Regularized Linear Regression

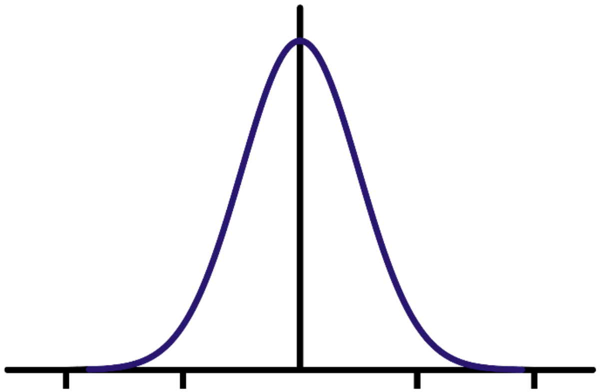
Signal Data Science

Motivation for regularization

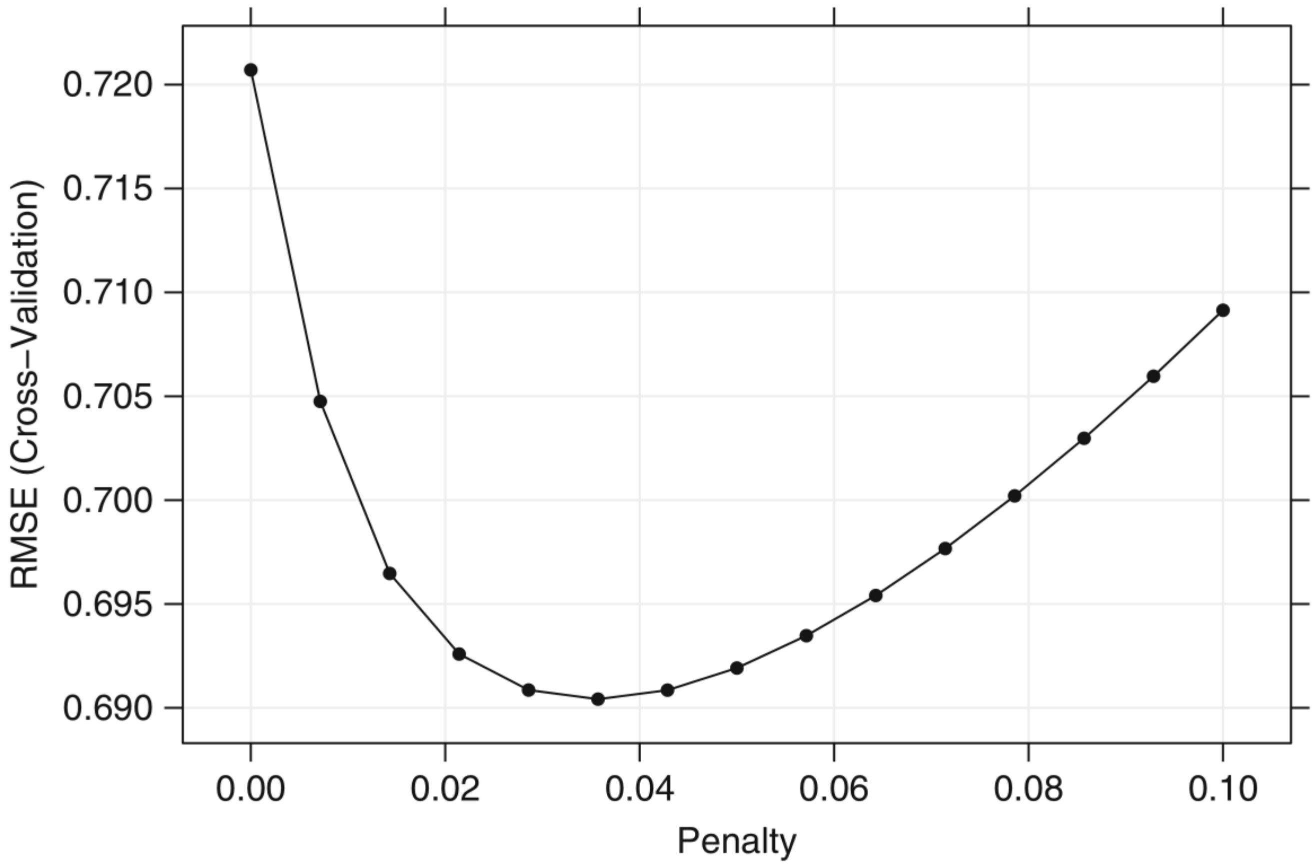
- MSE = (intrinsic error)^2 + (bias)^2 + variance
- Ordinary least squares is unbiased
  - Isn’t the model with the lowest MSE in general
- Introducing some bias can reduce variance
- Two major problems result in inflated coefficients
  - Collinearity among predictors
  - Overfitting

