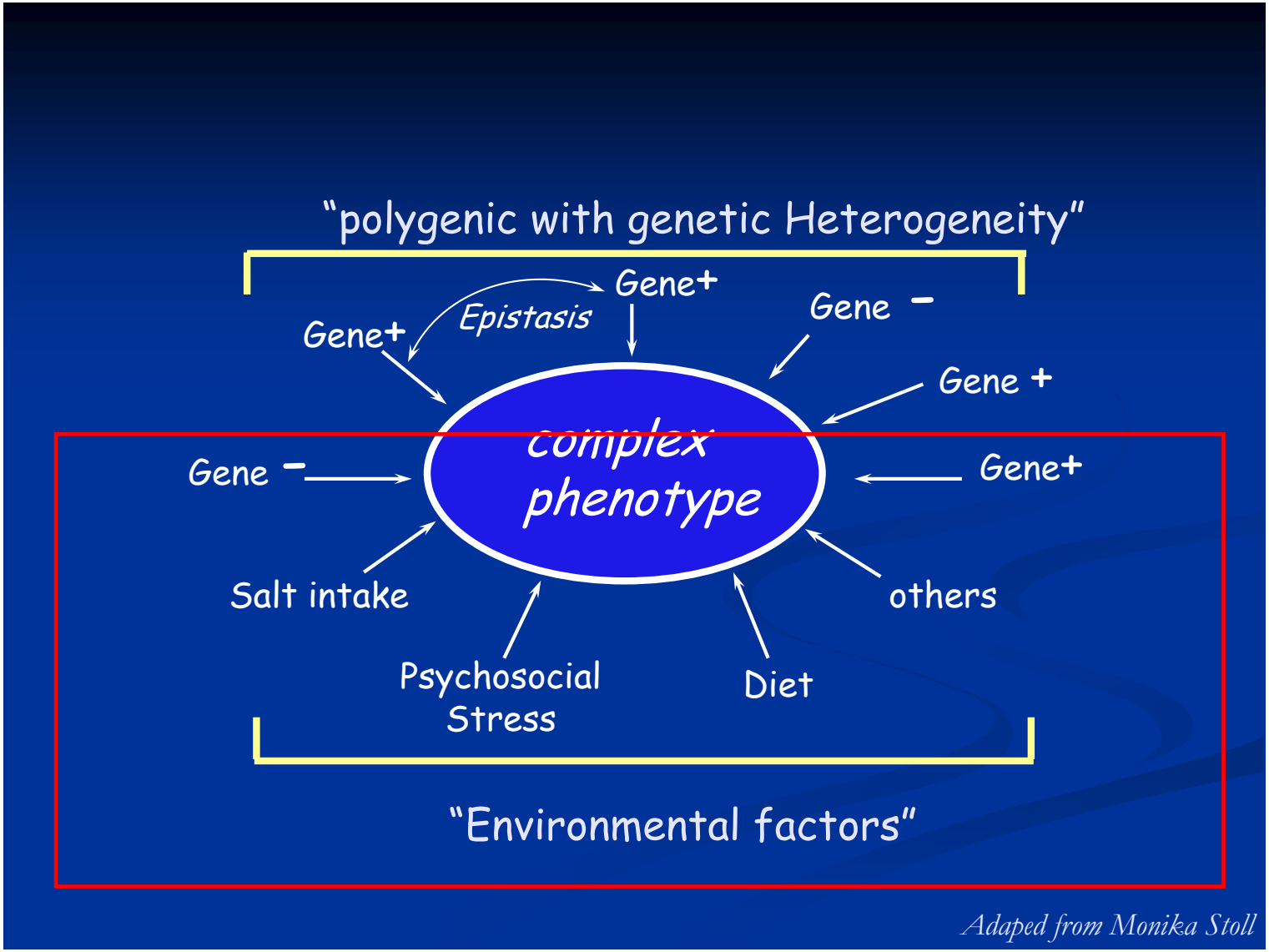
Robust integration of Multi-Omics Data for Gene-Environment Interactions

Xi Lu

Department of Statistics
Kansas State University

June 12, 2019

Heterogeneity of Complex Diseases



Background

- Nonlinear gene–environment interaction:
  - Approach: hypothesis testing.
  - Ma et. al.(2011); Wu and Cui (2013).
  - Not robust.
- Robust gene–environment interaction:
  - Approach: marginal penalization.
  - Shi et. al. (2014), Chai et. al. (2015).
  - Not able to identify nonlinear effects.
- Motivation:
  - Robust joint modelling approach to detect different types of gene–environment interactions in a unified framework.

