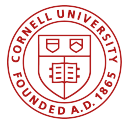


Variable Selection and Inference for High-Dimensional Longitudinal Omics with Heterogeneous Groups

Steve Broll

Format

- Background
- Obtaining estimators $\hat{\beta}$, σ^2 via autotuned group lasso
- Variable selection for longitudinal omics
- Inference for multiple treatment groups

Background - Variable Selection

Consider the usual linear regression framework: $Y = X\beta + \varepsilon$, $Y_i \in \mathbb{R}^1$, $X_i \in \mathbb{R}^p$, and errors $\varepsilon_i \sim N(0, \sigma^2 I)$ for $i = 1, \ldots, n$.

In high-dimensional setting ($p >> n$), we can use variants of the lasso[1] to perform variable selection (β-sparsity)

$$\min_{\beta \in \mathbb{R}^p} \quad \frac{1}{2n} ||Y - X\beta||_2^2 + \lambda ||\beta||_1.$$

Given predefined non-overlapping groups that partition our p variables, we can use group lasso[2] (group sparsity)

$$\min_{\beta \in \mathbb{R}^p} \quad \frac{1}{2n} ||Y - \sum_{g=1}^{G} X_g \beta_g||_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g} ||\beta_g||_2.$$

[1] Tibshirani, "Regression Shrinkage and Selection Via the Lasso", 1996.

[2] Yuan and Lin, "Model selection and estimation in regression with grouped variables", 2006.

Background - Inference after Variable Selection

The shrinkage towards zero results in a biased $\hat{\beta}$. There are several ways to "debias" $\hat{\beta}$; we focus on the following[3], related to the KKT conditions:

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

M is essentially a relaxed inverse of the sample covariance $\hat{\Sigma} = X^\top X/n$, and is designed to minimize the variance of $\hat{\beta}^u$ while also controlling non-Gaussianity and bias.

A similar correction has been applied to the group lasso and sparse group lasso[4].

[3] Javanmard and Montanari, "Confidence intervals and hypothesis testing for high-dimensional regression", 2014.

[4] Mitra and Zhang, "The benefit of group sparsity in group inference with de-biased scaled group Lasso", 2016; Cai, Zhang, and Zhou, "Sparse Group Lasso: Optimal Sample Complexity, Convergence Rate, and Statistical Inference", 2022.

Background - λ and σ^2

The tuning parameter λ determines the (group-)sparsity level of $\hat{\beta}$, and is typically selected via cross-validation.

Asymptotical results for high-dimensional linear models[5] suggest the optimal λ is proportional to $\sigma\sqrt{\log(p)/n}$ for lasso, and to $\sigma\sqrt{\log(G)/n}$ for group lasso.

Thus, estimating σ^2 is valuable both for tuning λ and for inference on debiased estimators.

[5] Bickel, Ritov, and Tsybakov, "Simultaneous Analysis of Lasso and Dantzig Selector", 2009; Meier, Van De Geer, and Bühlmann, "The Group Lasso for Logistic Regression", 2008.

Estimating σ^2

Methods that estimate $\hat{\sigma}^2$ exist, like the scaled (group-)lasso[6], and square root (group)-lasso[7].

For the group lasso, limited availability in R, and overall less study of these variants than the ordinary variant.

We propose a new method, autotune group lasso, that quickly estimates $\hat{\sigma}^2$ and $\hat{\beta}$ without cross-validation. Related to autotune lasso[8].

[6] Sun and Zhang, "Scaled sparse linear regression", 2012; Mitra and Zhang, "The benefit of group sparsity in group inference with de-biased scaled group Lasso", 2016.

[7] Belloni, Chernozhukov, and Wang, "Square-root lasso: pivotal recovery of sparse signals via conic programming", 2011; Bunea, Lederer, and She, "The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms", 2014.

[8] Sadhukhan et al., *Autotune: fast, accurate, and automatic tuning parameter selection for Lasso*, 2025.

Autotune for Group Lasso

Idea:

Update $\hat{\sigma}^2$, and consequently λ, alternately with β in a penalized negative Gaussian log-likelihood:

$$\min_{\beta,\sigma^2} = \frac{1}{2}\log\sigma^2 + \frac{1}{2n\sigma^2}||Y - X\beta||_2^2 + \lambda_0(\sqrt{p_g}\sum_{g=1}^{G}\left\|\beta_{(j)}\right\|_2).$$

When we alternate between β and σ^2 we hold one fixed and then have a convex function to minimize.

$\hat{\sigma}^2$ is a univariate parameter and converges quickly; then update $\hat{\beta}$ until convergence.

