Chapter 6: Linear Model Selection and Regularization

- As *p* (the number of predictors) comes close to or exceeds *n* (the sample size) standard linear regression is faced with problems.
- The variance of the estimates gets large and in the case of *p>n* no solution is possible.
- Reducing the number of predictors would then both improve the statistical properties of the regression estimates but would also simplify the model making its interpretation easier.

Main Topics

- Subset Selection: there are several approaches to reducing the number of predictor variables and then doing normal linear regression.
- Shrinkage: If we use all *p* predictors then some methods will shrink (also called *regularization*) the magnitude of the predictor. This may entail a small increase in bias with a large reduction in variance.
- Dimension reduction: We may create linear combinations of the *p* predictors or project them onto a subspace of smaller dimensionality. Both techniques will reduce the number of predictors prior to normal linear regression.

Subset Selection

- The best subset selection looks at all 2^p models using the following algorithm.
- (1) Let $\mathcal{M}_0$ be the null model with no parameters.
  (2) for *k*=2, ..,*p* fit all $\binom{p}{k}$ (=$\frac{p!}{(p-k)!k!}$) models. Pick the best ($\mathcal{M}_k$) based on the smallest *RSS* or largest R^2.
  (3) Select the best among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validation (*MSE*), C_p, AIC, BIC, or adjusted R^2.
- Using R^2 is OK at step (2) since all models have the same number of parameters.

Subset Selection

- For logistic-regression we can use the deviance in place of *RSS*. The deviance is -2 times the log-likelihood of the model. The smaller the better.
- The main drawback is the number of models that must be examined. For $p=20$ it is over one million.
- For least-squares models there are some shortcuts to fitting all possible models but it still becomes difficult for large p.
- Stepwise selection is computationally more efficient.

Stepwise Selection: forward selection

- Forward stepwise selection: this method starts with no predictors and add them one at a time.
  (1) Let $\mathcal{M}_0$ be the null model with no predictors
  (2) for *k*= 0, …,*p*-1, consider all *p*-*k* models by adding one parameter to $\mathcal{M}_k$. Choose the best model ($\mathcal{M}_{k+1}$) based on the smallest *RSS* or largest R^2.
  (3) Select the single best model among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validation (*MSE*), C_p, AIC, BIC, or adjusted R^2.
- As before all the models compared at step (2) have the same number of parameters so using *RSS* or R^2 is OK.

Stepwise Selection: forward selection

- The total number of models fitted is now only $1+p(p+1)/2$. So when $p=20$ we fit 211 not one million!
- We are not guaranteed to get the best model. If $p=3$, the best single variable model might be X_1, but the best model using 2 variables is X_2 plus X_3 which will be missed by forward selection.
- Although we can start the forward selection algorithm even if $p>n$ we can only go up to $\mathcal{M}_{n-1}$.

Stepwise Selection: backward selection

- Algorithm

  (1) Let $\mathcal{M}_p$ be the full model with all p predictors.

  (2) For k=p, p-1, …,1: fit all k models with one less predictor than used in $\mathcal{M}_p$. Choose the best model ($\mathcal{M}_{k-1}$) based on the smallest *RSS* or largest R^2.

  (3) Select the single best model among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validation (*MSE*), C_p, AIC, BIC, or adjusted R^2.
- The same number of models are fit as with forward selection. However, we must have p<n.

