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

First, Some Definitions

Omics data: Data from studies involving genomics, proteomics, metabolomics, microbiomics, etc.

RHEZ: rifampin (R), isoniazid (H), ethambutol (E), and pyrazinamide (Z), the standard combination of drugs used to treat tuberculosis

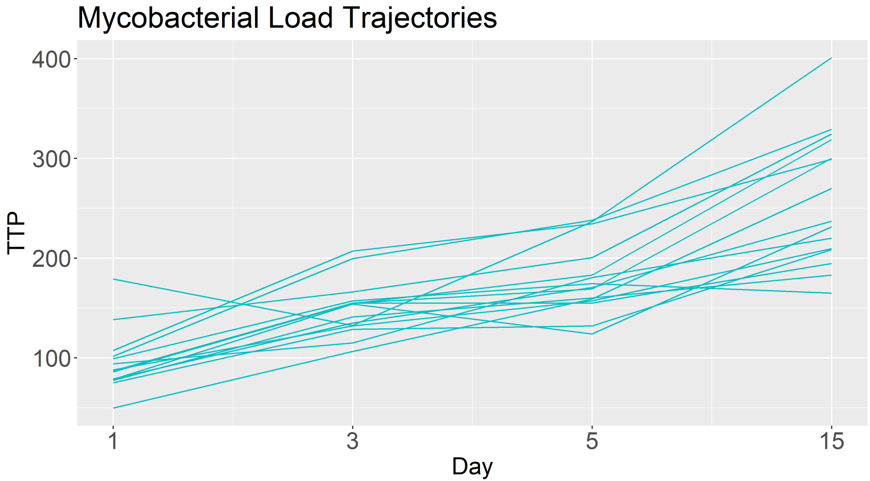
Mycobacterial Load: Measured quantity of mycobacteria, including the tuberculosis-causing pathogen

TTP: Time to Positivity, used to measure mycobacterial load, with smaller TTP values indicating higher load

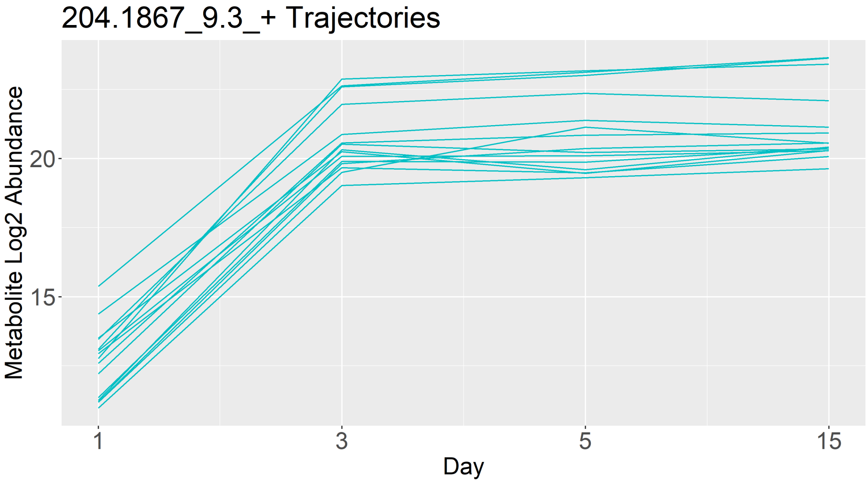
Motivation in Short

Clinician gives you a longitudinal clinical outcome, along with hundreds (or thousands) of longitudinal -omics variables, and asks which variables co-vary with the outcome?

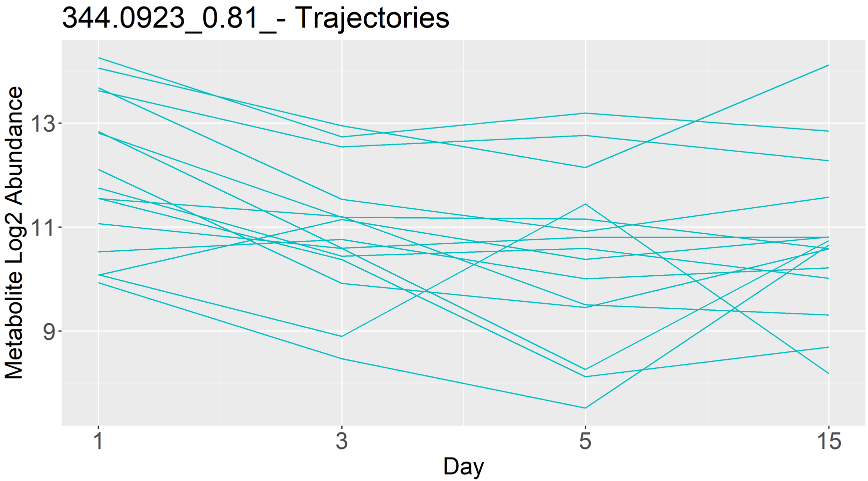
Motivation in Pictures



Motivation in Pictures



Motivation in Pictures



Motivation in More Words

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

We want to:

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

Tuberculosis Data

- 15 subjects, TB patients treated with RHEZ
- 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
- We also have microbiome and RNAseq data [1] for days 1 and 15 - more on this later

General Model Idea

Cornell University

- 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

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

And for each variable j we have

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

- 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

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

Moving X from Tensor to Matrix

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

Moving X from Tensor to Matrix

Now replace $\Delta\tilde{X}_{it}^{[j]}$ with row vector $\Delta\tilde{X}_{it} = |\Delta\tilde{X}_{it}^{[1]} \quad \Delta\tilde{X}_{it}^{[2]} \quad \dots \quad \Delta\tilde{X}_{it}^{[p]}|$

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

Network Penalty Via Correlation Matrix

We do not have a known graph associated with our metabolites, but we can construct one using our measured correlation matrix as a weighted adjacency matrix.

