Chapter 6

Penalised Regression Methods

We continue with the discussion of regression methods when X^tX is singular, or close to singular. So far, we have introduced PC regression and PLS regression, both of which are *shrinkage* methods. We now continue with other shrinkage methods.

6.1 Ridge Regression

If X^tX is close to singular, it can be made non-singular by adding a small constant k to the diagonal entries before taking the inverse: $((X^tX) + kI_r)^{-1}$ is used instead of $(X^tX)^{-1}$. The ridge regression estimator is therefore:

$$\widehat{\beta}_{rr} = (X^tX + kI)^{-1}X^tY = W(k)\widehat{\beta}_{OLS}$$

where

$$W(k) = (X^tX + kI)^{-1}X^tX.$$

Here $k \geq 0$. When $k > 0$, the estimator is *biased*. The parameter k is chosen to minimise Q_{res}, the error sum of squares. It can be characterised in the following two ways:

1. A ridge regression estimator is the solution of a penalised least squares problem. Specifically, it is the r-vector β that minimises

   $$Q_{res} := (Y - X\beta)^t(Y - X\beta)$$

   subject to the condition that $\|\beta\|^2 \leq c$ where c is a constant chosen by the user.

   In the two dimensional setting (where $\beta = \binom{\beta_1}{\beta_2}$), this may be represented pictorially: the level sets $(Y - X\beta)^t(Y - X\beta) = k$ for $k \geq 0$ are ellipses. The value of k, say $k(c)$ is chosen as the smallest k such that the ellipse intersects the circle $\|\beta\|^2 \leq c$ and estimate of $\binom{\beta_1}{\beta_2}$ is the value such that $(Y - X\beta)^t(Y - X\beta) = k(c)$ and $\|\beta\|^2 = c$.

Such constrained optimisation problems may be solved using the standard *Lagrange multiplier* technique. Consider the function

$$\phi(\beta) = (Y - X\beta)^t(Y - X\beta) - \lambda\beta^t\beta$$

where $\lambda \geq 0$ is a Lagrange multiplier. Differentiating with respect to β and setting the derivative equal to 0 gives

$$(X^tX + \lambda I)\widehat{\beta}_{rr}(\lambda) = X^tY \Rightarrow \widehat{\beta}_{rr}(\lambda) = (X^tX + \lambda I)^{-1}X^tY.$$

The value of λ is determined by the constraint that $\beta^t\beta \leq c$.

2. A ridge regression estimator is a shrinkage estimator that shrinks the OLS estimator towards zero.

   Consider the singular value decomposition of X as $U\Lambda^{1/2}V^t$ where U is $n \times n$ orthonormal, V is $r \times r$ orthonormal and $\Lambda^{1/2}$ is the diagonal matrix of eigenvalues: the values of Λ are ordered highest to lowest. Let $G = XV = U\Lambda^{1/2}$ so that $G^tG = \Lambda$, then

$$\begin{aligned}
\widehat{\beta}_{rr}(k) &= (X^tX + kI)^{-1}X^tY \\
&= (V\Lambda V^t + kVV^t)^{-1}V\Lambda^{1/2}U^tY \\
&= V(\Lambda + kI)^{-1}\Lambda^{1/2}U^tY \\
&= V(\Lambda + kI)^{-1}G^tY.
\end{aligned}$$

   Set $\alpha = V^t\beta$, so that $\beta = V\alpha$, then the canonical form of the regression model is:

$$Y = X\beta + \epsilon = G\alpha + \epsilon$$

   and the OLS estimator of α is

$$\widehat{\alpha}_{ols} = (G^tG)^{-1}G^tY = \Lambda^{-1}V^tX^tY.$$

   Set

$$\widehat{\alpha}_{rr} = V^t\widehat{\beta}_{rr} = (\Lambda + kI)^{-1}G^tY = (\Lambda + kI)^{-1}\Lambda\widehat{\alpha}_{ols}.$$

   Hence we have, for the jth component,

$$\widehat{\alpha}_{rr,j}(k) = \left(\frac{\lambda_j}{\lambda_j + k}\right)\widehat{\alpha}_{ols,j}.$$

Hence

$$\|\widehat{\beta}_{rr}(k)\|^2 = \|\widehat{\alpha}_{rr}(k)\|^2 = \sum_{j=1}^{k} \left(\frac{\lambda_j}{\lambda_j + k}\right) \widehat{\alpha}_{ols}^2$$

which is monotonically decreasing as a function of k. Hence $\|\widehat{\beta}_{rr}(k)\| < \|\widehat{\beta}_{ols}\|$ for all $k > 0$, hence the ridge estimator is a shrinkage estimator.

The Bias-Variance Trade-off The mean squared error of the ridge regression estimator is:

$$MSE(k) = \mathbb{E}[(\widehat{\beta}_{rr}(k) - \beta)^t(\widehat{\beta}_{rr}(k) - \beta)] = \text{Var}(k) + \text{Bias}^2(k).$$

It is straightforward to compute these terms:

$$\begin{aligned} \text{Var}(k) &= \text{tr}(\sigma^2(X^tX + kI)^{-1}X^tX(X^tX + kI)^{-1} \\ &= \sigma^2\text{tr}\{(\Lambda + kI)^{-1}\Lambda(\Lambda + kI)^{-1}\} \\ &= \sigma^2 \sum_{j=1}^{r} \frac{\lambda_j}{(\lambda_j + k)^2}. \end{aligned}$$

The bias is

$$\begin{aligned} \mathbb{E}[\widehat{\beta}_{rr}(k)] - \beta &= \{(X^tX + kI)^{-1}X^tX - I\}\beta \\ &= \{(V\Lambda V^t + kI)^{-1}V\Lambda V^t - I\}V\alpha \\ &= V\{(\Lambda + kI)^{-1}\Lambda - I\}\alpha. \end{aligned}$$

The bias-squared term is therefore

$$\text{Bias}^2(k) = k^2 \sum_{j=1}^{r} \frac{\alpha_j^2}{(\lambda_j + k)^2}$$

giving a total mean squared error of:

$$\text{MSE}(k) = \sum_{j=1}^{r} \frac{\sigma^2\lambda_j + k^2\alpha_j^2}{(\lambda_j + k)^2}$$

where λ_j is the jth largest eigenvalue of X^tX and σ^2 is the error variance.

