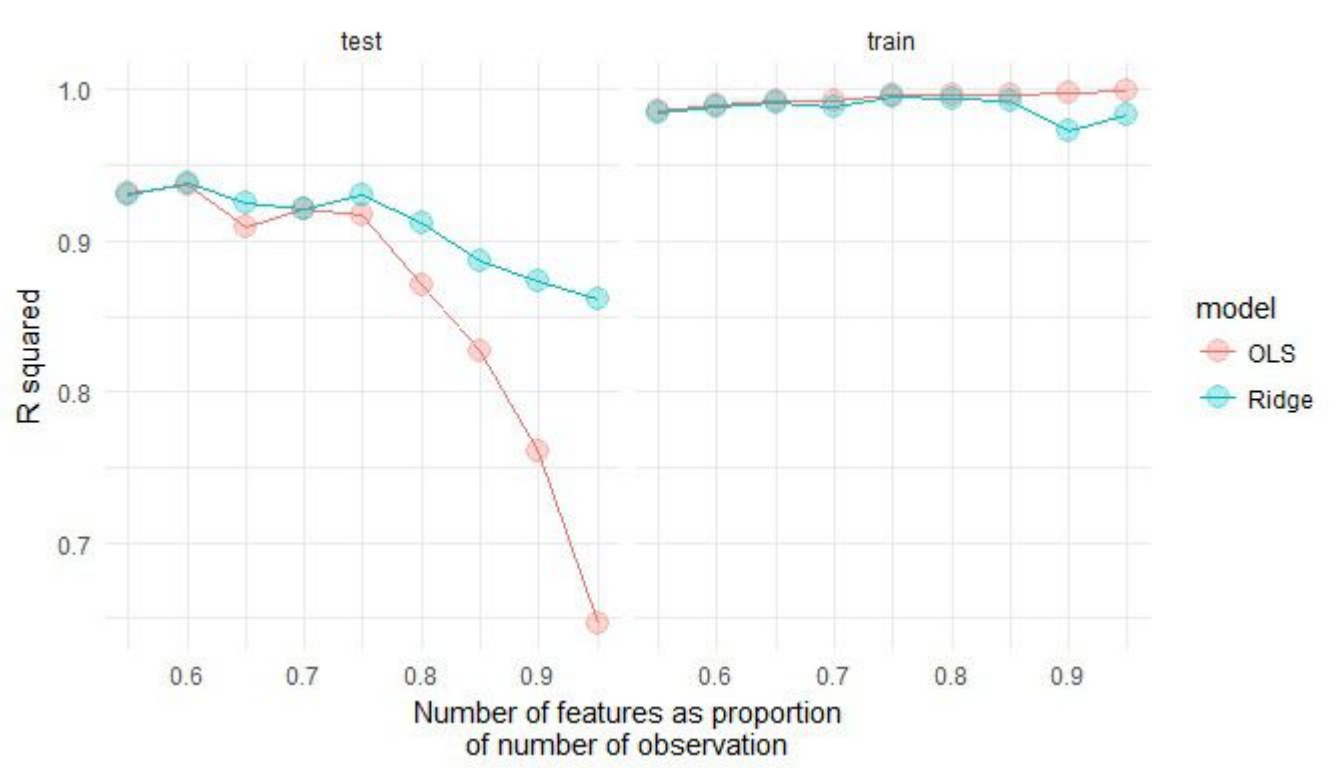
RIDGE REGRESSION VS ORDINARY LEAST SQUARES (OLS)

- Random predictors have Gaussian effects
  - From 55% to 95% of predictors have an effect
- Ridge regression does better in test data
  - Particularly when n is small
- OLS slightly better in training data

From:
https://drsimonj.svbtle.com/ridge-regression-with-glmnet

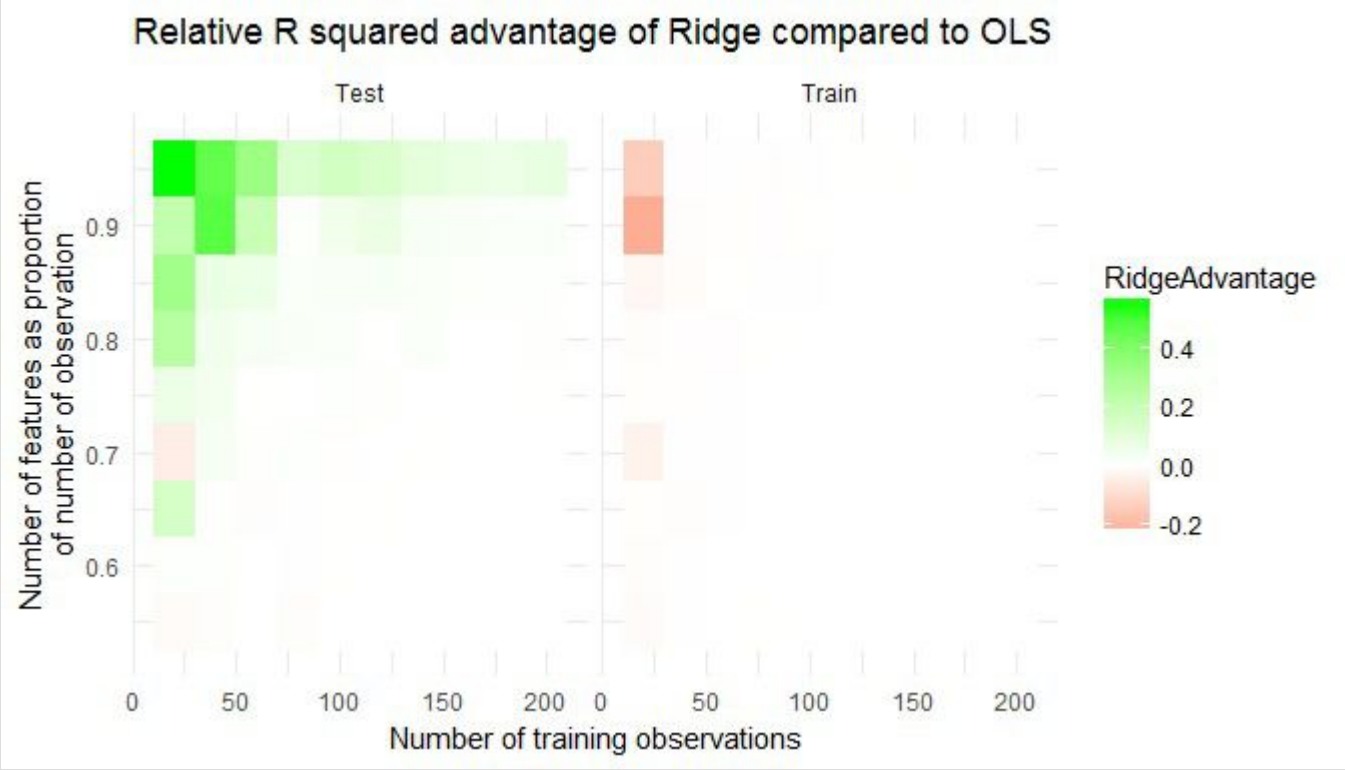
RIDGE REGRESSION VS ORDINARY LEAST SQUARES (OLS)



