Linear Methods for Regression
- Hastie – Chap – III

(part – B ; PRML – CS5691)

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.

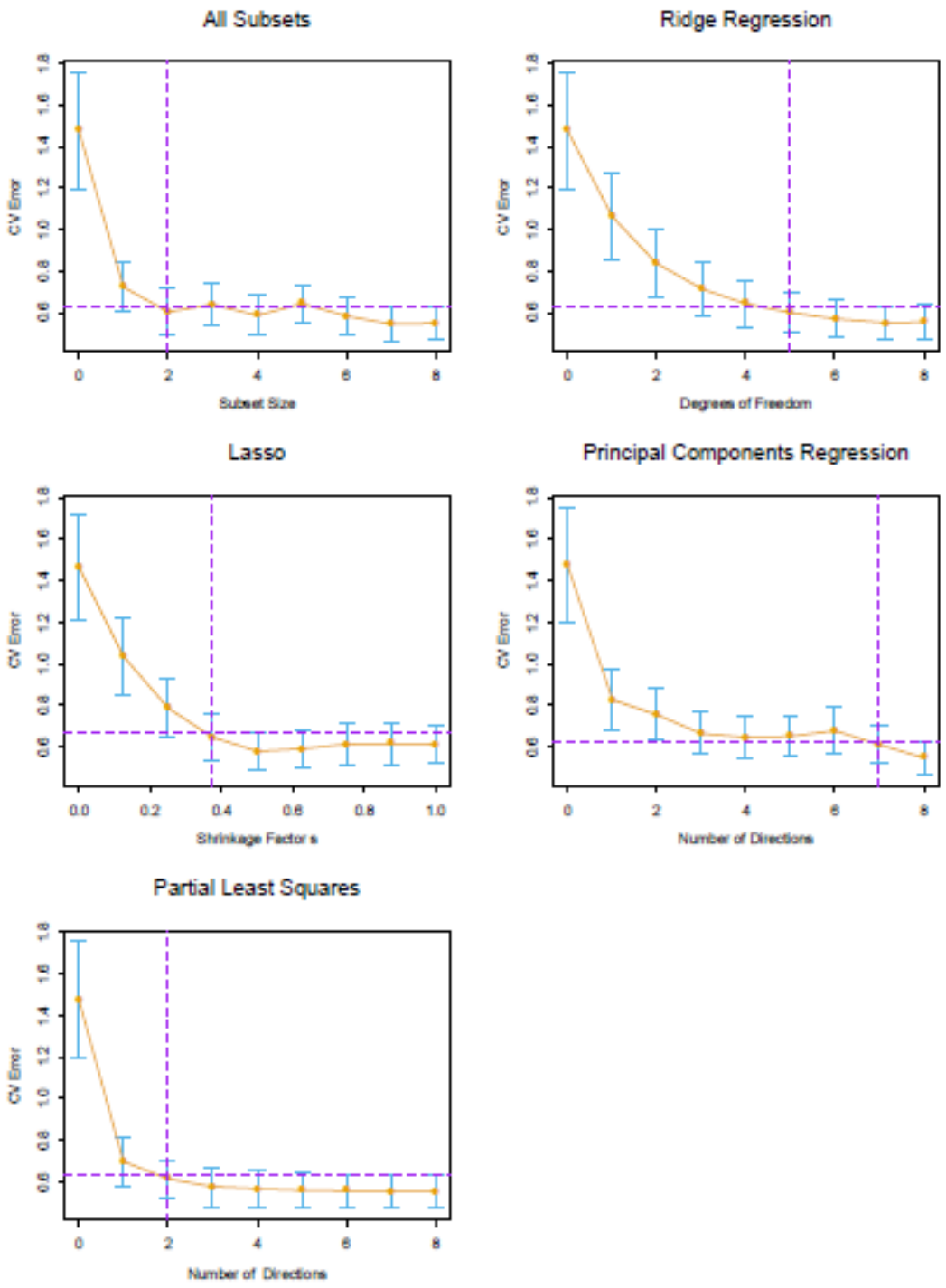


FIGURE 3.7. *Estimated prediction error curves and their standard errors for the various selection and shrinkage methods. Each curve is plotted as a function of the corresponding complexity parameter for that method. The horizontal axis has been chosen so that the model complexity increases as we move 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 purple 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 $\boldsymbol{\beta_0}$ by $\bar{y} = \frac{1}{N}\sum_1^N y_i$.

