

Inference for Network Constrained Variable Selection for High-Dimensional Longitudinal Omics Data with Heterogenous Treatment Groups

Steve Broll

March 2026

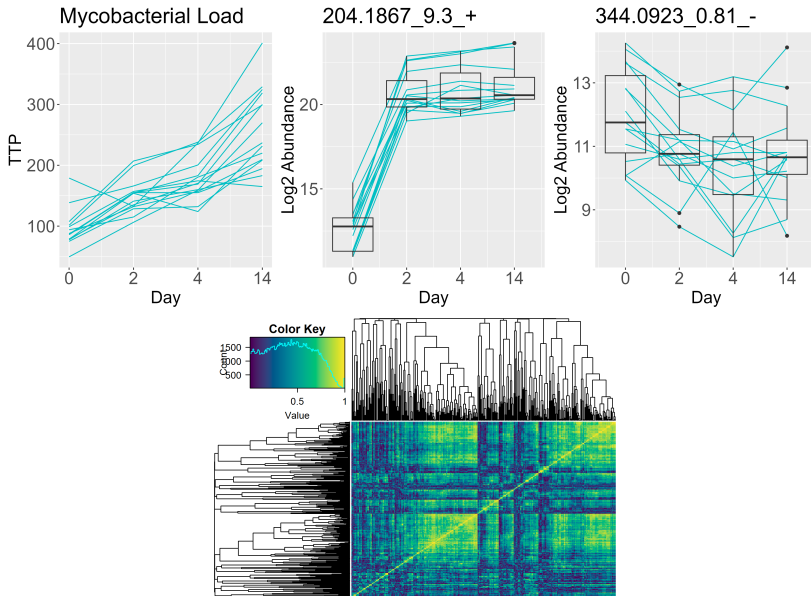
Format of Slides

- Brief review of the method 'PROLONG'
 - **p**enalized **r**egression for **o**utcome guided **l**ongitudinal **o**mic analysis with **n**etwork and **g**roup constraints
- Description of new methodological extension for pooling treatment groups
- Brief results and a discussion on areas of improvement

Original PROLONG - Setting and Motivation

- Outcome: continuous clinical outcome/phenotype/etc. measured at a few time points
 - For example, a measure of disease severity/progression
 - Time-varying with small n, T
- Variables: continuous omic abundances measured at the same time points
 - For example, serum or urinary metabolites
 - Time-varying with small n, T and high-dimensional with large $p \gg n$
- Motivation: finding which omic variables co-vary over time with the outcome
 - Selected variables can be used as biomarkers for early/late disease progression, or for biological insight

Motivating Data



Targeted Gaps

- There is extensive work in the omics setting, especially in differential expression/abundance between groups (binary/categorical outcome)
 - In contrast, we focus on time-varying continuous outcomes.
- Clinicians typically apply univariate linear mixed effects (LME) models or generalized estimating equations (GEEs), depending on the magnitude of n
- We propose a joint model leveraging correlation over time and between variables to select a sparse set of biomarkers
- We later provide a framework for uncertainty quantification and inference with two or more groups

General Idea

- First-difference the data, then stack our $T - 1$ values of X and Y so we have (for 4 time points)

$$Y = [Y_2 - Y_1 \quad Y_3 - Y_2 \quad Y_4 - Y_3]^T$$

And for each omic variable j we have

$$X_j = [X_{j2} - X_{j1} \quad X_{j3} - X_{j2} \quad X_{j4} - X_{j3}]^T$$

- Set up design matrix so that each first-differenced Y value is regressed on contemporaneous first-differenced values of X
 - (Optional:) Include all first-differenced values of X to account for potential lags
- Apply Laplacian and group lasso penalties to induce sparsity while utilizing correlation and group- structure

Vectorized $\tilde{Y}$

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

Matricized X (Univariate)

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

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

Matricized X (Multivariate)

