Leave One Out Cross Validation

Zheng Yuan

June 2019

1 PREDICTION ERROR

Consider a linear model

$$y = x^T\beta + e, \tag{1}$$

where y is a response variable, x is a predictor vector of dimension p, x^T denotes the transpose of x, β is a vector of p unknown parameters and e is a random error with mean 0 and variance σ^2. Since sometimes some of the components in β may be 0, a more compact expression would be

$$y = x_\alpha^T\beta_\alpha + e, \tag{2}$$

where α is a subset of d_α distinct positive integers that are less or equal than p and β_α is the subset of β whose components are indexed by the integers in α. Each subset α corresponds to a certain model, denoted by M_α. After knowing whether each component of β is 0 or not, all the models M_α can be classified into two categories:

- Category I: At least one nonzero component of β is not in β_α
- Category II: β_α contains all nonzero components of β

Suppose that we now split data set into two parts, $\{(y_i, x_i), i \in s\}$ and $\{(y_i, x_i), i \in s^c\}$, where s is a subset of $\{1, 2, ..., n\}$ containing n_ν integers and s^c is its complements containing n_s integers, so $n_\nu + n_s = n$. Then model M_α is fitted using $\{(y_i, x_i), i \in s^c\}$ and prediction error is evaluated using the validation data $\{(y_i, x_i), i \in s\}$.The average squared prediction error is

$$n_\nu^{-1}||y_s - \hat{y}_{\alpha,s^c}||^2 = n_\nu^{-1}||(I_{n_\nu} - Q_{\alpha,s})^{-1}(y_s - X_{\alpha,s}\hat{\beta}_\alpha||^2, \tag{3}$$

where $||u||^2 = u^Tu$ for a vector u, y_s is the n_ν vector consists of the components of y indexed by $i \in s$, $\hat{y}_{\alpha,s^c}$ is the prediction of y_s using $\{(y_i, x_i), i \in s^c\}$ by least squared method under model M_α, $Q_{\alpha,s} = X_{\alpha,s}(X_\alpha^TX_\alpha)^{-1}X_{\alpha,s}^T$,$\beta_\alpha$ is the least squares estimator of β_α using whole data set.

2 PROBABILITY OF SELECTING UNNECESSARY LARGE MODELS

2.1 Estimate prediction error by LOOCV

First, let us start from the CV(1) estimate of $\Gamma_{\alpha,n}$, the true prediction error. In fact, we can evaluate this with the help of projection matrix by a trick in algebra.
The CV(1) estimate of $\Gamma_{\alpha,n}$ can be written as

$$\begin{aligned}
\hat{\Gamma}_{\alpha,n}^{CV} &= \frac{1}{n}\sum_i [(1-w_{i\alpha})^{-1}(y_i - x_{i\alpha}^T\hat{\beta}_\alpha)]^2 \\
&= \frac{1}{n}\sum_i [1+2w_{i\alpha}+O(w_{i\alpha}^2)](y_i - x_{i\alpha}^T\hat{\beta}_\alpha)^2 \quad (By \quad (1-w_{i\alpha})^{-2} = 1+2w_{i\alpha}+O(w_{i\alpha}^2)) \\
&= \frac{1}{n}\sum_i (y_i - x_{i\alpha}^T\hat{\beta}_\alpha)^2 + \frac{1}{n}\sum_i [2w_{i\alpha}+O(w_{i\alpha}^2)](y_i - x_{i\alpha}^T\hat{\beta}_\alpha)^2 \\
&= \frac{1}{n}\sum_i r_{i\alpha}^2 + \frac{1}{n}\sum_i [2w_{i\alpha}+O(w_{i\alpha}^2)]r_{i\alpha}^2 \quad (where \quad r_{i\alpha} = y_i - x_{i\alpha}^T\hat{\beta}_\alpha) \\
&= \xi_{\alpha,n} + \phi_{\alpha,n}.
\end{aligned} \tag{4}$$

Here, we denote $\xi_{\alpha,n} = \frac{1}{n}\sum_i r_{i\alpha}^2$,$\phi_{\alpha,n} = \frac{1}{n}\sum_i [2w_{i\alpha}+O(w_{i\alpha}^2)]r_{i\alpha}^2$ and $w_{i\alpha}$ is the ith diagonal element of the projection matrix P_α.

