Linear Methods for Regression

Hastie – Chap - 3;

Introduction

- A linear regression model assumes that the regression function $E(Y\ |X)$ is linear in the inputs $X_1,\ldots,X_p$.

- They are simple and often provide an adequate and interpretable description of how the inputs affect the output. For prediction purposes they can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data.

Linear Regression Models and Least Squares

- Purpose: - to predict a real-valued output Y. The linear regression model has the form.

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \,. \qquad (3.1)$$

- The linear model either assumes that the regression function $E(Y|X)$ is linear, or that the linear model is a reasonable approximation. Here the β_j's are unknown parameters or coefficients, and the variables X_j can come from different sources:

- We have a set of training data $(x_1, y_1) \ldots (x_N, y_N)$ from which to estimate the parameters . Each $x_i = (x_{i1}\ x_{i2}\ \ldots\ x_{ip})^T$ is a vector of feature measurements for the i^{th} case. The most popular estimation method is $least\ squares$, in which we pick the coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ to minimize the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

$$= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \qquad (3.2)$$

- From a statistical point of view, this criterion is reasonable if the training observations (x_i, y_i) represent independent random draws from their population. Even if the x_i's were not drawn randomly, the criterion is still valid if the y_i's are conditionally independent given the inputs x_i.

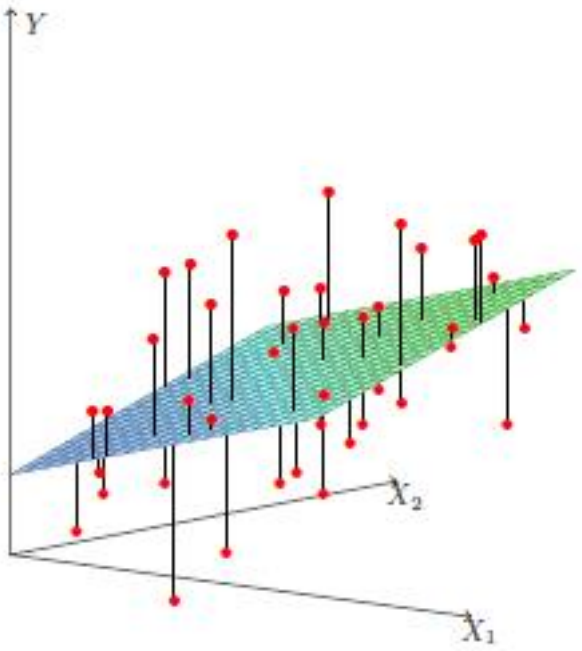


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Figure 3.1 illustrates the geometry of least-squares fitting in the (p+1)-dimensional space occupied by the pairs (X, Y).

- Figure 3.1 illustrates the geometry of least-squares fitting in the $\mathbb{R}^{p+1}$ −dimensional space occupied by the pairs (X, Y). Note that (3.2) makes no assumptions about the validity of model (3.1); it simply finds the best linear fit to the data. Least squares fitting is intuitively satisfying no matter how the data arise; the criterion measures the average lack of fit.

- How do we minimize (3.2)?

Denote by $\mathbf{X}$ the $N \times (p + 1)$ matrix with each row an input vector (with a 1 in the first position), and similarly let $\boldsymbol{y}$ be the N-vector of outputs in the training set. Then we can write the residual sum-of-squares as

$$RSS(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta). \qquad (3.3)$$

- This is a quadratic function in the p + 1 parameters. Differentiating with respect to we obtain

$$\frac{\partial RSS}{\partial \beta} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta)$$

$$\frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = \qquad (3.4)$$

- Assuming (for the moment) that $\boldsymbol{X}$ has full column rank, and hence $\boldsymbol{X}^T\boldsymbol{X}$ is positive definite, we set the first derivative to zero

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) = 0 \qquad (3.5)$$

- To obtain the unique solution

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \qquad (3.6)$$

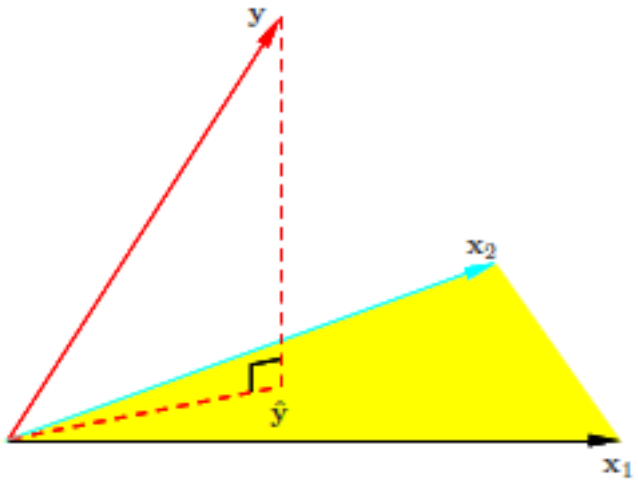


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector y is orthogonally projected onto the hyperplane spanned by the input vectors x_1 and x_2. The projection $\hat{y}$ represents the vector of the least squares predictions*

- The predicted values at an input vector x_0 are given by $\hat{f}(x_0) = (1 : x_0)^T \hat{\beta}$;the fitted values at the training inputs are

$$\hat{y} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y, \qquad (3.7)$$

where $\hat{y}_i = \hat{f}(x_i)$ .The matrix $\mathbf{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ appearing in equation (3.7) is sometimes called the "hat" matrix because it puts the hat on y.

- The hat matrix $\mathbf{H}$ computes the orthogonal projection, and hence it is also known as a projection matrix. It might happen that the columns of $\boldsymbol{X}$ are not linearly independent, so that $\mathbf{X}$ is not of full rank. for example, if two of the inputs were perfectly correlated, $(e.g., x_2 = 3\mathrm{x}_1)$.

