Econometrics **March 5, 2023**

Topic 12: Non-convex Learning

by Sai Zhang

Key points:

Disclaimer: *The note is built on Prof. Jinchi Lv's lectures of the course at USC, DSO 607, High-Dimensional Statistics and Big Data Problems.*

12.1 L0 Penalized Likelihood

Consider the model selection problem of choosing a parameter vector $\boldsymbol{\theta}$ that maximizes the penalized likelihood

$$\mathcal{L}_n(\boldsymbol{\theta}) - \lambda \|\boldsymbol{\theta}\|_0 \tag{12.1}$$

where the L_0-norm $\|c\dot{\|}_0$ denotes the **the number of nonzero components**, and $\lambda \geq 0$ is still the regularization parameter.

The L_0−penalized likelihood method is equivalent to **the best subset selection**

- given $\|\boldsymbol{\theta}_0\|_0 = m$, the solution to Problem 12.1 is the **best subset** that has the **largest** maximum likelihood among all subsets of size m
- then, choose the model size m among the p size-m best subsets ($1 \leq m \leq p$) by maximizing 12.1

hence it's a combinatorial problem, computationally complex.

L_0-Penalized Empirical Risk Minimization More generally, consider a unified approach of L_0−penalized empirical risk minimization for variable selection:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left\{ \hat{R}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_0 \right\} \tag{12.2}$$

where $\hat{R}(\boldsymbol{\theta})$ is the empirical risk function, which could be of different forms

- **negative log-likelihood loss**: equivalent t L_0-penalized likelihood
- **squared error (quadratic) loss**: L_0-penalized least squares
- selection via **RSS** (residual sum of squares): for the adjusted R^2

  $$R^2_{\text{adj}} = 1 - \frac{n-1}{n-d}\frac{RSS_d}{TSS}$$

  it's clear that $\max R^2_{\text{adj}} \Leftrightarrow \min \log\left(\frac{RSS_d}{n-d}\right)$, and since $\frac{RSS_d}{n} \simeq \sigma^2$, then

  $$n \log \frac{RSS_d}{n-d} \simeq \frac{RSS_d}{\sigma^2} + d + n(\log \sigma^2 - 1)$$

  which shows that adjusted R^2 method is approximately equivalent to 12.2 with $\lambda = 1/2$

- **generalized corss-validation** (GCV), **corss-validation** (CV)
- **risk inflation factor** (RIC): use $\lambda = \log p$, adjusting for the inflation of prediction risk caused by searching p variables[1]
- **AIC** ($\lambda = 1$), **BIC** ($\lambda = \frac{\log n}{2}$)

12.1.1 Properties of L0-Regularization Methods

risk bounds for model selection (Barron et al., 1999): for a family of models $\{S_m : m \in \mathcal{M}_p\}$, The penalty term generally takes the form of

$$\frac{\kappa L_m D_m}{n}$$

where

- κ: a positive constant
- $D_m = |S_m|$: the model dimension, account for the difficulty to estimate **within** the model S_m
- $L_m \geq 1$: a weight that satisfies: $\sum_{m\in\mathcal{M}_p} \exp(-L_m D_m) \leq 1$, accounting for the noise due to **the size** of the list of models

hence, in the linear model, the L_0-regularized estimator $\hat{\boldsymbol{\beta}}$ satisfies that

$$\mathbb{E}\left[n^{-1}\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0\|_2^2\right] \leq C \inf_{m\in\mathcal{M}_p}\left\{\min_{\boldsymbol{\beta}\in\text{model } S_m}\left[n^{-1}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}_0\|_2^2\right] + \frac{\kappa L_m D_m}{n}\right\}$$

where ***the tradeoff***: approximation error $n^{-1}\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0\|_2^2$, and the cost of searching $\frac{\kappa L_m D_m}{n}$

computational complexity L_0-regularization methods are appealing w.r.t. risk properties, but in high-dimensional settings, the computation is infeasible (combinatorial), and discontinuous, non-convex penalty function $\lambda\|\boldsymbol{\beta}\|_0$

12.1.2 Generalizations of L0-Regularization Methods

Consider continuous or convex relaxation of the L_0-regularization method

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\hat{R}(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda\left(|\beta_j|\right)\right\} \tag{12.3}$$

where, as in Problem 12.2

- $\hat{R}(\boldsymbol{\beta})$: the empirical risk function
- $p_\lambda(t), t \geq 0$: the nonnegative penalty function indexed by the regularization parameter $\lambda \geq 0$ with $p_\lambda(0) = 0$

[1]The $\log p$ is, once again, from the fact that for Gaussian random variables

$$\max_{1\leq j\leq p}|Z_i| \simeq \sqrt{2\log p}$$

for $(Z_1, \cdots, Z_p)' \sim \mathcal{N}(0, \mathbf{I}_p)$

Choices of penalty function In general, the choices of penalty function can be up for the researchers to decide. Fan and Li (2001) proposed 3 criteria for penalty function selection

- **Sparsity**: sets small estimated coefficients to 0, for *variable selection* and *reduction of model complexity*
- **Approximate unbiasedness**: nearly unbiased, especially when the true coefficient β_j is large
- **Continuity**: continuous in data to reduce instability in model selection

References

Andrew Barron, Birgé Lucien, and Massart Pascal. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.