- Random predictors have Gaussian effects
  - From 55% to 95% of predictors have an effect
- Ridge regression does better in test data
  - Particularly when p is large (compared to n)
- OLS slightly better in training data

From:
https://drsimonj.svbtle.com/ridge-regression-with-glmnet

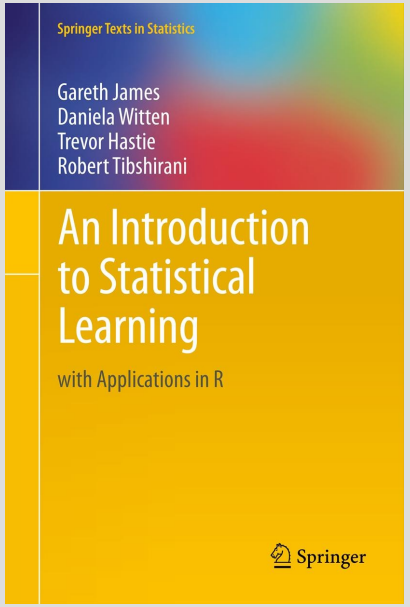
RIDGE REGRESSION VS ORDINARY LEAST SQUARES (OLS)



- Random predictors have Gaussian effects
  - From 55% to 95% of predictors have an effect
- Ridge regression does better in test data
  - Particularly when p is large and / or n is small
- OLS slightly better in training data
  - Overfits particularly when p is large and / or n is small

From:
https://drsimonj.svbtle.com/ridge-regression-with-glmnet

6.2 SHRINKAGE METHODS



Section 6.2

http://www-bcf.usc.edu/~gareth/ISL/

PENALIZED LIKELIHOOD FORMULATION

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

Ridge regression

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|$$

LASSO

How would you write AIC or BIC in this formulation?

CONSTRAINED MINIMIZATION FORMULATION

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

Ridge regression

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

LASSO

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s.$$

Best subset selection. LASSO provides efficient approximation for this

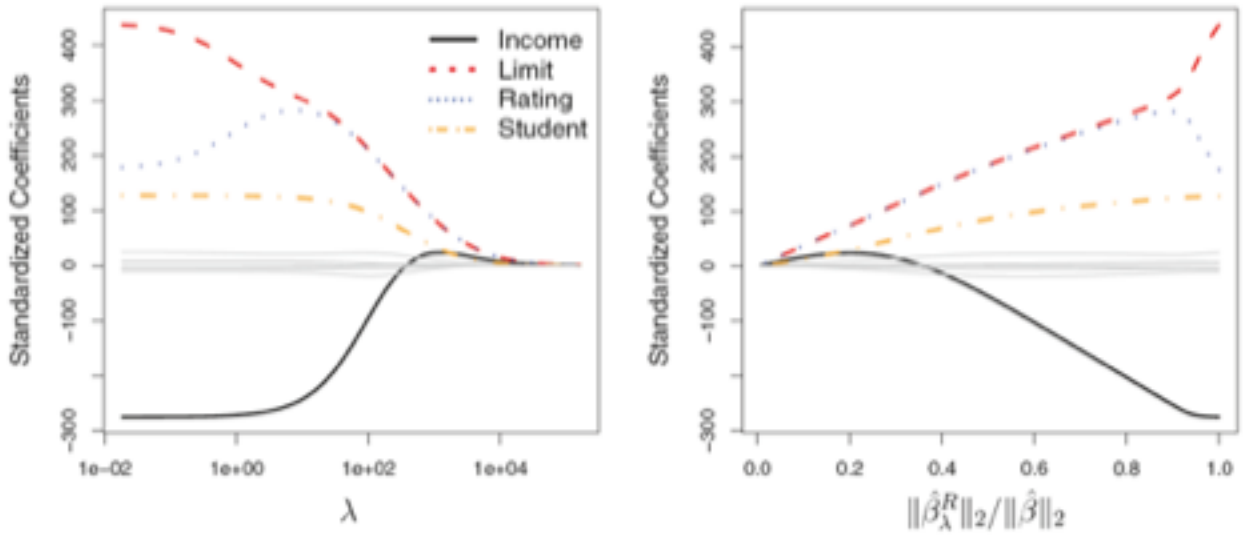


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

Ridge regression:
Coefficients shrunk but never to 0
Robust estimates but not sparse models

LASSO:
Coefficient become exactly 0
at some point
Sparse models good for
interpretability

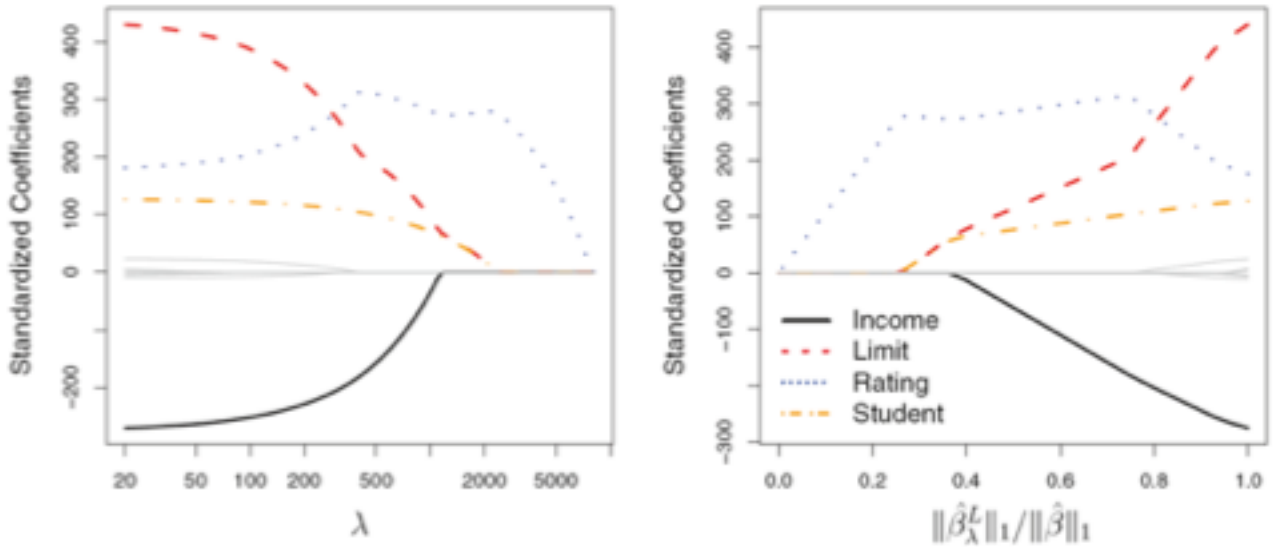


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

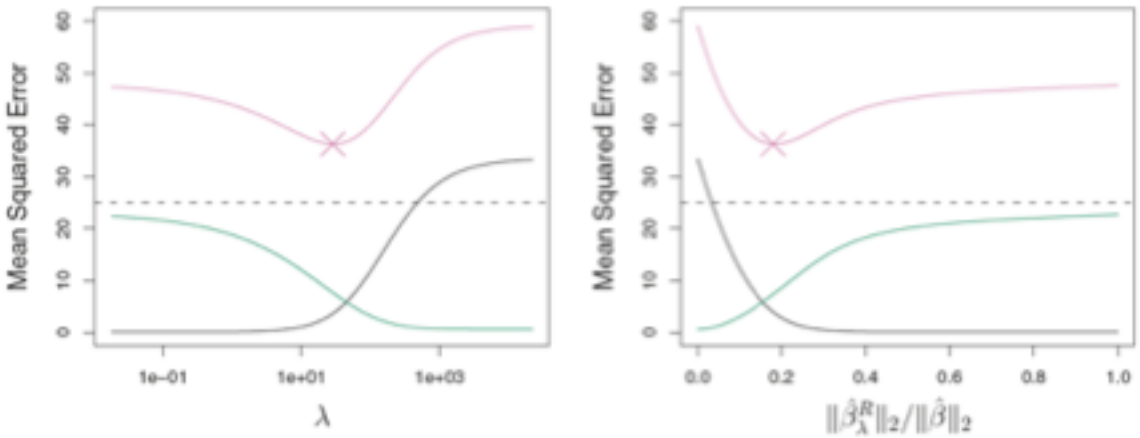


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$*. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Data where 45 predictors all have effects