- Read about the process of whitening

- 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 (Ex – 3.6 - Hastie).

- 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

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

Stack the (centered) observations into the rows of an $N \times p$ matrix $\mathbf{X}$. We construct the *singular value decomposition* of $\mathbf{X}$:

Sec. 14.5 – PP 535

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T. \tag{14.54}$$

This is a standard decomposition in numerical analysis, and many algorithms exist for its computation (Golub and Van Loan, 1983, for example). Here $\mathbf{U}$ is an $N \times p$ orthogonal matrix ($\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$) whose columns $\mathbf{u}_j$ are called the *left singular vectors*; $\mathbf{V}$ is a $p \times p$ orthogonal matrix ($\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$) with columns v_j called the *right singular vectors*, and $\mathbf{D}$ is a $p \times p$ diagonal matrix, with diagonal elements $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ known as the *sin-*

Here $\mathbf{U}$ and $\mathbf{V}$ are $N \times p$ and $p \times p$ orthogonal matrices, with the columns of $\mathbf{U}$ spanning the column space of $\mathbf{X}$, and the columns of $\mathbf{V}$ spanning the row space. $\mathbf{D}$ is a $p \times p$ diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ called the singular values of $\mathbf{X}$. If one or more values $d_j = 0$, $\mathbf{X}$ is singular.

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

$$\begin{aligned} \mathbf{X}\hat{\beta}^{\text{ls}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{U}^T\mathbf{y}, \end{aligned} \tag{3.46}$$

Strang - Sec. 6.3 – PP 367

Singular Value Decomposition: Any m by n matrix A can be factored into

$$A = U\Sigma V^{\mathrm{T}} = (\textbf{orthogonal})(\textbf{diagonal})(\textbf{orthogonal}).$$

The columns of U (m by m) are eigenvectors of AA^{T}, and the columns of V (n by n) are eigenvectors of $A^{\mathrm{T}}A$. The r singular values on the diagonal of Σ (m by n) are the square roots of the nonzero eigenvalues of both AA^{T} and $A^{\mathrm{T}}A$.

Hastie - Sec. 18.3.5 – PP 659

Given the $N \times p$ data matrix $\mathbf{X}$, let

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \quad (18.12)$$

$$= \mathbf{R}\mathbf{V}^T \quad (18.13)$$

be the singular-value decomposition (SVD) of $\mathbf{X}$; that is, $\mathbf{V}$ is $p \times N$ with orthonormal columns, $\mathbf{U}$ is $N \times N$ orthogonal, and $\mathbf{D}$ a diagonal matrix with elements $d_1 \geq d_2 \geq d_N \geq 0$. The matrix $\mathbf{R}$ is $N \times N$, with rows r_i^T.

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

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

- Note that $\mathbf{U}^T\boldsymbol{y}$ are the coordinates of $\boldsymbol{y}$ with respect to the orthonormal basis $\mathbf{U}$.

Now the ridge solutions are

$$\begin{aligned} X\hat{\beta}^{ridge} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{y} \\ &= \end{aligned} \tag{3.47}$$

- 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 $\mathbf{X}$.

- The *first principal component* direction v_1 has the property that $z_1 = Xv_1$ has the largest sample variance amongst all normalized linear combinations of the columns of X. Sample variance is easily seen to be

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

and in fact

$$\boldsymbol{z}_1 = \boldsymbol{X}v_1 = \boldsymbol{u}_1 d_1$$

- 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 (DoF):

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

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

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

- 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.

