Neural Computation **11**(4) (1999): 863-870.

# On Cross-Validation for Model Selection

**Isabelle Rivals**

**Léon Personnaz**

*Laboratoire d'Électronique, ESPCI, 10 rue Vauquelin, 75231 Paris Cedex 05, France.*

**In response to (Zhu and Rower, 1996), a recent communication (Goutte, 1997) established that leave-one-out cross-validation is not subject to the "no-free-lunch" criticism. Despite this optimistic conclusion, we show here that cross-validation has very poor performances for the selection of linear models as compared to classic statistical tests. We conclude that the statistical tests are preferable to cross-validation for linear as well as for non linear model selection.**

## 1. Introduction

Following the "no-free-lunch" theorems (Wolpert & Macready, 1995), an attempt was made in (Zhu and Rower, 1996) to demonstrate the inefficiency of leave-one-out cross-validation (LOO) on a simple problem, i.e. the problem of selecting the unbiased estimator of the expectation of a gaussian population between an unbiased and a highly biased one. A response to this attempt was given in (Goutte, 1997), where it was shown that the strict LOO procedure yields the expected results on this simple problem.

In this paper, we first give a probabilistic analysis of LOO scores. On this basis, and to complete the work done in (Goutte, 1997), we compare the selection performed by LOO between two estimators which are unbiased, but have a different variance, to that performed by statistical tests. Perspectives for non linear modeling are outlined.

## 2. Measure of model quality and leave-one-out cross-validation scores

We consider static modeling problems for the case of an input $n$-vector $\boldsymbol{x}$ and a random scalar output $y(\boldsymbol{x})$. We assume that a sample of $N$ input-output pairs $D_N = \{\boldsymbol{x}^k,\ y^k = y(\boldsymbol{x}^k)\}_{k=1 \text{ to } N}$ is available, and further that there exists an unknown regression function $\mu$ such that:

$$y\left(\boldsymbol{x}^k\right) = E\left[y\left(\boldsymbol{x}^k\right)\right] + w^k = \mu\left(\boldsymbol{x}^k\right) + w^k \tag{1}$$

where the $\left\{w^k\right\}$ are independent identically distributed (i.i.d.) random variables with zero expectation and variance $\sigma^2$ (homoscedasticity property)[1]. The problem is to find a parameterized function $f(\boldsymbol{x}, \boldsymbol{\theta}, D_N)$, $\boldsymbol{\theta} \in \mathbb{R}^q$, which is a good approximation of $\mu(\boldsymbol{x})$, and which will be denoted by $f_q^N(\boldsymbol{x})$. A natural measure of the quality of $f_q^N(\boldsymbol{x})$ as an estimator of $\mu(\boldsymbol{x})$ is the local mean-squared error (LMSE) at $\boldsymbol{x}$:

$$\begin{aligned} LMSE\left(f_q^N(\boldsymbol{x})\right) &= E\left[\left(y(\boldsymbol{x}) - f_q^N(\boldsymbol{x})\right)^2\right] \\ &= E\left[\left(y(\boldsymbol{x}) - \mu(\boldsymbol{x})\right)^2\right] + E\left[\left(f_q^N(\boldsymbol{x}) - \mu(\boldsymbol{x})\right)^2\right] \\ &= \sigma^2 + \left(E\left[f_q^N(\boldsymbol{x})\right] - \mu(\boldsymbol{x})\right)^2 + E\left[\left(f_q^N(\boldsymbol{x}) - E\left[f_q^N(\boldsymbol{x})\right]\right)^2\right] \end{aligned} \tag{2}$$

The expectations in (2) are taken over all possible samples, i.e. all possible values of the outputs for the $N$ input configurations $\left\{\boldsymbol{x}^k\right\}$. The second and third terms represent the squared bias and the variance of estimator $f_q^N(\boldsymbol{x})$, for a given input $\boldsymbol{x}$. An overall measure of the quality of $f_q^N$, its integrated mean-squared error (IMSE), is obtained by integrating bias and variance over $\boldsymbol{x}$:

$$IMSE\left(f_q^N\right) = \sigma^2 + \int \left(E\left[f_q^N(\boldsymbol{x})\right] - \mu(\boldsymbol{x})\right)^2 p(\boldsymbol{x})\, d\boldsymbol{x} + \int E\left[\left(f_q^N(\boldsymbol{x}) - E\left[f_q^N(\boldsymbol{x})\right]\right)^2\right] p(\boldsymbol{x})\, d\boldsymbol{x} \tag{3}$$

where $p(\boldsymbol{x})$ is the distribution of the inputs.

