MATH 569 Statistical Learning

Part II: Linear Methods of Regression

Maggie Cheng

ILLINOIS INSTITUTE OF TECHNOLOGY

Fig 3.1 Linear regression: minimizing RSS

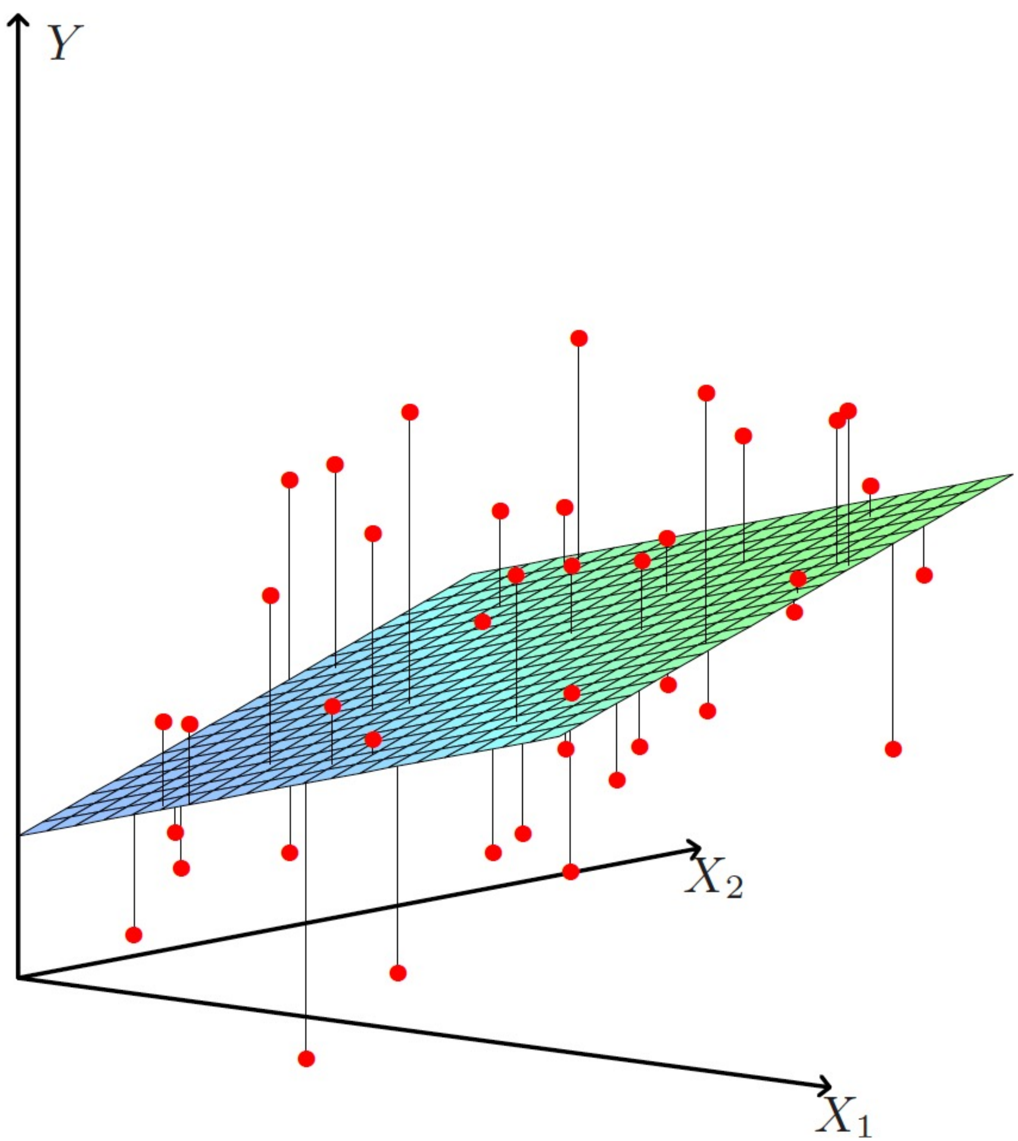
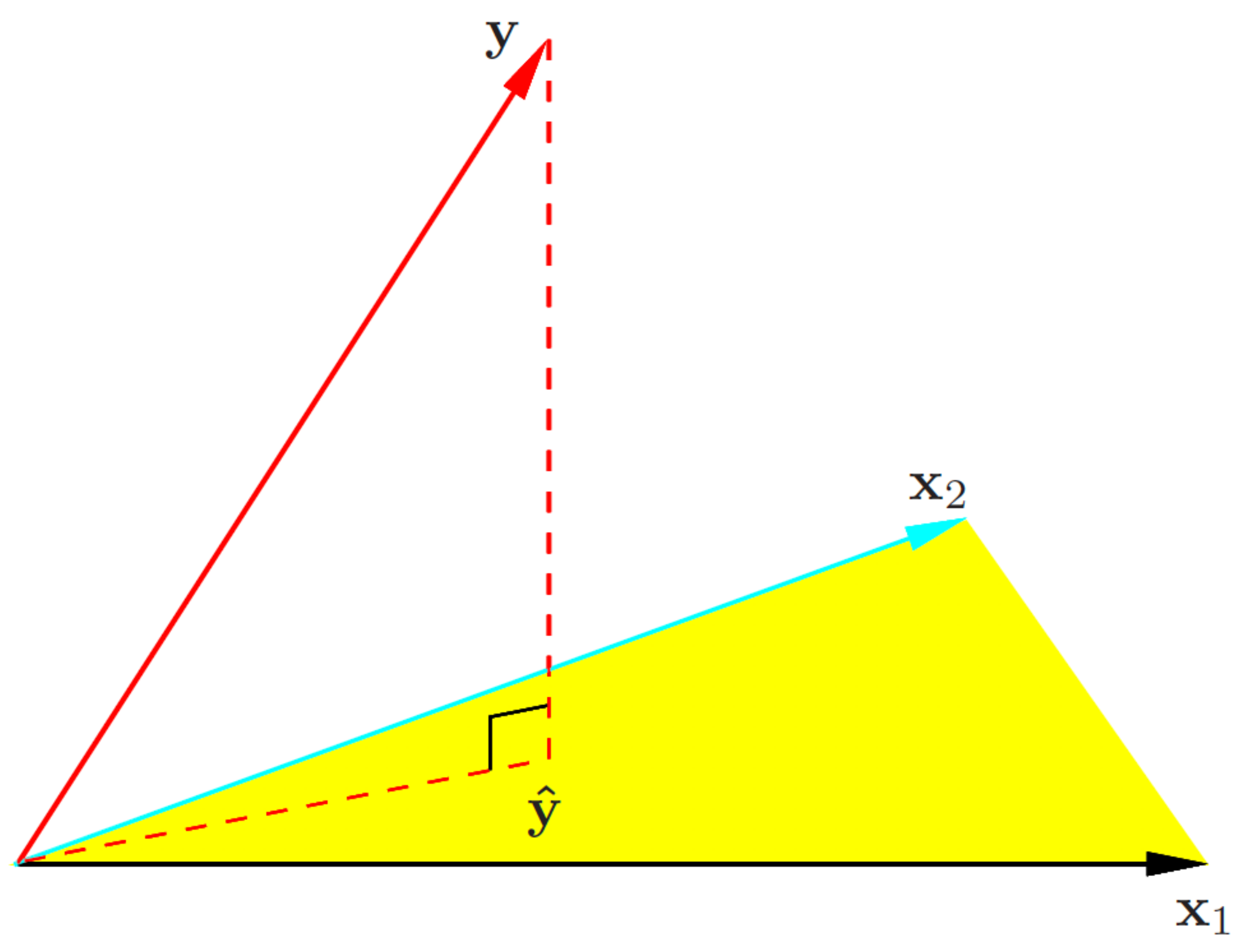


