Chapter 9
Variable Selection

Variable selection is very important in statistical modelling. We are frequently not only interested in using a model for prediction but also need to correctly identify the relevant variables, that is, to recover the correct model under given assumptions. It is known that under certain conditions, the ordinary least squares (OLS) method produces poor prediction results and does not yield a parsimonious model causing overfitting. Therefore the objective of the variable selection methods is to find the variables which are the most relevant for prediction. Such methods are particularly important when the true underlying model has a sparse representation (many parameters close to zero). The identification of relevant variables will reduce the noise and therefore improve the prediction performance of the fitted model.

Some popular regularisation methods used are the ridge regression, subset selection, L_1 norm penalisation and their modifications and combinations. Ridge regression, for instance, which minimises a penalised residual sum of squares using the squared L_2 norm penalty, is employed to improve the OLS estimate through a bias-variance trade-off. However, ridge regression has a drawback that it cannot yield a parsimonious model since it keeps all predictors in the model and therefore creates an interpretability problem. It also gives prediction errors close to those from the OLS model.

Another approach proposed for variable selection is the so-called "least absolute shrinkage and selection operator" (Lasso), aims at combining the features of ridge regression and subset selection either retaining (and shrinking) the coefficients or setting them to zero. This method received several extensions such as the Elastic net, a combination of Lasso and ridge regression or the Group Lasso used when predictors are divided into groups. This chapter describes the application of Lasso, Group Lasso as well as the Elastic net in linear regression model with continuous and binary response (logit model) variables.

W.K. Härdle, L. Simar, *Applied Multivariate Statistical Analysis*,
DOI 10.1007/978-3-662-45171-7_9

9.1 Lasso

Tibshirani (1996) first introduced Lasso for generalised linear models, where the response variable y is continuous rather than categorical. Lasso has two important characteristics. First, it has an L_1-penalty term which performs shrinkage on coefficients in a way similar to ridge regression, where an L_2 penalty is used.

Second, unlike ridge regression, Lasso performs variable subset selection driving some coefficients to exactly zero due to the nature of the constraint, where the objective function may touch the quadratic constraint area at a corner. For this reason, the Lasso is able to produce sparse solutions and is therefore able to combine good features of both ridge regression and subset selection procedure. It yields interpretable models and has the stability of ridge regression.

9.1.1 Lasso in the Linear Regression Model

The linear regression model can be written as follows:

$$y = \mathcal{X}\beta + \varepsilon,$$

where y is an $(n \times 1)$ vector of observations for the response variable, $\mathcal{X} = (x_1^\top, \dots, x_n^\top)^\top$, $x_i \in \mathbb{R}^p$, $i = 1, \dots, n$ is a data matrix of p explanatory variables, and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ is a vector of errors where $\mathsf{E}(\varepsilon_i) = 0$ and $\mathsf{Var}(\varepsilon_i) = \sigma^2$, $i = 1, \dots, n$.

In this framework, $\mathsf{E}\,(y|\mathcal{X}) = \mathcal{X}\beta$ with $\beta = \left(\beta_1, \dots, \beta_p\right)^\top$. Further assume that the columns of $\mathcal{X}$ are standardised such that $n^{-1}\sum_{i=1}^n x_{ij} = 0$ and $n^{-1}\sum_{i=1}^n x_{ij}^2 = 1$. The Lasso estimate $\hat{\beta}$ can then be defined as follows

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2 \right\}, \text{ subject to } \sum_{j=1}^{p} |\beta_j| \le s, \tag{9.1}$$

where $s \ge 0$ is the tuning parameter which controls the amount of shrinkage. For the OLS estimate $\hat{\beta}^0 = (\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top y$ a choice of tuning parameter $s < s_0$, where $s_0 = \sum_{j=1}^p |\hat{\beta}_j^0|$, will cause shrinkage of the solutions towards 0, and ultimately some coefficients may be exactly equal to 0. For values $s \ge s_0$ the Lasso coefficients are equal to the unpenalised OLS coefficients.

An alternative representation of (9.1) is:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{9.2}$$

with a tuning parameter $\lambda \geq 0$. As λ increases, the Lasso estimates are continuously shrunk toward zero. Then if λ is quite large, some coefficients are exactly zero. For $\lambda = 0$ the Lasso coefficients coincide with the OLS estimate. In fact, if the solution to (9.1) is denoted as $\hat{\beta}_s$ and the solution to (9.2) as $\hat{\beta}_\lambda$, then $\forall \lambda > 0$ and the resulting solution $\hat{\beta}_\lambda$ $\exists s_\lambda$ such that $\hat{\beta}_\lambda = \hat{\beta}_{s_\lambda}$ and vice versa which implies a one-to-one correspondence between these parameters. However, this does not hold if it is required that $\lambda \geq 0$ only and not $\lambda > 0$, because if, for instance, $\lambda = 0$, then $\hat{\beta}_\lambda$ is the same for any $s \geq \|\hat{\beta}\|_1$ and the correspondence is no longer one-to-one.

Geometrical Aspects in $\mathbb{R}^2$

The Lasso estimate under the least squares loss function solves a quadratic programming problem with linear inequality constraints. The criterion $\sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2$ yields the quadratic form objective function

$$(\beta - \hat{\beta}^0)^\top \mathcal{W}(\beta - \hat{\beta}^0) \tag{9.3}$$

with $\mathcal{W} = \mathcal{X}^\top \mathcal{X}$. For the special case when $p = 2$, $\beta = (\beta_1, \beta_2)^\top$, the resulting elliptical contour lines are centred around the OLS estimate and the linear constraints are represented by square (shaded area) shown in Fig. 9.1. The Lasso solution is the first place that the contours touch the square, and this sometimes occurs at a corner, corresponding to a zero coefficient. The nature of the Lasso shrinkage may not occur completely obvious. In the work by Efron, Hastie, Johnstone, and Tibshirani (2004) the Least Angle Regression (LAR) algorithm

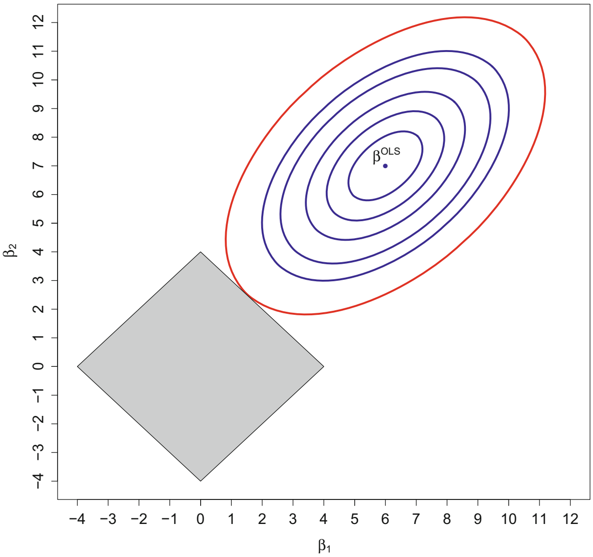

Fig. 9.1 Lasso in the general design case for $s = 4$ and OLS estimate $\hat{\beta}^0 = (6, 7)^\top$ MVAlassocontour

with a Lasso modification was described which computes the whole path of Lasso solutions and gives a better understanding of the shrinkage nature.