Two equivalent formulations

- Penalize large coefficients
  - Instead of minimizing sum of squared errors (SSE), minimize SSE + λ*sum(|coefficients|)
  - Or minimize SSE + λ*sum(|coefficients|^2)
- Impose Bayesian prior on coefficients
  - Use a Gaussian or Laplacian prior for coefficients being closer to 0



Penalty leads to lower RMSE

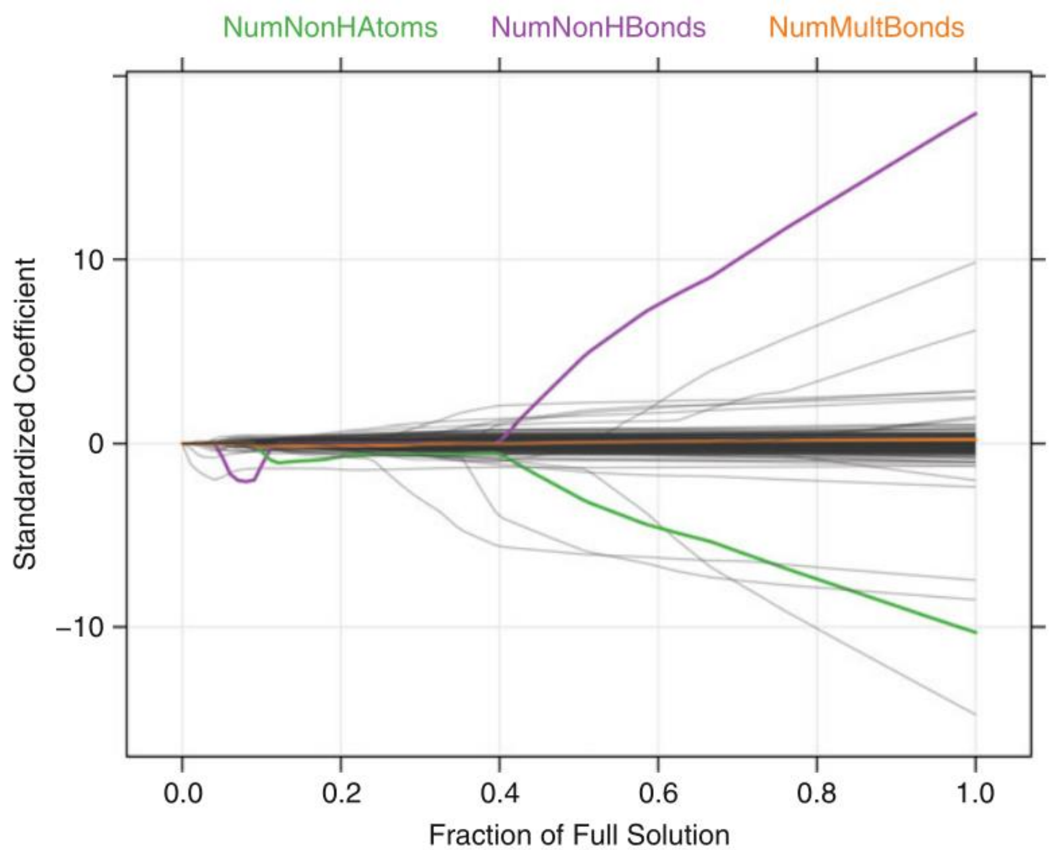


(Applied Predictive Modeling, p. 125)

