HiC Data Visualization and Statistical Analysis

Amal Agarwal
Advisors: Dr. Yu Zhang and Dr. Lingzhou Xue

Department of Statistics
Pennstate University

May 25, 2018

Outline

Outline

HiC data

HiC data structure:

- Two Cell types, "Gm12878" and "K562".
- 22 chromosomes and X chromosome in each cell type.
- Intensity matrix for each chromosome with 10K b.p. granularity.

Challenge:

- Intensity matrix is big!

HiC data

HiC data structure:

- Two Cell types, "Gm12878" and "K562".
- 22 chromosomes and X chromosome in each cell type.
- Intensity matrix for each chromosome with 10K b.p. granularity.

Challenge:

- Intensity matrix is big!

Outline

Regression Model

Notation:

- N: size of HiC matrix for one chromosome.
- n_b observations (banded values) for bandwidth parameter b.
- Consider only upper triangular elements (i.e. $j \geq i$).

Penalized Lasso Regression Model:

- $\text{Intensity}_{ij} = \alpha + \beta \times |i - j| + \sum_{k=1}^{N} \gamma_k \mathcal{I}(i \leq k \leq j) + \epsilon$

  subject to $\beta \leq 0, \gamma_k \leq 0, \forall k$ and $\sum_{k=1}^{N} \gamma_k \geq t$, where

  $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Regression Model

Notation:

- N: size of HiC matrix for one chromosome.
- n_b observations (banded values) for bandwidth parameter b.
- Consider only upper triangular elements (i.e. $j \geq i$).

Penalized Lasso Regression Model:

- $\text{Intensity}_{ij} = \alpha + \beta \times |i - j| + \sum_{k=1}^{N} \gamma_k \mathcal{I}(i \leq k \leq j) + \epsilon$

  subject to $\beta \leq 0, \gamma_k \leq 0, \forall k$ and $\sum_{k=1}^{N} \gamma_k \geq t$, where

  $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Regression Model

Equivalent formulation as Penalized Least Squares (PLS):

- $\min \left(\sum_{i=1}^{N} \sum_{j=i}^{\min(N, i+n_b)} \left[\left(\text{Intensity}_{ij} - \alpha - \beta - \sum_{k=1}^{N} \gamma_k \mathcal{I}(i \leq k \leq j) \right)^2 - \lambda \left(\sum_{k=1}^{N} \gamma_k \right) \right] \right)$

Challenges due to large N:

- For $N \sim 25K$ and $b \sim 200$, we have $n_b \sim 10M$.
- The design matrix becomes $\sim 10M \times 25K \implies 25B$ elements. Memory problem!
- Solving the PLS function is computationally expensive.

Regression Model

Equivalent formulation as Penalized Least Squares (PLS):

- $\min \left(\sum_{i=1}^{N} \sum_{j=i}^{\min(N, i+n_b)} \left[\left(\text{Intensity}_{ij} - \alpha - \beta - \sum_{k=1}^{N} \gamma_k \mathcal{I}(i \leq k \leq j) \right)^2 - \lambda \left(\sum_{k=1}^{N} \gamma_k \right) \right] \right)$

Challenges due to large N:

- For $N \sim 25K$ and $b \sim 200$, we have $n_b \sim 10M$.
- The design matrix becomes $\sim 10M \times 25K \implies 25B$ elements. Memory problem!
- Solving the PLS function is computationally expensive.

Our Approach

Coordinate Descent (CD) Algorithms:

- Used by "glmnet"
- Convergence is fast
- Closed form updates
- Active set strategy

Problem Formulation

Objective (generalized notation):

$$f_\lambda(\alpha, \beta, \gamma) = \sum_{i=1}^{n} \left[(y_i - \alpha - \beta z_i - \gamma^T \boldsymbol{x_i})^2 - \lambda \sum_{j=1}^{p} \gamma_j \right]$$

$$(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = \arg\min \left(f_\lambda(\alpha, \beta, \gamma) \right) \quad \begin{matrix} \beta \leq 0, \\ \gamma \leq 0 \end{matrix}$$

Note: Tuning parameter λ is assumed to be known here.

Naive CD Algorithm

Algorithm 1: Naive Coordinate Descent

Input: Tuning Parameter λ, Convergence threshold ϵ

Output: $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$

Initialization: Start with arbitrary $(\alpha^{(1)}, \beta^{(1)}, \gamma^{(1)})$

Compute $f_\lambda^{(1)} = f_\lambda(\alpha^{(1)}, \beta^{(1)}, \gamma^{(1)})$

Set $f_\lambda^{(0)} \leftarrow \infty$ and $t \leftarrow 1$

while $|f_\lambda^{(t)} - f_\lambda^{(t-1)}| > \epsilon$ **do**

  $\alpha^{(t+1)} \leftarrow \arg\min_\alpha(f_\lambda(\alpha, \beta^{(t)}, \boldsymbol{\gamma}^{(t)}))$

  $\beta^{(t+1)} \leftarrow \arg\min_\beta(f_\lambda(\alpha^{(t+1)}, \beta, \boldsymbol{\gamma}^{(t)}))$

  for $j \leftarrow 1$ **to** p **do**

    $\gamma_j^{(t+1)} \leftarrow \arg\min_{\gamma_j}(f_\lambda(\alpha^{(t+1)}, \beta^{(t+1)}, \gamma_1^{(t+1)}, \ldots, \gamma_{j-1}^{(t+1)}, \gamma_j, \gamma_{j+1}^{(t)}, \ldots, \gamma_p^{(t)}))$

  Compute $f_\lambda^{(t+1)} = f_\lambda(\alpha^{(t+1)}, \beta^{(t+1)}, \gamma^{(t+1)})$ and set $t \leftarrow t + 1$

Report

Check out the HiC report for:

- Active Set coordinate descent algorithm.
- Details about the closed form updates.

Outline

Future Directions

In process...

- Coding the Coordinate Descent algorithms.
- Exploring other stopping criteria.

Questions:

- How to design simulations that mimick HiC data?