Fig 3.2, $(Y - \hat{Y})$ orthogonal to the subspace spanned by the column vectors of X



Gram–Schmidt procedure for multiple regression

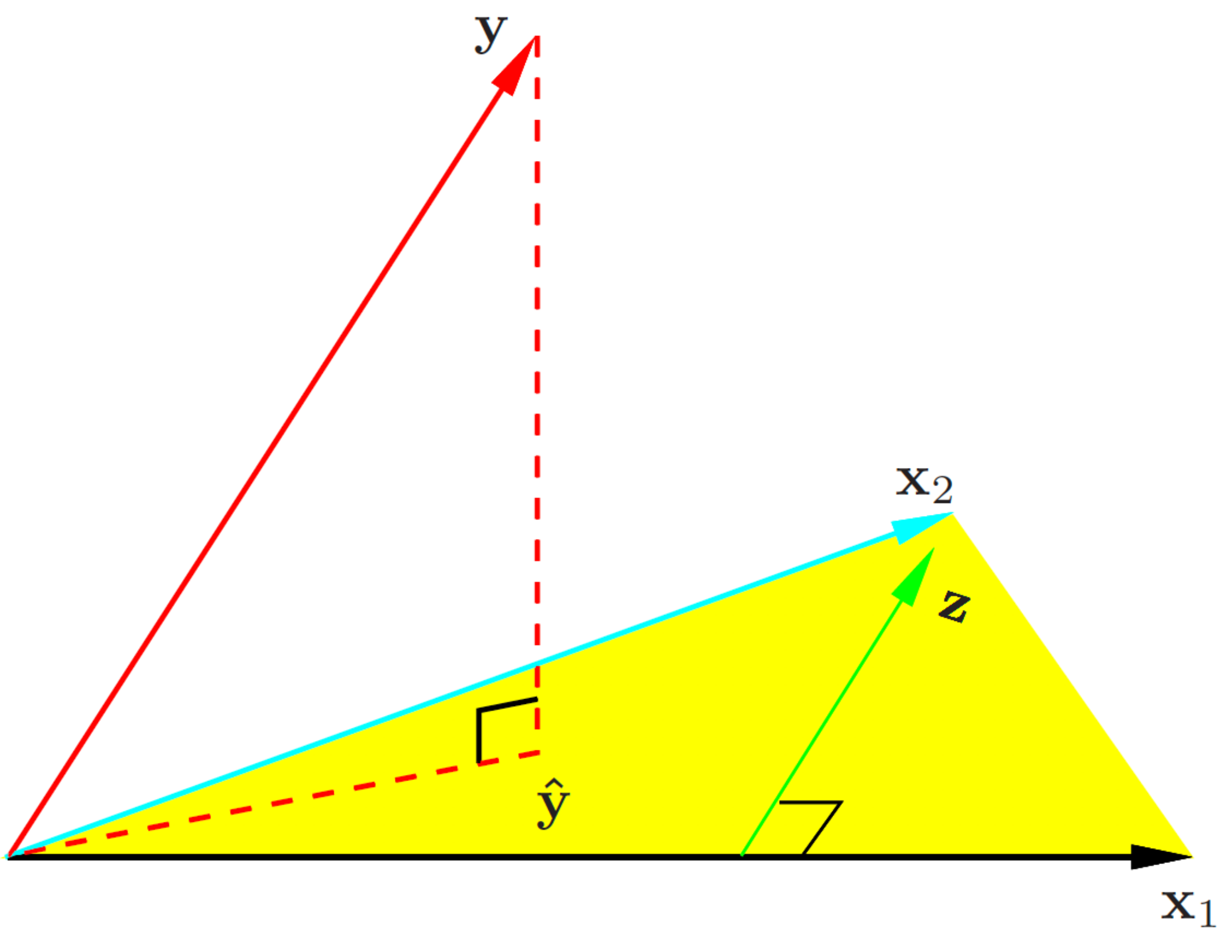


FIGURE 3.4. *Least squares regression by orthogonalization of the inputs. The vector* $\mathbf{x}_2$ *is regressed on the vector* $\mathbf{x}_1$*, leaving the residual vector* $\mathbf{z}$*. The regression of* $\mathbf{y}$ *on* $\mathbf{z}$ *gives the multiple regression coefficient of* $\mathbf{x}_2$*.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

Gram–Schmidt procedure for multiple regression

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 $\hat{\beta}_p$.