To do so, we need to construct an appropriate correlation matrix for our design matrix X.

Correlation Matrix

For any subject i, consider

$$\Delta X = \left|\Delta X^{[1]} \quad \cdots \quad \Delta X^{[p]}\right|_{p(T-1)}$$

$$\text{where} \qquad \Delta X^{[j]} = \left|\Delta X_1^{[j]} \quad \Delta X_2^{[j]} \quad \cdots \quad \Delta X_{(T-1)}^{[j]}\right|_{(T-1)}$$

Note that $\Delta X = vec(\mathcal{X})$, where $\mathcal{X}$ is a $(T-1) \times p$ matrix with correlation matrix $\tilde{R}$.

Correlation Matrix

The correlation matrix of ΔX is

$$\begin{bmatrix} \tilde{R}^{[11]} & \dots & \tilde{R}^{[1p]} \\ \vdots & \ddots & \vdots \\ \tilde{R}^{[p1]} & \dots & \tilde{R}^{[pp]} \end{bmatrix}_{p(T-1)\times p(T-1)}$$

where each block $\tilde{R}^{[j,k]}$ is a $(T-1)\times(T-1)$ correlation matrix for $\Delta X^{[j]}, \Delta X^{[k]}$.

Correlation Matrix

Now consider the correlation matrix $\mathcal{R}$ of $vec(\mathcal{X}^\top)$. This is also a $p(T-1) \times p(T-1)$ matrix but can be written in terms of blocks corresponding to pairs of time points of $\mathcal{X}$

$$\mathcal{R} = \begin{bmatrix} \mathcal{R}_{11} & \mathcal{R}_{12} & \cdots & \mathcal{R}_{1(T-1)} \\ \mathcal{R}_{21} & \mathcal{R}_{22} & \cdots & \mathcal{R}_{2(T-1)} \\ & & \vdots & \\ \mathcal{R}_{(T-1)1} & \mathcal{R}_{(T-1)2} & \cdots & \mathcal{R}_{(T-1)(T-1)} \end{bmatrix}_{p(T-1)\times p(T-1)}$$

Correlation Matrix

We can construct the correlation matrix R associated with our design matrix X using the blocks of $\mathcal{R}$

$$R = \left[\begin{array}{c|c|c|c} \mathcal{R}_{11} & 0 & \cdots & 0 \\ \hline 0 & \mathcal{R}_{(1:2)(1:2)} & \cdots & 0 \\ \hline 0 & 0 & \ddots & 0 \\ \hline 0 & \cdots & 0 & \mathcal{R} \end{array}\right]_{pT(T-1)/2 \times pT(T-1)/2}$$

We estimate R with $\hat{R}$ using observed correlations and use $\hat{R}$ as the weighted adjacency matrix for our graph.

Graph Laplacian

We define our graph G as having edges $e = (u \sim v)$ between columns u, v of X. These edges have weights

$$w(u,v) = |\hat{R}_{uv}|$$

The degree of each vertex is

$$d_u = \sum_{v \sim u} w(u,v) = \sum_{v \sim u} |\hat{R}_{uv}|$$

The normalized Laplacian matrix $\mathcal{L}$ for graph G is defined elementwise as

$$\mathcal{L}_{uv} \begin{cases} 1 - w(u,v)/d_u & \text{if } u = v \text{ and } d_u \neq 0 \\ -w(u,v)/\sqrt{d_u d_v} & \text{if } u \text{ and } v \text{ are adjacent} \\ 0 & \text{otherwise.} \end{cases}$$

Group Lasso + Laplacian Penalty

Li and Li (2008) introduced the network-constrained regularization criterion

$$L^* \left(\lambda_1, \lambda_2, \beta\right) = (Y - X\beta)^\top (Y - X\beta) + \lambda_1 |\beta|_1 + \lambda_2 \beta^\top \boldsymbol{\mathcal{L}} \beta$$

For each of our p metabolites we have $T(T-1)/2$ entries in our design matrix, and so we arrange these entries into p groups and use a network-constrained group lasso penalty

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

Where p_j is the size of group j, $T(T-1)/2$ in our case.

Model Computation

In order to minimize $L\left(\lambda_1, \lambda_2, \beta\right)$ we create an 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 $\mathcal{L} = \Gamma D \Gamma^\top$

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

We solve for β using group lasso then adjust by $1/\sqrt{1+\lambda_2}$ to get our estimate $\hat{\beta}$.