Ridge has lower MSE than LASSO because Model is not sparse and therefore LASSO is not as good as ridge regression

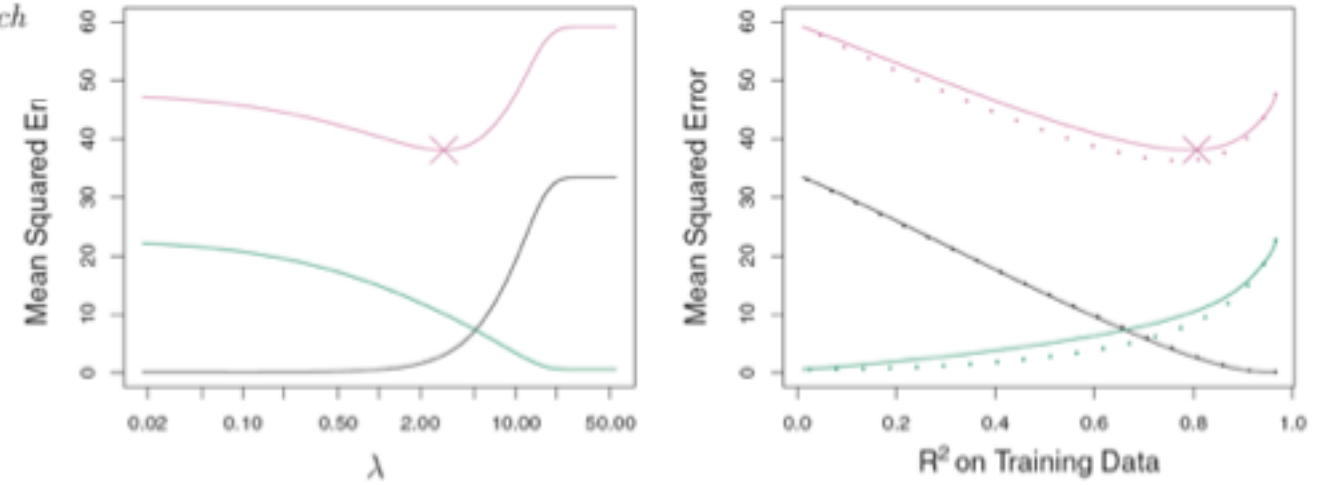


FIGURE 6.8. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their* R^2 *on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

ADVANTAGE OF LASSO IN SPARSE MODEL

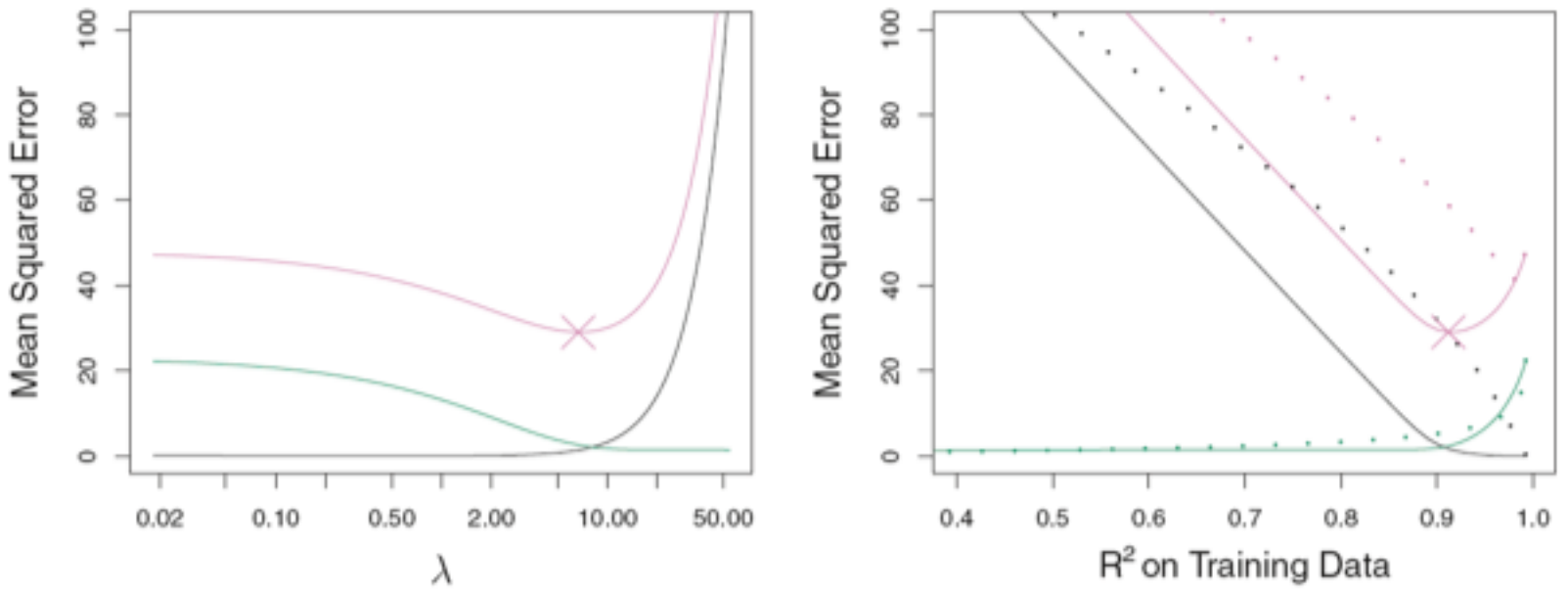


FIGURE 6.9. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their* R^2 *on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

