Ridge Regression

- Recall that the OLS fitting procedure estimates the beta coefficients using the values that minimize:

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

- Ridge regression coefficients are estimated by minimizing a slightly different quantity:

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

where $\lambda \geq$ arately.

Ridge Regression (cont.)

- Note that $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage.
- The idea of penalizing by the sum-of-squares of the parameters is also used in neural networks, where it is known as *weight decay*.
- An equivalent way to write the ridge problem is:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq t,$$

Ridge Regression (cont.)

- The effect of this equation is to add a shrinkage penalty of the form

$$\lambda \sum_{j=1}^{p} \beta_j^2,$$

where the tuning parameter λ is a positive value.

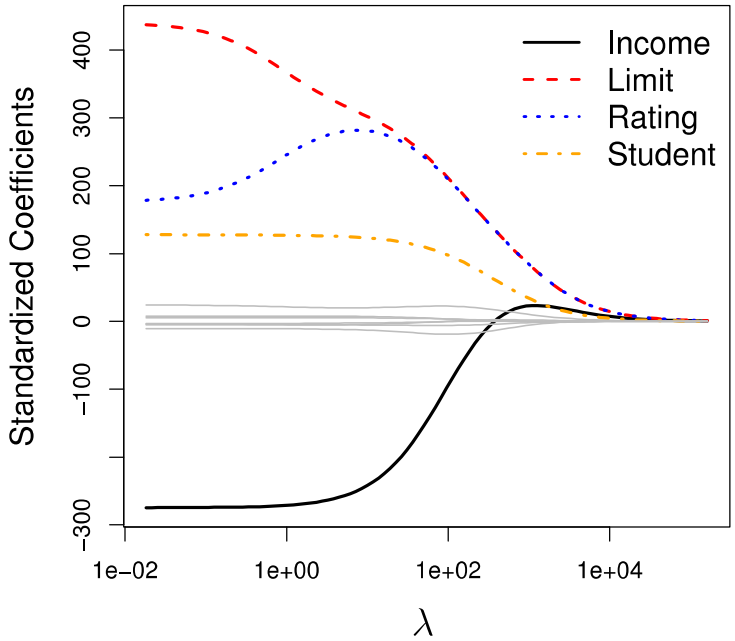
- This has the effect of shrinking the estimated beta coefficients towards zero. It turns out that such a constraint should improve the fit, because shrinking the coefficients can significantly reduce their variance.

- Note that when λ = 0, the penalty term as no effect, and ridge regression will procedure the OLS estimates. Thus, selecting a good value for λ is critical (can use cross-validation for this).

Ridge Regression (cont.)

- As λ increases, the standardized ridge regression coefficients shrinks towards zero.

- Thus, when λ is extremely large, then all of the ridge coefficient estimates are basically zero; this corresponds to the *null model* that contains no predictors.



Ridge Regression (cont.)

- The standard OLS coefficient estimates are *scale equi variant*.

- However, the ridge regression coefficient estimates c an change *substantially* when multiplying a given pre dictor by a constant, due to the sum of squared coef ficients term in the penalty part of the ridge regressi on objective function.

- Thus, it is best t $\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$ on after *standa rdizing the prec*

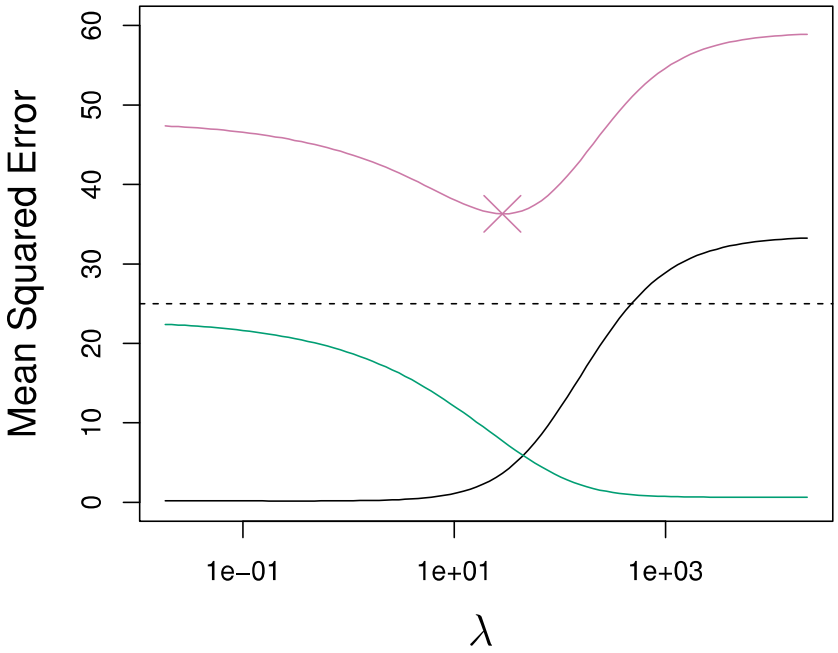
Ridge Regression (cont.)