The Model

- Consider the first level:
- Denote $Z = (Z_1, \ldots, Z_s)$ as the $n \times s$ data matrix of regulators (CNAs, microRNAs, methylations).
- $X = Z\eta + W$.
- $X = (X_1, \ldots, X_p)$ as the $n \times p$ data matrix of GEs, $W = (W_1, \ldots, W_p)$ is an $n \times p$ matrix of random errors.
- X contains 2 parts: regulated GEs and unregulated GEs.

The Model

- Consider the second level:
- Denote X^{reg} as the GEs regulated by Z and X^{unreg}as the GEs not regulated by Z, where $X^{reg} = (X_1, \ldots, X_{p_1})$, $X^{unreg} = (X_t, \ldots, X_{p_2})$, where $p_1 + p_2 - t = p$.
- y is the outcome variables with length n, $E = (E_1, \ldots, E_p)$ is $n \times q$ data matrix of environment factors, $\alpha = (\alpha_1, \ldots, \alpha_q)^{'}$
-

$$y = \sum_{l=1}^{q} \alpha_l E_l + \sum_{m=1}^{p_1} (\gamma_m X_m + \sum_{l=1}^{q} \beta_{ml} X_m E_l) + \sum_{m=t}^{p_2} (\gamma_m^{'} X_m + \sum_{l=1}^{q} \beta_{ml}^{'} X_m E_l) + \varepsilon \tag{1}$$

The LAD-LASSO Model

- For the first level, we use LAD-LASSO to find important gene expressions.
- The LAD loss function $L(\eta) = \sum_{i=1}^{p} |X_i - \sum_{j=1}^{s} Z_{ij}\eta_{ij}|$.
- We use LASSO to do penalization on the LAD loss function.

The AFT Model

- For the second level, we use AFT Model.
- Denote T as the logarithms of the survival time and C as the censoring time.
- The AFT model assumes that

$$
\begin{aligned}
T =& \sum_{l=1}^{q} \alpha_l E_l + \sum_{m=1}^{p_1} \gamma_m X_m + \sum_{m=1}^{p_1}\sum_{l=1}^{q} \beta_{ml} X_m E_l + \sum_{m=t}^{p_2} \gamma_m^{'} X_m + \sum_{m=t}^{p_2}\sum_{l=1}^{q} \beta_{ml}^{'} X_m E_l \\
&+\varepsilon \\
=& \sum_{l=1}^{q} \alpha_l E_l + \sum_{m=1}^{p_1} (\gamma_m X_m + \sum_{l=1}^{q} \beta_{ml} X_m E_l) + \sum_{m=t}^{p_2} (\gamma_m^{'} X_m + \sum_{l=1}^{q} \beta_{ml}^{'} X_m E_l) + \varepsilon \\
=& \sum_{l=1}^{q} \alpha_l E_l + \sum_{m=1}^{p_1} b_m^T U_m + \sum_{m=t}^{p_2} b_m^{'T} U_m + \varepsilon \\
=& \alpha E + b^T U + b^{'T} U^{'} + \varepsilon
\end{aligned}
\tag{2}
$$

The AFT Model

- where $b_m = (\gamma_m, \beta_{m1}, \ldots, \beta_{mq})^T$ and $U_m = (X_m, X_m E_1, \ldots, X_m E_q)^T$. Denote $\alpha = (\alpha_1, \ldots, \alpha_q)^T$, $b = (b_1^T, \ldots, b_{p_1}^T)^T$, $b^{'} = (b_t^{'T}, \ldots, b_{p_2}^{'T})^T$ and $U = (U_1^T, \ldots, U_{p_1}^T)^T$, $U^{'} = (U_t^T, \ldots, U_{p_2}^T)^T$. b_m and U_m represent all effects – main and interactions with respect to mth genetic variant.