Discussion: Subset Selection, Ridge Regression and the Lasso

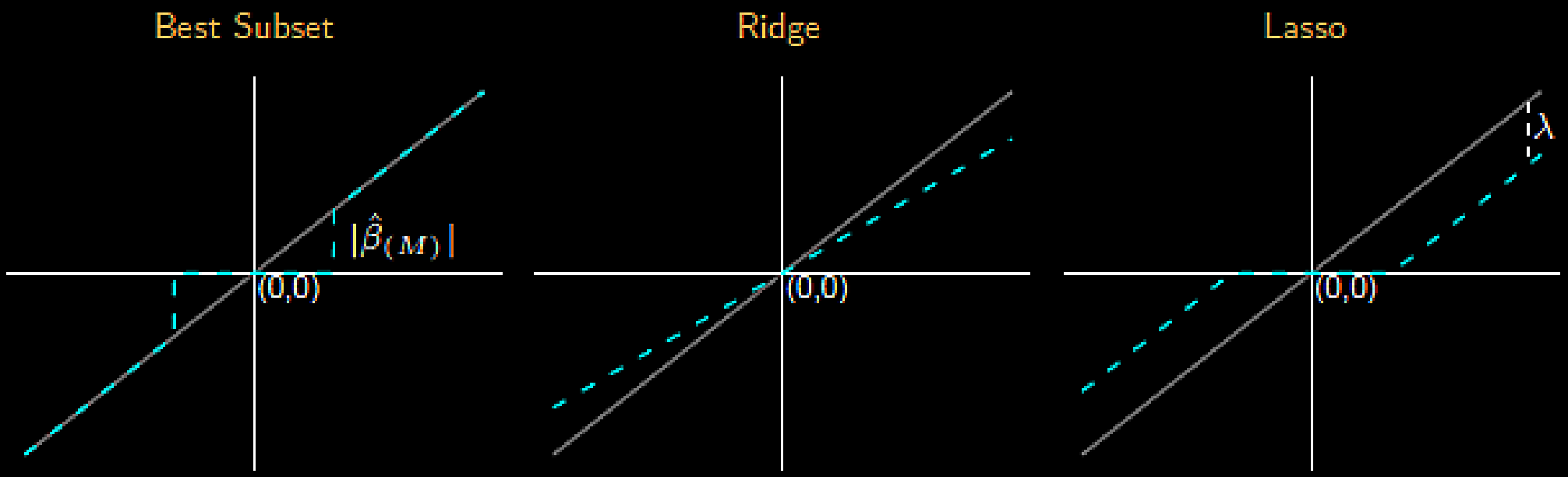
- In the case of an orthonormal input matrix $\mathbf{X}$ the three procedures have explicit solutions. Each method applies a simple transformation to the least squares estimate $\hat{\beta}_j$, as detailed in Table 3.4.

- 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.