2.2 Decomposition of $\xi_{\alpha,n}$

$$\begin{aligned}
\xi_{\alpha,n} &= \frac{1}{n}\sum_i r_{i\alpha}^2 \\
&= \frac{1}{n}(y-P_\alpha)^T(y-P_\alpha) \\
&= \frac{1}{n}y^T(I_n-P_\alpha)^T(I_n-P_\alpha)y \\
&= \frac{1}{n}(X\beta+e)^T(I_n-P_\alpha)^T(I_n-P_\alpha)(X\beta+e) \\
&= \frac{1}{n}(X\beta+e)^T(I_n-P_\alpha)(X\beta+e) \\
&= \frac{1}{n}e^T(I_n-P_\alpha)e + \frac{1}{n}\beta^T X^T(I_n-P_\alpha)X\beta + 2n^{-1}e^T(I_n-P_\alpha)X\beta \\
&= \frac{1}{n}e^T(I_n-P_\alpha)e + \triangle_{\alpha,n} + 2n^{-1}e^T(I_n-P_\alpha)X\beta.
\end{aligned} \tag{5}$$

Here, we denote $\frac{1}{n}\beta^T X^T(I_n-P_\alpha)X\beta$ as $\triangle_{\alpha,n}$ and P_α is the projection matrix (hat matrix) for M_α,i.e., $P_\alpha = X_\alpha(X_\alpha^T X_\alpha)^{-1}X_\alpha^T$
Thus, we obtain the decomposition of $\xi_{\alpha,n}$ into three parts: $\frac{1}{n}e^T(I_n-P_\alpha)e$, $\triangle_{\alpha,n}$ and $2n^{-1}e^T(I_n-P_\alpha)X\beta$.

2.3 $\xi_{\alpha,n}$ for M_α in Category II

For M_α in Category II, we have $X\beta = X_\alpha\beta_\alpha$, since β_α contains all nonzero components of β in this case.

Using this condition, now we have

$$\begin{aligned} \triangle_{\alpha,n} &= \frac{1}{n}\beta_\alpha^T X_\alpha^T (I_n - P_\alpha) X_\alpha \beta_\alpha \\ &= \frac{1}{n}\beta_\alpha^T X_\alpha^T X_\alpha \beta_\alpha - \frac{1}{n}\beta_\alpha^T X_\alpha^T P_\alpha X_\alpha \beta_\alpha, \end{aligned} \tag{6}$$

where

$$\begin{aligned} \frac{1}{n}\beta_\alpha^T X_\alpha^T P_\alpha X_\alpha \beta_\alpha &= \frac{1}{n}\beta_\alpha^T X_\alpha^T X_\alpha (X_\alpha^T X_\alpha)^{-1} X_\alpha^T X_\alpha \beta_\alpha \\ &= \frac{1}{n}\beta_\alpha^T X_\alpha^T X_\alpha \beta_\alpha \end{aligned} \tag{7}$$

Therefore,

$$\triangle_{\alpha,n} = 0. \tag{8}$$

Also,

$$\begin{aligned} \frac{1}{n}e^T(I_n - P_\alpha)X\beta &= \frac{1}{n}e^T(I_n - P_\alpha)X_\alpha\beta_\alpha \\ &= \frac{1}{n}e^T X_\alpha \beta_\alpha - \frac{1}{n}e^T P_\alpha X_\alpha \beta_\alpha \\ &= \frac{1}{n}e^T X_\alpha \beta_\alpha - \frac{1}{n}e^T X_\alpha (X_\alpha^T X_\alpha)^{-1} X_\alpha^T X_\alpha \beta_\alpha \\ &= \frac{1}{n}e^T X_\alpha \beta_\alpha - \frac{1}{n}e^T X_\alpha \beta_\alpha \\ &= 0 \end{aligned} \tag{9}$$

Therefore we have

$$\xi_{\alpha,n} = \frac{1}{n}e^T(I_n - P_\alpha)e$$

for all the models M_α in Category II.

2.4 Approximation for $\phi_{\alpha,n}$

From the notation above,

$$\begin{aligned} \phi_{\alpha,n} &= \frac{1}{n}\sum_i [2w_{i\alpha} + O(w_{i\alpha}^2)] r_{i\alpha}^2 \\ &= \frac{2}{n}\sum_i w_{i\alpha} r_{i\alpha}^2 + \frac{1}{n}\sum_i O(w_{i\alpha}^2) r_{i\alpha}^2 \end{aligned} \tag{10}$$

