DOUBLY PENALIZED LOGISTIC REGRESSION FOR GENOMEWIDE ASSOCIATION STUDIES WITH LINEARLY STRUCTURED GENETIC NETWORKS

Li X. and Wu T. T.

Department of Epidemiology and Biostatistics, University of Maryland College Park, MD

NSF Expeditions in Computing

INTRODUCTION

- Spatial effects should be taken into consideration when studying the association between genetic network and diseases
- This research aims to integrate linear structures of genetic networks into genomewide analysis studies (GWAS)
- Lasso penalized logistic regression is suited for continuous model selection for individual genes in case-control disease gene mapping, especially when the number of predictor variables far exceeds the number of observations. However, it totally ignores the network structures
- In order to incorporate the network structure, new penalty functions need to be designed

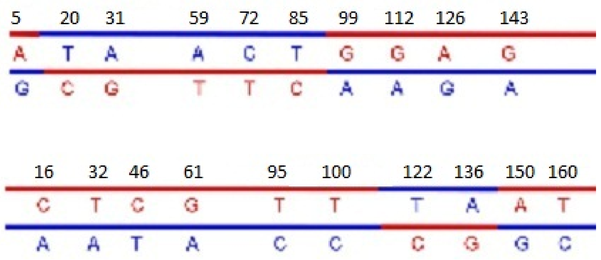


Figure 1: SNP location map in the Candida genome. The numbers are genomic locations showing different distances between SNPs.

OBJECTIVES

- To incorporate the linear structure of genetic networks into the model
- To achieve variable selection in individual genetic markers while considering the map distances of the marker in the genome
- To handle the underdetermined setting where the number of parameters far exceeds the number of observations
- *Objective Function*

$$f(\beta) = -L(\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} \omega_j |\beta_j - \beta_{j-1}|$$

where

  - $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$: vector of regression coefficients
  - $L(\beta)$: loglikelihood function of logistic regression
  - $|\beta_j|$: lasso (least absolute shrink and selection operater) penalty for individual SNP selection
  - $|\beta_j - \beta_{j-1}|$: fused lasso term for adjacent SNP pairs
  - λ_1 and λ_2: tuning parameters controlling for the strength of selection
  - $\omega_j = 1/(d_j - d_{j-1})$: weights for SNP pairs, where d_j is the map distance of SNP j

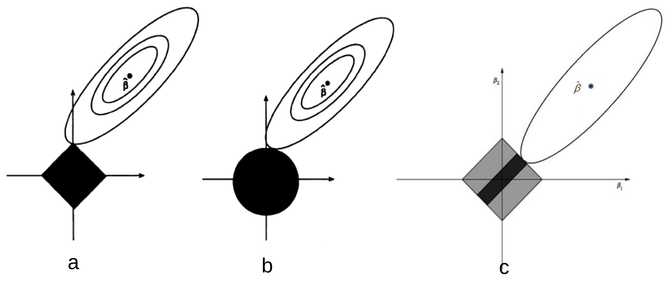


Figure 2: Estimation in (a) lasso, (b) ridge and (c) the fused lasso regression

- Fused lasso term $|\beta_j - \beta_{j-1}|$ penalizes differences of adjacent coefficients
- incorporates the linear structure of SNPs and encourages adjacent SNPs with closer distance to have similar values

METHODS

- *Reformulated Objective Function*

$$h_{L,\gamma}(\beta) = -\{L(\gamma) + (\beta - \gamma)^T L'(\gamma)\} + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} \omega_j |\beta_j - \beta_{j-1}| + \frac{L}{2} \|\beta - \gamma\|^2$$

- *EFLA steps*

  - Initialize: $\gamma_1 = \gamma_0, a_{-1} = 0, a_0 = 1, L = L_0$
  - Loop and update
    1. $b_i = \frac{a_{i-2}-1}{a_{i-1}}, s_i = \gamma_i + b_i(\gamma_i - \gamma_{i-1})$
    2. Find the smallest $L = L_{i-1}, 2L_{i-1}, \ldots$ such that $h(\gamma_{i+1}) \le h_{L,s_i}(\gamma_{i+1})$, where $\gamma_{i+1} = \text{argmin}_\theta h_{L,s_i}(\beta)$
    3. Set $L_i = L, a_{i+1} = \frac{1+\sqrt{1+4a^2{}_i}}{2}$
    4. End loop if

$$h(\gamma_k) - h(\gamma_{k+1}) \le \frac{2\max(2\tilde{L}, L_0)\|\gamma_0 - \gamma_{k+1}\|^2}{k^2}$$

    Where $\tilde{L}$ is the Lipschitz continuous gradient of $L(\beta)$
  - γ_k is the optimal solution

- *Determine Tuning* Parameters: Grid search on (λ_1, λ_2)

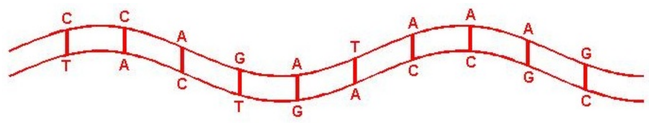
SIMULATED DATA ANALYSIS

- *Input Data*

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j, \text{ where}$$

  - $\beta = (2, 1, 0.5, 2, 0.5, 2.5, 0 \ldots, 0)$
  - Predictors are correlated

- *Set the Distances*
  - Equal distance ➔ no linear structure considered



  - Map distance ➔ linear structure considered

- ***Simulation Results***

		True Predictors						
Distance	(λ_1, λ_2)	x_1	x_2	x_3	x_4	x_5	x_6	Noise
$(n, p) = (150, 500)$								
Equal	(0.03,0.09)	50	50	50	50	50	50	2.84 (1.46)
Map	(0.07,0.06)	50	50	50	50	50	50	2.64 (1.12)
$(n, p) = (200, 1000)$								
Equal	(0.02,0.01)	50	50	50	50	50	50	9.32 (5.08)
Map	(0.04,0.02)	50	50	50	50	50	50	6.78 (3.58)

Table 1: Selection frequencies of true and noise predictors in 50 simulations

PANCREATIC CANCER DATA ANALYSIS

- Pancreatic cancer is the fourth leading cause of cancer death in the United States with a five-year survival rate of only 3%
- Pancreatic cancer data will be using this doubly penalized method
- The data was obtained from Biobank Japan at the Institute of Medical Science, The University of Tokyo as well as National Cancer Center Hospital.
- Cases=991, controls=5209, SNPS=420,236. Other covariates include age, gender, smoking status, etc.

KEY REFERENCE

- Liu, J., L. Yuan, et al. (2010). "An efficient algorithm for a class of fused lasso problems." KDD '10 Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (underline the proceeding)
- Tibshirani, R., M. Saunders, et al. (2005). "Sparsity and smoothness via the fused lasso." Journal of the Royal Statistical Society: Series B **67**(1): 91-108.
- Wu, T. T., Y. F. Chen, et al. (2009). "Genome-wide association analysis by lasso penalized logistic regression." Bioinformatics **25**(6): 714.
- Wu, T. T. and K. Lange (2008). "Coordinate descent algorithms for lasso penalized regression." The Annals of Applied Statistics **2**(1): 224-244.