ELSEVIER

Analytica Chimica Acta 476 (2003) 73–84

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

The Law of Mixtures method for multivariate calibration

L. Jin [a,b], J.A. Fernández Pierna [a], F. Wahl [c], P. Dardenne [d], D.L. Massart [a,*]

[a] *ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussels, Belgium*
[b] *Institute of Applied Chemistry, Nanchang University, 330047 Nanchang, PR China*
[c] *Institut Français du Pétrole (IFP), BP3, 69390 Vernaison, France*
[d] *Centre de Recherches Agronomiques de Gembloux, Département Qualité des Productions Agricoles,*
Chaussée de Namur 24, B-5030 Gembloux, Belgium

Received 28 May 2002; received in revised form 25 September 2002; accepted 8 October 2002

Abstract

The Law of Mixtures (LM) method is a new so-called topological method for multivariate calibration. It is shown to be a very good method to predict the response for new objects that are inside the convex hull determined by the calibration data set. A method is also proposed for those that are outside the convex hull.

© 2002 Elsevier Science B.V. All rights reserved.

Keywords: The Law of Mixtures (LM) method; Convex hull

1. Principle

The proposed Law of Mixtures (LM) method can be considered as a Nearest Neighbors method [1] or as what is sometimes called a topological model [2]. In this method, originally proposed by the *Institut Français du Pétrole* (IFP), the unknown sample (M) is first surrounded in the calibration data set space by a selection of k neighbors. In a second step, M is considered as a mixture of these k samples in order to calculate the value of its associated property using the Law of Mixtures.

In the example of Fig. 1, two explained variables, x_1 and x_2, describe the set of samples and three samples define a mixture. In that way all the samples in the triangle [M_1 M_2 M_3] can be expressed as a mixture of samples M_1, M_2 and M_3.

In Fig. 1, if M is a sample with an unknown response, one can write:

$$x_{1M} = \alpha_{M_1}x_{1M_1} + \alpha_{M_2}x_{1M_2} + \alpha_{M_3}x_{1M_3} \qquad (1)$$

$$x_{2M} = \alpha_{M_1}x_{2M_1} + \alpha_{M_2}x_{2M_2} + \alpha_{M_3}x_{2M_3} \qquad (2)$$

where α_{M1}, α_{M2} and α_{M3}, are the contribution of components M_1, M_2 and M_3, respectively in the mixture and x_{1i} and x_{2i} are the coordinates of each sample in the explained variables. The α-values have to satisfy:

$$\alpha_{M_1}, \alpha_{M_2}, \alpha_{M_3} > 0 \quad \text{and} \quad \alpha_{M_1} + \alpha_{M_2} + \alpha_{M_3} = 1 \qquad (3)$$

In the case we work with petroleum data for example, properties like aromatic carbon, hydrogen and aromatics show a linear Law of Mixtures. If y_M, y_{M1}, y_{M2}, and y_{M3} are the values of the property for samples M, M_1, M_2 and M_3, respectively:

$$y_M = \alpha_{M_1}y_{M_1} + \alpha_{M_2}y_{M_2} + \alpha_{M_3}y_{M_3}. \qquad (4)$$

The generalisation to more dimensions is immediate.

[*] Corresponding author. Tel.: +32-2-477-4734; fax: +32-2-477-4735.
E-mail address: fabi@vub.ac.be (D.L. Massart).

0003-2670/02/$ – see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0003-2670(02)01259-X

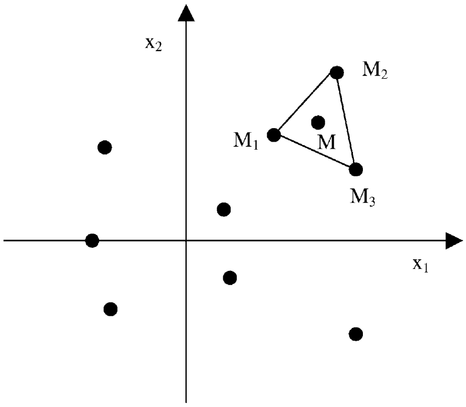

Fig. 1. x_1 and x_2 describe the set of samples. Sample M is considered a mixture of samples M_1, M_2 and M_3.

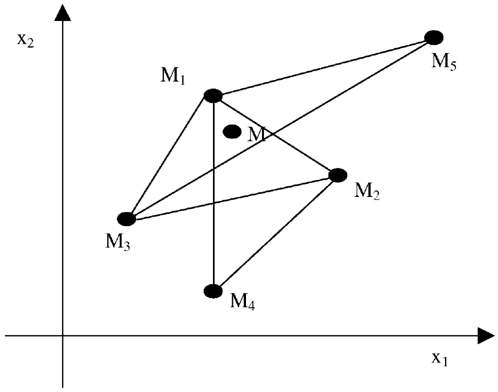

Fig. 2. In two dimensions several triangles can contain the sample M.