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



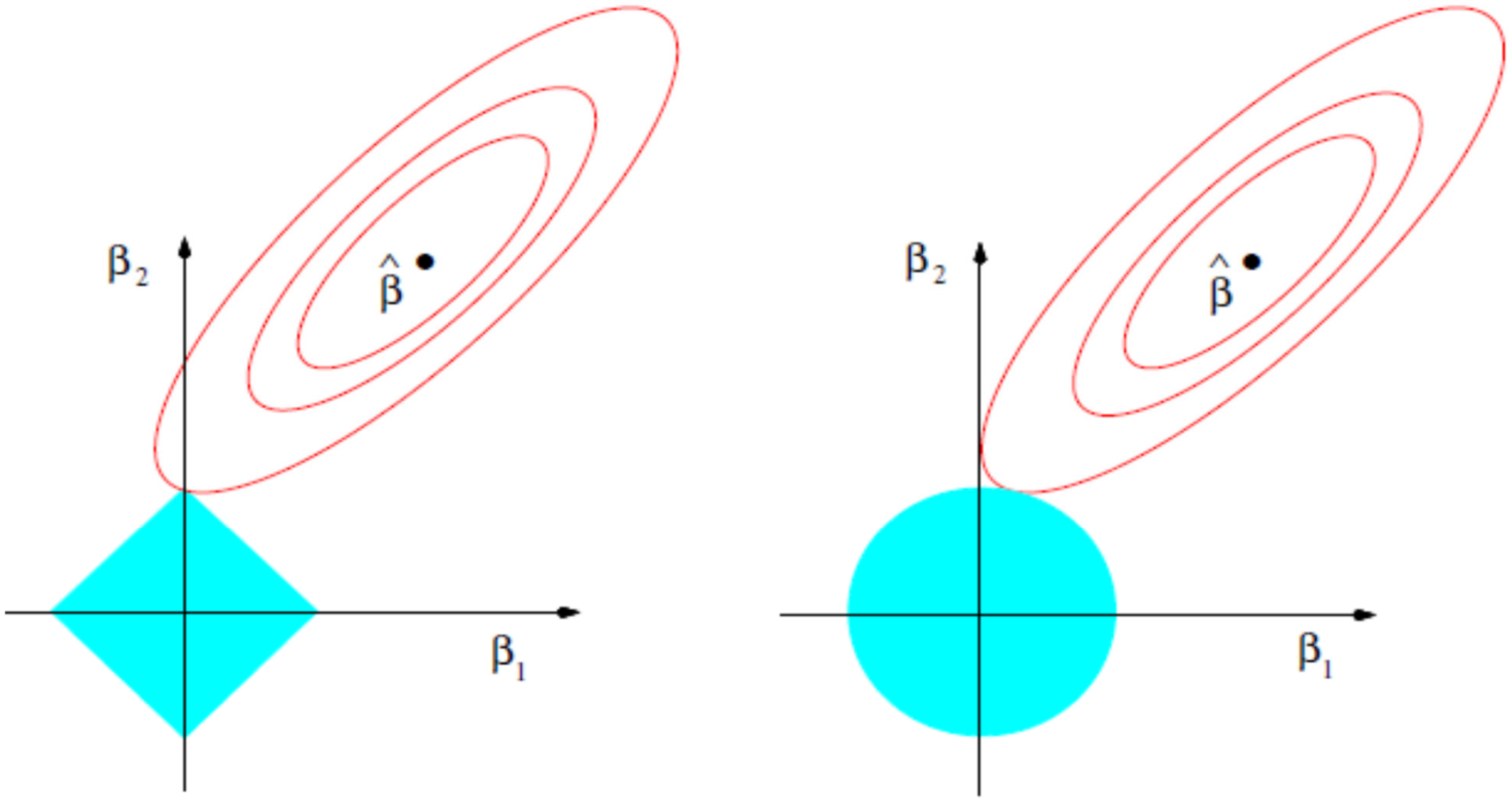


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are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

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

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.