Fig 3.9 Ridge regression shrinks the coefficients of the low-variance components more than the high-variance components

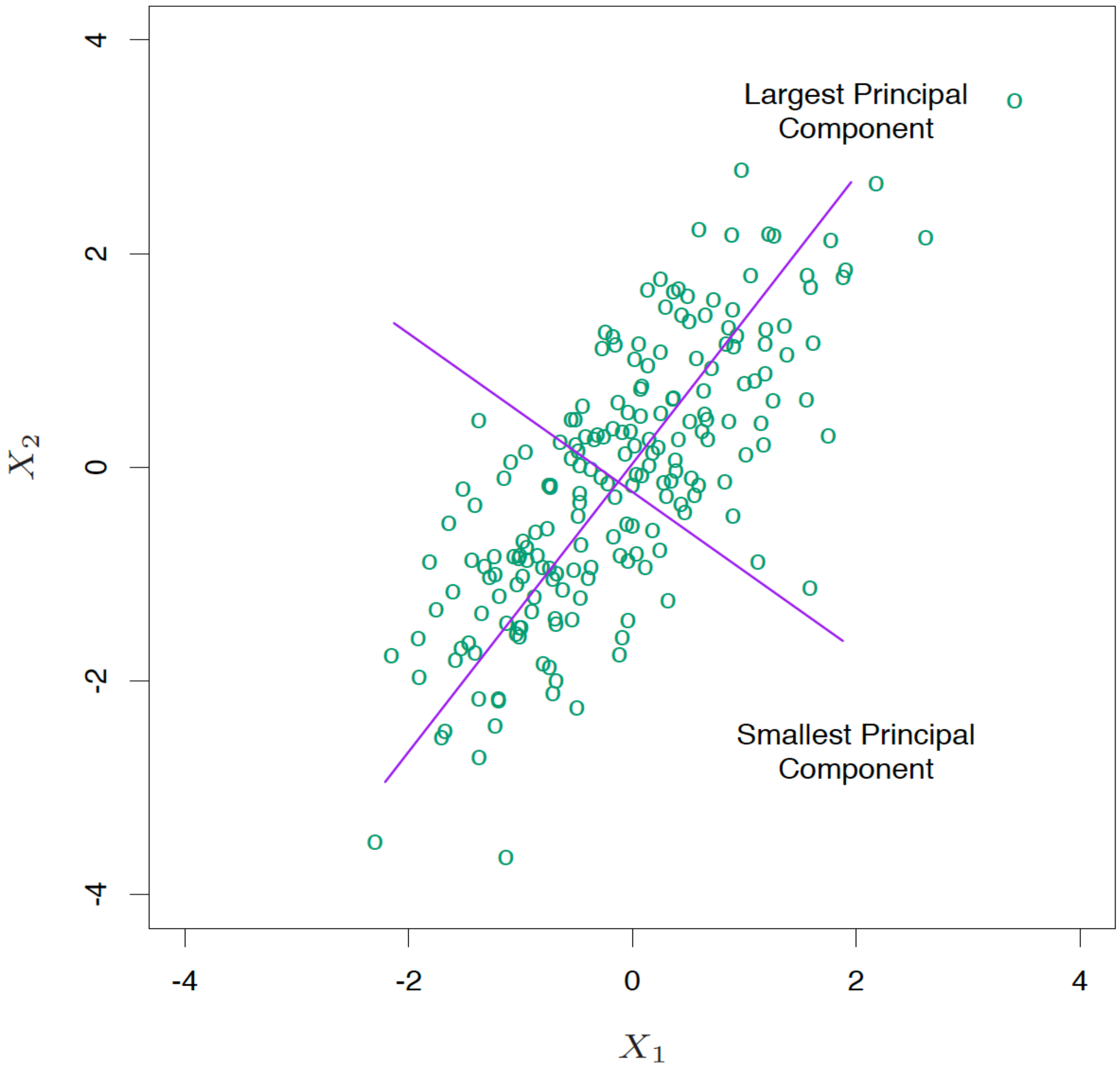


Fig 3.8 Ridge Regression

Ridge coefficients vs. df
df(lambda) monotonically
