Journal of Mathematical Chemistry Vol. 29, No. 3, 2001

Fast computation of cross-validated properties in full linear leave-many-out procedures

Emili Besalú

Institute of Computational Chemistry and Department of Chemistry, University of Girona, Catalonia, Spain
E-mail: emili@iqc.udg.es

Received 14 February 2001

This article is dedicated to Josep Martí, a scientist, a teacher, and a friend.

A general theorem which allows the fast and direct computation of predicted properties in a full multiple linear leave-many-out procedure is demonstrated by induction. The result allows the description of a general algorithm which only requires a single multiple linear regression calculation. From the data generated by this fitting, in a full leave-n-out procedure involving a set of m objects, the resolution of $\binom{m}{n}$ linear systems of equations of dimension $n \times n$ suffices to obtain all the sets of cross-validated properties.

KEY WORDS: leave-one-out, leave-many-out, leave-n-out theorem, cross-validation, multiple linear regression

1. Introduction

From early times it was necessary to describe numerical procedures in order to assess the confidence of prediction results arising from QSAR studies [1–9] or similar problems in correlation analysis. The most widely used techniques are related to the cross-validation paradigm. It is assumed by the scientific community that a result obtained by cross-validation possess some intrinsic robustness and even more if a Leave-Many-Out protocol has been considered [10]. In recent works, there are a wide range of systematic studies whose results are based on the well-known Leave-One-Out (LOO) protocol. Usually, the processes are based on Multiple Linear Regressions (MLRs) [11–19], Partial Least Squares or Principal Components [20–40] (generally coming from a CoMFA [41] study) and even on non-linear techniques as Artificial Neural Networks [42–45]. Some studies deal with other methodologies, as for example those based on taxonomy [46].

When QSAR models are obtained using MLRs, some practical advantages arise. If m molecules are represented by d descriptors, it is well known that in a linear LOO procedure it is not necessary to perform m fittings of dimension $(n-1) \times d$ [47]. This confers special features to the linear LOO procedure, but it is also claimed that the

0259-9791/01/0400-0191$19.50/0

related parameters overestimate the predictive model capabilities [48,49]. Leave-Many-Out protocols will be more adequate in order to obtain significant and optimal results. In fact, the formulation attached to a full linear LOO procedure can be generalized in order to obtain explicit expressions for the cross-validated properties which will be obtained when n molecules are being separated form the original group of m. That is, there is a *general* multiple linear Leave-n-Out (LnO) procedure formulation. This paper is restricted to the statement and demonstration of the related theorem.

2. The leave-n-out theorem (LnOT)

Consider a set of m molecules which are described by a set of $d < m - n$ independent descriptors ($0 < n < m$) collected in a rectangular matrix $\boldsymbol{X}$ of dimension $m \times d$. It is assumed that this matrix contains a constant column in order to allow the presence of a constant term in the linear model which will be constructed. Also, it is supposed that the molecular family has an attached vector of properties or dependent parameters, $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)^{\mathrm{T}}$. If a MLR fit is performed to the full data, some well-known algebraic parameters are obtained: the $m \times m$ predictions matrix (the so-called "hat" matrix),

$$\boldsymbol{H} = \{h_{ij}\} = \boldsymbol{X}\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^{\mathrm{T}}, \tag{1}$$

the vector of linear coefficients,

$$\boldsymbol{c} = \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \tag{2}$$

and the dependent values fitted by the linear model,

$$\boldsymbol{y}' = \left(y'_1, y'_2, \ldots, y'_m\right)^{\mathrm{T}} = \boldsymbol{X}\boldsymbol{c} = \boldsymbol{H}\boldsymbol{y}. \tag{3}$$

The predictions matrix is symmetric and the corresponding elements can be computed as

$$h_{ij} = h_{ji} = \boldsymbol{x}_i^{\mathrm{T}}\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}\boldsymbol{x}_j, \quad \forall i, j = 1, 2, \ldots, m, \tag{4}$$

where the terms $\boldsymbol{x}_i$ are column vectors, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^{\mathrm{T}}$, collecting the original independent descriptors attached to the ith molecule coming from the related row in matrix $\boldsymbol{X}$. Hence, form the geometrical point of view, the predictions matrix constitutes the Gram matrix of the set of vectors $\{\boldsymbol{x}_i\}$ in the space of metric $\boldsymbol{S} = \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}$.

Suppose that *n distinct* molecules are selected from the original data in order to build the cross-validated set $M^{(n)} = \{m_1, m_2, \ldots, m_n\}$. Consider the following system of linear equations:

$$\left\{\begin{pmatrix} h_{m_1m_1} & h_{m_1m_2} & \ldots & h_{m_1m_n} \\ h_{m_2m_1} & h_{m_2m_2} & \ldots & h_{m_2m_n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{m_nm_1} & h_{m_nm_2} & \ldots & h_{m_nm_n} \end{pmatrix} - \boldsymbol{I}_n\right\}\begin{pmatrix} \widehat{y}_{m_1}^{(n)} \\ \widehat{y}_{m_2}^{(n)} \\ \vdots \\ \widehat{y}_{m_n}^{(n)} \end{pmatrix} = \begin{pmatrix} t_{m_1}^{(n)} \\ t_{m_2}^{(n)} \\ \vdots \\ t_{m_n}^{(n)} \end{pmatrix}, \tag{5}$$

or

$$\left(\boldsymbol{H}^{(n)} - \boldsymbol{I}_n\right)\widehat{\boldsymbol{y}}^{(n)} = \boldsymbol{t}^{(n)}, \tag{6}$$

where $\boldsymbol{I}_n$ stands for the $n \times n$ unit matrix and the other definitions are obvious when comparing equations (5) and (6). The $t_{m_i}^{(n)}$ terms are defined by

$$t_{m_i}^{(n)} = -\sum_{j \notin M^{(n)}} h_{m_i j}\, y_j = \sum_{j \in M^{(n)}} h_{m_i m_j}\, y_{m_j} - y'_{m_i}, \quad \forall m_i \in M^{(n)}. \tag{7}$$

They correspond to the fitted value (3) attached to the molecule m_i and where the information relative to all the selected molecules in the set $M^{(n)}$ has been removed. The superscript used in (5)–(7) is used to explicitly show that the elements are related to a selected LnO procedure. Such a notation will be extensively used from now on.

