

Classifying Alcoholism from Electroencephalography Data using High-Dimensional Penalized Regression

Jong Won Lee and Nathaniel E. Helwig

Department of Psychology
School of Statistics
University of Minnesota

International Meeting of the Psychometric Society
University of Minnesota, Twin Cities
July 16, 2025



Motivating Problem

Alcoholics and controls are well-known to differ in magnitudes of P300

- Reduced P300 amplitude in occipital-parietal areas for alcoholics
- See Begleiter and Porjesz (1999) for more details

EEG data from the UCI Machine Learning Repository (Kelly et al., 2025)

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

Goal: identify the time points and channels that distinguish alcoholics

Details on EEG Data

$n = 122$ subjects

- 77 alcoholics
- 45 controls

$p = 14592$ features

- 256 time points
(1 sec at 256 Hz)
- 57-channel cap

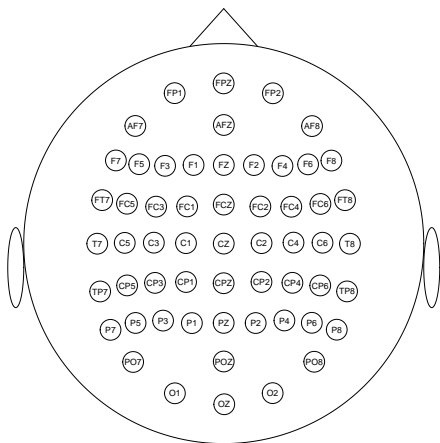


Figure 1: The 57 channel EEG cap.

Penalized Regression Framework

Consider the following definitions:

- $y_i \in \{-1, +1\}$: class label for the i -th observation
- $\mathbf{x}_i \in \mathbb{R}^p$: feature vector for the i -th observation
- $\eta(\mathbf{x}_i) = \boldsymbol{\beta}^\top \mathbf{x}_i$: prediction function

To estimate $\boldsymbol{\beta}$, we use a penalized regression approach

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n \mathcal{L}(y_i, \eta(\mathbf{x}_i)) + \lambda \sum_{j=1}^p \mathcal{P}_\alpha(\beta_j)$$

where

- $\mathcal{L}(\cdot, \cdot)$ is the loss function and $\mathcal{P}_\alpha(\cdot)$ is the penalty function
- $\lambda \geq 0$ and $\alpha \in [0, 1]$ are tuning parameters

Visualization of Loss Functions

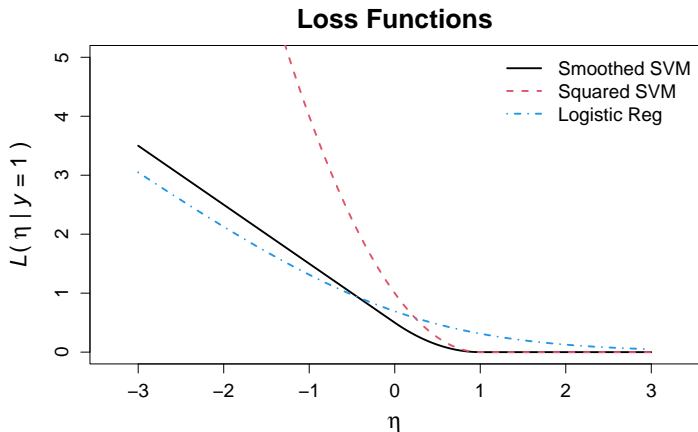


Figure 2: The three binary loss functions: logistic, smoothed SVM, squared SVM.

Visualization of Penalty Functions

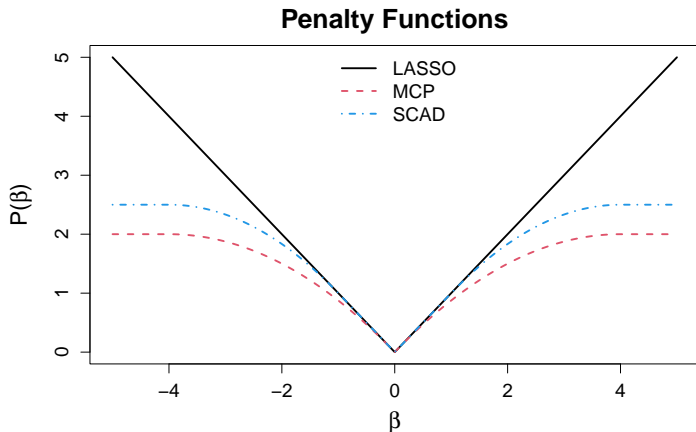


Figure 3: The three L1-based penalty functions: LASSO, MCP, SCAD.