- Then $\boldsymbol{X}^T\boldsymbol{X}$ is singular and the least squares coefficients $\hat{\beta}$ are not uniquely defined. However, the fitted values $\hat{y} = \boldsymbol{X}\hat{\beta}$ are still the projection of y onto the columns pace of $\boldsymbol{X}$; The non-full-rank case occurs most often when one or more qualitative inputs are coded in a redundant fashion.

- There is usually a natural way to resolve the non-unique representation, by recoding and/or dropping redundant columns in $\boldsymbol{X}$.

- Rank deficiencies can also occur in signal and image analysis, where the number of inputs p can exceed the number of training cases N. In this case, the features are typically reduced by filtering or else the fitting is controlled by regularization

- Assume that the observations y_i are uncorrelated and have constant variance σ^2, and that the x_i are fixed (non random). The variance–covariance matrix of the least squares parameter estimates is easily derived from (3.6) and is given by

$$Var(\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2. \qquad (3.8)$$

- Typically one estimates the variance σ^2 by.

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i{}^2). \qquad (3.8)$$

- The $N - p - 1$ rather than N in the denominator makes $\hat{\sigma}^2$ an unbiased estimate of σ^2: $E(\hat{\sigma}^2) = \hat{\sigma}^2$.

- The conditional expectation of Y is linear in $X_1,\ldots,X_p$. We also assume that the deviations of Y around its expectation are additive and Gaussian. Hence

$$Y = E(Y|X_1,\ldots,X_p) + \varepsilon$$

$$= \beta_0 + \sum_{j=1}^{p} X_j\beta_j + \varepsilon, \qquad (3.9)$$

where the error ε is a Gaussian random variable with expectation zero and variance σ^2, written $\varepsilon \sim \mathrm{N}(0,\sigma^2)$ .Under (3.9), it is easy to show that

$$\hat{\beta} \sim N(\beta, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2). \qquad (3.10)$$

- This is a multivariate normal distribution with mean vector and variance–covariance matrix as shown.

The Gauss–Markov Theorem

- One of the most famous results in statistics asserts that the **least squares estimates of the parameters $\boldsymbol{\beta}$ have the smallest variance among all linear unbiased estimates**.

- This observation will lead us to consider biased estimates such as ridge regression later. We focus on estimation of any linear combination of the parameters $\theta = a^T\beta$; for example, predictions $f(x_0) = x_0^T\beta$ are of this form.

- The least squares estimate of $a^T\beta$ is

$$\hat{\theta} = a^T\hat{\beta} = a^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \qquad (3.17)$$

- Considering $\boldsymbol{X}$ to be fixed, this is a linear function $\boldsymbol{c}_0^T\boldsymbol{y}$ of the response vector $\boldsymbol{y}$. If we assume that the linear model is correct, $\alpha^T\hat{\beta}$ is unbiased since

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

$$= a^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\beta$$

$$= a^T\beta. \qquad (3.18)$$

- The Gauss–Markov theorem states that if we have any other linear estimator $\bar{\theta} = \boldsymbol{c}^T\boldsymbol{y}$ that is unbiased for $a^T\beta$, that is, $E(\boldsymbol{c}^T\boldsymbol{y}) = a^T\beta$, then

$$Var(a^T\hat{\beta}) \leq Var(\boldsymbol{c}^T\boldsymbol{y}). \qquad (3.19)$$

- Consider the mean squared error of an estimator $\bar{\theta}$ in estimating θ:

$$MSE(\bar{\theta}) = E(\bar{\theta} - \theta)^2$$
$$= Var(\bar{\theta}) + [E(\bar{\theta}) - \theta]^2. \quad (3.20)$$

- The first term is the variance, while the second term is the squared bias. The Gauss-Markov theorem implies that the least squares estimator has the smallest mean squared error of all linear estimators with no bias.

Multiple Regression from Simple Univariate Regression

- The linear model (3.1) with $p > 1$ inputs is called the multiple linear regression model.

- Suppose first that we have a univariate model with no intercept, that is,

$$Y = X\beta + \varepsilon. \qquad (3.23)$$

- The least squares estimate and residuals are

$$\hat{\beta} = \frac{\sum_1^N x_i y_i}{\sum_1^N x_i^2},$$

$$r_i = y_i - x_i \hat{\beta}. \qquad (3.24)$$

Recollect from Method – 1 – LSQ:

$$m = \frac{N\sum_{i=1}^{N}(X_iY_i) - \sum_{i=1}^{N}X_i\sum_{i=1}^{N}Y_i}{DEN};$$

$$C = \frac{\sum_{i=1}^{N}Y_i\sum_{i=1}^{N}X_i^2 - \sum_{i=1}^{N}X_i\sum_{i=1}^{N}(X_iY_i)}{DEN};$$

where,

$$DEN = N\sum_{i=1}^{N}X_i^2 - (\sum_{i=1}^{N}X_i)^2$$

$$DEN' = N\sum_{i=1}^{N}X^2 ; C = 0;$$

$$\mathrm{m} = \frac{N\sum_{i=1}^{N}X_iY_i}{N\sum_{i=1}^{N}X^2} =;$$

- Convenient vector notation, we let $\mathrm{y} = (y_1, \ldots, y_N)^T$, $\mathrm{x} = (x_1, \ldots, x_N)^T$ and define

$$\langle \mathrm{x}, \mathrm{y} \rangle = \sum_{i=1}^{N} x_i y_i ,$$

$$= \mathrm{x}^T \mathrm{y}, \qquad (3.25)$$

- *Inner product* between x and y^1.Then we can write

$$\hat{\beta} = \frac{\langle \mathbf{x}, \boldsymbol{y} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} ,$$

$$r = \mathrm{y} - \mathrm{x}\hat{\beta}. \qquad (3.26)$$

- As we will see, this simple univariate regression provides the building block for multiple linear regression.