The LAR Algorithm and Lasso Solution Paths

The LAR algorithm may be introduced in the simple three-dimensional case as follows (assume that the number of covariates $p = 3$):

- first, standardise all the covariates to have mean 0 and unit length as well as make the response variable have mean zero;
- start with $\hat{\beta} = 0$;
- initialise the algorithm with the first two covariates: let $\mathcal{X} = (x_1, x_2)$ and calculate the prediction vector $\hat{y}_0 = \mathcal{X}\hat{\beta} = 0$;
- calculate $\overline{y}_2$ the projection of y onto $\mathcal{L}(x_1, x_2)$, the linear space spanned by x_1 and x_2;
- compute the vector of current correlations between the covariates $\mathcal{X}$ and the two-dimensional current residual vector: $C^{\hat{y}_0} = \mathcal{X}^\top(\overline{y}_2 - \hat{y}_0) = (c_1^{\hat{y}_0}, c_2^{\hat{y}_0})^\top$. According to Fig. 9.2, the current residual $\overline{y}_2 - \hat{y}_0$ makes a smaller angle with x_1, than with x_2, therefore $c_1^{\hat{y}_0} > c_2^{\hat{y}_0}$;
- augment $\hat{y}_0$ in the direction of x_1 so that $\hat{y}_1 = \hat{y}_0 + \hat{\gamma}_1 x_1$ with $\hat{\gamma}_1$ chosen such that $c_1^{\hat{y}_0} = c_2^{\hat{y}_0}$ which means that the new current residual $\overline{y}_2 - \hat{y}_1$ makes equal angles (is equiangular) with x_1 and x_2;
- suppose that another regressor x_3 enters the model: calculate a new projection $\overline{y}_3$ of y onto $\mathcal{L}(x_1, x_2, x_3)$;
- recompute the current correlations vector $C^{\hat{y}_1} = (c_1^{\hat{y}_1}, c_2^{\hat{y}_1}, c_3^{\hat{y}_1})^\top$ with $\mathcal{X} = (x_1, x_2, x_3)$, $\overline{y}_3$ and $\hat{y}_1$;
- augment $\hat{y}_1$ in the equiangular direction so that $\hat{y}_2 = \hat{y}_1 + \hat{\gamma}_2 u_2$ with $\hat{\gamma}_2$ chosen such that $c_1^{\hat{y}_1} = c_2^{\hat{y}_1} = c_3^{\hat{y}_1}$, then the new current residual $\overline{y}_3 - \hat{y}_2$ goes

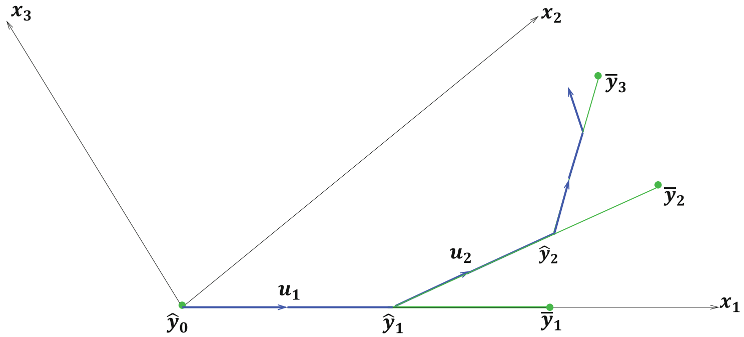

Fig. 9.2 Illustration of LARS algorithm

equiangularly between x_1, x_2 and x_3 (here u_2 is the unit vector lying along the equiangular direction $\hat{y}_2$);

- the three-dimensional algorithm is terminated with the calculation of the final prediction vector $\hat{y}_3 = \hat{y}_2 + \hat{\gamma}_3 u_3$ with $\hat{\gamma}_3$ chosen such that $\hat{y}_3 = \overline{y}_3$.

In the case of $p > 3$ covariates, $\hat{\gamma}_3$ would be smaller than $\overline{y}_3$ initiating another change of direction, as illustrated in Fig. 9.2.

In this setup, it is important that the covariate vectors x_1, x_2, x_3 are linearly independent. The LAR algorithm "moves" the variable coefficients to their least squares values. So the Lasso adjustment necessary for the sparse solution is that if a nonzero coefficient happens to return to zero, it should be dropped from the current ("active") set of variables and not be considered in further computations. The general LAR algorithm for p predictors can be summarised as follows.

Least Angle Regression Algorithm

1. The covariates are standardised to have mean 0 and unit length 1 and the response has mean 0:

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1; \quad j = 1, 2, \dots, p.$$

   The task is to construct the fit $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^\top$ by iteratively changing the prediction vector $\hat{y} = \sum_{j=1}^{p} x_j \hat{\beta}_j = \mathcal{X}\hat{\beta}$.

2. Denote $\mathcal{A}$ equal to a subset of the indices $\{1, 2, \dots, p\}$, begin with $\hat{y}_\mathcal{A} = \hat{y}_o = 0$ and calculate the vector of current correlations

$$\hat{c} = \mathcal{X}^\top (y - \hat{y}_\mathcal{A}).$$

3. Then review the current set $\mathcal{A} = \{j : |\hat{c}_j| = \hat{C}\}$ as the set of indices corresponding to the covariates with the greatest absolute current correlations, where $\hat{C} = \max_j\{|\hat{c}_j|\}$; let $s_j = \operatorname{sign}\{\hat{c}_j\}$ for $j \in \mathcal{A}$ and compute the matrix $\mathcal{X}_\mathcal{A} = (s_j x_j)_{j\in\mathcal{A}}$, the scalar $A_\mathcal{A} = (1_\mathcal{A}^\top \mathcal{G}_\mathcal{A}^{-1} 1_\mathcal{A})^{-\frac{1}{2}}$ with $\mathcal{G}_\mathcal{A} = \mathcal{X}_\mathcal{A}^\top \mathcal{X}_\mathcal{A}$ and $1_\mathcal{A}^\top$ being a vector of ones of length $|\mathcal{A}|$, and the so-called equiangular vector $u_\mathcal{A} = \mathcal{X}_\mathcal{A} w_\mathcal{A}$ with $w_\mathcal{A} = A_\mathcal{A} \mathcal{G}_\mathcal{A}^{-1} 1_\mathcal{A}$ which makes equal angles, each less than 90°, with the columns of $\mathcal{X}_\mathcal{A}$.

4. Calculate the inner product vector $a \overset{\text{def}}{=} \mathcal{X}^\top u_\mathcal{A}$ and the direction

$$\hat{\gamma} = \min_{j\in\mathcal{A}^c}{}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{A_\mathcal{A} - a_j}, \frac{\hat{C} + \hat{c}_j}{A_\mathcal{A} + a_j} \right\}$$

5. Define $\hat{d}$ to be the m-vector equaling $s_j w_{\mathcal{A}_j}$ for $j \in \mathcal{A}$ and zero elsewhere and $\gamma_j = -\hat{\beta}_j / \hat{d}_j$ yielding $\tilde{\gamma} = \min_{\gamma_j > 0} \{\gamma_j\}$
    (a) If $\tilde{\gamma} < \hat{\gamma}$, calculate the next LARS step as

    $$\hat{y}_{\mathcal{A}_+} = \hat{y}_{\mathcal{A}} + \tilde{\gamma} u_{\mathcal{A}}$$

    with $\mathcal{A}_+ = \mathcal{A} - \{\tilde{j}\}$.
    (b) Else: calculate the next step as

    $$\hat{y}_{\mathcal{A}_+} = \hat{y}_{\mathcal{A}} + \hat{\gamma} u_{\mathcal{A}}$$

6. Iterate until all p predictors have been entered, some of which are ultimately dropped from the active set $\mathcal{A}$.

This algorithm can be implemented on a grid from 0 to 1 of the standardised coefficients constraint s resulting in the complete paths of the Lasso coefficients and illustrating the nature of Lasso shrinkage.

