Lecture 5 – Ridge and Lasso Regression II

Martin Spindler

2016-04-26

Lasso Regression

$$\hat{\beta}(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} \left(||Y - X\beta||_2^2/n + \lambda ||\beta||_1 \right) (*)$$

$||Y - X\beta||_2^2 = \sum_{i=1}^n (Y_i - (X\beta)_i)^2$, $||\beta||_1 = \sum_{j=1}^p |\beta_j|$, $\lambda \geq 0$
penalisation parameter
$(*)$ is equivalent to

$$\hat{\beta}_{primal}(R) = \arg\min_{\beta \in \mathbb{R}^p} \left(||Y - X\beta||_2^2/n \right)$$

such that $||\beta||_1 \leq R$ with a one-to-one relation between R and λ.
This optimization problem is a convex problem (and hence efficient computation is possible.)

Lasso Regression

Key assumption: **sparsity**
The number of variables p can grow with the sample size and even be larger n, but the number of non-zero coefficients s is required to be smaller than n (but may also grow with the sample size).
Notation:

- β^0 true vector with components $\beta_j^0, j = 1, \ldots, p$
- $S_0 = \{j : \beta_j^0 \neq 0, j = 1, \ldots, p\}$, $s = |S|$
- $\hat{S} = \{j : \hat{\beta}_j \neq 0, j = 1, \ldots, p\}$

A glimpse on the theory

- convergence in prediction norm
- convergence in ℓ_p norm
- variable screening
- variable selection

Theory | Convergence in prediction norm

- Conditions:
- Restricted eigenvalue condition / compatability condition
- no condition on the non-zero coefficients
- Result:
$$||X(\hat{\beta} - \beta_0)||_2^2/n = O_P(s\log(p)/n)$$
- Interpretation

Theroy | Convergence in ℓ_p norm

- Conditions:
- Restricted eigenvalue condition / compatability condition
- no condition on the non-zero coefficients
- Result:
$$||\hat{\beta} - \beta_0||_q = O_P(s^{1/q}\sqrt{\log(p)/n})$$
  with $q \in \{1, 2\}$.
- Interpretation

Theory | Variable Screening

- Conditions
- Restricted eigenvalue condition
- beta-min condition: $\min_{j \in S} |\beta_j^0| >> C\sqrt{s \log(p)/n}$ (C some constant)
- Result:

  $$\mathbb{P}[S_0 \subset \hat{S}] \to 1$$

  $(p \geq n \to \infty)$
- Interpretation

Theory | Variable Selection

- Conditions:
- neighbourhood stability condition (equivalent to irrepresentable condition)
- beta-min condition
- Result:

$$\mathbb{P}[S_0 = \hat{S}] \to 1$$

Extensions

- Adaptive Lasso (Zou, 2006)
- Post-Lasso (Belloni & Chernozhukov, 2011)
- Elastic Net (Zou & Hastie, 2005)
- LAVA (Chernozhukov et al., 2015)
- Group Lasso

Adaptive Lasso (Zou, 2006)

- $\hat{\beta}_{adapt}(\lambda) = \arg\min_\beta \left(||Y - X\beta||_2^2/n + \lambda \sum_{j=1}^p \frac{|\beta_j|}{|\hat{\beta}_{init,j}|} \right)$
  where $\hat{\beta}_{init}$ is an initial estimator (e.g. Lasso from an initial stage)
- Intuition: $\hat{\beta}_{init,j} = 0$ leads to $\hat{\beta}_{adapt,j} = 0$
  $|\hat{\beta}_{init,j}|$ large $\Rightarrow$ small penalty
- Goal: Reduction of bias of Lasso

Post-Lasso (Belloni & Chernozhukov, 2011)

- $\hat{\beta}(\lambda) = \arg\min \left(||Y - X\beta||_2^2/n + \lambda|\beta|_1 \right)$
- $\hat{T} = supp(\hat{\beta}) = \{j \in \{1, \ldots, p\} : |\hat{\beta}_j| > 0\}$
- Post model selection estimator $\tilde{\beta}$ (Post-Lasso)

$$\tilde{\beta} = \arg\min_{\beta} ||Y - X\beta||_2^2/2 : \quad \beta_j = 0 \text{for each} j \in \hat{T}^C th$$

- Idea: Reduce bias by running OLS on the variables selected by Lasso in a first stage

Elastic Net (Zou & Hastie, 2005)

- Idea: Combination of ℓ_1- and ℓ_2-penalty
- ℓ_1-penalty: sparse model
- ℓ_2-penalty: enforcing grouping effect, stabilization regularization path, removes limit on number of selected variables

$$\hat{\beta} = \arg\min \left(||Y - X\beta||_2^2/n + \lambda_2||\beta||_2 + \lambda_1||\beta||_1 \right)$$

- $\hat{\beta}_{enet} = (1 + \lambda)(\hat{\beta})$

LAVA (Chernozhukov et al., 2015)

- Idea: $\theta = \underbrace{\beta} + \underbrace{\delta}$ = dense + sparse part
- $\hat{\theta} = \hat{\beta} + \hat{\delta}$
-

$$(\hat{\beta}, \hat{\delta}) = \arg \min_{(\beta', \delta')} \{ l(data, \beta + \delta) + \lambda_2 ||\beta||_2^2 + \lambda_1 ||\delta||_1$$

Group Lasso

- Motivation: with factor variables, one would like to choose if all categories or none of them should be included.
- $\mathcal{G}_1, \ldots, \mathcal{G}_q$ groups which partition the index set $\{1, \ldots, p\}$
- $\beta = (\beta_{\mathcal{G}_1}, \ldots, \beta_{\mathcal{G}_1})$, $\mathcal{G}_j = \{\beta_r, r \in \mathcal{G}_j$
- $\hat{\beta}(\lambda) = \arg\min_\beta Q_\lambda(\beta)$
- $$Q_\lambda(\beta) = 1/n||Y - X\beta||_2^2 + \lambda \sum_{j=1}^{q} m_j ||\beta_{\mathcal{G}_j}||_2$$

  $m_j = \sqrt{T_j}$, $T_j = |\mathcal{G}_j|$
- Either all variables in a group have either value zero or have a value different from zero. Selection of groups of variables (e.g. factors!)