The LnOT states that the solution vector of (5) gives the corresponding n property values which would be predicted if the set of n molecules had not been taken into account from the beginning and consequently did not enter into the MLR model construction. In other words, the solution vector $\widehat{\boldsymbol{y}}^{(n)}$ contains the corresponding *cross-validated* property values.

A full LnO procedure will be performed whenever equation (5) is solved for all the $\binom{m}{n}$ distinct choices of the cross-validation set $M^{(n)}$. The main idea underneath the LnOT allows to express it in another way:

Leave-*n*-Out theorem. A full linear LnO procedure involving a set of m objects represented by d independent descriptors can be performed by solving a *unique* $m \times d$ MLR and a set of $\binom{m}{n}$ linear systems of dimension $n \times n$.

The theorem states that in a full linear LnO procedure it is not necessary to perform $\binom{m}{n}$ MLRs of order $(m - n) \times d$. This result is specially useful in QSAR studies due to the fact that n is usually small compared to m.

3. Proof of the L*n*OT

The LnOT can be proved by induction. The main leading ideas will be given in this section and partial proofs and relevant details are described in the appendices below. This presentation structure has been chosen in order to achieve a clearer exposition. The two steps of an inductive proof follow.

3.1. The L*n*OT *applies for* $n = 1$.

When $n = 1$, the cross-validated set reduces to $M^{(1)} = \{m_1\}$ and system (5) returns a simple formula:

$$\widehat{y}_{m_1}^{(1)} = \frac{t_{m_1}^{(1)}}{h_{m_1 m_1} - 1}. \tag{8}$$

As it is demonstrated in appendix A, this corresponds to the LOO expression for a single molecule, the one labeled as m_1.

3.2. Assuming that the LnOT *applies for a problem of dimension n it also applies for a problem of dimension* $n+1$.

The explicit formulas for the LnOT are symmetric and homogeneous with respect to any molecule of the cross-validation set $M^{(n)}$. Then, the theorem can be proved without loss of generality for a particular molecule, the one labeled m_1, say. In other words, we will study how to obtain the cross-validated property value for an arbitrary molecule m_1 when the cross-validation set involves n and $n+1$ molecules. From equation (5), its cross-validated value arises from Cramer's rule:

$$\widehat{y}_{m_1}^{(n)} = \frac{T_n^{(n)}}{\Delta_n^{(n)}}, \tag{9}$$

where

$$\Delta_n^{(n)} = \begin{vmatrix} h_{m_1m_1} - 1 & h_{m_1m_2} & \dots & h_{m_1m_n} \\ h_{m_2m_1} & h_{m_2m_2} - 1 & \dots & h_{m_2m_n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{m_nm_1} & h_{m_nm_2} & \dots & h_{m_nm_n} - 1 \end{vmatrix} \tag{10}$$

and

$$T_n^{(n)} = \begin{vmatrix} t_{m_1}^{(n)} & h_{m_1m_2} & \dots & h_{m_1m_n} \\ t_{m_2}^{(n)} & h_{m_2m_2} - 1 & \dots & h_{m_2m_n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{m_n}^{(n)} & h_{m_nm_2} & \dots & h_{m_nm_n} - 1 \end{vmatrix}. \tag{11}$$

Expression (9) also applies when the set $M^{(n)}$ is expanded to contain $n+1$ *distinct* molecules and thus becomes $M^{(n+1)} = M^{(n)} \cup \{m_{n+1}\}$. The cross-validated property value for the molecule m_1 belonging to the set $M^{(n+1)}$ corresponds to a result of the leave-$(n+1)$-out procedure. This value can be computed using the expression (9) relative to a LnO test, but *hiding* the data of molecule number $n+1$. The process of data hiding consists into virtually set to zero the property value $y_{m_{n+1}}$ and the corresponding row in matrix $\boldsymbol{X}$, that is, to set $\boldsymbol{x}_{m_{n+1}} = (x_{m_{n+1}1}, x_{m_{n+1}2}, \dots, x_{m_{n+1}d})^{\mathrm{T}} = \mathbf{0}$. In fact, this corresponds to the design of a LnO calculation but using a *null* or *phantom* molecule entering into the linear model construction. Once the MLR coefficients are obtained and the linear model is applied over the cross-validated molecular set $M^{(n+1)}$, the predicted values exactly coincide with the ones coming from a leave-$(n+1)$-out test. As a consequence of these considerations expression (9) can be written as:

$$\widehat{y}_{m_1}^{(n+1)} = \frac{T_n^{(n+1)}}{\Delta_n^{(n+1)}} \tag{12}$$

in order to stress that the equation system is of order $n \times n$ but the result is equivalent to the one obtained in a leave-$(n+1)$-out procedure.

