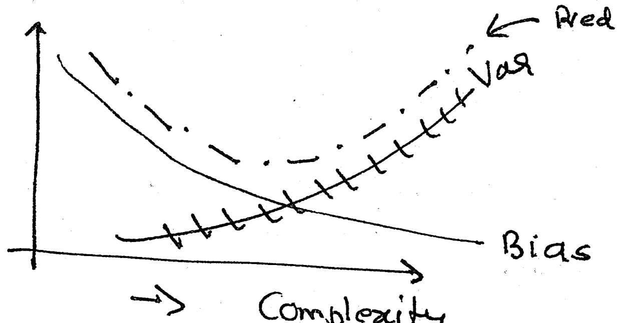
Regularization:

The Need: The Bias Variance trade-off.



- Var ++++
- → Bias^2
- — . — . — . Prediction Error.

With more terms, Local structure, Curvature may be picked up.

→ Coeff. Estimates suffer from high variance as more terms are included in the model.

- If β_js are unconstrained,
  - → They can explode,
  - → Susceptible to high variance.
- Scenarios where 'p' Nos of dimensions are far greater than Nos of rows 'data' available.
  e.g. Medical Data.

When X is high dimensional, covariates are super collinear.

Side Note: Spectral Decomposition.

Normal Matrix M s.t. $MM^T = M^T M$ the eigenspaces corresponding to different matrices are orthogonal to each other, though eigenvalues can still be complex.

$$Av = \lambda v$$

Linear Transform. Eigen vector. Eigen value.

Eigen vectors are values that a linear transformation 'A' merely elongates or shrinks

Let A be $(N \times N)$ matrix with 'N' linearly independent eigenvectors, q_i $(i = 1, \ldots N)$ Then A can be factorized as $A = Q \Lambda Q^{-1}$

where Λ is a diagonal matrix, whose diagonal elements are eigen vectors of A.

Q is a $(N \times N)$ matrix whose ith column is the eigen vector q_i

$$A^{-1} = Q \Lambda^{-1} Q^{-1}$$

We can also write

$$A = \sum_{j=1}^{p} \lambda_j v_j v_j^T$$

Inverse of A is then $A^{-1} = \sum_{j=1}^{p} \lambda_j^{-1} v_j v_j^T$

RHS is undefined if even a single $\lambda_j = 0$

Summary: Cols. of a high dimensional design matrix X are linearly dependent and this causes $X^T X$ to be Singular

Now $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$ cannot be evaluated.

Fix Proposed by Hoerl and Kennard 1970 → adhoc Fix

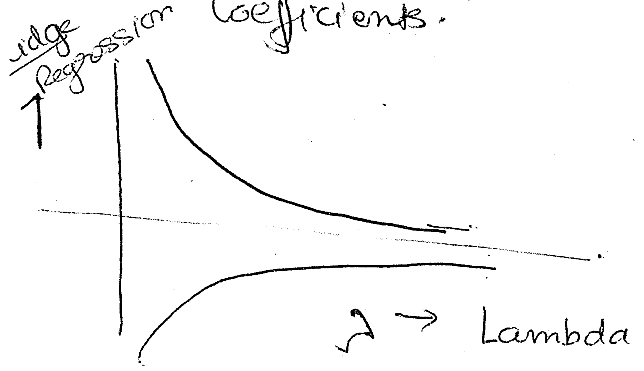
Replace $X^T X$ by $X^T X + \lambda I_{p \times p}$.

Proceed to define the ridge regression estimator

$$\hat{\beta}(\lambda) = (X^T X + \lambda I_{p \times p})^{-1} X^T y$$

Now Let us see how we can get the Ridge Regularization Path.

: Note] For every 'λ' we have a ridge Estimate of the Coefficients.



Proof of Ridge Estimator. $\hat{\beta}_{ridge} = (X^T X + \lambda I_{p \times p})^{-1} X^T y$

we have.

°] impose Ridge constraint.