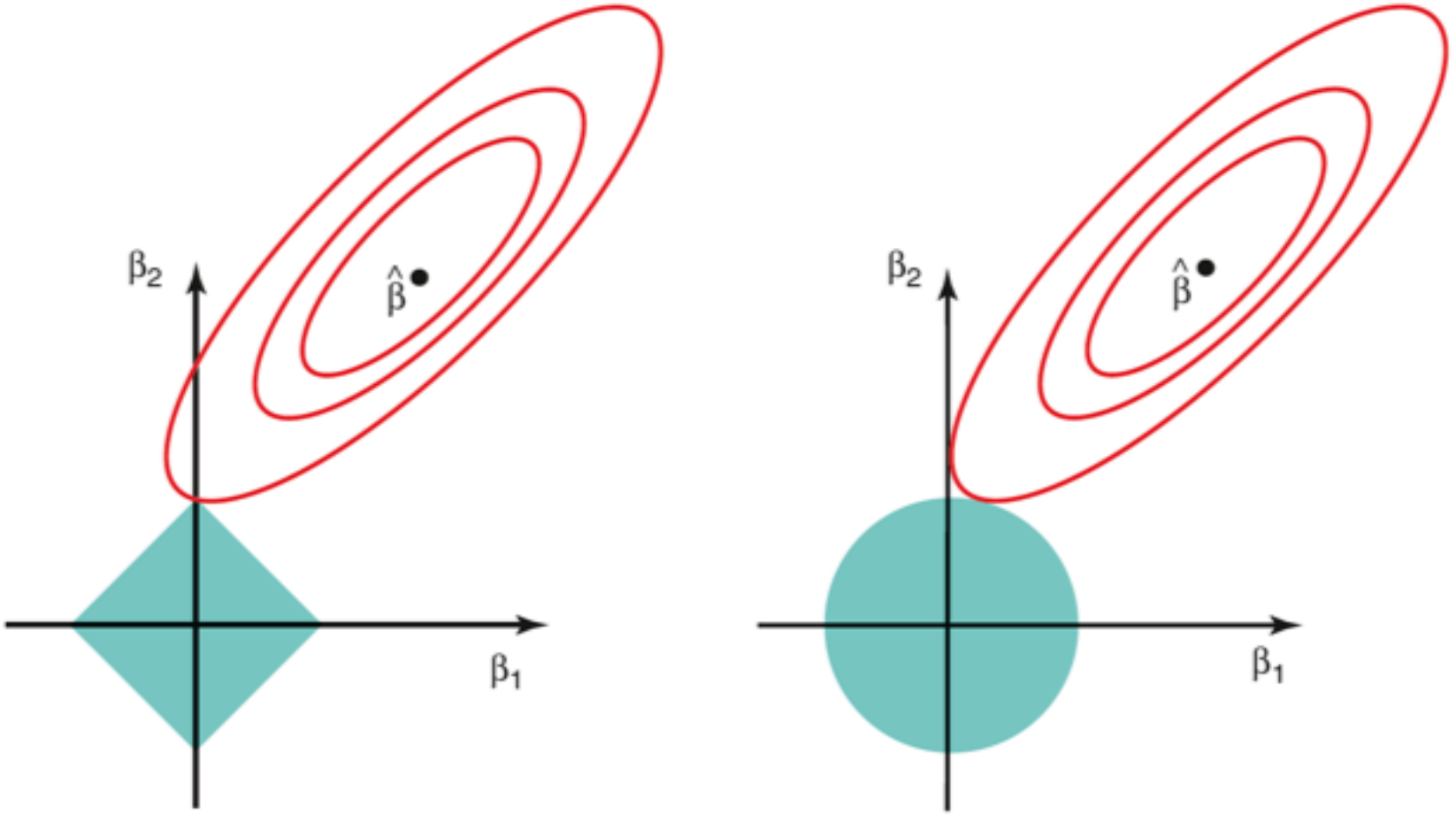
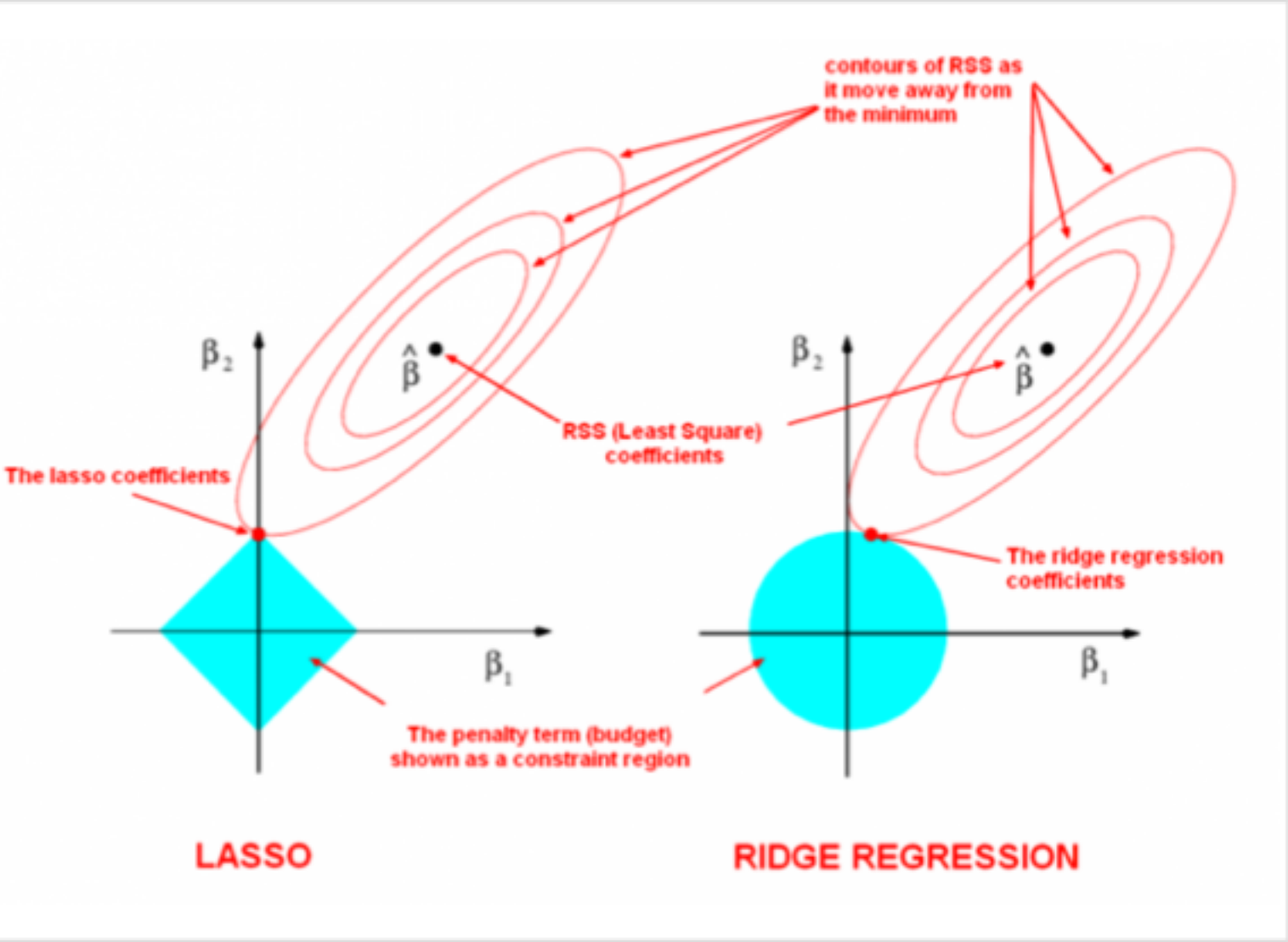


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Why does LASSO do variable selection and ridge does not?



