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:

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

Assumptions of Linear Regression Model

The regression model is linear in the coefficients and the error term

The error term has a population mean of zero

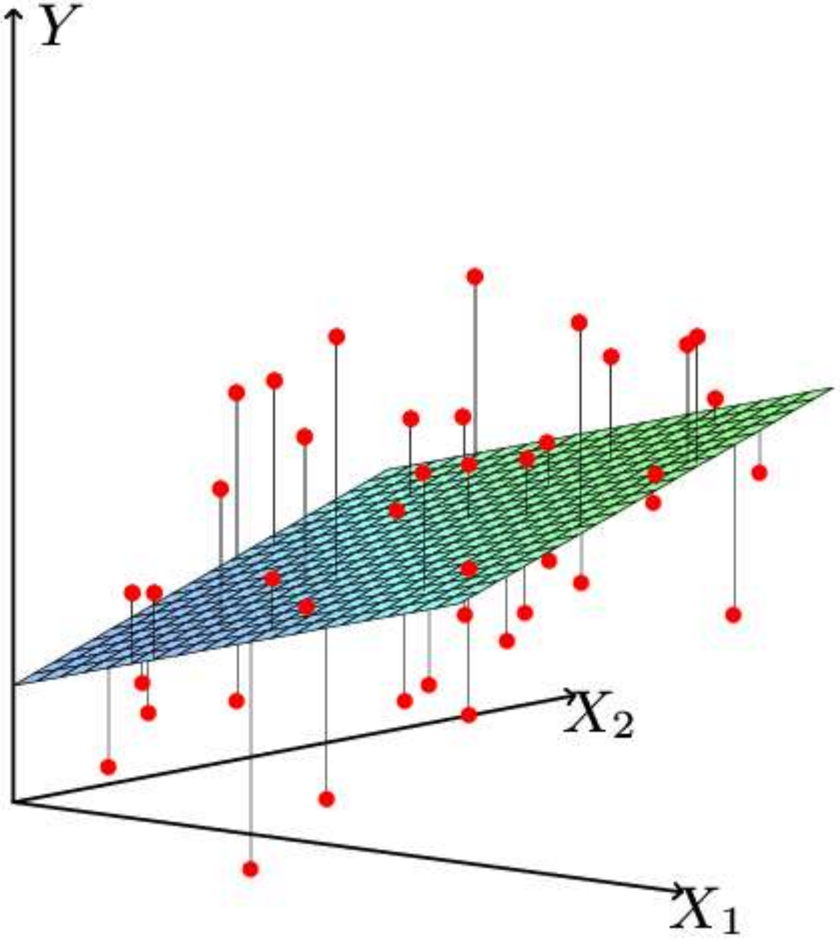
All independent variables are uncorrelated with the error term

Observations of the error term are uncorrelated with each other

The error term has a constant variance (no heteroscedasticity)

No independent variable is a perfect linear function of other explanatory variables

The error term is normally distributed


Y
X_2
X_1

Least Squares

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

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

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

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 β's: $\theta = a^T \beta$

The least squares estimate of θ 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):