- Suppose next that the inputs $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_p$ (the columns of the data matrix $\boldsymbol{X}$) are orthogonal; that is $\langle \mathrm{x}_\mathrm{j}, \mathrm{x}_k \rangle = 0$ for all $j \neq k$. Then it is easy to check that the multiple least squares estimates $\hat{\beta}_j$ are equal to $\langle \mathrm{x}_j, \mathrm{y} \rangle / \langle \mathrm{x}_j, \mathrm{x}_j \rangle$—the univariate estimates. In other words, when the inputs are orthogonal, they have no effect on each other's parameter estimates in the model.

- Orthogonal inputs occur most often with balanced, designed experiments (where orthogonality is enforced), but almost never with observational data.

- Hence we will have to orthogonalize them. Suppose next that we have an intercept and a single input x. Then the least squares coefficient of x has the form

$$\hat{\beta}_1 = \frac{\langle \mathrm{x} - \bar{x}\mathbf{1}, y \rangle}{\langle \mathrm{x} - \bar{x}\mathbf{1}, \mathrm{x} - \bar{x}\mathbf{1} \rangle}, \qquad (3.27)$$

- where $\bar{x} = \sum_i x_i / N$, and $\mathbf{1} = \mathrm{x}_0$, the vector of N ones.
- We can view the estimate (3.27) as the result of two applications of the simple regression(3.26). The steps are:

1. Regress x on $\mathbf{1}$ to produce the residual $z = \mathrm{x} - \bar{x}\mathbf{1}$;
2. Regress y on the residual $\boldsymbol{z}$ to give the coefficient $\hat{\beta}_1$.

Linear Methods for Regression

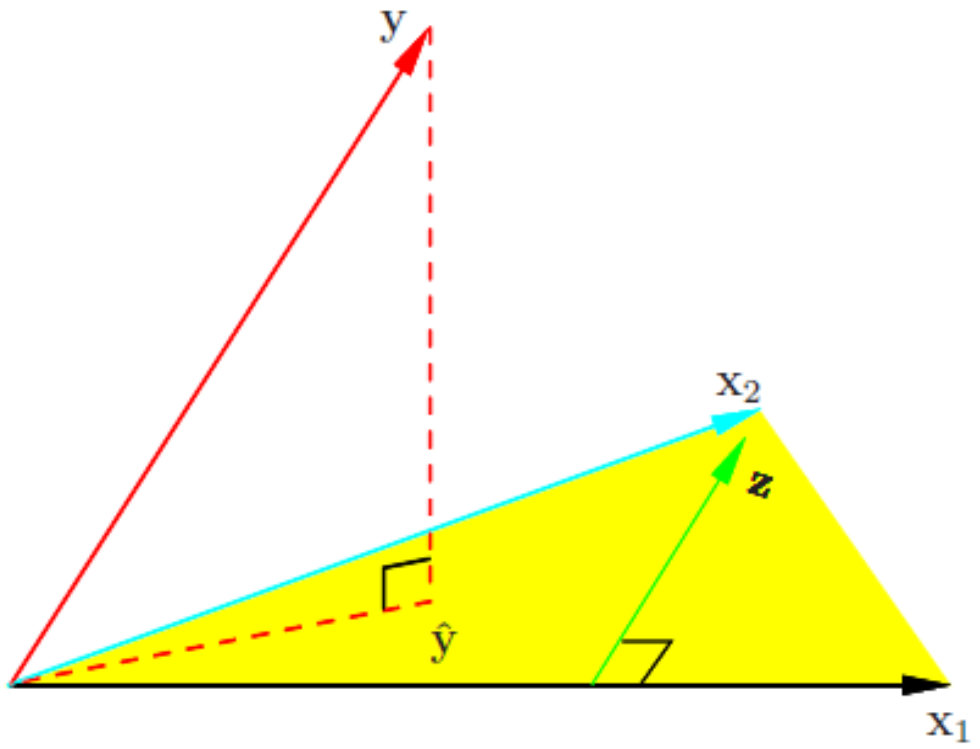


FIGURE 3.4. *Least squares regression by orthogonalization of the inputs. The vector* $\mathbf{x}_2$ *is regressed on the vector* $\mathbf{x}_1$*, leaving the residual vector* $\boldsymbol{z}$*. The regression of* $\boldsymbol{y}$ *on* $\boldsymbol{z}$ *gives the multiple regression coefficient of* $\mathbf{x}_2$*.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

- In this procedure, "regress b on a" means a simple univariate regression of b on a with no intercept, producing coefficient $\hat{\gamma} = \langle a, b \rangle / \langle a, a \rangle$ and residual vector $\mathbf{b} - \hat{\gamma}\mathbf{a}$. We say that $\boldsymbol{b}$ is adjusted for a, or is "orthogonalized" with respect to a.

- Step 1 orthogonalizes x with respect to $\mathrm{x}_0 = \mathbf{1}$. Step 2 is just a simple univariate regression, using the orthogonal predictors $\mathbf{1}$ and z. Figure 3.4 shows this process for two general inputs x_1 and x_2. The orthogonalization does not change the subspace spanned by x_1 and x_2, it simply produces an orthogonal basis for representing it.

- This recipe generalizes to the case of p inputs, as shown in Algorithm 3.1.Note that the inputs $\mathrm{z}_0, \ldots, \mathrm{z}_{\mathrm{j}-1}$ in step 2 are orthogonal, hence the simple regression coefficients computed there are in fact also the multiple regression coefficients.

Algorithm 3.1 *Regression by Successive Orthogonalization*.

