

# High-Dimensional Classification and Regression of Psychological Data

Nathaniel E. Helwig

Associate Professor of Psychology and Statistics  
Core Member of the Data Science Initiative  
University of Minnesota



Genes, Brains, and Behavior Seminar  
University of Minnesota, Twin Cities  
October 13, 2025

# Table of Contents

1. Overview of Statistical Learning Problem
2. ABGD Algorithm and **grpnet** R Package
3. Real Data Examples
  - Genes: Cancer GWAS
  - Brains: EEG Classification
  - Behavior: Math Performance
  - Brains + Behavior: ABCD Data
4. Conclusions and Future Directions

# Table of Contents

1. Overview of Statistical Learning Problem
2. ABGD Algorithm and `grpnet` R Package
3. Real Data Examples
  - Genes: Cancer GWAS
  - Brains: EEG Classification
  - Behavior: Math Performance
  - Brains + Behavior: ABCD Data
4. Conclusions and Future Directions

# Multiple and Generalized Nonparametric Regression

Suppose that  $Y$  follows an exponential family distribution.

Given predictors  $(X_1, \dots, X_p)$ , consider a MGNR model of the form

$$g(\mu) = f(X_1, \dots, X_p)$$

where

- $\mu$  is the (conditional) expectation of  $Y$  (given  $X_1, \dots, X_p$ )
- $g(\cdot)$  is a user-specified link function (monotonic and invertible)
- $f(\cdot)$  is an unknown “smooth” function of  $q \leq p$  predictors

Goal: estimate the unknown function  $f(\cdot)$  from training data.

# Challenge: Model Structure is Completely Unknown

**Variable Selection Problem:** which predictors should be included?

- Not all of the candidate predictors may be needed
- Want to determine which predictors are relevant (active)

**Term Selection Problem:** how should variables be included?

- Do the predictors combine in an additive fashion? Or interactive?
- Want to learn which main and/or interaction effects matter

**Smoothing Problem:** how smooth should each term be?

- Predictors may (and likely will) enter model in nonlinear fashion
- Want to avoid under-fitting (bias) and over-fitting (variability)

# Three-Pronged Approach

## Step 1: The Theory (Tensor Product Reproducing Kernels)

- Existing tensor product RKHS theory is not ideal for selection
- Develop improved representer theorem for tensor products

## Step 2: The Algorithm (Adaptively Bounded Gradient Descent)

- Existing algorithms are not easily adaptable (e.g., to different  $g$ )
- Develop versatile group elastic net algorithm for GLMs/GAMs

## Step 3: The Practice (Smoothing and Subspace Selection)

- Existing implementations only consider additive models
- Develop intuitive software for selecting main and interaction effects

# Group Elastic Net Regularized GLM/GAM

I consider a groupwise extension of the classic elastic net estimator:

$$\hat{\boldsymbol{\beta}}_{\lambda, \alpha} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{n} L(\boldsymbol{\beta} | \mathbf{D}) + \lambda P_{\alpha}(\boldsymbol{\beta}) \right\} \quad (1)$$

where the group elastic net penalty function has the form

$$P_{\alpha}(\boldsymbol{\beta}) = \sum_{k=1}^K \omega_k \left( \alpha \|\boldsymbol{\beta}_k\| + \frac{1}{2} (1 - \alpha) \|\boldsymbol{\beta}_k\|^2 \right) \quad (2)$$

which is a natural groupwise extension of the elastic net penalty.

- Helwig (2024) Stats paper for smoothing theory
- Helwig (2025) JCGS paper for computational theory
- Helwig (2025) **grpnet** R package for application

# Table of Contents

1. Overview of Statistical Learning Problem
2. ABGD Algorithm and **grpnet** R Package
3. Real Data Examples
  - Genes: Cancer GWAS
  - Brains: EEG Classification
  - Behavior: Math Performance
  - Brains + Behavior: ABCD Data
4. Conclusions and Future Directions

# ABGD Algorithm for GLMs with Group Elastic Net

---

**Algorithm 1** Adaptively bounded gradient descent algorithm for generalized linear models with group elastic net penalties.

---

- 1: Calculate  $\gamma_k = \text{maxeig} \left( \frac{1}{n} \sum_{i=1}^n w_i \mathbf{x}_{ik} \mathbf{x}_{ik}^\top \right)$  for  $k = 1, \dots, K$
  - 2: Initialize  $\beta_k = \mathbf{0}_{p_k}$  for all  $k = 1, \dots, K$
  - 3: **while**  $\max_j \frac{|\beta_j - \beta_{j(\text{old})}|}{1 + |\beta_j|} > \epsilon$  **do**
  - 4:    $\beta_{\text{old}} \leftarrow \beta$
  - 5:    $\delta = \min_i (d_i^2 v_i) = \max_i (1/d_i^2 v_i)$
  - 6:   **for**  $k = 1, \dots, K$  **do**
  - 7:      $\mathbf{g}_k = \frac{1}{n} \sum_{i=1}^n \frac{w_i}{d_i v_i} (y_i - \mu_i) \mathbf{x}_{ik}$
  - 8:      $\mathbf{b}_k = \beta_k + (\gamma_k \delta)^{-1} \mathbf{g}_k$
  - 9:      $\beta_k = \frac{1}{1 + (1 - \alpha) \lambda \omega_k (\gamma_k \delta)^{-1}} \left( 1 - \frac{\alpha \lambda \omega_k (\gamma_k \delta)^{-1}}{\|\mathbf{b}_k\|} \right)_+ \mathbf{b}_k$
  - 10:   **end for**
  - 11: **end while**
-

## Scope of Packages: Loss Functions

Table 1: Loss functions available in each package.

