

# Linear Regression Models

Based on Chapter 3 of  
Hastie, Tibshirani and Friedman

# Linear Regression Models

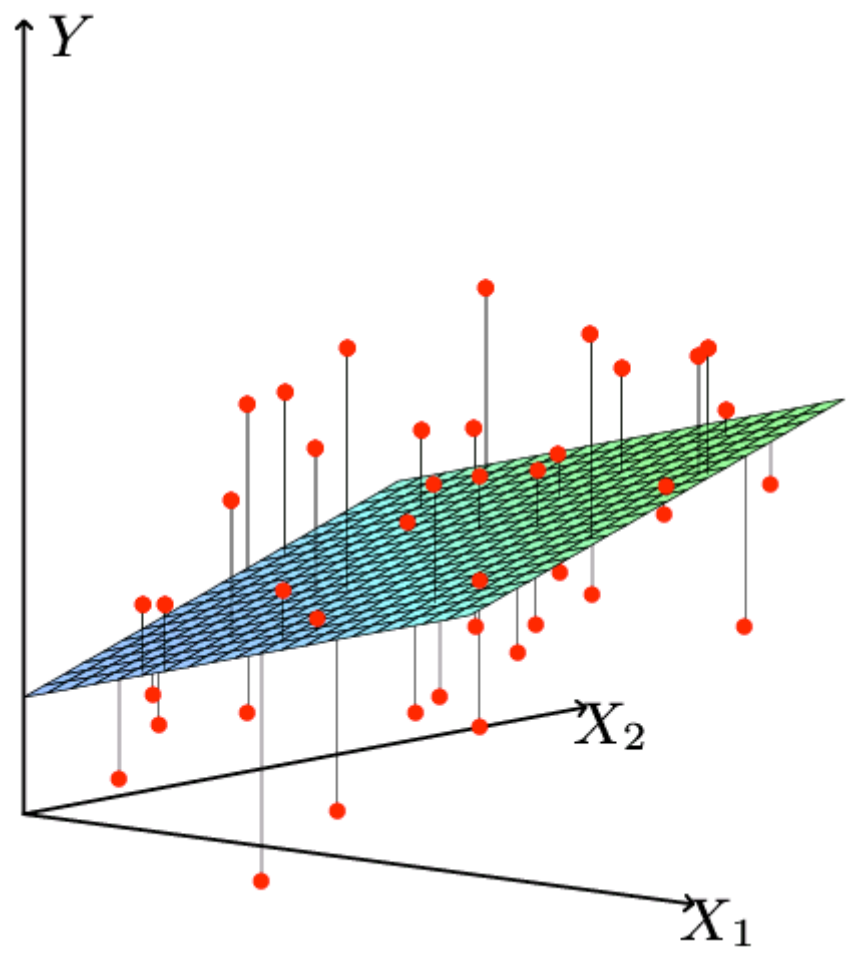
$$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j$$

Here the  $X$ 's might be:

- Raw predictor variables (continuous or coded-categorical)
- Transformed predictors ( $X_4 = \log X_3$ )
- Basis expansions ( $X_4 = X_3^2$ ,  $X_5 = X_3^3$ , etc.)
- Interactions ( $X_4 = X_2 X_3$ )

Popular choice for estimation is least squares:

$$RSS(\beta) = \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p X_j \beta_j)^2$$



# Least Squares

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

$$\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

$$\Rightarrow \hat{y} = X\hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{\text{hat matrix}} y$$

hat matrix

Often assume that the  $Y$ 's are independent and normally distributed, leading to various classical statistical tests and confidence intervals

# Gauss-Markov Theorem

Consider any linear combination of the  $\beta$ 's:  $\theta = a^T \beta$

The least squares estimate of  $\theta$  is:

$$\hat{\theta} = a^T \hat{\beta} = a^T (X^T X)^{-1} X^T y$$

If the linear model is correct, this estimate is unbiased ( $X$  fixed):

$$E(\theta) = E(a^T (X^T X)^{-1} X^T y) = a^T (X^T X)^{-1} X^T X \beta = a^T \beta$$

Gauss-Markov states that for any other linear unbiased estimator  $\tilde{\theta} = c^T y$ : i.e.,  $E(c^T y) = a^T \beta$ ,

$$\text{Var}(a^T \hat{\beta}) \leq \text{Var}(c^T y)$$

Of course, there might be a *biased* estimator with lower MSE...