The AFT Model

- We use subscripts i to denote the ith subject for the n independent subjects.
- Under right censoring, denote C_i as the censoring time and $\delta_i = 1\{T_i \leq C_i\}$ as the censoring indicator.
- We observe $(Y_i, \delta_i, E_{il}, U_{im})$ where (E_{il}, U_{im}) are the associated covariates with Y_i.
- Without loss of generality, assume that $(Y_i, \delta_i, E_{il}, U_{im})$s have been sorted according to Y_i in an ascending order.

The AFT Model

- Stute (1993) proposed the weighted least square estimation approach.
- Stute's estimator is the minimizer of the loss function

$$\sum_{i=1}^{n} d_{ni}(Y_i - \sum_{l=1}^{q} \alpha_l E_{il} - \sum_{m=1}^{p_1} b_m^T U_{im} - \sum_{m=t}^{p_2} b_m^{'T} U_{im})^2 \tag{3}$$

  where the Kaplan–Meier weights d_{ni} are defined as

$$d_{n1} = \frac{\delta_1}{n}, d_{ni} = \frac{\delta_i}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_j}, i = 2, \ldots, n \tag{4}$$

- One contaminated Y_i will lead to severely biased model estimation if $d_{ni} \neq 0$.

Robust Loss Function

- To accommodate the potential contamination in survival outcome, we propose the robust objective function.

- The weighted LAD loss function

$$L(\alpha, b) = \sum_{i=1}^{n} d_{ni} |Y_i - \sum_{l=1}^{q} \alpha_l E_{il} - \sum_{m=1}^{p_1} b_m^T U_{im} - \sum_{m=t}^{p_2} b_m^{'T} U_{im}| \qquad (5)$$

where (E_{il}, X_i, U_{im}) are the associated covariates with ordered Y_i's.

Issues and Solutions

- Nature of high–dimensionality.
- Not all genetic factors have interactions with the environment factors.
- Solution(selection):
  - Boosting;
  - Bayesian approaches;
  - **Penalization**.

Robust Penalization

- For the robust G×E model, consider the robust penalization with

$$Q_1(\alpha, b) = L(\alpha, b) + \lambda_1 \sum_{m=1}^{p} w_m \|B_m\|_2 + \lambda_2 \sum_{m=1}^{p} \sum_{l=1}^{q+1} w_{m,l} |B_{ml}| \tag{6}$$

  where w_m and $w_{m,l}$ are the adaptive weights corresponding to group and individual level penalties. And $(B_1, \ldots, B_p) = (b_1, \ldots, b_{p_1}, b_t^{'}, \ldots, b_{p_2}^{'})$.

- The rationale: group and individual level selection.

- LAD-SGL: Least Absolute Deviation-Sparse Group LASSO.

Computational Algorithms

- We first consider approximating $||B_m||_2$ by

$$\begin{aligned} ||B_m||_2 &\approx ||B_m^{(0)}||_2 + ||B_m^{(0)}||_2^{-1}|B_m^{(0)}|^\top(|B_m| - |B_m^{(0)}|) \\ &= ||B_m^{(0)}||_2^{-1}\sum_{l=1}^{q+1}|B_{m,l}^{(0)}||B_{m,l}| \end{aligned} \tag{7}$$

- With the above approximation, (6) changes to

$$\begin{aligned} &L(\alpha, B) + \lambda_1\sum_{m=1}^{p} w_m||B_m^{(s-1)}||_2^{-1}\sum_{l=1}^{q+1}|B_{m,l}^{(s-1)}||B_{m,l}| + \lambda_2\sum_{m=1}^{p}\sum_{l=1}^{q+1} w_{m,l}|B_{ml}| \\ &= L(\alpha, B) + \sum_{m=1}^{p}\{\lambda_1 w_m||B_m^{(s-1)}||_2^{-1}\sum_{l=1}^{q+1}|B_{m,l}^{(s-1)}||B_{m,l}| + \sum_{l=1}^{q+1}\lambda_2 w_{m,l}|B_{ml}|\} \end{aligned} \tag{8}$$

where $w_m = ||\tilde{B}_m||_2^{-2}$ with the initial value $\tilde{B}_m$.

