

Shrinkage Methods I

Statistical Data Mining I

Rachael Hageman Blair

Shrinkage Methods

- Subset Selection methods are a discrete process. They suffer a lot of flaws and can still exhibit high-variance and prediction error.
- Shrinkage methods are more continuous, and do not suffer as much from high variability.
- We will tackle a number of shrinkage methods:
 - (1) Ridge Regression
 - (2) The LASSO
 - (3) Least Angle Regression (LAR)
 - (4) Principal Components Regression
 - (5) Partial Least Squares

Shrinkage Methods: SVD Preliminaries

Singular Value Decomposition (SVD):

- Any matrix $X \in \mathbf{R}^{m \times n}$ with $m \geq n$ has a singular value decomposition (svd).
- In general: $X = UDV^T$
where,
 - $U = (u_1, u_2, \dots, u_p) \in \mathbf{R}^{n \times p}$, is an orthogonal matrix whose columns (singular vectors) form an orthonormal basis for the space spanned by the columns vectors of X .
 - $V = (v_1, v_2, \dots, v_p) \in \mathbf{R}^{p \times p}$, is an orthogonal matrix whose columns (singular vectors) form an orthonormal basis for the row vectors of X .
 - $D = (d_{11}, d_{22}, \dots, d_{pp}) \in \mathbf{R}^{p \times p}$ is a diagonal matrix containing the rank ordered singular values of X .

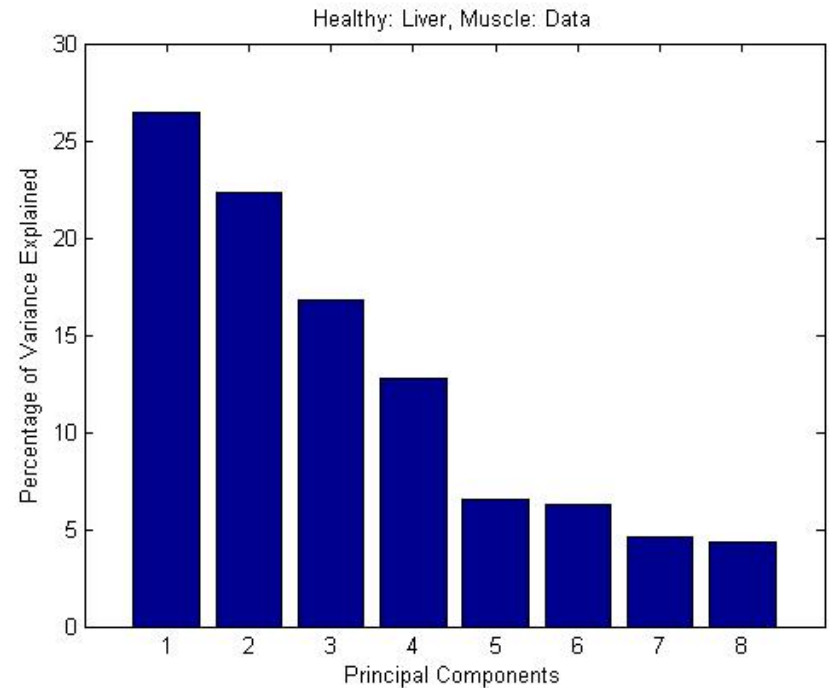
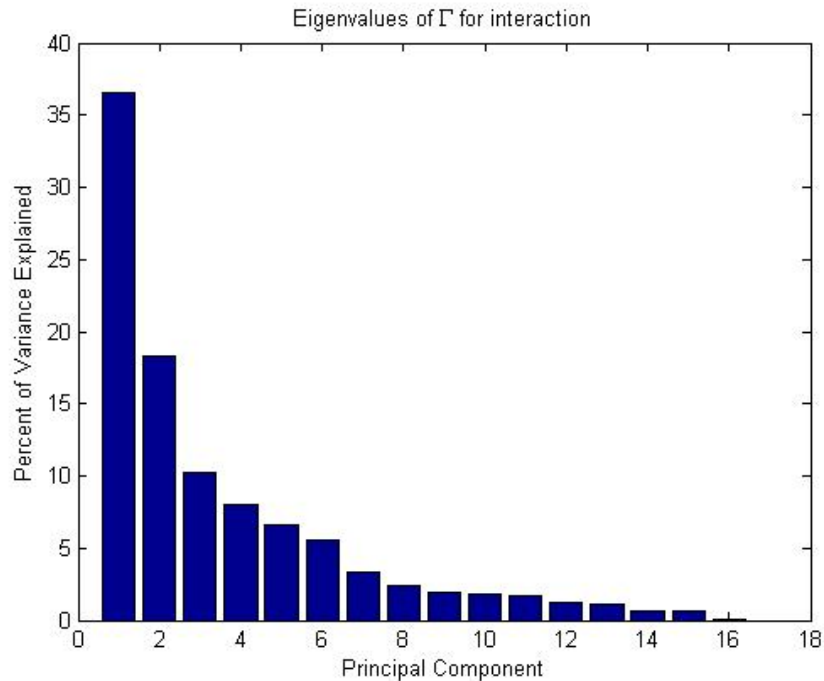
Shrinkage Methods: SVD Preliminaries