For other properties, it is sometimes possible to apply a transformation in order to find a linear relationship. For example, in the case of the density in petroleum products the inverse of the density follows a linear law:

$$\frac{1}{d} = \sum_{i=1}^{I} \alpha_i \frac{1}{d_i} \tag{5}$$

Sometimes the calculation of the mixture is not based on analytical expressions as in the case of the distillation.

2. Methodology

The LM method requires two steps. In the first step, a lattice is built in the PC space in order to be able to locate an unknown sample. With this aim an algorithm is developed which is explained below with an example in two dimensions. Then, in the second step, the prediction of the unknown sample is performed.

2.1. First step: building the lattice

In a classical k-Nearest Neighbours method one usually starts by choosing the number k of neighbours to predict or to classify an unknown sample M. In the LM method, this number k is equal to the number of dimensions $(n) + 1$ and the points are chosen such that the sample M is surrounded with the k selected points. This means that if $(\mathrm{M}_i)_{i=1,k}$ are the selected points, it is possible to find k coefficients $(\alpha_i)_{i=1,k}$ between 0 and 1 in such a way that $\mathrm{M} = \sum_{i=1,k} \alpha_i \mathrm{M}_i$ and $\sum_{i=1,k} \alpha_i = 1$.

In one dimension, two points are needed to surround an unknown sample. In two dimensions, at least three points, and in n dimensions, $n + 1$ points are needed.

In a given set of points, several groups of points can satisfy the previous criteria as is shown in Fig. 2, which shows an example in a two-dimensional space with mixtures of three components. Of course, mixtures of four components or more can also be considered.

Instead of calculating each time all the possible mixtures, it is more natural to fix from the beginning all the mixtures by building a lattice. In fact, there are $C_n^r = n!/(n-r)!r!$ mixtures with r elements to test in a data set with n samples and it becomes impossible to test all of them. Apart from the computational advantage, the geometry of the lattice yields predictions that are unique.

To build the lattice, we developed an algorithm in order to automatically cover the n-dimensional space with possible mixtures of $n+1$ samples. The tests that we performed give very good results. The obtained mixtures are $n + 1$ polyhedra with $n + 1$ vertices, where n is the number of independent variables, n represents the dimensionality of the calibration sample space. To build the lattice, one proceeds in the way illustrated by Fig. 3a–d for the case of two dimensions. The dimensions are selected as the most important PCs according to their correlation to y.