Algorithm Overview

- Start with a group-sparse, but non-zero, estimate $\hat{\beta}$.
  - i.e. $\lambda_{\text{init}} = \lambda_{\text{max}}/2$, where λ_{max} is smallest λ s.t. $\hat{\beta} = 0$.
- Rank the groups by the standard deviation of their partial residuals.
- Run sequential Wald tests to find significant groups.
- Update $\hat{\sigma}^2$ to be the residual variance with these significant groups as predictors.
- Update λ by scaling λ_{init} by $\hat{\sigma}^2/\text{Var}(Y)$
- Update β via block coordinate descent
- Repeat!

Some Notes on the Algorithm

The use of group-wise partial residuals from the alternating updates is key to the quick convergence of the algorithm.

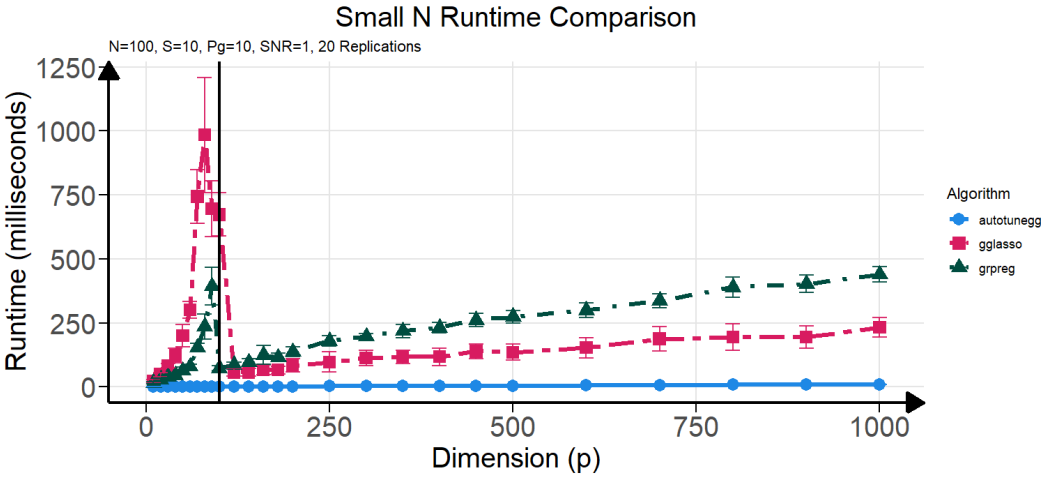
For updating β, we use the same block coordinate descent as in the grpreg R package[9]. Given the same λ, our default algorithm will give the same $\hat{\beta}$ estimate as grpreg. gglasso[10] instead uses block majorization descent, and will return a different estimate.

We can also apply a version of active set selection after convergence of $\hat{\sigma}^2$ that prioritizes the significant groups selected by the sequential tests.

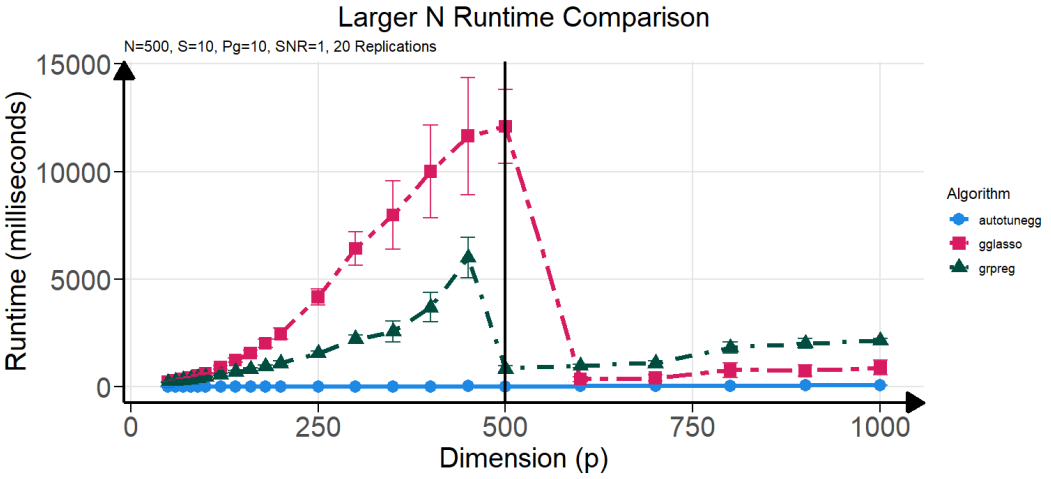
[9] Breheny and Huang, "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors", 2015.