decreases as lambda increases
(prostate cancer example)

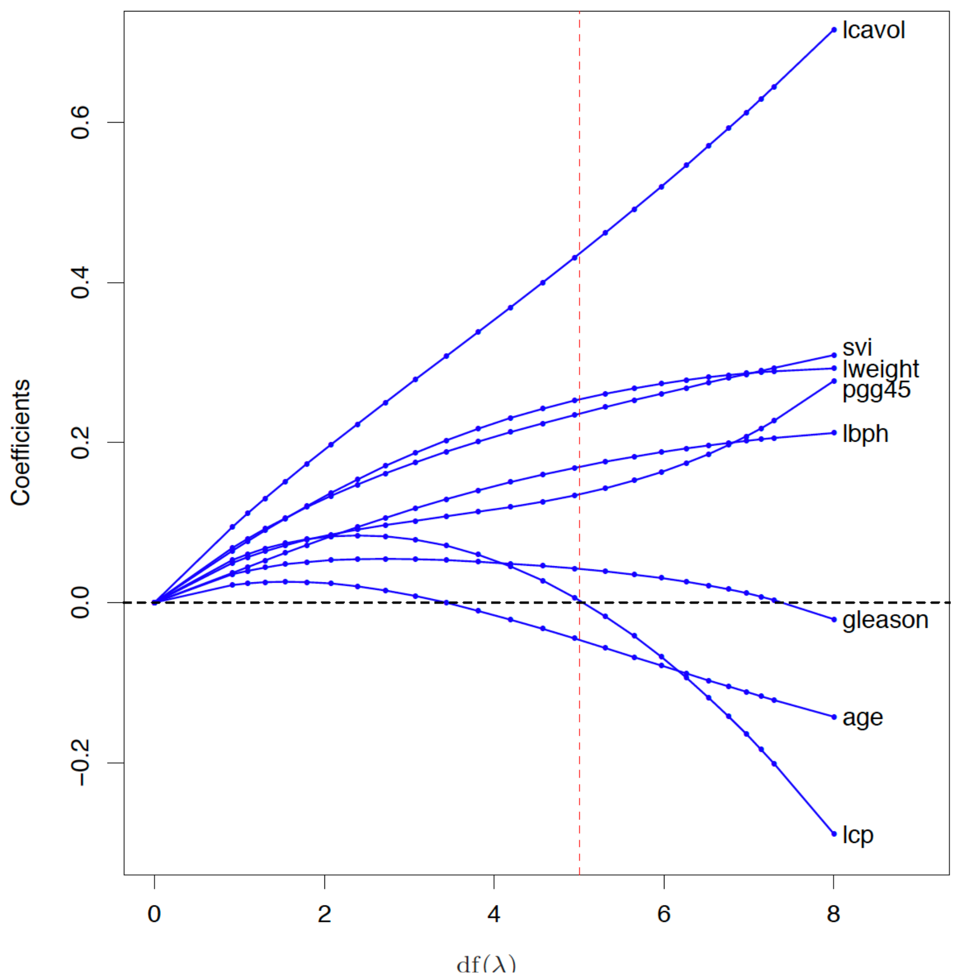


Fig 3.10 Lasso Regression

Lasso coefficients vs. the
standardized tuning parameter s

$$s = \frac{t}{\sum_{j=1}^{p} |\hat{\beta}_j|}$$

(prostate cancer example)