- It turns out that the OLS estimates generally have low bias but can be highly variable. In particular when n and p are of similar size or when $n < p$, then the OLS estimates will be extremely variable

- The penalty term makes the ridge regression estimates *biased* but can also substantially reduce variance

- As a result, there is a bias/variance trade-off.

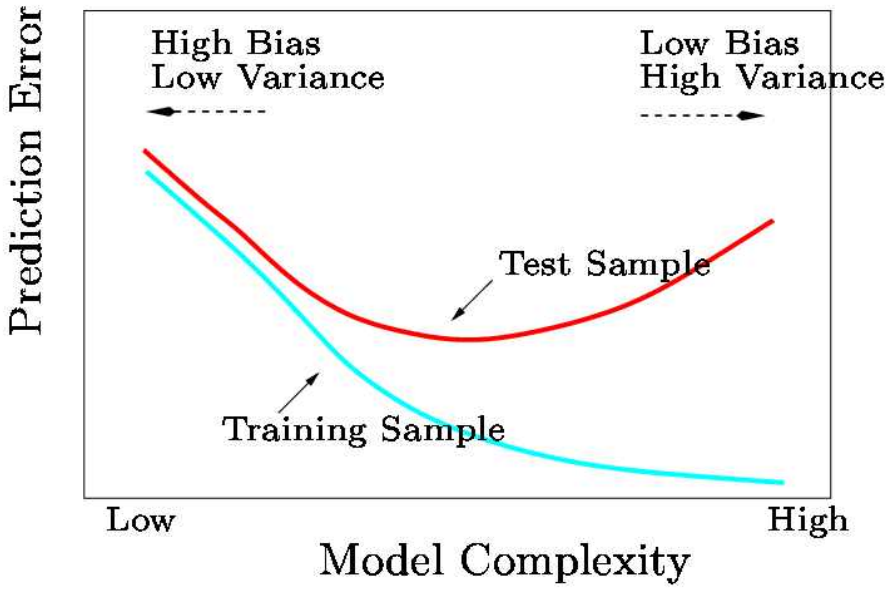
Ridge Regression (cont.)



- Black = Bias
- Green = Variance
- Purple = MSE

- Increased λ leads to inc reased bias but decreas ed variance

Ridge Regression (cont.)



- In general, the ridge regres sion estimates will be more biased than the OLS ones but have lower variance.
- Ridge regression will work best in situations where th e OLS estimates have high variance.

Ridge Regression (cont.)

Computational Advantages of Ridge Regression

- If p is large, then using the best subset selection app roach requires searching through enormous numbers of possible models.

- With ridge regression, for any given λ we only need t o fit one model and the computations turn out to be very simple.

- Ridge regression can even be used when $p > n$, a sit uation where OLS fails completely (i.e. OLS estimates do not even have a unique solution).

Ridge Regression (cont.)

- In matrix form:

$$\mathrm{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta,$$

the ridge regression solutions are easily seen to be

$$\hat{\beta}^{\mathrm{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

- The solution adds a positive constant to the diagonal of $\mathbf{X}^T\mathbf{X}$ before inv ersion (making the problem non-singular).

- The *singular value decomposition* (SVD) of the centered matrix **X** gives us some additional insight into the nature of ridge regression.

Ridge Regression (cont.)

