PROLONG

Penalized Regression on Longitudinal multi-Omics Data with Network and Group Lasso Constraints

Steve Broll [1]
Advised by Sumanta Basu, [1] Martin Wells [1], and Myung Hee Lee [2]

[1]Cornell University

[2]Weill Cornell Medicine

Motivation in Short

Clinician gives you a longitudinal clinical outcome, along with hundreds of longitudinal -omics variables, and asks

Which variables co-vary with the outcome?

More Details for Motivation

We have:

- Longitudinal measurements for some continuous phenotype and for -omics variables with only a few time points
- Large amount of variables with relatively small number of subjects

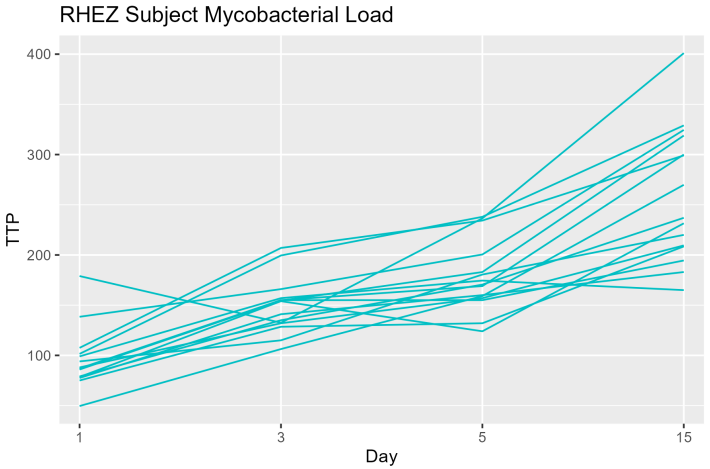
We want to:

- Identify -omics variables that co-vary with the phenotype
- Overcome time dependence, low signal, and high subject variability
- Incorporate correlation of the variables

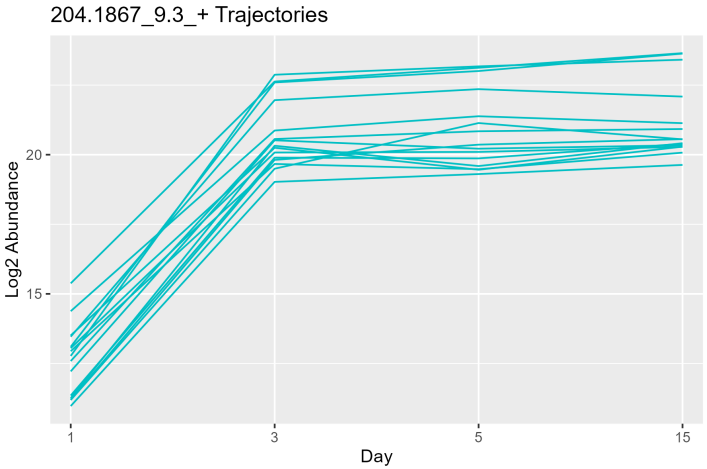
Tuberculosis Data

- 15 subjects, TB patients treated with RHEZ [rifampin (R), isoniazid (H), ethambutol (E), and pyrazinamide (Z)]
- Mycobacterial load measured by Time to Positivity (TTP)
- 352 metabolites with complete measurements for $>80\%$ of subjects, softImpute used for missing values
- 4 time points, days 1, 3, 5, 15
- Additionally, we have microbiome and RNAseq data [1] for days 1 and 15 - more on this later