Lasso and ridge regression

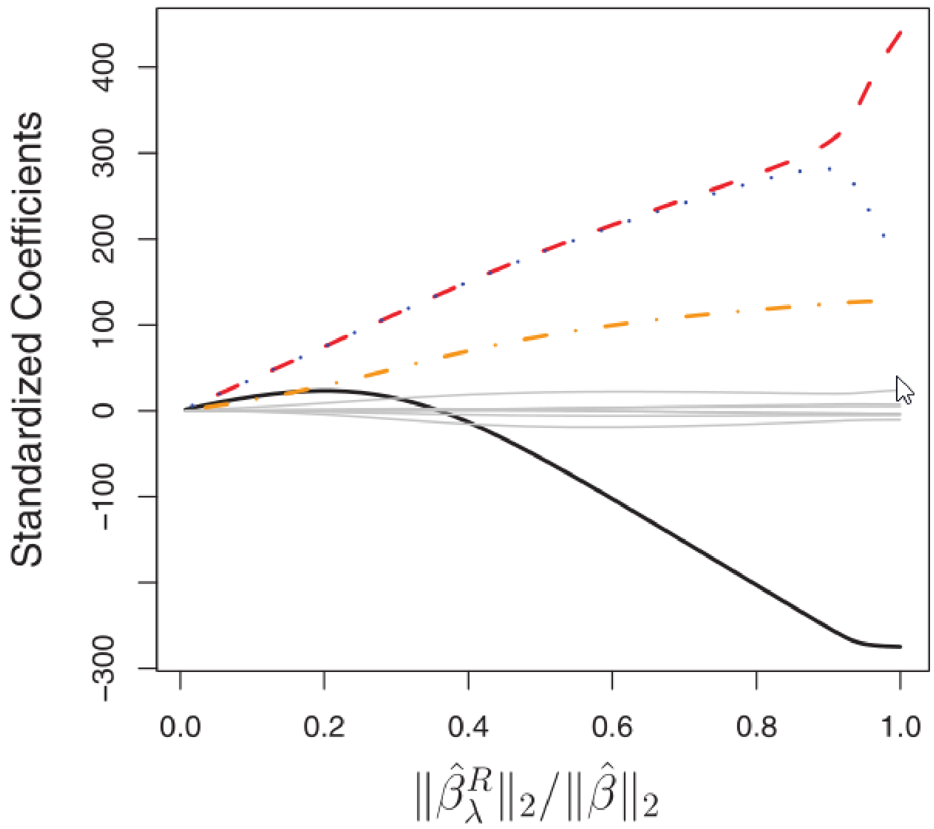
- Minimizing SSE + λ*sum(|coefficients|)
  - Called "LASSO" or "L^1 penalization"
  - Tends to shrink some coefficients to 0 and leave others
  - Easy to interpret
- Minimizing SSE + λ*sum(|coefficients|^2)
  - Called "ridge regression" or "L^2 penalization"
  - Tends to shrink coefficients uniformly
- "L^p penalization" comes from notion of *p*-norm
  - $|x|_p = (|x_1|^p + |x_2|^p + \ldots + |x_n|^p)^{1/p}$

L^1 coefficient shrinkage



(Applied Predictive Learning, p. 126)