The ridge parameter is chosen to minimise the mean squared error although this expression is not so useful.

Estimating the Ridge Parameter One way of estimating k is known as the *ridge trace*. This is a graphical display of all the components of the vector $\widehat{\beta}_{rr}(k)$ plotted on the same scatterplot against a range of values of k. The value of k to be used is estimated as the *smallest* value for which the trace stabilises for all coefficients.

The ridge trace is also used as a variable selection procedure. If an estimated regression coefficient changes sign in the graph of its ridge trace, this is taken to mean that the OLS estimator of that coefficient has an incorrect sign, so that the variable should not be included. However, this is heuristic and there are situations where a variable with good predictive properties can be eliminated by this method.

Cross validation has become the standard method, which we deal with shortly.

6.2 Estimating Prediction Error

Consider linear regression, where the assumption is that the errors are mean zero with variance σ^2 and that they are uncorrelated, but we make no further distributional assumptions. Without the assumption of Gaussianity, the prediction intervals computed under the assumption of Gaussianity are not valid.

The empirical prediction error, computed when the parameters are estimated on a learning set and the prediction error is then estimated on an independent 'test' set is the benchmark for evaluating the performance of a linear model.

We consider two possibilities for the X matrix: it may be that the explanatory variables are arrived at in a random manner, so that $(Y_i, X_{i.})_{i=1}^n$ may be considered as a collection of n independent identially distributed row vectors, or it may be that X is *designed*: the values of X are chosen by the user (and hence non-random) and Y is a response. We use the terms *random X case* and *non-random X case* to describe these two situations.

When the entire data set is sufficiently large, we can partition the data into learning sets (for learning the parameters) and test sets (for estimating the prediction error) and validation sets (to see if our prediction error, computed empirically using the test set) corresponds to the error for new observations.

Often, though, the data set is not large enough for this, or such a division may not be practical for other reasons. In such a situation, there are alternative methods.

6.2.1 Apparent Error Rate

Suppose that β has been estimated as $\widehat{\beta}$ and hence, for a new instantiation (y, x), the predicted value of y is $\widehat{y} = \widehat{\mu}(x) = x\beta$ (where x is taken as a row vector: β as a column vector).

Using squared error loss function, the loss incurred by predicting y by $\mu(x)$ is defined as:

$$L(y, \mu(x)) = (y - \mu(x))^2.$$

The loss in this case is therefore $(y - \widehat{\mu}(x))^2$. The *apparent* error rate, using the entire data set could therefore be estimated by:

$$\widehat{PE}(\widehat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\mu}(X_{i.}))^2 = \frac{Q_{\text{res}}}{n}.$$

This (of course) is misleadingly optimistic: firstly, we should divide by $n - r$ rather than n. This shouldn't affect the answer too much when $n >> r$. More importantly, $(y_i, X_{i.})$ is being used *both* to estimate $\widehat{\mu}$ *and then* as a 'new' observation. This (of course) is true both for random-X and fixed-X. We now consider 'cross validation', a technique which is useful for the random-X case, but is not appropriate for fixed-X.

6.2.2 Cross Validation

Suppose that the data set is an i.i.d. sample of n observations from the $r+1$ random vector (Y, X). If $n = 2m$, then split the data into two disjoint sets, each with m data points: compute $\widehat{\mu}$ based on data $m+1, \ldots, 2m$ and then estimate the prediction error by:

$$\widehat{PE} = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \widehat{\mu}(X_i))^2.$$

An overall prediction error can be estimated by switching the roles of the two sets to obtain another estimate and then averaging over the two.

This procedure may be generalised. Assume that $n = Vm$, where V is a small integer. In the above, we had $V = 2$. We split the data randomly into V disjoint subsets each of size m. We then use $V - 1$ of the subsets for learning the parameters and the remaining subset to estimate the prediction error. We can compute V estimates in this way and we average over them.

The most computationally intensive version of cross validation is 'leave one out', or LOO. In this setting, $\widehat{\mu}_i$ is computed by leaving out observation i and then the prediction error is estimated by:

$$\widehat{PE}_{LOO} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_{-i}(X_i))^2.$$

Now consider estimation of the model error ME, defined by:

$$ME = \mathbb{E}[(\widehat{\mu}(X) - \mu(X))^2] = (\beta - \widehat{\beta}_{OLS})^t \Sigma_{XX}(\beta - \widehat{\beta}_{OLS}) + (\widehat{\beta} - \beta, m_X)^2$$

where $\widehat{\mu}$ is the estimate of μ computed from the observations and the expectation is taken with respect to a new, independent copy of the r-vector X. Σ denotes covariance and m_X denotes expectation. We would like to estimate the model error. This is done quite simply by subtracting the variance estimate from the prediction error:

$$\widehat{ME} = \widehat{PE} - \widehat{\sigma}^2,$$

where $\widehat{\sigma}^2 = \frac{Q_{res}}{n-r}$, r is the rank of X.