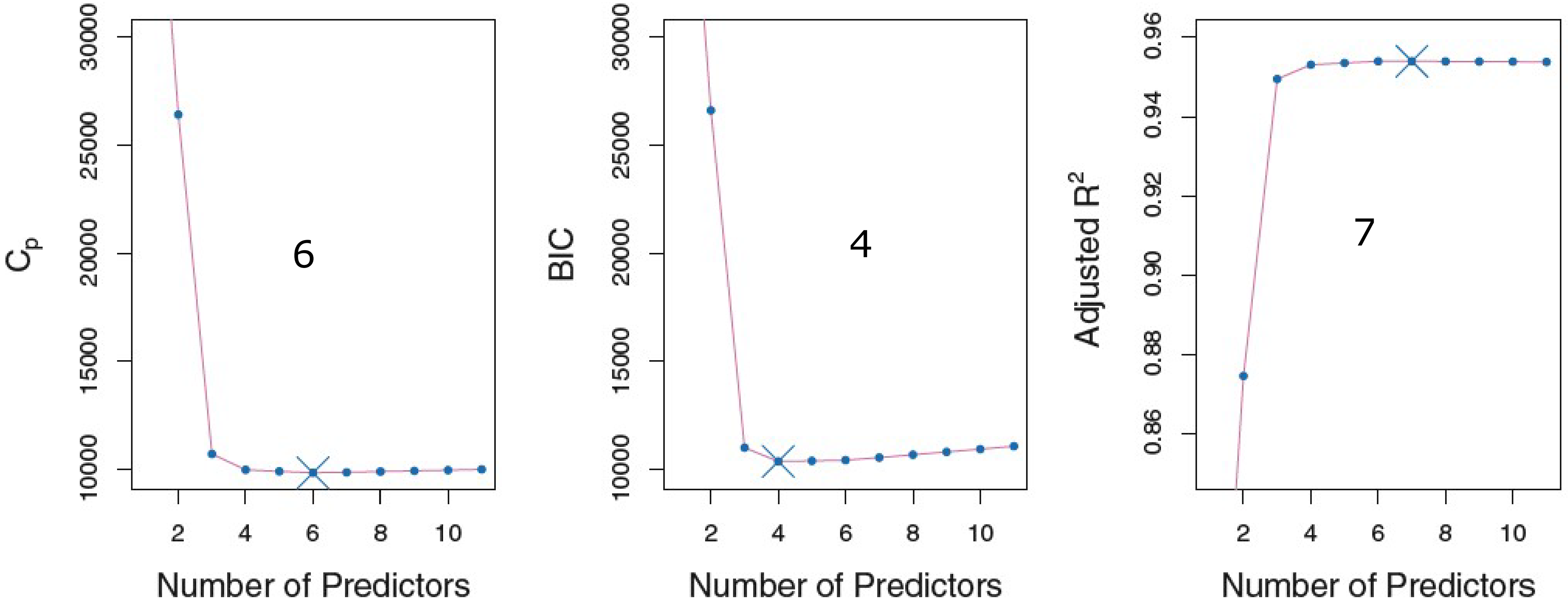
Choosing the Optimal Model

- We know the training MSE is an underestimate of the test MSE.
- Two different approaches,
  (1) Make adjustments to the training error to correct for the bias.
  (2) Directly estimate the test error with a validation set or cross-validation.
-

C_p, AIC, BIC, and Adjusted R^2

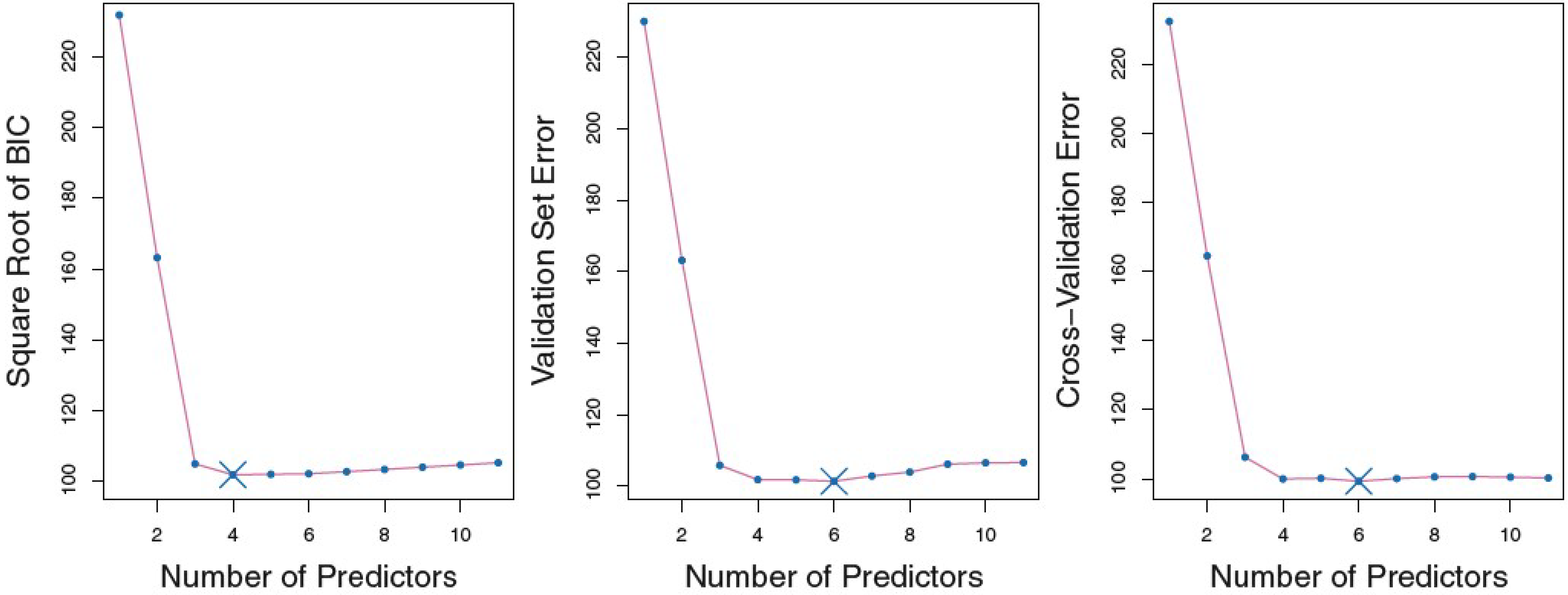
- Mallow's $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$, where *d* is the number of predictors
- AIC= $\frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$ for least squares AIC and Cp are proportional to each other.
- BIC = $\frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$, *n*>7 log(*n*)>2 so BIC will be greater that 2 and thus more conservation than C_p and AIC.
- Adjusted $R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$, the adjusted R^2 will no longer always increase with *d* like the R^2 does.
- Except for the Adjusted R^2 the other measures have a strong theoretical basis.

C_p, AIC, BIC, and Adjusted R^2



The best model is at the minimum of C_p and BIC (AIC) and the maximum of the adjusted R^2. For the credit data BIC indicates an optimum with fewer predictors than C_p.