[10] Yang, Zou, and Bhatnagar, *gglasso*, 2020.

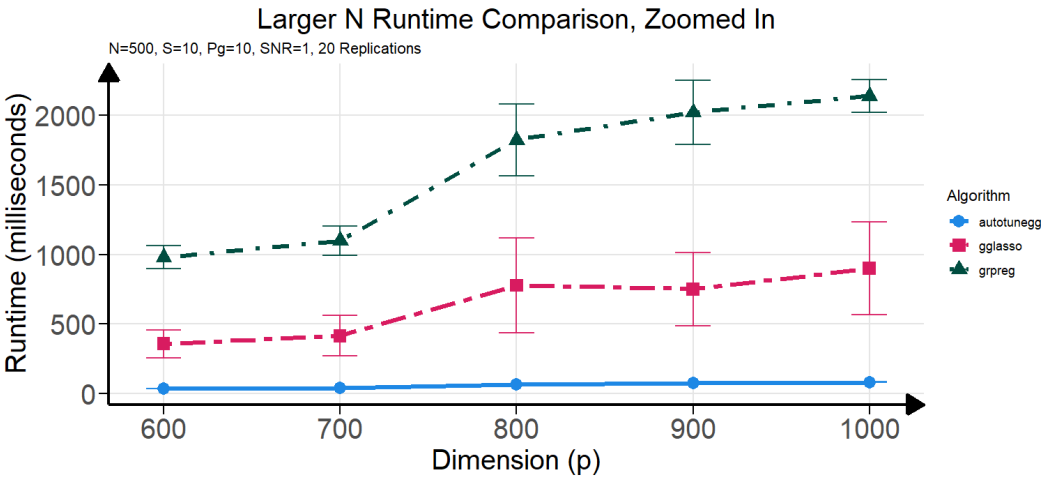
Runtime Comparison



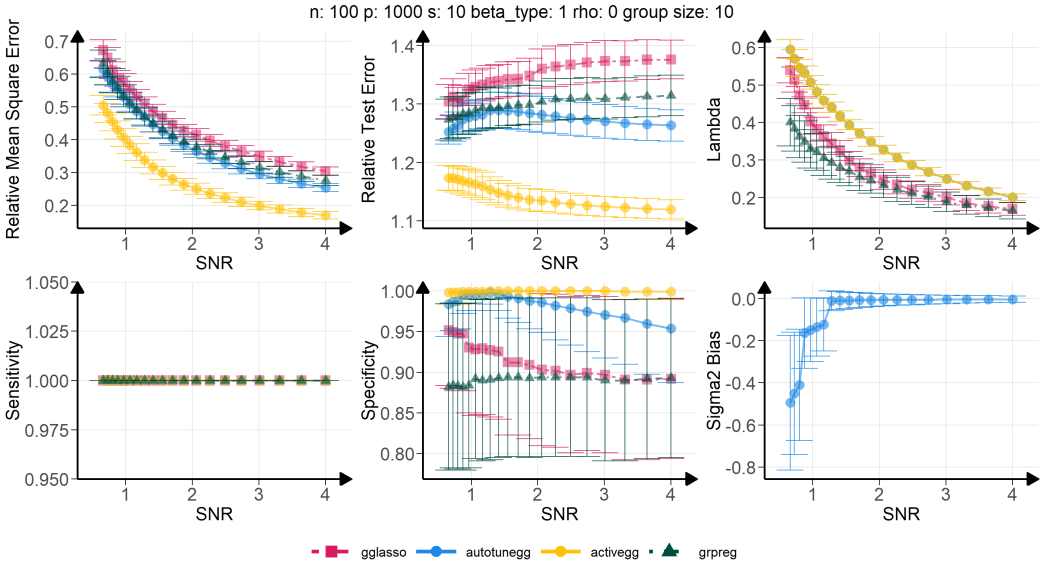
Runtime Comparison



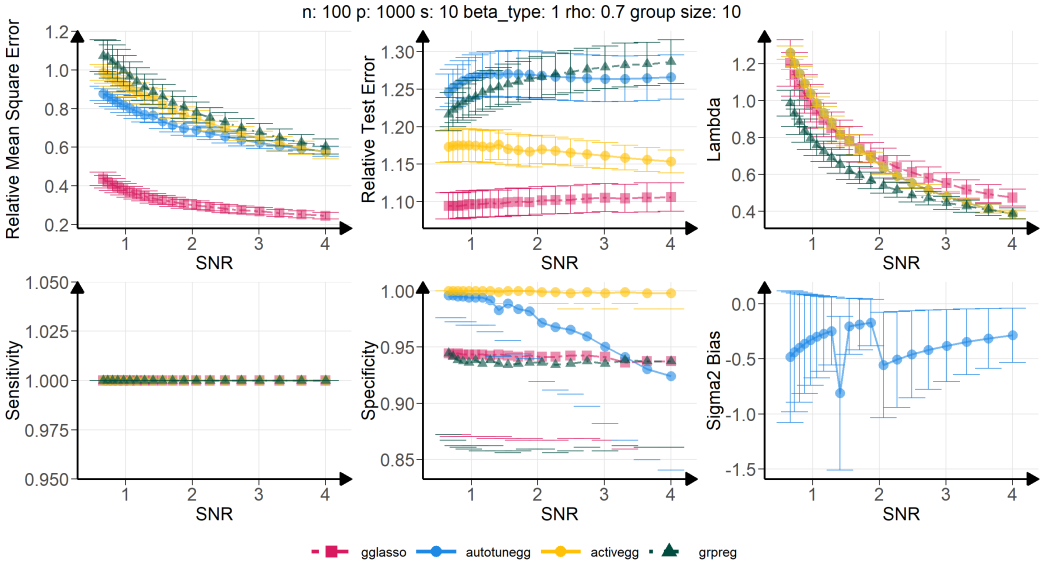
Runtime Comparison



Model Performance Comparison - Uncorrelated



Model Performance Comparison - Correlated



Application to Longitudinal Omics Data

- 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 $p >> n$
- 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

PROLONG Model[11]

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

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

  And for each omic variable j we have