- The SVD of the N x p matrix **X** has the form $\mathbf{X} = \mathbf{UDV}^T$
- Here, **U** and **V** are N x p and p x p orthogonal matrices, with the columns of **U** spanning the column space of **X**, and the columns of **V** spanning the row space.
- **D** is a p x p diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \cdots d_p \geq 0$ called singular values of **X**.
- If one or more values $d_j = 0$, **X** is singular.

Ridge Regression (cont.)

- Using SVD, we can write the OLS fitted vector as:

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\text{ls}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{U}^T\mathbf{y},\end{aligned}$$

- The ridge regression solutions are:

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\text{ridge}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\,\mathbf{D}(\mathbf{D}^2+\lambda\mathbf{I})^{-1}\mathbf{D}\,\mathbf{U}^T\mathbf{y} \\ &= \sum_{j=1}^{p}\mathbf{u}_j\frac{d_j^2}{d_j^2+\lambda}\mathbf{u}_j^T\mathbf{y},\end{aligned}$$

where $\mathbf{u}_j$ are the columns of **U**.

Ridge Regression (cont.)

- Like linear regression, ridge regression computes the coordinates of **y** with respect to the orthonormal basis **U**.
- It then *shrinks* these coordinates by the factor $\frac{d_j^2}{d_j^2+\lambda}$.
- This means that a greater amount of shrinkage is applied to the coordinates of basis vectors with smaller d_j^2.
- The SVD of the centered matrix **X** is another way of expressing t he *principal components* of the variables in **X**.

Ridge Regression (cont.)

- Thus, we have $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$, which is the *eigen decomposition* of $\mathbf{X}^T\mathbf{X}$.

- The eigenvectors v_j (columns of **V**) are also called the *principal components* directions of **X**.

- The first principal component direction v_1 has the property that $\mathbf{z}1 = \mathbf{X}v_1$ has the largest sample variance amongst all normalized linear combinations of the columns of **X**.

- The small singular values d_j correspond to directions in the column space of **X** having small variance, and ridge regression shrinks these directions the most.

The Lasso

- One significant problem of ridge regression is that the penalty term will never force any of the coefficients to be exactly zero.

- Thus, the final model will include all p predictors, which creates a challenge in model interpretation

- A more modern machine learning alternative is the *lasso*.

- The lasso works in a similar way to ridge regression, except it uses a different penalty term that shrinks some of the coefficients exactly to zero.

The Lasso (cont.)