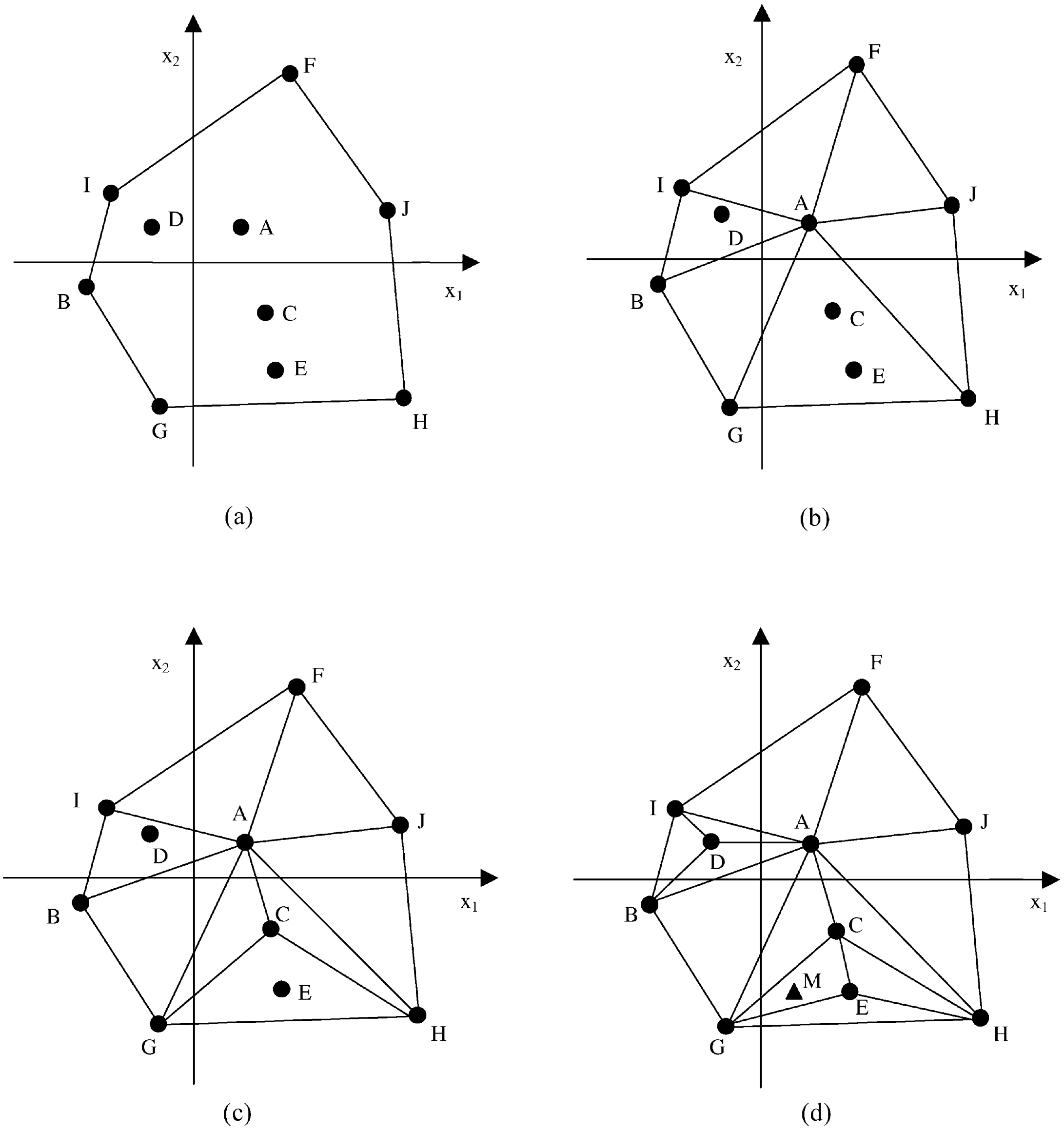

Fig. 3. (a–c) Steps to construct the lattice in two dimensions in the calibration set. (d) A new point M can be considered as a mixture of samples C, E and G.

(a) First a convex hull is constructed that envelops the data points [3,4].[1]

(b) One point is arbitrarily selected as an initial point. This point is linked to all the points in the convex hull, in such a way that a first lattice is built. Empirical tests show that good results are obtained when the selected point is the closest point to the gravity center (point A).

(c) An iterative procedure is carried out. Consider iteration *i*. Point *i* is found not to be a vertex of the *n*-polyhedron (point D in figure). To incorporate it in the lattice, one has to link it to all the ver-

[1] http://www.cse.unsw.edu.au/~lambert/java/3d/giftwrap.html.

tices of the n-polyhedron (triangle) that contains it.

Step (c) is repeated for all the objects inside the convex hull. A possible alternative (which in our case does not produce better results) is to repeat the steps (b) and (c) for each mesh in the lattice until there is no point left, that is not part of the lattice.

The difficulty of step (a) is to construct the convex hull. Step (b) is evident. In step (c) it is necessary to find the polyhedron which contains the new point i.

Several refinements are possible:

- A different initial point can be selected.
- The points in step (c) can be selected in a different way. For instance, for each n-polyhedron the gravity centre can be determined and step (b) is iterated for the whole data set. The tests that we performed did not show improved results and the algorithms are more complex.

2.2. *Second step: prediction of new samples*

For a new point M, the polyhedron (triangle for two dimensions) that contains it is found. This point M is considered as a mixture of the samples corresponding to the vertices of that polyhedron. The LM method is applied in order to determine the coefficients α of the mixture. From Eqs. (1)–(3), the α-values in two dimensions are defined as:

$$\alpha_{M_1}=\frac{(x_{2M}-x_{2M_2})(x_{1M_3}-x_{1M_2})+(x_{1M}-x_{1M_2})(x_{2M_2}-x_{2M_3})}{(x_{2M_1}-x_{2M_2})(x_{1M_3}-x_{1M_2})+(x_{1M_1}-x_{1M_2})(x_{2M_2}-x_{2M_3})} \quad (6)$$

$$\alpha_{M_2}=\frac{(x_{2M}-x_{2M_1})(x_{1M_3}-x_{1M_1})+(x_{1M}-x_{1M_1})(x_{2M_1}-x_{2M_3})}{(x_{2M_2}-x_{2M_1})(x_{1M_3}-x_{1M_1})+(x_{1M_2}-x_{1M_1})(x_{2M_1}-x_{2M_3})} \quad (7)$$

$$\alpha_{M_3}=\frac{(x_{2M}-x_{2M_2})(x_{1M_1}-x_{1M_2})+(x_{1M}-x_{1M_2})(x_{2M_2}-x_{2M_1})}{(x_{2M_3}-x_{2M_2})(x_{1M_1}-x_{1M_2})+(x_{1M_3}-x_{1M_2})(x_{2M_2}-x_{2M_1})} \quad (8)$$

For instance, in the case of Fig. 3d, these three equations are applied using point C as M_1, point E as M_2 and point G as M_3. These three points are the vertices of the polyhedron containing M and the value of the property for a new sample M can be obtained using Eq. (4).

The method of the Law of Mixtures is essentially an approximation method of the unknown function f that connects the property under study. In one dimension, the method works as an interpolation method of a function using several consecutive linear functions, as is shown in Fig. 4.