The kind of data manipulation described above produces the following changes over the numerical data employed [47]:

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \to \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{x}_{m_{n+1}}\boldsymbol{x}_{m_{n+1}}^{\mathrm{T}}$$

and it is straightforward to check that the following relationship holds [47]:

$$\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{x}_{m_{n+1}}\boldsymbol{x}_{m_{n+1}}^{\mathrm{T}}\right]^{-1} = \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1} + \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}\frac{\boldsymbol{x}_{m_{n+1}}\boldsymbol{x}_{m_{n+1}}^{\mathrm{T}}}{1 - h_{m_{n+1}m_{n+1}}}\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}. \tag{13}$$

Then, according to equation (4), this result can be used in a subsequent transformation of the predictive matrix elements:

$$\begin{aligned} h_{ij} &\to \boldsymbol{x}_i^{\mathrm{T}}\left[\boldsymbol{X}^{T}\boldsymbol{X} - \boldsymbol{x}_{m_{n+1}}\boldsymbol{x}_{m_{n+1}}^{\mathrm{T}}\right]^{-1}\boldsymbol{x}_j \\ &= \boldsymbol{x}_i^{\mathrm{T}}\left\{\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1} + \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}\frac{\boldsymbol{x}_{m_{n+1}}\boldsymbol{x}_{m_{n+1}}^{\mathrm{T}}}{1 - h_{m_{n+1}m_{n+1}}}\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}\right\}\boldsymbol{x}_j \\ &= \boldsymbol{x}_i^{\mathrm{T}}\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}\boldsymbol{x}_j + \boldsymbol{x}_i^{\mathrm{T}}\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}\frac{\boldsymbol{x}_{m_{n+1}}\boldsymbol{x}_{m_{n+1}}^{\mathrm{T}}}{1 - h_{m_{n+1}m_{n+1}}}\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}\boldsymbol{x}_j \end{aligned} \tag{14}$$

and, using equation (4) again, the transformation in equation (14) can be written in a better way as

$$h_{ij} \to h_{ij} + \frac{h_{m_{n+1}j}}{1 - h_{m_{n+1}m_{n+1}}}h_{i\,m_{n+1}}, \quad \forall i, j = 1, 2, \ldots, m. \tag{15}$$

Such a transformation conserves the symmetry of the new predictions matrix which has been generated by this procedure.

Simultaneously, the terms defined in equation (7) transform into

$$\begin{aligned} t_{m_i}^{(n)} &\to -\sum_{j \notin M^{(n)}}\left(h_{m_i j} + \frac{h_{m_i m_{n+1}}h_{m_{n+1}j}}{1 - h_{m_{n+1}m_{n+1}}}\right)y_j \\ &= -\sum_{j \notin M^{(n)}} h_{m_i j}y_j - \frac{h_{m_i m_{n+1}}}{1 - h_{m_{n+1}m_{n+1}}}\sum_{j \notin M^{(n)}} h_{m_{n+1}j}y_j. \end{aligned} \tag{16}$$

But, as now the value $y_{m_{n+1}} = 0$ holds, it can be written

$$t_{m_i}^{(n)} \to -\sum_{j \notin M^{(n+1)}} h_{m_i j}y_j - \frac{h_{m_i m_{n+1}}}{1 - h_{m_{n+1}m_{n+1}}}\sum_{j \notin M^{(n+1)}} h_{m_{n+1}j}y_j, \tag{17}$$

and the transformation (16) is finally expressed as

$$t_{m_i}^{(n)} \to t_{m_i}^{(n+1)} + \frac{t_{m_{n+1}}^{(n+1)}}{1 - h_{m_{n+1}m_{n+1}}}h_{m_i m_{n+1}}, \quad \forall m_i \in M^{(n)}. \tag{18}$$

According to the transformations (15) and (18), the determinants appearing in equation (12) bear the following structures:

$$\Delta_n^{(n+1)} = \begin{vmatrix} h_{m_1m_1} - 1 + \frac{h_{m_{n+1}m_1}}{1-h_{m_{n+1}m_{n+1}}} h_{m_1m_{n+1}} & h_{m_1m_2} + \frac{h_{m_{n+1}m_2}}{1-h_{m_{n+1}m_{n+1}}} h_{m_1m_{n+1}} & \cdots \\ h_{m_2m_1} + \frac{h_{m_{n+1}m_1}}{1-h_{m_{n+1}m_{n+1}}} h_{m_2m_{n+1}} & h_{m_2m_2} - 1 + \frac{h_{m_{n+1}m_2}}{1-h_{m_{n+1}m_{n+1}}} h_{m_2m_{n+1}} & \cdots \\ \vdots & \vdots & \ddots \\ h_{m_nm_1} + \frac{h_{m_{n+1}m_1}}{1-h_{m_{n+1}m_{n+1}}} h_{m_nm_{n+1}} & h_{m_nm_2} + \frac{h_{m_{n+1}m_2}}{1-h_{m_{n+1}m_{n+1}}} h_{m_nm_{n+1}} & \cdots \end{vmatrix}$$

$$\begin{vmatrix} h_{m_1m_n} + \frac{h_{m_{n+1}m_n}}{1-h_{m_{n+1}m_{n+1}}} h_{m_1m_{n+1}} \\ h_{m_2m_n} + \frac{h_{m_{n+1}m_n}}{1-h_{m_{n+1}m_{n+1}}} h_{m_2m_{n+1}} \\ \vdots \\ h_{m_nm_n} - 1 + \frac{h_{m_{n+1}m_n}}{1-h_{m_{n+1}m_{n+1}}} h_{m_nm_{n+1}} \end{vmatrix} \tag{19}$$