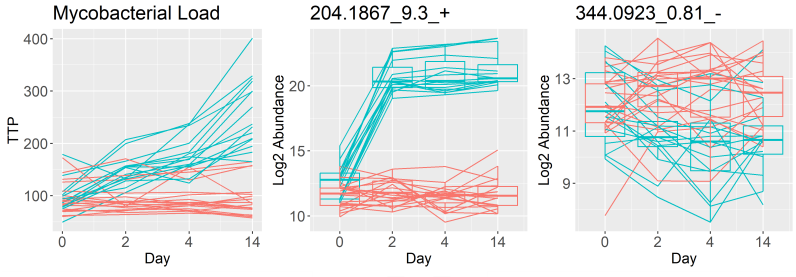
$$X_j = [X_{j2} - X_{j1} \qquad X_{j3} - X_{j2} \qquad X_{j4} - X_{j3}]^T$$

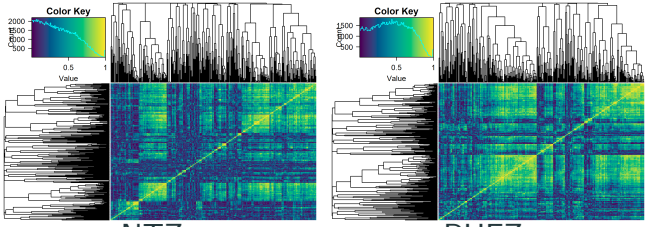
- Apply graph Laplacian and group lasso penalization:

$$||Y - X\beta||_2^2 + \underbrace{\lambda_1 \sum_{j=1}^{p} \left\|\boldsymbol{\beta}_{(j)}\right\|_2}_{\text{Induce Group Sparsity}} + \underbrace{\lambda_2 \beta^T L \beta}_{\substack{\text{Elastic Net-Like} \\ \text{Grouping Effect on} \\ \text{Correlated Predictors}}}$$

[11] Broll et al., "PROLONG: penalized regression for outcome guided longitudinal omics analysis with network and group constraints", 2025.