Validation Set and Cross-Validation

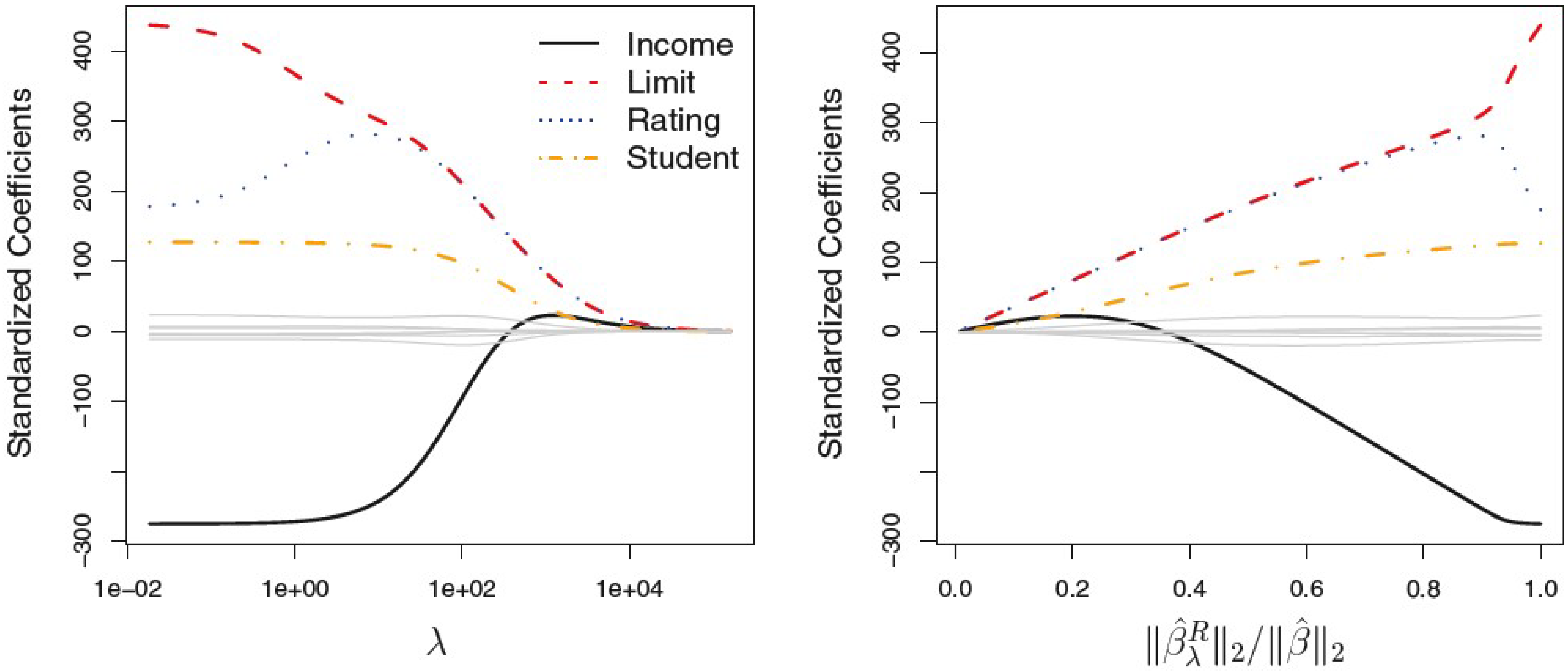


The same credit data which in this case gives the same optimum for the validation set and cross-validation. James et al. propose the 1 standard deviation rule. Calculate the standard deviation of the test MSE. After identifying the minimum see if plus 1 standard deviation includes the test MSE for fewer predictors

Shrinkage Methods: Ridge Regression

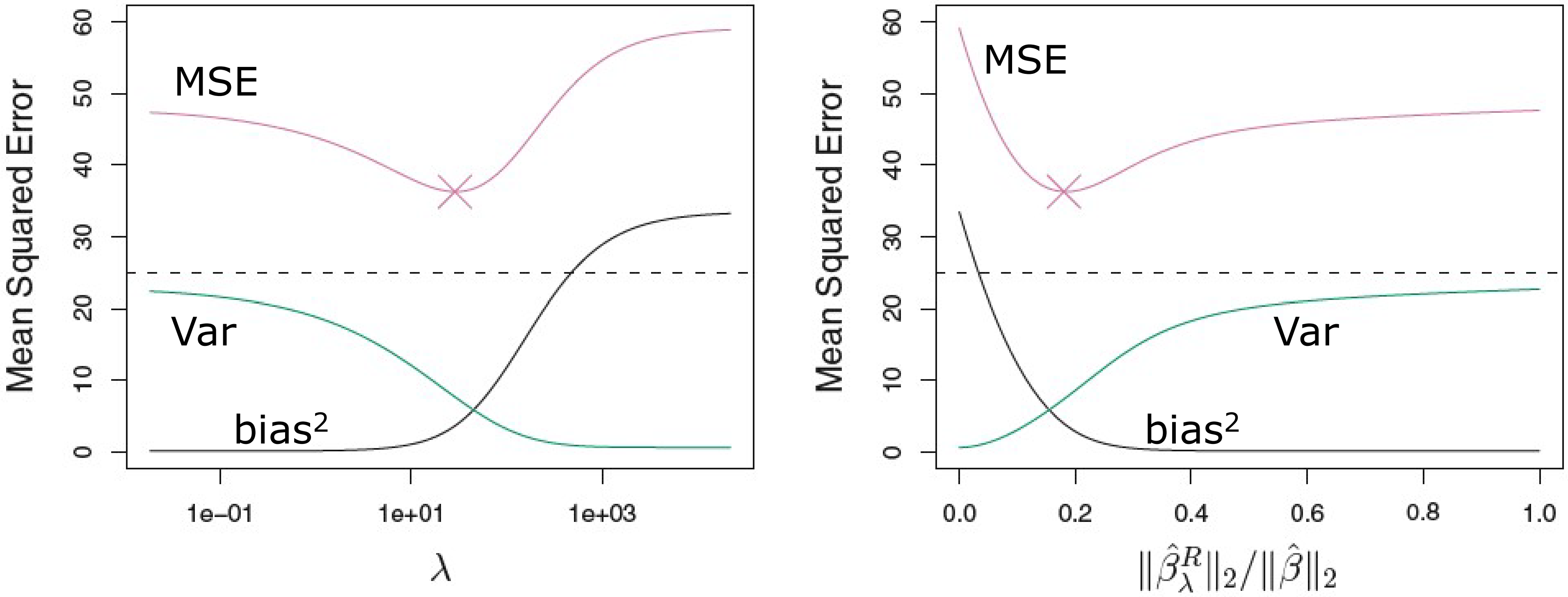
- Minimize $RSS + \lambda \sum_{j=1}^{p} \beta_j^2$
- λ is called the tuning parameter. $\lambda \sum_{j=1}^{p} \beta_j^2$ is called the shrinkage penalty.
- When λ =0, then the ridge estimators are just the normal least squares estimates.
- As $\lambda \rightarrow \infty$ the penalty grows and the ridge estimates approach 0.
- For each λ there is a different set of regression parameters, $\hat{\beta}_\lambda^R$.
- The ridge estimator, $\hat{\beta}_\lambda^R$, depends on both λ and the scale used to measure each feature. Thus, it is recommended that features be scaled by dividing each with their standard deviation.
- The penalty function does not include the intercept, β_0.
- James et al. don't talk about this directly but when *p>n* then there may be no unique solution to the ridge minimization formula.

