Highdimension Ordinary Least-squares Projection for Screening Variables

Demerits of Existing Variable Selection Methods

Penalized approach

- Can give non-consistent models if the irrepresentable condition on the design matrix violated

  (irrepresentable condition : the relevant variable may not be very correlated with the irrelevant variables)
- In HDLSS(High dimension low-sample size) settings, penalized approaches may not work
- Computation cost of penalizing methods for large-scale optimization is very high

SIS (Sure Independence Screening)

- Marginal Correlation Condition is often violated in HDLSS settings

  (MCC : Marginal correlations for the important variables must be bounded away from zero)

Highdimension Ordinary Least-squrares Projection for Screening variables

- Assumptions
  1. It follows linear regression assumptions
     - $Y = X\beta + \epsilon$
     - $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$
  2. dimension of variables p is much more higher than number of observations n $(p > n)$

     $\rightarrow$ XX' is invertible

- Algorithm
  1. Calculate $A = X'(XX')^{-1}$
  2. Calculate $\hat{\beta} = AY$
  3. Rank the componentes of $\hat{\beta}$ and select predictors x_j that satisfies $|\hat{\beta}_j| > \gamma$
  4. Perform data analysis with selected variables

- Properties
  1. it can be viewed as projection matrix to the rowspace of X

     $\hat{\beta} = AY = A(X\beta + \epsilon) = X'(XX')^{-1}X\beta + X'(XX')^{-1}\epsilon$

     which means HOLP uses the rowspace of X to capture β

2. This projection matrix $X'(XX')^{-1}X$ preserves the rank order of entries in β

   $\rightarrow$ which can makes variable screening possible by selecting top few $|\beta_j|$
3. Its computational complexity is $O(n^2p)$

   $\rightarrow$ in Unltra highdimensional assumptions, It is very computationally efficient
4. It Assymptotically has Sure Screenig property if we choose γ as

$$\frac{p\gamma_n}{n^{1-\tau-k}} \rightarrow 0 \text{ and } \frac{p\gamma_n\sqrt{\log n}}{n^{1-\tau-k}} \rightarrow \infty$$