- Let $X \in \mathbf{R}^{m \times n}$ be a data matrix, where each column is an observation of a real-valued random vector with mean 0 (mean-centered).
- The right singular vectors v_j are called **principal components directions** of X .
- The vector $z_1 = Xv_1 = d_1 v_1$ is the first **principal component**:
which carries the largest sample variance: $\text{var}(z_1) = \text{var}(Xv_1) = d_1^2 / N$.
- The normalized variable: $u_1 = \left(1/d_1\right) Xv_1$
is called the **normalized first principal component**.

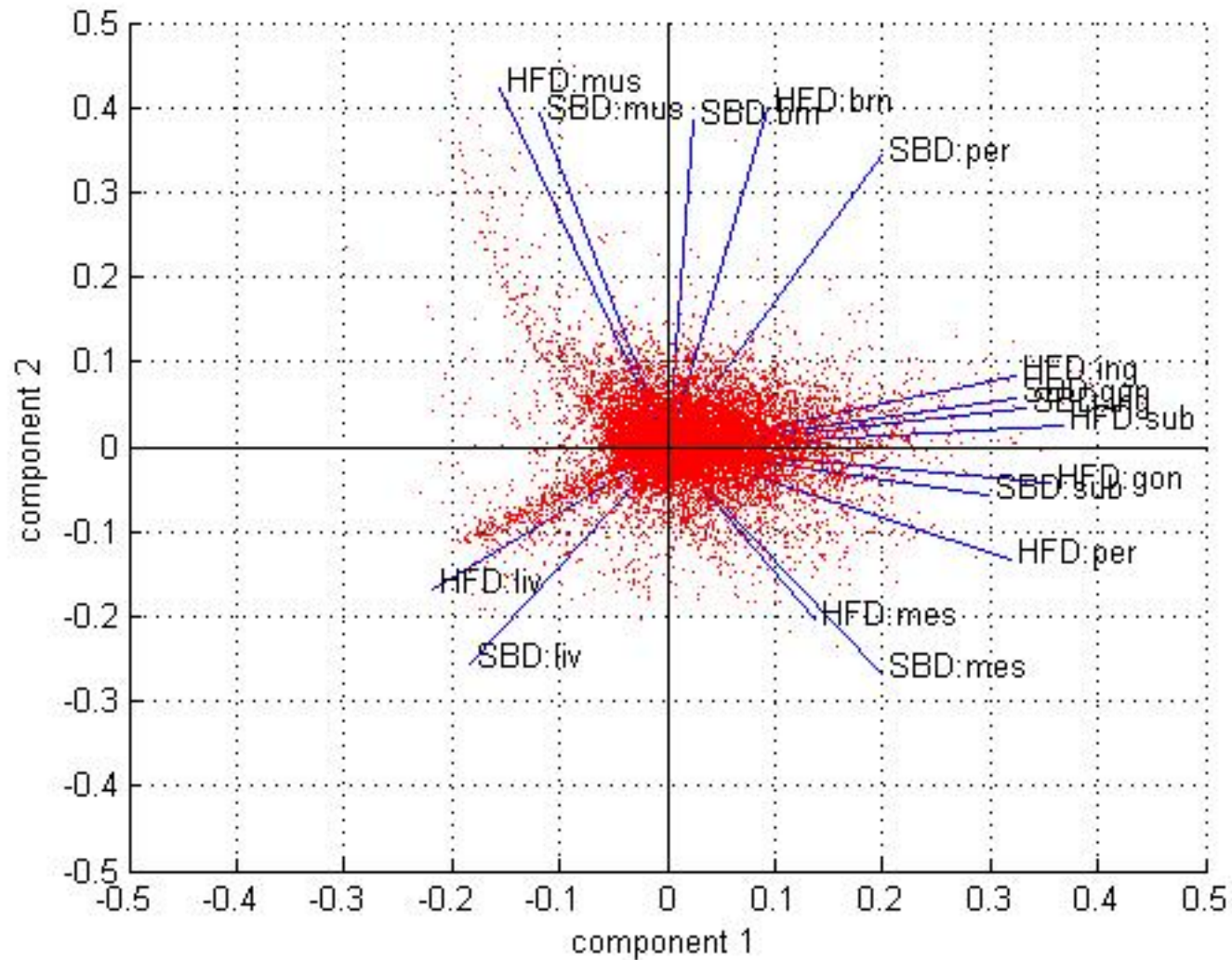
Shrinkage Methods: SVD Preliminaries

- Objective: Having determined the vector of largest sample variance, we usually want to go on and find the vector of second largest variance that is orthogonal to the first.
- This is done by computing the the vector of largest sample variance of the deflated data matrix $X - d_1 u_1 v_1^T$.
- Continuing this process will yield the ordered principle components.
- Principal component z_j is orthogonal to the z_1, \dots, z_{j-1} .