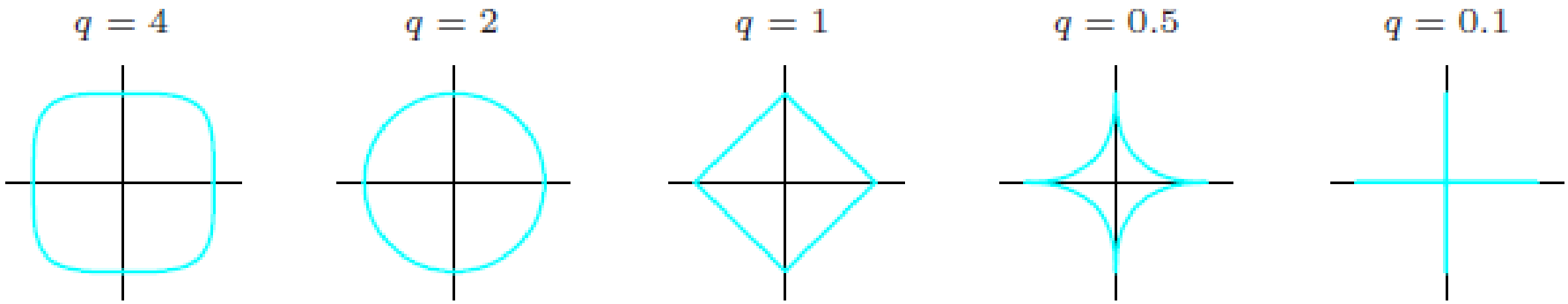
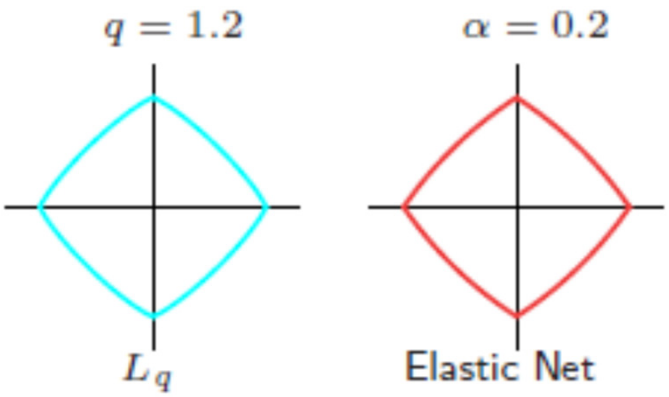


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.



- ***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*

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

The Grouped Lasso

- In some problems, the predictors belong to pre-defined groups; In this situation it may be desirable to shrink and select the members of a group together. The $grouped\ lasso$ is one way to achieve this. Suppose that the p predictors are divided into L groups, with p` the number in group ℓ. For ease of notation, we use a matrix X_ℓ to represent the predictors corresponding to the ℓth group, with corresponding coefficient vector β_ℓ.The grouped-lasso minimizes the convex criterion

- $$\min_{\beta \in R^p} \left(||y - \beta_0 \mathbf{1} - \sum_{\ell=1}^{L} \boldsymbol{X}_\ell \beta_\ell ||^2{}_2 + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell ||\beta_\ell||_2} \right), \qquad (3.80)$$

where the $\sqrt{p_\ell}$ terms accounts for the varying group sizes, and $|| \cdot ||_2$ is the Euclidean norm (not squared).

- Since the Euclidean norm of a vector β_ℓ is zero only if all of its components are zero, this procedure encourages sparsity at both the group and individual levels. That is, for some values of λ, an entire group of predictors may drop out of the model. This procedure was proposed by Bakin (1999) and Lin and Zhang (2006), and studied and generalized by Yuan and Lin (2007).

Further Properties of the Lasso

- A number of authors have studied the ability of the lasso and related procedures to recover the correct model, as N and p grow. Examples of this work include Knight and Fu (2000), Greenshtein and Ritov (2004), Tropp (2004), Donoho $(2006b)$, Meinshausen (2007), Meinshausen and B¨uhlmann (2006), Tropp (2006), Zhao and Yu (2006), Wainwright (2006), and Bunea et al. (2007).

Alternatively, one can modify the lasso penalty function so that larger coefficients are shrunken less severely; the *smoothly clipped absolute deviation* (SCAD) penalty of Fan and Li (2005) replaces $\lambda|\beta|$ by $J_a(\beta, \lambda)$, where

$$\frac{dJ_a(\beta,\lambda)}{d\beta} = \lambda \cdot \text{sign}(\beta)\Big[I(|\beta| \le \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda)\Big] \qquad (3.82)$$

for some $a \ge 2$. The second term in square-braces reduces the amount of shrinkage in the lasso for larger values of β, with ultimately no shrinkage as $a \to \infty$. Figure 3.20 shows the SCAD penalty, along with the lasso and