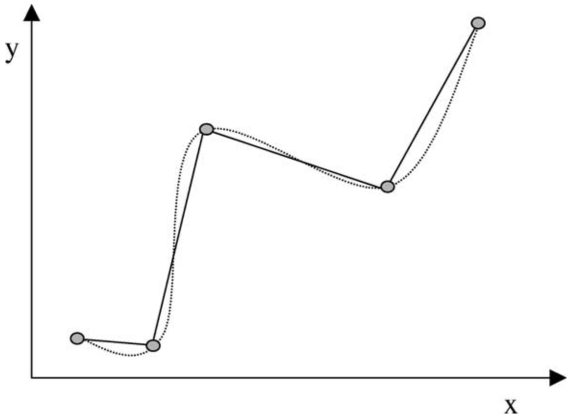

Fig. 4. Interpolation in one dimension using several consecutive linear functions.

It is possible to evaluate the approximation performed in each point. Suppose that a point M to be predicted is surrounded in the n-dimensional space with a polyhedron with $n+1$ vertices $M_1, M_2, \ldots, M_{n+1}$. $(x^i_j)_{j=1,n}$ are the co-ordinates of point M_i, $(x_j)_{j=1,n}$ the co-ordinates of point M, and $(\alpha_i)_{i=1,n+1}$ are the coefficients of the mixture.

We know from Eq. (4) that $\hat{y}=\sum_{i=1,n+1}\alpha_i y_i$ and we will replace in this equation each y by its value from the Taylor series.

$$y_i = y+\sum_j (x^i_j-x_j)\frac{\partial y}{\partial x_j} +\frac{1}{2}\sum_j (x^i_j-x_j)^2\frac{\partial^2 y}{\partial {x_j}^2}+\cdots \quad (9)$$

By substitution into the equation for $\hat{y}$, we obtain:

$$\hat{y} = y\left(\sum \alpha_i\right) + \sum_j \frac{\partial y}{\partial x_j}\left(\sum_i \alpha_i (x_j^i - x_j)\right) + \frac{1}{2}\sum_j \frac{\partial^2 y}{\partial x_j^2}\left(\sum_i \alpha_i (x_j^i - x_j)^2\right) + \cdots$$

Bearing in mind the equations of the mixture (Eqs. (1)–(3), the coefficient in terms of y is 1, the term with the first derivatives becomes zero and finally:

$$\hat{y} - y = \frac{1}{2}\sum_j \frac{\partial^2 y}{\partial {x_j}^2}\left(\sum_i \alpha_i (x_j^i - x_j)^2\right) + \cdots \qquad (10)$$

This expression explicitly shows that the approximation performed depends on the second derivates (more exactly on the Hessian matrix). It is, therefore, of the second order, while if the property y were modelised by a linear function, it would be of the first order. This quality of the approximation justifies the use of the LM method.

2.3. Predicting the response for the outliers

Outliers in prediction are samples that are inconsistent with the calibration data [5,6]. In LM, the prediction outliers are samples that are outside the convex hull determined by the calibration data. When outliers are present, the LM method as described in the preceding sections is not able to predict the response of these objects.

In order to estimate the response values for the objects that are outside the convex hull, the following steps are proposed. They are explained for the case of two dimensions, but can easily be generalised.

(a) For each outlier, one must find the two nearest neighbour objects belonging to the boundary of the calibration data set (Fig. 5). Together with the outlier they define a simplex (triangle) [7].

(b) Then, the outlier is reflected (mirrored) inside the convex hull through the centroid defined by the other two objects of the simplex on the face of the convex hull (Fig. 6). A new simplex inside the convex hull is obtained containing the reflected object.

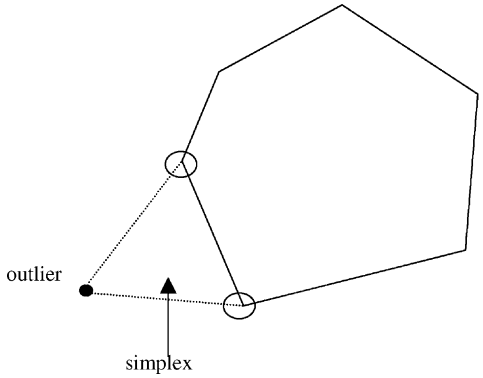

Fig. 5. Finding the two nearest objects on the boundary of the calibration data set. They defined together with the outlier a simplex.

(c) By means of the LM method the responses (y) for the centroid and for the reflected object can be calculated.

(d) In order to determine the response for the outlier, the difference between the response value for the centroid and the reflected object is computed and applied from the centroid to the outlier. The value for the outlier is determined by using the following expression:

$$y_{\text{reflected object}} - y_{\text{centroid}} = y_{\text{centroid}} - y_{\text{outlier}} \qquad (11)$$

where y represents the response value.

In the case of, e.g. three dimensions similar steps are necessary. However, the three nearest objects are used instead of the three nearest objects on the convex hull. The simplex is then defined also by objects not belonging to the boundary. The same expression as in the two-dimensional case is used to calculate the

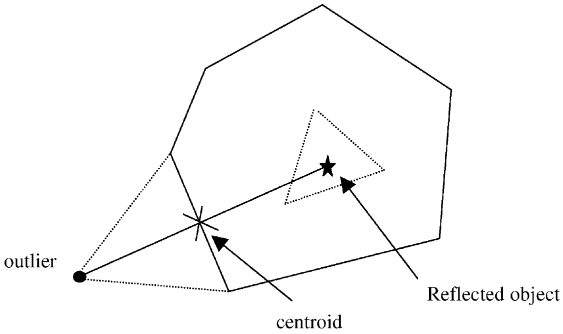

Fig. 6. Reflection of the outlier inside the convex hull.