and

$$T_n^{(n+1)} = \begin{vmatrix} t_{m_1}^{(n+1)} + \frac{t_{m_{n+1}}^{(n+1)}}{1-h_{m_{n+1}m_{n+1}}} h_{m_1m_{n+1}} & h_{m_1m_2} + \frac{h_{m_{n+1}m_2}}{1-h_{m_{n+1}m_{n+1}}} h_{m_1m_{n+1}} & \cdots \\ t_{m_2}^{(n+1)} + \frac{t_{m_{n+1}}^{(n+1)}}{1-h_{m_{n+1}m_{n+1}}} h_{m_2m_{n+1}} & h_{m_2m_2} - 1 + \frac{h_{m_{n+1}m_2}}{1-h_{m_{n+1}m_{n+1}}} h_{m_2m_{n+1}} & \cdots \\ \vdots & \vdots & \ddots \\ t_{m_n}^{(n+1)} + \frac{t_{m_{n+1}}^{(n+1)}}{1-h_{m_{n+1}m_{n+1}}} h_{m_nm_{n+1}} & h_{m_nm_2} + \frac{h_{m_{n+1}m_2}}{1-h_{m_{n+1}m_{n+1}}} h_{m_nm_{n+1}} & \cdots \end{vmatrix}$$

$$\begin{vmatrix} h_{m_1m_n} + \frac{h_{m_{n+1}m_n}}{1-h_{m_{n+1}m_{n+1}}} h_{m_1m_{n+1}} \\ h_{m_2m_n} + \frac{h_{m_{n+1}m_n}}{1-h_{m_{n+1}m_{n+1}}} h_{m_2m_{n+1}} \\ \vdots \\ h_{m_nm_n} - 1 + \frac{h_{m_{n+1}m_n}}{1-h_{m_{n+1}m_{n+1}}} h_{m_nm_{n+1}} \end{vmatrix}. \tag{20}$$

In appendix B it is shown that

$$(h_{m_{n+1}m_{n+1}} - 1)\Delta_n^{(n+1)} = \Delta_{n+1}^{(n+1)} \tag{21}$$

and, similarly, in appendix C it is also shown that a similar relationship holds for $T_n^{(n+1)}$:

$$(h_{m_{n+1}m_{n+1}} - 1)T_n^{(n+1)} = T_{n+1}^{(n+1)}. \tag{22}$$

In other words, the $n \times n$ determinants (19) and (20), properly multiplied by the same term, originate the ones which are homologous to (10) and (11) but extended to dimension $(n+1) \times (n+1)$. In this way, expression (9) is obtained again but expanded to the next dimension:

$$\widehat{y}_{m_1}^{(n+1)} = \frac{T_{n+1}^{(n+1)}}{\Delta_{n+1}^{(n+1)}}. \tag{23}$$

Similar expressions as (23) hold for *any cross-validated molecule* belonging to the set $M^{(n)}$. The predicted value for the hidden molecule $m_{n+1} \in M^{(n+1)}$ can be also obtained in the same manner because the attached data has no effect into the linear model construction. Thus, the relationship (5) also applies when n is replaced by $n+1$ and the LnOT is demonstrated.

Of course, the recursive substitution we are dealing with must be stopped: the expression (5) is valid for all the values of n ranging from 1 up to $m - d$. In a QSAR study, only those values of n which are small compared to m will generate useful data. Moreover, it would be only possible to obtain reliable results when all the descriptors are linearly independent, even in the MLR fitting procedure and when the cross-validation sets are being generated and system (5) is solved.

4. General algorithm and explicit expressions for particular cases

From the previous results, a practical and fast way to obtain predictions coming from a full linear LnO procedure is envisaged. The general algorithm to be followed is systematic and very simple:

1. Given the matrix $\boldsymbol{X}$, compute the predictions matrix $\boldsymbol{H}$ in equation (1).
2. Given the vector $\boldsymbol{y}$, compute the coefficient vector $\boldsymbol{c}$ in equation (2).
3. Obtain the vector of fitted data $\boldsymbol{y}'$ using equation (3).
4. Fix n, the number of leaving molecules.
5. Loop over all the molecular subsets $M^{(n)}$. For every subset:
   5.1. Solve the linear system (5).
   5.2. Keep the predicted values.

In order to optimize the algorithm presented above, the system of equations (5) can be explicitly solved for a particular value of n and then apply the specific obtained equations in the algorithm step number 5.1. The explicit formulas for $n = 1$, 2 and 3 are given next.

($n = 1$) In a LOO, equation (8) applies.

($n = 2$) In a leave-two-out, for a pair of molecules $M^{(2)} = \{m_1, m_2\}$, the mathematical formulas reduce to the two following practical terms:

$$\widehat{y}_{m_1}^{(2)} = \frac{1}{\Delta_2^{(2)}}\left[t_{m_1}^{(2)}\left(h_{m_2m_2} - 1\right) - t_{m_2}^{(2)}h_{m_1m_2}\right],$$
$$\widehat{y}_{m_2}^{(2)} = \frac{1}{\Delta_2^{(2)}}\left[t_{m_2}^{(2)}\left(h_{m_1m_1} - 1\right) - t_{m_1}^{(2)}h_{m_1m_2}\right],$$

where

$$\Delta_2^{(2)} = (h_{m_1m_1} - 1)(h_{m_2m_2} - 1) - h_{m_1m_2}^2.$$

($n = 3$) In a leave-three-out procedure, if $M^{(3)} = \{m_1, m_2, m_3\}$, the relevant mathematical formulas are