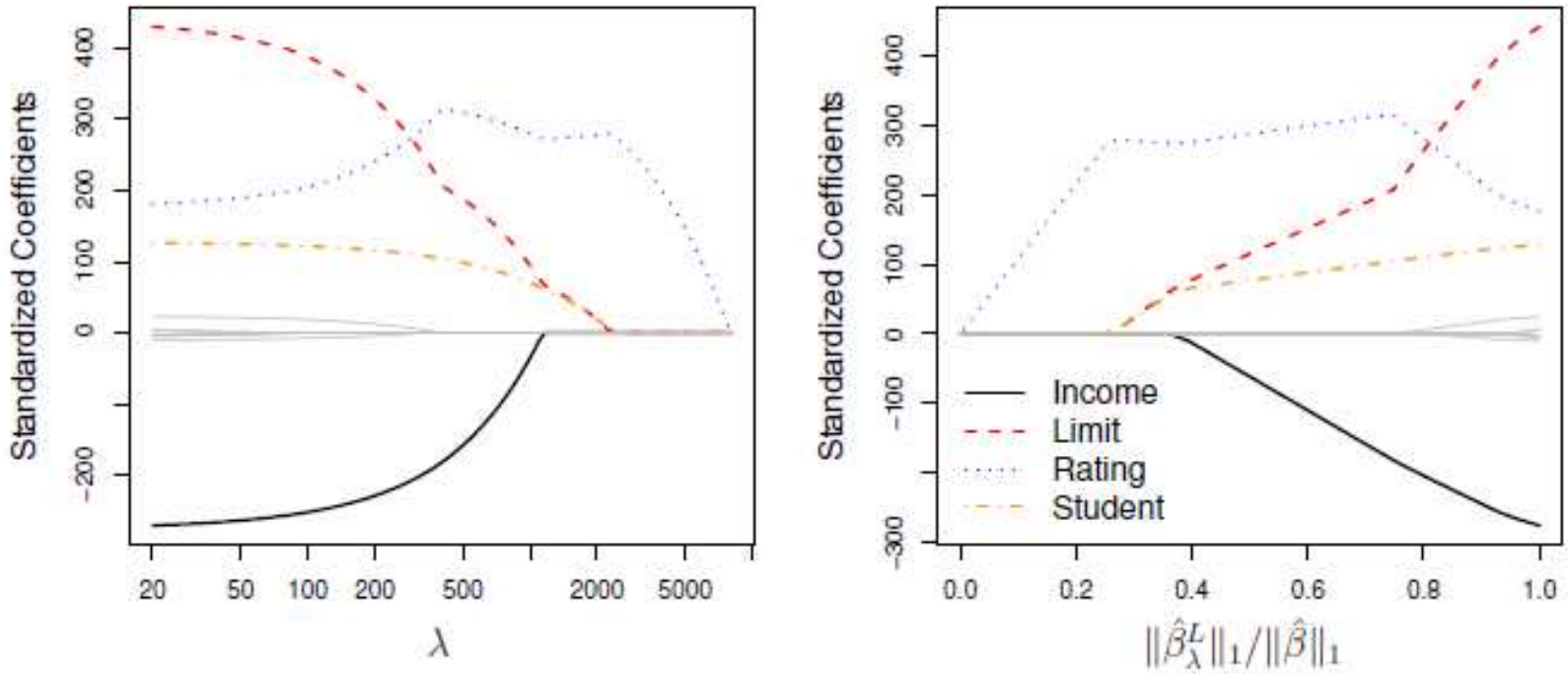
- The lasso coefficients minimize the quantity:

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

- The key difference from ridge regression is that the lasso uses an ℓ_1 penalty instead of an ℓ_2, which has the effect of forcing some of the coefficients to be exactly equal to zero when the tuning parameter λ is sufficiently large.

- Thus, the lasso performs variable/feature selection.

The Lasso (cont.)



- When λ = 0, then the lasso simply gives the OLS fit.
- When λ becomes sufficiently large, the lasso gives the null model in which all coefficient estimates equal zero.

The Lasso (cont.)

- One can show that the lasso and ridge regression co efficient estimates solves the problems:

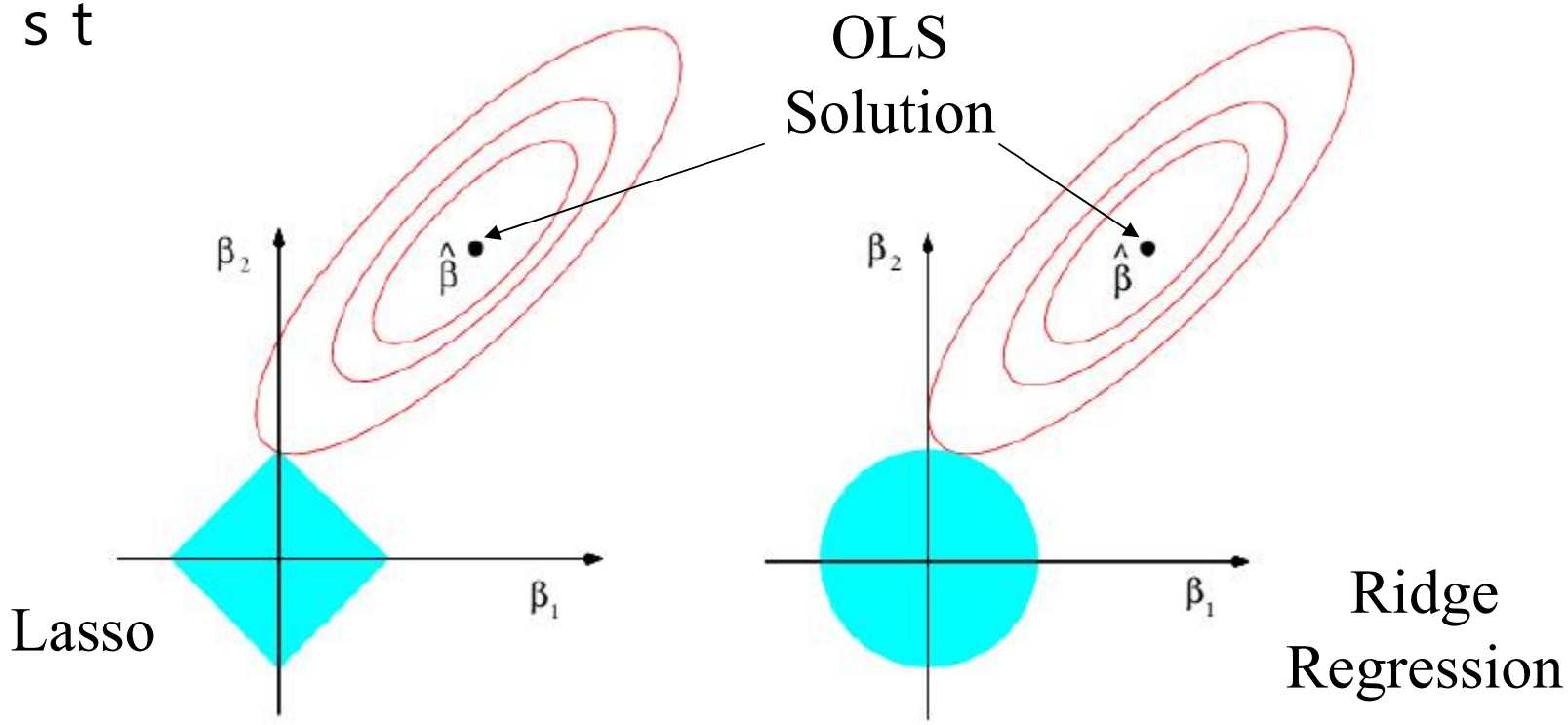
$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

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

The Lasso (cont.)

- The lasso and ridge regression coefficient estimates a re given by the first point at which an ellipse contact s t



Lasso vs. Ridge Regression

- The lasso has a major advantage over ridge regressio n, in that it produces simpler and more interpretable models that involved only a subset of predictors.
- The lasso leads to qualitatively similar behavior to rid ge regression, in that as λ increases, the variance dec reases and the bias increases.
- The lasso can generate more accurate predictions co mpared to ridge regression.
- Cross-validation can be used in order to determine w hich approach is better on a particular data set.