# bias-variance

For any estimator  $\tilde{\theta}$  :

$$\begin{aligned}\text{MSE}(\tilde{\theta}) &= E(\tilde{\theta} - \theta)^2 \\ &= E(\tilde{\theta} - E(\tilde{\theta}) + E(\tilde{\theta}) - \theta)^2 \\ &= E(\tilde{\theta} - E(\tilde{\theta}))^2 + E(E(\tilde{\theta}) - \theta)^2 \\ &= \text{Var}(\tilde{\theta}) + \underbrace{(E(\tilde{\theta}) - \theta)^2}_{\text{bias}}\end{aligned}$$

Note MSE closely related to prediction error:

$$E(Y_0 - x_0^T \tilde{\beta})^2 = E(Y_0 - x_0^T \beta)^2 + E(x_0^T \tilde{\beta} - x_0^T \beta)^2 = \sigma^2 + \text{MSE}(x_0^T \tilde{\beta})$$

# Too Many Predictors?

When there are lots of  $X$ 's, get models with high variance and prediction suffers. Three “solutions:”

1. Subset selection

Score: AIC, BIC, etc.

All-subsets + leaps-and-bounds,

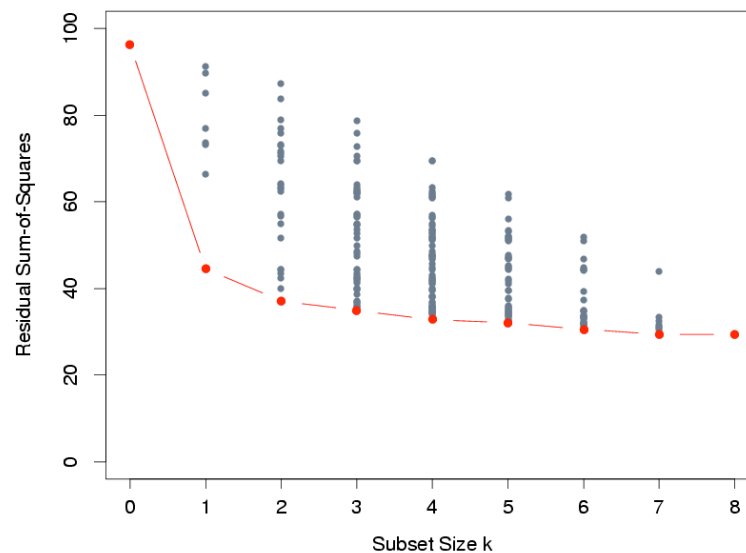
Stepwise methods,

2. Shrinkage/Ridge Regression

3. Derived Inputs

# Subset Selection

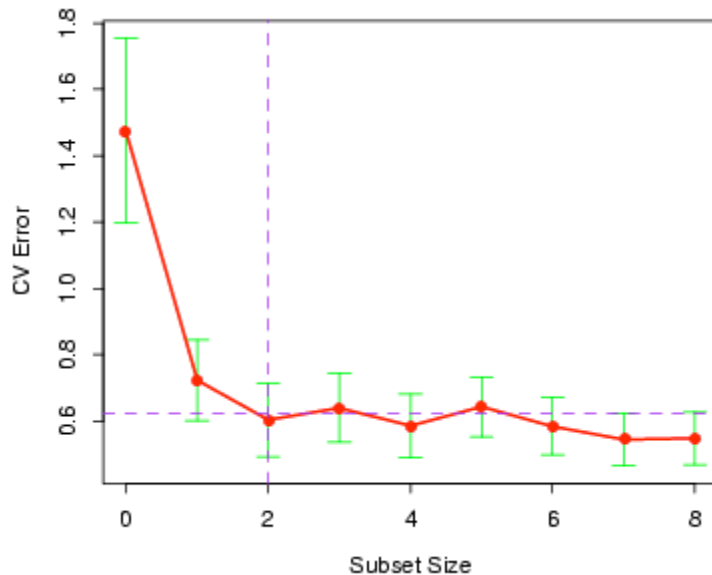
- Standard “all-subsets” finds the subset of size  $k$ ,  $k=1, \dots, p$ , that minimizes RSS:



- Choice of subset size requires tradeoff – AIC, BIC, marginal likelihood, cross-validation, etc.
- “Leaps and bounds” is an efficient algorithm to do all-subsets

# Cross-Validation

- e.g. 10-fold cross-validation:
  - Randomly divide the data into ten parts
  - Train model using 9 tenths and compute prediction error on the remaining 1 tenth
  - Do these for each 1 tenth of the data
  - Average the 10 prediction error estimates



“One standard error rule”

pick the simplest model within  
one standard error of the  
minimum

# Shrinkage Methods

- Subset selection is a discrete process – individual variables are either in or out
- This method can have high variance – a different dataset from the same source can result in a totally different model
- Shrinkage methods allow a variable to be partly included in the model. That is, the variable is included but with a shrunken co-efficient.