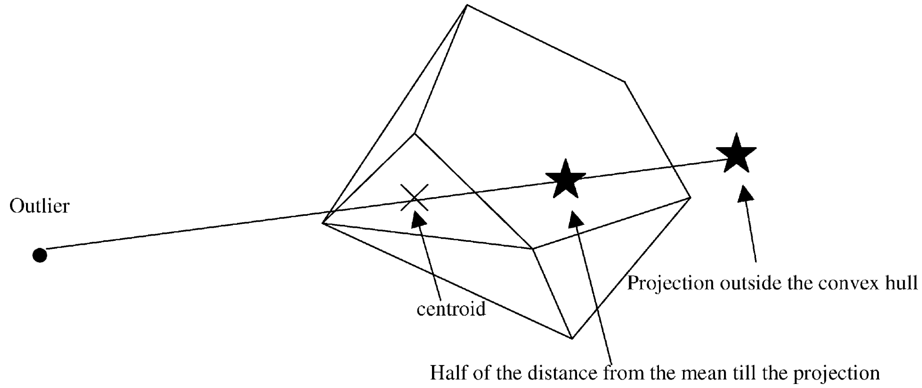
Outlier
centroid
Projection outside the convex hull
Half of the distance from the mean till the projection

Fig. 7. Reflection of the outlier outside the convex hull. Half of the distance between the centroid and the reflected point is used to find a simplex.

response value. Sometimes the reflection of the prediction outlier is also outside the convex hull, e.g. as in Fig. 7. In that case the reflection, y_{half}, is at half the distance between the centroid and the reflected object. In the case where taking half of the distance no simplex inside the convex hull is found, the procedure is repeated (using 1/3, 1/4, ... of the distance) until it does.

In order to obtain the response value, the following expression has to be used:

$$N(y_{\text{half}} - y_{\text{centroid}}) = y_{\text{centroid}} - y_{\text{outlier}} \qquad (12)$$

where N is 2 if 1/2 of the distance is used, 3 if 1/3, 4 if 1/4, ... Another possible situation is that where the reflection of the outlier is outside the convex hull and there is no simplex between this point and the centroid, i.e. all reflected points are always outside the lattice as is shown in Fig. 8. In these cases, the process is repeated with the first and the third neighbour objects in the boundary (in the case of two dimensions).

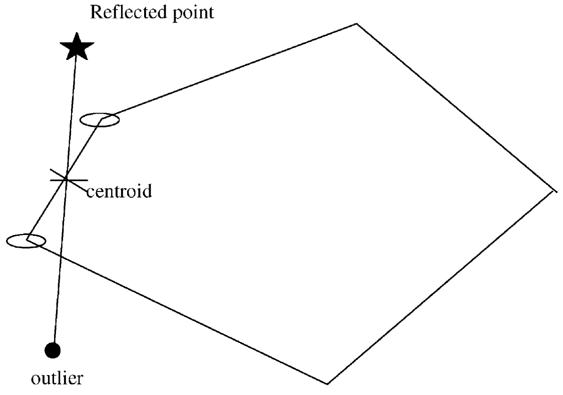

Fig. 8. Reflection outside the convex hull and no simplex between the centroid and the reflected points are found.

3. Results and discussion

Three NIR data sets were used. The first data set was received from the IFP and is called the Hydrogen data. It consists of 239 samples measured at 2128 wavelengths to determine the percentage of hydrogen in Gasoil. The second data set is called the Alfalfa data, received from the *Laboratori Agroalimentari de Cabrils* [8] (of the autonomous government of Cataluña, Spain). It consists of 305 samples of forages measured between 1108 and 2492 nm each 8 nm to determine the protein content.

The method was also applied to a much larger data set provided for the Agricultural Research Centre in Gembloux, Belgium. This data set, called here the Corn data, is used to determine the protein content in corn and consists of 1997 samples measured at 700 wavelengths.

3.1. Hydrogen data

This data set was split into a calibration set and a test set by using the duplex algorithm [9]. This method starts by selecting the two points furthest from each other and puts them both in a first set. Then the next two points furthest from each other are put in a second set, and the procedure is continued by alternatively placing pairs of points in the first or second set. In such a way 199 objects were used to build the model and 40 set aside for prediction. The calibration data were mean-centered. After pre-treatment, the mean of the calibration data was stored in order to center the 40 samples that must be predicted, using the same values.

PCR, PLS and the LM method are used as calibration methods. In order to select the number of components in PLS different techniques were applied. The first one is the cross-validation method leaving out one sample at a time. The RMSECV values were computed and plotted in function of the number of factors. In this data set, we obtained the minimal RMSECV when eight components were used. The randomisation test proposed by van der Voet [10] was also used. After comparing the predictive accuracy of the model using this test, five was found to be the optimal complexity. In the case of PCR, complexities of 10 and 5 are found with the minimal RMSECV and the randomisation test, respectively.