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

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

$$\mathrm{Var}(a^T \hat{\beta}) \le \mathrm{Var}(c^T y)$$

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

bias-variance

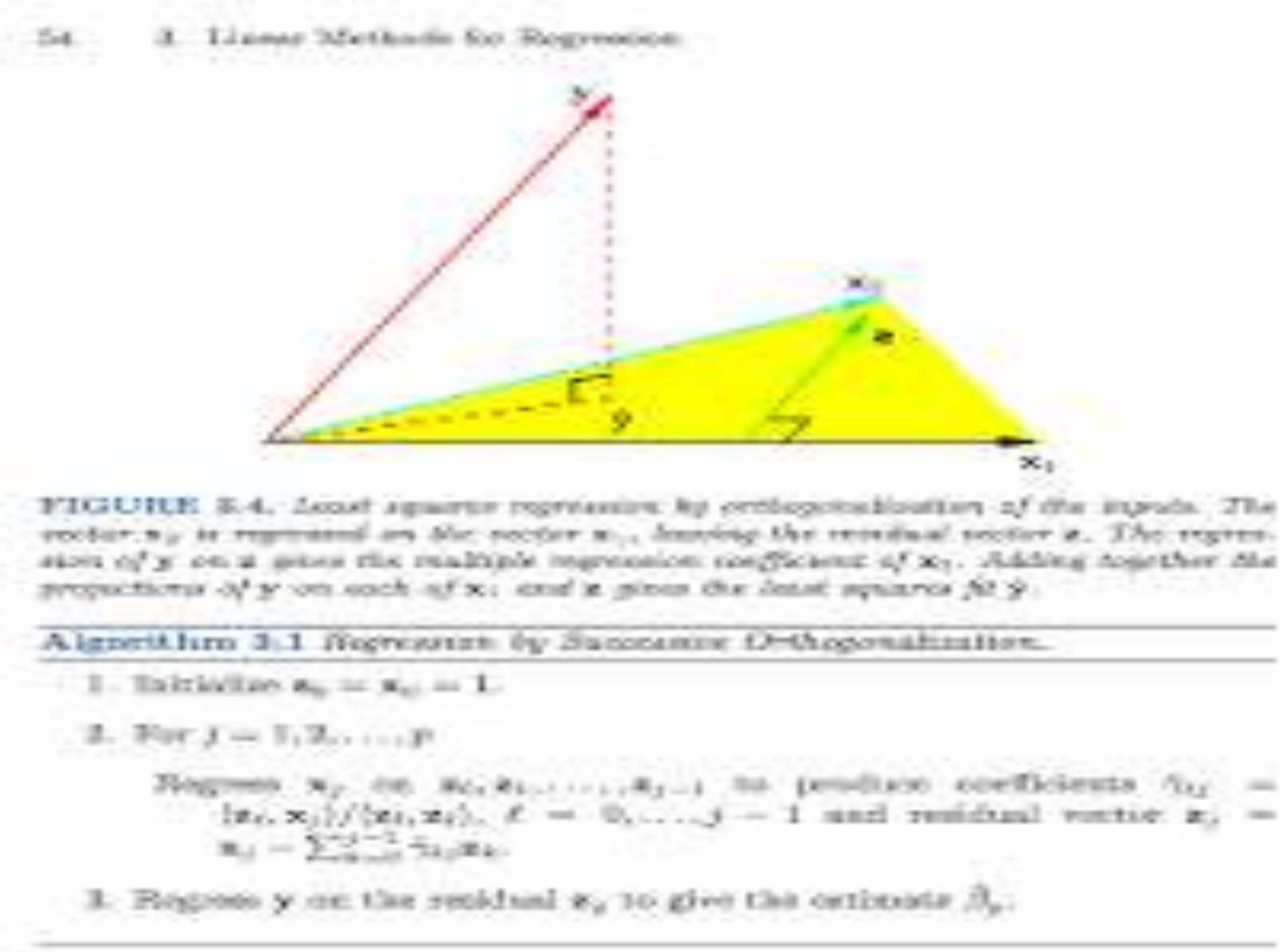
For any estimator $\tilde{\theta}$:

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

Note MSE closely related to prediction error:

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

Representation of Multivariate using Univariate



54 3. Linear Methods for Regression

FIGURE 3.4. *Least squares regression by orthogonalization of the inputs. The vector* $\mathbf{x}_2$ *is regressed on the vector* $\mathbf{x}_1$*, leaving the residual vector* $\mathbf{z}$*. The regression of* $\mathbf{y}$ *on* $\mathbf{z}$ *gives the multiple regression coefficient of* $\mathbf{x}_2$*. Adding together the projections of* $\mathbf{y}$ *on each of* $\mathbf{x}_1$ *and* $\mathbf{z}$ *gives the least squares fit* $\hat{\mathbf{y}}$*.*