Once the Lasso solution paths have been obtained, it is important to decide on a rule how to choose the "optimal" solution, or, equally, the regularisation parameter λ. There are several existing methods to do this and the most popular examples are the K-fold cross-validation, generalised cross-validation, Schwartz's (Bayesian) Information Criterion (BIC). All these methods can be viewed as degrees-of-freedom adjustments to the residual squared error (RSE) which underestimates the true prediction error

$$\mathrm{RSE} \overset{\mathrm{def}}{=} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Consider the generalised cross-validation statistic:

$$\mathrm{GCV}(\lambda) = n^{-1}\mathrm{RSE}_\lambda / \{1 - \mathrm{df}(\lambda)/n\}^2, \tag{9.4}$$

where RSE_λ is the residual sum of squares for the constrained fit with a particular regularisation parameter λ. An alternative is the BIC

$$\mathrm{BIC} = n \log(\hat{\sigma}^2) + \log(n) \cdot \mathrm{df}(\lambda) \tag{9.5}$$

with the estimation of error variance $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

The degrees of freedom of the predicted vector $\hat{y}$ in the Lasso problem with the linear Gaussian model with normally distributed errors having zero expectation and variance σ^2, written $\varepsilon_i \sim N(0, \sigma^2)$, can be defined as follows:

$$\mathrm{df}(\lambda) \overset{\mathrm{def}}{=} \sigma^{-2} \sum_{i=1}^{n} \mathsf{Cov}(\hat{y}_i, y_i), \tag{9.6}$$

which can actually be used for both linear and non-linear models. This expression for $\mathrm{df}(\lambda)$ can be viewed as a quantitative measure of the prediction error bias dependence on how much each y_i affects its fitted value $\hat{y}_i$. The estimate $\hat{\beta}$ minimising the GCV statistic can then be chosen. The following example shows how to compute $\mathrm{df}(\lambda)$.

Example 9.1 (Calculation of $\mathrm{df}(\lambda)$*)* As no closed-form solution exists for the Lasso problem, an approximation should be calculated. The constraint $\sum |\beta_j| \leq s$ can be rewritten as $\sum \beta_j^2/|\beta_j| \leq s$. Using the duality between the constrained and unconstrained problems and one-to-one correspondence between s and λ, the Lasso solution is computed as the ridge regression estimate

$$\hat{\beta} = (\mathcal{X}^\top \mathcal{X} + \lambda B^{-1})^{-1} \mathcal{X}^\top y,$$

where $B = \mathrm{diag}(|\hat{\beta}_j|)$. Then it follows that

$$\begin{aligned} \hat{y} &= \mathcal{X}\hat{\beta}, \\ &= \mathcal{X}(\mathcal{X}^\top \mathcal{X} + \lambda B^{-1})^{-1} \mathcal{X}^\top y. \end{aligned}$$

Then, to calculate $\mathsf{Cov}(\hat{y}_i, y_i)$, one could use $\mathsf{Cov}(\hat{y}_i, y_i) = \mathsf{Cov}(e_i^\top \hat{y}, e_i^\top y) = e_i^\top \mathsf{Cov}(\hat{y}, y) e_i$, where e_i is a vector where the i'th entry is 1 and the rest are zero. Furthermore, each entry in the sum of (9.6) can be calculated to be

$$\begin{aligned} \mathsf{Cov}(\hat{y}_i, y_i) &= e_i^\top \mathsf{Cov}(\hat{y}, y) e_i && (9.7) \\ &= e_i^\top \mathcal{X}(\mathcal{X}^\top \mathcal{X} + \lambda B^{-1})^{-1} \mathcal{X}^\top \mathsf{Cov}(y, y) e_i && (9.8) \\ &= \sigma^2 (\mathcal{X}^\top e_i)^\top (\mathcal{X}^\top \mathcal{X} + \lambda B^{-1})^{-1} (\mathcal{X}^\top e_i) && (9.9) \\ &= \sigma^2 x_i^\top (\mathcal{X}^\top \mathcal{X} + \lambda B^{-1})^{-1} x_i. && (9.10) \end{aligned}$$

Using the fact that (9.10) are scalars for all i as well as the properties of the trace of a matrix and matrix multiplication rules mentioned in Chap. 2, one obtains the final closed-form expression for the effective degrees of freedom in the Lasso problem:

$$\mathrm{df}(\lambda) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{tr}\left\{\sigma^2 x_i^\top (\mathcal{X}^\top \mathcal{X} + \lambda B^{-1})^{-1} x_i\right\}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \operatorname{tr}\left\{x_i x_i^\top (\mathcal{X}^\top \mathcal{X} + \lambda B^{-1})^{-1}\right\} \\
&= \operatorname{tr}\left\{\left(\sum_{i=1}^{n} x_i x_i^\top\right)(\mathcal{X}^\top \mathcal{X} + \lambda B^{-1})^{-1}\right\} \\
&= \operatorname{tr}\left\{\mathcal{X}^\top \mathcal{X}(\mathcal{X}^\top \mathcal{X} + \lambda B^{-1})^{-1}\right\} \\
&= \operatorname{tr}\left\{\mathcal{X}(\mathcal{X}^\top \mathcal{X} + \lambda B^{-1})^{-1}\mathcal{X}^\top\right\}.
\end{aligned}
$$

It should be noted that the formula for the effective degrees of freedom derived above is valid in the case of the underlying model with non-random regressors. When the random design is used and the set of nonzero predictors is not fixed, another estimator should be used.

Orthonormal Design Case

A computationally convenient special case is the so-called orthonormal design framework. In the orthonormal design case $\mathcal{X}^\top \mathcal{X}$ is a diagonal matrix that $\mathcal{X}^\top \mathcal{X} = \mathcal{I}$. Here the explicit Lasso estimate is

$$
\hat{\beta}_j = \operatorname{sign}\left(\hat{\beta}_j^0\right)\left(|\hat{\beta}_j^0| - \gamma\right)^+, \tag{9.11}
$$

$$
\gamma = \frac{\lambda}{2} \text{ subject to } \sum_{j=1}^{p} |\hat{\beta}_j| = s. \tag{9.12}
$$

The formula shows what was already mentioned in the beginning, namely that the Lasso estimate is a compromise between subset selection and ridge regression, the estimate is either shrunk by γ or is set to zero. As a consequence Lasso coefficients can take values between zero and $\hat{\beta}_j^0$.

Example 9.2 (Orthonormal Design Case for $p = 2$) Let $\hat{\beta} = \left(\widehat{\beta_1}, \widehat{\beta_2}\right)^\top$ w.l.o.g. be in the first quadrant, i.e. $\widehat{\beta_1} \geq 0$ and $\widehat{\beta_2} \geq 0$. This gives us the first condition. The orthonormal design ensures that the elliptical contour lines describe circles around the OLS estimate. Thus we get a linear function going through the point $\hat{\beta}^0$ and being orthogonal (if possible) to the first condition. Equalising both conditions

$$
\hat{\beta}_1 + \hat{\beta}_2 = s \tag{9.13}
$$

$$
\hat{\beta}_2 = \hat{\beta}_1 + \left(\hat{\beta}_2^0 - \hat{\beta}_1^0\right) \tag{9.14}
$$

the Lasso estimate can now be accurately determined:

$$\hat{\beta}_1 = \left(\frac{s}{2} + \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2} \right)^+ \tag{9.15}$$

$$\hat{\beta}_2 = \left(\frac{s}{2} - \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2} \right)^+ . \tag{9.16}$$

For cases in which $\left(\frac{s}{2} + \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2} \right) \leq 0$ or $\left(\frac{s}{2} - \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2} \right) \leq 0$ the corresponding Lasso estimates will always be zero as the position of the $\hat{\beta}_1^0$ and corresponding contour lines do not make it possible to get the orthogonality condition mentioned above. Let $\hat{\beta}^0 = (6, 7)^\top$ and tuning parameter $s = 4$. In this case the Lasso estimator is given by, as shown in Fig. 9.3:

$$\hat{\beta}_1 = \frac{4}{2} + \frac{6 - 7}{2} = 1.5, \tag{9.17}$$

$$\hat{\beta}_2 = \frac{4}{2} - \frac{6 - 7}{2} = 2.5. \tag{9.18}$$

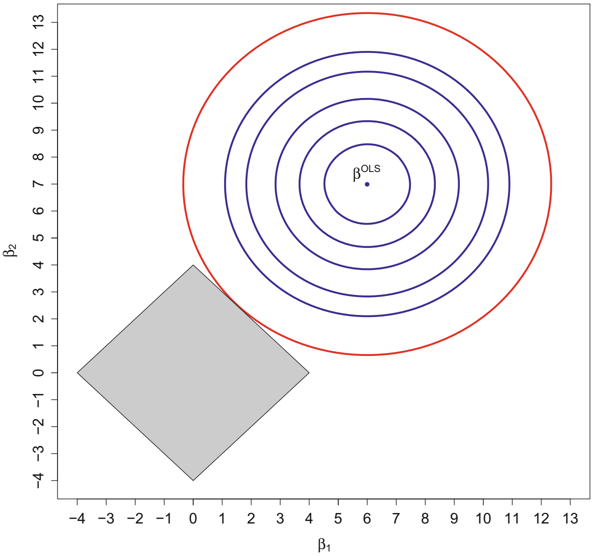

Fig. 9.3 Lasso in the orthonormal design case for $s = 4$ and OLS estimate $\hat{\beta}^0 = (6, 7)^\top$ MVAlassocontour