Under the condition

$$\lim_{n\to\infty} \max_{i\le n} \quad w_{i\alpha} = 0,$$

we have

$$\frac{1}{n}\sum_i O(w_{i\alpha}^2) r_{i\alpha}^2 = o(n^{-1})$$

On the other hand,

$$\begin{aligned}
\frac{2}{n}\sum_i w_{i\alpha} r_{i\alpha}^2 &= \frac{2}{n}\sum_i w_{i\alpha}(\sigma^2 + o(1)) \\
&= \frac{2\sigma^2}{n}\sum_i w_{i\alpha} + o(n^{-1}) \\
&= \frac{2\sigma^2}{n} tr(P_\alpha) + o(n^{-1}) \\
&= \frac{2\sigma^2}{n} tr(X_\alpha (X_\alpha^T X_\alpha)^{-1} X_\alpha^T) + o(n^{-1}) \\
&= \frac{2\sigma^2}{n} tr(X_\alpha^T X_\alpha (X_\alpha^T X_\alpha)^{-1}) + o(n^{-1}) \\
&= \frac{2\sigma^2}{n} tr(I_{d_\alpha}) + o(n^{-1}) \\
&= \frac{2}{n} d_\alpha \sigma^2 + o(n^{-1}).
\end{aligned} \tag{11}$$

Therefore,

$$\phi_{\alpha,n} = \frac{2}{n} d_\alpha \sigma^2 + o(n^{-1}), \tag{12}$$

and we obtain an approximation for $\phi_{\alpha,n}$

$$\phi_{\alpha,n} \approx \frac{2}{n} d_\alpha \sigma^2. \tag{13}$$

2.5 Two lemmas on projection matrix P_α

Lemma 1. *$\alpha \in \mathscr{A}$ is a subset of d_α distinct positive integers that are less or equal than p, β_α is the subset of β whose components are indexed by the integers in α and similar as X_α. M_α is the model corresponds to α. Let P_α be the projection matrix for M_α, i.e., $P_\alpha = X_\alpha (X_\alpha^T X_\alpha)^{-1} X_\alpha^T$.Then the eigenvalues of P_α can only be 0 or 1.*

Proof. Let λ be the eigenvalue of P_α. Then there exists $v \in R^n$ such that

$$\lambda v = P_\alpha v$$

which implies

$$P_\alpha(\lambda v) = P_\alpha(P_\alpha v)$$

Note that

$$P_\alpha(P_\alpha v) = P_\alpha v = \lambda v$$

and

$$P_\alpha(\lambda v) = \lambda(P_\alpha v) = \lambda^2 v$$

which follows

$$\lambda v = \lambda^2 v.$$

By solving this, λ can only be 0 or 1. □

Lemma 2. *Under the notations in lemma 1, if $e_0 \sim N(0, I_n)$ is a standard normal random vector, then $e_0^T P_\alpha e_0 \sim \chi^2_{d_\alpha}$, where d_α is the number of elements in α.*

Proof. First, let $\lambda_1,\ \lambda_2, ..., \lambda_n$ be the eigenvalues of P_α. Using the proposition of eigenvalues, we have

$$\sum_i \lambda i = tr(P_\alpha). \tag{14}$$

On the other hand, by the relation between matrix trace and rank, we have

$$tr(P_\alpha) = rank(P_\alpha) = d_\alpha. \tag{15}$$

Putting (14) and (15) together,

$$\sum_i \lambda i = d_\alpha. \tag{16}$$

From lemma 1, λ_i are only be 0 or 1, so there are d_α 1's and $n - d_\alpha$ 0's in λ_i's. Without loss of generality, we assume that the first d_α eigenvalues equal to 1 and the others are 0. So we obtain the eigen decomposition for P_α:
there exists an R^{n*n} matrix V satisfying $V^T V = I_n$ and

$$P_\alpha = V^T D_\alpha V, \tag{17}$$

where D_α is an R^{n*n} diagonal matrix whose first d_α elements on diagonal are 1 and the others are 0.
Now with this decomposition,

$$\begin{aligned} e_0^T P_\alpha e_0 &= e_0^T (V^T D_\alpha V) e_0 \\ &= (V e_0)^T D_\alpha (V e_0) \\ &= \sum_{i=1}^{d_\alpha} (V e_0)^2_{(i)}, \end{aligned} \tag{18}$$

where $(V e_0)_{(i)}$ denotes the ith elements of vector $V e_0$.
Note that $V e_0 \sim N(0, I_n)$, so $(V e_0)_{(i)}$'s are independent standard normal random variables, $i = 1, 2, ..., d_\alpha$. Therefore,

$$e_0^T P_\alpha e_0 = \sum_{i=1}^{d_\alpha} (V e_0)^2_{(i)} \sim \chi^2_{d_\alpha}, \tag{19}$$

□

2.6 Empirical probability of selecting an unnecessary complex model

After getting everything ready, now it's the time to put them together to prove the last theorem.