1. Initialize $z_0 = x_0 = \mathbf{1}$.

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

   Regress x_j on $z_0, z_1, \ldots, , z_{j-1}$ to produce coefficients:

$$\hat{\gamma}_{\ell j} = \frac{\langle \mathbf{z}_\ell, \mathbf{x}_\ell \rangle}{\langle \mathbf{z}_\ell, \mathbf{z}_\ell \rangle}, \ell = 0, \ldots, j-1$$

   and residual vector

$$z_j = x_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} z_k .$$

3. Regress y on the residual z_p to give the estimate $\hat{\beta}_p$.

- The result of this algorithm is

$$\hat{\beta}_p = \frac{\langle \mathrm{zp},\ \mathrm{y} \rangle}{\langle \mathrm{z_p}, \mathrm{z_p} \rangle} \qquad (3.28)$$

- Re-arranging the residual in step 2, we can see that each of the $\mathbf{x}_j$ is a linear combination of the z_k, $k \leq j$.

- Since the $\boldsymbol{z}_j$ are all orthogonal, they form a basis for the column space of $\mathbf{X}$, and hence the least squares projection onto this subspace is $\hat{\mathrm{y}}$.

- Since z_p alone involves x_p (with coefficient 1), we see that the coefficient (3.28) is indeed the multiple regression coefficient of y on x_p.

- *The multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of x_j on $\boldsymbol{y}$, after x_j has been adjusted for* $\mathrm{x}_0, \mathrm{x}_1, \ldots, \mathrm{x}_{j-1}, \mathrm{x}_{j+1} \ldots, \mathrm{x}_p$.

- Algorithm 3.1 is known as the *Gram–Schmidt* procedure for multiple regression,. We can obtain from it not just $\hat{\beta}_p$, but also the entire multiple least squares fit,

- We can represent step 2 of Algorithm 3.1 in matrix form:

$$\mathbf{X} = \mathbf{Z}\mathbf{\Gamma}, \tag{3.30}$$

where $\mathbf{Z}$ has as columns the z_j (in order), and $\mathbf{\Gamma}$ is the upper triangular matrix with entries $\hat{\gamma}_{kj}$.

- Introducing the diagonal matrix $\mathbf{D}$ with jth diagonal entry $D_{jj} = \| \mathrm{z}_j \|$, we get

$$\mathbf{X} = \mathbf{Z}\mathbf{D}^{-1}\mathbf{D}\boldsymbol{\Gamma}$$

$$= \mathbf{Q}\mathbf{R}, \qquad (3.31)$$

the so-called QR decomposition of $\mathbf{X}$. Here $\mathbf{Q}$ is an $N \times (p+1)$ orthogonal matrix, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, and $\mathbf{R}$ is a $(p+1) \times (p+1)$ upper triangular matrix.

- The $\boldsymbol{QR}$ decomposition represents a convenient orthogonal basis for the column space of $\boldsymbol{X}$. It is easy to see, for example, that the least squares solution is given by

$$\hat{\beta} = \mathbf{R}^{-1}\mathbf{Q}^T\boldsymbol{y}, \qquad (3.32)$$

$$\hat{\boldsymbol{y}} = \mathbf{Q}\mathbf{Q}^T\boldsymbol{y}. \qquad (3.33)$$

- Equation (3.32) is easy to solve as $\boldsymbol{R}$ is upper triangular

Multiple Outputs

- Suppose we have multiple outputs $Y_1, Y_2, \ldots, Y_K$ that we wish to predict from our inputs $X_0, X_1, X_2, \ldots, X_p$. We assume a linear model for each output

$$Y_k = \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk} + \varepsilon_k \qquad (3.34)$$

$$= f_k(X) + \varepsilon_k. \qquad (3.35)$$

- With N training cases we can write the model in matrix notation

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}. \qquad (3.36)$$

- Here $\mathbf{Y}$ is the $N \times K$ response matrix, with ik entry y_{ik}, $\mathbf{X}$ is the $N \times (p+1)$ input matrix, $\mathbf{B}$ is the $(p+1) \times K$ matrix of parameters and $\mathbf{E}$ is the $N \times K$ matrix of errors.

- A straightforward generalization of the univariate loss function (3.2) is

$$RSS(\mathbf{B}) = \sum_{k=1}^{K}\sum_{i=1}^{N}(y_{ik} - f_k(x_i))^2 \qquad (3.37)$$

$$= tr[(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})]. \qquad (3.38)$$

- The least squares estimates have exactly the same form as before

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \qquad (3.39)$$

If the errors $\varepsilon = (\varepsilon_1, \ldots\ldots, \varepsilon_K)$ in (3.34) are correlated; if $Cov(\varepsilon) = \Sigma$, then the multivariate weighted criterion:

$$\mathrm{RSS}(\mathbf{B};\boldsymbol{\Sigma}) = \sum_{i=1}^{N}(y_i - f(x_i))^T\boldsymbol{\Sigma}^{-1}(y_i - f(x_i)) \qquad (3.40)$$

Shrinkage Methods

- By retaining a subset of the predictors and discarding the rest, subset selection produces a model that is interpretable and has possibly lower prediction error than the full model. However, because it is a discrete process—variables are either retained or discarded—it often exhibits high variance, and so doesn't reduce the prediction error of the full model. Shrinkage methods are more continuous, and don't suffer as much from high variability.