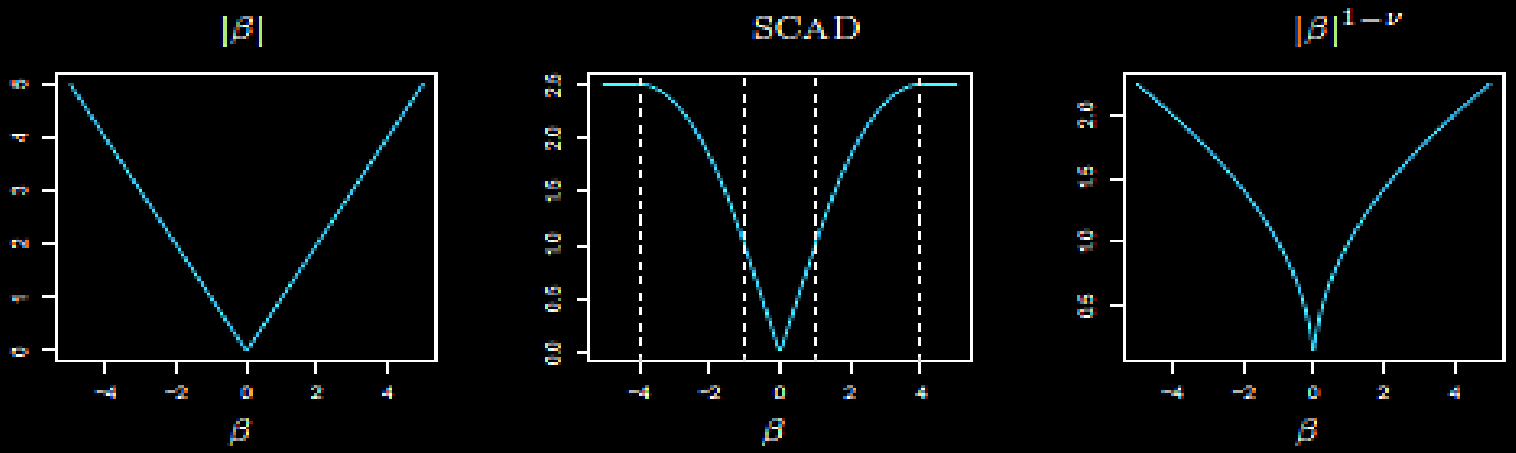


FIGURE 3.20. *The lasso and two alternative non-convex penalties designed to penalize large coefficients less. For SCAD we use* $\lambda = 1$ *and* $a = 4$*, and* $\nu = \frac{1}{2}$ *in the last panel.*

$|\beta|^{1-\nu}$. However this criterion is non-convex, which is a drawback since it makes the computation much more difficult. The *adaptive lasso* (Zou, 2006) uses a weighted penalty of the form $\sum_{j=1}^{p} w_j|\beta_j|$ where $w_j = 1/|\hat{\beta}_j|^\nu$, $\hat{\beta}_j$ is the ordinary least squares estimate and $\nu > 0$. This is a practical approximation to the $|\beta|^q$ penalties ($q = 1-\nu$ here) discussed in Section 3.4.3. The adaptive lasso yields consistent estimates of the parameters while retaining the attractive convexity property of the lasso.

Other Competing methods:

LAR

PLS

PCR

FSW

FS_0

Computational Considerations

- Least squares fitting is usually done via the Cholesky decomposition of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ or a $\boldsymbol{QR}$ decomposition of $\boldsymbol{X}$. With N observations and p features, the Cholesky decomposition requires $p^3 + Np^2/2$ operations, while the $\boldsymbol{QR}$ decomposition requires Np^2 operations. Depending on the relative size of N and p, the Cholesky can sometimes be faster; on the other hand, it can be less numerically stable (Lawson and Hansen, 1974). Computation of the lasso via the $\boldsymbol{LAR}$ algorithm has the same order of computation as a least squares fit.

Discussion: A Comparison of the Selection and Shrinkage Methods

- To summarize, $\boldsymbol{PLS}$, $\boldsymbol{PCR}$ and **ridge regression** tend to behave similarly.

- Ridge regression may be preferred because it shrinks smoothly, rather than in discrete steps. Lasso falls somewhere between ridge regression and best subset regression, and enjoys some of the properties of each.