6.2.3 Bootstrap

Estimating prediction error may also be carried out using bootstraps. We use *unconditional* bootstrap in the random-X case and a *conditional* bootstrap for the fixed-X case.

Firstly, sample n times *with replacement* from the original sample to get a random-X bootstrap sample. Obtain the OLS estimate using this sample. Then estimate the prediction error by averaging over $(Y - \widehat{\mu}(X))^2$ for those points (Y, X) which are *not included* in the original bootstrap sample.

6.2.4 Conditional Bootstrap

The *conditional* bootstrap for the fixed-X case operates by sampling *with replacement* from the *residuals* obtained from OLS regression using the complete sample. Firstly, the regression is performed, the residuals computed and $\widehat{\sigma}^2$ computed. Having re-sampled n errors $e_1^*, \ldots, e_n^*$, we compute n 'new' responses

$$Y_i^* = \widehat{\mu}(X_i) + e_i^* \qquad i = 1, \ldots, n$$

where $\widehat{\mu}$ is obtained by OLS regression using the complete sample. Then Y^* is regressed on X to get the bootstrapped estimator of the regression coefficients β^*. Under this bootstrap sampling scheme, $\sqrt{n}(\beta^* - \widehat{\beta})$ has, approximately, the same distribution as $\sqrt{n}(\widehat{\beta} - \beta)$.

6.2.5 Cross Validation for Ridge Regression

We have already seen 'leave one out' cross validation in study of the UBC (unbiased cross validation) density estimator. The *V-fold Cross-validation* algorithm for Ridge Regression and choosing the ridge parameter k is as follows:

1. Standardise each column of X so that it has mean 0 and standard deviation 1.

2. Partition the data into V learning and test sets, where $V = 4, 10$ or n. For example, with LOO (leave one out) there are n learning /test sets, each learning set constructed by leaving out one of the variables; the test set is the variable left out.

3. Choose possible values of k, equally spaced, say $k_1, \ldots, k_N$.

4. For $i = 1, \ldots, N$ and $v = 1, \ldots, V$

   - Use the vth learning set to compute the coefficients $\widehat{\beta}_v(k_i)$

- Obtain an estimate of the prediction error $\widehat{PE}_v(k_i)$ by using $\widehat{\beta}_v(k_i)$ on the vth test set.

5. For $i = 1, \ldots, N$
   - Average the V prediction error estimates to get an overall estimate of the prediction error.
   - Plot the overall estimate against k_i.
6. Choose the value of k that minimises the estimated average prediction error.

6.3 Regularised Regression

Both ridge regression and variable selection have their advantages and disadvantages. It would therefore be useful to construct a hybrid of these which takes the best properties of both methods.

Consider the general form of the *penalised least squares* criterion: minimising

$$\phi(\beta) = (Y - X\beta)^t(Y - X\beta) + \lambda p(\beta)$$

for a given penalty p and regularisation parameter λ. For example: let

$$p_q(\beta) := \sum_{j=1}^{r} |\beta_j|^q$$

Similarly to ridge regression, we can choose λ so that the minimising β solves the standard least squares problem subject to the constraint $\sum_j |\beta_j|^q \leq c$ for a fixed $c > 0$. If we subscript the ϕ function by q, then ϕ_q is a *smooth* convex function for $q > 1$ and convex for $q = 1$.

Ridge regression corresponds to $q = 2$. The ridge regression estimator is that point on the elliptical contours of $Q_{\text{res}}(\beta)$ centred at $\widehat{\beta}_{OLS}$ which first touches the hypersphere $\sum_j \beta_j^2 \leq c$. The parameter c controls the size of the hypersphere and hence how much $\widehat{\beta}_{OLS}$ is shrunk towards the origin.

When $q \neq 2$, the penalty is no longer rotationally invariant. The most interesting case is $q < 2$ where the penalty function collapses towards the coordinate axes. It not only shrinks the coefficients towards zero; it also sets some of them to be zero.

The case of $q = 1$ produces the so-called LASSO (Least Absolute Shrinkage and Selection Operator). Therefore, the LASSO method finds a β which minimises

$$\phi(\beta) := (Y - X\beta)^t(Y - X\beta) + \lambda \sum_{j=1}^{r} |\beta_j|.$$

The OLS regression coefficients are shrunk towards the origin, with the value of c (equivalently λ) controlling the amount of shrinkage. At the same time, it also behaves like a variable selection technique.For a given value of $c > 0$, only a subset of the coefficient estimates $\widehat{\beta}_j : j = 1, \ldots, r$ will have non-zero values. A coefficient value will be exactly zero when one of the elliptical contours of the function

$$Q_{res}(\beta) = Q_{res}(\widehat{\beta}_{OLS}) + (\beta - \widehat{\beta}_{OLS})^t X^t X (\beta - \widehat{\beta}_{OLS})$$

touches a corner of the diamond shaped penalty function.

6.4 The Garotte

This tecnhique only works when $\widehat{\beta}_{OLS}$ exists. Let W be a diagonal matrix with nonnegative weights along the diagonal. The problem is to find the weights which minimise:

$$\phi(w) = (Y - XW\widehat{\beta}_{OLS})^t (Y - XW\widehat{\beta}_{OLS})$$

subject to one of the following two constraints:

1. $w_j \geq 0$ for each j, $\sum_k w_j \leq c$
2. $\sum_j w_j^2 \leq c$.

As c is decreased, more of the w_j become 0, thus eliminating those variables. The non-zero estimates shrink towards 0.