LASSO produces sparsity

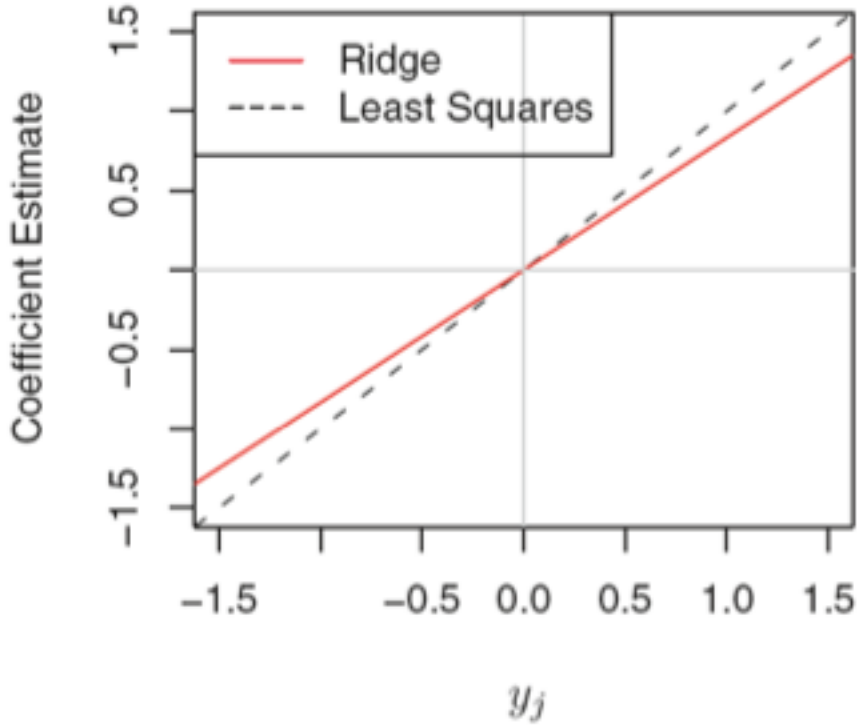
In high dimensions, with LASSO, we have edges and corners that make a diamond.

When the likelihood surface of a given value approaches the diamond, it is likely to hit the diamond at an edge or a corner where some/many coordinates are 0 (leading to some/many 0 coefficients).

RR does not produce sparsity

RR has a spherical budget region so there is no preference for points on coordinate axes to be the ones that hit the likelihood function at the largest value among all points in the region.

https://gerardnico.com/wiki/data_mining/lasso



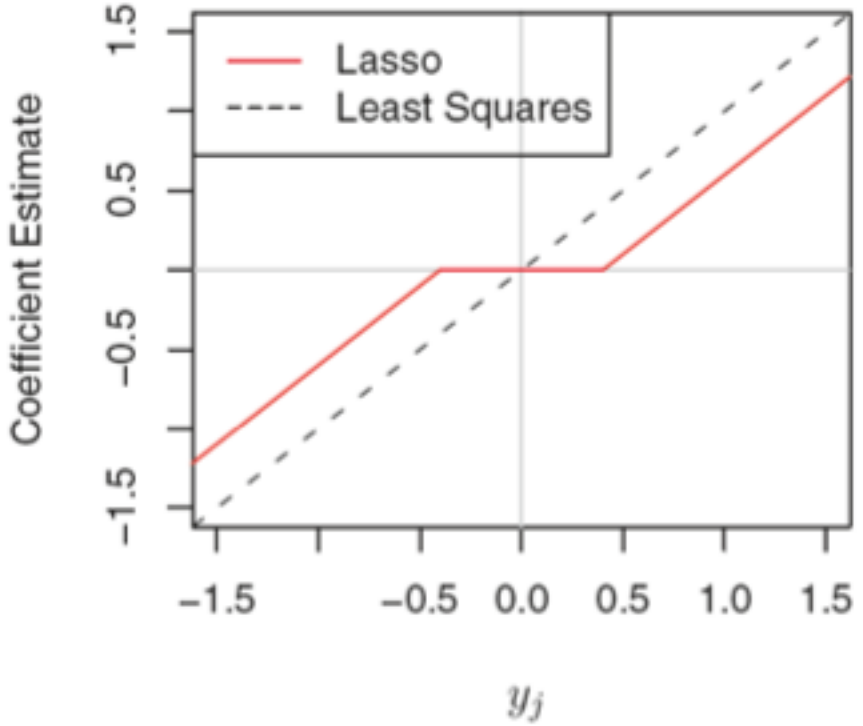


FIGURE 6.10. *The ridge regression and lasso coefficient estimates for a simple setting with $n = p$ and $\mathbf{X}$ a diagonal matrix with 1's on the diagonal.* Left: *The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates.* Right: *The lasso coefficient estimates are soft-thresholded towards zero.*

For orthogonal variables methods do simple actions

LASSO: Soft-thresholding

Ridge: Constant proportional shrinkage

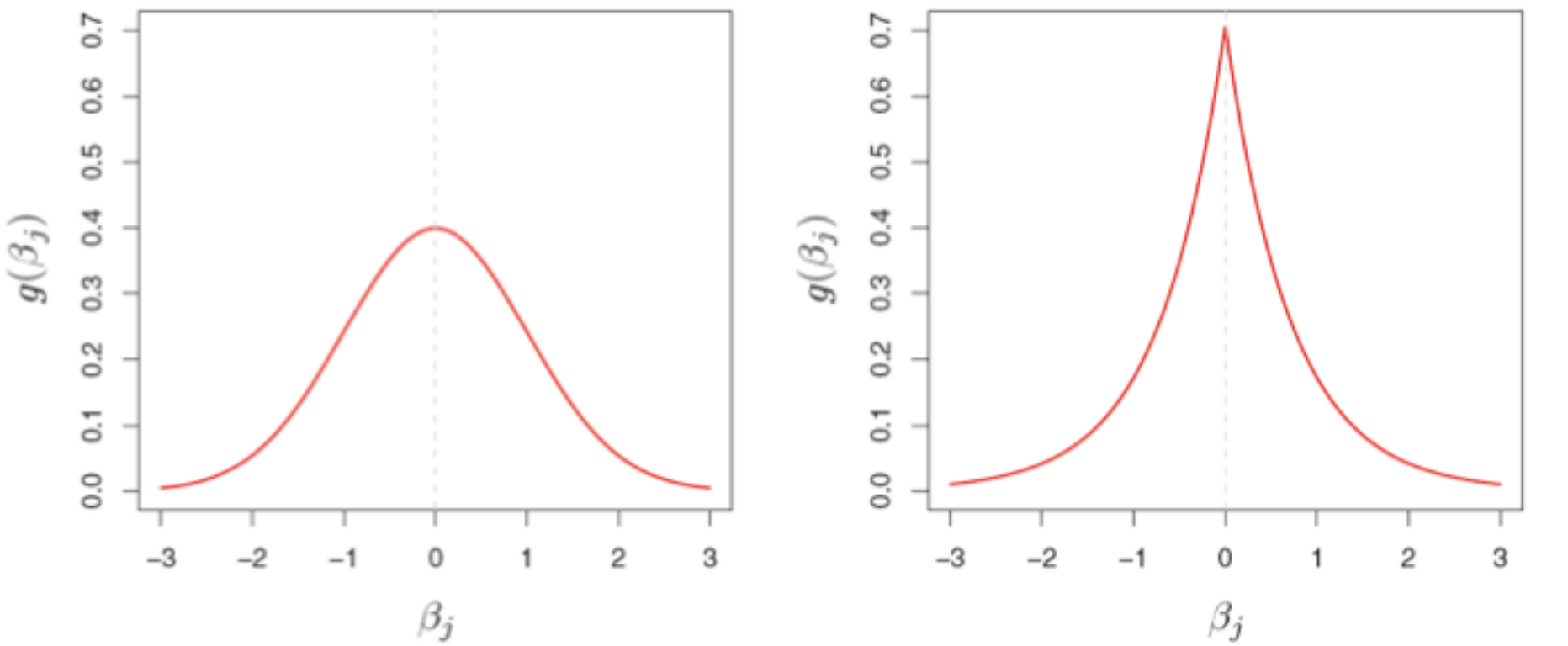


FIGURE 6.11. Left: *Ridge regression is the posterior mode for β under a Gaussian prior.* Right: *The lasso is the posterior mode for β under a double-exponential prior.*

Penalized likelihood is simply the posterior distribution in Bayesian statistics.