# Ridge Regression

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

subject to:  $\sum_{j=1}^p \beta_j^2 \leq s$


Equivalently:

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

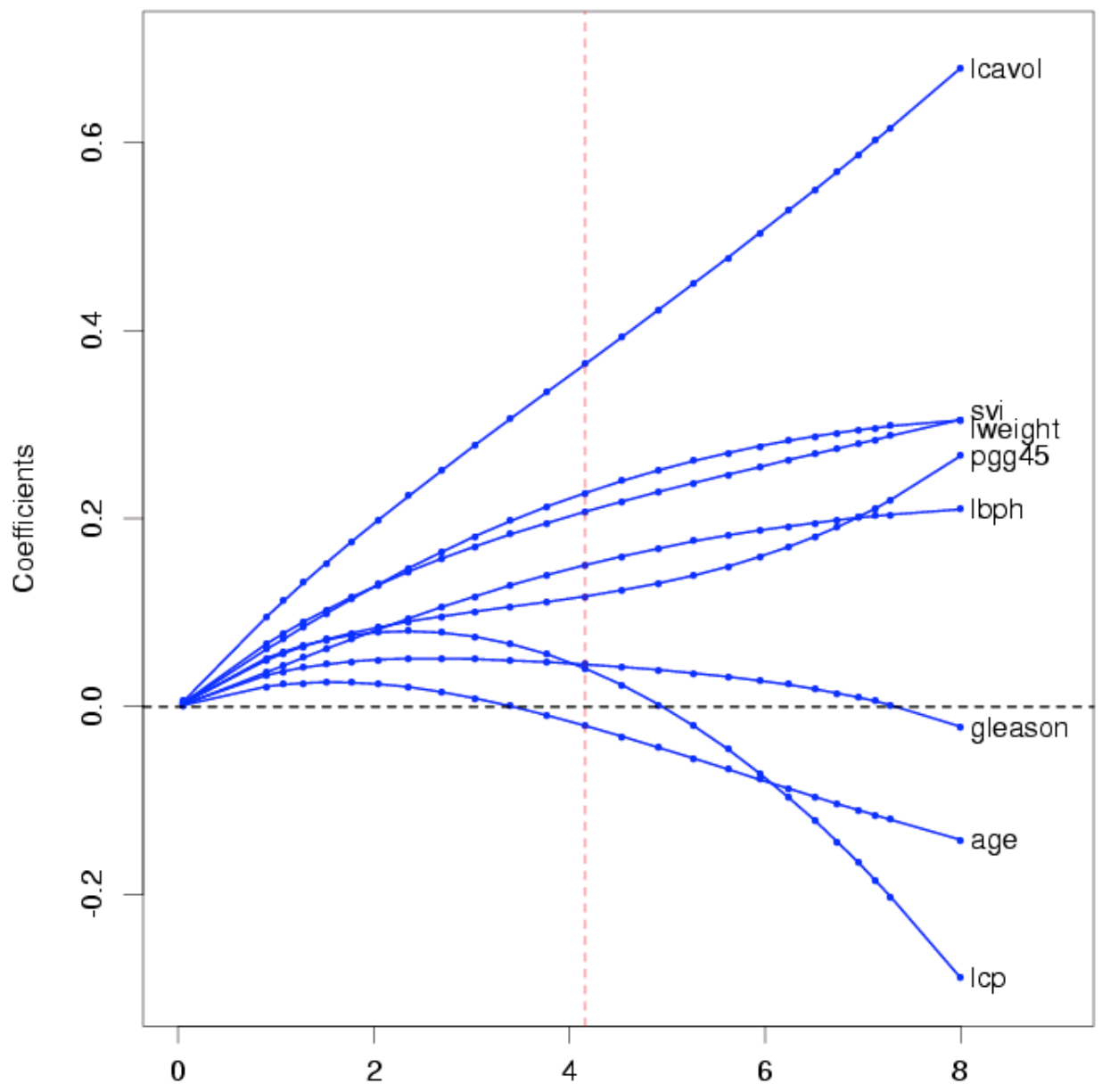
This leads to:

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

works even when  
 $X^T X$  is singular



Choose  $\lambda$  by cross-validation. Predictors should be centered.