Model Hyperparameters

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

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

- λ_1 is the tuning parameter for our group lasso penalty
- λ_2 is the tuning parameter for the network penalty

Hyperparameter Selection

λ_2 is selcted via MLE, using the following optimization problem from Steiner et al.[3]

$$n \ln \left[\|Y\|_2^2 - Y^\top X \left(B_\lambda + X^\top X \right)^{-1} X^\top Y \right] + \ln \left| B_\lambda + X^\top X \right| - \ln |B_\lambda|$$

$$\text{where} \quad B_\lambda = \lambda_2 \mathcal{L} + \lambda_R I$$

After minimizing over both λ_2 and our nuisance hyperparameter λ_R, which is needed because $\mathcal{L}$ is non-invertible, we add $\lambda_R I$ to $\mathcal{L}$ before computing $\mathcal{S}$ via $\Gamma D \Gamma^\top$ decomposition.

Hyperparameter Selection

λ_1 is selected via a custom cross-validation on the artificial dataset $(\mathcal{X}, \mathcal{Y})$ generated using λ_2 and λ_R.

Splitting our rows randomly into folds would mean some subjects would have their time points split between training and test data and $\mathcal{S}$ would be similarly divided.

Instead, we always keep $\mathcal{S}^\top$ intact, and in each of our 5 folds use all rows associated with 3 of the subjects as our test data.

Nice Properties of this Penalty

- Every variable has multiple entries in the model, but the group lasso penalty gives either all zero or all non-zero coefficients for each variable, helping interpretability
- If two variables are highly correlated, and one is a strong enough predictor to be selected, the other variable is more 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 Models
- Wald tests 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.

Performance in Simulations

Uncorrelated Simulation Scheme

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

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

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

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

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

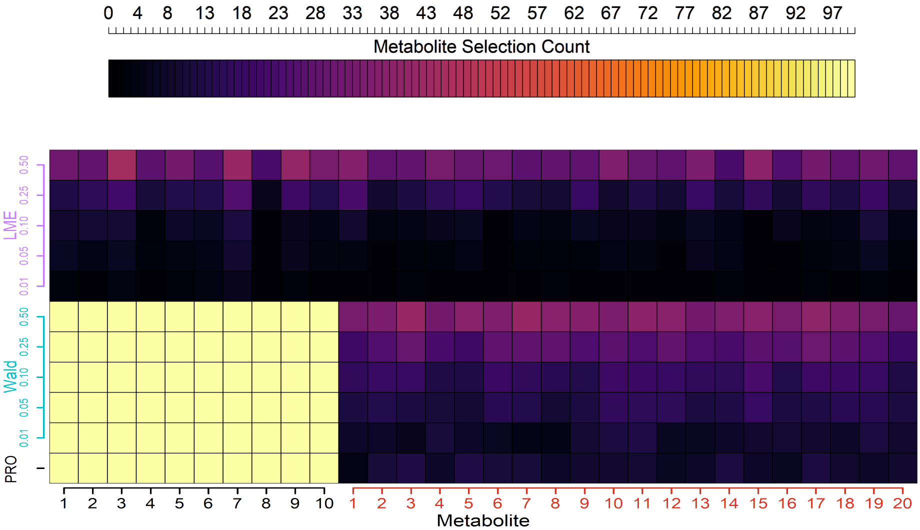
$$y_4 \sim N(y_3 + \beta(x_4 - x_3) + \beta(x_3 - x_2) + \beta(x_2 - x_1), 5)$$