Ridge uses Gaussian prior for effects

LASSO uses double Exponential prior for effects

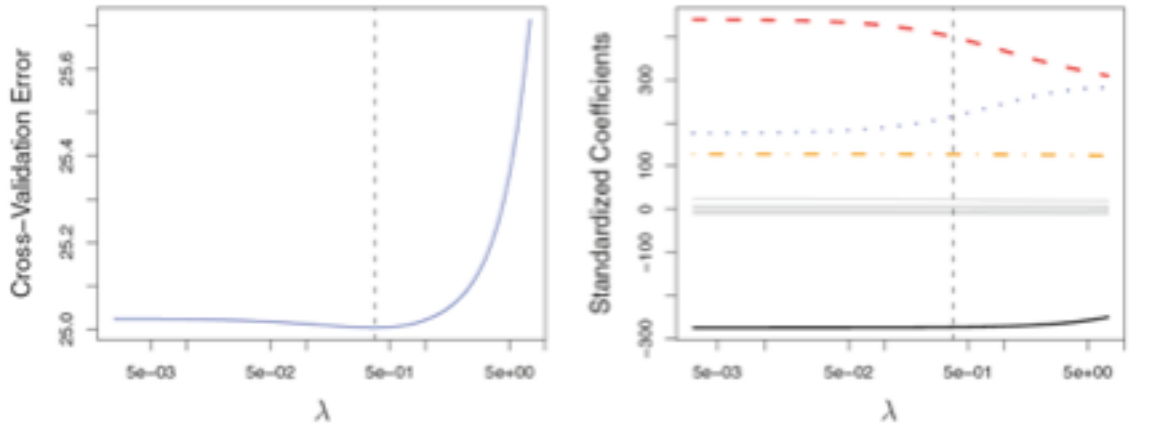


FIGURE 6.12. Left: *Cross-validation errors that result from applying ridge regression to the* `Credit` *data set with various value of* λ. Right: *The coefficient estimates as a function of* λ. *The vertical dashed lines indicate the value of* λ *selected by cross-validation.*

Cross-validation is the key to choose lambda for both methods.

Ridge regression and LASSO are flexible families of regression models that adapt their bias-variance compromise to the data through lambda value, aiming to the smallest MSE.

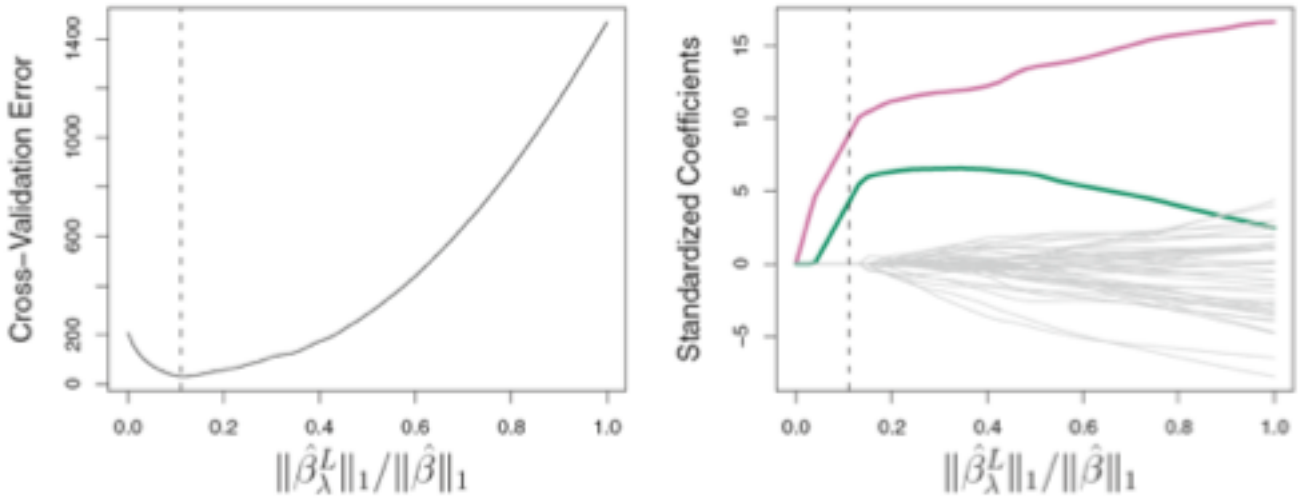


FIGURE 6.13. Left: *Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 6.9.* Right: *The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.*

OTHER PENALTIES (ESL P.72-73)

- Different exponents q outside q=1 (LASSO) and q=2 (ridge) give different penalties
- Elastic net penalty combines LASSO and ridge penalties by linear weighting by a given parameter value alpha (alpha=1 is LASSO, alpha = 0 is ridge)
- Enet includes variable selection property (making some coefficients zero) from LASSO while penalties with q>1 do not have such a property

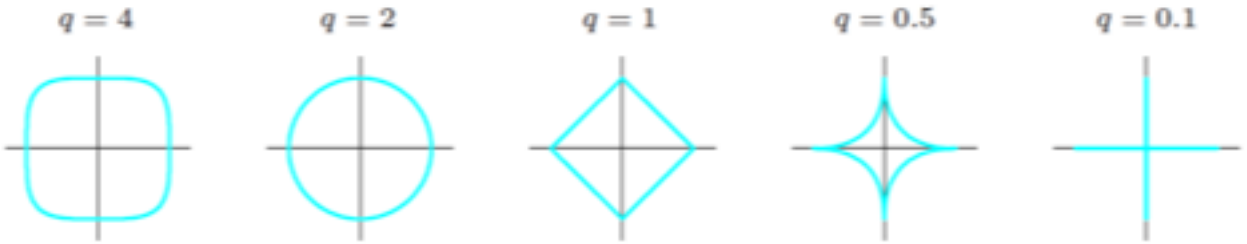
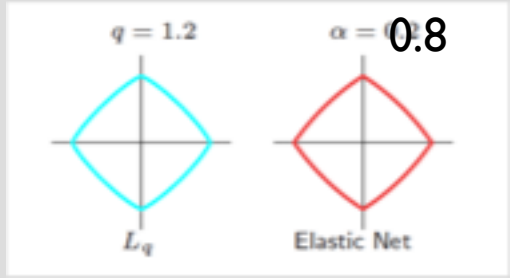


