Least-squares refitted Lasso

Miranda Fix

Statistical Machine Learning

September 30, 2014

The Setup

Suppose we have the model

$$Y = X\beta^* + \sigma\epsilon,$$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{nxp}$, $\beta^* \in \mathbb{R}^p$ and the noise vector $\epsilon \in \mathbb{R}^n$ has associated noise level $\sigma > 0$.
Define the active set by

$$S := \left\{ j \in \{1, \dots, p\} : \beta^*_j \neq 0 \right\}$$

The sparsity level is $s := |S|$.
Consider $p \approx n$ or $p \gg n$ but $s < n, p$

The Setup

The initial estimators are

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ g(||Y - X\beta||_2^2) + \lambda ||\beta||_1 \right\} \tag{1}$$

The LS refitted estimators are

$$\overline{\beta}_{\hat{S}} := \arg\min_{\xi \in \mathbb{R}^{|\hat{S}|}} ||Y - X_{\hat{S}}\xi||_2^2 \ , \ \ \overline{\beta}_{\hat{S}^c} := 0 \tag{2}$$

Results

- For both prediction and estimation, LS refitting can be beneficial (especially when the level of correlation in X is low). However, LS refitting can be disadvantageous if both the design matrix is highly correlated and the sparsity level is considerably larger than 1.
- LS refitting can be advisable for estimation but exhibits better performances for prediction.

[INTERLUDE: PRETTY PICTURES]

Theorem

Theorem (Lederer, 2013)

With probability one, the LS refitted estimator (2) *relates to the initial estimator* (1) *as follows:*

$$\overline{S} = \hat{S}$$

$$||\overline{\beta} - \hat{\beta}||_q = ||(X_{\hat{S}}^T X_{\hat{S}})^{-1} sign(\hat{\beta}_{\hat{S}})||_q \frac{\lambda}{2g'(||Y - X\hat{\beta}||_2^2)}$$

$$||X\overline{\beta} - X\beta^*||_2^2 - ||X\hat{\beta} - X\beta^*||_2^2 \leq ||(X_{\hat{S}}^T X_{\hat{S}})^{-1} sign(\hat{\beta}_{\hat{S}})||_1 \frac{\lambda\sigma||X_{\hat{S}}^T \epsilon||_\infty}{g'(||Y - X\hat{\beta}||_2^2)}$$

A Criterion

$$F(\hat{S}) := \frac{1}{|\hat{S}|} \left| \left\{ j \in \hat{S} : sign(\hat{\beta}_j) \neq sign(((X_{\hat{S}}^T X_{\hat{S}})^{-1} sign(\hat{\beta}_{\hat{S}}))_j) \right\} \right| \in [0, 1]$$

If $F(\hat{S}) \leq c$ then use the LS-refitted estimator $\overline{\beta}$. Otherwise, use the initial estimator $\hat{\beta}$.

The choice of c depends on whether we are interested in prediction or estimation (e.g. $c = 0.4$ for prediction and $c = 0.2$ for estimation – these are based on heuristic arguments that use the KKT conditions of Lasso).

References

Johannes Lederer (2013)
Trust, but verify: benefits and pitfalls of least-squares refitting in high dimensions
arXiv:1306.0113v1 [stat.ME]