$$\Delta_3^{(3)} = (h_{m_1m_1} - 1)(h_{m_2m_2} - 1)(h_{m_3m_3} - 1) - h_{m_2m_3}^2(h_{m_1m_1} - 1) - h_{m_1m_3}^2(h_{m_2m_2} - 1) - h_{m_1m_2}^2(h_{m_3m_3} - 1) + 2h_{m_1m_2}h_{m_1m_3}h_{m_2m_3}$$

and

$$\widehat{y}_{m_1}^{(3)} = \frac{1}{\Delta_3^{(3)}}\left\{t_{m_1}^{(3)}\left[(h_{m_2m_2} - 1)(h_{m_3m_3} - 1) - h_{m_2m_3}^2\right] + t_{m_2}^{(3)}\left[h_{m_1m_3}h_{m_2m_3} - h_{m_1m_2}(h_{m_3m_3} - 1)\right] + t_{m_3}^{(3)}\left[h_{m_1m_2}h_{m_2m_3} - h_{m_1m_3}(h_{m_2m_2} - 1)\right]\right\}.$$

This last equation must be repeated performing two cyclic substitutions: $\{m_1 \to m_2, m_2 \to m_3, m_3 \to m_1\}$ and $\{m_1 \to m_3, m_2 \to m_1, m_3 \to m_2\}$.

Other sets of solutions attached to higher values of n can be obtained from symbolic mathematical programs, such as Mathematica [50].

At the end of the algorithm, every molecule has an attached pool of $\binom{m-1}{n-1}$ predicted (*cross-validated*) values. It is recommendable then to perform a statistical study of the generated LnO data. A possible option consists into fitting every data subset to a Gaussian distribution. If the fitting is acceptable, the expected value of the Gaussian distribution (which will almost coincide with the mean of the data subset) can be taken as the representative for the cross-validated value of the respective molecule.

The correlation coefficient found between cross-validated representatives and the experimental property values gives the r_{cv} statistic. From the same data it can be also computed the $q^{(2)}$ term [51–55], nevertheless it seems preferable to use the r_{cv}^2 statistic [56]. Also, apart of the standard statistical tests [57], it is also recommended to consider alternative methods for the evaluation of the statistical significance of the obtained correlation [8].

5. Application to the search of an optimal set of descriptors

In QSAR studies it is very common to search for an optimal set of representative descriptors. The algorithm described in previous section can be conceived as to be a tool which gives a statistic, r_{cv}, attached to a set of them, those defining the matrix $\boldsymbol{X}$. It is obvious that the algorithm can be repeated as many times as sets of descriptors are tested in a QSAR project. Hence, the whole procedure furnishes a new method for selecting the best set of descriptors according to the criteria of maximization of the value r_{cv}. This will be presumably the most immediate utility of the LnOT. In this case, in order to speed up the process, it is not only recommended to implement the practical expressions outlined in previous section but also not to perform the gaussian fitting described above. It is faster to obtain the correlation coefficients directly from the set of arithmetic mean values attached to every series of molecular cross-validated values.

6. Conclusions

The leave-n-out theorem has been demonstrated. As a consequence, general and explicit expressions attached to full linear leave-many-out cross-validation processes have been given. The formulation will allow to easily construct computer codes oriented to be applied to QSAR problems and other fields.

Acknowledgements

Some comments and remarks of Prof. R. Carbó-Dorca are greatly acknowledged. Thanks are given to Prof. J. Martí for checking some equations with Mathematica. The CICYT Research Project SAF2000-0223-01, the Fundació Maria Francisca de Roviralta and a University of Girona grant for the investigation project (session date 3/00) are also acknowledged.

Appendix A. Demonstration of the LOO expression

The present demonstration constitutes a variant of the one which can be found in reference [47] related to the $q^{(2)}$ statistic calculation. Consider that the parameters of equations (1)–(3) are known for the full molecular data set. In order to obtain the cross-validated value for an arbitrary molecule, m_1, it is necessary to hide in equations (1) and (2) the related information as it is explained in the main body of this paper. In order to virtually set to zero the data of molecule m_1, two transformations have to be considered:

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \rightarrow \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{x}_{m_1}\boldsymbol{x}_{m_1}^{\mathrm{T}} \tag{A.1}$$

and

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \rightarrow \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{x}_{m_1}y_1. \tag{A.2}$$

Then, following a similar notation as employed in equation (2), the coefficients of the linear model become

$$\boldsymbol{c} \to \boldsymbol{c}_{m_1}^{(1)} = \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{x}_{m_1}\boldsymbol{x}_{m_1}^{\mathrm{T}}\right]^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{x}_{m_1}y_1\right). \tag{A.3}$$

By a simple application of the model to the data, the vector derived in equation (A.3) allows the computation of the LOO cross-validated property value for the molecule number m_1:

$$\widehat{y}_{m_1}^{(1)} = \boldsymbol{x}_{m_1}^{\mathrm{T}}\boldsymbol{c}_{m_1}^{(1)}. \tag{A.4}$$

Equation (13) can be rewritten now as

$$\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{x}_{m_1}\boldsymbol{x}_{m_1}^{\mathrm{T}}\right]^{-1} = \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1} + \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}\frac{\boldsymbol{x}_{m_1}\boldsymbol{x}_{m_1}^{\mathrm{T}}}{1 - h_{m_1m_1}}\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}. \tag{A.5}$$

From equations (26)–(28) and definition (4), it is straightforward to obtain

$$\widehat{y}_{m_1}^{(1)} = \frac{h_{m_1m_1}y_{m_1} - \boldsymbol{x}_{m_1}^{\mathrm{T}}\boldsymbol{c}}{h_{m_1m_1} - 1}. \tag{A.6}$$

In this expression, the numerator corresponds to the term $t_{m_1}^{(1)}$ defined in equation (7) and the whole quotient can be identified as expression (8).