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.



$$\text{penalty}_{\text{enet}}^{\alpha}(\boldsymbol{\beta}) = \sum_{i=1}^{p}((1-\alpha)\beta_i^2 + \alpha|\beta_i|)$$

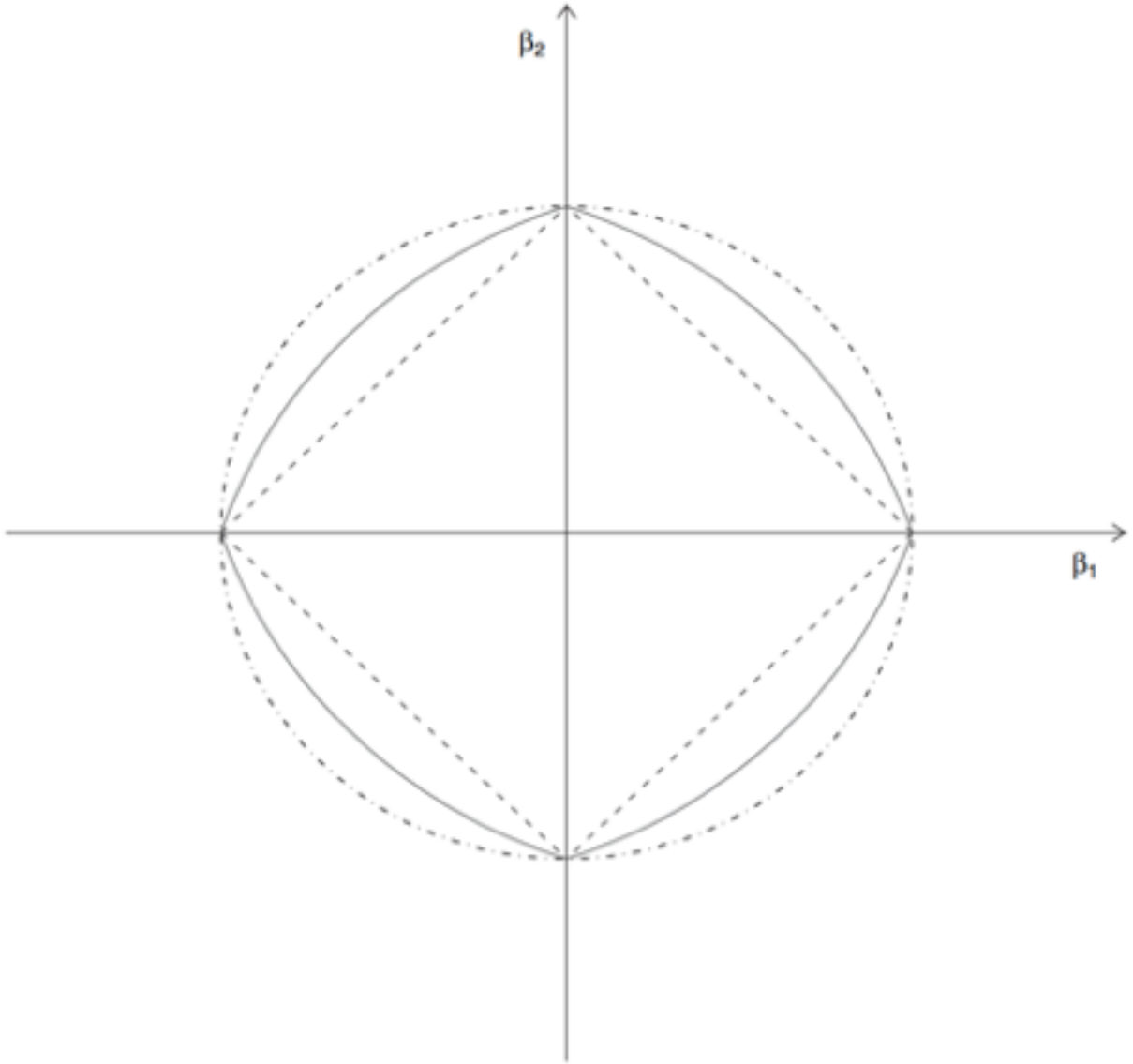
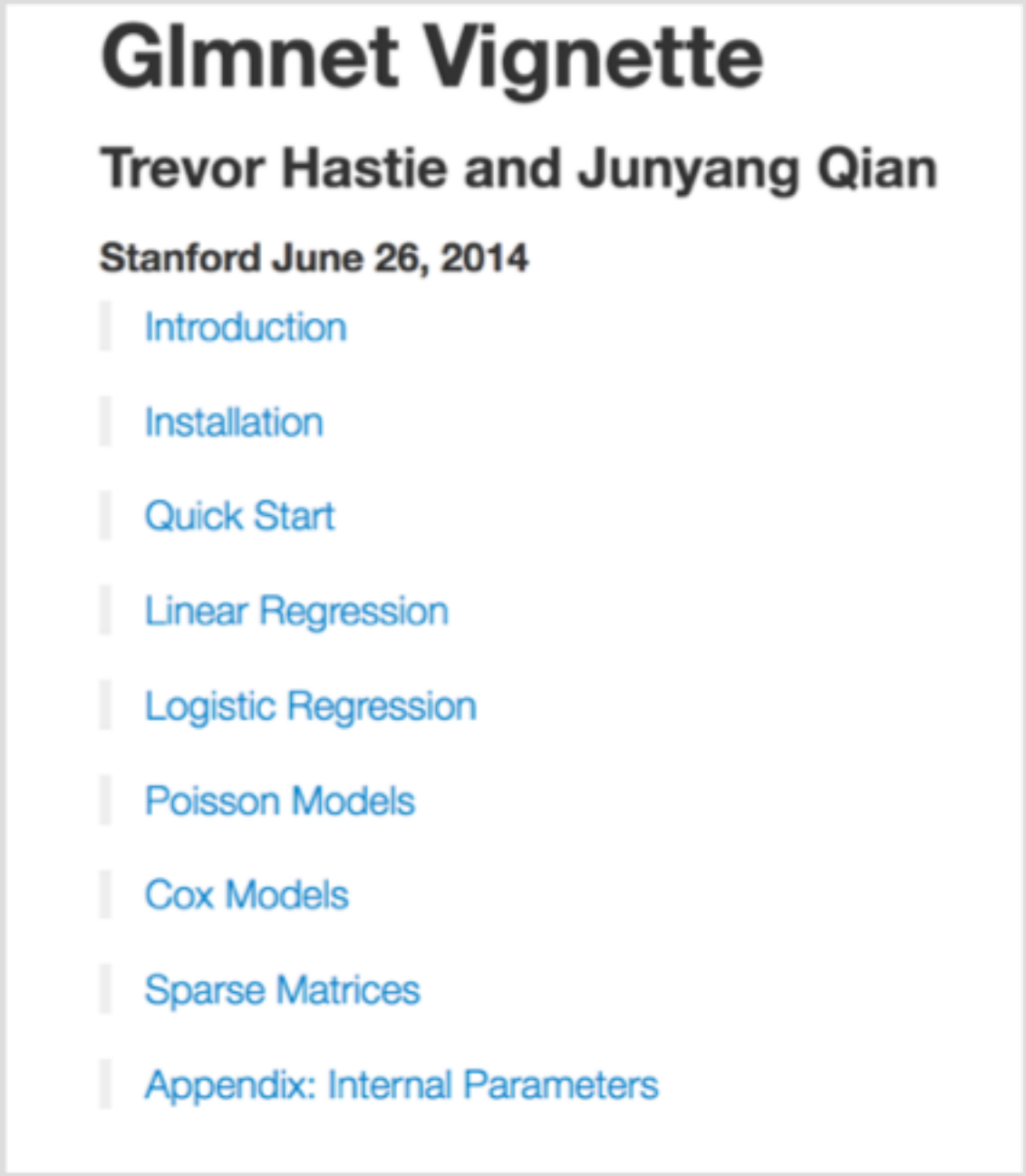