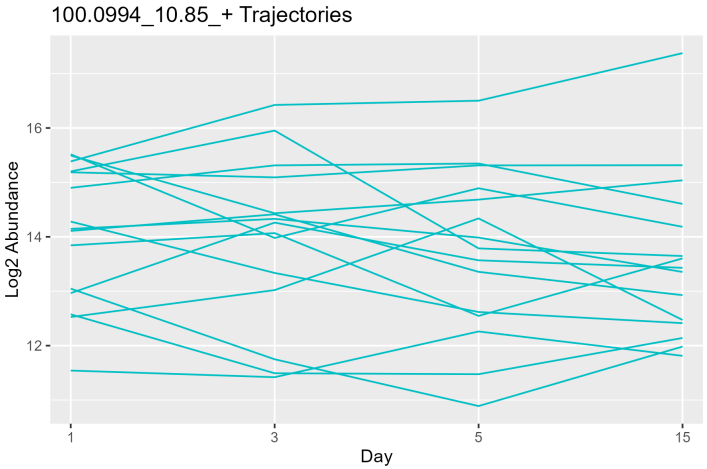
TB Clinical Outcome



TB Example Metabolite 1



TB Example Metabolite 2



General Model Idea

- Take first difference of the data to deal with observed temporal dependence
- Stack our $t-1$ first differenced value of X and Y so we have

$$Y = |Y_4 - Y_3 \qquad Y_3 - Y_2 \qquad Y_2 - Y_1|^T$$

And for each variable j we have

$$X_j = |X_{j4} - X_{j3} \qquad X_{j3} - X_{j2} \qquad X_{j2} - X_{j1}|^T$$

- 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

Vectorized Y

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

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

Moving X from Tensor to Matrix

Cornell University

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

$$\rightarrow X^{[j]} = \left[\begin{array}{c|c|c|c} \begin{matrix} \Delta\tilde{X}^{[j]}_{11} \\ \vdots \\ \Delta\tilde{X}^{[j]}_{n1} \end{matrix} & \multicolumn{1}{c|}{0} & 0 & \multicolumn{1}{c}{0} \\ \hline 0 & \begin{matrix} \Delta\tilde{X}^{[j]}_{11} & \Delta\tilde{X}^{[j]}_{12} \\ \vdots & \\ \Delta\tilde{X}^{[j]}_{n1} & \Delta\tilde{X}^{[j]}_{n2} \end{matrix} & 0 & 0 \\ \hline 0 & 0 & \ddots & 0 \\ \hline 0 & 0 & 0 & \begin{matrix} \Delta\tilde{X}^{[j]}_{11} & \cdots & \Delta\tilde{X}^{[j]}_{1(T-1)} \\ & \vdots & \\ \Delta\tilde{X}^{[j]}_{n1} & \cdots & \Delta\tilde{X}^{[j]}_{n(T-1)} \end{matrix} \end{array}\right]_{n(T-1) \times T(T-1)/2}$$

Group Lasso Laplacian Penalty

Given our stacked response vector Y and design matrix X we seek to minimize

$$(Y - X\beta)^T(Y - X\beta) + \lambda_1 \sum_{j=1}^{p} \left\| \boldsymbol{\beta}_{(j)} \right\|_2 + \lambda_2 \beta^T L \beta,$$

- λ_1 is the tuning parameter for our group lasso penalty, where each group j corresponds to all of the representations in the design matrix of the jth variable
- λ_2 is the tuning parameter for the network penalty
- L is the Laplacian matrix for the weighted graph where the edge weights between each pair of variables are their absolute correlation

Nice Properties of this Penalty

- Each variable is represented multiple times in the model, but the group lasso penalty results in either all zero or all non-zero coefficients for the representations of each variable, helping interpretability
- If two variables are highly correlated, and one is a strong enough predictor to be selected, the other variable is likely to be selected as well
- If two variables are identical, either both will be selected and have the same coefficient or neither will be selected

Models Compared

- Linear Mixed Effects Model
- Wald test on the Δ scale with each $X^{[j]}$
- PROLONG