Algorithm 3.1 *Regression by Successive Orthogonalization.*

```
1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$.
2. For $j = 1, 2, \ldots, p$
     Regress $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{j-1}$ to produce coefficients $\hat{\gamma}_{\ell j} = \langle \mathbf{z}_\ell, \mathbf{x}_j \rangle / \langle \mathbf{z}_\ell, \mathbf{z}_\ell \rangle$, $\ell = 0, \ldots, j-1$ and residual vector $\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k$.
3. Regress $\mathbf{y}$ on the residual $\mathbf{z}_p$ to give the estimate $\hat{\beta}_p$.
```

Too Many Predictors?

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

1. Subset selection

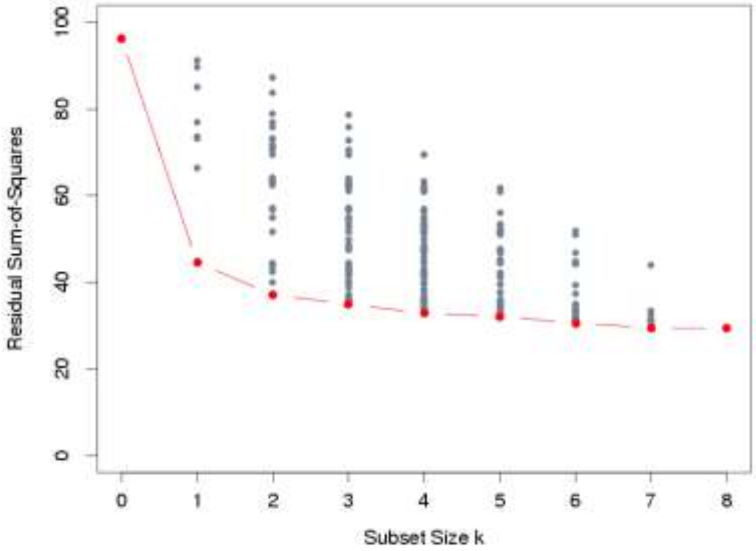
   Score: AIC, BIC, etc.
   All-subsets + leaps-and-bounds,
   Stepwise methods,

2. Shrinkage/Ridge Regression

3. Derived Inputs

Subset Selection

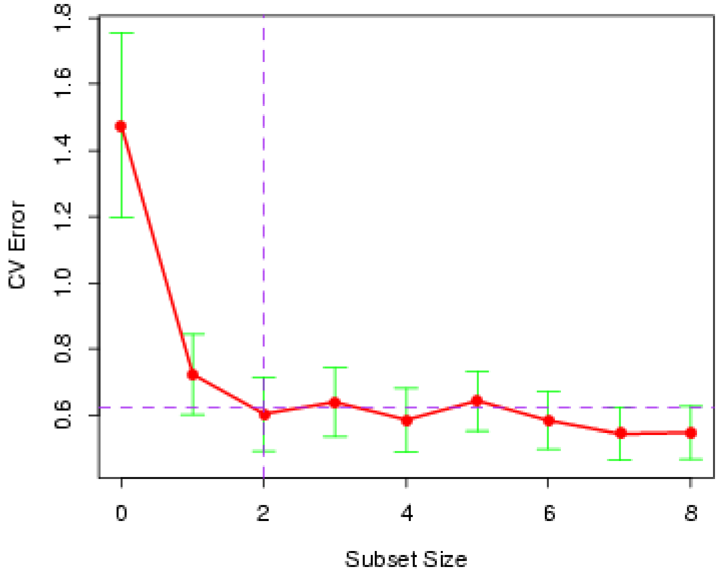
•Standard "all-subsets" finds the subset of size k, k=1,…,p, that minimizes RSS:



•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:

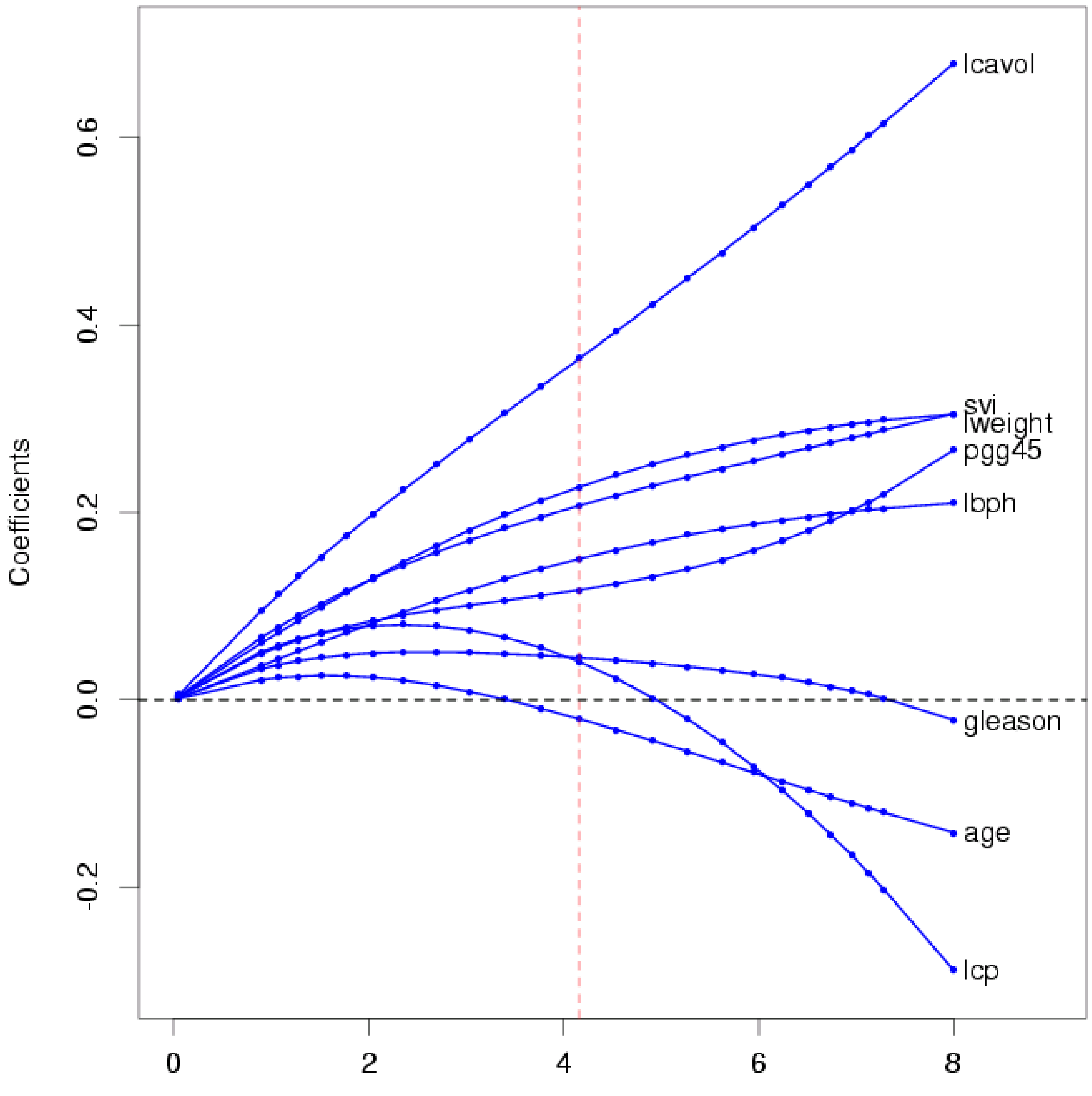
$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left(\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

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 λ by cross-validation. Predictors should be centered.