Another technique to determine the complexity of the model consists in applying the Monte-Carlo cross-validation (MCCV) method [11–14]. The MCCV method is an asymptotically consistent method to determine the number of components in calibration. It is based on the same principle as the leave-one-out cross-validation, but instead of leaving only one point out (nv $=$ 1), subsets of different sizes (nv $\gg$ 2) are left out during the calibration. These nv samples are then used for the validation. This procedure is repeated F times (F usually equals twice the number of objects in the calibration data). It is applied with 1, 2, 3, ... components kept in the PLS method and the RMSE is calculated each time. The number of components corresponding to the minimum of RMSE is selected. This procedure shows that the optimal number of factors when 150 samples are left out ($F = 2 \times 199$) for this data set is five as is shown in Fig. 9.

In this data set, the first five important PCs selected according to their correlation to y are 1st, 3rd, 4th, 6th and 2nd PC. In the LM method, when more dimensions are added more objects are detected as outliers. For this data set, five outliers are detected in the prediction set when working in two dimensions which represents 12.5% of the prediction objects (Fig. 10), 16 outliers in three dimensions (40%), 20 in four dimensions (50%) and 27 outliers in five dimensions (67.5% of the objects).

Another possibility is to split the data randomly. The duplex splitting we applied is preferred for

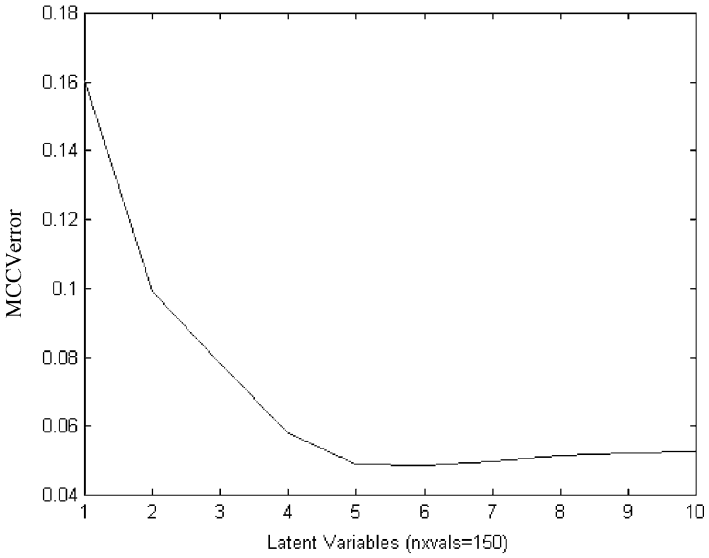

Fig. 9. Hydrogen data: MCCV error vs. number of PLS variables.

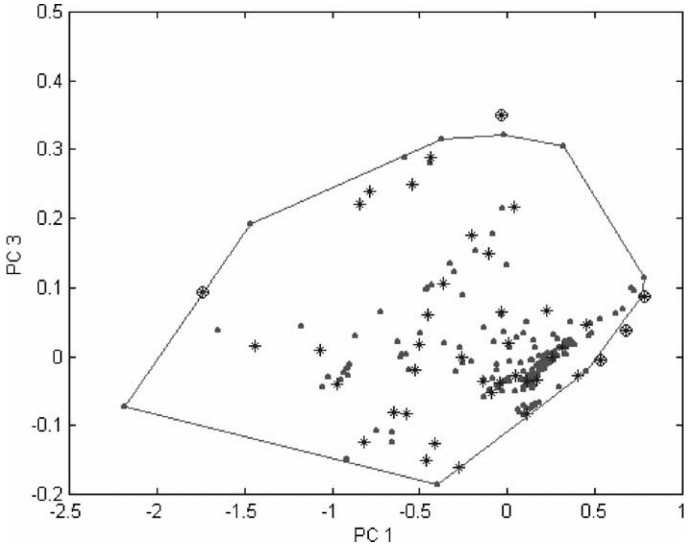

Fig. 10. Hydrogen data: PC1–PC3 (five outliers are detected) where ● are the calibration objects, ✳ are the prediction objects and ⊕ are the objects detected as outliers.

calibration with regression methods, because it yields a representative data set biased towards extreme samples, i.e. it includes relatively more extreme samples in the smallest data set, here the validation set. By random splitting some of the extreme points are included in the calibration data set and the number of outliers indeed decreases. The number of outliers working in two dimensions is then 1 (2.5%), 4 in three dimensions (10%), 10 in four dimension (25%) and 14 in five dimensions (35%).

The results applying the different calibration techniques in 2–5 dimensions are shown in Table 1 after removal of the outliers detected in the LM method (using the duplex method) also when performing PCR and PLS. This means that for instance the results for five dimensions are obtained using only 32.5% of the prediction objects. In that table, RMSECV represents the cross-validation error, and RMSEP describes the models predictive ability.

