Variable Selection in High Dimensional Time-varying Effect Model: based on Iterative Shrinkage Thresholding Algorithm (ISTA)

Yujie Liao

Department of Statistics

Table of contents

Motivation

- Variable selection in high-dimensional time-varying effect model

$$Y_i = \sum_{k=1}^{K} \beta_k(t_i) X_{ik} + \epsilon_i$$

- iid $(Y_i, \boldsymbol{X}_i, t_i)$, $i = 1, \cdots, n$ (not longitudinal case for simplicity)
- General method: approximate $\beta_k(t) \approx \sum_{l=1}^{L} \beta_{kl} B_l(t)$
- Interests: $\boldsymbol{\beta}_k^* = (\beta_{k1}, \cdots, \beta_{kL})^T$ and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, \cdots, \beta_K^{*T})^T$
- Select the non-zero $\beta_k(t)$ by minimizing

$$Q(\boldsymbol{\beta}^*) = \frac{1}{2n} \sum_{i=1}^{n} \{Y_i - \sum_{k-1}^{K} \sum_{l=1}^{L} \beta_{kl} B_l(t_i) X_{ik}\}^2 + g(\boldsymbol{\beta}^*)$$

$$= \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \beta^{*T} \boldsymbol{X}_i^*)^2 + g(\boldsymbol{\beta}^*)$$

where $\boldsymbol{X}_i^* = (X_{i1} \boldsymbol{B}(t_i)^T, \cdots, X_{ik} \boldsymbol{B}(t_i)^T)$, $g(\beta^*)$ is the penalty function (mentioned later).
- Two challenges: $Q(\boldsymbol{\beta}^*) = f(\boldsymbol{\beta}^*) + g(\boldsymbol{\beta}^*)$

Iterative Shrinkage Thresholding Algorithm (ISTA)

- Target: $Q(\beta) = f(\beta) + g(\beta)$, where $f(\cdot)$ is smooth (probably complicated) and $g(\cdot)$ is non-smooth
- Local isotropic quadratic approximation:
  1. $Q(\beta) \approx Q_A(\beta) = f_A(\beta|\boldsymbol{\beta}_{r-1}) + g(\beta)$, where $f_A(\beta|\boldsymbol{\beta}_{r-1})$ is

     $$f(\boldsymbol{\beta}_{r-1}) + f^{'}(\boldsymbol{\beta}_{r-1})^T(\beta - \boldsymbol{\beta}_{r-1}) + \frac{1}{2}(\beta - \boldsymbol{\beta}_{r-1})^T \frac{\partial^2 f(\boldsymbol{\beta}_{r-1})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}(\beta - \boldsymbol{\beta}_{r-1})$$

  2. By Raleigh-Ritz Theorem: $\forall \boldsymbol{x} \neq \mathbf{0} \in \mathbb{R}^n$, $\boldsymbol{A}: n \times n$ Hermitian matrix,

     $$\lambda_{min}(\boldsymbol{A})\boldsymbol{x}^T\boldsymbol{x} \leq \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \leq \lambda_{max}(\boldsymbol{A})\boldsymbol{x}^T\boldsymbol{x}$$

     If $\{\frac{\partial^2 f(\boldsymbol{\beta}_{r-1})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\}^{-1} \approx s_r \boldsymbol{I}$, then

     $$f_A(\beta) = f(\boldsymbol{\beta}_{r-1}) + f^{'}(\boldsymbol{\beta}_{r-1})^T(\beta - \boldsymbol{\beta}_{r-1}) + \frac{1}{2s_r}\|\beta - \boldsymbol{\beta}_{r-1}\|^2$$

  3. Goal:
     $\boldsymbol{\beta}_r = \arg\min_\beta Q_A(\beta|\boldsymbol{\beta}_{r-1}, s_r) = \arg\min_\beta\{f_A(\boldsymbol{\beta}_r|\boldsymbol{\beta}_{r-1}, s_r) + g(\boldsymbol{\beta}_r)\}$

Why interesting?