$\mathrm{df}(\lambda)$ ← 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

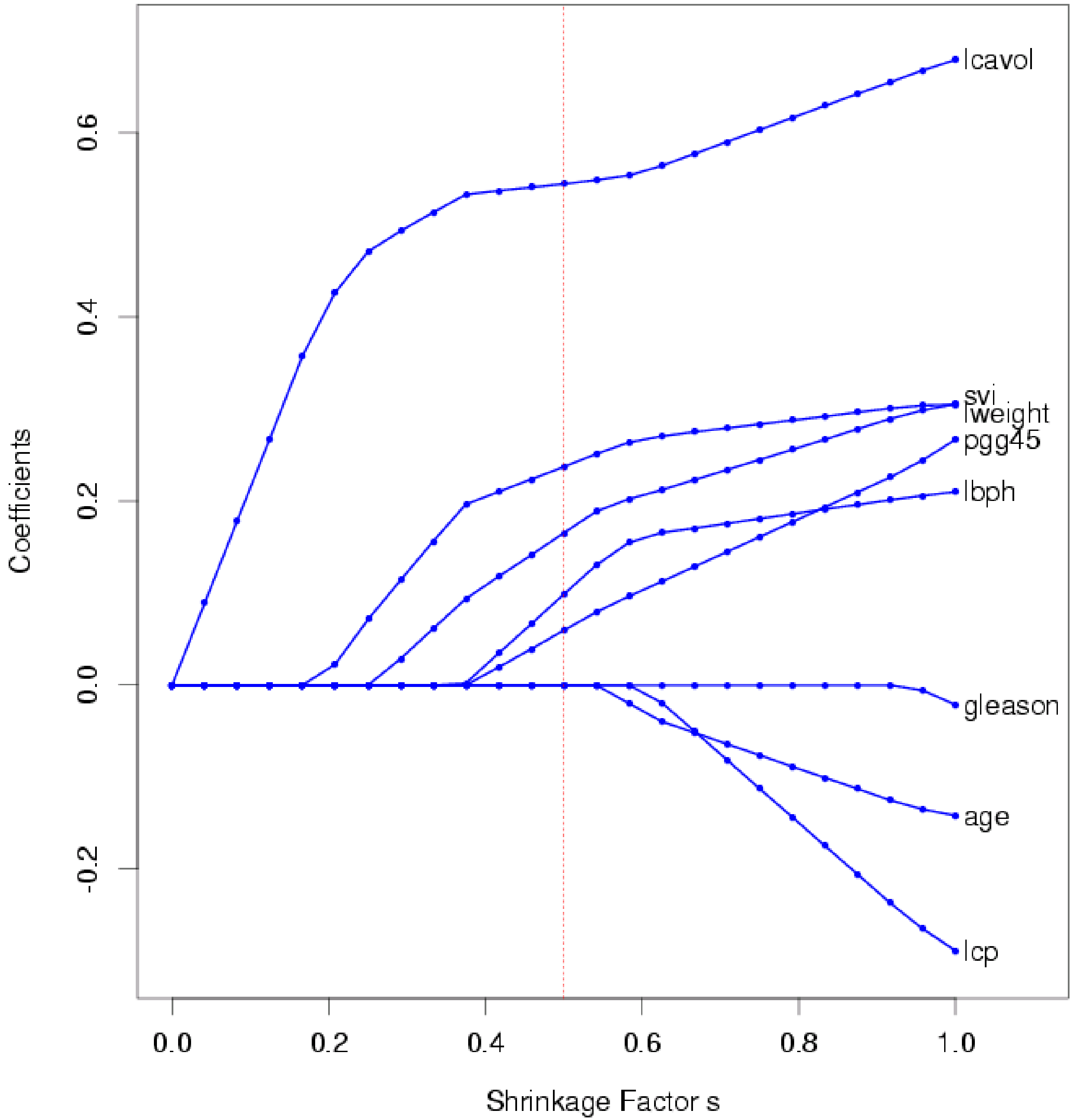
$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