In the following simulations, the univariate models are evaluated at different FDR thresholds and compared to PROLONG

$$x_1 \sim N(\mu, \Sigma_X); \quad \mu \sim U(10, 20), \Sigma_X = \mathsf{diag}(\sigma_1, \ldots, \sigma_p), \sigma_j \sim U(1, 2)$$

$$x_2 \sim x_1 + N(d\mu, \Sigma_X); \qquad d\mu = (5, \ldots, 10, 0, \ldots, 0)$$

$$x_t \sim x_{t-1} + N(0, \Sigma_X) \quad t \in 3, 4$$

$$y_1 \sim N(15, 5); \qquad y_2 = N(y_1 + \beta(x_2 - x_1), 5)$$

$$y_3 \sim N(y_2 + \beta(x_3 - x_2) + \beta(x_2 - x_1), 5)$$

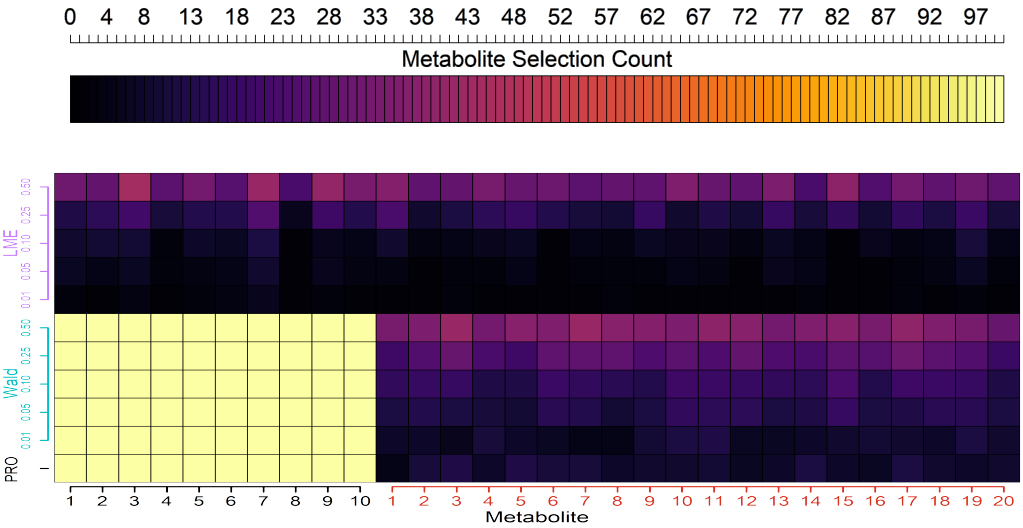
$$y_4 \sim N(y_3 + \beta(x_4 - x_3) + \beta(x_3 - x_2) + \beta(x_2 - x_1), 5)$$