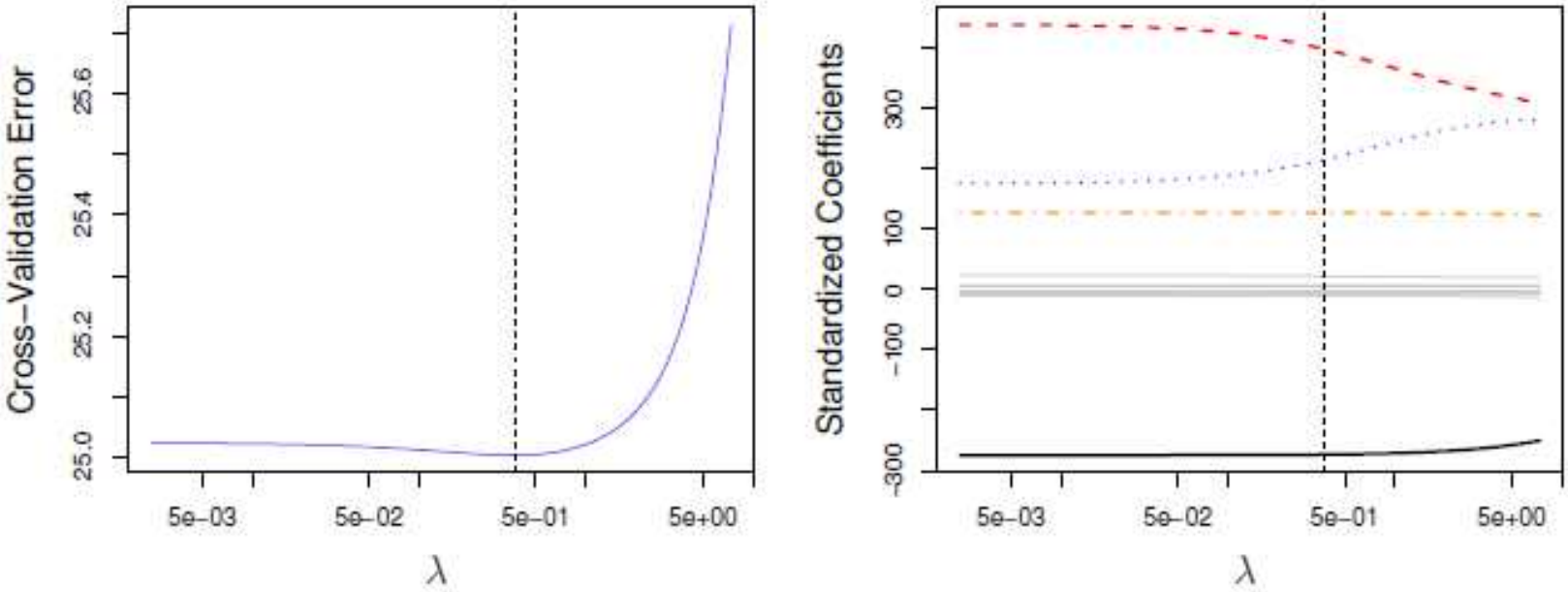
Selecting the Tuning Parameter λ

- As for subset selection, for ridge regression and lasso we require a method to determine which of the models under consideration in best; thus, we required a method selecting a value for the tuning parameter λ or equivalently, the value of the constraint *s*.

- Select a grid of potential values; use cross-validation to estimate the error rate on test data (for each value of λ) and select the value that gives the smallest error rate.

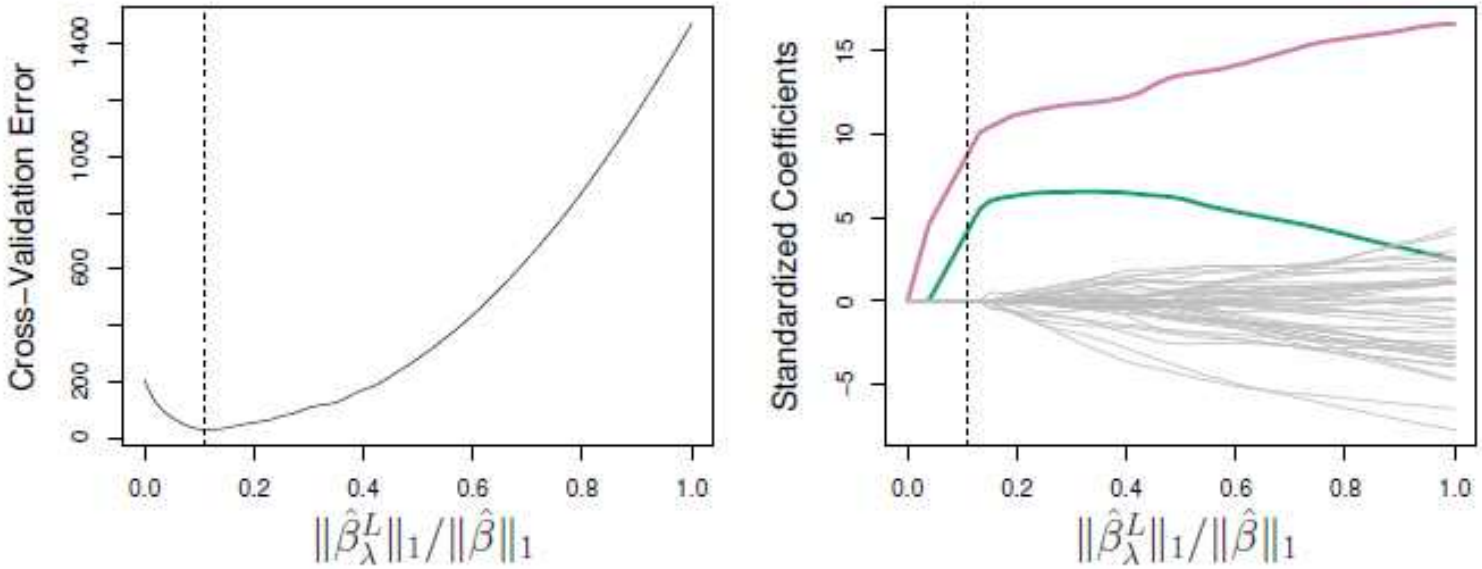
- Finally, the model is re-fit using all of the variable observations and the selected value of the tuning parameter λ.