Shrinkage Methods: Ridge Regression



$||\beta||_2$ is called the l2 norm and equals $\sqrt{\sum_{j=1}^{p}\beta_j^2}$. So, the *x*-axis can be thought of as a measure of the relative amount of shrinkage, which decreases to the right until equal to 1 which is no shrinkage. Standardized coefficients are derived from features that have been scaled (use the scale function in R).

Why does ridge regression work?



Using simulated data with n=50 and p=45, the MSE (top purple? line) for the ridge estimator, the squared bias (black) and the variance (green) are shown. The LSE show a very large variance which is decreased substantially by the ridge estimator. Ridge regression does not eliminate predictors, at best they get assigned very small coefficients.

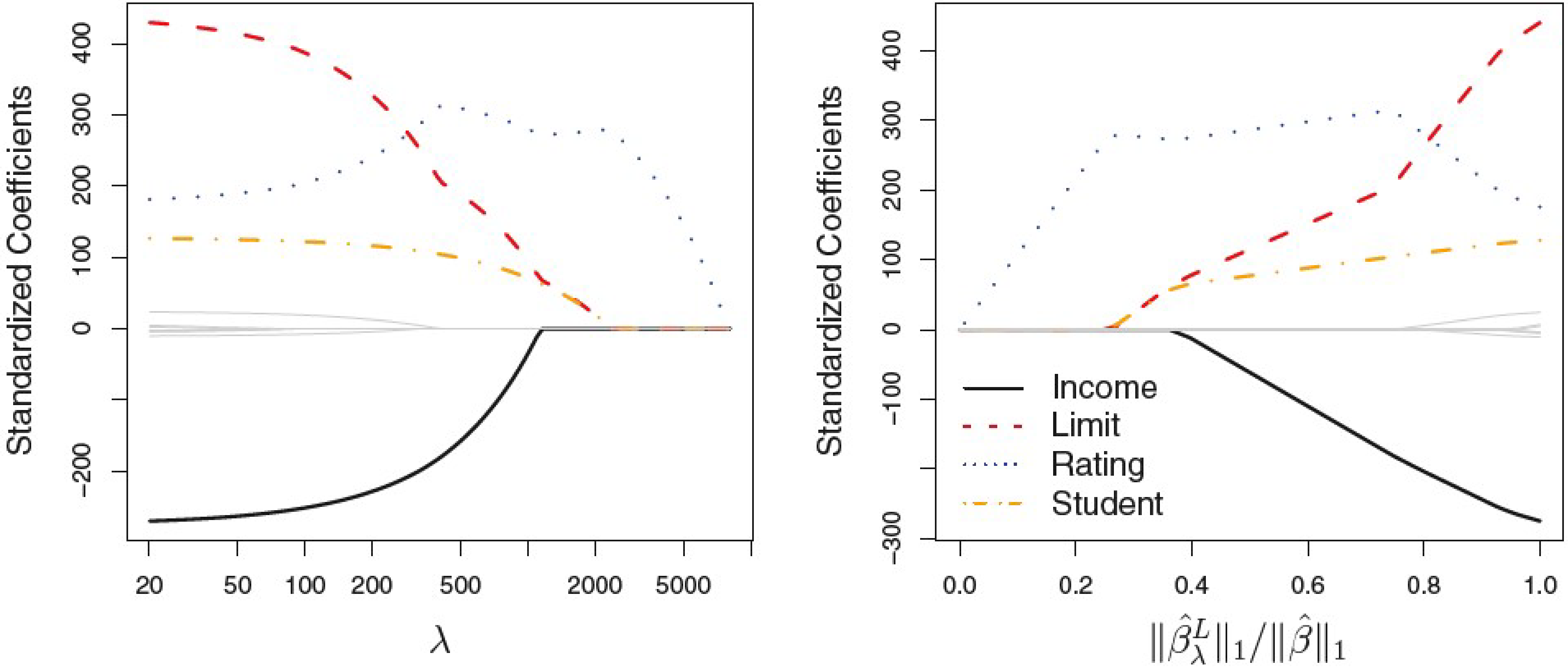
The Lasso

- The lasso can set some predictor coefficients to 0 and thus effectively aid with variable selection. The penalty function uses an l_1 norm instead of an l_2 norm squared. The $\hat{\beta}_{\lambda}^{L}$ lasso coefficients satisfy,