$L(\beta \mathbf{D})$	<code>gglasso</code>	<code>grplasso</code>	<code>grpnet</code>	<code>grpreg</code>
Gaussian	✓	✓	✓	✓
Multivariate Gaussian			✓	
Binomial / Logistic	✓	✓	✓	✓
Multinomial			✓	
Poisson		✓	✓	✓
Negative Binomial			✓	
Gamma			✓	
Inverse Gaussian			✓	
Squared Hinge	✓		✓	
Smoothed Hinge	✓		✓	
Cox PH				✓

# Scope of Packages: Usability Features

Table 2: Usability features of each package.

Features	gglasso	grplasso	grpnet	grpreg
$\alpha$ and $\gamma$ tuning			✓	
Parallelization			✓	
Default method	✓		✓	✓
Formula method		✓	✓	
Orthogonalization		✓	✓	✓
Standardization			✓	
Additive splines			✓	✓
Tensor product splines			✓	
MCP and SCAD			✓	✓

# Table of Contents

1. Overview of Statistical Learning Problem
2. ABGD Algorithm and `grpnet` R Package
3. Real Data Examples
  - Genes: Cancer GWAS
  - Brains: EEG Classification
  - Behavior: Math Performance
  - Brains + Behavior: ABCD Data
4. Conclusions and Future Directions

# Overview of Gene Expression Cancer RNA-Seq Dataset

Gene Expression Cancer RNA-Seq dataset from UCI Machine Learning Repository (Fiorini, 2016; Kelly et al., 2025)

- From the PANCAN (Pan-Cancer Atlas) project
- <https://www.synapse.org/#!Synapse:syn4301332>

Standard GWAS dataset (samples  $\times$  gene expressions)

- $n = 801$  patients with one type of cancer
- $m = 5$  different types of cancer
- $p = 20531$  gene expressions

**Goal:** classify the cancer types from the gene expressions\*

---

\*Adapted from Helwig (2025)

## Response and Predictor Variables

The response variable has the following distribution:

- BRCA (breast cancer): 300 patients
- COAD (colon cancer): 78 patients
- KIRC (kidney cancer): 146 patients
- LUAD (lung cancer): 141 patients
- PRAD (prostate cancer): 136 patients

267 genes were excluded due to having a standard deviation of zero

- Implies no measurable variation on the gene
- $p = 20264$  remaining genes with useful data

# Overview of Data Splitting and Analysis

Multinomial regression classifying cancer type from gene expressions

- Consider all main effects of all 20264 genes
- Use (spectral) smoothing spline with 5 df (Helwig, 2024)
- Note: design matrix has 101320 columns after expansion

Use 80/20 (training/testing) data splitting procedure:

- Fit/tune model on training data; evaluate on testing data
- Repeat process 10 times to investigate stability of solution

For each sample of training data:

- 10-fold CV to tune  $\lambda$  with  $\alpha = 1$  (using `lambda.1se` for prediction)
- Compare LASSO, MCP, and SCAD

# Results: Prediction Accuracy

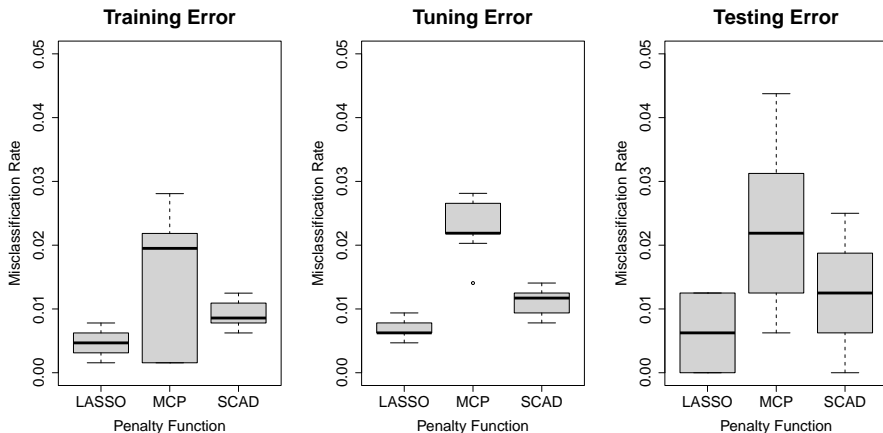


Figure 1: Cancer GWAS: prediction accuracy.