The LOO score of estimator $f_q^N$ is an empirical IMSE (Efron & Tibshirani, 1993) (Goutte, 1997):

$$s_{LOO}\left(f_q^N\right) = \frac{1}{N}\sum_{j=1}^{N}\left(y\left(\boldsymbol{x}^j\right) - f\left(\boldsymbol{x}^j, \boldsymbol{\theta}, \left\{\boldsymbol{x}^k, y^k\right\}_{\substack{k=1 \text{ to } N \\ k\neq j}}\right)\right)^2 \tag{4}$$

$s_{LOO}\left(f_q^N\right)$ is an estimator of $IMSE\left(f_q^N\right)$. We characterize the bias of this estimator in the following sections.

### 3. Leave-one-out cross-validation for the selection among estimators of a constant

The problem of (Goutte, 1997) is to select between two estimators of the expectation of a gaussian population ($\mu$ = constant, $\sigma^2 = 1$, $N = 16$) using LOO. When the output expectation does not depend on an external input, it is easily shown that the LOO score of any estimator $f_1^N$ is biased[2]:

$$E\left[s_{LOO}\left(f_1^N\right)\right] = IMSE\left(f_1^{N-1}\right) \neq IMSE\left(f_1^N\right) \tag{5}$$

The estimators considered in (Goutte, 1997) are the unbiased sample mean, and the highly biased sample maximum; their IMSE are very different (cf. Table 1). Thus, even if the LOO scores of the

---

[1] Scalars are denoted by lowercase letters, e.g. $y$ and the $\left\{y^k\right\}$; vectors are denoted by boldface lowercase letters, e.g. the $n$-vectors $\boldsymbol{x}$ and the $\left\{\boldsymbol{x}^k\right\}$; matrices are denoted by uppercase letters, e.g. the input matrix $X$ (see section 4).

[2] This result can be generalized to $m$-fold cross-validation ($m$ divides $N$): $E\left[s_{m\text{-}fold\,CV}\left(f_1^N\right)\right] = IMSE\left(f_1^{N-\frac{N}{m}}\right)$.

mean and of the maximum are biased, their bias is small as compared to the difference between the two IMSE, and the LOO procedure always selects the mean. Table 1 gives the IMSE of the estimators, and the bias of their LOO scores. The bias of the LOO score of the mean equals $\frac{\sigma^2}{N\,(N-1)}$; for the computation of the expectation and variance of the maximum estimator, see for example [Pugatchev, 1982].

| $f_1^{16}$ | $IMSE\left(f_1^{16}\right)$ | $E\left[s_{LOO}\left(f_1^{16}\right)\right] - IMSE\left(f_1^{16}\right)$ |
|---|---|---|
| mean | 1.0625 | $4.1667\ 10^{-3}$ |
| maximum | 4.4137 | $-\ 9.9298\ 10^{-2}$ |

Table 1. IMSE and bias of the LOO scores of the sample mean and maximum ($\mu$ = constant, $\sigma^2 = 1$, $N = 16$).

To conclude, (Goutte, 1997) only showed that LOO does not make a wrong choice for a trivial problem where any reasonable method would not make the wrong choice either. In more realistic settings of linear or nonlinear process modeling however, it is often necessary to select an estimator among several estimators of decreasing complexity (estimators linear in the parameters with a decreasing number of parameters, such as polynomials or radial basis functions, or neural networks with a decreasing number of hidden neurons). If the selection concerns unbiased estimators with different variances, it is important to be able to select the estimator with the smallest variance: this is exactly the problem solved by statistical tests. We therefore tackle the model selection problem in the next section, and, in section 5, we give an illustration which leads to more pessimistic conclusions about LOO than the preceding example.