$df(\lambda)$  ← effective number of  $X$ 's

# Ridge Regression = Bayesian Regression

$$y_i \sim N(\beta_0 + x_i^T \beta, \sigma^2)$$

$$\beta_j \sim N(0, \tau^2)$$

same as ridge with  $\lambda = \sigma^2 / \tau^2$

# The Lasso

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

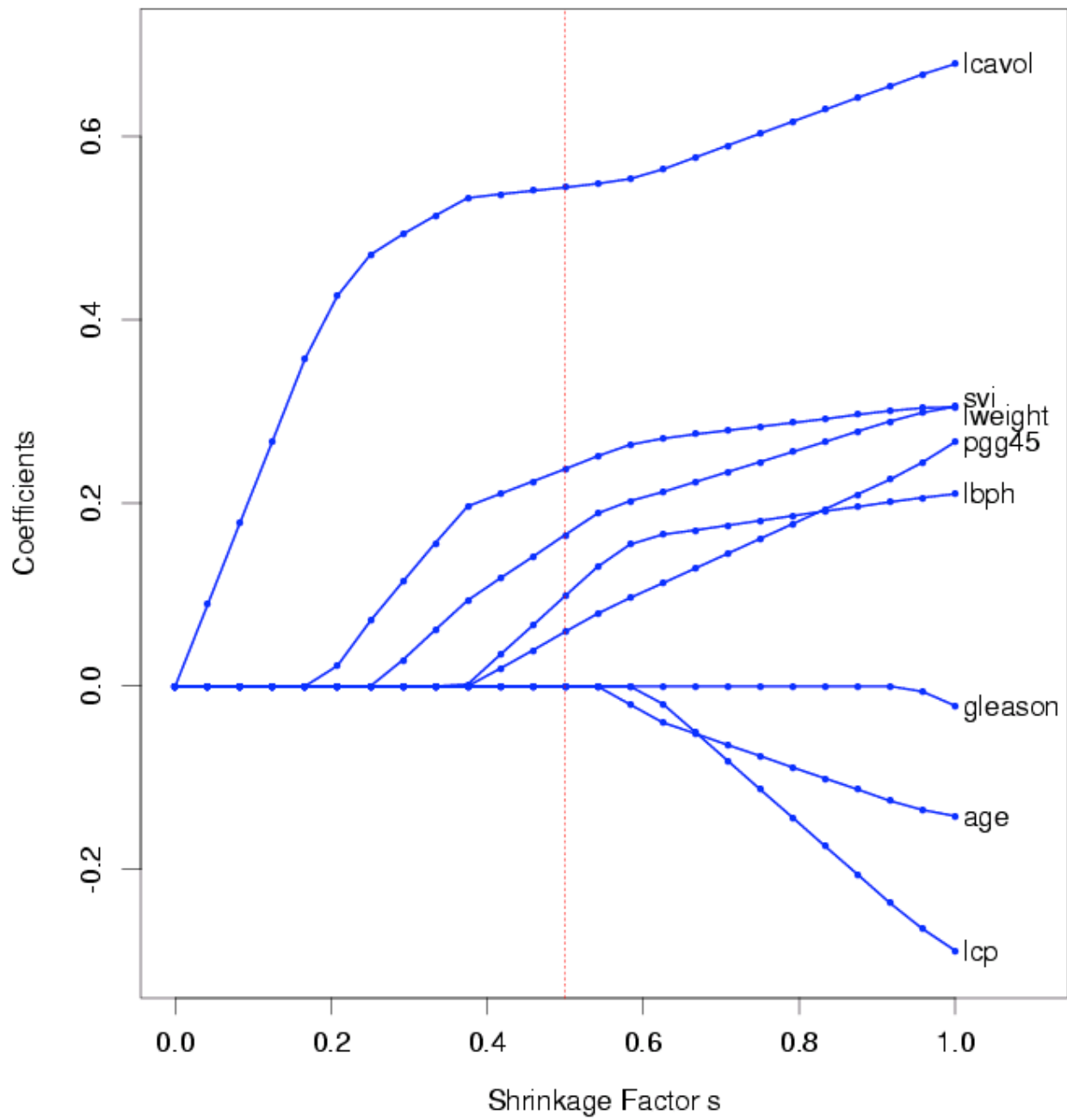
$$\text{subject to: } \sum_{j=1}^p |\beta_j| \leq s$$

Quadratic programming algorithm needed to solve for the parameter estimates. Choose  $s$  via cross-validation.

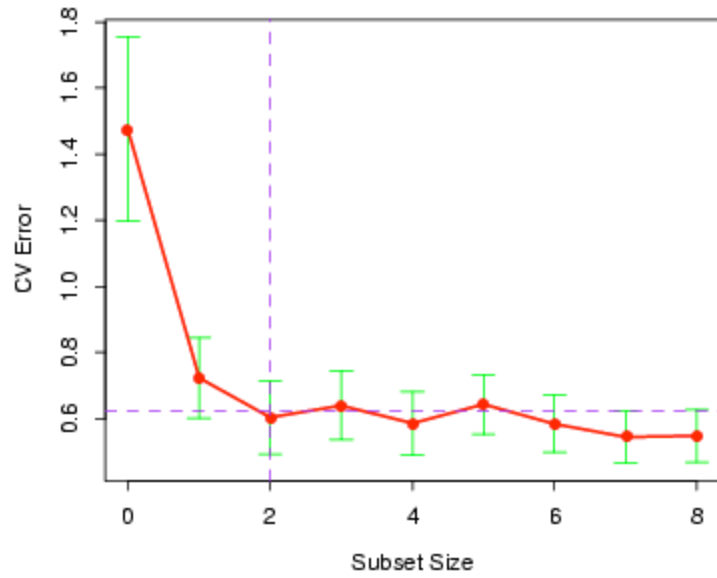
---

$$\tilde{\beta} = \arg \min_{\beta} \left( \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|^q \right)$$