In these regression techniques, the regularisation parameter λ presents a compromise between how well the function fits the data and the size of the coefficients. The value of λ is determined by V-fold cross validation.

6.5 Least-Angle Regression

This is an automatic variable-selection method, which may be used in situations where $r >> n$.

Assume that the variables have been centred and standardised, so that there is no β_0 coefficient, $\sum_i X_{ij} = 0$ and $\sum_i X_{ij}^2 = 1$ for each $j = 1, \ldots, r$. The 'output' variable has mean zero: $\sum_i Y_i = 0$. We consider two algorithms, the *forward stagewise* and *LARS*

Let $\widehat{c}$ denote the *current* correlations between the columns of X and the *current* residual vector $r = Y - \widehat{\mu}$, where $\widehat{\mu}$ is the *current* estimated mean value.

The Forwards Stagewise Algorithm

1. Initialise $\widehat{\beta} = 0$ so that $\widehat{\mu} = 0$. Set $r = Y$.
2. Find the covariate vector $X_{.j_1}$ most highly correlated with Y ($j_1 = \text{argmax}_j |\widehat{c}_j|$).
3. Update $\widehat{\beta}_{j_1} \leftarrow \widehat{\beta}_{j_1} + \epsilon \text{sgn}(\widehat{c}_{j_1})$ where ϵ is a small constant that controls the step length.
4. Update $\widehat{\mu} \leftarrow \widehat{\mu} + \delta_{j_1} X_{.j_1}$ and $r \leftarrow r - \delta_{j_1} X_{.j_1}$.
5. Return to 2., repeat loop until convergence.

The LARS Algorithm The LARS algorithm improves on this:

1. Initialise $\widehat{\beta} = 0$ so that $\widehat{\mu} = 0$ and $r = Y$. Start with the 'active' set of indices A as the empty set.

2. Find the covariate vector $X_{.j_1}$ most highly correlated with r. Set $A \leftarrow A \cup \{j_1\}$. Add $X_{.j_1}$ to the regression model.

3. Move $\widehat{\beta}_{j_1}$ in the direction $\text{sgn}(\widehat{c}_{j_1})$ (choosing ϵ as in forward stagewise) *until some other covariate vector, say $X_{.j_2}$ has the same correlation with r as $X_{.j_1}$*. The new active set is $A \leftarrow A \cup \{j_2\}$ and $X_{.j_2}$ is added to the model.

4. Update r and move $(\widehat{\beta}_{j_1}, \widehat{\beta}_{j_2})$ in the joint OLS direction when r is regressed against $(X_{.j_1}, X_{.j_2})$, until a third covariate vector $X_{.j_3}$ has the same correlation with r as the first two variables. Add j_3 to A and add X_{j_3} to the regression model.

5. After k LARS steps, $A = \{j_1, \ldots . j_k\}$. Let $\widehat{\mu}_A$ denote the current LARS estimate, where exactly k coefficients $(\widehat{\beta}_{j_1}, \ldots, \widehat{\beta}_{j_k})$ are non-zero. The current vector of correlations is: $\widehat{c} = X^t(Y - \widehat{\mu}_A)$.

6. Continue until all r covariates have been added to the model and $\widehat{c} = 0$. THis is the OLS solution.

The R package **lars** includes a C_p type statistic as a stopping rule for choosing a suitable LARS model.

Tutorial: (Written Exercises)

1. **PLS (Partial Least Squares) Regression** Recall the PLS algorithm from lectures. Show that if $X^tX = I$ (where the columns of X have been centred and standardised) then the algorithm stops after a single step.

2. Consider the regression problem where the $n \times r$ matrix X has been centred and standardised; there are p explanatory variables. Let $S = \frac{1}{n}X^tX$. For a deterministic vector α, the notation $\mathrm{Var}(X\alpha)$ therefore denotes $\alpha^tS\alpha$.

   (a) For PC regression, show that the mth principal component direction v_m solves:

   $$\max_{\alpha} \mathrm{Var}(X\alpha)$$

   subject to: $\|\alpha\| = 1$ and $\alpha^tSv_j = 0$ for $j = 1, \ldots, j-1$.

   (b) Show that the mth PLS direction $\widetilde{\phi}_m$ (i.e. $Z_m = \sum_{j=1}^{r} \widetilde{\phi}_{mj}X_{\cdot j}^{(m)}$) solves:

   $$\max_{\alpha} \mathrm{Corr}^2(Y, X\alpha)\mathrm{Var}(X\alpha)$$

   subject to: $\|\alpha\| = 1$, $\alpha^tS\widetilde{\phi}_l = 0$, $l = 1, \ldots, m-1$.

   (c) Hence show that PLS directions are a compromise between OLS estimation and PC directions.

3. Consider the problem $Y = X\beta + \epsilon$ where X is an $n \times p+1$ matrix, with $X_{\cdot 0} = \mathbf{1}_n$ ($\mathbf{1}_n$ is the n-vector with each entry 1; the first column of X, labelled column 0, is a column of 1s). Let $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^t$. Let $\widetilde{X}$ denote the $n \times p$ matrix which is obtained by centring columns $1, \ldots, p$ of X. That is:

   $$(\mathbf{0}|\widetilde{X}) = (I_n - \mathbf{1}_n\mathbf{1}_n^t)X$$

   where I_n denotes the $n \times n$ identity matrix. Show the relation between $\widehat{\beta}_{rr}$ and $\widehat{\beta}_{rr}^{(c)}$ where

   $$\widehat{\beta}_{rr} = \mathrm{argmin}_{\beta}\{(Y - X\beta)^t(Y - X\beta) + k\|\beta\|^2\}$$