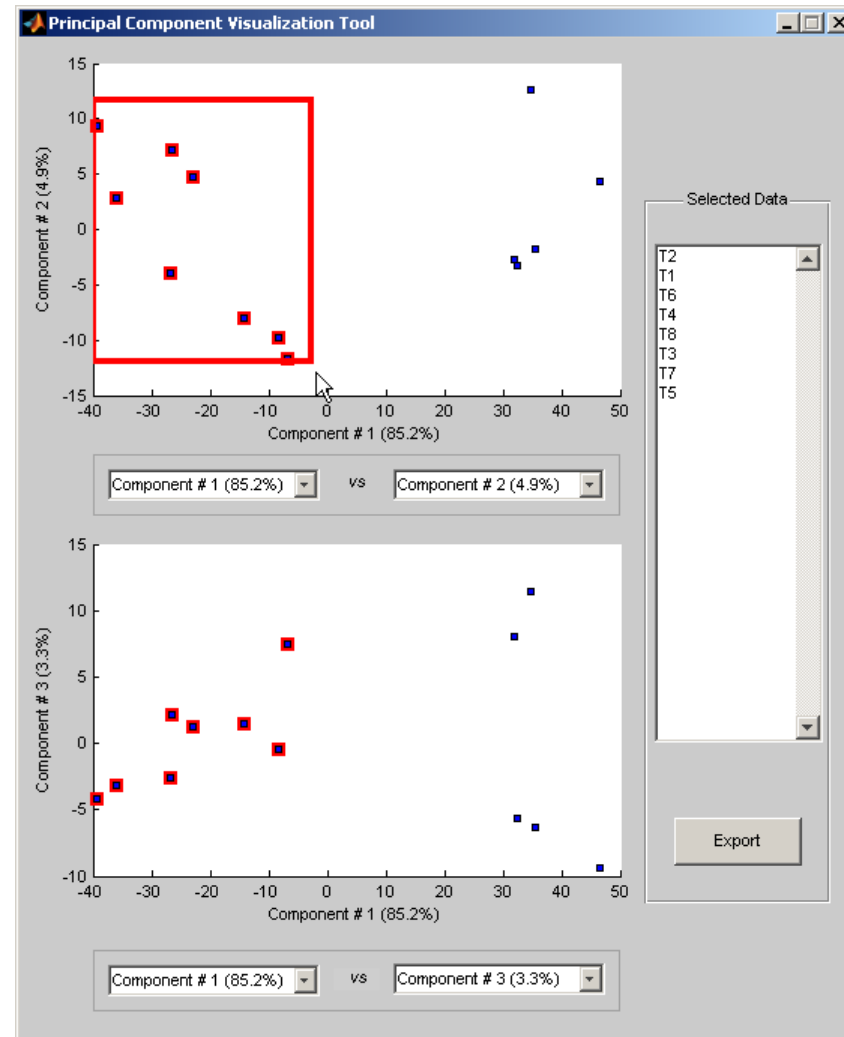
Shrinkage Methods: SVD Preliminaries



Shrinkage Methods: SVD Preliminaries




Shrinkage Methods: SVD Preliminaries



Shrinkage Methods: Ridge Regression

- Ridge regression shrinks the regression coefficients by imposing a penalty on their size.

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

penalty 

- The coefficients are shrunk towards zero and each other.
- The larger λ , the stiffer the penalty and the larger the shrinkage.
- Note: When the variables are correlated, their coefficients can become poorly determined and exhibit high variance.

Shrinkage Methods: Ridge Regression

- An alternative representation:

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

such that $\sum_{j=1}^p \beta_j^2 \leq t.$



L2 ridge penalty

which makes explicit a size constraint on the parameters.

Shrinkage Methods: Ridge Regression

- In matrix form, we can formulate the ridge regression problem:


$$RSS(\lambda) = (y - X\beta)^T (y - X\beta) + \lambda\beta^T \beta.$$

- The solution is given as:

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

Shrinkage Methods: Ridge Regression

Geometric Interpretation:

- Consider the fitted response: $\hat{y} = X\hat{\beta}^{ridge}$
$$= X(X^T X + \lambda I)^{-1} X^T y$$
$$= UD(D^2 + \lambda I)^{-1} DU^T y$$
$$= \sum_{j=1}^p u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y.$$


shrinkage factor
- Shrinks the coordinates with respect to the orthonormal basis formed by the principal components.
- Coordinates relating to the principal components with a smaller variance are shrunk more.

Shrinkage Methods: Ridge Regression

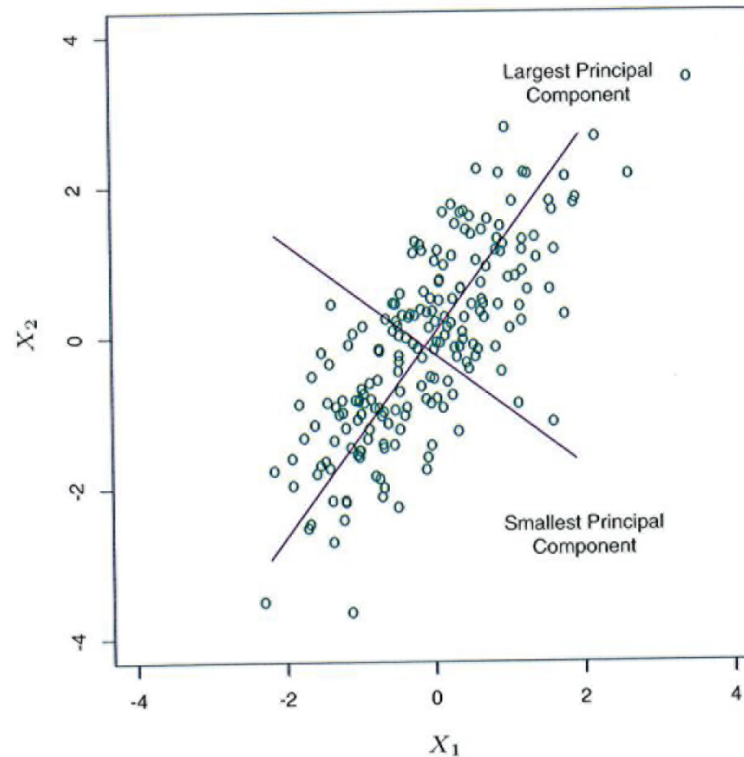


FIGURE 3.9. Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects \mathbf{y} onto these components, and then shrinks the coefficients of the low-variance components more than the high-variance components.