# Results: Variable Selection

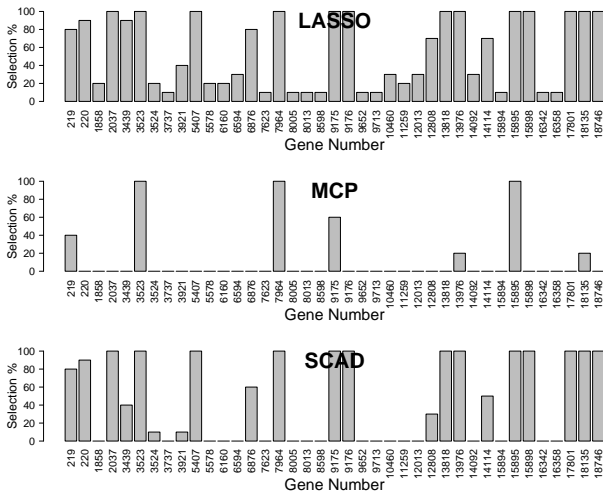


Figure 2: Cancer GWAS: variable selection.

# Results: Variable Importance

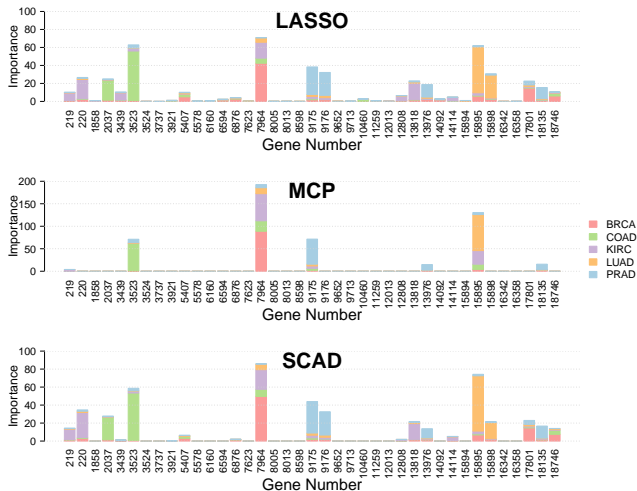


Figure 3: Cancer GWAS: variable importance.

## Results: Model Predictions

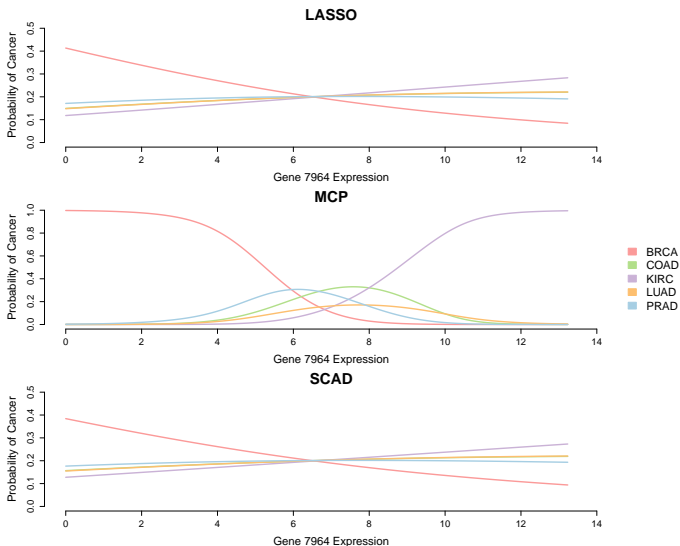


Figure 4: Cancer GWAS: model predictions.

# Overview of EEG Database

EEG Database - UCI Machine Learning Repository (Kelly et al., 2025)

- From Henri Begleiter's neurodynamics lab (Begleiter, 1995)
- Using version of data on GitHub from Mohammed et al. (2020)

Visual stimulus ERPs for one second at 256 Hz

- $n = 122$  subjects ( $n_a = 77$  alcoholics and  $n_c = 45$  controls)
- $p = 14592$  features (256 time points  $\times$  57-channel cap)

**Goal:** identify the time points and channels that distinguish alcoholics<sup>†</sup>

---

<sup>†</sup>Joint work with Jong Won Lee

# Response and Predictor Variables

Response is class label:

$$y_i \in \{-1, +1\}$$

Predictors are ERPs:

- 57 electrodes
- 256 time points

ERP features were concatenated into vector of length  $p = 14592$

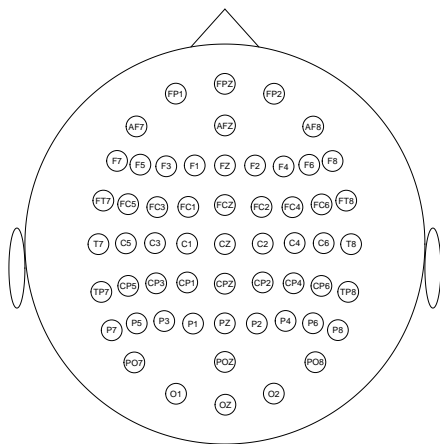


Figure 5: The 57 channel EEG cap.