Theorem 1. *Suppose that*

$$\lim_{n\to\infty}\max_{i\leq n} \quad w_{i\alpha}=0$$

holds, where $w_{i\alpha}$ is the ith diagonal element of the projection matrix, we have the following conclusions:

(1) If M_α is in Category II, then

$$\hat{\Gamma}_{\alpha,n}=\frac{1}{n}e^Te-\frac{1}{n}e^TP_\alpha e+\frac{2}{n}d_\alpha\sigma^2+o(n^{-1}). \tag{20}$$

(2) The empirical probability of selecting model M_α is

$$Pr(M_\alpha \textit{ is preferable to } M_\star \textit{ by the } CV(1))=Pr(2k<\chi^2_k), \tag{21}$$

where k is the difference between the dimension of M_α and that of true model $M_\star$, namely, $k=d_\alpha-d_\star$, χ^2_k here denotes a chi-square random variable with k as its degree of freedom. Since we only consider the models in Category II, $k\geq 0$ always holds.

Proof. Putting all the results in 2.1-2.4 together,

$$\begin{aligned}\hat{\Gamma}_{\alpha,n}&=\xi_{\alpha,n}+\phi_{\alpha,n}\\&=\frac{1}{n}e^T(I_n-P_\alpha)e+\frac{2}{n}d_\alpha\sigma^2+o(n^{-1})\\&=\frac{1}{n}e^Te-\frac{1}{n}e^TP_\alpha e+\frac{2}{n}d_\alpha\sigma^2+o(n^{-1}),\end{aligned} \tag{22}$$

the first conclusion can be derived immediately. Now let us focus on the second conclusion.

$$\begin{aligned}&Pr(M_\alpha \textit{ is preferable to } M_\star \textit{ by the } CV(1))\\&=Pr(\hat{\Gamma}^{CV}_{\alpha,n}<\hat{\Gamma}^{CV}_{\star,n})\\&=Pr(2n^{-1}d_\alpha\sigma^2-n^{-1}e^TP_\alpha e<2n^{-1}d_\star\sigma^2-n^{-1}e^TP_\star e)\quad (\textit{by conclusion } (1))\\&=Pr(2(d_\alpha-d_\star)\sigma^2<e^T(P_\alpha-P_\star)e)\\&=Pr(2(d_\alpha-d_\star)\sigma^2<\sigma^2e_0^T(P_\alpha-P_\star)e_0)\quad (\textit{where } e=\sigma^2e_0 \textit{ and thus } e_0\sim N(0,I_n))\\&=Pr(2(d_\alpha-d_\star)<e_0^T(P_\alpha-P_\star)e_0)\end{aligned} \tag{23}$$

Note that

$$e_0^T P_\alpha e_0 \sim \chi^2_{d_\alpha}$$

and

$$e_0^T P_\star e_0 \sim \chi^2_{d_\star}$$

which follows

$$e_0^T (P_\alpha - P_\star) e_0 \sim \chi^2_{d_\alpha - d_\star} \tag{24}$$

since $e_0^T(P_\alpha - P_\star)e_0$ is independent of $e_0^T P_\star e_0$

Therefore,

$$\begin{aligned} & Pr(M_\alpha \; is \; preferable \; to \; M_\star \; by \; the \; CV(1)) \\ &= Pr(2(d_\alpha - d_\star) < \chi^2_{d_\alpha - d_\star}) \\ &= Pr(2k < \chi^2_k). \end{aligned} \tag{25}$$

□

2.7 A simple simulation study

One simple simulation verifying the theorem above is to just set the true model as the null model. In this case, we set response variable $y \sim N(3, 1)$, i.e.,

$$y = x_0 + e$$

, where $x_0 = 3$ and $e \sim N(0, \sigma^2)$. Besides, we create a redundant variable $x_1 \sim U(10, 20)$.

In each time of simulation, with the distributions above, we can generate a data set of size 1000. Then we fit a null and a one-variable linear regression respectively using the data we generate. After this, we utilize leave one out cross validation to do the model selection.

It turns out that among 1000 times simulations, there are about 16.7% of them in which leave one out cross validation selects the unnecessary large model (the one-variable model) rather than the null model, which coincides with $Pr(2 < \chi^2_2) = 0.157$.(In this case, k=1).