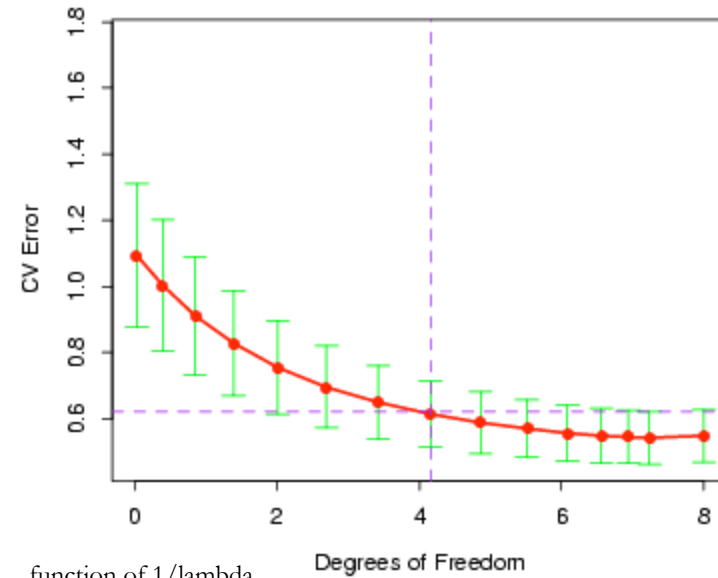
$q=0$ : var. sel.  
 $q=1$ : lasso  
 $q=2$ : ridge  
Learn  $q$ ?