Shrinkage Methods: Ridge Regression

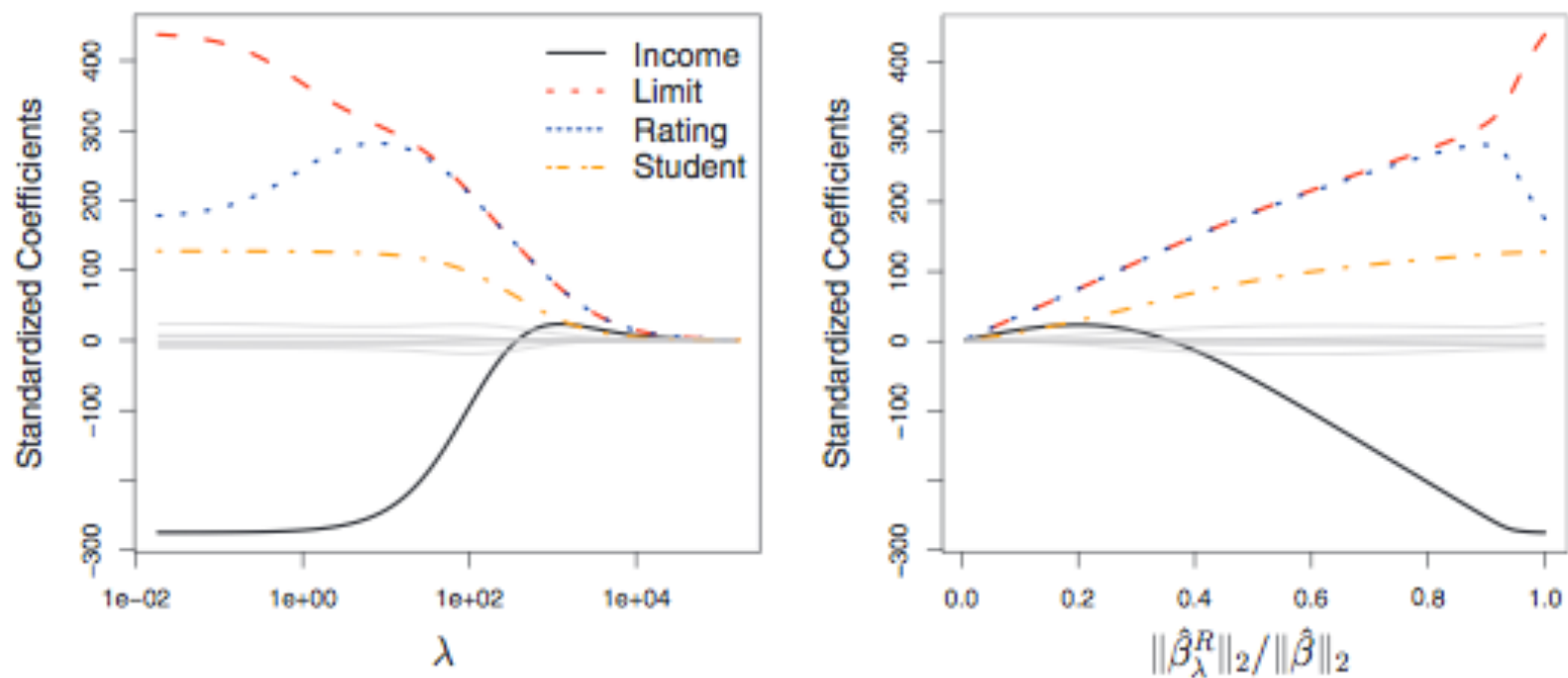


FIGURE 6.4. The standardized ridge regression coefficients are displayed for the **Credit** data set, as a function of λ and $\|\hat{\beta}_\lambda^R\|_2 / \|\hat{\beta}\|_2$.

