Parameter Selection Algorithms for High-dimensional Linear Models

Zhaoxue Tong

Penn State University
Department of Statistics

Outline

Background

High-dimensional Sparse Linear Model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

n: # of observations.
p: # of predictors.
$\boldsymbol{X}$: $n \times p$ design matrix, $\boldsymbol{p} > \boldsymbol{n}$, each row $X_i \sim N(0, \Sigma)$.
$\boldsymbol{Y}$: $n \times 1$ response vector.
$\boldsymbol{\beta}$: $p \times 1$ coefficient vector, a **small** subset of β_i's is non-zero.
$\boldsymbol{\epsilon}$: noise, $var(\boldsymbol{\epsilon}) = \sigma^2 < \infty$.

Goal: Consistently estimate $\boldsymbol{\beta}$ and recover its support.
Methods:

- Penalization-based methods: LASSO, SCAD;
- Penalization-free methods: LAT, RAT (Wang et al. 2016).

Advantages of LAT and RAT:

- Theory: **Random** design model; **highly correlated** predictors; **general** noise; **ultra-high** dimensional setting ($\ln p = o(n)$).
- Computation: **Parallelizable** for large p; **non-iterative**.

Goals

More rigorous comparison of LASSO, SCAD, LAT and RAT in terms of

- Coefficient estimation;
- Variable selection;
- Runtime.

LASSO & SCAD

- Least Absolute Shrinkage and Selection Operator (LASSO)
  - Estimates β by minimizing

$$L(\boldsymbol{\beta}) = \frac{1}{2n}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_1.$$

  - **A**lternating **D**irection **M**ethod of **M**ultipliers algorithm.
- Smoothly Clipped Absolute Deviation (SCAD)
  - Estimates β by minimizing

$$L(\beta_1, ..., \beta_p) = \frac{1}{2n}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|).$$

  - **C**yclic **C**oordinate **D**escent algorithm.
  - Select the tuning parameter $\hat{\lambda} = \arg\min_\lambda \text{HBIC}(\lambda)$.

LAT & RAT

Idea: Pre-selection + Hard thresholding + OLS

Motivation of Pre-selection:

OLS estimator : $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

Ridge estimator: $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \lim_{\lambda\to 0}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y};$$
$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda\boldsymbol{I}_n)^{-1}\boldsymbol{Y}, \quad \forall p, n, \lambda > 0.$$

High-dimensional version of the OLS:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(HD)} &= \lim_{\lambda\to 0}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda\boldsymbol{I}_n)^{-1}\boldsymbol{Y} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{Y}\\ &= \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{\epsilon} = \Phi\boldsymbol{\beta} + \boldsymbol{\eta}.\end{aligned}$$

If $\boldsymbol{\beta}$ is **sparse** and Φ is **diagonally dominant**, we can use $\hat{\boldsymbol{\beta}}^{(HD)}$ for **dimension reduction**.

LAT & RAT

Algorithm 1: Least-squares Adaptive Thresholding (LAT)

Input: $\{X_i, Y_i\}_{i=1}^{n}, d = \lceil n/\ln(n) \rceil, \delta = .2$

Output: Estimated β

Stage 1: Pre-selection

1 : $\hat{\boldsymbol{\beta}}^{(HD)} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{Y}$. Denote the model corresponding to the d largest $|\hat{\beta}_j^{(HD)}|$'s as $\tilde{M}_d$.

Stage 2: Hard thresholding

2 : $\hat{\boldsymbol{\beta}}^{(OLS)} = (\boldsymbol{X}_{\tilde{M}_d}^T \boldsymbol{X}_{\tilde{M}_d})^{-1} \boldsymbol{X}_{\tilde{M}_d}^T \boldsymbol{Y}$;

3 : $\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-d)$;

4 : $\bar{C} = (\boldsymbol{X}_{\tilde{M}_d}^T \boldsymbol{X}_{\tilde{M}_d})^{-1}$;

5 : Hard threshold $\hat{\boldsymbol{\beta}}^{(OLS)}$ by MEAN$(\sqrt{2\hat{\sigma}^2 \bar{C}_{ii} \ln(4d/\delta)})$. Denote the refined model as $\hat{M}$;

Stage 3: Refinement

6 : $\hat{\boldsymbol{\beta}}_{\hat{M}} = (\boldsymbol{X}_{\hat{M}}^T \boldsymbol{X}_{\hat{M}})^{-1} \boldsymbol{X}_{\hat{M}}^T \boldsymbol{Y}$;

7 : $\hat{\boldsymbol{\beta}}_i = 0, \forall i \notin \hat{M}$;

return $\hat{\boldsymbol{\beta}}$

Non-iterative!

Ridge Adaptive Thresholding (RAT): replace $\bar{C}$ with its ridge version $(\boldsymbol{X}_{\tilde{M}_d}^T \boldsymbol{X}_{\tilde{M}_d} + rI_d)^{-1}$.

Numerical Results

Experiment		LAT	RAT	LASSO	SCAD
1	RMSE	0.4835	0.4835	0.8718	0.0251
	# FPs	0.0650	0.0650	0.0350	0.0000
	# FNs	0.2300	0.2300	0.0000	0.0000
	Time	4.2	4.0	20.7	61.2
2	RMSE	0.0565	0.0575	1.9849	0.0521
	# FPs	0.0350	0.0350	0.0350	0.0000
	# FNs	0.0000	0.0000	0.0000	0.0000
	Time	3.3	3.3	23.3	274.7
3	RMSE	23.6069	9.4343	9.4710	7.0446
	# FPs	0.7500	0.1800	0.0000	0.0200
	# FNs	1.1750	1.1700	0.0050	0.0050
	Time	10.0	9.6	208.7	252.1
4	RMSE	0.0045	0.0045	0.1526	0.0044
	# FPs	0.0100	0.0100	0.0600	0.0000
	# FNs	0.0000	0.0000	0.0000	0.0000
	Time	4.2	3.8	15.7	325.4
5	RMSE	0.0268	4.1056	0.2427	0.0218
	# FPs	0.0800	0.0050	0.1250	0.0000
	# FNs	0.0000	0.9750	0.0000	0.0000
	Time	4.8	4.6	21.0	80.5

Table 1: Results for $(n, p) = (200, 1000)$

- Report
  - root mean squared error (RMSE) $||\hat{\beta} - \beta||_2$;
  - false negatives (#FN);
  - false positives (#FP);
  - walltime.
- Coefficient estimation: 1.SCAD 2.LAT/RAT;
- Variable selection: 1.SCAD 2.LASSO;
- Runtime: 1.LAT/RAT 2.LASSO.

Conclusion

1. Coefficient estimation: SCAD;
2. Variable selection: SCAD;
3. Runtime: LAT.