$$RSS + \lambda \sum_{j=1}^{p} \left| \beta_j \right|$$

- As with the ridge estimates as λ gets larger the coefficients shrink towards 0 but now some may equal be 0. Thus, we say the lasso yields sparse models.
- By convex duality you can shown when $p>n$ there can be at most ***n*** non-zero lasso coefficients! (see Rosset & Zhu, 2007. Piecewise linear regularization paths. Ann. Stat. 35:1012-1030)
- When $p>n$ there may not be a unique solution.

The Lasso



Credit data. The number of predictors in the final model is a function of λ. In the right figure as you move to the right "Rating" is the first variable to come into the model followed by "Student" and "Limit".

The Lasso

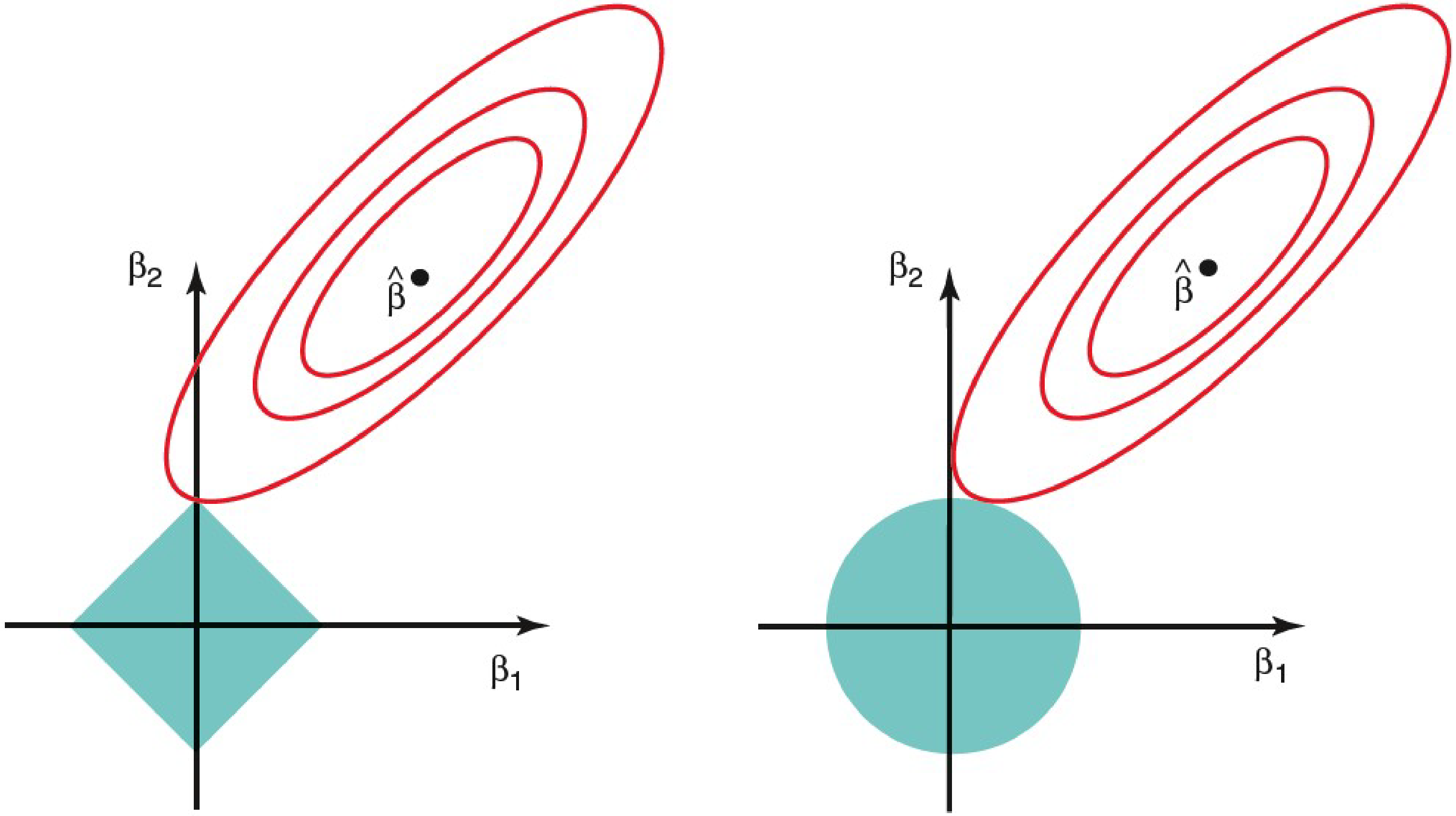
- An alternative way to write solutions for the ridge and lasso estimates are,

$$\underset{\beta}{minimize}\, RSS\ subject\ to\ \sum_{j=1}^{p} \beta_j^2 \leq s$$

$$\underset{\beta}{minimize}\, RSS\ subject\ to\ \sum_{j=1}^{p} \left|\beta_j\right| \leq s$$