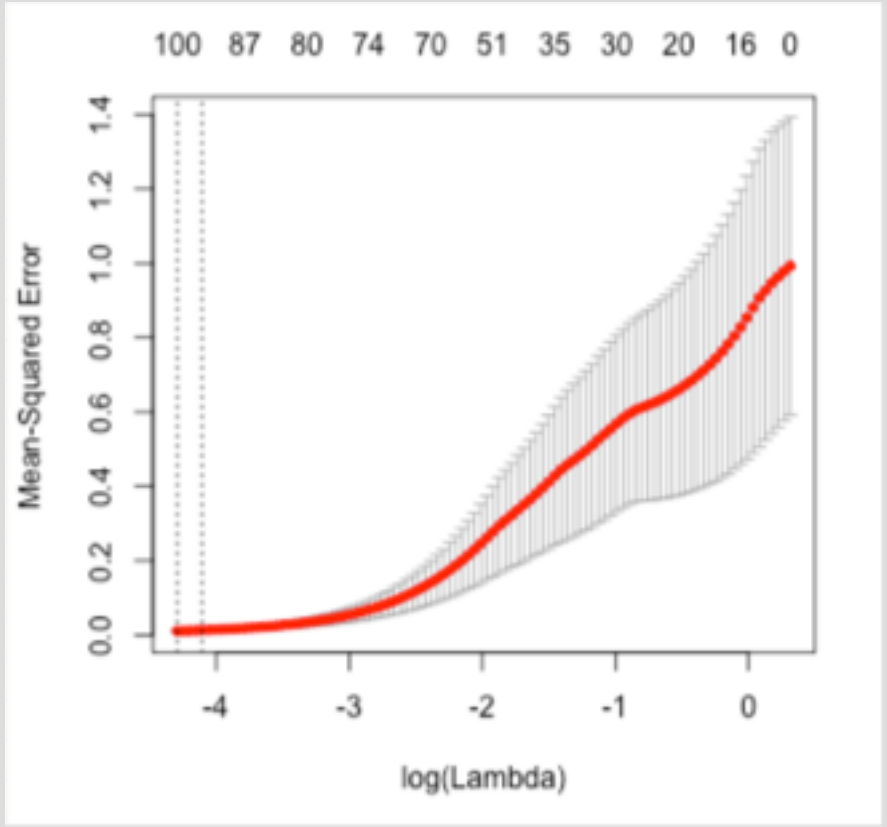
Fig. 1. Two-dimensional contour plots (level 1) (· - - - - -, shape of the ridge penalty; -------, contour of the lasso penalty; ———, contour of the elastic net penalty with $\alpha = 0.5$): we see that singularities at the vertices and the edges are strictly convex; the strength of convexity varies with α

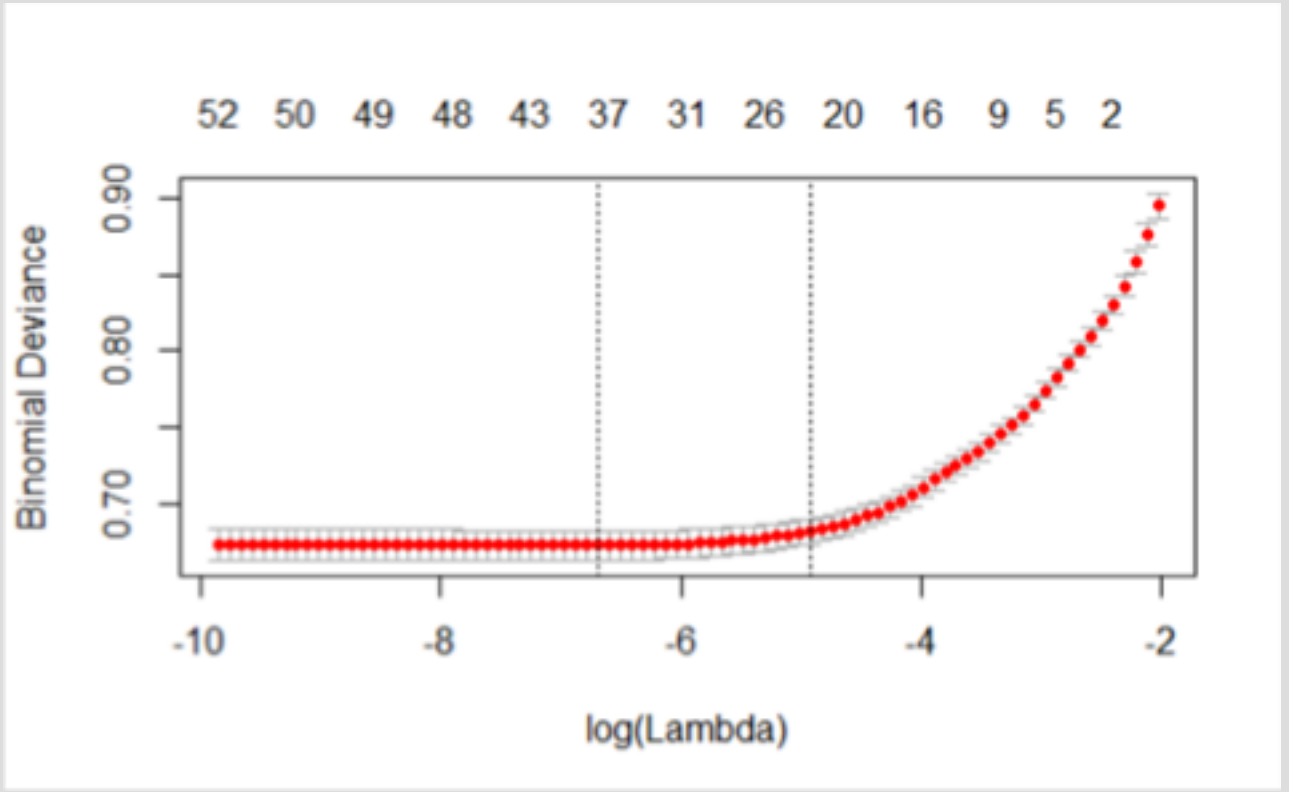


- GLMNET package
  - Does elastic **net** penalized regression for most common generalized linear models (**GLM**s)
  - Includes ridge regression (alpha = 0), LASSO (alpha =1) and linear model as special cases
  - Very fast
- Read from the beginning of the Glmnet vignette to the end of the linear regression part before you do exercises 4

https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

CV.GLMNET OUTPUT





What do these plots say?