Ridge Regression

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

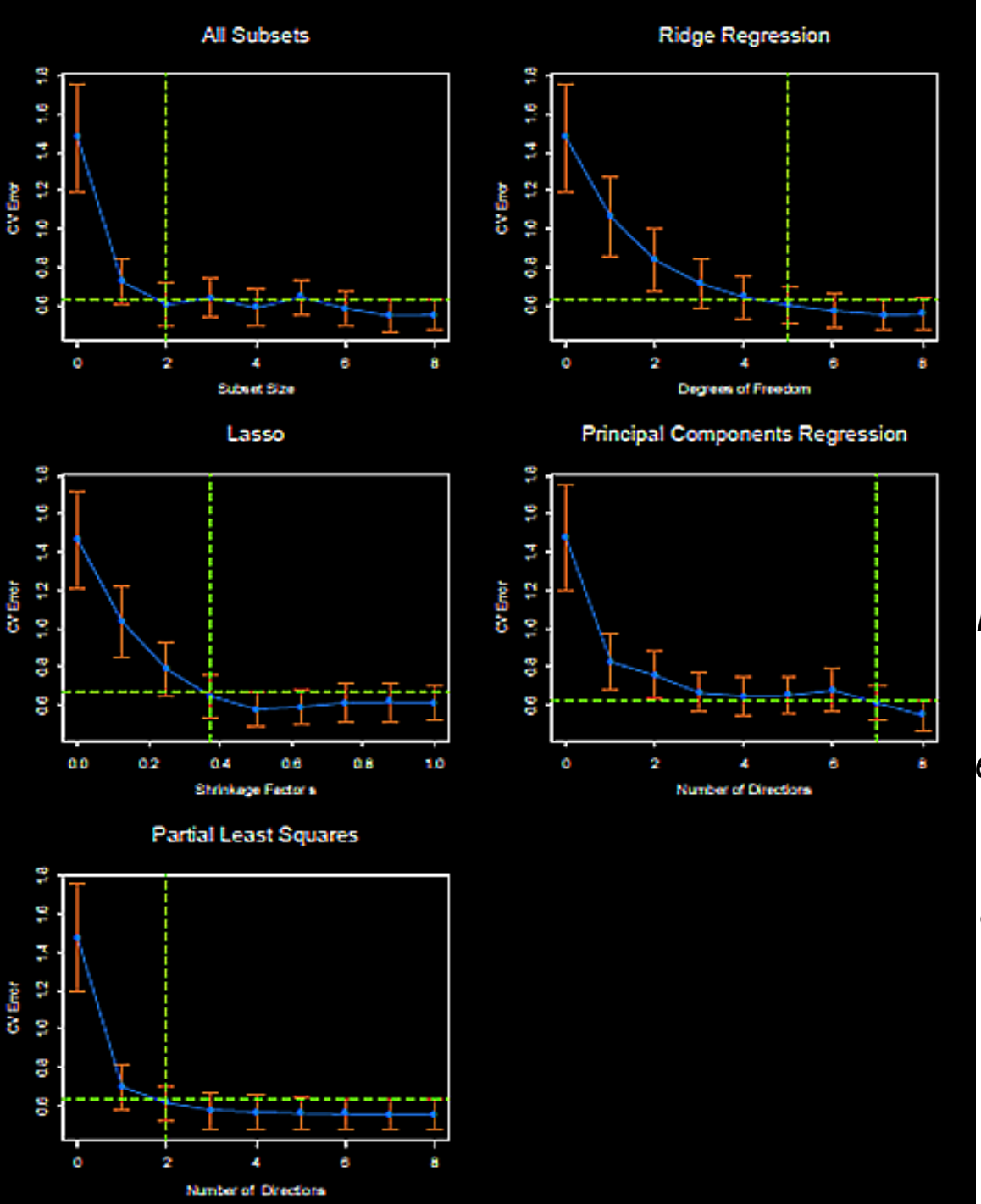


FIGURE 3.7. *Estimated prediction error curves and their standard errors for the various selection and shrinkage methods. Each curve is plotted as a function of the corresponding complexity parameter for that method. The horizontal axis has been chosen so that the model complexity increases as we move from left to right. The estimates of prediction error and their standard errors were obtained by tenfold cross-validation; full details are given in Section 7.10. The least complex model within one standard error of the best is chosen, indicated by the green vertical broken lines.*

- The ridge coefficients minimize a penalized residual sum of squares,

$$\hat{\beta}^{ridge} = argmin_{\beta}\left\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\right\}. \quad (3.41)$$

- Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. The coefficients are shrunk toward zero (and each other).

- An equivalent way to write the ridge problem is

$$\hat{\beta}^{ridge} = argmin_{\beta} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2, \quad (3.42)$$

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

- Which makes explicit the size constraint on the parameters. There is a one to-one correspondence between the parameters λ in (3.41) and t in (3.42). When there are many correlated variables in a linear regression model, their coefficients can become poorly determined and exhibit high variance. A wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin. By imposing a size constraint on the coefficients, as in (3.42), this problem is alleviated.

- The ridge solutions are not equivariant under scaling of the inputs, and so one normally standardizes the inputs before solving (3.41). The solution to (3.41) can be separated into two parts, after reparametrization using centered inputs: each x_{ij} gets replaced by $x_{ij} - \bar{x}_j$. We estimate β_0 by $\bar{y} = \frac{1}{N}\sum_1^N y_i$.

- The remaining coefficients get estimated by a ridge regression without intercept, using the centered x_{ij} . Henceforth we assume that this centering has been done, so that the input matrix $\mathbf{X}$ has p (rather than $p + 1$) columns.

- Writing the criterion in (3.41) in matrix form,

$$RSS(\lambda) = (\boldsymbol{y} - \mathbf{X}\beta)^T(\boldsymbol{y} - \mathbf{X}\beta) + \lambda\beta^T\beta, \qquad (3.43)$$

- The ridge regression solutions are easily seen to be

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\,\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y}, \qquad (3.44)$$

- Where $\mathbf{I}$ is the $p \times p$ identity matrix. Notice that with the choice of quadratic penalty $\beta^T\beta$, the ridge regression solution is again a linear function of y. The solution adds a positive constant to the diagonal of $\boldsymbol{X}^T\boldsymbol{X}$ before inversion.