Analysis Procedure

Use 70/30 train/test split with 10 replications to assess stability

- Balance split based on class proportions
- 5-fold cross validation to tune λ and α

For each training sample, consider all combinations of . . .

- 3 loss functions (logistic, smoothed SVM, squared SVM)
- 3 L1-based penalty functions (LASSO, MCP, SCAD)

Implement analyses using **grpnet** R package (Helwig, 2025).

Misclassification Error

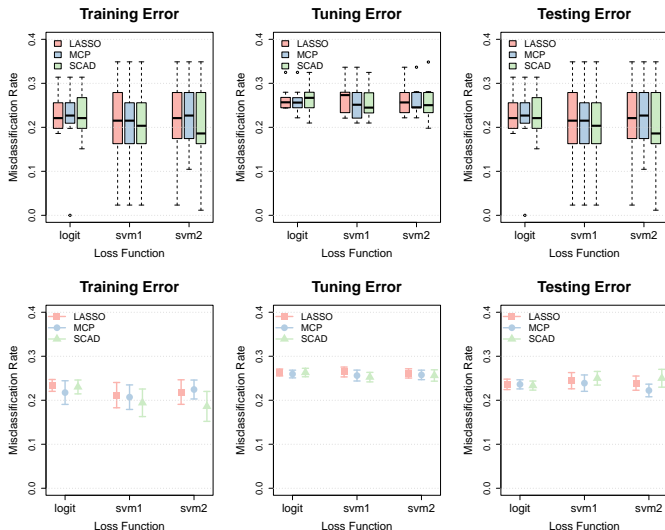


Figure 4: Misclassification rate boxplots (top) and average \pm one SE (bottom).

Variable Selection

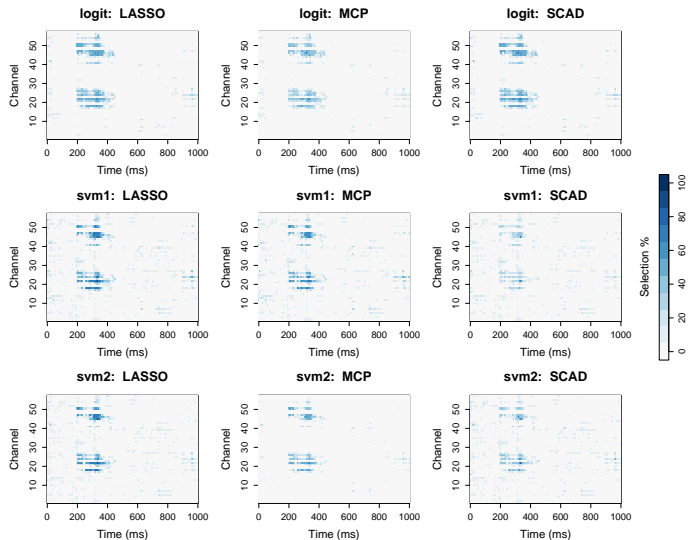


Figure 5:
Variable selection percentage for each time-by-channel.

Variable Importance: Heatmap

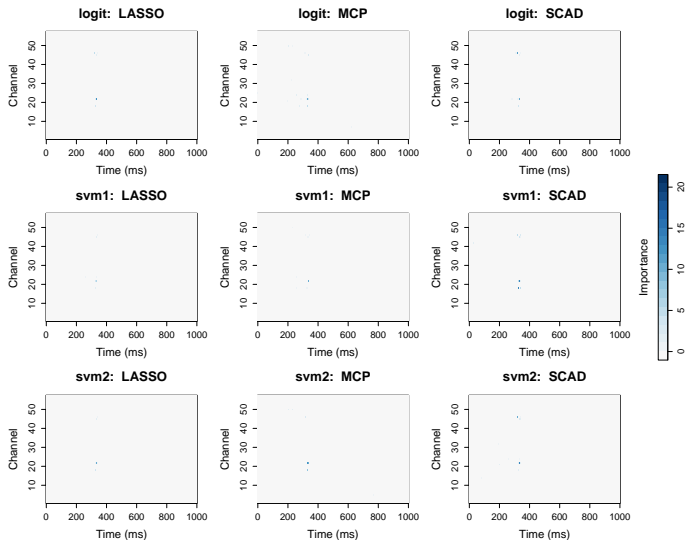


Figure 6:
Average variable importance for each time-by-channel.