Computational Algorithms

- With the assistance of slack variables, this optimization problem can be casted as a linear programming problem:

$$\begin{aligned} \underset{\xi, B_m}{\text{minimize}} \quad & \frac{1}{n}\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-) + \lambda_1 \sum_{l=1}^{q+1} c_{ml}(B_{ml}^+ + B_{ml}^-) \\ \text{subject to} \quad & \xi_i^+ - \xi_i^- = Y_i - \sum_{l=1}^{q} \alpha_l E_{il} - \sum_{m=1}^{p} B_m^T U_{im}; i = 1, \ldots, n, \\ & \xi_i^+ \geq 0, \xi_i^- \geq 0; i = 1, \ldots, n, \end{aligned} \tag{9}$$

where

$c_{ml} = \lambda_1 w_m ||B_m^{(s-1)}||_2^{-1} |B_{m,l}^{(s-1)}| + \lambda_2 w_{ml}, l = 1, \ldots, q+1$

Computational Algorithms

- With fixed tunings, the GCD algorithm proceeds as follows.

  (1) Initialize $\tilde{\alpha}$ and $\tilde{B}$ in the penalized robust objective function using LASSO.

  (2) At the kth iteration, compute $B_m^{(k)}$ via the linear programming problem for $m = 1, \ldots, p$.

  (3) Update $\alpha^{(k)}$ by minimizing the weighted LAD loss function after fixing B as $B^{(k)}$.

  (4) Iterate steps (2) and (3) until convergence.

Simulation

- Data generating model:
  level 1 : $X = Z\eta + W$
  level 2 : $y = \sum_{l=1}^{q} \alpha_l E_l + \sum_{m=1}^{p_1} (\gamma_m X_m + \sum_{l=1}^{q} \beta_{ml} X_m E_l) + \sum_{m=t}^{p_2} (\gamma_m^{'} X_m + \sum_{l=1}^{q} \beta_{ml}^{'} X_m E_l) + \varepsilon$
- n=1000 and $p = 200$.
- The coefficients:
  - For first level, the coefficients of X_1 and X_3 are nonzero.
  - Coefficients of first 10 environment factors are 1.8.
  - The coefficients of first 10 gene expressions are generated from $Unif[1.8, 2.2]$.
  - For $G \times E$ interactions, the coefficients of interactions between X_1, X_3, X_5 and $E_1.E_2, E_3, E_4, E_5$ are generated from $Unif[1.8, 2.2]$.
  - All the rest of the coefficients are set as 0.

Simulation

- Simulate Z as the regulators from multivariate normal distribution and E as environment factors from multivariate normal distribution.
- X as the gene expressions generate from $Z\eta + W$.
- The random errors are generated from:
  $(1)N(0,1)$(Error 1); $(2)0.8N(0,1) + 0.2\text{Cauchy}(0,1)$ (Error 2); (3) $0.7N(0,1) + 0.3\text{Cauchy}(0,1)$ (Error 3);
- Three approaches
  - A1: level 1 : LAD LASSO, level 2: robust SGL model;
  - A2: level 1: LASSO, level 2: non–robust LASSO survival model;
  - A3: level 1: LAD-LASSO, level 2: robust LASSO survival model.

Identification results

Table: Identification results for simulation data. mean(sd) based on 100 replicates. TP/FP: true/false positives.

