PROLONG

Penalized Regression On Longitudinal Omics data with Network and Group lasso constraints

sb2643@cornell.edu

Steve Broll[1] advised by Martin T. Wells[1] Sumanta Basu[1] and Myung-Hee Lee[2]

[1] *Cornell University*
[2] *Weill Cornell Medicine*

Introduction

We have:

- Longitudinal measurements for some continuous outcome of interest and for -omics variables with only a few time points
- Large amount of variables with a relatively small number of subjects (p >> n)

We want to:

- Identify -omics variables that co-vary with the outcome
- Overcome time dependence, low signal, and high between-subject variability
- Incorporate correlation of the variables into the model penalty

General Model Idea

- Take first difference of the data to deal with observed temporal dependence
- Stack our $t-1$ first differenced values of Y
- Set up design matrix so that each first differenced Y value is regressed on all prior first differenced values of X to account for potential lags
- Apply network and group lasso penalties to induce sparsity while utilizing correlation and inherent group structure

Reshaping the Data

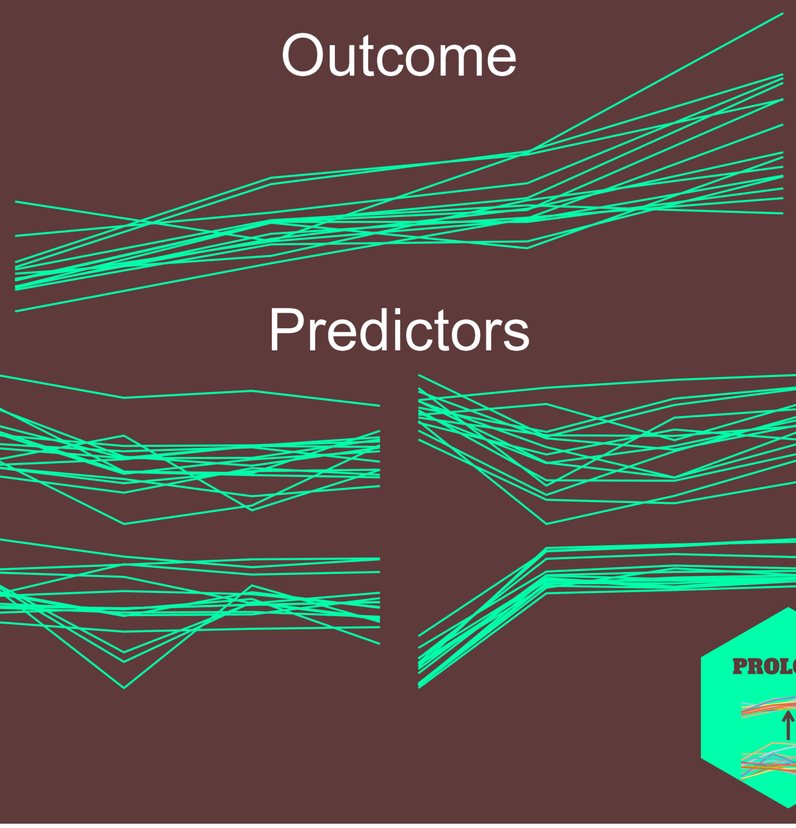
$$\tilde{Y} = \begin{bmatrix} \tilde{Y}_{11} & \cdots & \tilde{Y}_{1T} \\ & \vdots & \\ \tilde{Y}_{n1} & \cdots & \tilde{Y}_{nT} \end{bmatrix} \rightarrow \begin{bmatrix} \Delta\tilde{Y}_{11} & \cdots & \Delta\tilde{Y}_{1(T-1)} \\ & \vdots & \\ \Delta\tilde{Y}_{n1} & \cdots & \Delta\tilde{Y}_{n(T-1)} \end{bmatrix}$$

$$\rightarrow Y = \begin{bmatrix} \Delta\tilde{Y}_{11} & \cdots & \Delta\tilde{Y}_{n1} & \cdots & \Delta\tilde{Y}_{1(T-1)} & \cdots & \Delta\tilde{Y}_{n(T-1)} \end{bmatrix}^\top$$

$$\tilde{X}^{[j]} = \begin{bmatrix} \tilde{X}^{[j]}_{11} & \cdots & \tilde{X}^{[j]}_{1T} \\ & \vdots & \\ \tilde{X}^{[j]}_{n1} & \cdots & \tilde{X}^{[j]}_{nT} \end{bmatrix} \rightarrow \begin{bmatrix} \Delta\tilde{X}^{[j]}_{11} & \cdots & \Delta\tilde{X}^{[j]}_{1(T-1)} \\ & \vdots & \\ \Delta\tilde{X}^{[j]}_{n1} & \cdots & \Delta\tilde{X}^{[j]}_{n(T-1)} \end{bmatrix}$$