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

SNR ranging from 1 to 2 in targets

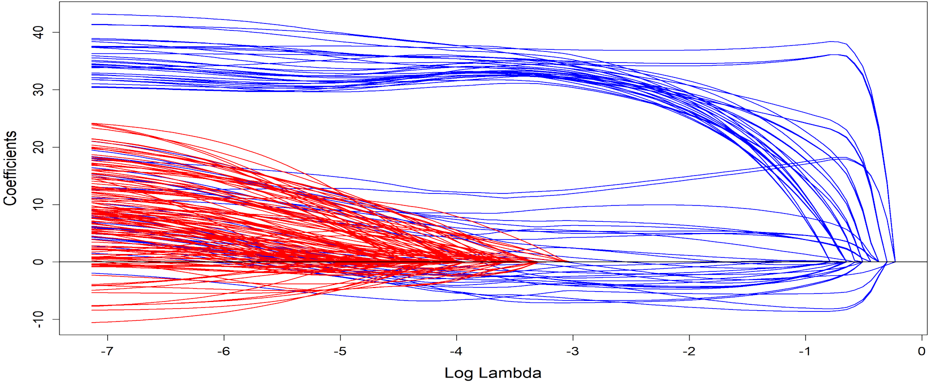
Performance in Simulations

Uncorrelated Simulated Variables



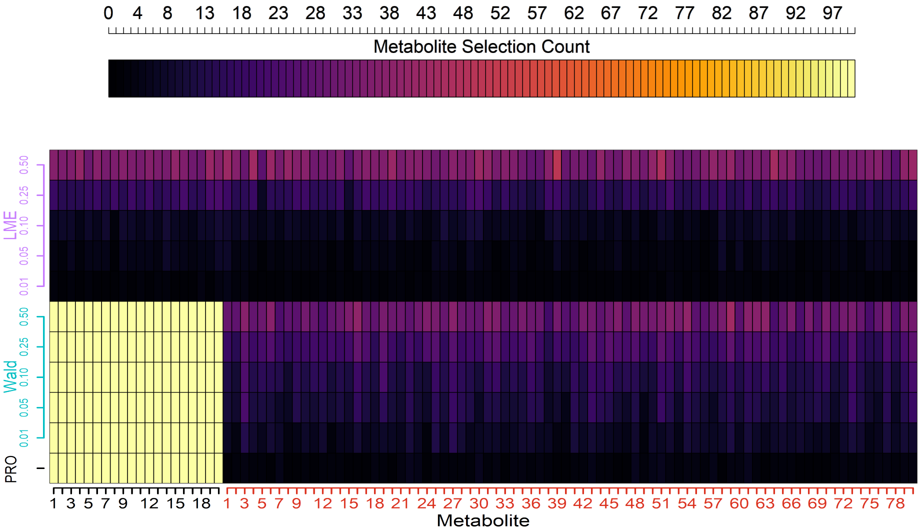
Performance in Simulations

Uncorrelated Simulated Variables



Performance in Simulations
Uncorrelated Simulated Variables

Cornell University



Performance in Simulations

Correlated Simulation Scheme

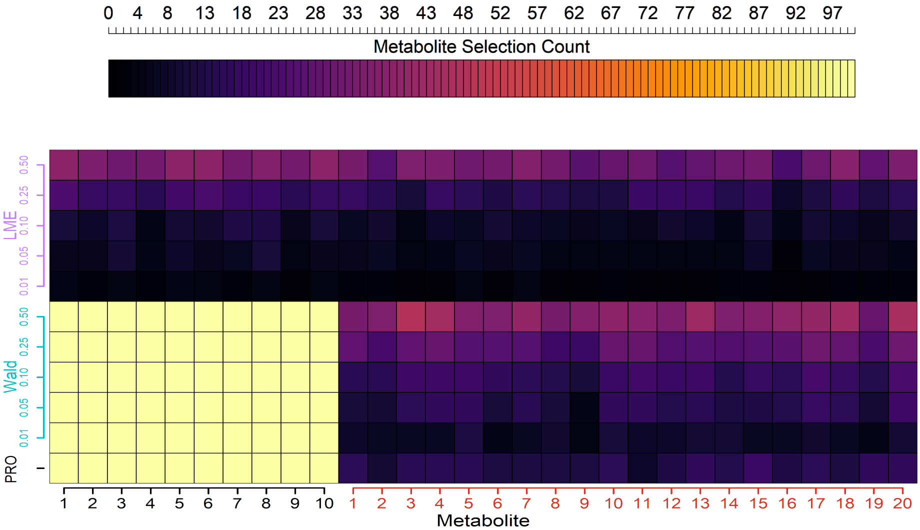
Same as previous scenario, but with

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

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 $(-1, 1)$.