Variable Importance: Line Plot

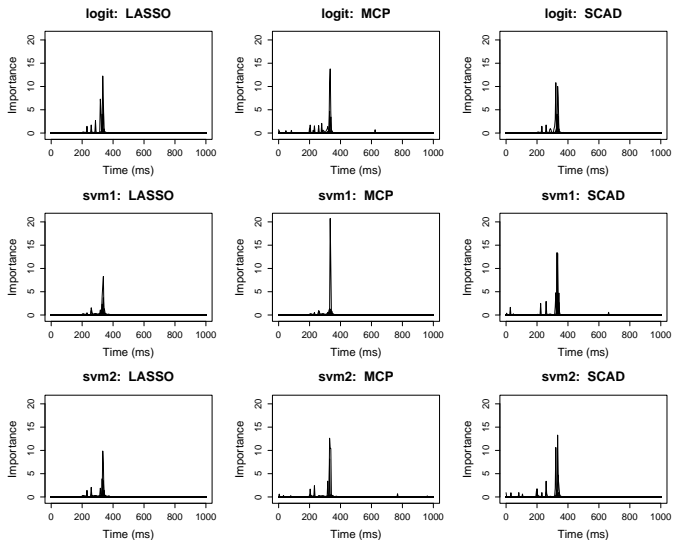


Figure 7:
Average variable importance by time for each channel.

Variable Importance: EEG Cap

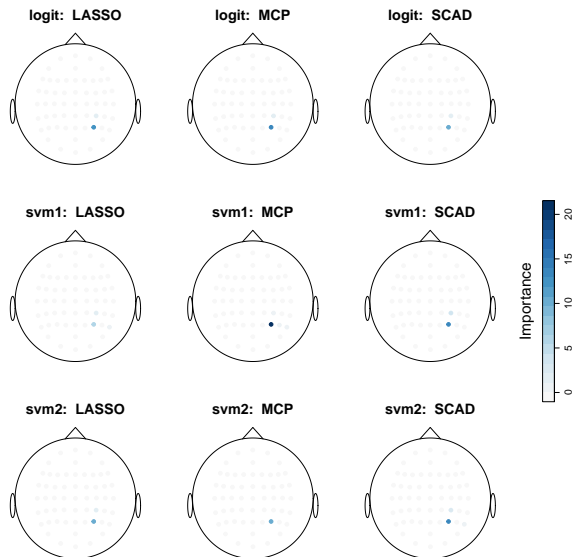


Figure 8: Average variable importance at time point 333 ms for each channel. The distinguishing activity is localized at electrode P4.

Conclusions and Extensions

Primary Findings:

- P4 activation around 333 ms distinguished alcoholics from controls
- Several small effects in the 200-400 ms region are consistently selected

Secondary Findings:

- All three loss functions produced similar classification performance
- All three penalty functions produced similar selection performance

Future Directions:

- Incorporate spatiotemporal information into classification features
- Apply approach to other types of neuroimaging data, e.g., fMRI

References

- Begleiter, H. (1995). EEG Database. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5TS3D>.
- Begleiter, H. and B. Porjesz (1999). What is inherited in the predisposition toward alcoholism? a proposed model. *Alcoholism: Clinical and Experimental Research* 23, 1125–1135.
- Helwig, N. E. (2025). *grpnet: Group Elastic Net Regularized GLMs and GAMs*. R package version 1.0.
- Kelly, M., R. Longjohn, and K. Nottingham (2025). The UCI Machine Learning Repository. <https://archive.ics.uci.edu>.
- Mohammed, S., D. K. Dey, and Y. Zhang (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.