# Overview of Data Splitting and Analysis

Binary classification using three loss functions:

- logit: logistic regression (binomial family)
- svm1: quadratically smoothed hinge
- svm2: squared hinge

Use 70/30 (training/testing) data splitting procedure:

- Fit/tune model on training data; evaluate on testing data
- Repeat process 10 times to investigate stability of solution

For each sample of training data:

- 5-fold CV to tune  $\lambda$  and  $\alpha$  (using `lambda.1se` for prediction)
- Compare LASSO, MCP, and SCAD

# Results: Prediction Accuracy

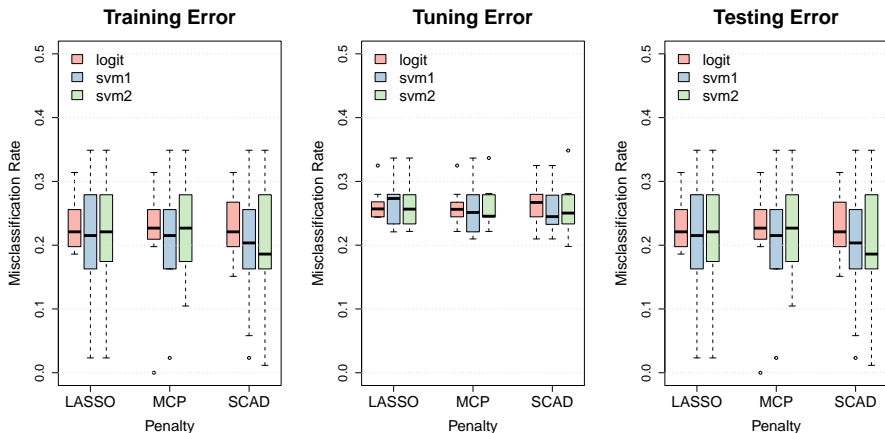


Figure 6: EEG Database: prediction accuracy.

# Results: Variable Selection

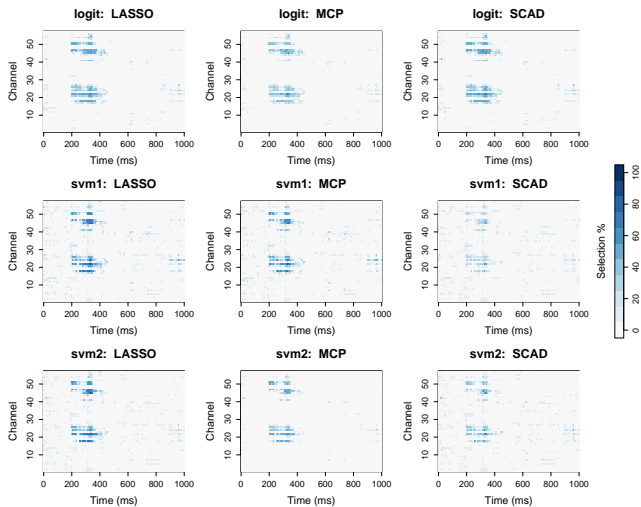


Figure 7: EEG Database: variable selection.

# Results: Variable Importance (Heatmap)

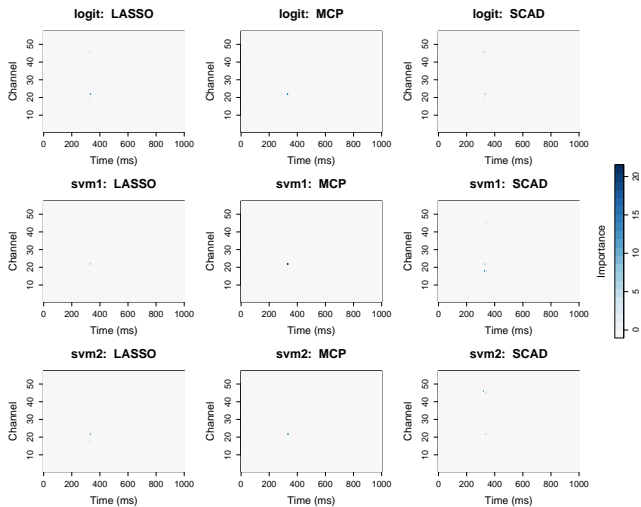


Figure 8: EEG Database: variable importance by time and channel.