### All Subsets



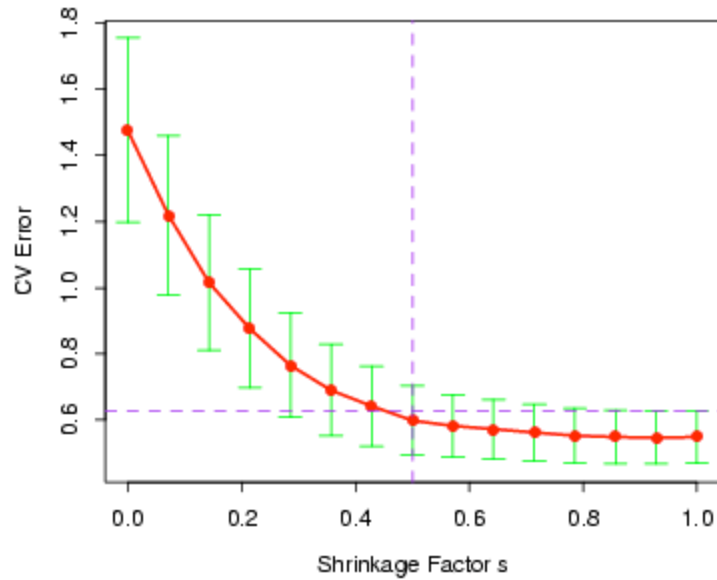
### Ridge Regression



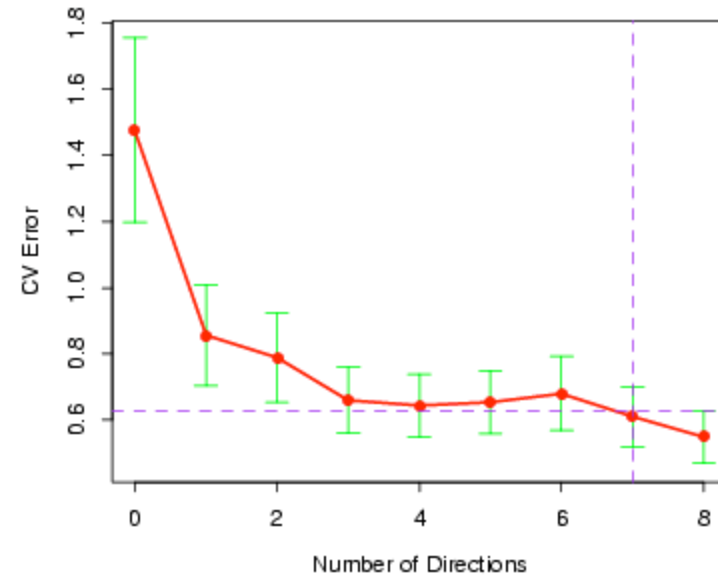
function of  $1/\lambda$

Degrees of Freedom

### Lasso



### Principal Components Regression



# Principal Component Regression

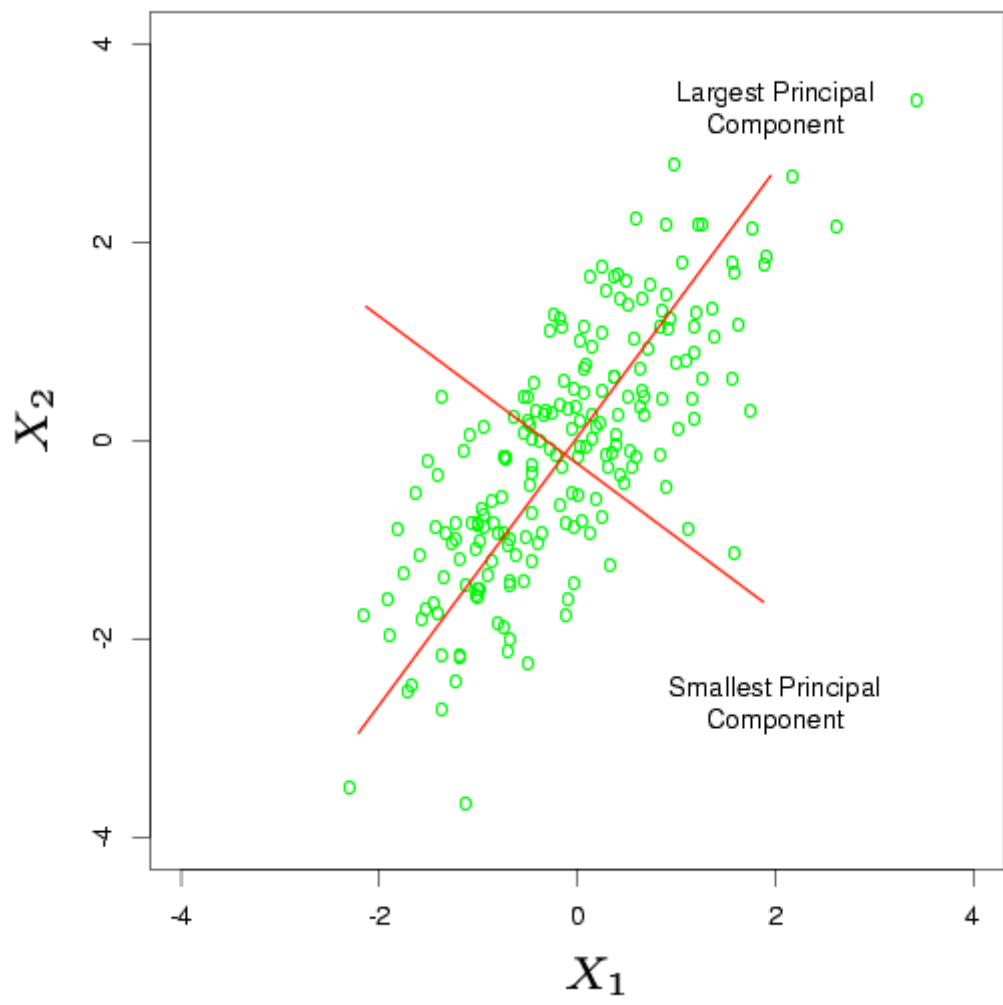
Consider a an eigen-decomposition of  $X^T X$  (and hence the covariance matrix of  $X$ ):

$$X^T X = VD^2V^T$$

The eigenvectors  $v_j$  are called the *principal components* of  $X$   
 $D$  is diagonal with entries  $d_1 \geq d_2 \geq \dots \geq d_p$

$Xv_1$  has largest sample variance amongst all normalized linear combinations of the columns of  $X$  ( $\text{var}(Xv_1) = \frac{d_1^2}{N}$ )

$Xv_k$  has largest sample variance amongst all normalized linear combinations of the columns of  $X$  subject to being orthogonal to all the earlier ones