Minimize $\sum_{i=1}^{n} (y_i - \beta^T x_i)^2$ subject to $\sum_{j=1}^{p} \beta_j^2 \le t$

°] Rewriting the following penalized sum of squares

$$PRSS(\beta)_{l_2} = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{n} \beta_j^2$$

$$= (y - X\beta)^T (y - X\beta) + \lambda \|\beta\|^2$$

Now for Min Error] $\frac{\partial}{\partial \beta} PRSS(\beta/l_2) = -2X^T(y - X\beta) + 2\lambda \|\beta\|$

$$= -2X^T y + 2\lambda\beta + 2X^T X\beta = 0$$

or $(\lambda I + X^T X)\beta = X^T y$

or $\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y$

Proving that $\hat{\beta}_{ridge}$ is biased.

We have

$$\hat{\beta}_{ridge} = (X^TX + \lambda I_p)^{-1} X^T y \quad \text{— (1)}$$

Let $R = X^TX$

Thus we can re-write (1) as $(R + \lambda I_p)^{-1} R \underbrace{R^{-1} X^T y}$

$$= (R(I_p + \lambda I R^{-1}))^{-1} R \hat{\beta}_{OLS}$$

$$= [I + \lambda R^{-1}]^{-1} (R^{-1} R) \hat{\beta}_{OLS}$$

$$= [I_p + \lambda R^{-1}]^{-1} \hat{\beta}_{OLS}.$$

$$\neq \hat{\beta}_{OLS}$$

Why Does the adhoc Fix work?

$X = UDV^T$] Consider Singular Value Decomposition.

$$X_{mn} = U_{mm} D_{mn} \underbrace{V^T_{nn}}$$

orthogonal

orthogonal

From Eig (XX^T)

diagonal Matrix

From Eig ~~(X^TX)~~ (X^TX)

Rewrite OLS estimator in terms of Singular Values.

we have $X = UDV^T$

$X^T = (UDV^T)^T = (DV^T)^T U^T = V D^T U^T$

Thus $X^TX = VD^TU^TUDV^T$ Now $U^TU = I$ (orthogonal)

Thus we have $(X^TX)\cancel{\hat{\beta}_{OLS}} = VD^TDV^T$

Note: orthogonal Property
$V^TV = I$
$VV^T = I$

Let $A = VD^2$, $B = V^T$

Then $(X^TX)^{-1} = (AB)^{-1} = B^{-1}A^{-1} = (V^T)^{-1}(VD^2)^{-1}$
$= VD^{-2}V^T$

Thus $\hat{\beta}_{OLS} = (X^TX)^{-1}X^Ty$

$= VD^{-2}\underbrace{V^T V}D^TU^Ty$

$\hat{\beta}_{OLS} = V\underbrace{D^{-2}D^T}U^Ty$ → Singular Values For $\hat{\beta}_{OLS}$.

Now Look at Ridge Regression:

$\hat{B}(\lambda) = (X^TX + \lambda I)^{-1}X^Ty$

$= (VD^2V^T + \lambda I)^{-1}(VDU^T)y$

$= V(D^2 + \lambda I)^{-1}\underbrace{V^TV}DU^Ty$

$\hat{\beta}_{ridge} = V\underbrace{(D^2+\lambda I)^{-1}D}_{\text{Singular values.}}U^Ty$

$\hat{\beta}_{OLS} = VD^{-2}D^TU^Ty$

$\hat{\beta}_{OLS}$] Singular values $D^{-2}D^T$ say d_{jj}^{-1}

For Ridge $\dfrac{d_{jj}}{(d_{jj}^2 + \lambda)}$

Thus Ridge Penalty Shrinks the Effects of Singular Values.

Representing M.V. data

Vector

'n' observations of 'p' variables

?ectors of dimension 'p'

$y_2 \ldots y_n$

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ip} \end{pmatrix} \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_p \end{pmatrix}$$

Matrix $\Rightarrow$ n rows $\times$ p cols.

$$\begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & & \\ y_n & & y_{np} \end{bmatrix} \qquad y_1 = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1p} \end{pmatrix}, \quad y_1' = (y_{11} \cdots y_{1p})$$

$$Y = \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{pmatrix} \qquad \bar{y}' = \frac{1}{n} j' Y$$

where $j = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{pmatrix}_{n \times 1}$

Thus $\bar{y} = \frac{1}{n} Y' j$

the mean vector

r of 'p' means is given as

$\cancel{\frac{1}{n} Y j}$ $\boxed{\bar{y} = \frac{1}{n} Y' j}$

ance matrix.