$$
\begin{aligned}
\beta_r &= \arg\min_{\beta}\{f^{'}(\beta_{r-1})^T\beta + \frac{1}{2s_r}\left\|\beta - \beta_{r-1}\right\|^2 + g(\beta_r)\} \\
&= \arg\min_{\beta}\left[\frac{1}{2s_r}\left\|\beta - \{\beta_{r-1} - s_r f'(\beta_{r-1})\}\right\|^2 + g(\beta_r)\right] \\
&= \arg\min_{\beta}\{\frac{1}{2s_r}\left\|\beta - \tilde{\beta}_r\right\|^2 + g(\beta_r)\} \quad \text{where} \quad \tilde{\beta}_r = \beta_{r-1} - s_r f'(\beta_{r-1})
\end{aligned}
$$

- Pros:
    1. Standard variable selection subject function
    2. Not depend on $f(\cdot)$ at all, only on $g(\cdot)$
    3. $f(\cdot)$ can be complicated, we only need to consider its gradient
    4. closed form for many important functions (e.g. lasso $\rightarrow$ component-wise $\rightarrow$ soft-thresholding rule)

Also called Proximal Gradient Descent

- Define $prox_{g,s}(\boldsymbol{z}) = \arg\min_{\boldsymbol{x}} \frac{1}{2s} \|\boldsymbol{x} - \boldsymbol{z}\|^2 + g(\boldsymbol{x})$
- Choose initialize $\boldsymbol{x}_0$, repeat

$$\boldsymbol{x}_r = prox_{g,s_r}\{\boldsymbol{x}_{r-1} - s_r f'(\boldsymbol{x}_{r-1})\}, \quad r = 1, 2, \cdots$$

- Then $\boldsymbol{x}_r = \boldsymbol{x}_{r-1} - s_r \cdot G_{s_r}(\boldsymbol{x}_{r-1})$,
  where G_s is the generalized gradient of f,

$$G_s(\boldsymbol{x}) = \frac{\boldsymbol{x} - prox_{g,s}(\boldsymbol{x} - s \cdot g'(\boldsymbol{x}))}{s}.$$

Another Challenge

- $Q(\boldsymbol{\beta}^*) = \frac{1}{2n} \left\| \boldsymbol{Y} - \boldsymbol{X}^* \boldsymbol{\beta}^* \right\|^2 + g(\boldsymbol{\beta}^*)$, $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, \cdots, \boldsymbol{\beta}_K^{*T})^T \in \mathbb{R}^{K \times L}$
- $\boldsymbol{\beta}_k^* = (\beta_{k1}, \cdots, \beta_{kL})^T \in \mathbb{R}^L$ is associated with group k, since $\beta_k(t) \approx \sum_{l=1}^{L} \beta_{kl} B_l(t)$
- Time-varying model: not penalty on each component of $\boldsymbol{\beta}^* \in \mathbb{R}^{K \times L}$
- Group variable selection: need coefficients in a group to be in or out of the model at the same time $\to$ sparsity between groups, not within groups
- Group LASSO, group SCAD, other constraints: $\sum_{k=1}^{K} p_\lambda(\|\boldsymbol{\beta}_k^*\|_2)$ not component-wise, but group-wise
- Solution:
  $\min_{\boldsymbol{\beta}^*} f(\boldsymbol{\beta}^*) = \frac{1}{2n} \left\| \boldsymbol{Y} - \boldsymbol{X}^* \boldsymbol{\beta}^* \right\|^2$ s.t. $\tau(\{k : \|\boldsymbol{\beta}_k^*\|_2 > 0\}) \leq m$

ISTA with Constraints

$$\begin{aligned}
\beta^*_{k,r} &= \arg\min_{\beta^*_k} f_A(\beta^*_k|\beta^*_{k,r-1}) \\
&= \arg\min_{\beta^*_k}\{f(\beta^*_{k,r-1}) + f^{'}(\beta^*_{k,r-1})^T(\beta^*_k - \beta^*_{k,r-1}) + \frac{1}{2s_r}\left\|\beta^*_k - \beta^*_{k,r-1}\right\|^2\} \\
&= \arg\min_{\beta^*}\{\frac{1}{2s_r}\left\|\beta^*_k - \tilde{\beta}^*_{k,r}\right\|^2\} = h_A(\beta^*_k),
\end{aligned}$$

where $\tilde{\beta}^*_r = \beta^*_{r-1} - s_r f'(\beta^*_{r-1})$, for $\forall k = 1, \cdots, K$,

s.t. $\tau(\{k : \|\beta^*_k\|_2 > 0\}) \leq m$.