# Results: Variable Importance (Line Plots)

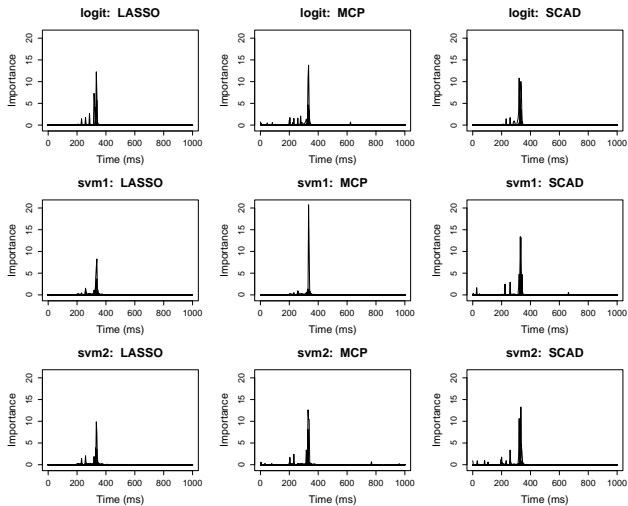
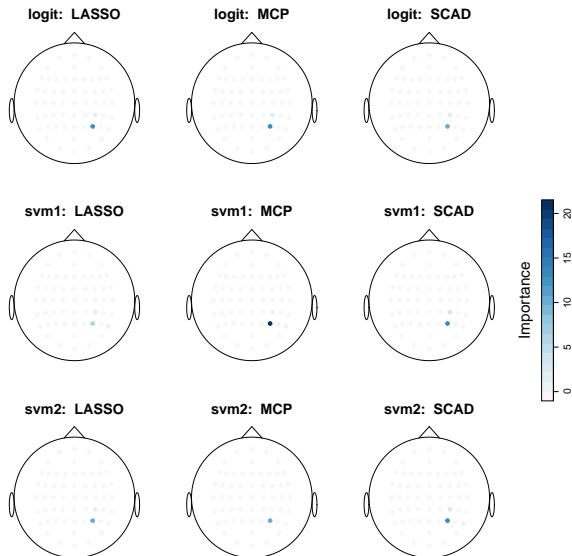


Figure 9: EEG Database: variable importance by time for each channel.

# Results: Variable Importance (EEG cap)



**Figure 10:** EEG Database: average variable importance at time point 333 ms for each channel. The distinguishing activity is localized at electrode P4.

# Results: Model Predictions

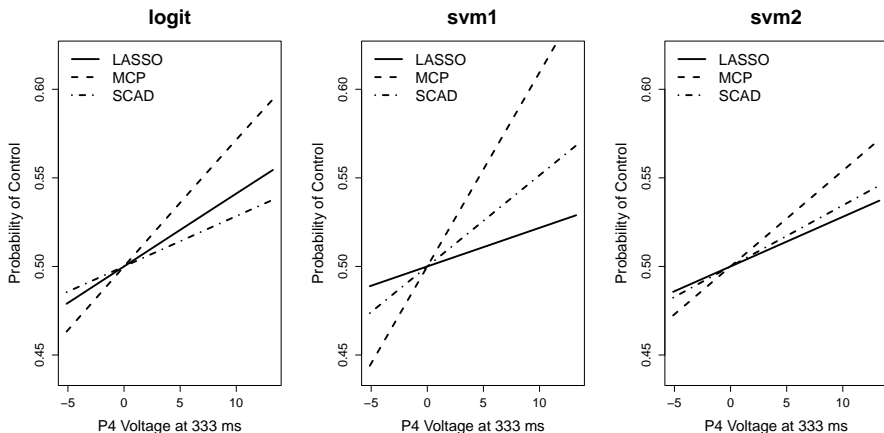


Figure 11: EEG Database: model predictions.

# Overview of Student Performance Dataset

**Student Performance** dataset from UCI Machine Learning Repository

- Published by Cortez (2008) from Cortez & Silva (2008)
- Obligatory UCI citation (Kelly et al., 2025)

Focused on the math performance dataset

- $n = 395$  Portuguese secondary school students
- $m = 3$  math exam scores (G1, G2, G3)
- $p = 30$  sociodemographic features

**Goal:** predict the exams scores from the sociodemographic features<sup>‡</sup>

---

<sup>‡</sup>Adapted from Helwig (2021)

# Response and Predictor Variables

Table 3: Math Performance: predictor variables.

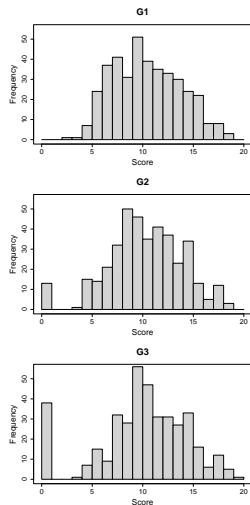


Figure 12: Math Performance: response variables.

Variable	R Class	Spline Type	Range/Levels
school	character	Nominal	GP = Gabriel Pereira, MS = Mousinho da Silveira
sex	character	Nominal	F = female, M = male
age	integer	Cubic	[15, 22]
address	integer	Nominal	U = urban, R = rural
famsize	character	Nominal	LE3 = 3 or less, GT3 = 4 or more
Pstatus	character	Nominal	T = together, A = apart
Medu	integer	Ordinal	0, 1, 2, 3, 4
Fedu	integer	Ordinal	0, 1, 2, 3, 4
Mjob	character	Nominal	teacher, health, services, home, other
Fjob	character	Nominal	teacher, health, services, home, other
reason	character	Nominal	home, reputation, course, other
guardian	character	Nominal	mother, father, other
traveltime	integer	Ordinal	1 = <15m, 2 = 15-30m, 3 = 30-60m, 4 = >60m
studytime	integer	Ordinal	1 = <2hr, 2 = 2-5hr, 3 = 5-10hr, 4 = >10hr
failures	integer	Ordinal	0, 1, 2, 3
schoolsup	character	Nominal	no, yes
famsup	character	Nominal	no, yes
paid	character	Nominal	no, yes
activities	character	Nominal	no, yes
nursery	character	Nominal	no, yes
higher	character	Nominal	no, yes
internet	character	Nominal	no, yes
romantic	character	Nominal	no, yes
famrel	integer	Ordinal	1 = very bad, ..., 5 = very good
freetime	integer	Ordinal	1 = very low, ..., 5 = very high
goout	integer	Ordinal	1 = very low, ..., 5 = very high
Dalc	integer	Ordinal	1 = very low, ..., 5 = very high
Walc	integer	Ordinal	1 = very low, ..., 5 = very high
health	integer	Ordinal	1 = very bad, ..., 5 = very good
absences	integer	Cubic	[0, 75]