$$\beta = (1/3, 1/3, \ldots, 0, \ldots, 0)$$

SNR ranging from 1 to 2 in targets

Performance in Simulations

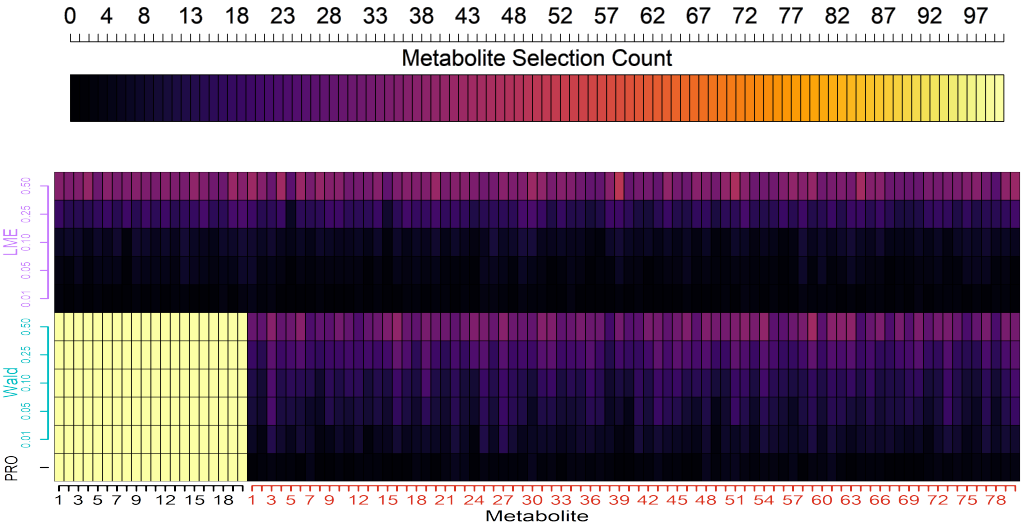
Uncorrelated Simulated Variables



Performance in Simulations

Uncorrelated Simulated Variables

Cornell University



Same as previous scenario, but with

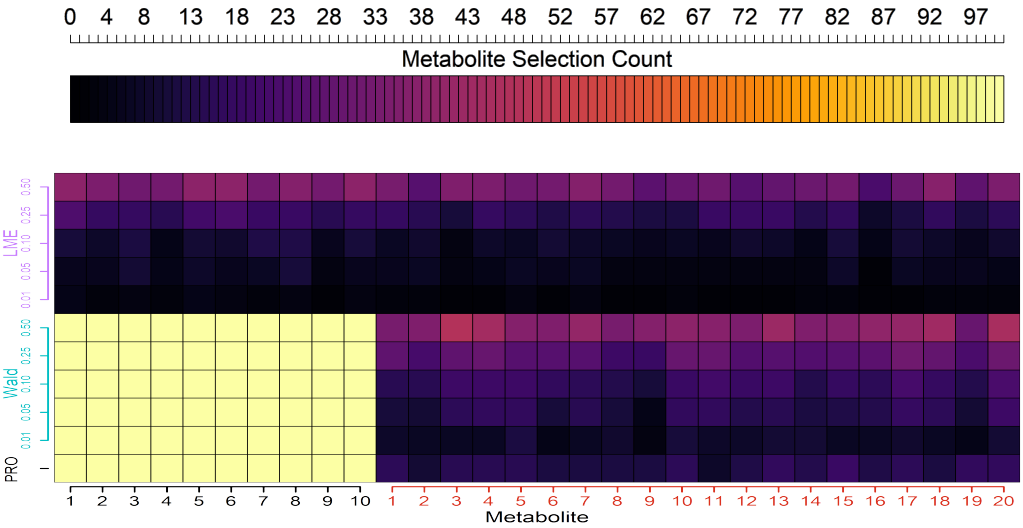
$$\Sigma_X = \begin{bmatrix} \Sigma_C & 0 \\ 0 & \Sigma_\epsilon \end{bmatrix}$$

where Σ_C generated so that the variances are in the same range as in Σ_ϵ and the covariances correspond to correlations uniformly drawn from $(-1, 1)$