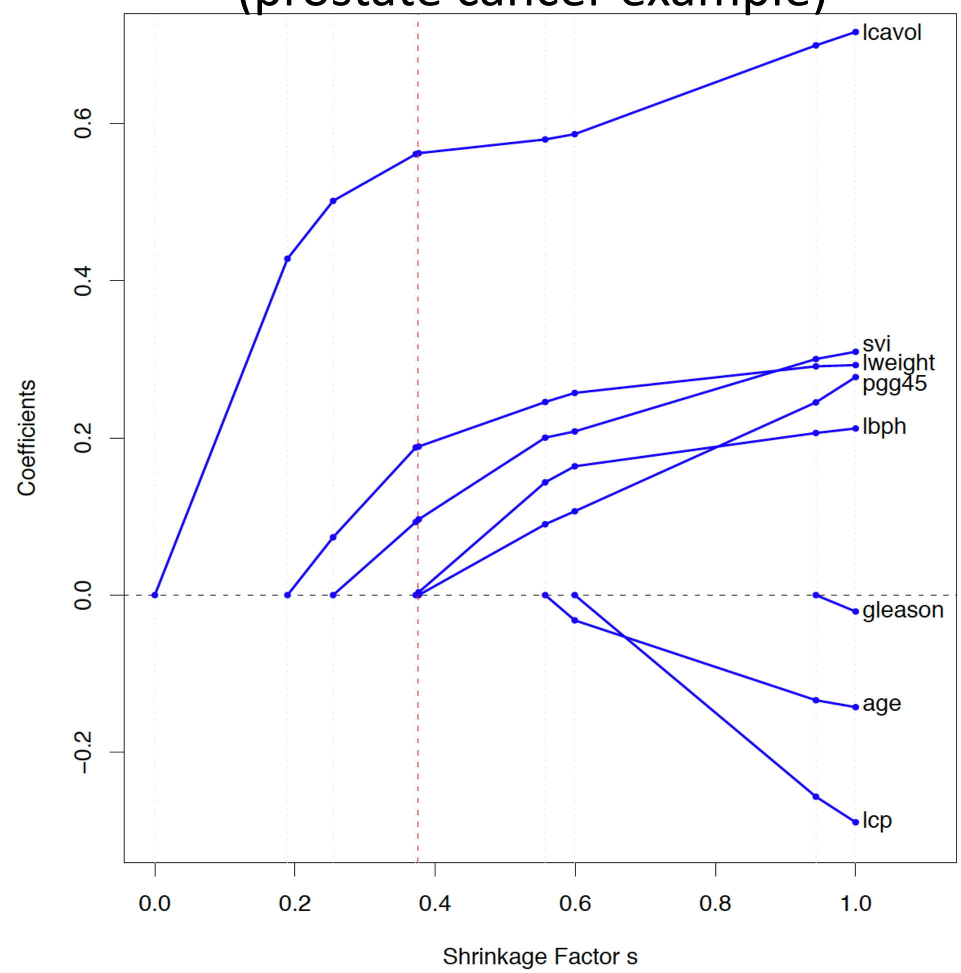


Table 3.4   With orthonormal input matrix X