Appendix B. Demonstration of equation (21)

The determinant (19) can be interpreted as to be the one appearing in equation (10) but with a term $h_{m_{n+1}m_j}h_{m_im_{n+1}}/(1 - h_{m_{n+1}m_{n+1}})$ $(i, j = 1, 2, \ldots, n)$ added to every element. In this way, every column is split into two terms. Using the property of the determinants consisting into that the linear combinations among the columns expand the corresponding linear combinations among the determinants, the term (19) originates the sum of 2^n determinants. These 2^n terms can be classified into three kinds:

- First, there is only one determinant collecting all the original columns appearing in expression (10).
- On the other hand, there are exactly n determinants bearing $n - 1$ of the original columns and a different one, the column number j $(j = 1, 2, \ldots, n)$, having the structure of the following vector:

$$\frac{h_{m_{n+1}m_j}}{1 - h_{m_{n+1}m_{n+1}}}\boldsymbol{h}_{m_{n+1}}, \tag{B.1}$$

  where

$$\boldsymbol{h}_{m_{n+1}} = (h_{m_1m_{n+1}}, h_{m_2m_{n+1}}, \ldots, h_{m_nm_{n+1}})^{\mathrm{T}}. \tag{B.2}$$

- Finally, in all the remaining $2^n - (n + 1)$ determinants, at least two different columns have the structure of expression (B.1). But once common factors are

extracted, both columns became equal to expression (B.2). Hence, the whole set of $2^n - (n+1)$ determinants vanish.

In this way, determinant (19) expands only a sum of $n+1$ selected determinants. One can symbolize this sum as follows:

$$\Delta_n^{(n+1)} = \left|\bar{1}, 2, 3, \ldots, n\right| + \left|1, \bar{2}, 3, \ldots, n\right| + \cdots + \left|1, 2, 3, \ldots, \bar{n}\right| + |1, 2, 3, \ldots, n|, \tag{B.3}$$

where the numbers specify columns and the bar used in the first n terms indicates that the selected column bears the same structure as expression (B.1). Then, in each of the first n determinants a common scalar, $h_{m_{n+1}m_j}$ $(j = 1, 2, \ldots, n)$ is found. Multiplying the whole expression (B.3) by the term $h_{m_{n+1}m_{n+1}} - 1$ then

$$\begin{aligned}(h_{m_{n+1}m_{n+1}} - 1)\Delta_n^{(n+1)} = &-h_{m_{n+1}m_1}|\boldsymbol{h}_{m_{n+1}}, 2, 3, \ldots, n| - h_{m_{n+1}m_2}\left|1, \boldsymbol{h}_{m_{n+1}}, 3, \ldots, n\right| \\ &- \cdots - h_{m_{n+1}m_n}|1, 2, 3, \ldots, n-1, \boldsymbol{h}_{m_{n+1}}| \\ &+ (h_{m_{n+1}m_{n+1}} - 1)|1, 2, 3, \ldots, n|. \end{aligned} \tag{B.4}$$

In all the $n \times n$ determinants appearing in (B.4), the column ordering almost corresponds to the one appearing in the following $(n+1) \times (n+1)$ determinant:

$$\Delta_{n+1}^{(n+1)} = \begin{vmatrix} h_{m_1m_1} - 1 & h_{m_1m_2} & \ldots & h_{m_1m_{n+1}} \\ h_{m_2m_1} & h_{m_2m_2} - 1 & \ldots & h_{m_2m_{n+1}} \\ \vdots & \vdots & \ddots & \vdots \\ h_{m_{n+1}m_1} & h_{m_{n+1}m_2} & \ldots & h_{m_{n+1}m_{n+1}} - 1 \end{vmatrix}. \tag{B.5}$$

The exception is found only in one column placed on every one of the first n determinants. This selected column corresponds to equation (B.2) and, according to the structure in the determinant (B.5), it must be moved to the rightmost position. In each case, this column translation requires $n - j$ swaps. Once these permutations are performed, an additional sign equal to $(-1)^{n-j}$ shall be attached to every term. Consequently, expression (B.4) can be conveniently written now as

$$\begin{aligned}(h_{m_{n+1}m_{n+1}} - 1)\Delta_n^{(n+1)} = (-1)^n \big\{&(-1)^0 h_{m_{n+1}1}|2, 3 \ldots, n, \boldsymbol{h}_{m_{n+1}}| \\ &+ (-1)^1 h_{m_{n+1}2}|1, 3, \ldots, n, \boldsymbol{h}_{m_{n+1}}| \\ &+ \cdots + (-1)^{n-1} h_{m_{n+1}m_n}|1, 2, 3, \ldots, n-1, \boldsymbol{h}_{m_{n+1}}| \\ &+ (-1)^n (h_{m_{n+1}m_{n+1}} - 1)|1, 2, 3, \ldots, n|\big\}. \end{aligned} \tag{B.6}$$

Such expression coincides with the Laplace development [58] of the determinant (B.5) by its last row. The terms $\{h_{m_{n+1}m_j}\}$ $(j = 1, 2, \ldots, n)$ and $h_{m_{n+1}m_{n+1}} - 1$ can be identified as the 1×1 minors. In this way, equation (21) is demonstrated.