As a first conclusion one can say that LM gives good predictions for those objects that are inside the limits of the calibration samples, but too many outliers are out of those limits (although, it should be remembered that is due to the way we split the data with duplex method). In order to solve the problem, the proposed methodology to estimate the response values for the outliers is applied. Tables 2 and 3 show the responses of the outliers using two and three dimensions (5 and 16 outliers), respectively. Most of these responses are calculated using Eq. (11) but sometimes the reflected point is outside the boundary. In these cases Eq. (12) is applied.

Table 1
Hydrogen data: results using different dimensions when outliers were removed

Dimensions	Number of outliers removed	Value	PCR	PLS	LM
2	5 (12.5)	RMSECV, RMSEP	0.1077, 0.1479	0.0987, 0.1339	0.085, 0.1122
3	16 (40)	RMSECV, RMSEP	0.0706, 0.0647	0.0893, 0.0939	0.0628, 0.0724
4	20 (50)	RMSECV, RMSEP	0.0543, 0.0663	0.0621, 0.062	0.0412, 0.0501
5	27 (67.5)	RMSECV, RMSEP	0.0519, 0.0751	0.0487, 0.0694	0.0368, 0.0549

Values in parentheses are in percentage.

Table 2
Hydrogen data: response of the outliers using the simplex method and the LM method in two dimensions[a]

Outliers (index)	Y_{ref}	y (centroid)	y (reflection)	y (outlier)
1	10.92	10.85	10.80	10.90
2	14.37	14.25	14.19	14.31
4	12.44	12.38	12.17	12.59
21	14.21	14.19	14.27	14.11
38	14.4	14.19	14.12	14.26

[a] Y_{ref} are the values obtained with the reference method for each outlier, y (centroid) the responses of the centroid defined by the two nearest boundary objects, y (reflection) the responses of the reflected objects and y (outlier) are the estimated values.

From Table 2 one can see that the estimation of the response of the outliers is a good approximation to the real value (Y_{ref}). When all objects, outliers and points within the convex hull, are considered the RMSEP becomes 0.1115, almost the same as when only the points inside the lattice are used.

Also in three dimensions most of the outliers are well predicted and therefore the RMSEP (0.0888) is close to the value obtained when only the points inside the convex hull are included.

Table 3
Hydrogen data: response of the outliers using the simplex method and the LM method in three dimensions[a]

Outliers (index)	Y_{ref}	y (centroid)	y (reflection)	y (outlier)
1	10.92	10.97	11.00	10.93
2	14.37	14.25	14.17	14.34
3	12.71	12.72	12.67	12.77
4	12.44	12.63	12.83	12.44
11	11.04	11.26	11.25	11.35
13	11.9	12.11	12.18	12.00
17	13.86	13.74	13.71	13.81
19	11.6	12.00	12.27	11.73
20	13.28	13.06	12.72	13.39
21	14.21	13.94	13.74	14.14
27	12.16	12.33	12.51	12.15
30	12.19	12.01	11.96	12.10
32	12.92	12.92	12.97	12.87
34	13.36	13.38	13.62	13.15
37	14.08	13.91	13.68	14.13
38	14.4	14.13	14.01	14.26

[a] Y_{ref} are the values obtained with the reference method for each outlier, y (centroid) the responses of the centroid defined by the two nearest objects, y (reflection) the responses of the reflected objects and y (outlier) are the estimated values.

Table 4
Hydrogen data: comparison of the RMSEP value using the LM method with the value from PCR and PLS in 2–5 dimensions for all objects (outliers + points within the convex hull)

Dimensions	PLS	PCR	LM
2	0.1397	0.1541	0.1115
3	0.1174	0.0855	0.0888
4	0.0748	0.0728	0.0836
5	0.0635	0.0680	0.0915

The same study is also performed in four and five dimensions and using all these results, a new comparison with PLS and PCR this time for the whole prediction set was performed. The results are shown in Table 4.

From Table 4, it is obvious that the error obtained using LM is better than the one obtained using PCR and PLS working in two dimensions. In the case of three dimensions, the results obtained using LM are better than the result from PLS and similar to that from PCR. In four and five dimensions, the results are also shown in the table. In these two cases some outliers are not well predicted using LM, mainly the objects that are very far away from the boundary, but the RMSEP value is still reasonable.

3.2. Alfalfa data

This data set was split by the providers of the data into two independent data sets, a calibration set (205 samples) and a test set (100 samples). All the data were mean-centered.

In PCR the minimum RMSECV is obtained using 11 components and the randomization test gives 10 components. For the PLS regression the minimal RMSECV is obtained when 18 components are used, the randomization test gives 16 as the optimal complexity and the MCCV method shows that one can consider five latent variables.The first five important PCs are the 5th, 1st, 8th, 14th and 10th and LM is applied in the data space defined by these PCs.

