Subset Selection with Shrinkage

Xiaoqian Wang

2023-02-16

Mazumder, R., Radchenko, P., & Dedieu, A. (2022). Subset selection with shrinkage: Sparse linear modeling when the SNR is low. Operations Research.

Hazimeh, H., & Mazumder, R. (2020). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. Operations Research, 68(5), 1517-1537.

1 Best Subset Selection

Suppose that the data are generated from a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where matrix $\mathbf{X}$ is deterministic and the elements of $\boldsymbol{\epsilon} \in \mathbb{R}^n$ are independent $N\left(0, \sigma^2\right)$.

$$\underset{\boldsymbol{\beta}}{\operatorname{minimize}} \quad \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k$$

1. Computationally infeasible

2. Signal to noise ratio (SNR) $\longrightarrow$ **Overfit**

    1. Measured by $\|\boldsymbol{\beta}^*\|_1 / \sigma$, $\|\boldsymbol{\beta}^*\|_2 / \sigma$, $\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2 / \|\boldsymbol{\epsilon}\|_2^2$.
    2. When SNR is low, $\hat{\boldsymbol{\beta}}$ is instable.
    3. The impossibility of variable selection when the signal is weak.
    4. The un-regularized fit can be improved by shrinkage when σ is large.

So. it's note a right approach when the noise level is high.

Q: How to fix the problem?

- Continuous shrinkage methods

    1. Lasso and ridge regression trade off an increase in bias with a decrease in variance.
    2. The estimated models are **denser** than those produced by best subset selection.

- Sparsity & Shrinkage

2 Subset Selection with Shrinkage

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \underbrace{\lambda\|\boldsymbol{\beta}\|_q}_{\text{Shrinkage}} \quad \text{s.t.} \quad \underbrace{\|\boldsymbol{\beta}\|_0 \leq k}_{\text{Sparsity}}.$$

- separate out the effects of shrinkage and sparsity.

2.1 Mixed Integer Optimization formulations

$$\begin{aligned}
&\text{minimize } \tfrac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_q \\
&\text{s.t.} \quad -\mathcal{M}z_j \leq \beta_j \leq \mathcal{M}z_j, j \in [p]; \\
&\mathbf{z} \in \{0,1\}^p; \\
&\textstyle\sum_j z_j = k
\end{aligned}$$

This can be written as follows:

$$\begin{aligned}
&\text{minimize } u/2 + \lambda v \\
&\text{s.t. } \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \leq u \\
&\|\boldsymbol{\beta}\|_q \leq v \\
&-\mathcal{M}z_j \leq \beta_j \leq \mathcal{M}z_j, j \in [p]; \\
&\mathbf{z} \in \{0,1\}^p; \\
&\textstyle\sum_j z_j = k
\end{aligned}$$

Computational performance of MISOCO solvers (Gurobi, for example) is found to improve by adding structural implied inequalities, or cuts, to the basic formulation.

A structured version of the above formulation with additional implied inequalities (cuts) for improved lower bounds is:

$$\begin{aligned}
&\text{minimize}\, u/2 + \lambda v \\
&\text{s.t. } \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \leq u \\
&\|\boldsymbol{\beta}\|_q \leq v \\
&-\mathcal{M}_j z_j \leq \beta_j \leq \mathcal{M}_j z_j, j \in [p] \\
&z_j \in \{0,1\}, j \in [p] \\
&\textstyle\sum_j z_j = k \\
&\text{- } \mathcal{M}_i \leq \beta_i \leq \mathcal{M}_i, i \in [p] \\
&\text{- } \overline{\mathcal{M}}_i^- \leq \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \leq \overline{\mathcal{M}}_i^+, i \in [n] \\
&\|\boldsymbol{\beta}\|_1 \leq \mathcal{M}_{\ell_1}
\end{aligned}$$

- $\mathcal{M}_i, i \in [p]$ denote bounds on β_i 's.
- $-\overline{\mathcal{M}}_i^-, \overline{\mathcal{M}}_i^+$ denote bounds on the predicted values $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle$ for $i \in [n]$.
- $\mathcal{M}_{\ell_1}$ denotes an upper bound on the ℓ_1-norm of the regression coefficients $\|\boldsymbol{\beta}\|_1$.

Consider the following extended family of L_0-based estimators. That is, L_0L_q-regularized regression problems of the form:

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda_0\|\beta\|_0 + \lambda_q\|\beta\|_q^q$$

where $q \in \{1, 2\}$ determines the type of the additional regularization (i.e., L_1 or L_2).