon^\top \mathbb{X}(\hat{\beta} - \beta^*)/n + \lambda||\beta^*||_1 \tag{1}$$

Note that this comes from simply rewriting the inequality $||Y - \mathbb{X}\hat{\beta}||_2^2/n + \lambda||\hat{\beta}||_1 \leq ||Y - \mathbb{X}\beta^*||_2^2/n + \lambda||\beta^*||_1$.

Since the $2\epsilon^\top \mathbb{X}(\hat{\beta} - \beta^*)/n$ term involves the measurement error, we call it the **empirical process** part of the problem. The $\lambda||\beta^*||_1$ is then the deterministic part. Our goal is to choose λ so that the deterministic part outweighs the empirical process part.

We can bound the empirical process by

$$2\epsilon^\top \mathbb{X}(\hat{\beta} - \beta^*)/n \leq (\max_{1 \leq j \leq p} 2|\epsilon^\top x_j|/n)||\hat{\beta} - \beta^*||_1 \tag{2}$$

Let us introduce the set $\mathcal{T} = \left\{\max_{1 \leq j \leq p} 2|\epsilon^\top x_j|/n \leq \lambda_0\right\}$. Assume that $\mathbb{X}$ is standardized: $\hat{\Sigma} = \mathbb{X}^\top \mathbb{X}/n$ has 1's on the diagonal. We can show that for an appropriate value of λ_0, the set $\mathcal{T}$ has large probability. In particular, if $\lambda_0 = 2\sigma\sqrt{(t^2 + 2logp)/n}$ and $\epsilon \sim N(0, \sigma I)$, then

$$\mathbb{P}(\mathcal{T}) \geq 1 - 2e^{-t^2/2} \tag{3}$$

(See Homework 2 solutions for a proof under the assumption of Gaussian errors. Note that normality is not essential, we just need some degree of control over the tails.)

Now let us introduce another set, the active set $S_* = \left\{j : \beta_j^* \neq 0\right\}$. We can show that on $\mathcal{T}$ with $\lambda \geq 2\lambda_0$,

$$2||\mathbb{X}(\hat{\beta} - \beta^*)||_2^2/n + \lambda||\hat{\beta}_{S_*^c}||_1 \leq 3\lambda||\hat{\beta}_{S_*} - \beta_{S_*}^*||_1 \tag{4}$$

(See Homework 2 solutions for a proof of this inequality.)

Compatibility condition

We will require a certain compatibility of ℓ_1-norms and ℓ_2-norms for the theory to progress. In equation (4) above, we would like to incorporate the $||\hat{\beta}_{S_*} - \beta_{S_*}^*||_1$ term on the right-hand side into the $2||\mathbb{X}(\hat{\beta} - \beta^*)||_2^2/n$

term on the left-hand side. By the Cauchy-Schwartz inequality, we can replace the ℓ_1-norm with the ℓ_2-norm using $||\hat{\beta}_{S_*} - \beta^*_{S_*}||_1 \leq \sqrt{s_*}||\hat{\beta}_{S_*} - \beta^*_{S_*}||_2$, where $s* = |S_*|$ is the size of the active set.

This brings us to the ℓ_2 world, where we now need to relate $||\hat{\beta}_{S_*} - \beta^*_{S_*}||_2$ to $||\mathbb{X}(\hat{\beta} - \beta^*)||_2$ Recalling from equation (4) that on $\mathcal{T}$, $||\hat{\beta}_{S^c_*}||_1 \leq 3||\hat{\beta}_{S_*} - \beta^*_{S_*}||_1$, we restrict ourselves to only those β that satisfy this.

This gives us the **compatibility condition:** We say that the compatibility condition is met for the set S_* if for some $\phi_0 > 0$, and for all β satisfying $||\beta_{S^c_*}||_1 \leq 3||\beta_{S_*}||_1$, it holds that

$$||\beta_{S_*}||_1^2 \leq (\beta^\top \hat{\Sigma} \beta) s_* / \phi_0^2 \tag{5}$$

When does this compatibility condition hold? Notice that if we replace $||\beta_{S_*}||_1^2$ in equation (5) by its upper bound $s_*||\beta_{S_*}||_2^2$, the condition is similar to a condition on the smallest eigenvalue of $\hat{\Sigma}$. Due to the restriction $||\beta_{S^c_*}||_1 \leq 3||\beta_{S_*}||_1$, it is weaker than imposing non-zero eigenvalues. If the cardinality s_* were known (not possible in practice), it would be sufficient to check the inequalities for all sets S with cardinality s_*. This is know as the **restricted eigenvalue assumption**.

Another related notion is that of **restricted isometry**. A restricted isometry doesn't deform angles too much over relevant parts of the space, i.e. $\exists \delta > 0$ such that for all interesting β

$$(1 - \delta)||\beta||_2^2 \leq ||\mathbb{X}\beta||_2^2 \leq (1 + \delta)||\beta||_2^2.$$

This basically says the covariates are nearly uncorrelated, which is a very strong assumption. Dr. Wasserman argues that restricted isometry is not a reasonable assumption for regression. For more on this topic see http://normaldeviate.wordpress.com/2012/08/07/rip-rip-restricted-isometry-property-rest-in-peace/

Suppose that the compatibility condition holds for S_* with compatibility constant ϕ_*. Then on $\mathcal{T}$ and for $\lambda \geq 2\lambda_0$ we get the oracle inequality

$$||\mathbb{X}(\hat{\beta} - \beta_*)||_2^2/n + \lambda||\hat{\beta} - \beta_*||_1 \leq 4\lambda^2 \frac{S_*}{\phi_*} \tag{6}$$

(See Homework 2 solutions for a proof.)

Equation (6) clearly implies that $||\mathbb{X}(\hat{\beta} - \beta_*)||_2^2 \leq 4n\lambda^2 \frac{S_*}{\phi_*}$ and $\lambda||\hat{\beta} - \beta_*||_1 \leq 4\lambda^2 \frac{S_*}{\phi_*}$.

Combining equations (3) and (6) we can form a PAC bound: if $\lambda \geq 4\sigma\sqrt{(t^2 + 2logp)/n}$, then with probability at least $1 - 2e^{-t^2/2}$, $||\mathbb{X}(\hat{\beta} - \beta_*)||_2^2/n + \lambda||\hat{\beta} - \beta_*||_1 \leq 4\lambda^2 \frac{S_*}{\phi_*}$.

We could also define $\lambda := 4\hat{\sigma}\sqrt{(t^2 + 2logp)/n}$, where $\hat{\sigma}^2$ is an estimator of σ^2. Then with probability at least $1 - 2e^{-t^2/2} - \mathbb{P}(\hat{\sigma} \leq \sigma)$ we have $||\mathbb{X}(\hat{\beta} - \beta_*)||_2^2/n + \lambda||\hat{\beta} - \beta_*||_1 \leq 4\lambda^2 \frac{S_*}{\phi_*}$.

3 HARNESS

The basic idea is that we split our data into two parts: D_1 and D_2. We use D_1 to obtain $\hat{\beta}$ with any variable selection method, then we use D_2 to do distribution free inference. There isn't much to add to the lecture slides for this section, but see Homework 2 solutions for an example of the HARNESS in practice.

References

[1] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer, 2011.