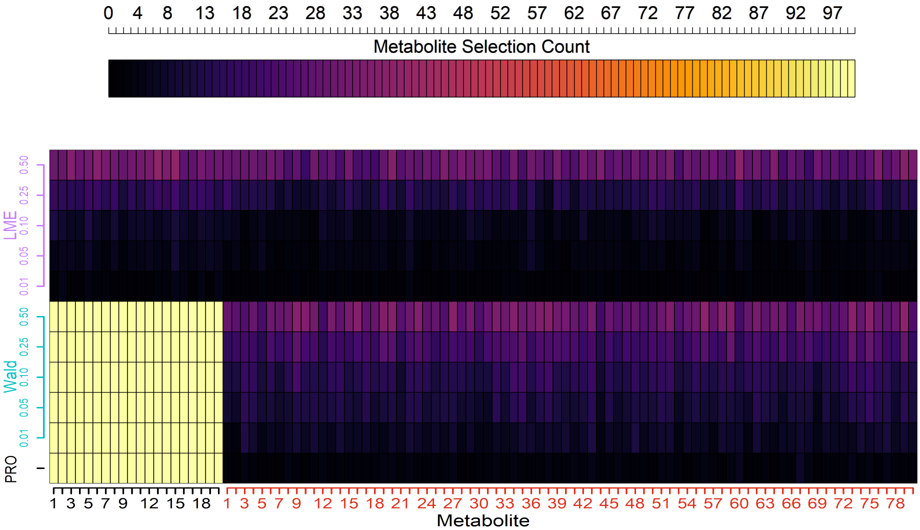
Performance in Simulations

Correlated Simulated Variables



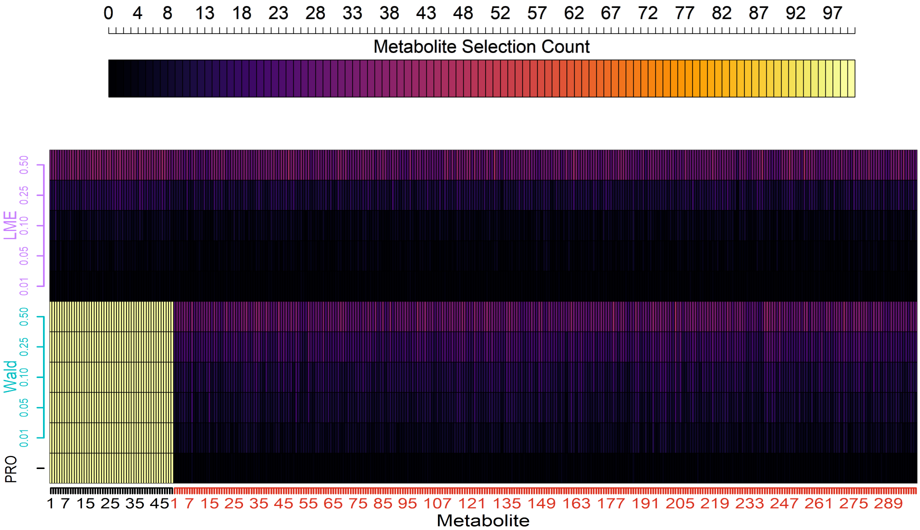
Performance in Simulations

Correlated Simulated Variables



Performance in Simulations

Correlated Simulated Variables



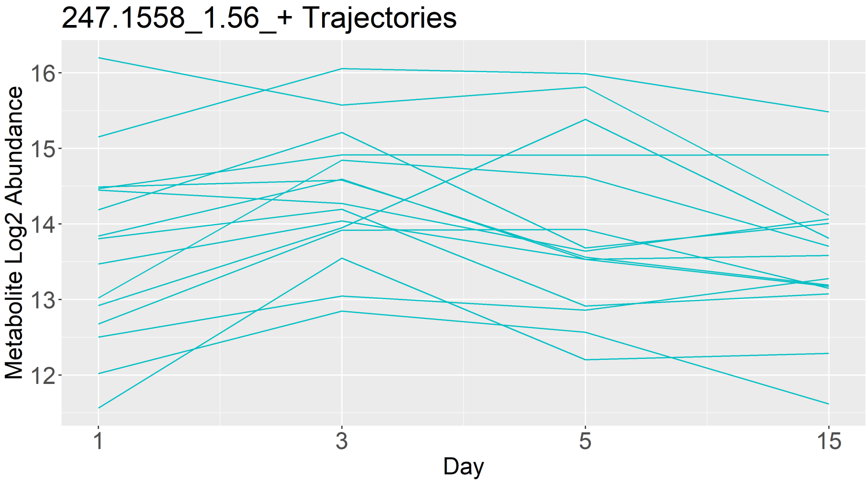
Performance with Real Data

- Univariate mixed effect models do not pick up a single metabolite from our 352 at an FDR of 0.05, and still only picks out one at 0.5
- Univariate Delta Wald tests pick 116 metabolites at an FDR of 0.05
- PROLONG selects 45 metabolites, including targets identified by our clinician collaborators and during our EDA

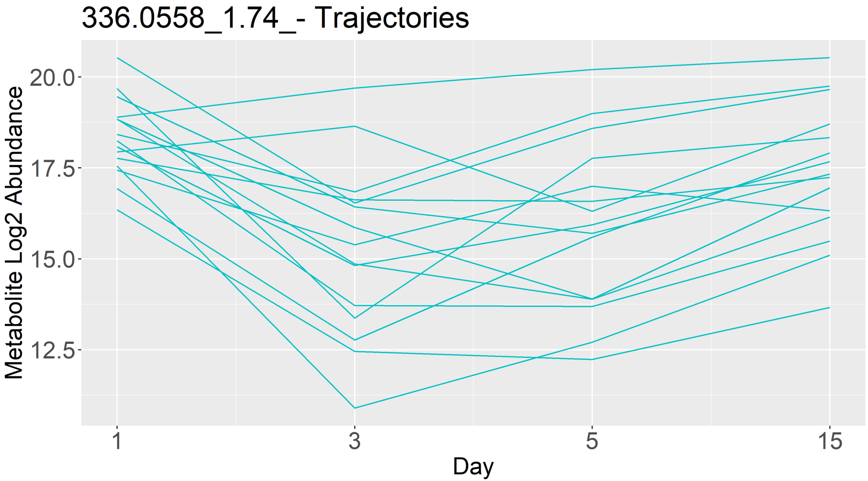
Summary

- PROLONG gets high sensitivity and specificity in simulations while 'competitor' mixed effects models perform terribly across the board
- Univariate Wald version performs best at lower dimensions, but PROLONG improves significantly as dimension and sparsity increase and performs better than Wald at 100 predictors and beyond
- Preprocessing is needed to get the data into the right structure, but model hyper-parameters are automatically selected, reducing the burden on clinicians/biostatisticians using this method