L^2 coefficient shrinkage

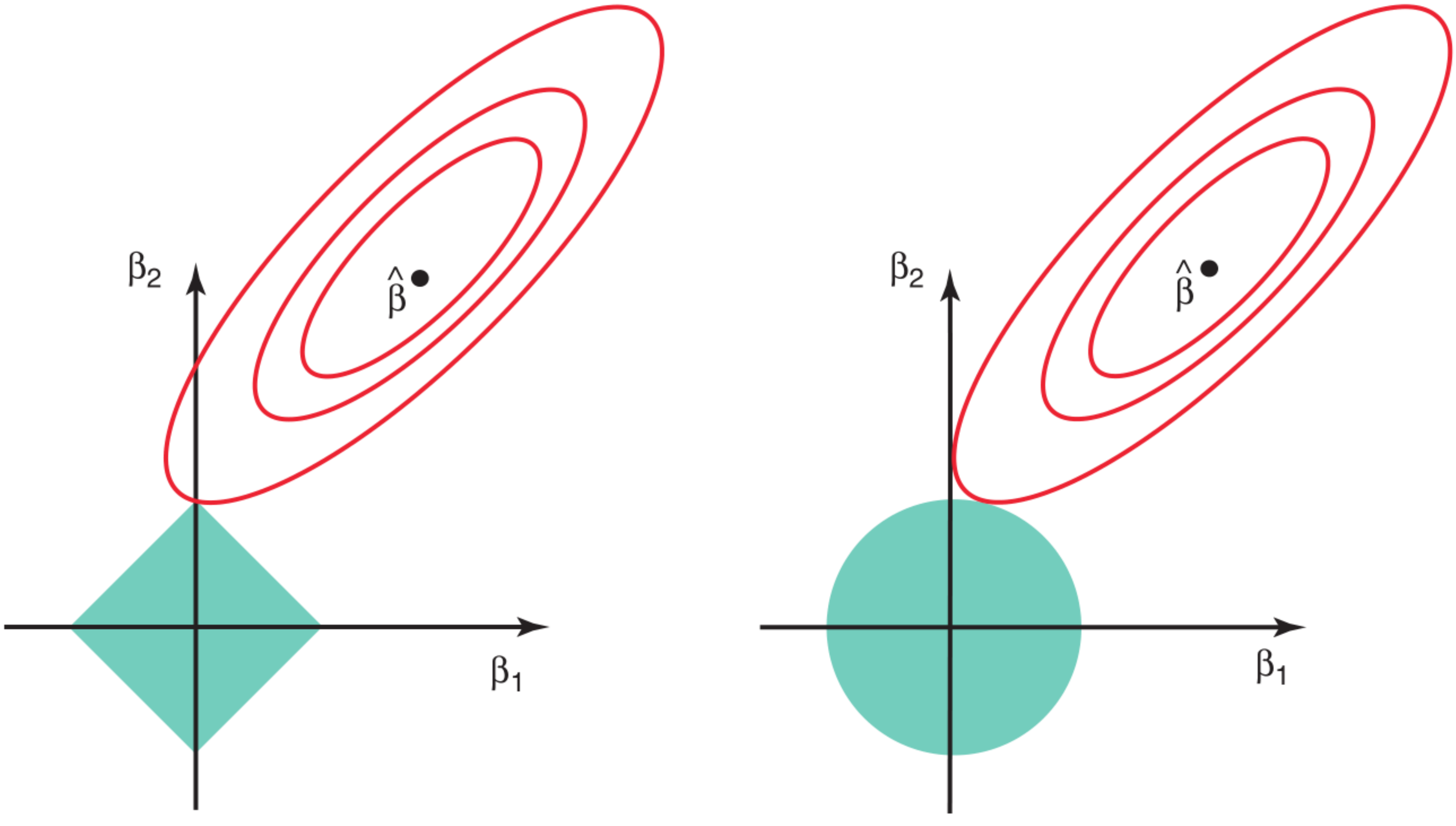


(Introduction to Statistical Learning, p. 216)