In terms of λ, the Lasso solution (9.11) in the orthonormal design case can be calculated in a usual unconstrained minimisation problem. Note that in this case the least squares solution is given by

$$\hat{\beta}^0 = (\mathcal{X}^\top \mathcal{X})^{-1} \mathcal{X}^\top y = \mathcal{X}^\top y.$$

Then the minimisation problem is written as

$$\begin{aligned}
\hat{\beta} &= \arg \min_{\beta \in \mathbb{R}^p} \|y - \mathcal{X}\beta\|_2^2 + \lambda \|\beta\|_1 \\
&= \arg \min_{\beta \in \mathbb{R}^p} (y - \mathcal{X}\beta)^\top (y - \mathcal{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \\
&= \arg \min_{\beta \in \mathbb{R}^p} - 2y^\top \mathcal{X}\beta + \beta^\top \beta + \lambda \sum_{j=1}^{p} |\beta_j| \\
&= \arg \min_{\beta \in \mathbb{R}^p} - 2\hat{\beta}^{0\top} \beta + \beta^\top \beta + \lambda \sum_{j=1}^{p} |\beta_j| \\
&= \arg \min_{\beta \in \mathbb{R}^p} \sum_{j=1}^{p} \left(-2\hat{\beta}_j^0 \beta_j + \beta_j^2 + \lambda |\beta_j| \right).
\end{aligned}$$

The objective function can now be minimised by separate minimisation of its jth element. To solve

$$\min_{\beta} (-2\hat{\beta}^0 \beta + \beta^2 - \lambda |\beta|), \tag{9.19}$$

where the index j was dropped for simplicity, let's first assume that $\hat{\beta}^0 > 0$, then $\beta \geq 0$, because a lower value for the objective function may be obtained by changing the sign. Then the solution for the modified problem

$$\min_{\beta} (-2\hat{\beta}^0 \beta + \beta^2 + \lambda \beta) \tag{9.20}$$

is, obviously, $\hat{\beta} = \hat{\beta}^0 - \gamma$, where $\gamma = \lambda/2$, as in (9.11). To ensure the sign consistency for this case, one could see that the solution is

$$\hat{\beta} = (\hat{\beta}^0 - \gamma)^+ = \text{sign}(\hat{\beta}^0)(|\hat{\beta}^0| - \gamma)^+. \tag{9.21}$$

Now let us take $\hat{\beta}^0 \leq 0$, then $\beta \leq 0$ as well and the solution for the new problem

$$\min_{\beta} (-2\hat{\beta}^0 \beta + \beta^2 - \lambda \beta) \tag{9.22}$$

is $\hat{\beta} = \hat{\beta}^0 + \gamma$, but the sign consistency requires that

$$\begin{aligned}\hat{\beta} &= (\hat{\beta}^0 + \gamma)^- \\ &= -(-\hat{\beta}^0 - \gamma)^+ \\ &= \operatorname{sign}(\hat{\beta}^0)(|\hat{\beta}^0| - \gamma)^+.\end{aligned}$$

As the solutions are the same in both cases, the expression $\operatorname{sign}(\hat{\beta}^0)(|\hat{\beta}^0| - \gamma)^+$ is indeed the solution to the original Lasso problem.

General Lasso Solution

For a fixed $s \geq 0$ the Lasso estimation problem is a least squares problem subjected to 2^p linear inequality constraints as there are 2^p different possible signs for $\beta = \left(\beta_1, \dots, \beta_p\right)^\top$. Lawson and Hansen (1974) suggested solving the least squares problem subject to a general linear inequality constraint $G\beta \leq h$ where $G(m \times p)$ corresponds to the $m = 2^p$ constraints and $h = s1_m$. As m could be very large, this procedure is not very fast computationally. Therefore Lawson and Hansen (1974) introduced the inequality constraints sequentially in their algorithm, seeking a feasible solution.

Let $g(\beta) = \sum_{i=1}^n \left(y_i - x_i^\top \beta\right)^2$ and let $\delta_k, k = 1, \dots, 2^p$, be column vectors of p-tuples of the form $(\pm 1, \dots, \pm 1)$. It follows that the linear inequality condition can be equivalently described as $\delta_k^\top \beta \leq s,\ k = 1, \dots, 2^p$. Now let $E = \{k | \delta_k^\top \beta = s\}$ the equality set, m_E the number of elements of E and $G_E = \left(\delta_k^\top\right)_{k \in E}$ a matrix whose rows are all δ_k's for $k \in E$. Now the algorithm works as follows, see Tibshirani (1996):

1. Find OLS estimate $\hat{\beta}^0$ and let $\delta_{k_0} = \operatorname{sign}(\hat{\beta}^0)$, $E = \{k_0\}$.
2. Find $\hat{\beta}$ to minimise $g(\beta)$ subject to $G_E \beta \leq s1_{m_E}$.
3. If $\sum_{j=1}^p |\hat{\beta}_j| \leq s$ the computation is complete.
4. If $\sum_{j=1}^p |\hat{\beta}_j| > s$ add k to the set E where $\delta_k = \operatorname{sign}(\hat{\beta})$ and go back to step 2.
5. The final iteration is a solution to the original problem.

As the number of steps is limited by $m = 2^p$, the algorithm has to converge in finite time. The average number of iterations in practice is between $0.5p$ and $0.75p$.

Example 9.3 Let us consider the car data set (Table 22.3) where $n = 74$. We want to study in-what way the price (X_1) depends on the 12 other variables $(X_2), \dots, (X_{13})$, which are represented by $j = 1, 2, \dots, 12$, using Lasso regression. In Fig. 9.4 one can clearly see that coefficients become nonzero one at a time, that means the variables enter the regression equation sequentially as the scaled shrinkage parameter $\hat{s} = s / \|\hat{\beta}^0\|_1$ increases, in order $j = 6, 11, 9, 3, \dots$

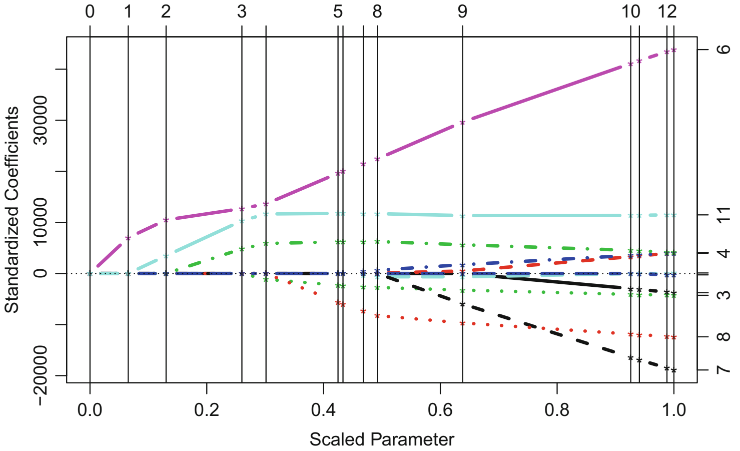

Fig. 9.4 Lasso estimates of standardised regression $\hat{\beta}_j$ for car data with $n = 74$ and $p = 12$ MVAlassoregress

(representing $X_7, X_{12}, X_{10}, X_4, \ldots$), hence the L_1 penalty results in variable selection and the variables which are most relevant are shrunk less. In this example, an optimal $\hat{s}$ can be found such that the fitted model gives the smallest residual (see Exercise 9.3).