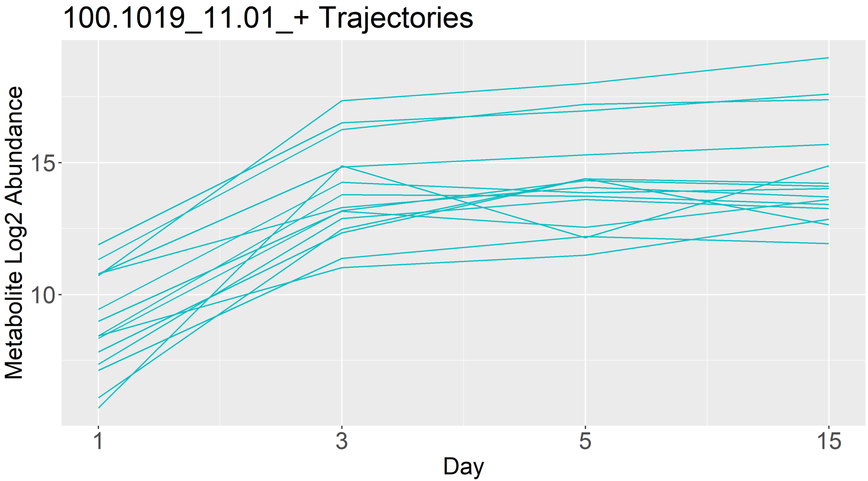
Some Selected Metabolites



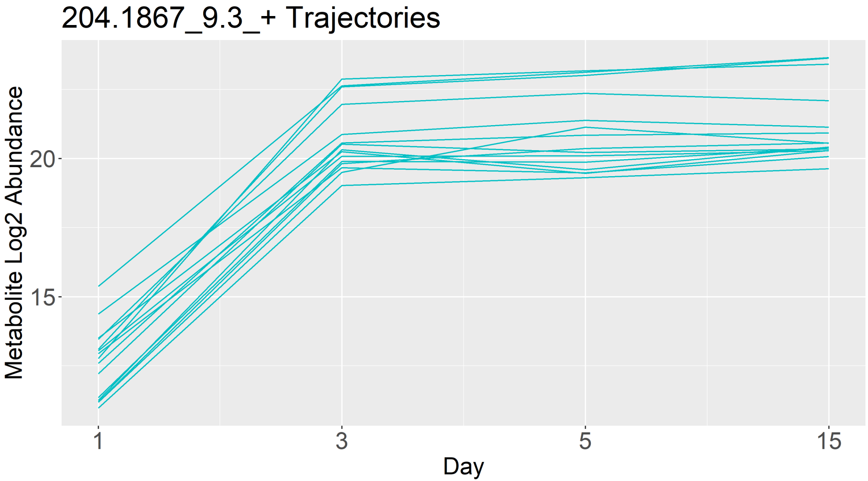
Some Selected Metabolites



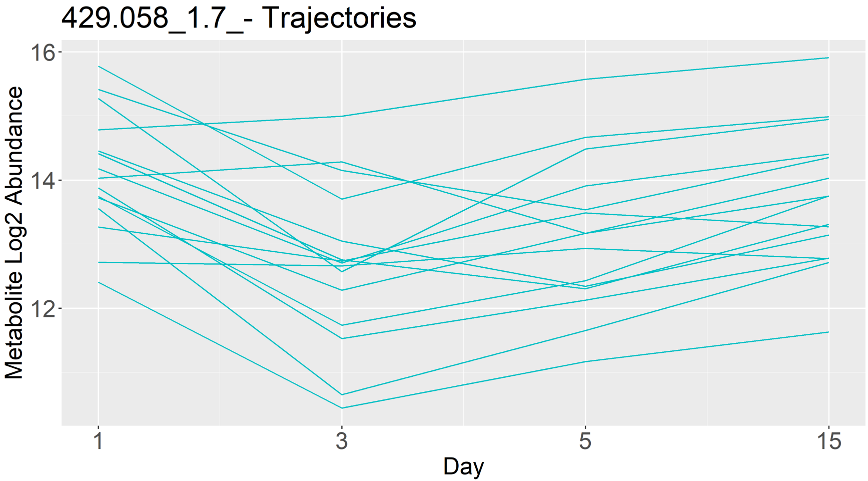
Some Selected Metabolites



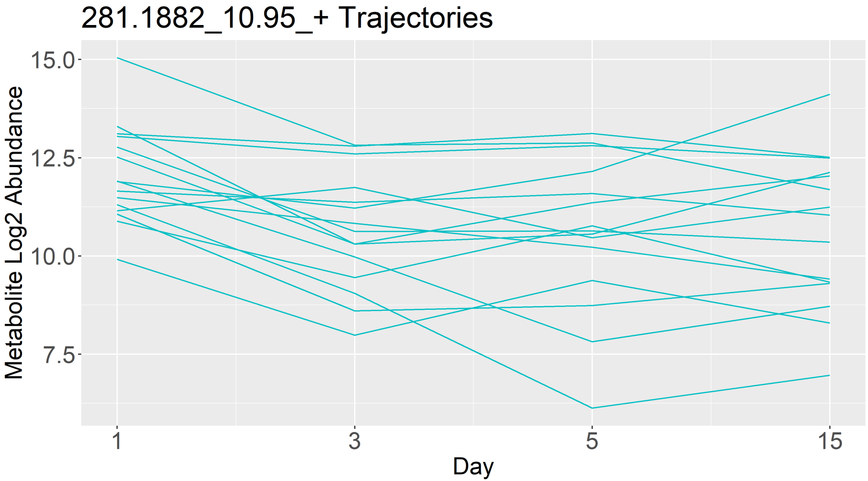
Some Selected Metabolites



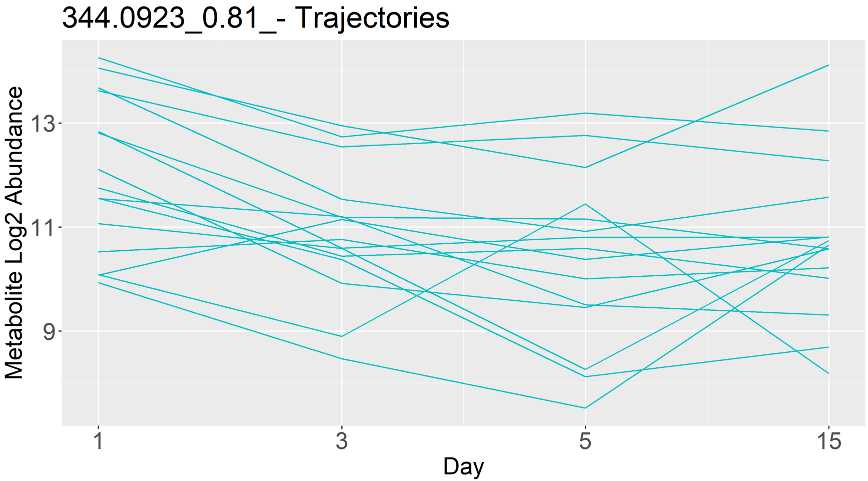
Some Selected Metabolites



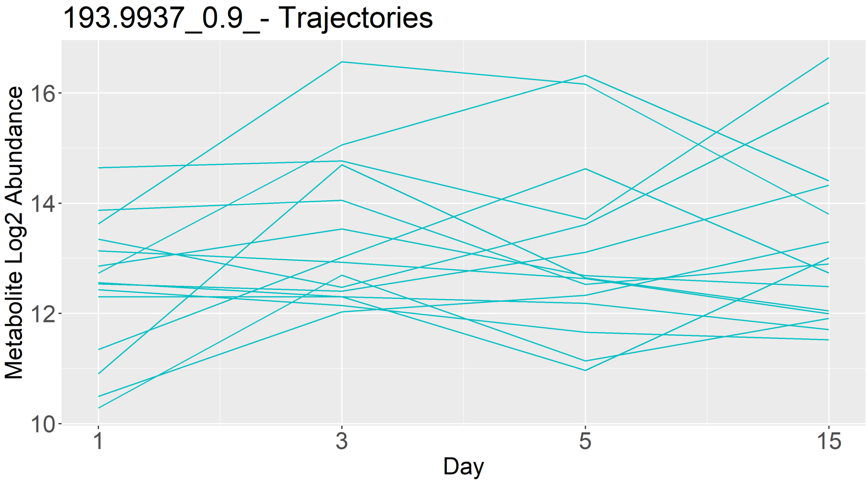
Some Selected Metabolites



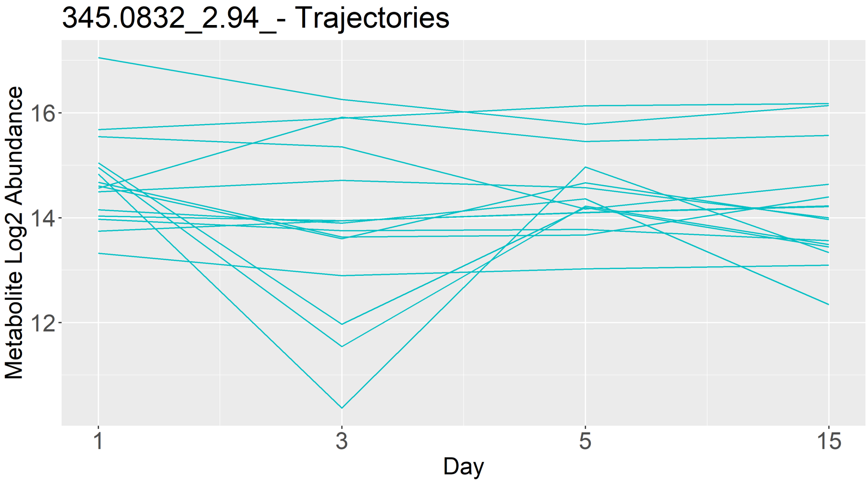
Some Selected Metabolites