Note. Medu and Fedu: 0 = none, 1 = primary, 2 = 5th-9th, 3 = secondary, 4 = higher

# Overview of Data Splitting and Analysis

Multivariate regression predicting exams (G1, G2, G3) from features

- Consider all main effects (30) and two-way interactions (435)
- Use tensor product (spectral) smoothing splines (Helwig, 2024)
- Note: design matrix has 2837 columns after expansion

Use 80/20 (training/testing) data splitting procedure:

- Fit/tune model on training data; evaluate on testing data
- Repeat process 10 times to investigate stability of solution

For each sample of training data:

- 10-fold CV to tune  $\lambda$  and  $\alpha$  (using `lambda.1se` for prediction)
- Compare LASSO, MCP, and SCAD

# Results: Prediction Accuracy

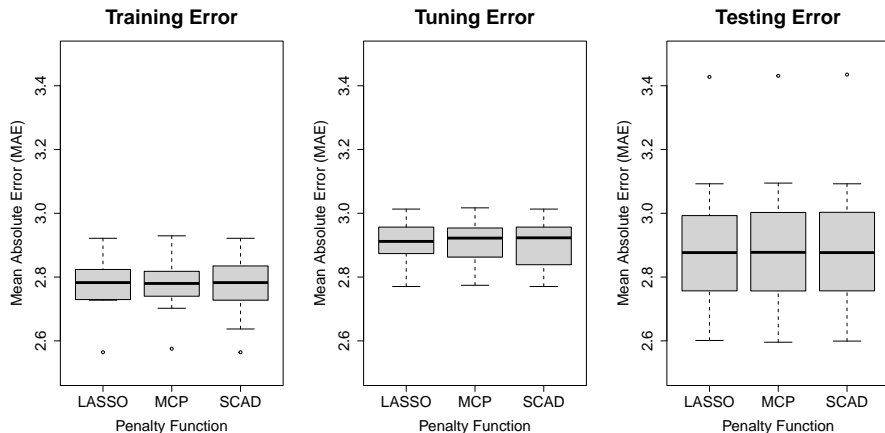


Figure 13: Math Performance: prediction accuracy.

# Results: Variable Selection

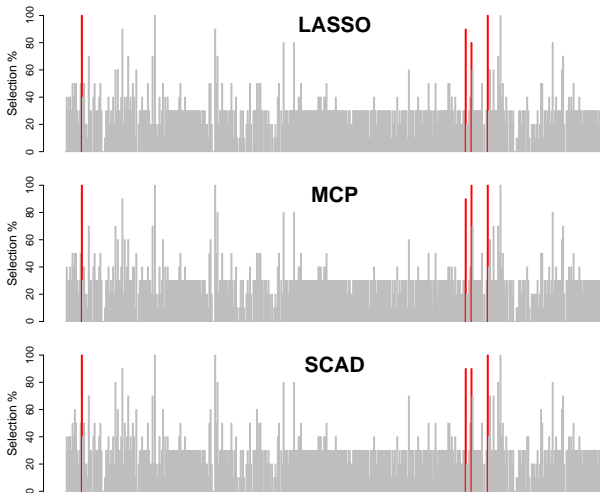


Figure 14: Math Performance: variable selection.

# Results: Variable Importance

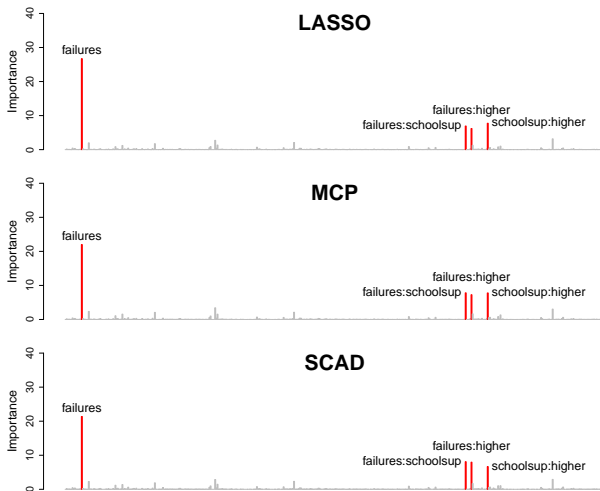


Figure 15: Math Performance: variable importance.