**4. Leave-one-out cross-validation versus statistical tests for the selection of linear models**

We deal with the particular case of *linear* static modeling problems, i.e. there exists an unknown parameter $n$-vector $\boldsymbol{\theta}_0$ such that the regression function can be written as:

$$\mu\,(\boldsymbol{x}) = \boldsymbol{x}^T\ \boldsymbol{\theta}_0 \tag{6}$$

We consider $f_n^N\,(\boldsymbol{x},\ \boldsymbol{\theta}_{LS},\ D_N)$, the least squares (LS) estimator of the regression, denoted by $f_n^N\,(\boldsymbol{x})$:

$$f_n^N\,(\boldsymbol{x}) = \boldsymbol{x}^T\,\boldsymbol{\theta}_{LS} = \boldsymbol{x}^T\left(X^T X\right)^{-1}\,X^T\,\boldsymbol{y} \tag{7}$$

where $\boldsymbol{y} = \left[y^1\ y^2\ \ldots\ y^N\right]^T$, $\boldsymbol{x}^k = \left[x_1^k\ x_2^k\ \ldots\ x_n^k\right]^T$, and $X = \left[\boldsymbol{x}^1\ \boldsymbol{x}^2\ \ldots\ \boldsymbol{x}^N\right]^T$ is the $(N, n)$ input matrix whose columns are assumed to be linearly independent. The estimator $f_n^N(\boldsymbol{x})$ is unbiased, and its LMSE is[3]:

$$LMSE\left(f_n^N(\boldsymbol{x})\right) = \sigma^2 + \sigma^2\ \boldsymbol{x}^T \left(X^T X\right)^{-1}\ \boldsymbol{x} \tag{8}$$

The IMSE of $f_n^N$ thus equals:

$$IMSE\left(f_n^N\right) = \sigma^2 \left(1 + E\left(trace\left(\boldsymbol{x}\ \boldsymbol{x}^T \left(X^T X\right)^{-1}\right)\right)\right) \tag{9}$$

Let us make the weak assumption that the components of the input vector are uncorrelated, with covariance $K(\boldsymbol{x}) = \sigma_x^2\ I_n$; then:

$$IMSE\left(f_n^N\right) = \sigma^2 \left(1 + \sigma_x^2\ trace\left(\left(X^T X\right)^{-1}\right)\right) \tag{10}$$

The inputs of the data set being drawn from the same distribution, and in order to have a simple expression for (10), we will consider the case where:

$$X^T X = N\ \sigma_x^2\ I_n \tag{11}$$

i.e. the case where the $n$ columns of $X$ (regressor vectors) are orthogonal and $\forall k$, $\sum_{i=1}^{N} \left(x_i^k\right)^2 = N\ \sigma_x^2$. We have then:

$$IMSE\left(f_n^N\right) = \sigma^2 \left(1 + \frac{n}{N}\right) \tag{12}$$

The IMSE depends only on the variance of the noise, on the size of the training set, and on the number of parameters.

Let us now consider the expectation of the LOO score[4] of $f_n^N$. We obtain:

$$E\left[s_{LOO}\left(f_n^N\right)\right] = \sigma^2 \left(1 + \frac{1}{N} \sum_{k=1}^{N} \frac{P_{kk}}{1 - P_{kk}}\right) > \sigma^2 \left(1 + \frac{n}{N}\right) \tag{13}$$

where the $\{P_{kk}\}$ are the diagonal elements of $P = X\left(X^T X\right)^{-1}\ X^T$, the orthogonal projection matrix[5] on the range of $X$. The LOO score is thus a *biased* estimator of (12).

Suppose that we want to choose between (6) and a submodel of (6) with $n' < n$ inputs, i.e. we want to decide whether:

$$\boldsymbol{\theta}_0 = \begin{bmatrix} \boldsymbol{\theta}_0' \\ \mathbf{0} \end{bmatrix} \tag{14}$$

---

[3] Note that expression (8) cannot be used to compute the expectation of the square of the residuals; their expectation is: $E\left[\left(r^k\right)^2\right] = \sigma^2 - \sigma^2 \left(\boldsymbol{x}^k\right)^T \left(X^T X\right)^{-1}\ \boldsymbol{x}^k \quad k = 1$ to $N$.