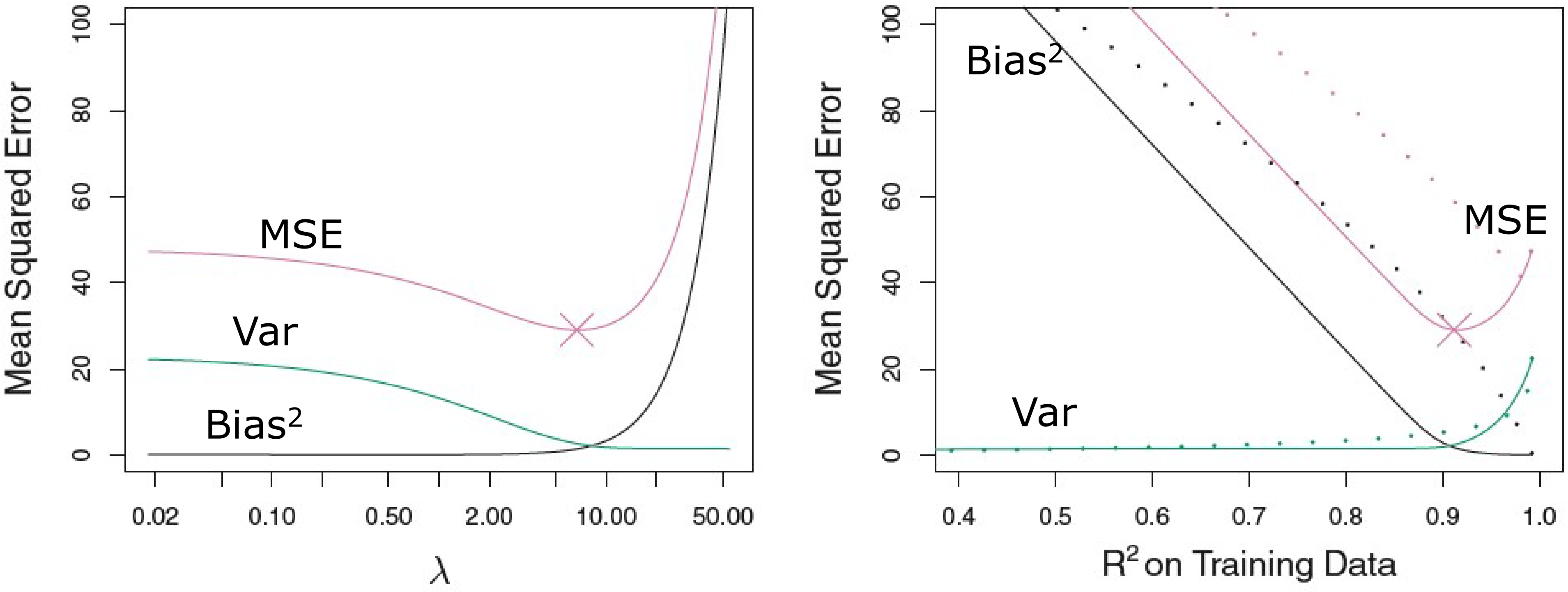
- For every value of λ there is a corresponding value of *s*.

The Lasso



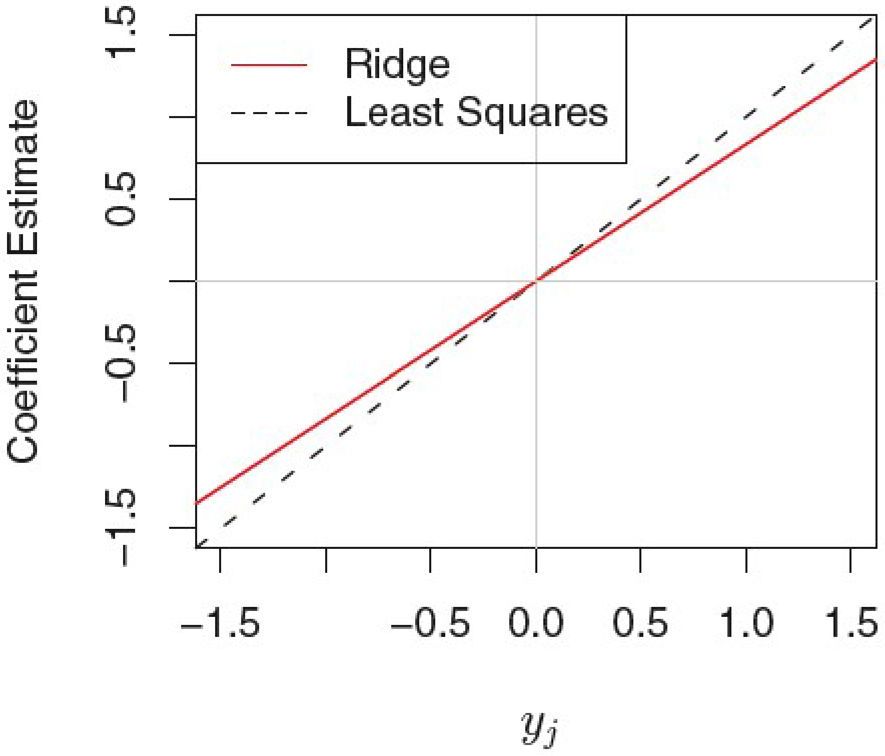
The regions demarcated by *s* for the lasso (left) and ridge estimators (right) are where the solutions must reside. $\hat{\beta}$ is the least squares estimate. The ellipses are regions of constant RSS and get larger as you move away from $\hat{\beta}$. The solutions for the lasso will often hit a vertex of the region which results in one or more parameters being set to 0.