   and $\widehat{\beta}_{rr}^{(c)} = (\widehat{\beta}_{0;rr}^{(c)}|\widehat{\beta}_{1;rr}^{(c)t})^t$ satisfies:

   $$\widehat{\beta}_{rr}^{(c)} = \mathrm{argmin}_{\beta^{(c)}}\{(Y - \beta_0^{(c)}\mathbf{1}_n - \widetilde{X}\beta_1^{(c)})^t(Y - \beta_0^{(c)}\mathbf{1}_n - \widetilde{X}\beta_1^{(c)}) + k(\beta_0^{(c)2} + \|\beta_1^{(c)}\|^2)\}$$

   where $\beta^{(c)}$ is a $p+1$ vector; $\beta^{(c)} = (\beta_0^{(c)}|\beta_1^{(c)t})^t$ and $\beta_1^{(c)}$ is a p vector.

4. Consider a *Bayesian* approach to the regression problem. Let β be a p-vector of parameters and let $\pi(\beta)$ (the prior distribution over the parameters) be $N(0, \tau I_p)$ where $\tau > 0$ and I denotes the identity matrix. Suppose that the conditional distribution of Y (the n-vector of observations) given β is: $Y|\beta \sim N(X\beta, \sigma^2 I)$.

   Compute the posterior distribution $\pi(\beta|Y)$ (up to proportionality) and show that the mean (and mode) of the posterior is the same as the regression estimate of the parameter vector β.

   What is the ridge parameter k in terms of τ and σ^2?

5. Consider the following technique to turn a ridge regression problem into an OLS problem. Augment the $n \times r$ matrix X by adding in r additional rows; these rows form an $r \times r$ matrix $\sqrt{k}I$. Call the resulting matrix X^*. Augment the n-vector Y by adding in r zeros and call the resulting vector Y^*. Show that the ridge regression estimator is obtained by regressing Y^* against X^*.

   This gives a way of carrying out ridge regression, although much of the output will be irrelevant for the original problem of regressing Y on X.

6. Suppose we run a ridge regression, with ridge parameter k, where $X = (\mathbf{1}_n|x)$; $x = (x_1, \ldots, x_n)^t$ is an n vector (i.e. there is one explanatory variable) and obtain parameter estimates $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)^t$. Suppose we now run a ridge regression using $X^* = (\mathbf{1}_n|x|x)$ (i.e. we introduce another variable which is exactly the same). Show that the parameter estimates are: $\beta^* = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_1)^t$.

7. Consider the LASSO estimator for the coefficients β for a model $Y = X\beta + \epsilon$. That is: β minimises $(Y - X\beta)^t(Y - X\beta)$ subject to the constraint that $\sum |\beta| \leq t$ for some $t > 0$. Suppose that the fitted value for variable X_j is $\widehat{\beta}_j = a$. Suppose that we now augment the set of variables by X^* which is an identical copy of X_j. Suppose that $\widehat{\gamma}^*$ and $\widehat{\gamma}_j$ are the parameter estimates for X^* and X_j respectively in the enlarged model. Describe the set of solutions for the enlarged model, in terms of solutions for the original model, for fixed parameter t.

8. Consider the problem of minimising

$$L(\beta) := (Y - X\beta)^t(Y - X\beta)$$

   subject to the constraint $\sum_j |\beta_j| \leq t$. Suppose that this is obtained by minimising

$$L(\beta) + \lambda \sum_j |\beta_j|$$

   (λ is the value such that the solution of the constrained optimisation problem is the minimiser of the Lagrange functional).

   For each $j = 1, \ldots, r$, set $\beta_j = \beta_j^+ - \beta_j^-$ where $\beta_j^+ = \max(\beta_j, 0)$ and $\beta_j^- = -\min(\beta_j, 0)$.

(a) Show that the solution to the problem can be obtained by finding the minimum of the functional:

$$\mathcal{F}(\beta) = (Y - X\beta)^t(Y - X\beta) + \lambda \sum_j (\beta_j^+ + \beta_j^-) - \sum_j \lambda_j^+ \beta_j^+ - \sum_j \lambda_j^- \beta_j^-$$

subject to conditions: $\lambda_j^+ \beta_j^+ = 0$, $\lambda_j^- \beta_j^- = 0$, so that the solution is obtained by solving:

$$\begin{cases} \frac{\partial}{\partial \beta_j} L(\beta) + \lambda - \lambda_j^+ = 0 \\ -\frac{\partial}{\partial \beta_j} L(\beta) + \lambda - \lambda_j^- = 0 \\ \lambda_j^+ \beta_j^+ = 0 \qquad \lambda_j^- \beta_j^- = 0 \qquad j = 1, \ldots, r \end{cases}$$

(b) Show that $\left| \frac{\partial}{\partial \beta_j} L(\beta) \right| \leq \lambda$ for all $j = 1, \ldots r$ and that one of the following three scenarios holds for the optimal β:

i) $\lambda = 0 \Rightarrow \frac{\partial}{\partial \beta_j} L(\beta) = 0$ for $j = 1, \ldots, r$

ii)

$$\beta_j^+ > 0, \quad \lambda > 0 \quad \Rightarrow \quad \lambda_j^+ = 0, \quad \frac{\partial}{\partial \beta_j} L(\beta) = -\lambda < 0, \quad \beta_j^- = 0$$

iii)

$$\beta_j^- > 0, \quad \lambda > 0 \Rightarrow \lambda_j^- = 0, \quad \frac{\partial}{\partial \beta_j} L(\beta) = \lambda > 0, \quad \beta_j^+ = 0.$$

(c) Hence show that for any 'active' predictor with $\beta_j \neq 0$, either ($\frac{\partial}{\partial \beta_j} L(\beta) = -\lambda$ and $\beta_j > 0$) or ($\frac{\partial}{\partial \beta_j} L(\beta) = \lambda$ and $\beta_j < 0$).