$$\rightarrow X^{[j]} = \left[\begin{array}{c|ccc|c|ccc} \Delta\tilde{X}^{[j]}_{11} & & & & & & & \\ \vdots & & 0 & & 0 & & 0 & \\ \Delta\tilde{X}^{[j]}_{n1} & & & & & & & \\ \hline & \Delta\tilde{X}^{[j]}_{11} & \cdots & \Delta\tilde{X}^{[j]}_{12} & & & & \\ 0 & & \vdots & & 0 & & 0 & \\ & \Delta\tilde{X}^{[j]}_{n1} & \cdots & \Delta\tilde{X}^{[j]}_{n2} & & & & \\ \hline 0 & & 0 & & \ddots & & 0 & \\ \hline & & & & & \Delta\tilde{X}^{[j]}_{11} & \cdots & \Delta\tilde{X}^{[j]}_{1(T-1)} \\ 0 & & 0 & & 0 & & \vdots & \\ & & & & & \Delta\tilde{X}^{[j]}_{n1} & \cdots & \Delta\tilde{X}^{[j]}_{n(T-1)} \end{array}\right]$$

Given a longitudinal, continuous clinical outcome, PROLONG can select correlated, longitudinal -omics predictors for high-dimensional data



Penalties

- Group lasso is used to account for the fact that each variable is represented multiple times in the model
- The network-constraint via Laplacian matrix $\mathcal{L}$ allows us to incorporate the pairwise absolute correlations between variables as graph edge weights

$$L(\lambda_1, \lambda_2, \beta) = (Y - X\beta)^\top (Y - X\beta) + \lambda_1 \sum_{j=1}^{p} p_j \left\| \beta^{[j]} \right\|_2 + \lambda_2 \beta^\top \mathcal{L} \beta$$

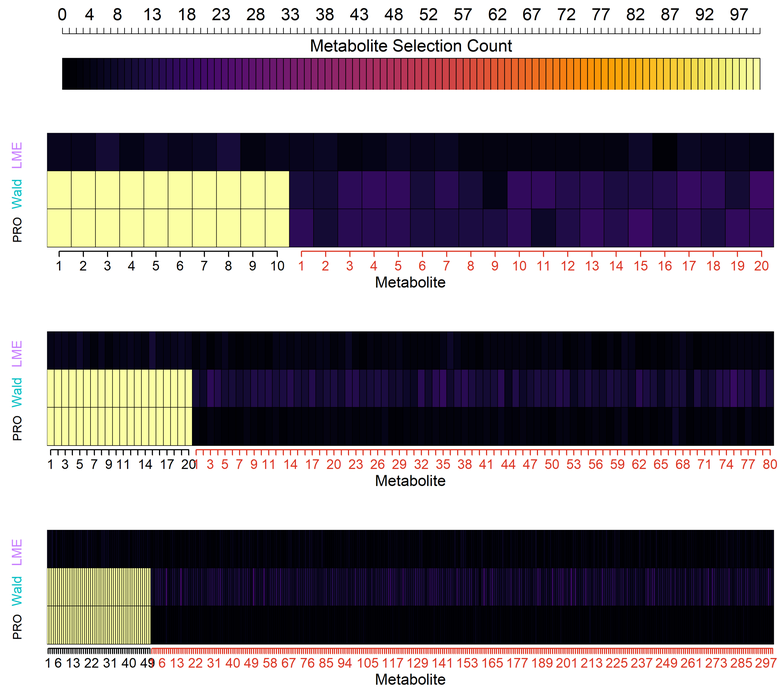
- To minimize $L(\lambda_1, \lambda_2, \beta)$ we create artificial dataset $(\mathcal{Y}, \mathcal{X})$ by appending a 0-vector to Y and $\mathcal{S}^\top$ to X, where $\mathcal{S} = \Gamma D^{1/2}$ given $\mathcal{L} = \Gamma D \Gamma^\top$

$$\mathcal{X} = (1+\lambda_2)^{-1/2} \begin{bmatrix} X \\ \sqrt{\lambda_2}\mathcal{S}^\top \end{bmatrix}, \qquad \mathcal{Y} = \begin{bmatrix} Y \\ 0 \end{bmatrix}$$

- We solve for β using group lasso then adjust by $1/\sqrt{1+\lambda_2}$ to get our estimate $\hat{\beta}$
- λ_2 is selected via MLE, λ_1 via cross-validation

Simulations

- 15 subjects and 4 time points like our motivating data
- Each change in Y only depends on $X_{j2} - X_{j1}$ for our target variables and has no relation to the noise variables
- Coefficients and variances are chosen so that the total signal-to-noise ratio (SNR) ranges incrementally from 1 to 2
- We evaluate models by their sensitivity and specificity across 100 simulations in each scenario
- We compare PROLONG, Wald tests using the same reshaped first-differenced data at an FDR threshold of 0.05, and standard longitudinal mixed effects models at an FDR threshold of 0.05



Real Data

- Using PROLONG, we selected 45 metabolites out of the 352 in the dataset
- All selected metabolites were identified as targets by our collaborators or via EDA

Future Work

- Extension to other continuous -omics variables is immediate
- Further investigation into microbiome integration
- Incorporation of RNA-seq variables