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

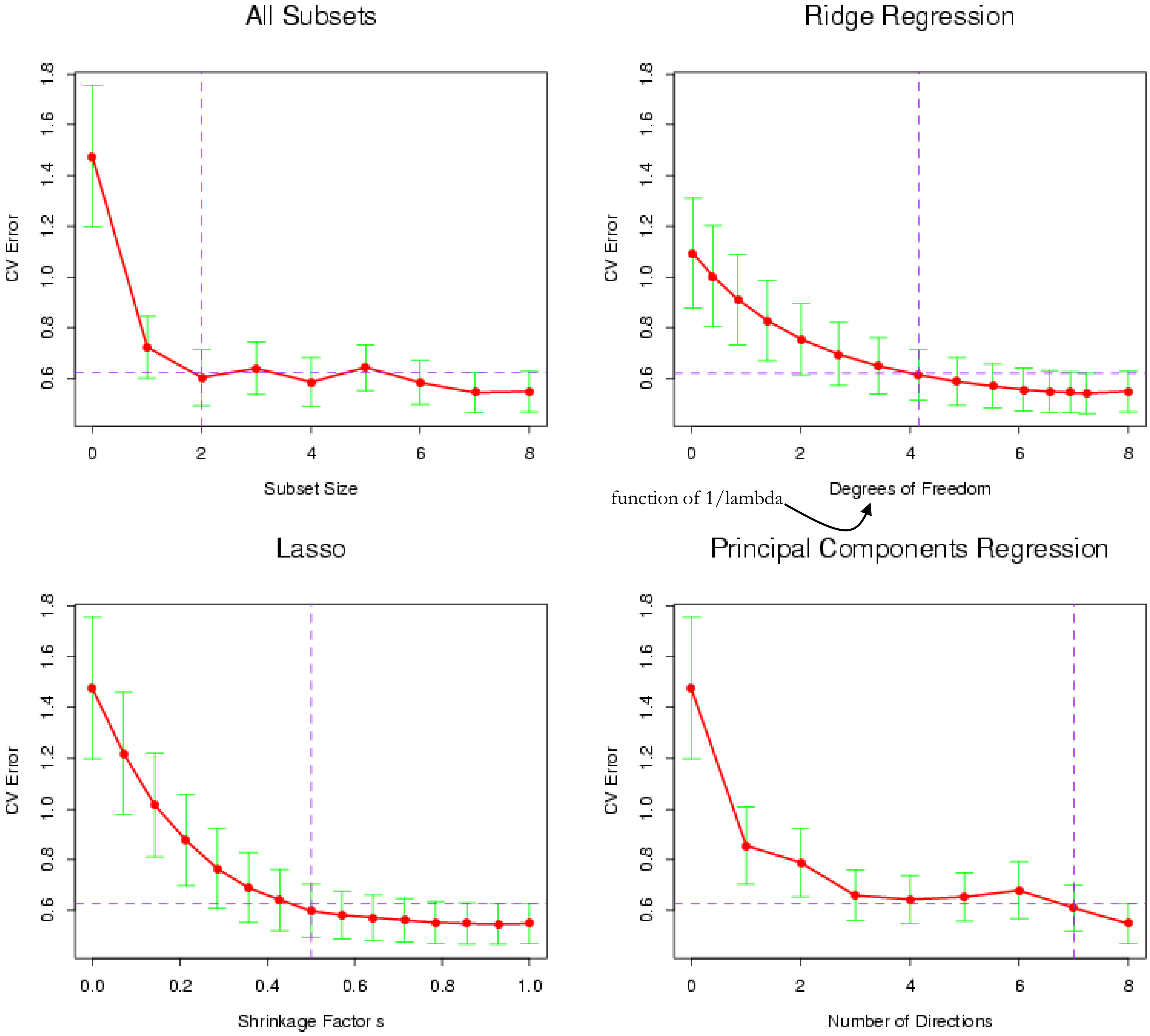
Quadratic programming algorithm needed to solve for the parameter estimates. Choose *s* via cross-validation.