Example Heterogeneous-Group Data



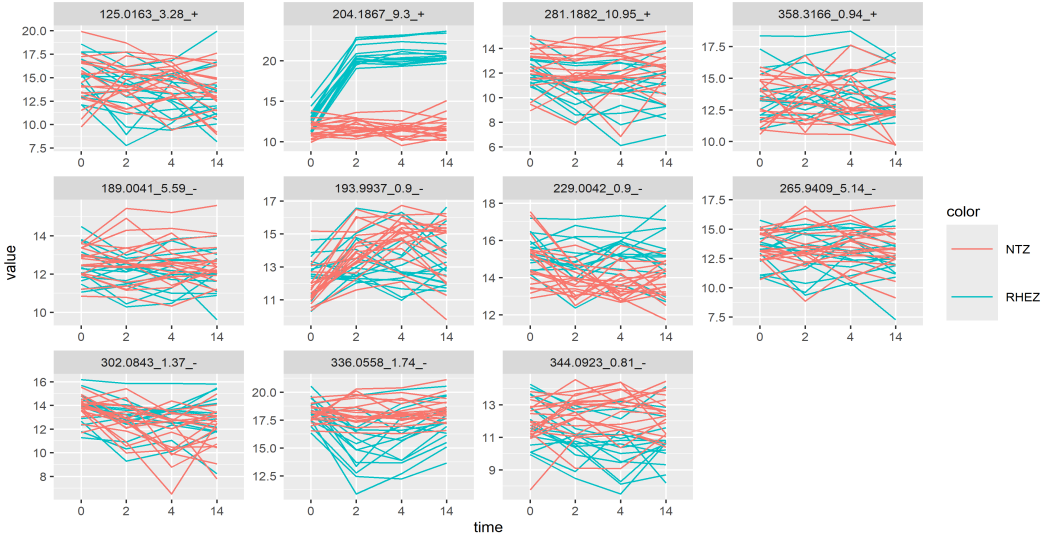


NTZ RHEZ

Multi-Group Extension - General Idea

- Fit a separate PROLONG model to each group, allowing coefficients and hyperparameters to vary between heterogenous groups
  - If $\lambda, \hat{\sigma}^2$ are turned across groups, low signal groups can lead to over-sparsifying higher signal groups.
- Construct debiased estimator $\hat{\beta}^u$.
- Using autotuned $\hat{\sigma}^2$, test all (grouped) coefficients for each omic variable, or test individual coefficients (e.g. specific time points)

Inference Selected Variable Trajectories



Simulation Results

Debiased PROLONG procedure obtains desired FDR control - LME and GEE do not.

FDR control, as expected leads to power loss compared to pure variable selection.

Our previous MLE-based estimator led to a consistently high estimation of $\hat{\sigma}^2$, leading to over conservative tests. Newer autotune $\hat{\sigma}^2$ estimator appears to be much more accurate - more testing needed on empirical bias.

Belloni, A., V. Chernozhukov, and L. Wang. **"Square-root lasso: pivotal recovery of sparse signals via conic programming".** In: *Biometrika* 98.4 (2011), pp. 791–806.

Bickel, Peter J., Ya'acov Ritov, and Alexandre B. Tsybakov. **"Simultaneous Analysis of Lasso and Dantzig Selector".** In: *The Annals of Statistics* 37.4 (2009), pp. 1705–1732.

Breheny, Patrick and Jian Huang. **"Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors".** eng. In: *Statistics and Computing* 25.2 (Mar. 2015), pp. 173–187.

Broll, Steven et al. **"PROLONG: penalized regression for outcome guided longitudinal omics analysis with network and group constraints".** In: *Bioinformatics* 41.4 (Mar. 2025), btaf099.

Bunea, Florentina, Johannes Lederer, and Yiyuan She. **"The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms".** In: *IEEE Transactions on Information Theory* 60.2 (2014), pp. 1313–1325.

Cai, T. Tony, Anru R. Zhang, and Yuchen Zhou. **"Sparse Group Lasso: Optimal Sample Complexity, Convergence Rate, and Statistical Inference".** In: *IEEE Transactions on Information Theory* 68.9 (2022), pp. 5975–6002.

Javanmard, Adel and Andrea Montanari. **"Confidence intervals and hypothesis testing for high-dimensional regression".** In: *J. Mach. Learn. Res.* 15.1 (Jan. 2014), 2869–2909.

Meier, Lukas, Sara Van De Geer, and Peter Bühlmann. **"The Group Lasso for Logistic Regression".** In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70.1 (Feb. 2008), pp. 53–71.

Mitra, Ritwik and Cun Hui Zhang. **"The benefit of group sparsity in group inference with de-biased scaled group Lasso".** In: *Electronic Journal of Statistics* 10.2 (2016), pp. 1829–1873.

Sadhukhan, Tathagata et al. **Autotune: fast, accurate, and automatic tuning parameter selection for Lasso.** 2025.

Sun, Tingni and Cun-Hui Zhang. **"Scaled sparse linear regression".** In: *Biometrika* 99.4 (Dec. 2012), pp. 879–898.

Tibshirani, Robert. **"Regression Shrinkage and Selection Via the Lasso".** In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (Jan. 1996), pp. 267–288.

Yang, Yi, Hui Zou, and Sahir Bhatnagar. **gglasso: Group Lasso Penalized Learning Using a Unified BMD Algorithm.** Mar. 2020.

Yuan, Ming and Yi Lin. **"Model selection and estimation in regression with grouped variables".** en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (Feb. 2006), pp. 49–67.

Thank you!

Email: stevebroll@cornell.edu