9.1.2 Lasso in High Dimensions

The problem with the algorithm by Tibshirani to calculate the Lasso solutions is that it is initialised from an OLS solution of the unconstrained problem which does not correspond to the true model. Another problem is that for the case of $p > n$, this computation is infeasible. Therefore it may be optimal to start with a small initial guess for β and iterate through a different kind of an algorithm to obtain the Lasso solutions. Such an algorithm is based on the properties of the Lasso problem as a convex programming one. Osborne et al. (2000) showed that the original Lasso estimate problem (9.1) can be rewritten as:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2} (y - \mathcal{X}\beta)^\top (y - \mathcal{X}\beta) \overset{\text{def}}{=} \frac{1}{2} r^\top r, \quad \text{subject to} \quad s - \|\beta\|_1 \geq 0, \tag{9.23}$$

where $r \overset{\text{def}}{=} (y - \mathcal{X}\beta)$. Let $\mathcal{J} = \{i_1, \ldots, i_p\}$ be the set of indices such that $|(\mathcal{X}^\top r)_{i_j}| = \|\mathcal{X}^\top r\|_\infty$, for $j = 1, \ldots, p$; so indices in $\mathcal{J}$ correspond to nonzero elements of β. Also let P be the permutation matrix that permutes the elements of the coefficient vector β so that the first elements are the nonzero elements:

$\beta = P^\top (\beta_\mathcal{J}, 0)^\top$. Denote $\theta_\mathcal{J} = \text{sign}(\beta_\mathcal{J})$ be equal to 1 if the corresponding element of $\beta_\mathcal{J}$ is positive and -1 otherwise. Further denoting $f(\beta) = (y - \mathcal{X}\beta)^\top (y - \mathcal{X}\beta)$ the following optimisation algorithm is based on the local linearisation of (9.1) around β:

$$\hat{\beta} = \arg\min_h f(\beta + h), \quad \text{subject to} \quad \theta_\mathcal{J}^\top(\beta_\mathcal{J} + h_\mathcal{J}) \le s \quad \text{and} \quad h = P^\top (h_\mathcal{J}, 0)^\top, \tag{9.24}$$

the solution for which can be shown to be equal to

$$h_\mathcal{J} = (\mathcal{X}_\mathcal{J}^\top \mathcal{X}_\mathcal{J})^{-1}\{\mathcal{X}_\mathcal{J}^\top(y - \mathcal{X}_\mathcal{J}\beta_\mathcal{J}) - \mu\theta_\mathcal{J}\},$$

where

$$\mu = \max\left\{0, \frac{\theta_\mathcal{J}^\top(\mathcal{X}_\mathcal{J}^\top \mathcal{X}_\mathcal{J})^{-1}\mathcal{X}_\mathcal{J}^\top y - s}{\theta_\mathcal{J}^\top(\mathcal{X}_\mathcal{J}^\top \mathcal{X}_\mathcal{J})^{-1}\theta_\mathcal{J}}\right\}.$$

The procedure as a whole is implemented as shown in the "Lasso solution-path optimisation" algorithm. As shown in the algorithm, indices may enter and leave the set $\mathcal{J}$, which makes the Lasso problem similar to other subset selection techniques. Moreover, one can compute the whole path of Lasso solutions for $0 \le s \le s_0$, each time taking the solution for the previous s as a starting point for the next one.

9.1.3 Lasso in Logit Model

The Lasso model can be extended to generalised linear models, one of the most common of which is the logistic regression (logit) model. Coefficients in the logit model have probabilistic interpretation. In the logit model, the linear predictor $\mathcal{X}\beta$ is related to the conditional mean μ of the response variable y via the logit link $\log\{\mu/(1-\mu)\}$. As the response variable is binary, it is binomial-distributed and $\mu = p(x_i)$. Therefore, as defined in (9.25), the logit model for $y \in \{0, 1\}$ of $(n \times 1)$ observations on a binary response variable and $x_i = (x_{i1}, \ldots, x_{ip})^\top$ is,

$$\log\left\{\frac{p(x_i)}{1 - p(x_i)}\right\} = \sum_{j=1}^p \beta_j x_{ij},$$

where

$$p(x_i) = \mathrm{P}(y_i = 1 \mid x_i) = \frac{\exp(\sum_{j=1}^p \beta_j x_{ij})}{1 + \exp(\sum_{j=1}^p \beta_j x_{ij})}. \tag{9.25}$$

Algorithm Lasso solution-path optimisation

```
procedure FIND(optimal $\beta$)
    Choose initial $\beta$ and $\mathcal{J}$ (e.g. $\beta \leftarrow 0$, $\mathcal{J} \leftarrow \emptyset$)
    repeat
        Solve (9.23) to obtain $h$
        Set $\hat{\beta} \leftarrow \beta + h$
        if sign$(\hat{\beta}_{\mathcal{J}}) = \theta_{\mathcal{J}}$ then
            Obtain the solution $\beta = \hat{\beta}$
        else
            repeat
                Find the smallest $\gamma$, $0 < \gamma < 1$, $k \in \mathcal{J}$ such that $0 = \beta_k + \gamma h_k$
                Set $\beta = \beta + \gamma h$
                Set $\theta_k = -\theta_k$
                Solve (9.23) again to obtain a new $h$
                if $\theta_{\mathcal{J}}^\top(\beta_{\mathcal{J}} + h_{\mathcal{J}}) \le s$ then
                    $\hat{\beta} = \beta + h$
                else
                    Update $\mathcal{J} \leftarrow \mathcal{J}_{-k}$
                    Recompute $\beta_{\mathcal{J}}, \theta_{\mathcal{J}}, h$
                end if
            until sign$(\hat{\beta}_{\mathcal{J}}) = \theta_{\mathcal{J}}$
        end if
        Compute $\hat{v} \leftarrow \mathcal{X}^\top \hat{r} / \|\mathcal{X}_{\mathcal{J}}^\top \hat{r}\|_\infty = P^\top(\hat{v}_1, \hat{v}_2)^\top$          ▷ here $\hat{r} = y - \mathcal{X}\hat{\beta}$
        if $-1 \le (\hat{v}_2)_\imath \le 1$ for $1 \le \imath \le p - |\mathcal{J}|$ then
            $\hat{\beta}$ is a solution
        else
            Find $\jmath$ such that $|(\hat{v}_2)_\jmath|$ is maximised
            Update $\mathcal{J} \leftarrow (\mathcal{J}, \jmath)$
            Update $\hat{\beta}_{\mathcal{J}} \leftarrow (\hat{\beta}_{\mathcal{J}}, 0)^\top$
            Update $\theta_{\mathcal{J}} \leftarrow (\theta_{\mathcal{J}}, \text{sign}(\hat{v}_2)_\jmath)^\top$
        end if
        Set $\beta \leftarrow \hat{\beta}$
    until $-1 \le (\hat{v}_2)_\imath \le 1$ for $1 \le \imath \le p - |\mathcal{J}|$
end procedure
```

The Lasso estimate for the logit model is obtained by solving the following optimisation problem:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} g\left(-y_i x_i^\top \beta\right) \right\}, \text{ subject to } \sum_{j=1}^{p} |\beta_j| \le s, \tag{9.26}$$

with tuning parameter $s \ge 0$ and log-loss function $g(u) = \log\{1 + \exp(u)\}$. An alternative representation of the Lasso estimate $\hat{\beta}$ in the logit model is:

$$\arg\min_{\beta} \left\{ \sum_{i=1}^{n} g\left(-y_i x_i^\top \beta\right) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{9.27}$$