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

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



Coefficients
0.6
0.4
0.2
0.0
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
Shrinkage Factor s
lcavol
svi
lweight
pgg45
lbph
gleason
age
lcp



function of 1/lambda

Principal Component Regression

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

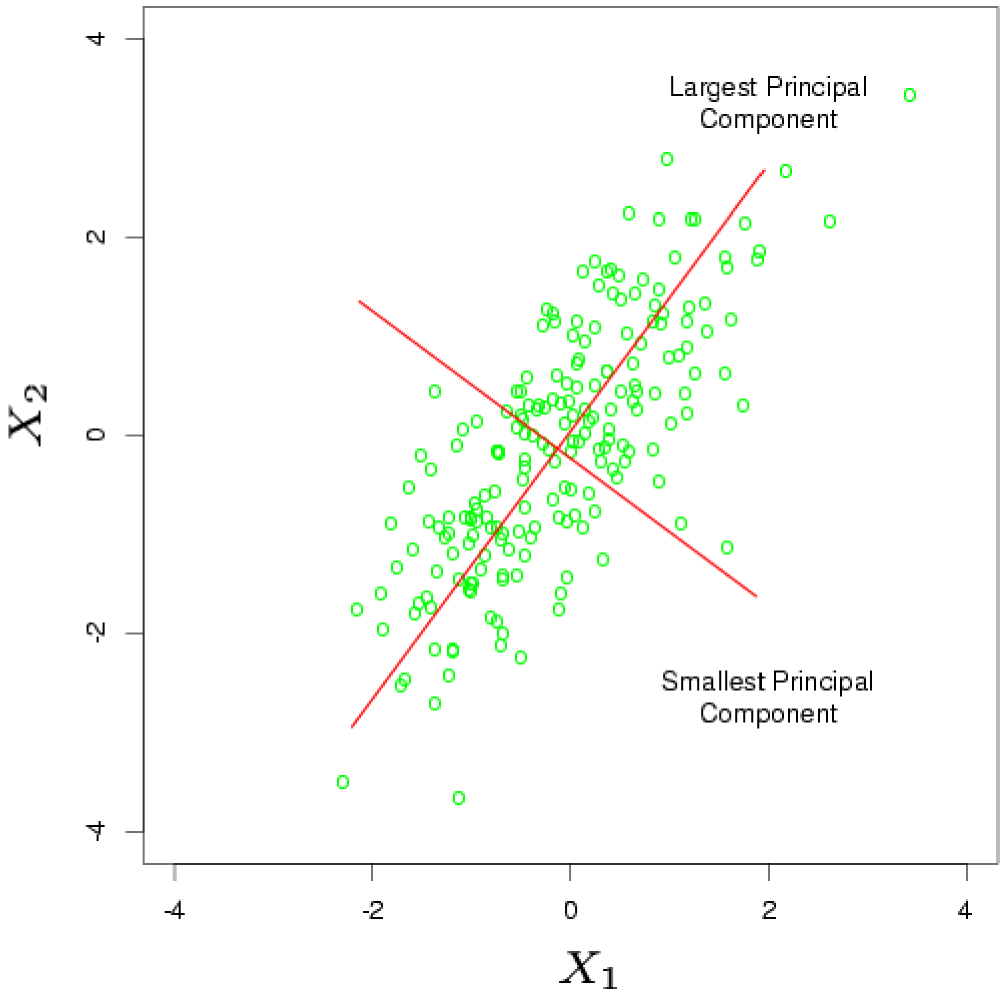
$$X^T X = VD^2V^T$$

The eigenvectors v_j are called the *principal components* of X

D is diagonal with entries $d_1 \geq d_2 \geq \ldots \geq d_p$

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

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


Largest Principal Component
Smallest Principal Component
X_2
X_1
4
2
0
-2
-4
-4
-2
0
2
4

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