Performance in Simulations

Correlated Simulated Variables

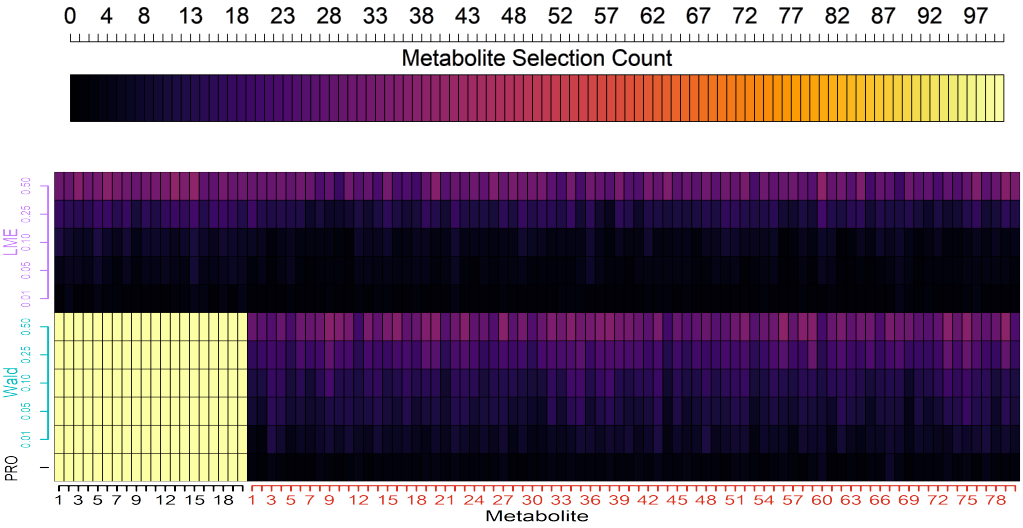
Cornell University



Performance in Simulations

Correlated Simulated Variables

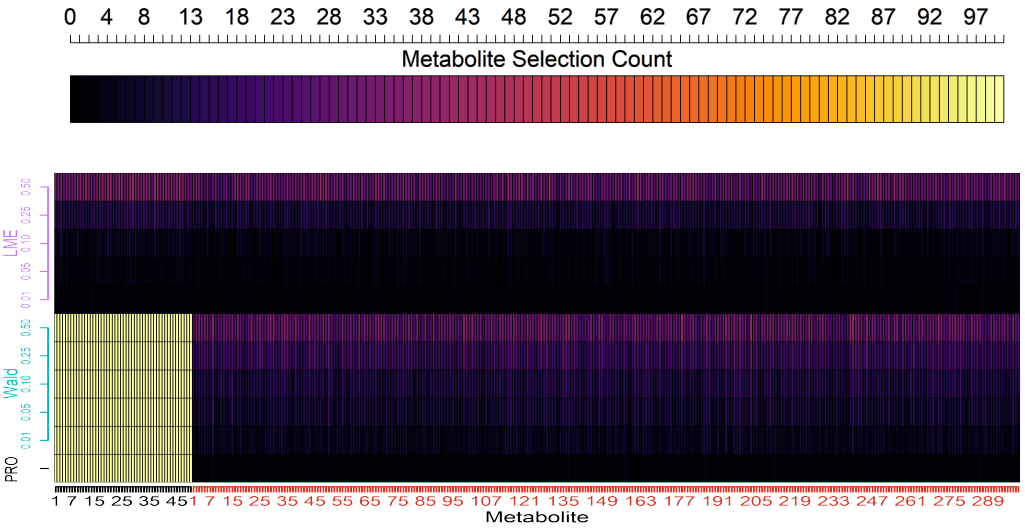
Cornell University



Performance in Simulations

Correlated Simulated Variables

Cornell University



Performance with Real Data

- Univariate methods don't pick up a single metabolite from our 352 even with an FDR of 0.5
- PROLONG selects 45 metabolites, including targets identified by our clinician collaborators and during our EDA

Summary

- High sensitivity and specificity in simulations
- Group lasso + network penalty model is slightly less sensitive at some λ_2 values but much more specific than regular lasso + network penalty
- Limited preprocessing necessary
- Stable across choice of λ_2, λ_1 can be chosen with usual MSE cross-validation for lasso and group lasso or with a grid search using AIC/BIC, Mallow's C_p, etc

Microbiome and Other -Omic Extension

- Extension to other continuous omics variables is immediate
- Our current work is incorporating the relative abundances of 282 microbiome species measured at the first and last time points

Challenges with Microbiome Data

- Zero Inflation
- Compositional data - relative abundances are used instead of raw counts
- Estimating correlation within microbiome and between microbiome and metabolites
- Subset of time points for clinical outcome and metabolomic variables
- High between-subject variation

Microbiome Composition at Class Level

Proposed Model

We propose incorporating the compositional data directly into the same model framework along with the metabolomic variables by using the radial transformation [2]

$$\frac{x}{||x||_2}$$

Additional investigation is needed to determine if Pearson's correlation using the radial transformed data is adequate for the purposes of our network constraint.

References

[1] Wipperman, M.F., Bhattarai, S.K., Vorkas, C.K. et al. Gastrointestinal microbiota composition predicts peripheral inflammatory state during treatment of human tuberculosis. Nat Commun 12, 1141 (2021).

[2] Park, Junyoung, et al. "Kernel Methods for Radial Transformed Compositional Data with Many Zeros." International Conference on Machine Learning. PMLR, (2022).