Project 3 High Dimensional Data Analysis

Boyi Guo

BST 765

2021-04-21

Objective

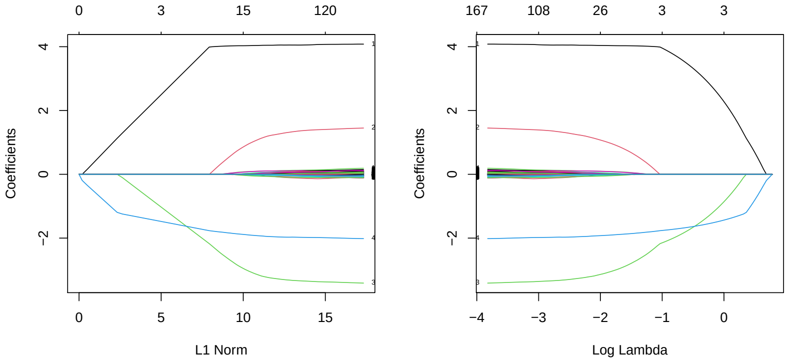
- To introduce the concept *solution path*
- To understand the behavior of Ridge penalty, LASSO penalty, Minimax Concave Penalty(MCP), Spike-and-slab LASSO, EMVS
- To demonstrate data standardization matters in penalized model

Solutin Path

- Variable selection via regularization/penalization is continuous process
  - in comparison to step-wise selection
  - coefficient estimate is a function of tuning parameter, e.g. Ridge regression
- Solution path plots the coefficient estimate across different values of tuning parameter

Solution Path Example

Two forms for tuning parameter of LASSO: L1 Norm $\sum |\beta_i|$ VS Shrinkage parameter λ



Simulation Study

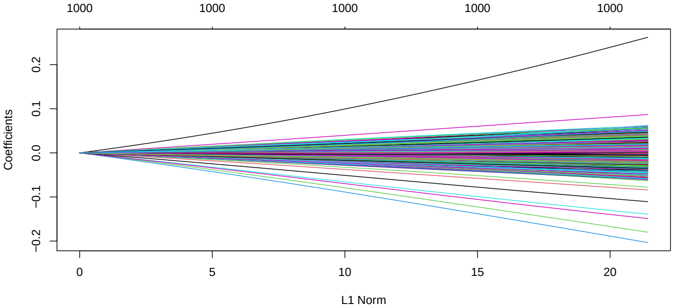
- High dimension setting ($p >> n$)
- Highly correlated predictors
- Sparse signal (4/1000 active predictor)
- Examine the solution path

$$
\begin{aligned}
i &= 1, \ldots, 200 \\
X_i &\sim N_{1000}(\mathbf{0}, \Sigma_{AR(0.8)}) \\
\boldsymbol{\beta} &= \begin{pmatrix} 4 & 2 & -4 & -2 & \underbrace{0 \quad \cdots \quad 0}_{996} \end{pmatrix}^T \\
y_i &= \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \epsilon_i \sim N(0, 1)
\end{aligned}
$$

- One iteration, 10-fold cross-validation

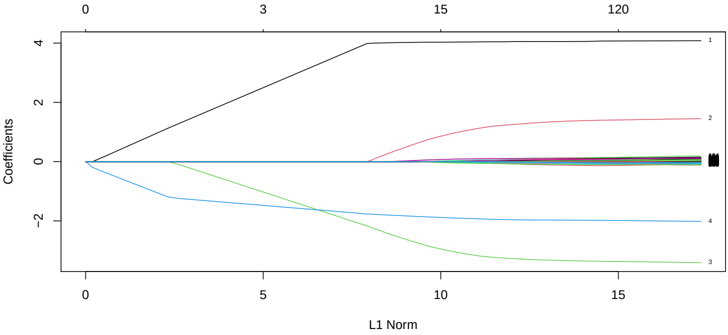
Ridge

- Designed to solve collinearity problem
- Doesn't work well for high-dimensional setting as the coefficients doesn't shrink to zero
  - extremely biased estimates