# Principal Component Regression

PC Regression regresses on the first  $M$  principal components where  $M < p$

Similar to ridge regression in some respects – see HTF, p.66

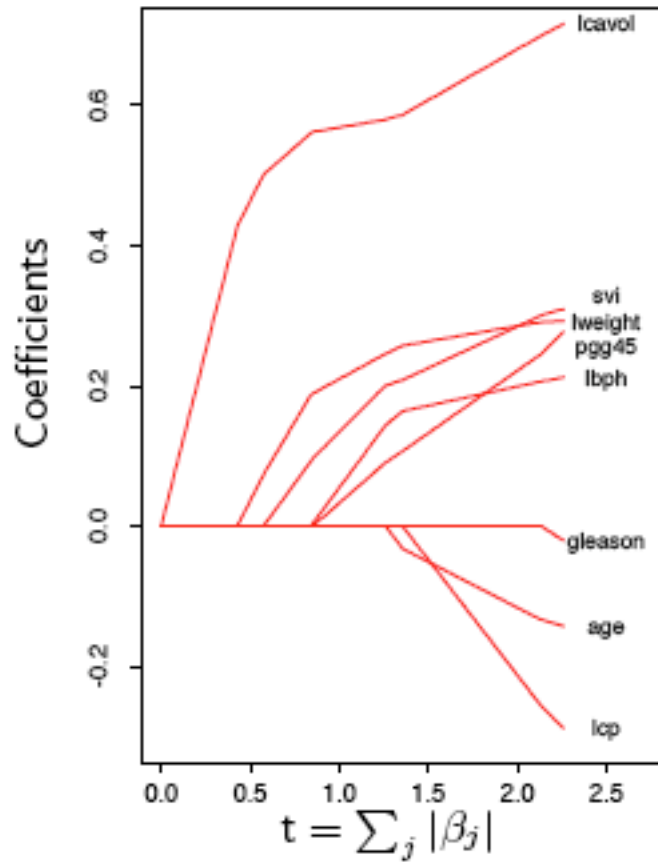
# Forward Stagewise Regression

(Forward Stagewise = Least Squares Boosting)

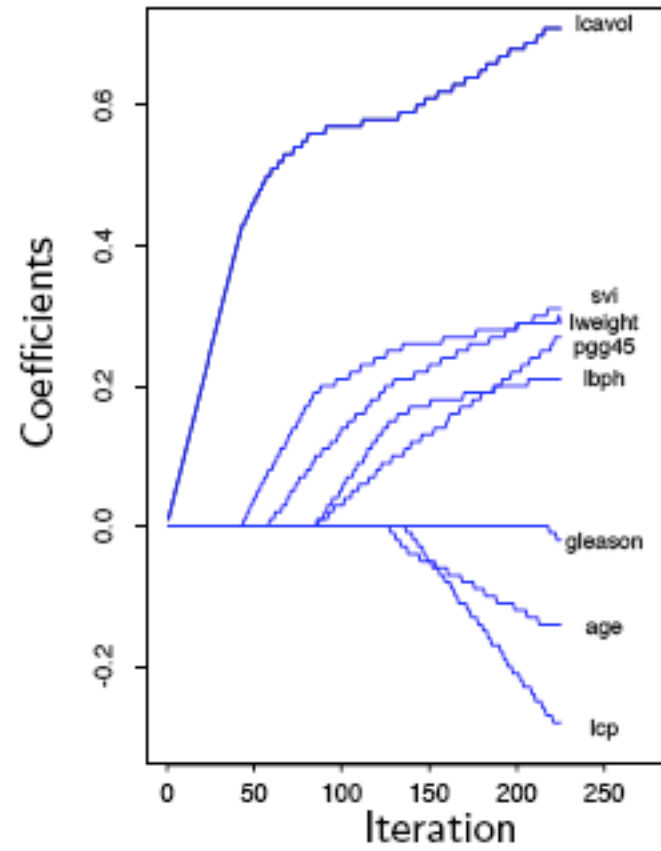
1. Initialize: standardize predictors, center  $y$ ,  
 $r = y, \beta_1 = \dots = \beta_p = 0$
2. Repeat many times
  - ▶ Find the predictor  $x_j$  most correlated with  $r$
  - ▶  $\delta = \epsilon \text{sign}(r \cdot x_j)$
  - ▶  $\hat{\beta}_j \leftarrow \hat{\beta}_j + \delta$
  - ▶  $r \leftarrow r - \delta x_j$

# Prostate Cancer Data

Lasso



Forward Stagewise



## Similarity:

Are LASSO and infinitesimal forward stagewise identical?

- ▶ With orthogonal predictors, yes.
- ▶ Otherwise similar.

Least Angle Regression provides explanation, and fast implementation.

# Stepwise, Forward Stagewise, Least Angle

## Stepwise regression:

- ▶ Pick predictor most correlated with  $y$
- ▶ Bring predictor completely into model (full LS fit)

## Forward stagewise:

- ▶ Pick predictor most correlated with  $y$
- ▶ Increment coefficient for predictor

## Least Angle Regression:

- ▶ Pick predictor most correlated with  $y$
- ▶ Bring predictor into model only to extent it is better than others
- ▶ Move in least-squares direction until another variable is as correlated

## Least Angle Regression — LAR

*Like a “more democratic” version of forward stepwise regression.*

1. Start with  $r = y$ ,  $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p = 0$ . Assume  $x_j$  standardized.
2. Find predictor  $x_j$  most correlated with  $r$ .
3. Increase  $\beta_j$  in the direction of  $\text{sign}(\text{corr}(r, x_j))$  until some other competitor  $x_k$  has as much correlation with current residual as does  $x_j$ .
4. Move  $(\hat{\beta}_j, \hat{\beta}_k)$  in the joint least squares direction for  $(x_j, x_k)$  until some other competitor  $x_\ell$  has as much correlation with the current residual
5. Continue in this way until all predictors have been entered. Stop when  $\text{corr}(r, x_j) = 0 \forall j$ , i.e. OLS solution.