## Results: Model Predictions

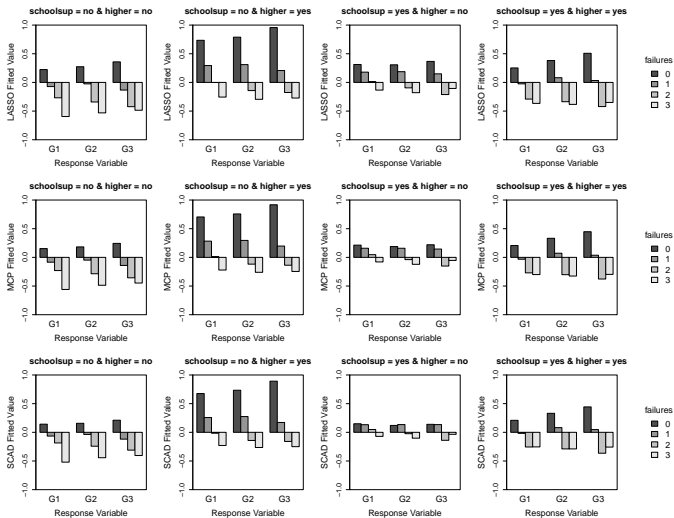


Figure 16: Math Performance: model predictions.

# Overview of ABCD Study Data

ABCD Study - baseline sample of 9–11 year olds

- Brain: resting-state connectivity (rs-fMRI) between 12 networks
- Behavior: Attention Problems subscale from CBCL

Considered all pairwise connections (including self-connections) and covariates: site, age, sex, race, income, education, alcohol, drugs

- $n = 7979$  subjects ages 9–11 years old
- $p = 86$  features (8 covariates + 78 rs-fMRI)

**Goal:** predict attention problems from resting-state connectivity<sup>§</sup>

---

<sup>§</sup>Adapted from Duffy & Helwig (2024)

# Response and Predictor Variables

CBCL Attention Problems Distribution

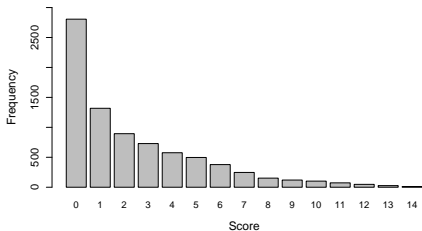


Figure 17: The attention problem subscores (from the CBCL).

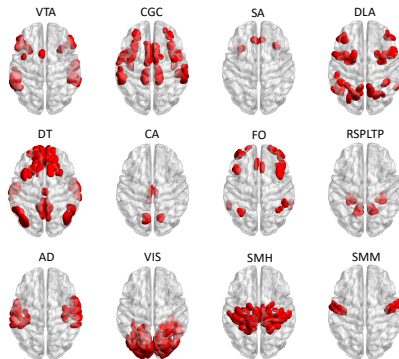


Figure 18: The 12 functional brain networks (from the Gordon parcellation).

# Overview of Data Splitting and Analysis

Poisson loss function (b/c Attention Problems are count data):

- Consider only the 86 main effects
- Use additive (spectral) smoothing splines (Helwig, 2024)
- Note: design matrix has 433 columns after expansion

Use 80/20 (training/testing) data splitting procedure:

- Fit/tune model on training data; evaluate on testing data
- Repeat process 10 times to investigate stability of solution

For each sample of training data:

- 10-fold CV to tune  $\lambda$  and  $\alpha$  (using `lambda.1se` for prediction)
- Compare LASSO, MCP, and SCAD

# Results: Prediction Accuracy

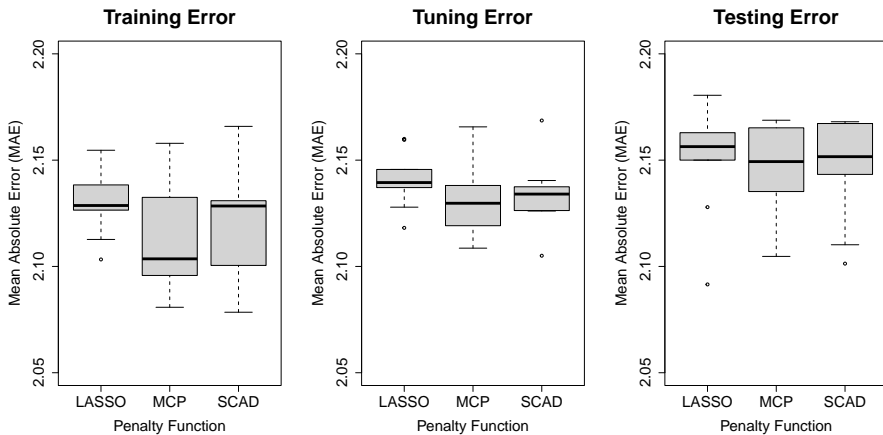


Figure 19: ABCD Data: prediction accuracy.