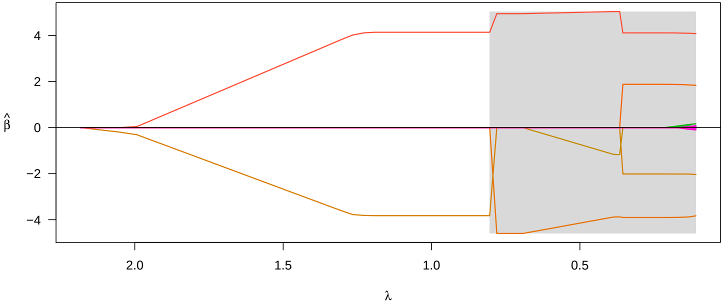
LASSO

- Assumption: signals are sparse, i.e. small amount of non-zero coefficient
- LASSO include the "truth" as an subset $\beta \subseteq \hat{\beta}_{LASSO}$
- Cross-validated model select more than 20 predictors



MCP

- Fancier LASSO
- Less biased estimates, faster coefficient Stabilization
- Extra parameter γ

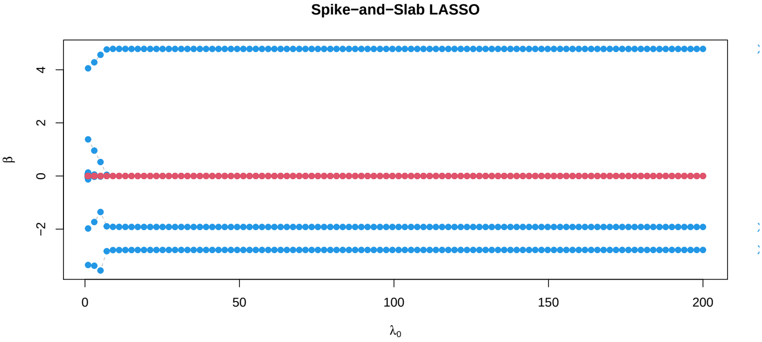


EMVS

- Spike-and-slab Mixture Normal Prior with Mximum A Posteri Estimate
- More complicated variable selection, depending on a soft threshold

Spike and Slab Lasso

- Spike-and-slab Mixture Double exponential Prior with Maximum A Posteri Estimate

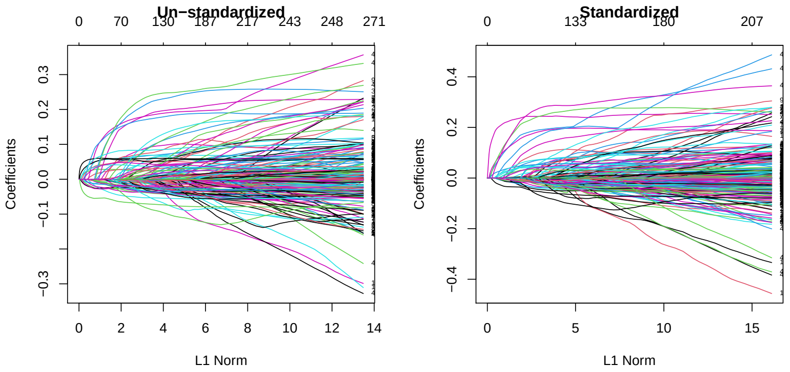


SNP data

Characteristic	**Overall**, N = 709	**Male**, N = 307	**Female**, N = 40
Age	58 (47, 68)	56 (46, 66)	59 (48, 69)
HB_CAT			
low	65 (9.2%)	29 (9.4%)	36 (9.0%)
med	383 (54%)	158 (51%)	225 (56%)
high	261 (37%)	120 (39%)	141 (35%)
firstbleed	100 (14%)	48 (16%)	52 (13%)
T1	503 (199, 749)	421 (169, 749)	561 (235, 749)
T2r	511 (210, 753)	429 (181, 752)	564 (240, 754)

LASSO models

Un-standardized design matrix VS Standardized



Closing Remarks

- Know when to standardize your data before model fitting
  - Most of time but not all the time
- 1-SE rule when selecting tuning parameter
- Know when and why to use validation/nested validation
  - Model selection VS Model assessment
  - To estimate in-sample error / extra-sample error