[4] The LOO error is extensively analyzed in (Antoniadis & al., 1992), and briefly in (Efron & Tibshirani, 1993).

[5] Properties of the $(N, N)$ projection matrix $P$, with $\text{rank}(P) = n$: a) $\sum_{k=1}^{N} P_{kk} = n$; b) $0 \le P_{kk} \le 1 \quad k = 1$ to $N$ .

where $\boldsymbol{\theta}_0'$ is a *n'*-vector. If the null hypothesis (14) is true, then the variance of $f_{n'}^N$ is smaller than that of $f_n^N$, and thus $IMSE\left(f_{n'}^N\right) < IMSE\left(f_n^N\right)$. But, since the LOO score is a biased estimator of the IMSE, it is likely that LOO will not lead to a correct choice.

By comparison, a statistical test is based on *unbiased* estimations of the noise variance $\sigma^2$ through the residuals of both models when (14) holds. If the null hypothesis is true, and if the gaussian assumption can be made, a Fisher variable can be constructed. The decision to reject the null hypothesis with a risk $\alpha\%$ of rejecting it while it is true will be taken when:

$$\frac{RSS^{\,2} - RSS'^{\,2}}{RSS^{\,2}} \frac{N-n}{n-n'} > F_{N-n}^{n-n'}(\alpha\%) \tag{15}$$

where $RSS^2$ and $RSS'^2$ denote the values of the residual sums of squares of the estimators, and $F_{N-n}^{n-n'}(\alpha\%)$ is the value for which the Fisher cumulative distribution with $n - n'$ and $N–n$ degrees of freedom equals $1–\alpha$ .

**5. Illustrative example**

We consider the modeling of simulated processes:

$$y^k = \begin{bmatrix} 1 & x^k \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} + w^k = a_0 + b_0\, x^k + w^k \quad k=1 \text{ to } N \tag{16}$$

with 1) $a_0 = 1, b_0 = 0,$ or 2) $a_0 = 1, b_0 = 1,$ and for different values of $N$. We want to choose between the two following estimators, or models:

$$f_1^N(x) = a_1^N \; (n' = 1) \tag{17}$$

$$f_2^N(x) = a_2^N + b_2^N x \;\; (n = 2) \tag{18}$$

where $a_1^N$, $a_2^N$ and $b_2^N$ denote the LS estimators of the parameters ($a_1^N$ is the sample mean). For each sample size $N$, we choose equally spaced inputs $\{x^k\}$ such that $X^T X = N I_2$ (according to relation (11) with $\sigma_x^2 = 1$, and inputs thus in roughly $[-\sqrt{3}; \sqrt{3}]$); a million samples, i.e. outputs $\{y^k\}$, are simulated. The selection between $f_1^N$ and $f_2^N$ is performed on each sample with LOO and with statistical tests.

1) We first consider the process with $a_0 = 1$, $b_0 = 0$: both estimators $f_1^N$ and $f_2^N$ are unbiased, but $f_1^N$ has a smaller variance. Almost by definition, the frequency over a million samples of the rejection of the null hypothesis reaches the risk taken, as shown in Table 2. The frequency of selection by LOO of the large model $f_2^N$ decreases with $N$, due to the decrease of the bias of the LOO scores (see Table 3, where the biases are computed with expressions (12) and (13) involving the values $\{P_{kk}\}$).

But with 16% of selection of $f_2^N$ for very large $N$, LOO still performs poorly as compared to a statistical test. These results do not vary with the value of $\sigma^2$.

| | Frequency of wrong selection of $f_2^N(x) = a_2^N + b_2^N x$ against $f_1^N(x) = a_1^N$ obtained over $10^6$ samples. | |
|---|---|---|
| $N$ | LOO | Test with risk 1% / 5% |
| 10 | 19.7 % | 1.0 % / 5.0 % |
| 20 | 17.8 % | 1.0 % / 5 .0 % |
| 30 | 17.0 % | 1.0 % / 5.0 % |
| 100 | 16.1 % | 1.0 % / 5.0 % |
| 1000 | 15.8 % | 1.0 % / 5.0 % |