# Results: Variable Importance

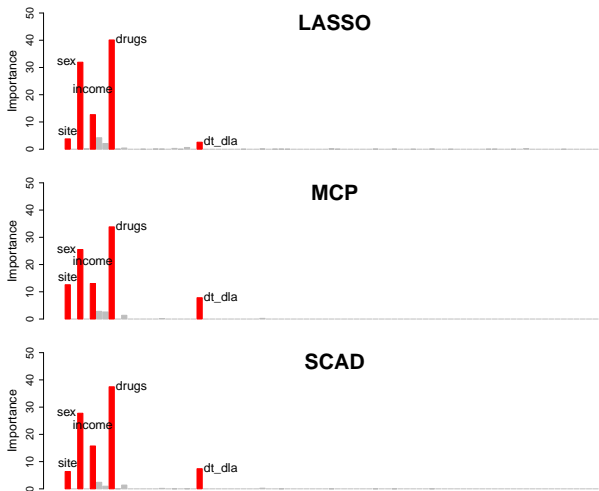


Figure 21: ABCD Data: variable importance (dt\_dla = default mode  $\times$  dorsal lateral attention).

# Results: Model Predictions

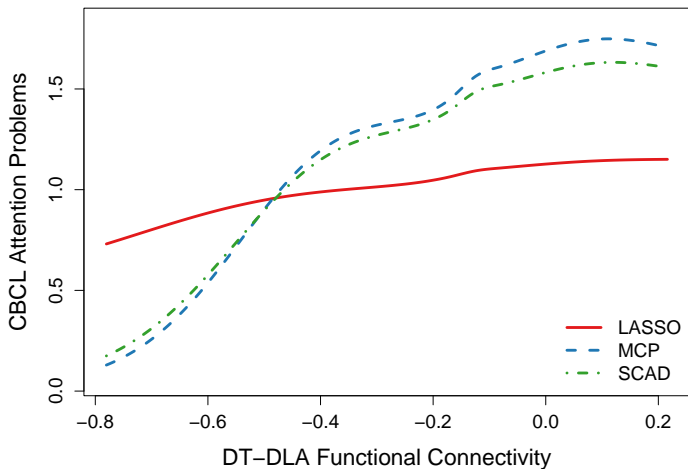


Figure 22: ABCD Data: model predictions.

# Table of Contents

1. Overview of Statistical Learning Problem
2. ABGD Algorithm and `grpnet` R Package
3. Real Data Examples
  - Genes: Cancer GWAS
  - Brains: EEG Classification
  - Behavior: Math Performance
  - Brains + Behavior: ABCD Data
4. Conclusions and Future Directions

## Take-Home Points and Extensions

With HD data, model selection and interpretation are a challenge

- Which terms to consider? Nonlinearities? Interactions?
- Too burdensome to consider all available options...

GRPNET provides interpretable machine learning for HD regression

- Finds main and interaction effects without prior assumptions
- Blends strengths of parametric and nonparametric regression

Extensions for ordinal and multivariate outcomes:

- Ordinal outcomes are coming soon (using group penalized POLR)
- Multivariate outcomes with user-specified components is TBD...

# References and Funding

## References:

- Begleiter, H. (1995). *EEG Database*. UCI Machine Learning Repository. (DOI: <https://doi.org/10.24432/C5TS3D>)
- Cortez, P. (2008). *Student Performance*. UCI Machine Learning Repository. (DOI: <https://doi.org/10.24432/C5TG7T>)
- Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In A. Brito & J. Teixeira (Eds.), *Proceedings of 5th FUTURE BUSINESS TECHNOLOGY conference (FUBUTEC 2008)* (p. 5-12). Porto, Portugal: EUROSIS.
- Duffy, K. A., & Helwig, N. E. (2024). Resting-state functional connectivity predicts attention problems in children: Evidence from the abcd study. *NeuroSci*, 5(4), 445–461. doi: 10.3390/neurosci5040033
- Fiorini, S. (2016). *gene expression cancer RNA-Seq*. UCI Machine Learning Repository. (DOI: 10.24432/C5R88H)
- Helwig, N. E. (2021). Spectrally sparse nonparametric regression via elastic net regularized smoothers. *Journal of Computational and Graphical Statistics*, 30(1), 182-191. doi: 10.1080/10618600.2020.1806855
- Helwig, N. E. (2024). Precise tensor product smoothing via spectral splines. *Stats*, 7(1), 34-53. doi: 10.3390/stats7010003
- Helwig, N. E. (2025). grpnet: Group elastic net regularized GLMs and GAMs [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=grpnet> (R package version 1.0) doi: <https://doi.org/10.32614/CRAN.package.grpnet>
- Helwig, N. E. (2025). Versatile descent algorithms for group regularization and variable selection in generalized linear models. *Journal of Computational and Graphical Statistics*, 34(1), 239-252. doi: 10.1080/10618600.2024.2362232
- Kelly, M., Longjohn, R., & Nottingham, K. (2025). *The UCI Machine Learning Repository*. <https://archive.ics.uci.edu>.
- Mohammed, S., Dey, D. K., & Zhang, Y. (2020). Classification of high-dimensional electroencephalography data with location selection using structured spike-and-slab prior. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(5), 465-481. doi: <https://doi.org/10.1002/sam.11477>

Funded by a sabbatical award from Minnesota and NIH grants:

R01EY030890, R01MH115046, U01DA046413