Selecting the Tuning Parameter λ: Credit Data Example



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

Selecting the Tuning Parameter λ: Simulated Data Example



Left: *Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Slide 39.* Right: *The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.*

Dimension Reduction

- The methods we have discussed so far have involved fitting linear regression models, via OLS or a shrunken approach, using the original predictors.

- We now explore a class of approaches that *transform* the predictors and then fit an OLS model using the transformed variables.

- We refer to these techniques as *dimension reduction* methods.

Partial Least Squares

- PCR identifies linear combinations, or *directions*, that best represents the predictors.
- These directions are identified is an *unsupervised* way, since the response *Y* is not used to help determine the principal component directions.
- That is, the response does not *supervise* the identification of the principal components.
- PCR suffers from a potentially serious drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

Partial Least Squares (cont.)

- Like PCR, *partial least squares* (PLS) is a dimension reducti on method, which first identifies a new set of features Z_1, …,Z_M that are linear combinations of the original features.
- Then PLS fits an OLS linear model using these *M* new feat ures.
- Unlike PCR, PLS identifies these new features in a *supervis ed* way; PLS makes use of the response *Y* in order to iden tify new features that not only approximate the old featur es well, but also that are *related to the response*.
- The PLS approach attempts to find directions that help ex plain both the response and the predictors.

Partial Least Squares (cont.)

- After standardizing the p predictors, PLS computes t he first partial least squares direction Z_1 by setting e ach ϕ_{1j} in

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j$$

equal to the coefficient from the simple linear regressi on of Y onto X_j.

- One can show that this coefficient is proportional to the correlation between Y and X_j.

Partial Least Squares (cont.)

- Hence, in computing $Z_1 = \sum_{j=1}^{p} \phi_{1j} X_j$, PLS places the highest weight on the variables that are most strongly related to the response.
- Subsequent directions are found by taking residuals and then repeating the above prescription.
- As with PCR, the number M of PLS directions used in PLS is a tuning parameters that is typically chosen by cross-validation.
- While the supervised dimension reduction of PLS can reduce bias, it also has the potential to increase variance.

Partial Least Squares (cont.)

Algorithm 3.3 *Partial Least Squares.*

1. Standardize each $\mathbf{x}_j$ to have mean zero and variance one. Set $\hat{\mathbf{y}}^{(0)} = \bar{y}\mathbf{1}$, and $\mathbf{x}_j^{(0)} = \mathbf{x}_j,\ j = 1, \ldots, p$.

2. For $m = 1, 2, \ldots, p$

   (a) $\mathbf{z}_m = \sum_{j=1}^{p} \hat{\varphi}_{mj}\mathbf{x}_j^{(m-1)}$, where $\hat{\varphi}_{mj} = \langle \mathbf{x}_j^{(m-1)}, \mathbf{y} \rangle$.

   (b) $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$.

   (c) $\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m$.

   (d) Orthogonalize each $\mathbf{x}_j^{(m-1)}$ with respect to $\mathbf{z}_m$: $\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - [\langle \mathbf{z}_m, \mathbf{x}_j^{(m-1)} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle]\mathbf{z}_m,\ j = 1, 2, \ldots, p$.

3. Output the sequence of fitted vectors $\{\hat{\mathbf{y}}^{(m)}\}_1^p$. Since the $\{\mathbf{z}_\ell\}_1^m$ are linear in the original $\mathbf{x}_j$, so is $\hat{\mathbf{y}}^{(m)} = \mathbf{X}\hat{\beta}^{\text{pls}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.