Shevade and Keerthi (2003) developed a simple and efficient algorithm to solve the optimisation in (9.27) based on the Gauss–Seidel method using coordinate-wise descent approach. The algorithm is asymptotically convergent and easy to implement. Firstly, define the following terms,

$$\begin{aligned} u_i &= -y_i x_i^\top \beta, \\ F_j &= \sum_{i=1}^{n} \frac{\exp(u_i)}{\exp(1+u_i)} y_i x_{ij}. \end{aligned} \tag{9.28}$$

The first order optimality conditions for (9.27) are:

$$\begin{aligned} F_j &= 0 && \text{if} \quad j = 0, \\ F_j &= \lambda && \text{if} \quad \beta_j > 0,\ j > 0, \\ F_j &= -\lambda && \text{if} \quad \beta_j < 0,\ j > 0, \\ -\lambda \le F_j &\le \lambda && \text{if} \quad \beta_j = 0,\ j > 0. \end{aligned}$$

A new variable is defined

$$\begin{aligned} v_j &= |F_j| && \text{if} \quad j = 0, \\ &= |\lambda - F_j| && \text{if} \quad \beta_j > 0,\ j > 0, \\ &= |\lambda + F_j| && \text{if} \quad \beta_j < 0,\ j > 0, \\ &= \psi_j && \text{if} \quad \beta_j = 0,\ j > 0. \end{aligned}$$

where $\psi_j = \max\{(F_j - \lambda), (-\lambda - F_j), 0\}$. Thus, the first-order optimality conditions can be written as

$$v_j = 0 \quad \forall j. \tag{9.29}$$

It is difficult to obtain exact optimality condition, so the stopping criterion for (9.27) is defined as follows (for some small ε),

$$v_j \le \varepsilon \quad \forall j. \tag{9.30}$$

To write the algorithm, let us define $I_z = \{j : \beta_j = 0, j > 0\}$ and $I_{nz} = \{j : \beta_j \ne 0, j > 0\}$ for sets of zero estimates and sets of nonzero estimates, respectively, and $I = I_z \cup I_{nz}$. The algorithm consists of two loops. The first loop runs over the variables in I_z to choose the maximum violator, v. In the second loop W is optimised with respect to β_v, therefore the set I_{nz} is modified and maximum violator in I_{nz} is obtained. The second loop is repeated until no violators are found in I_{nz}. The algorithm alternates between the first and second loop until no violators exist in both I_z and I_{nz}.

Algorithm Lasso in logit model

```
procedure FIND(optimal Lasso estimate $\hat{\beta}$)
    Set $\beta_j = 0$ for all $j$
    while an optimality violator exists in $I_z$ do
        Find the maximum violator $(v)$ in $I_z$
        repeat
            Optimise $W$ with respect to $\beta_v$
            Find the maximum violator (v) in $I_{nz}$
        until no violator exists in $I_{nz}$
    end while
end procedure
```

Another way to obtain the lasso estimate in the logit model is by maximising the likelihood function of logit model with lasso constraint. The log-likelihood function of logit model is written as

$$\log L(\beta) = \sum_{i=1}^{n} [y_i \log p(x_i) + (1 - y_i) \log\{1 - p(x_i)\}] . \tag{9.31}$$

Suppose $\ell(\beta) = \log L(\beta)$, with $\beta = (\beta_1, \dots, \beta_p)^\top$, the Lasso estimates are obtained by maximising the penalised log likelihood for logit model as follows

$$\hat{\beta} = \arg\max_{\beta} \left\{ n^{-1} \sum_{i=1}^{n} \ell(\beta) \right\} , \text{ subject to } \sum_{j=1}^{p} |\beta_j| \le s. \tag{9.32}$$

It can solved by a general non-linear programming procedure or by using iteratively reweighted least squares (IRLS). Friedman, Hastie, and Tibshirani (2010) developed an algorithm to solve the problem in (9.32). An alternative representation of the Lasso problem is defined as follows:

$$\hat{\beta} = \arg\max_{\beta} \left\{ n^{-1} \sum_{i=1}^{n} \ell(\beta) - \lambda \sum_{j=1}^{p} |\beta_j| \right\} . \tag{9.33}$$

Example 9.4 Following Example 9.3, the price (X_1) of car data set (Table 22.3) has average 6,192.28. We now define a new categorical variable which takes the value 0 if $X_1 \le 6{,}000$ and otherwise is equal to 1. We want to study in what way the price (X_1) depends on the 12 other variables $(X_2, \dots, X_{13})$ using Lasso in logit model.

In Fig. 9.5 one can see that coefficients' dynamics depends on the shrinkage parameter $s = \|\hat{\beta}(\lambda)\|_1$, the L_1 norm of estimated coefficients. An optimal s can be chosen such that the fitted model gives the smallest residual (see Exercise 9.4).

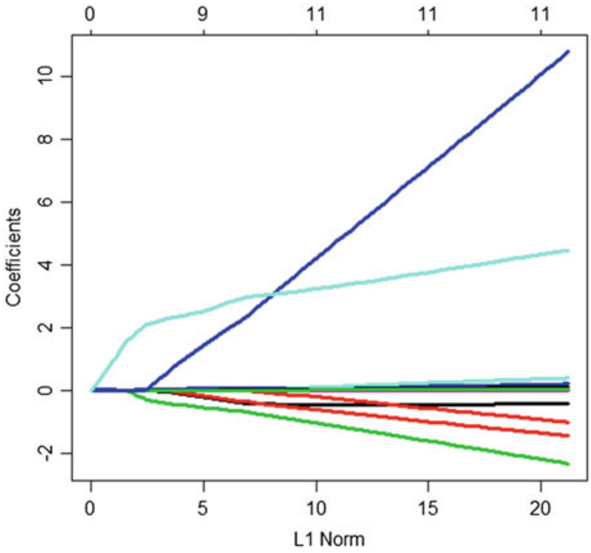

Fig. 9.5 Lasso estimates $\hat{\beta}_j$ of logit model for car data with $n = 74$ and $p = 12$ `MVAlassologit`

9.2 Elastic Net

Although the Lasso is widely used in variable selection, it has several drawbacks. Zou and Hastie (2005) stated that:

1. if $p > n$, the Lasso selects at most n variables before it saturates;
2. if there is a group of variables which has very high correlation, then the Lasso tends to select only one variable from this group;
3. for usual $n > p$ condition, if there are high correlations between predictors, the prediction performance of the Lasso is dominated by ridge regression, see Tibshirani (1996).

Zou and Hastie (2005) introduced the Elastic net which combines good features of the L_1-norm and L_2-norm penalties. The Elastic net is a regularised regression method which overcomes the limitations of the Lasso. This method is very useful when $p \gg n$ or there are many correlated variables. The advantages are: (1) a group of correlated variables can be selected without arbitrary omissions, (2) the number of selected variables is no longer limited by the sample size.

9.2.1 Elastic Net in Linear Regression Model

We describe the Elastic net in linear regression model. For simplicity reason we assume that the x_{ij} are standardised such that $\sum_{i=1}^{n} x_{ij} = 0$ and $n^{-1}\sum_{i=1}^{n} x_{ij}^2 = 1$. The Elastic net penalty $P_\alpha(\beta)$ leads to the following modification of the problem to obtain the estimator $\hat{\beta}$

$$\arg\min_{\beta} \left\{ (2n)^{-1} \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2 + \lambda P_\alpha(\beta) \right\}, \tag{9.34}$$

where

$$\begin{aligned} P_\alpha(\beta) &= \frac{1}{2}(1-\alpha)\,\|\beta\|_2^2 + \alpha\,\|\beta\|_1 \\ &= \sum_{j=1}^{p} \left\{ \frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right\}. \end{aligned} \tag{9.35}$$

The penalty $P_\alpha(\beta)$ is a compromise between ridge regression and the Lasso. If $\alpha = 0$ then the criterion is the ridge regression and if $\alpha = 1$ the method will be the Lasso. Practically, for small $\varepsilon > 0$, the Elastic net with $\alpha = 1 - \varepsilon$ performs like the Lasso, but removes degeneracies and erratic variable selection behaviour caused by extreme correlation. Given a specific λ, as α increases from 0 to 1, the sparsity of the Elastic net solutions increases monotonically from 0 to the sparsity of the Lasso solutions.