Appendix C. Demonstration of equation (22)

As the demonstration for a particular molecule is performed here, in the first column of equation (20) appears the transformed term according to equation (18). If

the demonstration is carried out for another molecule $m_i \in M^{(n)}$, this column structure would be reproduced in the corresponding place.

As transformations (15) and (18) bear the same structure, also the columns of determinants (20) and (19) do. When considering the first column, the terms $h_{m_{n+1}m_1}$ must be replaced here by the constant factor $t^{(n+1)}_{m_{n+1}}$. For this column, and only for this column, this implies a particular redefinition of equation (B.1). Apart from this characteristic, equation (22) demonstration is analogous to the previous one as outlined in appendix B. From the expansion of determinant (20) among the linear combinations of the columns, a sum of 2^n determinants arises. For the same reason as in the previous case of appendix B, only $n + 1$ are, in general, different from zero. This sum, once multiplied by the term $h_{m_{n+1}m_{n+1}} - 1$, corresponds to the Laplace expansion of the determinant

$$T^{(n+1)}_{n+1} = \begin{vmatrix} t^{(n+1)}_{m_1} & h_{m_1m_2} & \dots & h_{m_1m_{n+1}} \\ t^{(n+1)}_{m_2} & h_{m_2m_2} - 1 & \dots & h_{m_2m_{n+1}} \\ \vdots & \vdots & \ddots & \vdots \\ t^{(n+1)}_{m_{n+1}} & h_{m_{n+1}m_2} & \dots & h_{m_{n+1}m_{n+1}} - 1 \end{vmatrix} \tag{C.1}$$

by its last row.

References

[1] S. Wold, *Quant. Struct.-Act. Relat.* 10 (1991) 191–193.
[2] D.M. Allen, *Technometrics* 16 (1974) 125–127.
[3] S. Wold, *Technometrics* 20 (1978) 397–405.
[4] M.A. Pleiss and S.H. Unger, in: *Quantitative Drug Design*, ed. C.A. Ramsden, Comprehensive Medicinal Chemistry, Vol. 4; *The Design of Tests Series and the Significance of QSAR Relationship*, ed. C. Hansch (Pergamon Press, Oxford, 1990).
[5] M. Clark and R.D. Cramer III, Quant. Struct.-Act. Relat. 12 (1993) 137–145.
[6] R.D. Cramer III, J.D. Bunce, D.E. Patterson and I. Frank, Quant. Struc.-Act. Relat. 7 (1988) 18–25.
[7] S. Clementi and S. Wold, How to choose the proper statistical method, in: *Chemometric Methods in Molecular Design*, ed. H. van de Waterbeemd (VCH, Weinheim, 1995).
[8] J. Pecka and R. Ponec, J. Math. Chem. 27 (2000) 13–22.
[9] G. Klopman and A.N. Kalos, J. Comput. Chem. 5 (1985) 492–506.
[10] P. Constans and J.D. Hirst, J. Chem. Inf. Comput. Sci. 40 (2000) 452–459.
[11] D. Robert, L. Amat and R. Carbó-Dorca, J. Chem. Inf. Comput. Sci. 39 (1999) 333–344.
[12] A.R. Katritzky, S. Sild and M. Karelson, J. Chem. Inf. Comput. Sci. 38 (1998) 840–844.
[13] V. Nguyen-Cong and B.M. Rode, J. Chem. Inf. Comput. Sci. 36 (1996) 114–117.
[14] J.M. Luco and F.H. Ferretti, J. Chem. Inf. Comput. Sci. 37 (1997) 392–401.
[15] D. Amić, D. Davidović-Amić, D. Bešlo, B. Lucić and N. Trinajstić, J. Chem. Inf. Comput. Sci. 37 (1997) 581–586.
[16] J.D. Gough and L.H. Hall, J. Chem. Inf. Comput. Sci. 39 (1999) 356–361.
[17] M. Lobato, L. Amat, E. Besalú and R. Carbó-Dorca, Quant. Struct.-Act. Relat. 16 (1997) 465–472.
[18] R. Carbó-Dorca, L. Amat, E. Besalú, X. Gironés and D. Robert, Quantum molecular similarity: theory and applications to the evaluation of molecular properties, biological activities and toxicity, in: *The Fundamentals of Molecular Similarity* eds. R. Carbó-Dorca, X. Gironés and P.G. Mezey, (Kluwer Academic, Dordrecht, 2000).