Table 2. Selection using LOO versus statistical tests (process $y^k = 1 + w^k$ $k$=1 to $N$ ).

| $N$ | $\frac{IMSE\left(f_2^N\right)}{\sigma^2}$ | $\frac{E\left[s_{LOO}\left(f_2^N\right)\right] - IMSE\left(f_2^N\right)}{\sigma^2}$ | $\frac{IMSE\left(f_1^N\right)}{\sigma^2}$ | $\frac{E\left[s_{LOO}\left(f_1^N\right)\right] - IMSE\left(f_1^N\right)}{\sigma^2}$ |
|---|---|---|---|---|
| 10 | 1.200 | $6.65\ 10^{-2}$ | 1.100 | $1.11\ 10^{-2}$ |
| 100 | 1.020 | $4.94\ 10^{-4}$ | 1.010 | $1.01\ 10^{-4}$ |
| 1000 | 1.002 | $4.81\ 10^{-6}$ | 1.001 | $1.00\ 10^{-6}$ |

Table 3. IMSE and bias of the LOO scores (process $y^k = 1 + w^k$ $k$=1 to $N$ ).

2) We next consider the process with $a_0 = 1$, $b_0 = 1$: estimator $f_1^N$ is biased. As in (Goutte, 1997), since one of the estimators has a large bias, LOO and the tests always select the unbiased estimator, provided $\sigma^2 \leq 0.3$ (for larger values of $\sigma^2$ and small $N$, the signal-to-noise ratio becomes very large, and model (16) becomes meaningless with the numerical values we have chosen).

***Numerical results in the general case***

We have considered the particular case where the input matrix is chosen according to (11) since the LOO bias can be calculated in this case. Nevertheless, the results of Table 2 are almost the same when the inputs $\{x^k\}$ are different for each simulated data set, uniformly chosen in $[-\sqrt{3}; \sqrt{3}]$. We then obtain the following percentages of the wrong selection of the large model: 19.9% ($N = 10$), 16.1% ($N = 100$), 15.8% ($N = 1000$).

Risk is relevant in statistical tests, but it is important to stress that there is no notion of risk in the choice of the model with the smallest LOO score. Thus, even in cases where the LOO score might be unbiased, this procedure leads frequently to inappropriate decisions.

## 6. Conclusion

In the linear case, even for large $N$, LOO does not perform well as compared to statistical tests. Furthermore, when $N$ is large, the gaussian hypothesis is no longer necessary for a statistical test to be valid; there is then no advantage in performing LOO.

Since even for linear estimators, LOO performs poorly for small $N$, it is extremely unlikely that it would perform better in the case of nonlinear estimators, like neural networks. Furthermore, LOO becomes very time-consuming or even untractable for large $N$ since it requires (at least) $N$ nonlinear optimizations. Also, when $N$ is large, the curvature of the expectation surface of a non linear model becomes small (Seber, 1989) (Antoniadis & al., 1992), thus statistical tests similar to those for linear models can be performed successfully, by assuming only homoscedasticity.

We draw the conclusion that, even though LOO is not subject to the "no-free-lunch" criticism as pointed out in (Goutte, 1997), statistical tests are strongly preferred to LOO, provided that the (linear or nonlinear) model has the properties required for the statistical tests to be valid.


## Acknowledgements

We thank Howard Gutowitz, whose insightful comments improved the clarity of this paper.


## References


Antoniadis A., Berruyer J., Carmona R. (1992). *Régression non linéaire et applications*, Economica.

Efron B., Tibshirani R. J. (1993). *An introduction to the bootstrap*, Chapman & Hall.

Goutte C. (1997). Note on free lunches and cross-validation. *Neural Computation* ***9*** (6), 1245-1249.

Pugatchev, V. (1982). *Théorie des probabilités et statistiques*, Éditions Mir.

Seber G. A. F., Wild C. J. (1989). *Nonlinear regression*, John Wiley & Sons.

Wolpert D. H., Macready, W. G. (1995). *The mathematics of search* (Tech. Rep. No. SFI-TR-95-02-010). Santa Fe: Santa Fe Institute.

Zhu H., Rohwer R. (1996). No free lunch for cross validation. *Neural Computation* ***8*** (7), 1421-1426.