3 LOOCV PREDICTION ERROR DIFFERENCE TEST

3.1 Two models comparison

Suppose $M_\alpha \subset M_\lambda$ are two models in Category II, with dimensions of d_α and d_λ. Then by Theorem 1, we have decomposition of two prediction errors respectively,

$$\begin{aligned} \hat{\Gamma}_{\alpha,n} &= \frac{1}{n} e^T e - \frac{1}{n} e^T P_\alpha e + \frac{2}{n} d_\alpha \sigma^2 + o(n^{-1}) \\ \hat{\Gamma}_{\lambda,n} &= \frac{1}{n} e^T e - \frac{1}{n} e^T P_\lambda e + \frac{2}{n} d_\lambda \sigma^2 + o(n^{-1}), \end{aligned} \tag{26}$$

which gives us

$$\hat{\Gamma}_{\alpha,n} - \hat{\Gamma}_{\lambda,n} = \frac{1}{n} e^T P_\lambda e - \frac{1}{n} e^T P_\alpha e + \frac{2}{n} d_\alpha \sigma^2 - \frac{2}{n} d_\lambda \sigma^2. \tag{27}$$

After some rearranging and transformation, it shows

$$\frac{n}{\sigma^2}(\hat{\Gamma}_{\alpha,n} - \hat{\Gamma}_{\lambda,n}) - 2(d_\alpha - d_\lambda) = e_0^T (P_\lambda - P_\alpha) e_0, \tag{28}$$

or

$$(\frac{n}{\sigma^2}\hat{\Gamma}_{\alpha,n} - 2d_\alpha) - (\frac{n}{\sigma^2}\hat{\Gamma}_{\lambda,n} - 2d_\lambda) = e_0^T (P_\lambda - P_\alpha) e_0, \tag{29}$$

where $e = \sigma^2 e_0$ *and thus* $e_0 \sim N(0, I_n)$.
Similar as it is shown in the proof of Theorem 1,

$$e_0^T (P_\lambda - P_\alpha) e_0 \sim \chi^2_{d_\lambda - d_\alpha},$$

which follows

$$[(\frac{n}{\sigma^2}\hat{\Gamma}_{\alpha,n} - 2d_\alpha) - (\frac{n}{\sigma^2}\hat{\Gamma}_{\lambda,n} - 2d_\lambda)] \sim \chi^2_{d_\lambda - d_\alpha} \tag{30}$$

This provides a new way to utilize the leave one out cross validation prediction error for model selection.
Recall the procedure of likelihood ratio test. Suppose we have a set of models, usually not all of the same dimension, and want to decide which of them fits a data set best. For the Wilks test, suppose that we had an m-dimensional model H_0 included in a d-dimensional model H_1, where m < d. The maximum of the likelihood over H_1 would always be at least as large, and usually larger, than over H_0 because of the inclusion. But, if the maximum likelihood over H_0 was not too much smaller than over H_1, then in the test, H_0 is not rejected.
Here leave one out cross validation prediction error shares the similar asymptotic property with likelihood ratio test. So it is natural to consider constructing a similar test based on the difference of prediction error.
For the detailed procedure, suppose that we had an d_α-dimensional model H_0 included in a d_λ-dimensional model H_1, where $d_\alpha < d_\lambda$. The leave one out cross validation prediction error over H_1 would always be at least as small, and usually smaller, than over H_0 because of the inclusion. The key point is to detect how small it is compared to that of the null hypothesis model.
So the procedure works as following:
we reject M_α if

$$(\frac{n}{\sigma^2}\hat{\Gamma}_{\alpha,n} - 2d_\alpha) - (\frac{n}{\sigma^2}\hat{\Gamma}_{\lambda,n} - 2d_\lambda) > \chi^2_{1-\gamma}(d_\lambda - d_\alpha), \tag{31}$$

where $\chi^2_{1-\gamma}(d_\lambda - d_\alpha)$ is the $(1-\gamma)$ quantile for χ^2 random variable with degree of freedom as $(d_\lambda - d_\alpha)$, otherwise we accept the null hypothesis model or the simpler model.

3.2 The essence of the test

Cross validation, as its name indicates, selects the model with the best average predictive ability calculated based on all (or some) different ways of data splitting. It is known to many statisticians that the leave one out cross validation(the cross validation with $n = 1$ is asymptotically incorrect and is too conservative in the sense that it tends to select and unnecessarily large model.
Clearly, the computational complexity of cross validation method increases as n_ν increases. This is why the simplest cross validation with $n_\nu = 1$ has been the main focus of researchers' attention over the 30 years. Thus the main goal is try to improve the disadvantage of leave one out cross validation in order to enjoy its computational convenience, especially in the linear model case.
The traditional leave one out cross validation method is just simply to select the model with the smallest prediction error, which inevitably selects the more complex mode. But in our loocv prediction error difference test, we not only know whether one prediction error is larger than the other, but also need to determine "how large" it is compared with other one and the criterion is based on a χ square distribution. By doing this, we will finally obtain a $(1 - \gamma)$ level test, which means that we only incorrectly choose the unnecessarily large model with probability γ given the simpler model is true and that γ can be decided by ourselves.