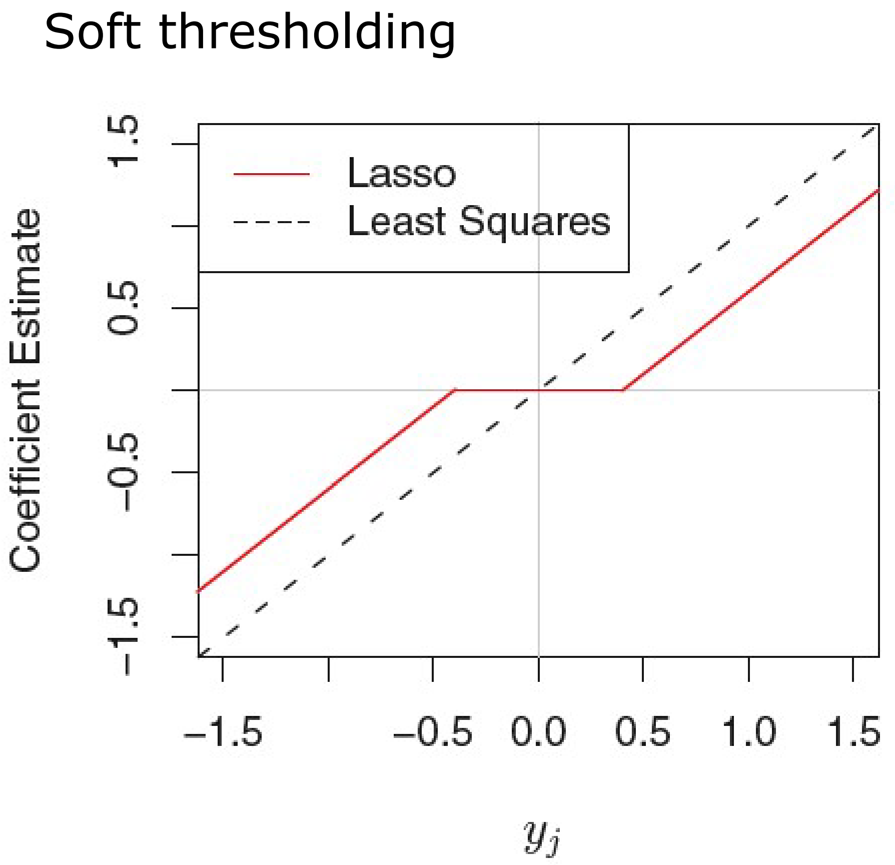
The Lasso



This simulation has p=45, n=50, but now only 2 of the predictors are related to the response. On the right are the lasso (solid) and ridge (dashed) estimator properties. The lasso outperforms the ridge estimators in this case since the ridge estimator will always maintain some estimate for every feature even if they are really zero.

