Homework assignments Week 13 (Students should submit their homework before 21 p.m. on December 25, 2021.)

Programming work

5. Variable selection via lasso estimation (6%)

In this programming work we will build an ADMM-based iterative scheme to find a lasso estimate of regression coefficients in the following linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

The lasso estimate for $\boldsymbol{\beta}$ is defined as

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 \right\},$$

where $\lambda \geq 0$ is a user-specified tuning parameter.

To build an iterative scheme to find $\widehat{\beta}^{\text{lasso}}$, you need to:

1. Reformulate the lasso estimation problem in an ADMM form and derive the corresponding augmented Lagrangian and therefore the iterative scheme.

2. The second line of the ADMM iterative scheme should be

$$\boldsymbol{\beta}^{r+1} = \arg\min_{\boldsymbol{\beta}} \left\{ \lambda||\boldsymbol{\beta}||_1 + \frac{m\rho}{2} \left|\left| \boldsymbol{\beta} - \frac{1}{m}\sum_{k=1}^{m} \boldsymbol{\theta}_k^{r+1} - \frac{1}{m}\sum_{k=1}^{m} \boldsymbol{\alpha}_k^r \right|\right|_2^2 \right\}.$$

   The jth element of $\boldsymbol{\beta}^{r+1}$ has a closed form representation

$$(\boldsymbol{\beta})_j^{r+1} = \text{sign}(v_j)\left(|v_j| - \frac{\lambda}{m\rho} \right)_+, \tag{1}$$

   where v_j is the jth element of

$$\mathbf{v} = \frac{1}{m}\sum_{k=1}^{m} \boldsymbol{\theta}_k^{r+1} + \frac{1}{m}\sum_{k=1}^{m} \boldsymbol{\alpha}_k^r.$$

   Here (1) is called the **soft thresholding operator**.

3. Specify the scale parameter ρ in the iterative scheme. You can use whatever way you like to specify ρ.

4. Specify (a) a **stopping criterion**, (b) a **tolerance** for the error and (c) the **maximum number of iterations** for stopping the iterative scheme. Here you are allowed to use whatever way you like to specify the **stopping criterion** like the following one:

   Some measure on error $\leq$ tol OR The number of iterations $>$ max_iter.

   However, I strongly suggest you to use the **primal residual** and the **dual residual** stated in the course slides to specify the stopping criterion. In addition, the tolerance for the error and the maximum number of iterations should be

$$\begin{aligned} \text{tol} &= 5 \times 10^{-3}, \\ \text{max\_iter} &= 500. \end{aligned}$$

5. To run such an iterative scheme, you need to specify what the size of the subsample (the size of the data block) is in order to execute the first line of the iterative scheme. In addition, since the loss function is a squared l_2-norm loss, there should be a closed form representation for the first line of the iterative scheme.

6. The lasso estimation has a tuning parameter λ. You need to decide what value λ should have in order to yield a good estimate for the regression coefficients. However, once you decide the value of λ, it should not be changed during the iteration.

Data generation: We let the number of observations $n = 3 \times 10^6$ and the number of **true covariates** $p^{\text{true}} = 5$. We use the following model to generate the data of the true covariates and the value of $\mathbf{y}$:

$$\begin{aligned} \boldsymbol{\beta}^{\text{true}} &= (\beta_{100}, \beta_{200}, \beta_{300}, \beta_{400}, \beta_{500}) = (-2, -2, 2, 2, -2), \\ \mathbf{x}_i &= (x_{i100}, x_{i200}, x_{i300}, x_{i400}, x_{i500}), \\ x_{ij} &\sim \text{Normal}(0, 1) \text{ for } i = 1, 2, \cdots, 3 \times 10^6 \text{ and } j = 100, 200, 300, 400, 500, \\ y_i &= -2x_{i100} - 2x_{i200} + 2x_{i300} + 2x_{i400} - 2x_{i500} + \epsilon_i \text{ for } i = 1, 2, \cdots, 3 \times 10^6, \end{aligned}$$

where $\epsilon_i \sim \text{Normal}(0, 0.5)$. After generating the data of the true model part, you need to generate data for the **redundant covariates**, which is covariates with indices $\{1, 2, \cdots, 500\} \setminus \{100, 200, 300, 400, 500\}$. We assume redundant covariate $x_{ij} \sim \text{Normal}(0, 1)$. Now with true covariates and redundant covariates you need to form a covariate matrix $\mathbf{X}$ in which the 100th, 200th, 300th, 400th and 500th columns are the true covariates, and the rest of columns are the redundant covariates. Now with $(\mathbf{y}, \mathbf{X})$, you can find the lasso estimate by running the iterative scheme.

Tasks: Report plots of the following two settings:

1. Fix tuning parameter λ at 4 different values you like and run the iterative scheme under the 4 different values of λ separately. Produce 4 plots according to the 4 different values of λ with the following format: The x-axis is the number of iterations r and the y-axis is the Euclidean norm of the **primal residual** and **dual residual** of the iterative scheme;

2. Select 20 different values of λ from the interval $[0.001, 5]$ and run the iterative scheme under the 20 different values of λ separately. Collect the values of $\widehat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)$. Produce a trace plot of $\widehat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)$ with the following format: The x-axis is the value of λ and the y-axis is $\widehat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)$. Since we have $p = 500$, there should be 500 such trace lines for $\widehat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)$. Use red color to draw the trace lines for the 100th, 200th, 300th, 400th and 500th elements of $\widehat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)$, and gray color to draw the trace line for the rest of elements in $\widehat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)$.