(d) Assuming the predictors are standardised, relate λ to the correlation between the jth predictor and the residuals.

(e) Suppose that the set of active predictors is unchanged for $\lambda_0 \geq \lambda \geq \lambda_1$. Show that there is a vector γ_0 such that

$$\widehat{\beta}(\lambda) = \widehat{\beta}(\lambda_0) - (\lambda - \lambda_0)\gamma_0 \qquad \lambda \in [\lambda_1, \lambda_0].$$

Thus the LASSO path is linear as λ ranges from λ_0 to λ_1.

Answers

1. Let us follow through the algorithm in the setting where $X^tX = I$. First, the variables are centred and standardised - we assume this is done.

   Step 1 Regress $Y^{(k-1)}$ on $X_{.j}^{(k-1)}$ for each j to get $\widehat{\beta}_{k-1,j} = \frac{\text{Cov}(X_{.j}^{(k-1)}, Y^{(k-1)})}{\text{Var}(X_{.j}^{(k-1)})}$. Since $\text{Var}(X_{.j}^{(0)}) = 1$, this gives:

   $$\widehat{\beta}_j = \text{Cov}(X_{.j}, Y)$$

   Now take the weighted average; $Z_k \propto \sum_{j=1}^{r} \text{Cov}(X_{.j}^{(k-1)}, Y^{(k-1)}) X_{.j}^{(k-1)}$. This gives

   $$Z = \sum_{j=1}^{r} \widehat{\beta}_j X_{.j}$$

   and since $X^tX = I$ and X is centred (so it has mean 0),

   $$\text{Var}(Z) = \sum_{j=1}^{r} \widehat{\beta}_j^2$$

   Now regress $Y^{(k-1)}$ on Z_k to get the OLS coefficient

   $$\widehat{\theta}_k = \frac{\text{Cov}(Z_k, Y^{(k-1)})}{\text{Var}(Z_k)}.$$

   Here $\text{Cov}(Z_k, Y^{(k-1)}) = \sum_{j=1}^{r} \widehat{\beta}_j \text{Cov}(X_{.j}, Y) = \sum_{j=1}^{r} \widehat{\beta}_j^2$ so

   $$\widehat{\theta} = 1$$

   Now, $Y^{(1)}$ is the residual vector;

   $$Y^{(1)} = Y - Z$$

   and now we remove from $X_{.j}$ its projection onto Z; do an OLS to get coefficient

   $$\widehat{\phi}_{kj} = \frac{\text{Cov}(Z_k, X_{.j}^{(k-1)})}{\text{Var}(Z_k)}$$

   so that, using $X^tX = I$ (which gives orthogonality),

   $$\widehat{\phi}_j = \frac{\widehat{\beta}_j}{\sum_i \widehat{\beta}_i^2}$$

   $$X_{.j}^{(1)} = X_{.j} - \frac{\widehat{\beta}_j}{\sum_i \widehat{\beta}_i^2} Z$$

   so that

$$\begin{aligned}\mathrm{Cov}(X_{.j}^{(1)}, Y^{(1)}) &= \mathrm{Cov}(X_{.j}, Y) - \mathrm{Cov}(X_{.j}, Z) - \frac{\widehat{\beta}_j}{\sum_i \widehat{\beta}_i^2}\mathrm{Cov}(Y, Z) + \frac{\widehat{\beta}_j}{\sum_i \widehat{\beta}_i^2}\mathrm{Var}(Z) \\ &= \widehat{\beta}_j - \widehat{\beta}_j - \widehat{\beta}_j + \widehat{\beta}_j \\ &= 0 \qquad \forall j = 1, \ldots, r.\end{aligned}$$

The termination occurs when all these covariances are 0.

2. (a) Straight from the definition: The PCR directions are the eigenvectors of $X^tX = P\Lambda P^t$ where P is orthonormal. Since the columns of X have been centralised (so that they have mean 0), therefore:

$$\mathrm{Var}(X\alpha) = \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{r} X_{ij}\alpha_j)^2 = \frac{1}{n}\alpha^t X^t X\alpha = \alpha^t P\Lambda P\alpha.$$

Now, we can remove the constraint $\|\alpha\| = 1$ by maximising $\frac{\mathrm{Var}(X\alpha)}{\|\alpha\|^2}$ without any constraint on $\|\alpha\|$. Maxima occur at zeroes of:

$$\frac{\partial}{\partial\alpha_i}\frac{\mathrm{Var}(X\alpha)}{\|\alpha\|^2} = \frac{\frac{2}{n}\sum_j (X^tX)_{ij}\alpha_j}{\|\alpha\|^2} - 2\frac{\mathrm{Var}(X\alpha)}{\|\alpha\|^4}\alpha_i$$

Hence any maximising α satisfying $\|\alpha\| = 1$ satisfies:

$$\frac{1}{n}(X^tX)\alpha = \mathrm{Var}(X\alpha)\alpha$$

so that solutions α are eigenvectors of $\frac{1}{n}(X^tX)$ with corresponding eigenvalues $\mathrm{Var}(X\alpha)$. The first direction v_1 maximises $\mathrm{Var}(X\alpha)$ subject to the constraint $\|\alpha\| = 1$. Subsequent directions v_m solve the same maximisation problem subject to the constraint $\alpha' S v_j = 0$ for $j = 1, \ldots, m-1$.