The Elastic net optimisation problem can be represented as the usual Lasso problem, using modified $\mathcal{X}$ and y vectors, as shown in the following example.

Example 9.5 To turn the Elastic net optimisation problem into the usual Lasso one, one should first augment y with p additional zeros to obtain $\tilde{y} = (y, 0)^\top$. Then, augment $\mathcal{X}$ with the multiple of the $p \times p$ identity matrix $\sqrt{\lambda\alpha}\mathcal{I}$ to get $\tilde{\mathcal{X}} = \left(\mathcal{X}^\top, \sqrt{\lambda\alpha} I\right)^\top$. Next, define $\tilde{\lambda} = \lambda(1-\alpha)$ and solve the original Lasso minimisation problem in terms of the new input $\tilde{y}$, $\tilde{\mathcal{X}}$ and $\tilde{\lambda}$. This new problem is equivalent to the original Elastic net problem:

$$\begin{aligned} \|\tilde{y} - \tilde{\mathcal{X}}\beta\|_2^2 + \tilde{\lambda}\|\beta\|_1 &= \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} \mathcal{X}\beta \\ \sqrt{\lambda\alpha}\mathcal{I}\beta \end{bmatrix} \right\|_2^2 + \lambda(1-\alpha)\|\beta\|_1, \\ &= \|y - \mathcal{X}\beta\|_2^2 - \lambda\alpha\|\beta\|_2^2 + \lambda\|\beta\|_1 - \lambda\alpha\|\beta\|_1, \\ &= \|y - \mathcal{X}\beta\|_2^2 + \lambda\left\{\alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1\right\}, \end{aligned}$$

which is equivalent to the original Elastic net problem.

We follow the idea of Friedman et al. (2010) who used a coordinate descent algorithm to solve the optimisation problem in (9.34). Let us suppose to have estimates $\tilde{\beta}_k$ for $k \neq j$. Then we optimise (9.34) partially with respect to β_j by computing the gradient at $\beta_j = \tilde{\beta}_j$, which only exists if $\tilde{\beta}_j \neq 0$. Having the soft-thresholding operator $S(z, \gamma)$ as

$$\operatorname{sign}(z)\,(|z| - \gamma)^+ = \begin{cases} z - \gamma & \text{if} \quad z > 0 \quad \text{and} \quad \gamma < |z|, \\ z + \gamma & \text{if} \quad z < 0 \quad \text{and} \quad \gamma < |z|, \\ 0 & \text{if} \qquad \gamma \geq |z|. \end{cases} \tag{9.36}$$

it can be shown that the coordinate-wise update has the following form

$$\widetilde{\beta_j} = \frac{S\left\{n^{-1}\sum_{i=1}^{n} x_{ij}\left(y_i - \tilde{y}_i^{(j)}\right), \lambda\alpha\right\}}{1 + \lambda(1 - \alpha)}, \tag{9.37}$$

where $\tilde{y}_i^{(j)} = \sum_{k \neq j} x_{ik}\tilde{\beta}_k$ is a fitted value which excludes the contribution x_{ij}, therefore $y_i - \tilde{y}_i^{(j)}$ is partial residual for fitting β_j.

The algorithm computes the least square estimate for the partial residual $y_i - \tilde{y}_i^{(j)}$, then applies the soft-thresholding rule to perform the Lasso contribution to the penalty $P_\alpha(\beta)$. Afterwards, a proportional shrinkage is applied to ridge penalty. There are several methods used to update the current estimate $\tilde{\beta}$. We describe the simplest updating method, the so-called naive update.

The partial residual can be rewritten as follows:

$$\begin{aligned} y_i - \tilde{y}_i^{(j)} &= y_i - \hat{y}_i + x_{ij}\widetilde{\beta_j} \\ &= r_i + x_{ij}\widetilde{\beta_j}, \end{aligned} \tag{9.38}$$

with $\widehat{y_i}$ being the current fit and r_i the current residual. As x_j is standardised, therefore

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}\left(y_i - \tilde{y}_i^{(j)}\right) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}r_i + \widetilde{\beta_j}. \tag{9.39}$$

Note that the first term on the right-hand side of the new partial residual is the gradient of the loss with respect to β_j.

9.2.2 *Elastic Net in Logit Model*

The Elastic net penalty can similarly be applied to the logit model. Recall the log-likelihood function of the logit model in (9.31),

$$\log L(\beta) = \sum_{i=1}^{n} \left[y_i \log p\,(x_i) + (1 - y_i)\log\{1 - p\,(x_i)\}\right].$$

Penalised log-likelihood for the logit model using Elastic net has the following form

$$\max_{\beta} \left\{ n^{-1} \sum_{i=1}^{n} \ell(\beta) - \lambda P_{\alpha}(\beta) \right\}, \tag{9.40}$$

with $\ell(\beta) = \log L(\beta)$. The solution of (9.40) can be found by means of a Newton algorithm. For a fixed λ and given a current parameter $\tilde{\beta}$, the quadratic approximation (Taylor expansion) is updated about current estimates $\tilde{\beta}$ as follows:

$$\ell_Q(\beta) = -(2n)^{-1} \sum_{i=1}^{n} w_i (z_i - x_i^{\top}\beta)^2 + C(\tilde{\beta})^2, \tag{9.41}$$

where working response and weight, respectively, are:

$$\begin{aligned} z_i &= x_i^{\top}\tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)\{1 - \tilde{p}(x_i)\}}, \\ w_i &= \tilde{p}(x_i)\{1 - \tilde{p}(x_i)\}. \end{aligned}$$

A Newton update is obtained by minimising $\ell_Q(\beta)$.

Friedman et al. (2010) proposed similar approach creating an outer loop for each value of λ, which computes a quadratic approximation in (9.41) about current estimates $\tilde{\beta}$. Afterwards, a coordinate descent algorithm is used to solve the following penalised weighted least squares problem (PWLS)

$$\min_{\beta} \left\{ -\ell_Q(\beta) + \lambda P_{\alpha}(\beta) \right\}. \tag{9.42}$$

This inner coordinate descent loop continues until the maximum change in (9.42) is less than a very small threshold.

9.3 Group Lasso

The Group Lasso was first introduced by Yuan and Lin (2006) and was motivated by the fact that the predictor variables can occur in several groups and one could want a parsimonious model which uses only a few of these groups. That is, assume that there are K groups and the vector of coefficients is structured as follows

$$\beta^G = (\beta_1^{\top}, \dots, \beta_K^{\top})^{\top} \in \mathbb{R}^{\sum_k p_k},$$

where p_k is the coefficient vector dimension of the kth group, $k = 1, \dots, K$. A sparse set of groups is produced, although within each group either all entries of β_k,

$k = 1, \dots, K$, a corresponding element of the whole vector β^G are zero or all of them are nonzero. The Group Lasso problem can be formulated in general as

$$\arg \min_{\beta \in \mathbb{R}^{\sum_k p_k}} n^{-1} \left\| y - \sum_{k=1}^{K} \mathcal{X}_k \beta_k \right\|_2^2 + \lambda \sum_{k=1}^{K} \sqrt{p_k} \|\beta_k\|_2, \tag{9.43}$$

where $\mathcal{X}_k$ is the kth component of the matrix $\mathcal{X}$ with columns corresponding to the predictors in the group k, β_k is the coefficient vector for that group and p_k is the cardinality of the group, i.e. the size of the coefficient vector which serves as a balancing weight in the case of widely differing group sizes. It is obvious that if groups consist of single elements, i.e. $p_k = 1\ \forall k$, then the Group Lasso problem is reduced to the usual Lasso one.