- This makes the problem nonsingular, even if $\boldsymbol{X}^T\boldsymbol{X}$ is not of full rank, and was the main motivation for ridge regression when it was first introduced in statistics (Hoerl and Kennard, 1970).

- Traditional descriptions of ridge regression start with definition (3.44). We choose to motivate it via (3.41) and (3.42), as these provide insight into how it works.

- Ridge regression can also be derived as the mean or mode of a posterior distribution, with a suitably chosen prior distribution. In detail, suppose $y_i \sim N(\beta_0 + x_i^T\beta, \sigma^2)$, and the parameters β_j are each distributed as $N(0, \tau^2)$, independently of one another. Then the (negative) log-posterior density of β, with τ^2 and σ^2 assumed known, is equal to the expression in curly braces in (3.41), with $\lambda = \sigma^2/\tau^2$ Thus the ridge estimate is the mode of the posterior distribution; since the distribution is Gaussian, it is also the posterior mean.

- The $singular\ value\ decomposition\ (SVD)$ of the centered input matrix $\mathbf{X}$ gives us some additional insight into the nature of ridge regression. The SVD of the $N \times p$ matrix $\mathbf{X}$ has the form

$$\mathbf{X} = \mathbf{UDV}^T \qquad (3.45)$$

- Using the singular value decomposition we can write the least squares fitted vector as

$$\boldsymbol{X}\hat{\beta}^{ls} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{3.46}$$
$$= \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{y},$$

- After some simplification. Note that $\mathbf{U}^T\boldsymbol{y}$ are the coordinates of $\boldsymbol{y}$ with respect to the orthonormal basis $\mathbf{U}$. Note also the similarity with (3.33); $\mathbf{Q}$ and $\mathbf{U}$ are generally different orthogonal bases for the column space of $\mathbf{X}$ ($Exercise$ 3.8).
  Now the ridge solutions are

$$X\hat{\beta}^{ridge} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{3.47}$$
$$= \boldsymbol{U}\,\boldsymbol{D}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}\,\boldsymbol{U}^T\boldsymbol{y}$$
$$= \sum_{j=1}^{p} \boldsymbol{u}_j \frac{d_j^2}{d_j^2+\lambda}\,\boldsymbol{u}_j^T\boldsymbol{y},$$

- Where the $\boldsymbol{u}_j$ are the columns of $\mathbf{U}$. Note that since $\lambda \geq 0$, we have $d_j^2/(d_j^2 + \lambda) \leq 1$. Like linear regression, ridge regression computes the coordinates of y with respect to the orthonormal basis $\mathbf{U}$. It then shrinks these coordinates by the factors $d_j^2/(d_j^2 + \lambda)$ .This means that a greater amount of shrinkage is applied to the coordinates of basis vectors with smaller d_j^2.

- What does a small value of d_j^2 mean? The SVD of the centered $\mathbf{X}$ is another way of expressing the $principal\ components$ of the variables in $\mathbf{X}$. The sample covariance matrix is given by $\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X}/N$, and from (3.45) we have

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T, \qquad (3.48)$$

- Which is the $eigen\ decomposition$ of $\boldsymbol{X}^T\boldsymbol{X}$ (and of $\boldsymbol{S}$, up to a factor N).The eigenvectors v_j (columns of $\boldsymbol{V}$) are also called the $principal\ components$ (or Karhunen–Loeve) directions of $\mathbf{X}$. Sample variance is easily seen to be and in fact $\boldsymbol{z}_1 = \boldsymbol{X}v_1 = \boldsymbol{u}_1 d_1$

$$Var(\boldsymbol{z}_1) = Var(\boldsymbol{X}v_1) = \frac{d_1^2}{N}, \qquad (3.49)$$

- Subsequent principal components $\boldsymbol{z}_j$ have maximum variance d_j^2/N, subject to being orthogonal to the earlier ones. Conversely the last principal component has minimum variance. Hence the small singular values d_j correspond to directions in the column space of $\mathbf{X}$ having small variance, and ridge regression shrinks these directions the most.

- In Figure 3.7 we have plotted the estimated prediction error versus the quantity

$$df(\lambda) = tr[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T],$$

$$= tr(\mathbf{H}_\lambda)$$

$$= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda} . \qquad (3.50)$$

- This monotone decreasing function of λ is the $effective\ degrees\ of\ freedom$ of the ridge regression fit. Usually in a linear-regression fit with p variables, the degrees-of-freedom of the fit is p, the number of free parameters. The idea is that although all p coefficients in a ridge fit will be non-zero, they are fit in a restricted fashion controlled by λ . Note that $df(\lambda) = p$ when $\lambda = 0$ (no regularization) and $df(\lambda) \to 0$ as $\lambda \to \infty$.

The Lasso

- The lasso is a shrinkage method like ridge, with subtle but important differences. The lasso estimate is defined by

$$\hat{\beta}^{lasso} = argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$subject\ to\ \sum_{j=1}^{p} |\beta_j| . \qquad (3.51)$$