$$S_{jk} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ \vdots & & & \\ s_{p1} & \cdots & \cdots & s_{pp} \end{bmatrix}$$

l elements

$$s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2 = s_j^2$$

$$= \frac{1}{n-1}\left(\sum_i y_{ij}^2 - n\bar{y}_j^2\right)$$

ple cov. of jth, kth var.

$$= \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)(y_{ik}-\bar{y}_k)$$

$$= \frac{1}{n-1}\left(\sum_i y_{ij}y_{ik} - n\bar{y}_j\bar{y}_k\right)$$

$$\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})'$$

$(p\times n)(n\times p) = (p\times p)$

Thus. $(y_i-\bar{y})' = (y_{i1}-\bar{y}_1,\ y_{i2}-\bar{y}_2,\ \ldots\ y_{ip}-\bar{y}_p)$

$= \Big[$ p cols → , ↓ n Rows $\Big]$

Now $y_{ij}y_{ik}$ → Product of cols of j, k of y

$n\times p$ $\quad y'y \to y_{ij}y_{ik}$.

also $\bar{y} = \frac{y'j}{n}$ $\quad \bar{y}' = \frac{j'y}{n}$

$$\therefore\ n\bar{y}_j\bar{y}_k = y'\left(\frac{J}{n}\right)y$$

$$S = \frac{1}{n-1}\left[y'y - y'\left(\frac{J}{n}\right)y\right] = \frac{1}{n-1}y'\left(I-\frac{1}{n}J\right)y$$

7.2 The PRESS statistic

We mentioned in Section 5.7.1 that cross-validation can provide a reliable estimate of the algorithm generalization error G_N. The disadvantage of such an approach is that it requires the training process to be repeated l times, which sometimes means a large computational effort. However, in the case of linear models there exists a powerful statistical procedure to compute the leave-one-out cross-validation measure at a reduced computational cost (Fig.7.5). It is the PRESS (Prediction Sum of Squares) statistic [5], a simple formula which returns the leave-one-out (l-o-o) as a by-product of the parametric identification of $\hat{\beta}$ in Eq. 8.1.39.

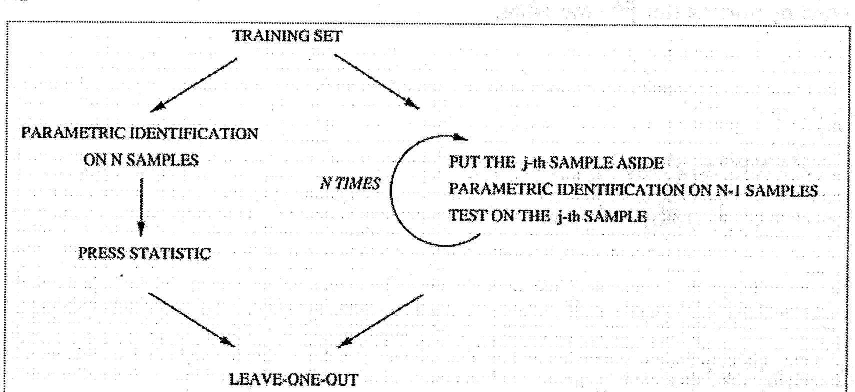


Figure 7.5: *Leave-one-out for linear models. The leave-one-out error can be computed in two equivalent ways: the slowest way (on the right) which repeats N times the training and the test procedure; the fastest way (on the left) which performs only once the parametric identification and the computation of the PRESS statistic.*

Consider a training set D_N in which for N times

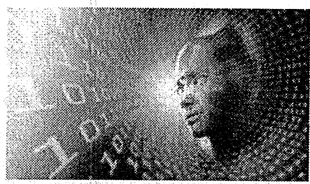
1. we set aside the j^{th} observation $\langle x_j, y_j \rangle$ from the training set,
2. we use the remaining $N-1$ observations to estimate the linear regression coefficients $\hat{\beta}^{-j}$,
3. we use $\hat{\beta}^{-j}$ to predict the target in x_j.