[19] M. Pompe and M. Nović, J. Chem. Inf. Comput. Sci. 39 (1999) 59–67.
[20] O. Ivanciuc, J. Chem. Inf. Comput. Sci. 40 (2000) 1412–1422.
[21] T.J. Hou, Z.M. Li, Z. Li, J. Liu and X.J. Xu, J. Chem. Inf. Comput. Sci. 40 (2000) 1002–1009.
[22] L.M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R.M. Blair, W.S. Branham, S.L. Dial, C.L. Moland and D.M. Sheehan, J. Chem. Inf. Comput. Sci. 41(1) (2001) 186–195.
[23] W. Tong, D.R. Lowis, R. Perkins, Y. Chen, W.J. Welsh, D.W. Goddette, T.W. Heritage and D.M. Sheehan, J. Chem. Inf. Comput. Sci. 38 (1998) 669–677.
[24] B. Hoffman, S.J. Cho, W. Zheng, S. Wyrick, D.E. Nichols, R.B. Mailman and A. Tropsha, J. Med. Chem. 42 (1999) 3217–3226.
[25] E.U. Ramos, W.H.J. Vaes, H.J.M. Verhaar and J.L.M. Hermens, J. Chem. Inf. Comput. Sci. 38 (1998) 845–852.
[26] M.G. Albuquerque, A.J. Hopfinger, E.J. Barreiro and R.B. de Alencastro, J. Chem. Inf. Comput. Sci. 38 (1998) 925–938.
[27] T. Sulea, T.I. Oprea, S. Muresan and S.L. Chan, J. Chem. Inf. Comput. Sci. 37 (1997) 1162–1170.
[28] S. Rao, R. Aoyama, M. Schrag, W.F. Trager, A. Rettie and J.P. Jones, J. Med. Chem. 43 (2000) 2789–2796.
[29] B.R. Sadler, S.J. Cho, K.S. Ishaq, K. Chae and K.S. Korach, J. Med. Chem. 41 (1998) 2261–2267.
[30] S.S. Kulkarni and V.M. Kulkarni, J. Chem. Inf. Comput. Sci. 39 (1999) 1128–1140.
[31] A.K. Debnath, J. Chem. Inf. Comput. Sci. 38 (1998) 761–767.
[32] S.J. Cho, A. Tropsha, M. Suffness, Y.-C. Cheng and K.-H. Lee, J. Med. Chem. 39 (1996) 1383–1395.
[33] D. Robert, X. Gironés and R. Carbó-Dorca, Polycycl. Aromat. Comp. (ISPAC17) (in press).
[34] A. Gallegos, D. Robert, X. Gironés and R. Carbó-Dorca, J. Comput.-Aided Molec. Design (in press).
[35] X. Gironés, A. Gallegos and R. Carbó-Dorca, J. Chem. Inf. Comput. Sci. 40 (2000) 1400–1407.
[36] X. Gironés, A. Gallegos and R. Carbó-Dorca, J. Comput.-Aided Molec. Design (submitted).
[37] D. Robert, X. Gironés and R. Carbó-Dorca, J. Chem. Inf. Comput. Sci. 40 (2000) 839–846.
[38] X. Gironés, L. Amat, R. Carbó-Dorca and D. Robert, J. Comput.-Aided Molec. Design 14 (2000) 477–485.
[39] X. Gironés, L. Amat and R. Carbó-Dorca, SAR & QSAR Envir. Research 10 (1999) 545–556.
[40] C.L. Waller and M.P. Bradley, J. Chem. Inf. Comput. Sci. 39 (1999) 345–355.
[41] R.D. Cramer III, D.E. Patterson and J.D. Bunce, J. Am. Chem. Soc. 110 (1988) 5959.
[42] S.-S. So, S.P. van Helden, V.J. van Geerestein, and M. Karplus, J. Chem. Inf. Comput. Sci. 40 (2000) 762–772.
[43] M. Nović, Z. Nikolovska-Coleska and T. Solmajer, J. Chem. Inf. Comput. Sci. 37 (1997) 990–998.
[44] I.V. Tetko, A.E.P. Villa and D.J. Livingstone, J. Chem. Inf. Comput. Sci. 36 (1996) 794–803.
[45] A.F. Duprat, T. Huynh and G. Dreyfus, J. Chem. Inf. Comput. Sci. 38 (1998) 586–594.
[46] W. Zheng and A. Tropsha, J. Chem. Inf. Comput. Sci. 40 (2000) 185–194.
[47] D.C. Montgomery and E.A. Peck, *Introduction to Linear Regression Analysis* (Wiley, New York, 1992).
[48] S.-S. So, S.P. van Helden, V.J. van Geerestein and M. Karplus, J. Chem. Inf. Comput. Sci. 40 (2000) 762–772.
[49] J. Shao, J. Am. Stat. Assoc. 88 (1993) 486–494.
[50] *Mathematica 2.2 for Windows* (Wolfram Research, Champaign, IL, 1993).
[51] H. Kubinyi, U. Abraham, Practical problems in PLS analyses, in: *3D QSAR in Drug Design*, ed. H. Kubinyi (ESCOM, Leiden, 1993) Appendix B.
[52] M. Baroni, G. Constantino, G. Cruciani, D. Riganelli, R. Valigi and S. Clementi, Quant. Struct.-Act. Relat. 12 (1993) 9–20.
[53] S. Clementi, G. Cruciani, D. Riganelli and R. Valigi, GOLPE: Merits and drawbacks in 3D-QSAR, in: *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*, eds. F. Sanz, J. Giraldo and F. Manaut (Prous Pub., Barcelona, 1995).

[54] H. van de Waterbeemd, Chemometric methods used in drug discovery, in: *Structure-Property Correlations in Drug Research*, ed. H. van de WaterBeemd (Academic Press, San Diego, CA, 1996).
[55] R.D. Cramer III, S.A. Depriest, D.E. Patterson and P. Hecht, The developing practice of comparative molecular field analysis, in: *3D QSAR in Drug Design*, ed. H. Kubinyi (ESCOM, Leiden, 1993).
[56] L. Amat, E. Besalú, R. Carbó-Dorca and R. Ponec, J. Chem. Inf. Comput. Sci. (submitted).
[57] L. Sachs, *Applied Statistics. A Handbook of Techniques*, Springer Series in Statistics (Springer, New York, 1982).
[58] I.M. Vinogradov (ed.), *Encyclopaedia of Mathematics* (Reidel-Kluwer Academic Publishers, Dordrecht, 1987).