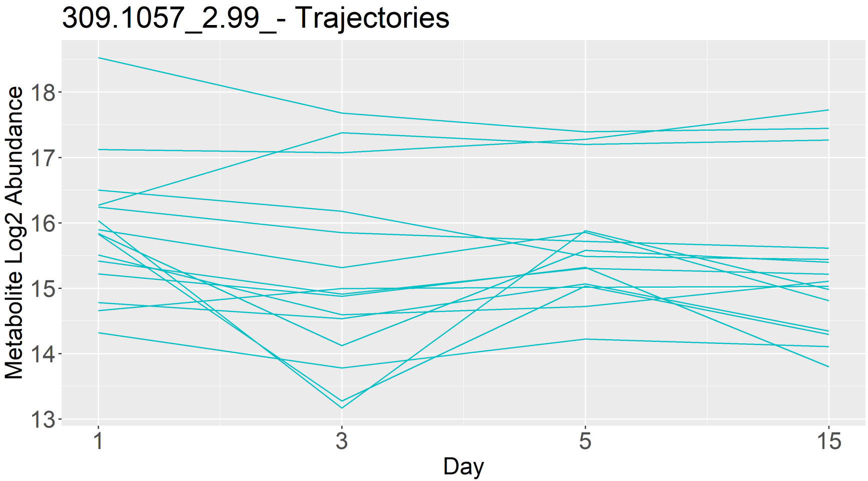
Some Selected Metabolites



Some Selected Metabolites



Some Selected Metabolites



Microbiome and Multi-Omic Extension

- Extension to other continuous omics variable types is immediate
- Our current work involves incorporating the relative abundances of 282 microbiome species measured at the first and last time points
- RNA-seq data also measured at the first and last time points will be integrated after completion of the microbiome project