Shrinkage Methods: Ridge Regression

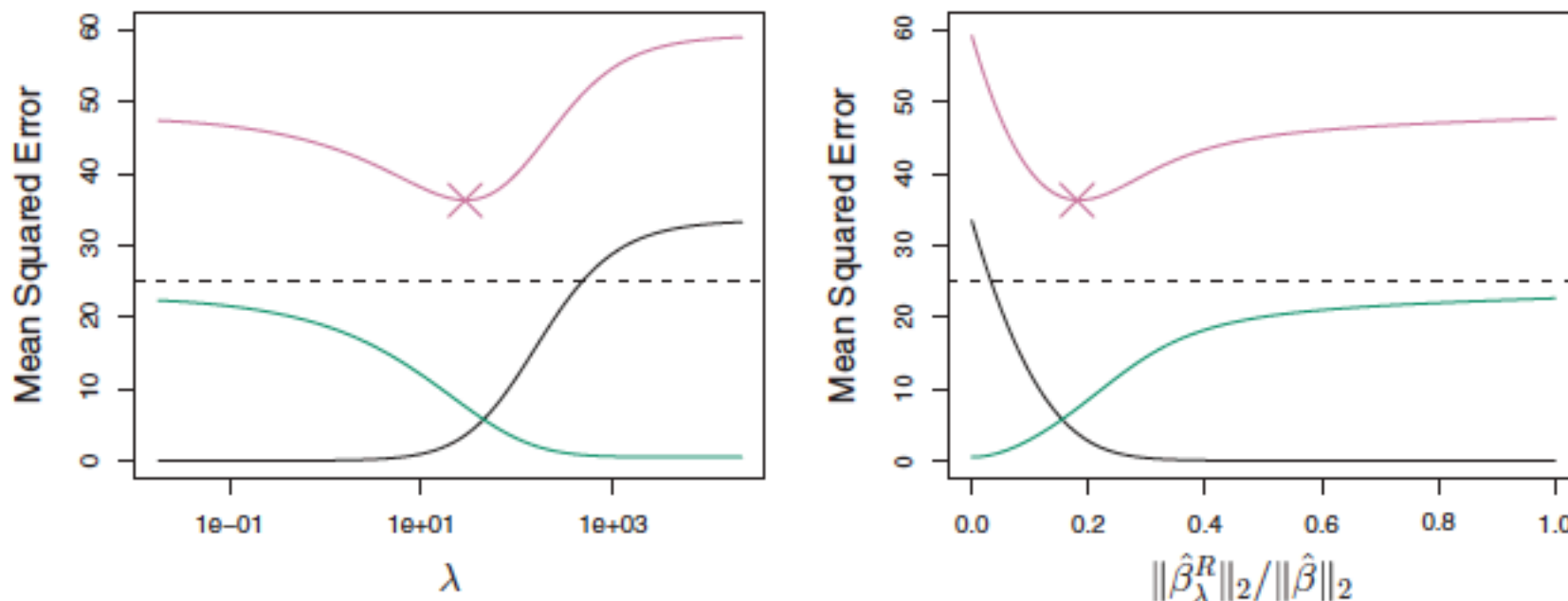


FIGURE 6.5. Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}_\lambda^R\|_2 / \|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses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

Conclusions

- OLS is the best unbiased estimate, but can have high variance.
- We are willing to often trade some bias, for a reduction in variance, we can achieve this through subset selection and shrinkage.
Why? We want smaller EPE.
- Subset selection is rather heuristic.
- Ridge regression shrinks the components that contribute least to the variation of the response.
- Comparing subset selection and ridge (verbal).

Desirable to shrink coefficients (as in ridge) and drop some predictors (as in subset selection).