The leave-one-out residual is

$$e_j^{\text{loo}} = y_j - \hat{y}_j^{-j} = y_j - x_j^T \hat{\beta}^{-j}$$

The PRESS statistic is an efficient way to compute the l-o-o residuals on the basis of the simple regression performed on the whole training set. This allows a fast cross-validation without repeating N times the leave-one-out procedure. The PRESS procedure can be described as follows:

Book information



About this book
Feedback on this book

Buy a printed copy

Gianluca Bontempi
Souhaib Ben Taieb

Statistical foundations of machine learning

- Introduction
- Foundations of probability
- Classical parametric estimation
- Nonparametric estimation and testing
- Statistical supervised learning
- The machine learning procedure
- Linear approaches
  - Linear regression
  - The PRESS statistic
  - The weighted least-squares
  - Discriminant functions for classification
- Nonlinear approaches
- Model averaging approaches
- Feature selection
- Conclusions
- Bibliography
- Appendix

1. we use the whole training set to estimate the linear regression coefficients *ebe*. This procedure is performed only once on the N samples and returns as by product the Hat matrix (see Section 7.1.13)

$$H = X(X^TX)^{-1}X^T$$

2. we compute the residual vector e, whose j^{th} term is $e_j = y_j - x_j^T\hat{\beta}$,

3. we use the PRESS statistic to compute e_j^{loo} as

$$e_j^{\text{loo}} = \frac{e_j}{1 - H_{jj}}$$

where H_{jj} is the j^{th} diagonal term of the matrix H.

Note that 7.2.3 is not an approximation of 7.2.1 but simply a faster way of computing the leave-one-out residual e_j^{loo}.

Let us now derive the formula of the PRESS statistic.

Matrix manipulations show that

$$X^TX - x_jx_j^T = X_{-j}^TX_{-j}$$

where $X_{-j}^TX_{-j}$ is the X^TX matrix obtained by putting the j^{th} row aside.

Using the relation B.5.1 we have

$$\begin{aligned}(X_{-j}^TX_{-j})^{-1} &= (X^TX - x_jx_j^T)^{-1}\\ &= (X^TX)^{-1} + \frac{(X^TX)^{-1}x_jx_j^T(X^TX)^{-1}}{1 - H_{jj}}\end{aligned}$$

and

$$\begin{aligned}\hat{\beta}^{-j} &= (X_{-j}^TX_{-j})^{-1}X'_{-j}y_{-j}\\ &= \left[(X^TX)^{-1} + \frac{(X^TX)^{-1}x_jx_j^T(X^TX)^{-1}}{1 - H_{jj}}\right]X_{-j}^Ty_{-j}\end{aligned}$$

where y_{-j} is the target vector with the j^{th} sample set aside.

From 7.2.1and 7.2.6 we have

$$\begin{aligned}e_j^{\text{loo}} &= y_j - x_j^T\hat{\beta}^{-j}\\ &= y_j - x_j^T\left[(X^TX)^{-1} + \frac{(X^TX)^{-1}x_jx_j^T(X^TX)^{-1}}{1 - H_{jj}}\right]X_{-j}^Ty_{-j}\\ &= y_j - x_j^T(X^TX)^{-1}X_{-j}^Ty_{-j} - \frac{H_{jj}x_j^T(X^TX)^{-1}X_{-j}^Ty_{-j}}{1 - H_{jj}}\\ &= \frac{(1 - H_{jj})y_j - x_j^T(X^TX)^{-1}X_{-j}^Ty_{-j}}{1 - H_{jj}}\\ &= \frac{(1 - H_{jj})y_j - x_j^T(X^TX)^{-1}(X^Ty - x_jy_j)}{1 - H_{jj}}\\ &= \frac{(1 - H_{jj})y_j - \hat{y}_j + H_{jj}y_j}{1 - H_{jj}}\\ &= \frac{y_j - \hat{y}_j}{1 - H_{jj}} = \frac{e_j}{1 - H_{jj}}\end{aligned}$$

where $X_{-j}^T y_{-j} + x_j y_j = X^T y$ and $x_j^T (X^T X)^{-1} X^T y = \hat{y}_j$.

Thus, the leave-one-out estimate of the local mean integrated squared error is:

$$\hat{G}_{\text{loo}} = \frac{1}{N} \sum_{i=1}^{N} \{ \frac{y_i - \hat{y}_i}{1 - H_{ii}} \}^2$$

‹ 7.1.16 The PSE and the FPE up 7.3 The weighted least-squares ›