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

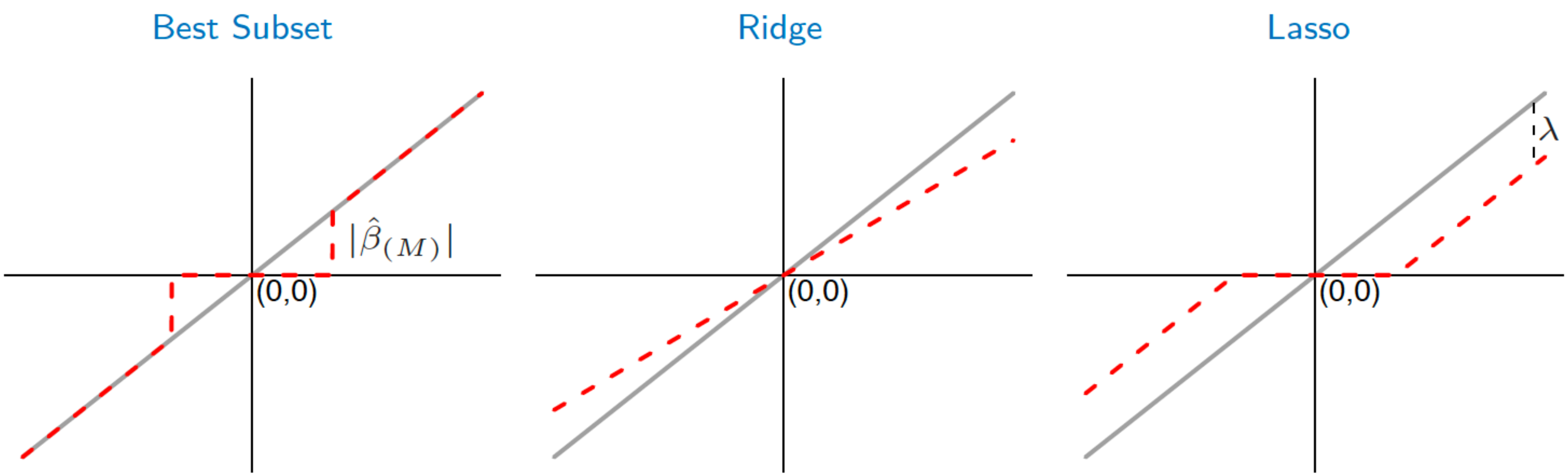
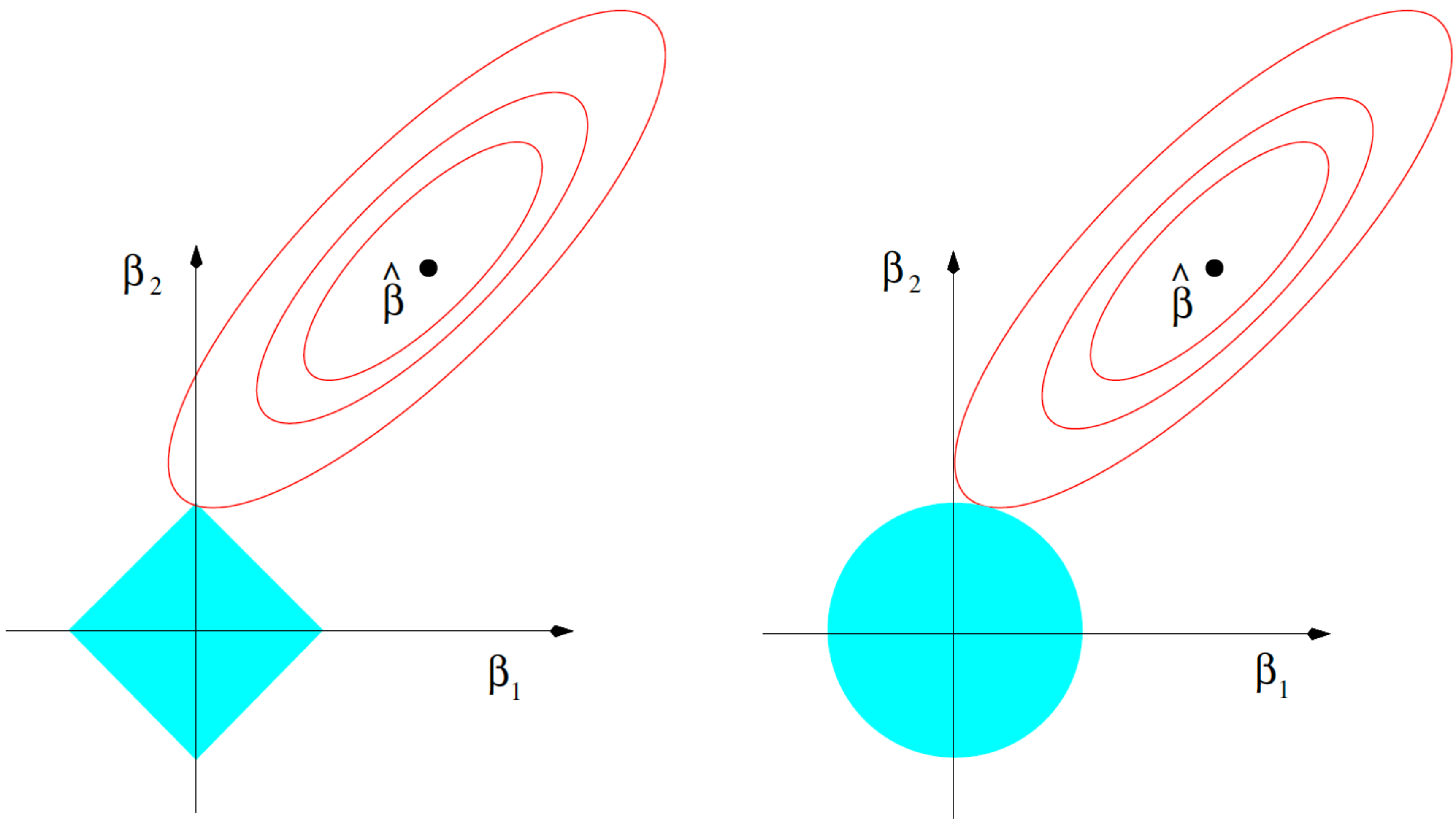