Now replace each $\Delta \tilde{X}_{it}^{[j]}$ with row vector

$$\Delta \tilde{X}_{it} = [\Delta \tilde{X}_{it}^{[1]} \Delta \tilde{X}_{it}^{[2]} \dots \Delta \tilde{X}_{it}^{[p]}]$$

$$\rightarrow X = \begin{bmatrix} \Delta \tilde{X}_{11} & & 0 & 0 & 0 \\ \vdots & & & & \\ \Delta \tilde{X}_{n1} & & 0 & 0 & 0 \\ \hline 0 & \Delta \tilde{X}_{11} & \Delta \tilde{X}_{12} & 0 & 0 \\ & \vdots & & & \\ 0 & \Delta \tilde{X}_{n1} & \Delta \tilde{X}_{n2} & 0 & 0 \\ \hline 0 & 0 & \ddots & 0 & 0 \\ \hline 0 & 0 & 0 & \Delta \tilde{X}_{11} & \dots & \Delta \tilde{X}_{1(T-1)} \\ & & & \vdots & & \\ & & & \Delta \tilde{X}_{n1} & \dots & \Delta \tilde{X}_{n(T-1)} \end{bmatrix}_{n(T-1) \times pT(T-1)/2}$$

Penalization

Given our first-differenced and stacked response vector Y , and our first-differenced and stacked design matrix X , we seek to minimize

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

- λ_1 is the tuning parameter for our group lasso penalty
- 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 (data derived)

Novelties of PROLONG

- First-differencing and block-design matrix
- Group Lasso + Laplacian Penalty

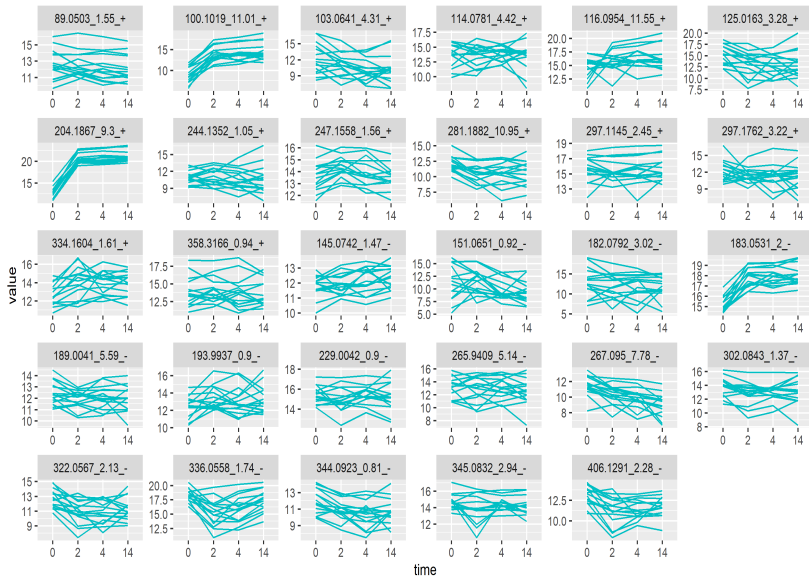
Why First-Differencing?

- Capture trends over time, de-emphasize baseline value (intercept)
- Controls first-order time dependence in the outcome
- Remove any time-invariant effects, simplifying the model parameters
- Analogous to paired testing

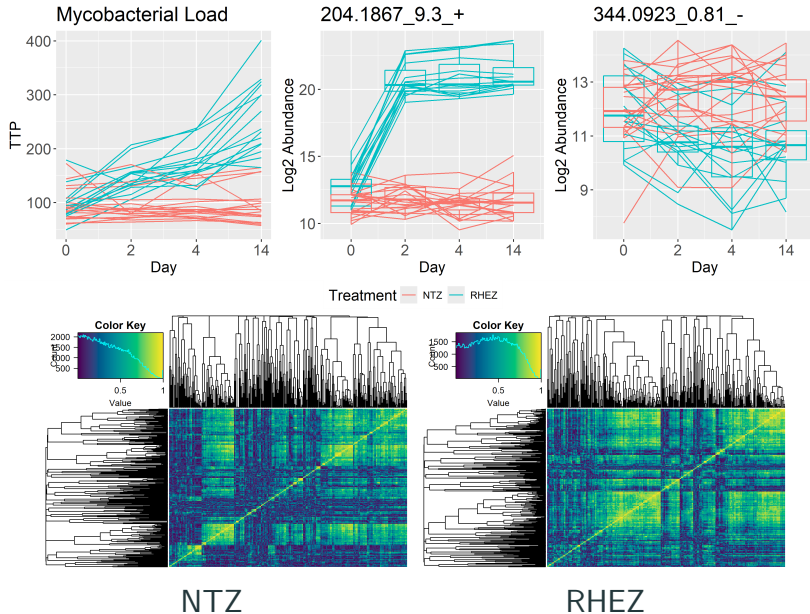
Why Laplacian Penalization?