Lasso and Ridge

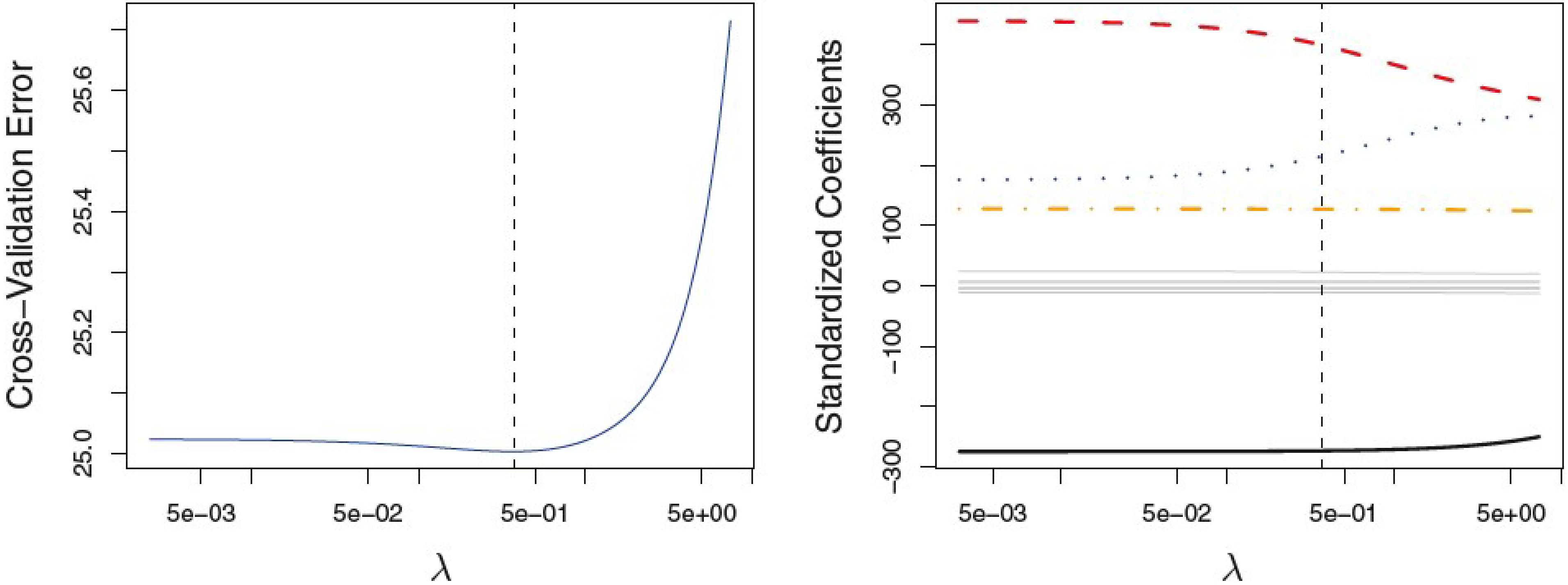




Consider a simple model, no intercept, *n*=*p*, **X** a diagonal matrix =**I**. Then the ridge solution is $\hat{\beta}_j^R = y_j/(1+\lambda)$ and the lasso is $\hat{\beta}_j^L = \begin{cases} y_j - \frac{\lambda}{2} if\ y_j > \lambda/2 \\ y_j + \frac{\lambda}{2} if\ y_j < -\lambda/2 \\ 0\ if\ \left|y_j\right| \le \lambda/2 \end{cases}$

Ridge estimators are shrunk by the same proportion while lasso estimators are shrunk towards zero by the same amount and when close to zero are shrunk exactly to 0.

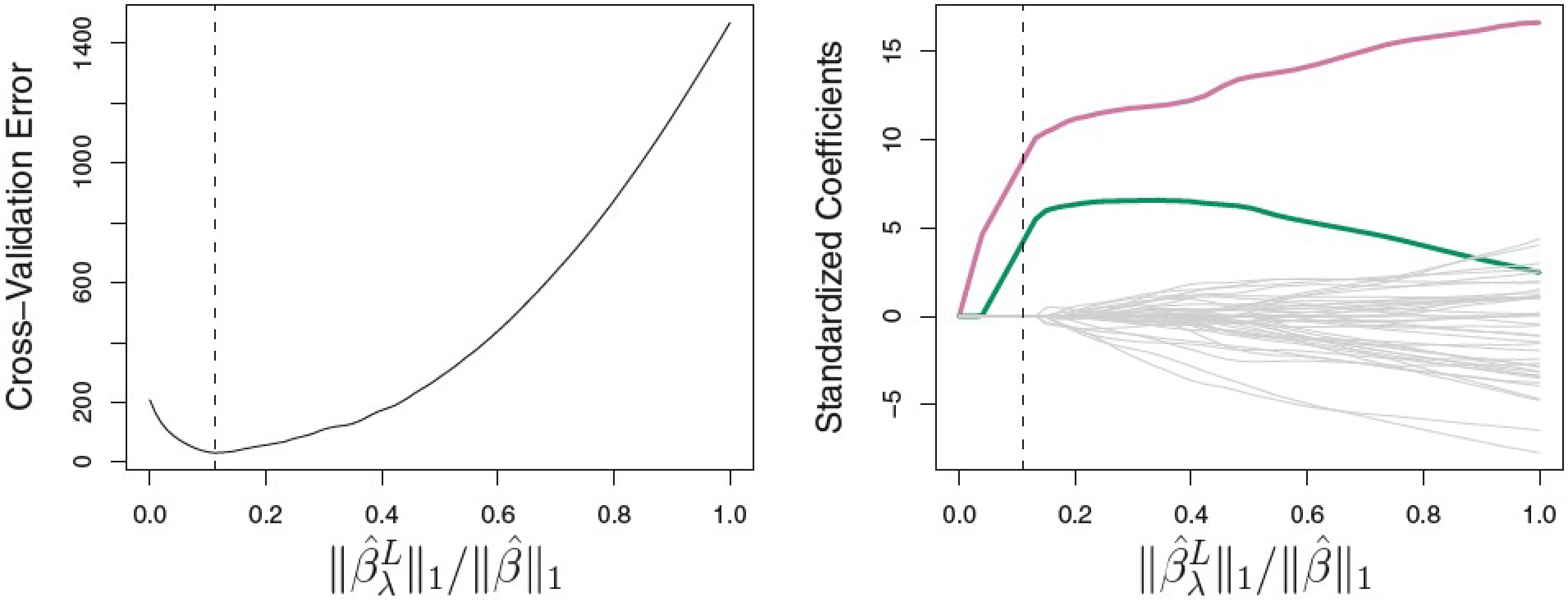
Choosing λ



Using the leave-one-out cross validation ridge regression was applied to the credit data. The optimal λ is small and results in a modest reduction in the MSE and magnitude of the coefficients. Perhaps the original least square estimates are not that bad.

ALGORITHM: choose a grid of λ values and then use cross validation to find the λ that gives the minimum MSE. Then at that λ use the entire data set to estimate the coefficients.

Choosing λ



Lasso applied to the simulated data with p=45 but only two that affect the outcome. Now the optimal λ results in two non-zero coefficients which were the two that affect the outcome.