Fig 3.11 Lasso (left) and ridge (right)
Red: contours of RSS
Blue: constraint regions



Generalize lasso and ridge: Bayes estimates with different priors

q=0: subset selection
q=1: lasso (smallest q such that the constraint region is convex)
q=2: ridge

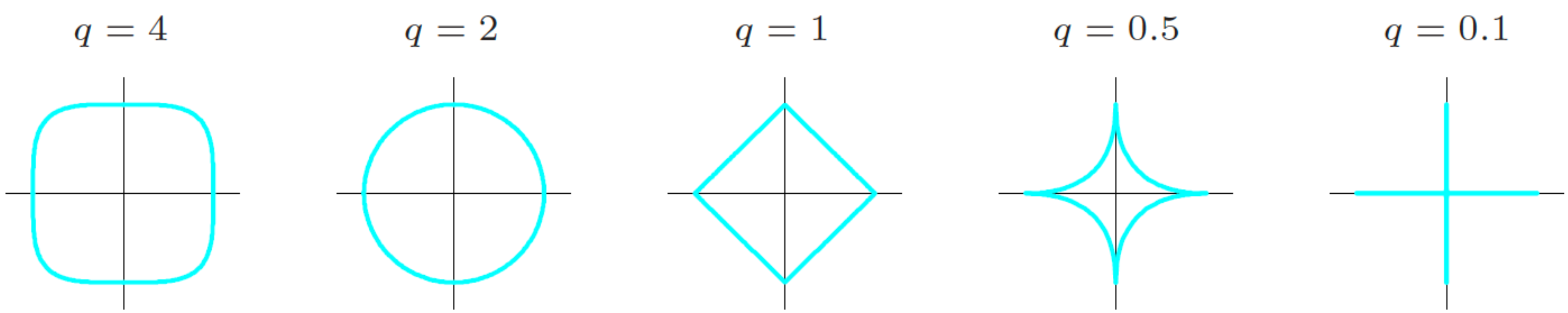


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.