		Total		No Interactions		Linear Interaction	
		TP	FP	TP	FP	TP	FP
Error 1	A1	14.65(1.14)	9.15(1.89)	5.95(0.51)	3.15(0.82)	8.7(0.73)	6(1.45)
	A2	17.55(5.03)	3.15(2.18)	5.4(1.76)	0.15(0.48)	12.15(3.55)	3(2.10)
	A3	10.6(5.35)	22.45(15.67)	3.95(1.79)	3.2(2.94)	6.65(3.74)	19.25(13.28)
Error 2	A1	11.6(4.97)	10.25(8.44)	4.65(2.23)	3.5(2.78)	6.95(2.83)	6.75(5.81)
	A2	11.95(3.45)	14.25(31.93)	2.6(1.46)	0.7(2.05)	9.35(2.3)	13.55(29.93)
	A3	10.65(2.41)	23.15(15.93)	3.75(0.72)	2.55(1.96)	6.9(2.07)	20.6(14.39)
Error 3	A1	14.2(2.3)	9.4(3.8)	6.4(0.52)	2.6(1.18)	7.73(2.15)	6.8(3.17)
	A2	6.4(2.37)	8.1(19.5)	0.5(0.83)	0.35(0.99)	5.9(1.05)	7.75(18.54)
	A3	7.15(1.67)	16.7(15.77)	3.4(1.05)	1.95(1.7)	3.75(2.07)	14.75(15.28)

Summary

- A_1 outperforms all the other two approaches;
- Robust approaches performs well when heavy–tailed errors exist;
- Similar patterns have been observed across differents scenarios.

Applications to Lung Cancer Data

- Data from TCGA-LUSC(The Cancer Genome Atlas Program-Lung Squamous Cell Carcinoma)
- Top 200 genes are chosen for downstream analysis (n=344).
- Response variable: time to death.
- Four environment factors: pathologic tumor stage, gender, race, and smoking pack year.

Results

Table: Analysis of the lung cancer data using approach A1.

Gene	Main Effects	Stage	Gender	Race	Smoking
COL5A3				-0.011	
PRRX2		-0.029			
MUCL1		-0.098			
STK40					-0.206
PARD6G	0.098	-0.515			
RPTN					-0.230
IBSP		1.080			
RNASE7		-0.071			
WBP2NL		0.233			

- The coefficients of environment factors pathologic tumor stage, gender, race, smoking pack year are -0.209, -0.436, 0.012, -0.545.

Results

Table: Analysis of the lung cancer data using approach A2.

Gene	Main Effects	Stage	Gender	Race	Smoking
PYGB	-0.012				
LHX8	-0.006				
ENTPD6	-0.121				
TRIM55					-0.033
TPPP3	-0.030				
PAX1	-0.124				
PLEKHA6				-0.041	
EDN2					-0.069
PRRX2		-0.172			
ARHGEF18					-0.018
STK40					-0.012
PARD6G		0.1777			
ZNF532				-0.020	
RPTN				-0.006	
FHDC1		0.032			
PHPT1				0.102	
ART3			-0.028		

- The coefficients of environment factors pathologic tumor stage, gender, race, smoking pack year are -0.234, 0.137, -0.180, -0.048.

Results

Table: Analysis of the lung cancer data using approach A3.

Gene	Main Effects	Stage	Gender	Race	Smoking
PYGB	-0.615				
TREM1	0.074	-0.026			
ENTPD6	0.027				
NACC2				0.073	
GZF1		0.032			
PLEKHA6				-0.053	
FKBP8					-0.002
ANGPT2					-0.343
UBE4B	0.062				
MIER2				0.072	-0.016
WBP2NL				-0.004	
TEX14	-0.004				

- The coefficients of environment factors pathologic tumor stage, gender, race, smoking pack year are -0.373, 0.216, -0.024, -0.173.

Summary

- Propose a robust SGL penalization approach to detect gene-environment interactions.
- A flexible robust-parametrics modelling of complex interaction effects.
- Extensive simulation studies under different settings indicate the advantage of the method over the alternatives.
- The findings in case study are important for generating biological hypothesis for future lab validation.

Thank you for your attention!