The computation of the Group Lasso solution involves calculating the necessary and sufficient subgradient KKT conditions for $\hat{\beta}^G = (\hat{\beta}_1^\top, \dots, \hat{\beta}_K^\top)^\top$ to be a solution for (9.43)

$$-\mathcal{X}_k^\top \left(y - \sum_{k=1}^{K} \mathcal{X}_k \beta_k \right) + \frac{\lambda \beta_k \sqrt{p_k}}{\|\beta_k\|} = 0, \tag{9.44}$$

if $\beta_k \neq 0$; otherwise, for $\beta_k = 0$, it holds that

$$\left\| \mathcal{X}_k^\top \left(y - \sum_{l \neq k} \mathcal{X}_l \hat{\beta}_l \right) \right\| \leq \lambda \sqrt{p_k}. \tag{9.45}$$

Expressions (9.44) and (9.45) allow to calculate the solution, the so-called update step which can be used to implement an iterative algorithm to solve the problem (9.43). The solution resulting from the KKT conditions is readily shown to be the following:

$$\hat{\beta}_k = \left\{ \left(\lambda \sqrt{p_k} \|\hat{\beta}_k\|^{-1} + \mathcal{X}_k^\top \mathcal{X}_k \right)^{-1} \right\}^+ \mathcal{X}_k^\top \hat{r}_k, \tag{9.46}$$

where the residual $\hat{r}_k$ is defined as $\hat{r}_k \overset{\text{def}}{=} y - \sum_{l \neq k} \mathcal{X}_l \hat{\beta}_l$. As a special (orthonormal) case, when $\mathcal{X}_l^\top \mathcal{X}_l = \mathcal{I}$, the solution is simplified to the $\hat{\beta}_k = (\lambda \sqrt{p_k} \|\hat{\beta}_k\|^{-1} + 1)\mathcal{X}_k^\top \hat{r}_k$. To obtain a full solution to this problem, Yuan and Lin (2006) suggest using a blockwise coordinate descent algorithm which iteratively applies the estimate (9.46) to $k = 1, \dots, K$.

Meier, van de Geer, and Bühlmann (2008) extended the Group Lasso to the case of logistic regression and demonstrated convergence of several algorithms for the computation of the solution as well as outlined consistency results for the Group Lasso logit estimator. The general setup for that model involves a binary response variable $y_i \in \{0, 1\}$ and K groups predictor variable $x_i = (x_{i1}^\top, \dots, x_{ik}^\top)^\top$, both x_i and y_i are i.i.d., $i = 1, \dots, n$. Then the logistic linear regression model may be written as before:

$$\log\left\{\frac{p(x_i)}{1 - p(x_i)}\right\} = \eta(x_i) \overset{\text{def}}{=} \beta_0 + \sum_{k=1}^{K} x_{ik}^\top \beta_k, \tag{9.47}$$

where the conditional probability $p(x_i) = \mathrm{P}(y_i = 1|x_i)$. The Group Lasso logit estimator $\hat{\beta}$ then minimises the objective function

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^{p+1}} \left\{-\ell(\beta) + \lambda \sum_{k=1}^{K} \sqrt{p_k} \|\beta_k\|_2\right\}, \tag{9.48}$$

where $\ell(\cdot)$ is the log-likelihood function

$$\ell(\beta) = \sum_{i=1}^{n} y_i \eta(x_i) - \log[1 + \exp\{\eta(x_i)\}].$$

The problem is solved through a group-wise minimisation of the penalised objective function by, for example, the block-coordinate descent method.

Example 9.6 The Group Lasso results can be illustrated by an application to the MEMset Donor dataset of human donor splice sites with a sequence length of 7 base pairs. The full dataset (training and test parts) consists of 12.623 true ($y_i = 1$) and 269.155 false ($y_i = 0$) human donor sites. Each element of data represents a sequence of DNA within a window of the splice site which consists of the last three positions of the exon and first 4 positions of the intron; so the strings of length 7 are made up of 4 characters A, C, T, G and therefore the predictor variables are 7 factors, each having 4 levels. False splice sites are sequences on the DNA which match the consensus sequence at position four and five. Figure 9.6 shows how the Group Lasso does shrinkage on the level of groups built by DNA letters.

As is seen from Example 9.6, the solution to the Group Lasso problem yields a sparse solution only regarding the "between" case, that is, it excludes some of the groups from the model but then all coefficients in the remaining groups are nonzero. To ensure both the sparsity of groups and within each group, Simon, Friedman, Hastie, and Tibshirani (2013) proposed the so-called "sparse Group Lasso" which uses a more general penalty which yields sparsity an both inter- and intragroup level. The sparse Group Lasso estimate solves the problem

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\| y - \sum_{k=1}^{K} \mathcal{X}_k \beta_k \right\|_2^2 + \lambda_1 \sum_{k=1}^{K} \|\beta_k\|_2 + \lambda_2 \|\beta\|_1, \tag{9.49}$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_K)^\top$ is the entire parameter vector.

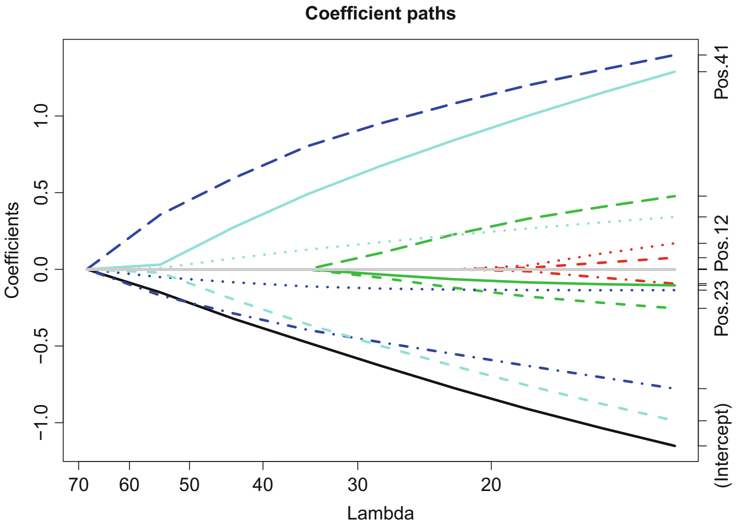

Fig. 9.6 Lasso estimates of standardised regression $\hat{\beta}_j$ for car data with $n = 74$ and $p = 12$ MVAgrouplasso

Summary

↪ Lasso gives a sparse solution. Lasso estimate combines best of both ridge regression and subset regression.

↪ If there is a group of variables which has very high correlation, then the Lasso tends to select only one variable from the group.

↪ The LARS algorithm computes the whole path of Lasso solutions and is feasible for the high-dimensional case $p \gg n$.

↪ Elastic net combines good features of L_1-norm and L_2-norm penalties.

↪ The Elastic net is very useful when $p \gg n$ or there are many correlated variables.

↪ The Sparse Group Lasso can perform shrinkage both on inter- and intragroup level.

9.4 Exercises

Exercise 9.1 *Derive the explicit Lasso estimate in (9.11) for the orthonormal design case.*

Exercise 9.2 *Compare Lasso orthonormal design case for $p = 2$ graphically to ridge regression, i.e. to the problem $\hat{\beta} = \operatorname{argmin}\left\{\sum_{i=1}^{n}\left(y_i - x_i^{\top}\beta\right)^2\right\}$ subject to $\sum_{j=1}^{p} {\beta_j}^2 \leq s$.*
Why does Lasso produce variable selection and ridge regression does not?

Exercise 9.3 *Optimise the value of s such that the fitted model in Example 9.3 produces the smallest residual.*

Exercise 9.4 *Optimise the value of s such that the fitted model in Example 9.4 produces the smallest residual.*