sliced_block_by_block

Xuelong Wang

2017-11-14

Problem

Assume we want to use the SIR method, but the data is too large to be loaded to the memory.

Solution

We can read the data blocks by blocks, and recorded the sufficient statistcs for each block and each slice. At the end, we can add them together.

Some details

Goal

$$x_i = \Sigma_{xx}^{-1/2}(x_i - \bar{x}), \quad V = \sum_{h=1}^{H} p m_h m_h^T$$

Statistics

$\bar{x}$ the mean vector of the x
Σ_{xx} sample covaraince
m_h mean for each slice
p propotion for each slice

sufficient statistics by block

- Assume that the data set is splited into B block, each block is represted as X_b, and we have

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_B \end{bmatrix}$$

- For each X_b, it may contains observations from each slice, thus we may also have

$$X_b = \begin{bmatrix} X_{b1} \\ \vdots \\ X_{bh} \end{bmatrix}$$

Calcuate Σ_{xx}

$$\Sigma_{xx} = \frac{1}{n}(X^T X - \bar{X}^T \bar{X})$$

Since

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_b \end{bmatrix},$$

thus

$$X^T X = \sum_{b=1}^{B} X_b^T X_b$$

$$\bar{X} = \frac{1}{n} \sum_{b=1}^{B} \mathbb{1}_{n_b}^T X_b$$

So all the values can be calcuated by adding the values from all the blocks

Calcuate the slice mean m_h

The notations used for calculating m_h is a little bit of tedious, it invovles two layers of subscribes

- one layer is for the block b
- one layer is for the slice h

Aassume no block, and calculate m_h in matrix format

$$\tilde{m}_h = \frac{\mathbb{1}_{n_h}^T}{n_h} \left[X_h - \bar{X}\right] \Sigma_{xx}^{-1/2},$$

For each slice data X_h, we break down further with blocks

$$X_h = \begin{bmatrix} X_{h1} \\ \vdots \\ X_{hb} \end{bmatrix}, \mathbb{1}_{n_h} = \begin{bmatrix} \mathbb{1}_{n_{h1}} \\ \vdots \\ \mathbb{1}_{n_{hb}} \end{bmatrix}$$

Then the slice mean can be written in the following way

$$\begin{aligned} \tilde{m}_h &= \frac{\mathbb{1}_{n_h}^T}{n_h} \left[X_h - \bar{X}\right] \Sigma_{xx}^{-1/2} \\ &= \frac{1}{n_h} \left(\left[\mathbb{1}_{n_{h1}}^T \dots \mathbb{1}_{n_{hb}}^T\right] \begin{bmatrix} X_{h1} \\ \vdots \\ X_{hb} \end{bmatrix} - \bar{X} \right) \Sigma_{xx}^{-1/2} \\ &= \frac{1}{n_h} \sum_{b=1}^{b=B} \left(\mathbb{1}_{n_{hb}}^T X_{hb} - \bar{X}\right) \Sigma_{xx}^{-1/2} \end{aligned}$$

The sufficient statistcs for each block b and slice h

$$S_{hb} = \mathbb{1}_{n_{hb}}^T X_{hb},\ n_{bh},\ X_b^T X_b$$

Based those statistics we can calculate following

$$n = \sum_b \sum_h n_{hb}, \quad \bar{x} = \frac{\sum_b \sum_h S_{hb}}{n}$$