- Ridge-like component makes it comparable to elastic net, but with a stronger grouping effect
 - Elastic net adds penalty $\lambda_2 \|\beta\|_2^2 = \lambda_2 \beta^\top I \beta \rightarrow \lambda_2 \beta^\top L \beta$
- Nudges coefficients for correlated variables together (smoothing over dependence network)
- Alleviates numerical challenges typical to lasso-type problems in presence of multicollinearity

Variables Selected by PROLONG



Heterogeneous Group Data



Multi-Group Extension - General Idea

- Fit a separate PROLONG model to each group, allowing coefficients and hyperparameters to vary between heterogeneous groups
 - If hyperparameters are turned across groups, low signal groups can lead to over-sparsifying higher signal groups
- Optionally move from group lasso $\rightarrow$ sparse group lasso, which may have support set recovery benefits
- Construct a debiased estimator
- Estimate σ^2 with MLE from Laplacian ridge-like formulation
- Test all coefficients for each omic variable simultaneously

Debiasing Lasso-type Estimators

The KKT conditions for estimator $\hat{\beta}^n$, which describe necessary conditions for finding an optimal solution $\hat{\beta}^n$ without violating our constraints, give

$$\frac{X^\top(Y - X\hat{\beta}^n)}{n} = \lambda_1 s$$

where s is a subgradient of the lasso-type penalty at $\hat{\beta}_n$.

We add a term proportional to the subgradient to compensate for the downward bias introduced by the shrinkage.

$$\hat{\beta}^u = \hat{\beta}^n + \frac{MX^\top(Y - X\hat{\beta}^n)}{n},$$

Here, M is a matrix designed to 'decorrelate' the columns of X , and controls both the bias and variance of $\hat{\beta}^u$.

Purpose of M

Set $\hat{\Sigma} \equiv (\mathbf{X}^\top \mathbf{X}) / n$.

for $i = 1, 2, \dots, p$ **do**

Let m_i be a solution of the convex program:

$$\text{minimize } m^\top \hat{\Sigma} m \quad \text{subject to } \left\| \hat{\Sigma} m - e_i \right\|_\infty \leq \mu$$

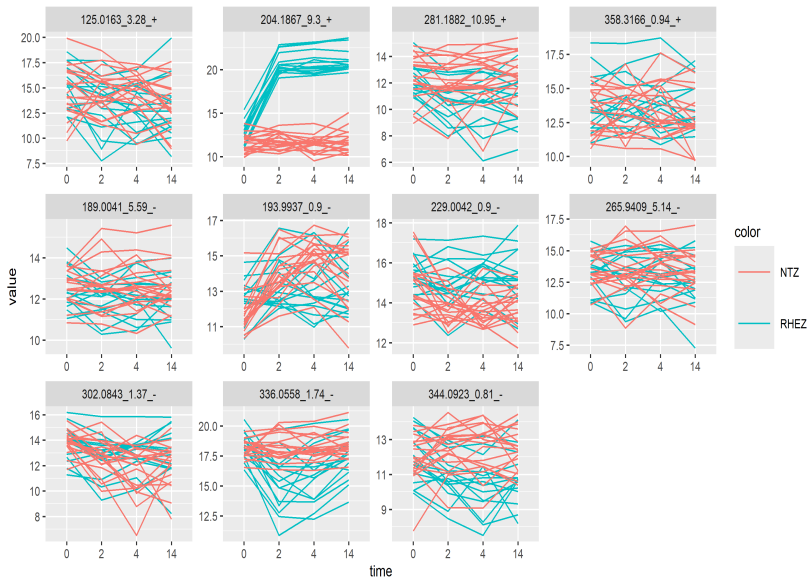
where $e_i \in \mathbb{R}^p$ is the vector with one at the i -th position and zero everywhere else. $\triangleright$ Set $M = I_{p \times p}$ if not feasible