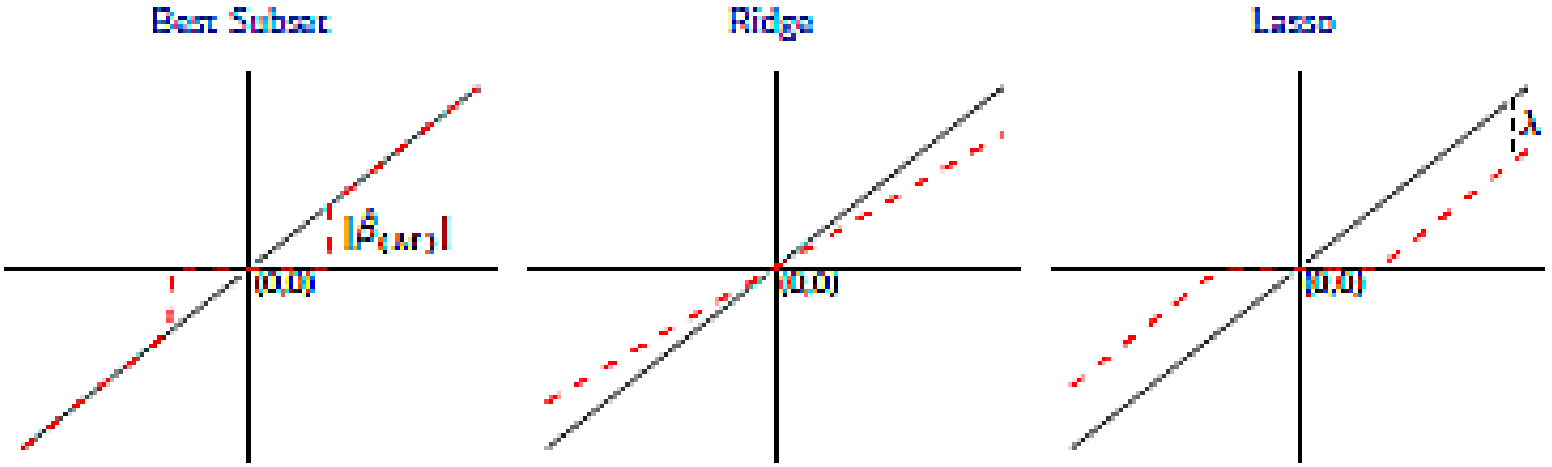
- Write the lasso problem in the equivalent Lagrangian form

$$\hat{\beta}^{lasso} = argmin_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \qquad (3.52)$$

- Notice the similarity to the ridge regression problem (3.42) or (3.41): the L_2 ridge penalty $\sum_1^p \beta_j^2$ is replaced by the L_1 lasso penalty $\sum_1^p |\beta_j|$ .Thus the lasso does a kind of continuous subset selection. If t is chosen larger than $t_0 = \sum_1^p |\hat{\beta}_j|$ (where $\hat{\beta}_j = \hat{\beta}_j^{ls}$, the least squares estimates), then the lasso estimates are the $\hat{\beta}_j$'s. On the other hand, for $t = t_0/2$ say, then the least squares coefficients are shrunk by about 50% on average.

TABLE 3.4. *Estimators of β_j in the case of orthonormal columns of $\mathbf{X}$. M and λ are constants chosen by the corresponding techniques; sign denotes the sign of its argument (± 1), and x_+ denotes "positive part" of x. Below the table, estimators are shown by broken red lines. The 45° line in gray shows the unrestricted estimate for reference.*

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	$\mathrm{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



Discussion: Subset Selection, Ridge Regression and the Lasso

- In the case of an orthonormal input matrix $\mathbf{X}$ the three procedures have explicit solutions. Each method applies a simple transformation to the least squares estimate $\hat{\beta}_j$, as detailed in Table 3.4.

- Ridge regression does a proportional shrinkage. Lasso translates each coefficient by a constant factor λ, truncating at zero. This is called "soft thresholding,". Best-subset selection drops all variables with coefficients smaller than the M^{th} largest; this is a form of "hard-thresholding."

- Back to the no orthogonal case; some pictures help understand their relationship. Figure 3.11 depicts the lasso $(left)$ and ridge regression $(right)$ when there are only two parameters. The residual sum of squares has elliptical contours, centered at the full least squares estimate.