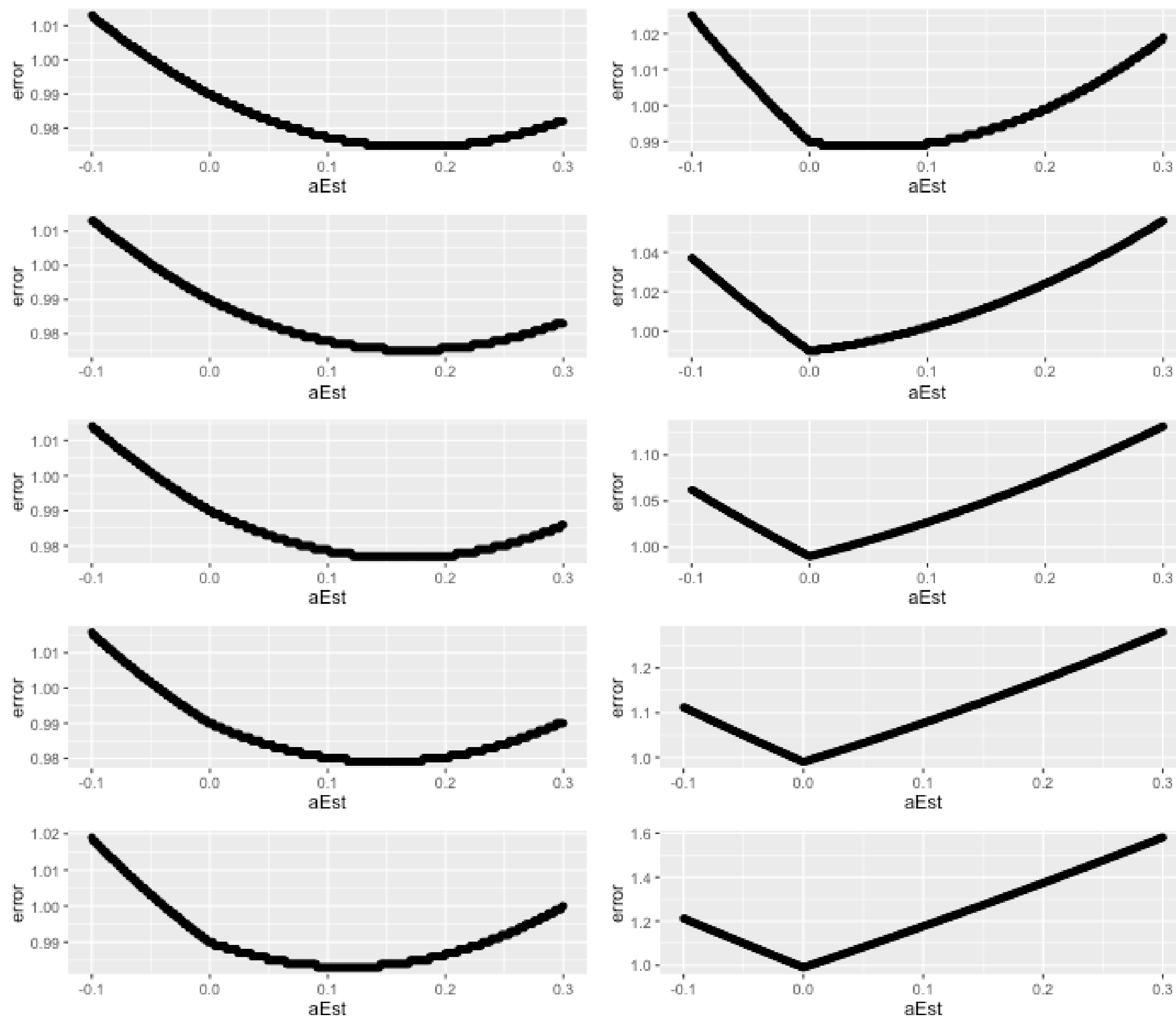
Duality of optimization

- Two equivalent mathematical formulations
  - Minimizing SSE + λ*sum(|coefficients|)
  - Minimizing SSE *subject to* sum(|coefficients|) ≤ $s_1(\lambda)$
- Same for ridge regression
  - Minimizing SSE + λ*sum($|\text{coefficients}|^2$)
  - Minimizing SSE *subject to* sum($|\text{coefficients}|^2$) ≤ $s_2(\lambda)$

Visual intuition



L^1 regularization



L^2 regularization