(b) The mth PLSR direction is $\widetilde{\phi}_m$; these are the coefficients such that $Z_m = \sum_{j=1}^{r}\widetilde{\phi}_{mj}X_{.j}$.

Now consider the constraints: $Y^{(m-1)}$ is he projection of Y onto the space that is spanned by $X\widetilde{\phi}_m, \ldots X\widetilde{\phi}_r$ so finding the mth direction is equivalent to maximising the expression with $Y^{(m-1)}$. Then $X^{(m-1)}$ is the projection of X onto the space spanned by $Z_m, \ldots, Z_r$. Therefore, with the constraints, the problem is equivalent to finding the unit vector α which maximises $\mathrm{Corr}(Y^{(m-1)}, X^{(m-1)}\alpha)^2\mathrm{Var}(X^{(m-1)}\alpha)$.

Now, $\mathrm{Corr}(Y, X\alpha) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{r} Y_i X_{ij}\alpha_j}{\sqrt{\mathrm{Var}(Y)\mathrm{Var}(X\alpha)}}$ and the directions α satisfy $\nabla\mathcal{F}(\alpha) = 0$ where

$$\mathcal{F}(\alpha) = \frac{\mathrm{Cov}(Y, X\alpha)^2}{\|\alpha\|^2}.$$

$$\frac{\partial}{\partial \alpha_j}\mathcal{F}(\alpha) = \frac{2\text{Cov}(Y, X\alpha)\sum_{i=1}^{n} Y_i X_{ij}}{\|\alpha\|^2} - 2\frac{\text{Cov}(Y, X\alpha)^2\alpha_j}{\|\alpha\|^4}$$

so that, for $\|\alpha\| = 1$, α satisfies:

$$\alpha = \frac{1}{\text{Cov}(Y, X\alpha)} X^t Y.$$

This gives the result;

$$\widetilde{\phi}_{mj} \propto \text{Cov}(Y^{(m-1)}, X_{.j}^{(m-1)}).$$

From the expression in the question, we see that the object to be maximised is $\text{Var}(X\alpha)$, just as for PCR, but multiplied by $\text{Corr}(Y, X\alpha)^2$, the component which connects the vector α with the vector Y.

3. Simply match up the terms. Let $X = (\mathbf{1}_n | X_1)$ so that:

$$(\mathbf{1}_n | X_1)\widehat{\beta}_{rr} = (\mathbf{1}_n | \widetilde{X})\widehat{\beta}_{rr}^{(c)}$$

then

$$\widehat{\beta}_{rr;0}\mathbf{1}_n + X_1\widehat{\beta}_{rr;1} = \widehat{\beta}_{rr;0}^{(c)}\mathbf{1}_n + (I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t)X_1\widehat{\beta}_{rr;1}^{(c)}.$$

$$\widehat{\beta}_{rr;0}\mathbf{1}_n + X_1\widehat{\beta}_{rr;1} = \mathbf{1}_n(\widehat{\beta}_{rr;0}^{(c)} + \sum_{j=1}^{r}\overline{x}_{.j}\widehat{\beta}_{rr;1j}^{(c)}) + X_1\widehat{\beta}_{rr;1}^{(c)}$$

This gives: $\widehat{\beta}_{rr;1}^{(c)} = \widehat{\beta}_{rr;1}$ and therefore $\widehat{\beta}_{rr;0}^{(c)} = \widehat{\beta}_{rr;0} - \sum_{j=1}^{r}\overline{x}_{.j}\widehat{\beta}_{rr;1j}$.

4.

$$\pi(\beta|y) \propto \pi(\beta)p(y|\beta) = \frac{1}{(2\pi)^{p/2}\tau^{p/2}}\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\tau}|\beta|^2 - \frac{1}{2\sigma^2}(Y - X\beta)^t(Y - X\beta)\right\}$$

(a) **Maximum Posterior Estimate** Maximising posterior is equivalent to minimising

$$(Y - X\beta)^t(Y - X\beta) + \frac{\sigma^2}{\tau}|\beta|^2$$

so it is equivalent to a ridge parameter of $k = \frac{\sigma^2}{\tau}$.

(b) **Mean Posterior Estimate** In the case of the Gaussian distribution, the *mean posterior* and *maximum posterior* estimates are the same.

We'll now show this - and in the process we'll recover the posterior distribution for β.

$$\pi(\beta|Y) \propto \exp\left\{-\frac{1}{2\tau}Q\right\}$$

where

$$Q = (Y - X\beta)'(Y - X\beta) + \frac{\sigma^2}{\tau}|\beta|^2 = \beta'\left(X'X + \frac{\sigma^2}{\tau}\right)\beta - 2\beta'X'Y + Y'Y$$

Let $A = X'X + \frac{\sigma^2}{\tau}I$. This is symmetric and invertible and positive definte. For $A = PDP^t$, let $D^{1/2}$ denote the diagonal matrix with positive square roots of elements of D and let $A^{1/2} = PD^{1/2}P^t$ then

$$Q = \beta'A^{1/2}A^{1/2}\beta - 2\beta'A^{1/2}A^{-1/2}X'Y + Y'XA^{-1}X'Y - Y'XA^{-1}X'Y + Y'Y,$$

which is

$$Q = (\beta'A^{1/2} - Y'XA^{-1/2})(A^{1/2}\beta - A^{-1/2}X'Y) + Y'(I - XA^{-1}X')Y = (i) + (ii).$$