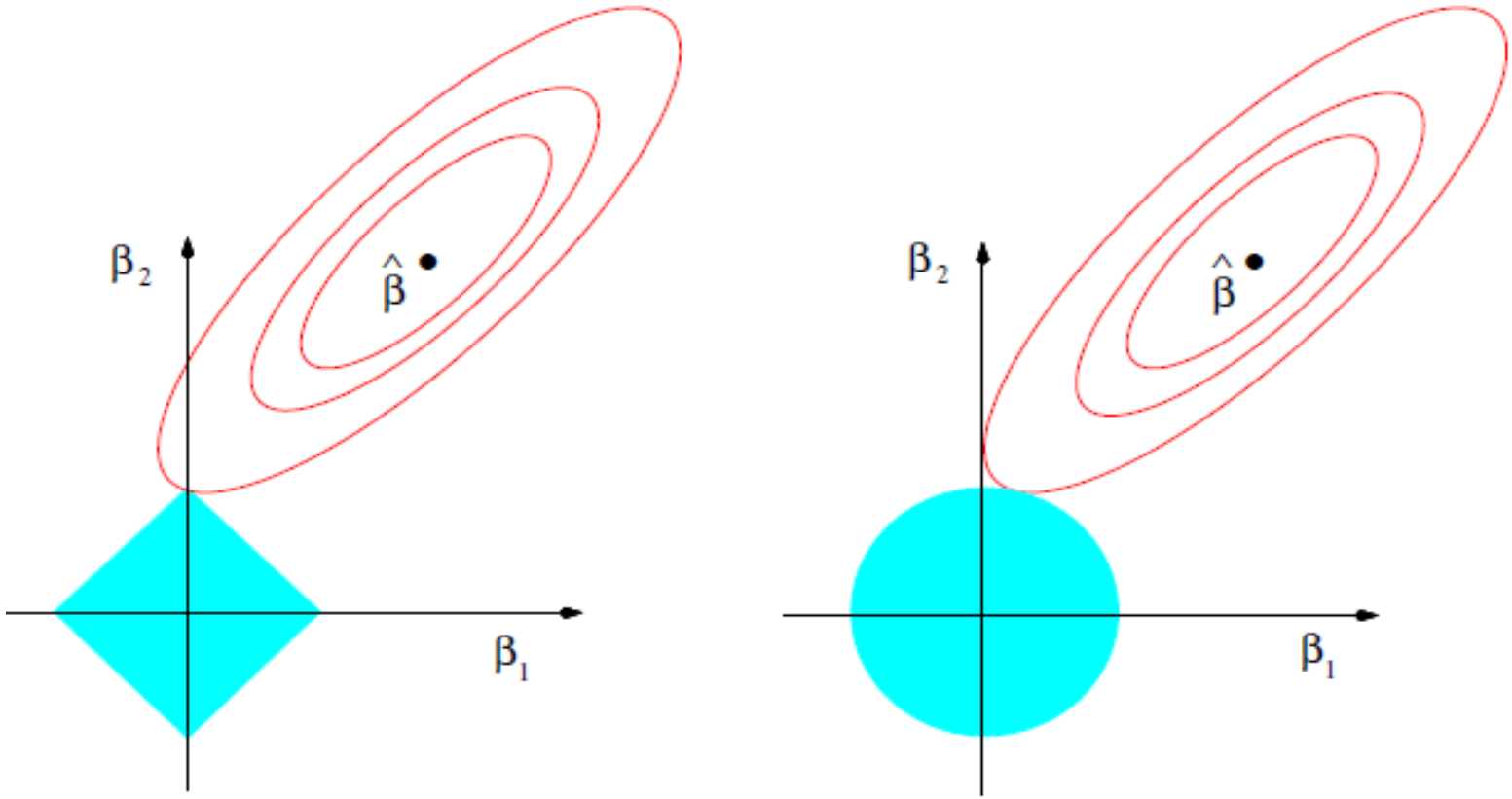


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

- Region for ridge regression is the disk $\beta_1^2 + \beta_2^2 \leq t$, while that for lasso is the diamond $|\beta_1| + |\beta_2| \leq t$. Both methods find the first point where the elliptical contours hit the constraint region. Unlike the disk, the diamond has corners; if the solution occurs at a corner, then it has one parameter β_j equal to zero. When $p > 2$, the diamond becomes a rhomboid, and has many corners, flat edges and faces; there are many more opportunities for the estimated parameters to be zero. Consider the criterion

- $\tilde{\beta} = argmin_\beta \left\{ \sum_{i=1}^{N} (yi - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}.$ (3.53)

- for $q \geq 0$. The contours of constant value of $\sum_j |\beta_j|^q$ are shown in Figure 3.12, for the case of two inputs.

- Thinking of $|\beta_j|^q$ as the log-prior density. The case $q = 1\ (lasso)$ is the smallest q such that the constraint region is convex; non-convex constraint regions make the optimization problem more difficult. In this view, the lasso, ridge regression and best subset selection are Bayes estimates with different priors. They are derived as posterior modes, that is, maximizers of the posterior.

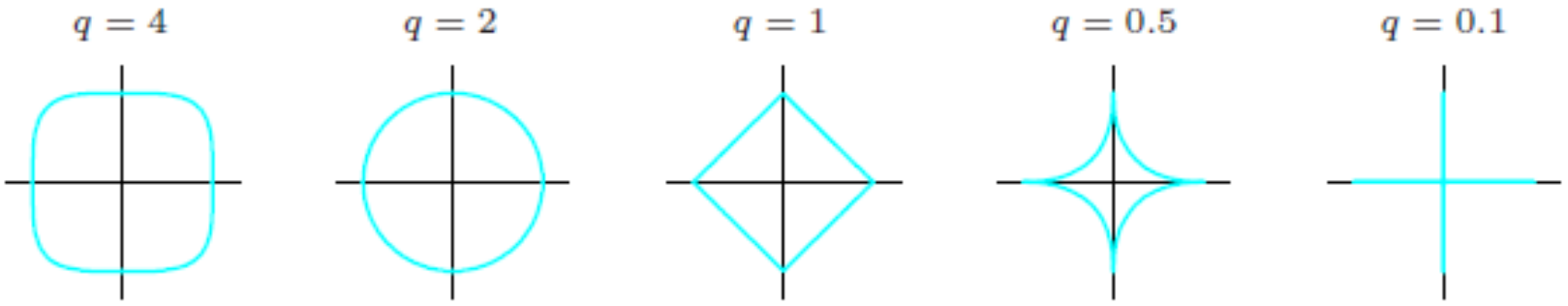
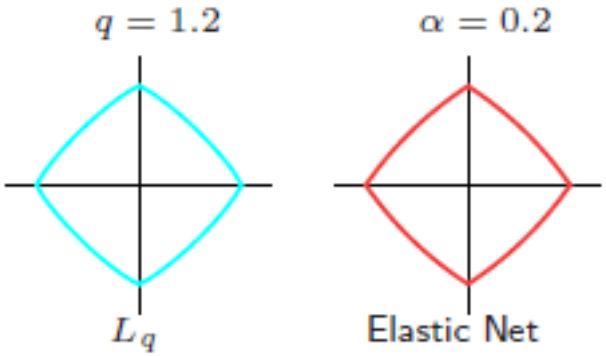


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.



- ***FIGURE 3.13.*** *Contours of constant value of* $\sum_j |\beta_j|^q$ *for* $q = 1.2$ *(left plot), and the elastic-net penalty* $\sum_j (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$ *(right plot).* Although visually very similar, the elastic

- *Zou and Hastie* (2005) *introduced the* $elastic\ net$ *penalty*

$$\lambda \sum_{j=1}^{p} (\alpha\beta_j^2 + (1-\alpha)|\beta_j|), \qquad (3.54)$$