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

Zero Inflation

Zero inflation is always a problem with microbiome data, and there are two primary strategies to reduce the impact of the excess zeros before transforming the data:

- We can aggregate to different taxa, reducing the proportion of zeros but potentially reducing the usefulness of the model results. Aggregating from sub-species to species is likely fine, but clinicians may find results after aggregating to class or phylum to not be very helpful
- We can exclude species that are not found in at least one time point for some minimum number of subjects. If our model selects species that are only present in only one or two subjects we would have a hard time interpreting and justifying those selections

Zero Inflation

We currently aggregate to the species level and require at least 3 subjects with a non-zero abundance. Increasing that number from 3 could help interpretability but potentially lose useful information, as we drop to 154 species present in at least 5 subjects, 92 species present in at least 8 subjects, and 68 species present in at least 10.

Compositional Data

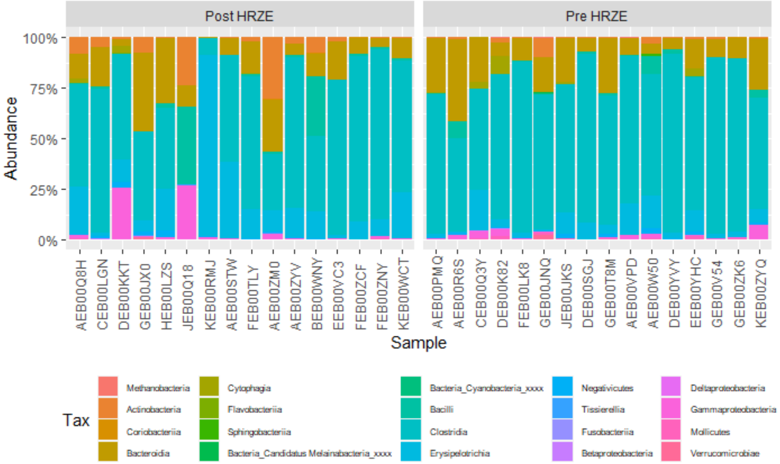
Compositional data consists of positive values with a sum (to 1) constraint, thus its points lie on a simplex. This leads to several issues with analysis, including spurious negative correlations.

The centered-log-ratio (CLR) transformation sends the data to real space without the simplex constraint, which allows for multivariate analysis and is generally easier to work with

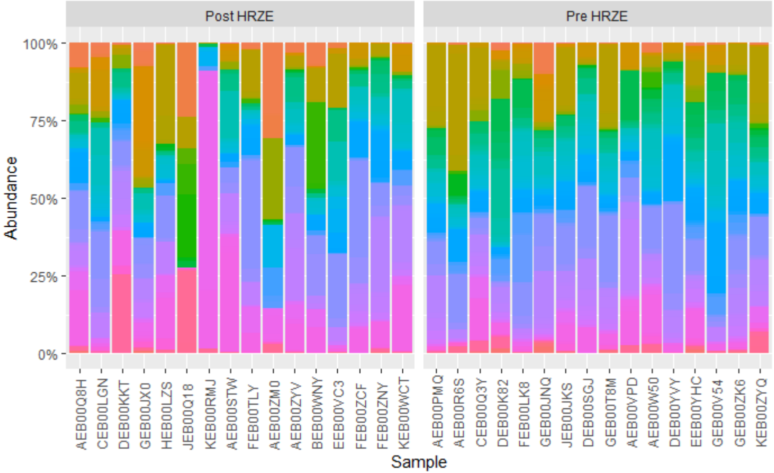
$$clr(x) := (\ln x_i - \frac{1}{D} \sum_{j=1}^{D} \ln x_j)_i$$

However, this transformation requires replacing the zeros, and results from statistical models and tests can change based on the zero-replacement strategy employed, hurting interpretability. [4]

Microbiome Composition at Class Level



Microbiome Composition at Species 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 [4]:

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

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

More investigation is also needed to determine the minimum acceptable number of subjects with non-zero abundances in at least one time point for a given species.

Preliminary Results

PROLONG selects 56 metabolites and 49 microbiome species. The higher number of metabolites selected could be explained by correlation between some metabolites and the stronger microbiome variables.

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]Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics, 24(9):1175–1182, (2008)

[3]Aleksandra Steiner, Kausar Abbas, Damian Brzyski, Kewin Paczek, Timothy W. Randolph, Joaquín Goñi, and Jaroslaw Harezlak. Incorporation of spatial- and connectivity-based cortical brain region information in regularized regression: Application to Human Connectome Project data. Frontiers in Neuroscience, 16, (2022)

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

Thank You!