- Controlling $|M\hat{\Sigma} - I|_\infty$, the maximum absolute entry-wise difference, controls the bias of $\hat{\beta}^u$
- Minimizing the diagonal elements of $M\hat{\Sigma}M$ controls the variance of $\hat{\beta}^u$

Wald Test

- We want to test all coefficients for a single omic variable j across all time points and treatment groups
- We apply a Wald test with $H_0 : \beta_j = \mathbf{0}; H_A : \beta_j \neq \mathbf{0} \quad \forall j$.
- Our covariance matrix is block diagonal, with one block for each treatment group a : $M_a \hat{\Sigma}_a M_a$

Inference Selected Variable Trajectories

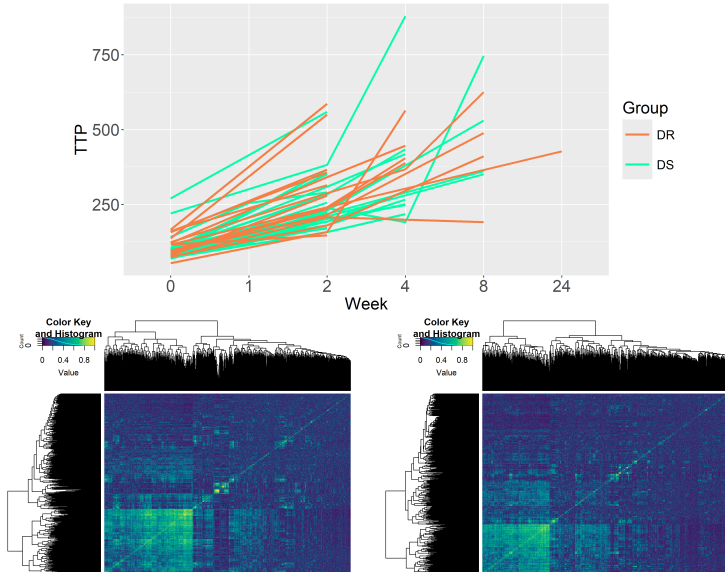


Simulation Results

- Debiased PROLONG procedure obtains desired FDR control - LME and GEE do not
- FDR control, as expected leads to significant power loss
- Our estimates $\hat{\sigma}^2$ are consistently too high, leading to over-conservative tests

Secondary Data Set

Mycobacterial Load Trajectories



- Generalizing approach so that # time points T_n can vary
- Thresholding absolute correlation to reduce complexity
- Tweaking algorithm for M to better account for group structure
- I'm working in parallel on a group lasso algorithm that
 - automatically tunes sparsity parameter λ_1 (no CV)
 - estimates σ^2 while correcting for bias from shrinkage
 - significantly improves computation time when p near n

References



Adel Javanmard and Andrea Montanari.

Confidence intervals and hypothesis testing for high-dimensional regression.

J. Mach. Learn. Res., 15(1):2869–2909, January 2014.
ISSN 1532-4435.



Caiyan Li and Hongzhe Li.

Network-constrained regularization and variable selection for analysis of genomic data.

Bioinformatics, 24(9):1175–1182, May 2008.
ISSN 1367-4803.

doi: 10.1093/bioinformatics/btn081.

Thank You!

Email: `stevebroll@cornell.edu`

R package available via Github:

`https://github.com/stevebroll/prolong`

PROLONG Manuscript available via **Bioinformatics!**