`lars` : Efron and Hastie (S-PLUS and R)

- ▶ Linear regression

`glm` : Park and Hastie (R)

- ▶ GLM and Cox Proportional Hazards

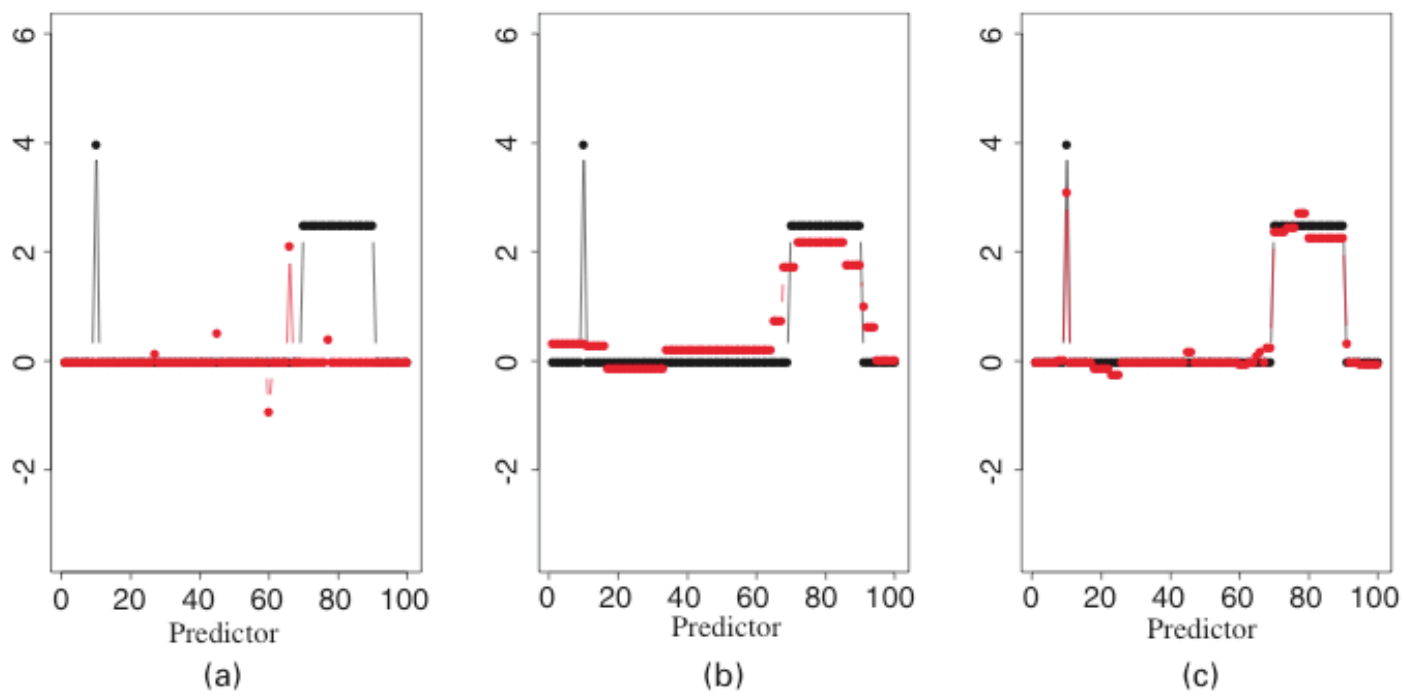
Methods: `plot`, `print`, `predict`, `cv`, `coef`

# Fused Lasso

- If there are many correlated features, lasso gives non-zero weight to only one of them
- Maybe correlated features (e.g. time-ordered) should have similar coefficients?

$$\hat{\beta} = \arg \min \left\{ \sum_i \left( y_i - \sum_j x_{ij} \beta_j \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^p |\beta_j| \leq s_1 \text{ and } \sum_{j=2}^p |\beta_j - \beta_{j-1}| \leq s_2$$



**Fig. 4.** Simulated example with only two areas of non-zero coefficients (black points and lines; red points, estimated coefficients from each method): (a) lasso,  $s_1 = 4.2$ ; (b) fusion,  $s_2 = 5.2$ ; (c) fused lasso,  $s_1 = 56.5$ ,  $s_2 = 13$

# Group Lasso

- Suppose you represent a categorical predictor with indicator variables
- Might want the set of indicators to be in or out

regular lasso:

$$\hat{\beta}_\lambda = \arg \min_{\beta} \|Y - X\beta\|_2^2 + \lambda \sum_{i=1}^p |\beta_i|$$

group lasso:

$$\hat{\beta}_\lambda = \arg \min_{\beta} \|Y - X\beta\|_2^2 + \lambda \sum_{g=1}^G \|\beta_{\mathcal{I}_g}\|_2$$