The first part is equal to:

$$(i) = (\beta' - Y'XA^{-1})A(\beta - A^{-1}X'Y).$$

This gives us everything, since $\pi(\beta|Y)$ is a density (and hence integrates up to 1 so that

$$\pi(\beta|Y) = \frac{1}{(2\pi\tau)^{p/2}|A|^{1/2}}\exp\{-\frac{1}{2\tau}(\beta' - Y'XA^{-1})A(\beta - A^{-1}X'Y)\}$$

so the mean posterior estimate is:

$$\mathbb{E}[\beta|Y] =: \widehat{\beta}_{MEP} = A^{-1}X'Y = (X'X + \frac{\sigma^2}{\tau}I)^{-1}X'Y$$

which is the same as the MAP, which is $\widehat{\beta}_{rr}$ with ridge parameter $k = \frac{\sigma^2}{\tau}$.

5. The OLS estimator for the augmented problem is the minimiser of

$$(Y^* - X^*\beta)^t(Y^* - X^*\beta) = Y^tY - 2Y^tX\beta + \beta^t(X^tX + kI)\beta$$

and is therefore the solution to the ridge regression problem with ridge parameter k.

6. $X = (\mathbf{1}_n|x|x)$ hence the estimates solve:

$$\begin{pmatrix} n+k & n\overline{x} & n\overline{x} \\ n\overline{x} & n\overline{x^2}+k & n\overline{x^2} \\ n\overline{x} & n\overline{x^2} & n\overline{x^2}+k \end{pmatrix} \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} n\overline{y} \\ n\overline{xy} \\ n\overline{xy} \end{pmatrix}$$

This equation has unique solution and the result is clear from this formulation; directly from this (subtract second row from third) $k(\widehat{\beta}_2 - \widehat{\beta}_1) = 0$. Let $\widehat{\gamma}_0$, $\widehat{\gamma}_1$ denote the estimated parameters for the problem with $X = (\mathbf{1}_n|x)$, then (clearly) $\widehat{\gamma}_0 = \widehat{\beta}_0$ and $\widehat{\beta}_1 = \widehat{\beta}_2 = \frac{\widehat{\gamma}_1}{2}$.

7. Rearrange X so that the jth column is the last column and denote the jth column by x. Let X_1 denote the matrix such that $X = (X_1|x)$. Then the minimisation problem is: find $\beta_1^t, \gamma_j, \gamma_j^*$ which minimise

$$(Y - X_1\beta_1 - x(\gamma_j + \gamma^*))^t(Y - X_1\beta_1 - x(\gamma_j + \gamma^*))$$

subject to the constraint

$$\sum_i |\beta_{1i}| + |\gamma_j| + |\gamma^*| \leq t.$$

This is the same optimisation problem as the original, with $\beta_j = \gamma_j + \gamma^*$,

so the solutions of interest are: β_1 same as before and $a = \gamma_j + \gamma^*$ so that:

$$(\widehat{\gamma}^*, \widehat{\gamma}_j) = ((1-\alpha)a, \alpha a) : |\alpha||a| + |1-\alpha||a| + \sum_i |\beta_{1i}| \leq t.$$

In particular, if $|a| + \sum_i |\beta_{1i}| = t$, then $\alpha \in [0, 1]$.

8. (a) The Lagrange multiplier technique states: find β which minimises

$$\mathcal{L}(\beta) = (Y - X\beta)^t(Y - X\beta) + \nu \sum(\beta_i^+ + \beta_i^-) + \sum \nu_i^+ \beta_i^+ + \sum \nu_i^- \beta_i^-$$

where for each i, $\beta_i = \beta_i^+ - \beta_i^-$ and the values of ν_i^+, ν_i^- are chosen to satisfy the constraints of $\beta_i^+ \geq 0$, $\beta_i^- \geq 0$. Then:

$$0 = 2(X^tX)_{j\cdot}\beta + 2Y^tX_{\cdot j} + (\nu + \nu_j^+)\mathbf{1}(\beta_j > 0) - (\nu + \nu_j^-)\mathbf{1}(\beta_j < 0)$$

and matching this up with the minimiser of the Lagrangian of the stated problem gives that a valid choice of ν, ν_i^+, ν_i^- is:

$$\nu = \lambda, \qquad \nu_i^+ = 0 \quad \beta_i > 0, \qquad \nu_i^- = 0 \quad \beta_i < 0$$

as required.

(b) The fact that $\lambda_j^+ \geq 0$ and $\lambda_j^- \geq 0$ for each j follows from the fact that the minimiser exists. The fact that one of these three possibilities holds is then clear.

(c) This follows directly from the previous part.

(d) Recall that $\nabla L(\beta) = -2X'(Y - X\beta)$. The residuals are: $R = Y - X\widehat{\beta}$ and, using $\frac{\partial}{\partial \beta_j} L(\beta) = -\lambda$ for $\beta_j > 0$ and λ for $\beta_j < 0$, we have

$$2(X^t R)_j = -\lambda \text{sgn}(\beta_j).$$

Hence, for each j, using the fact that the columns of X have been centred and standardised:

$$\text{Corr}(X_{\cdot,j}, R) = -\frac{\lambda}{2n\sqrt{\text{Var}(R)}} \text{sgn}(\beta_j)$$

(e) For the active predictors,

$$2(X_{\cdot,j}, Y) - \sum_k \sum_l X_{kj} X_{kl} \widehat{\beta}_l(\lambda) = \begin{cases} \lambda \\ -\lambda \end{cases}$$

depending on the sign of the active predictor. Hence linear, hence the result.