- When $\hat{\beta}^*_k \neq 0$, then $\hat{\beta}^*_k = \tilde{\beta}^*_k \quad \Rightarrow \quad h_{A1}(\beta^*_k) = 0$
  When $\hat{\beta}^*_k = 0 \quad \Rightarrow \quad h_{A2}(\beta^*_k) = \frac{1}{2s_r}\left\|\tilde{\beta}^*_k\right\|^2$
- Set $\beta^*_k = 0$ if $h_{A2}(\beta^*_k) - h_{A1}(\beta^*_k) = \frac{1}{2s_r}\left\|\tilde{\beta}^*_k\right\|^2$ is small,
  i.e. $\left\|\tilde{\beta}^*_k\right\|^2$ is small
- Let $g_k = \tilde{\beta}^{*T}_k \tilde{\beta}^*_k$, sort g_i so that $g_{(1)} \geq g_{(2)} \geq \cdots g_{(K)}$.
  By hard-thresholding rule, $\hat{\beta}^*_k = \tilde{\beta}^*_k \boldsymbol{I}\{g_k > g_{(m+1)}\}$

Relation with MM Algorithm: by backtracking

- $f(\beta^*) \to f_A(\beta^*|\beta^*_{r-1}) =$
  $f(\beta^*_{r-1}) + f^{'}(\beta^*_{r-1})^T(\beta^* - \beta^*_{r-1}) + \frac{1}{2s_r}\left\|\beta^* - \beta^*_{r-1}\right\|^2$
- Check two conditions of MM algorithm:
  1. $f_A(\beta^*_{r-1}|\beta^*_{r-1}) = f(\beta^*_{r-1})$
  2. Majorizing-minimization: choose s_r by backtracking rule
     - 2.1 Set $s_0 > 0$, $0 < \delta < 1$, β^*_0.
     - 2.2 Find the smallest non-negative integer i_r s.t. with $s = \delta^{i_r} s_{r-1}$,

$$f(\beta^*_{r,s}) \le f_A(\beta^*_{r,s}|\beta^*_{r-1}, s),$$

i.e. $f(\beta^*_{r,s}) \le f(\beta^*_{r-1}) + f^{'}(\beta^*_{r-1})^T(\beta^* - \beta^*_{r-1}) + \dfrac{1}{2s}\left\|\beta^* - \beta^*_{r-1}\right\|^2$,

where $\beta^*_{k,r,s} \leftarrow \hat{\beta}^*_{k,r,s} = \tilde{\beta}^*_{k,r,s}\boldsymbol{I}\{g_k > g_{(m+1)}\}$ is a function of s.

     - 2.3 Set $s_r = \delta^{i_r} s_{r-1} \quad \Rightarrow \quad \hat{\beta}^*_{k,r,s_r}$

Simulation

- $(t_i^{'}, \boldsymbol{X}_i^{'}) \sim N_{K+1}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (\sigma_{ij})$
  $\sigma_{ij} = 1$ if $i = j$; $\sigma_{ij} = \rho$ if $i \neq j$; $\rho = 0.5$
- $t_i = \Phi(t_i^{'})$, where $\Phi(\cdot)$ is the CDF of $N(0, 1)$
- $L = 5,\ Rep = 1000,\ K = 800,\ n = 200,\ m = [\frac{n^{4/5}}{log(n^{4/5})}]$
- True coefficient functions:
  $\beta_1(t) = 0.95cos(\frac{\pi t}{2}) + 3.36, \quad \beta_2(t) = 1.5sin(2\pi t) + 3,$
  $\beta_3(t) = -2t^2 - 4, \quad \beta_4(t) = 0.52t + t^3 + 2.9$
- Criteria: P_k: the proportion of submodels $\hat{\mathcal{M}}$ with size m that contain X_k among Rep repetitions
- Results:

P_1	P_2	P_3	P_4
1	0.94	0.88	1