As in the previous data set, the number of outliers increases with the number of dimensions. In two dimensions, 12 objects are detected as outliers in the test set (12% of the objects); in 3, 18 (18%); in 4, 29 (29%) and in 5, 39 (39%).

Table 5 shows the RMSECV and RMSEP when the different techniques are applied and the same outliers

Table 5
Alfalfa data: results using different dimensions when outliers were removed

Dimensions	Number of outliers removed	Value	PCR	PLS	LM
2	12 (12)	RMSECV, RMSEP	1.1807, 1.2373	2.2005, 2.1559	1.2037, 1.1981
3	18 (18)	RMSECV, RMSEP	1.0771, 1.0595	1.3798, 1.3898	1.1073, 0.9755
4	29 (29)	RMSECV, RMSEP	1.0116, 1.0387	1.1561, 1.1552	0.9314, 0.8917
5	39 (39)	RMSECV, RMSEP	0.9455, 0.8954	1.1067, 0.976	0.9133, 0.9696

Values in parentheses are in percentage.

are deleted from the test data set. For LM, the best prediction is obtained when four dimensions are used; the difference is not large between three and four components, so that it was decided that indeed the 14th and the 10th PCs should not be included.

Table 5 shows that the results using PCR are always better than the results using PLS and that the LM method performs at least as well as PCR. Table 6 shows the responses of the outliers using three dimensions (18 outliers).

For two dimensions, the response values for most of the outliers are well predicted. Table 7 shows the comparison of the RMSEP value using the LM method with the value from PCR and PLS in 2–4 dimensions for all objects in prediction.

Table 6
Alfalfa data: response of the outliers using the simplex method and the LM method in three dimensions[a]

Outliers (index)	Y_{ref}	y (centroid)	y (reflection)	y (outlier)
32	27.51	23.86	18.86	28.86
48	18.19	17.06	16.68	17.80
69	27.16	23.44	20.92	25.95
82	19.79	17.51	17.18	18.17
84	19.7	19.10	19.20	18.18
86	14.43	18.17	18.24	17.43
87	19.84	19.10	19.05	19.68
89	12.9	19.103	20.20	8.18
90	18.74	16.231	16.03	16.64
91	15.69	16.231	17.23	14.24
92	16.58	16.204	16.27	16.08
93	16.93	16.231	15.99	16.47
94	16.97	16.231	16.24	16.86
95	15.61	19.331	24.77	8.45
96	10.78	17.782	24.65	10.91
98	18.4	19.296	19.88	18.13
99	20.06	16.231	15.99	18.65
100	20.67	19.103	18.77	19.76

[a] Y_{ref} are the values obtained with the reference method for each outlier, y (centroid) the responses of the centroid defined by the two nearest objects, y (reflection) the responses of the reflected objects and y (outlier) are the estimated values.

Table 7
Alfalfa data: comparison of the RMSEP value using the LM method with the value from PCR and PLS in 2–4 dimensions for all objects (outliers + points within the convex hull)

Dimensions	PLS	PCR	LM
2	2.4346	1.4441	1.2932
3	1.6819	1.2321	1.3372
4	1.3645	1.3740	1.8534

In two dimensions, the results from the LM method are better than the results from the PLS and PCR methods. In the case of three dimensions, the results are better than the values obtained using PLS and comparable with the results using PCR. The best LM model uses two components and it is of the same order as the best PCR model (with three PCs) and the best PLS model (with four components).

3.3. Corn data

This data set was split into two subsets by using random selection: 1700 objects are used for calibration and 297 objects for prediction. All the data were mean-centered.

For the PCR method, the minimum RMSECV is obtained when 13 components are considered. Eleven was found to be the optimal complexity after comparing the predictive accuracy of the model using the randomization test between 13 and other complexities. For the PLS method, a complexity of 15 is obtained when the minimal RMSECV and the randomization test are used. Because the first seven PCs explained 99.86% of the total variance, here we considered those seven dimensions in the LM method.

Table 8
Corn data: results using different dimensions when outliers were removed

Dimensions	Number of outliers removed	Value	PCR	PLS	LM
2	0 (0)	RMSECV, RMSEP	0.9219, 0.9176	1.1698, 1.1796	1.092, 1.1372
3	3 (1.01)	RMSECV, RMSEP	0.8221, 0.8184	0.9781, 1.0125	1.0187, 1.0428
4	18 (6)	RMSECV, RMSEP	0.7572, 0.7117	0.9079, 0.8972	0.7648, 0.7672
5	34 (11.45)	RMSECV, RMSEP	0.6875, 0.6443	0.8255, 0.8179	0.6824, 0.6719
6	47 (15.82)	RMSECV, RMSEP	0.6299, 0.5803	0.6983, 0.6960	0.6281, 0.6098
7	69 (23.23)	RMSECV, RMSEP	0.5897, 0.5386	0.6090, 0.5827	0.5852, 0.5836

Values in parentheses are in percentage.

Table 9
Corn data: response of the outliers using the simplex method and the LM method in three dimensions[a]

Outliers (index)	Y_{ref}	y (outlier)
27	10.02	9.63
47	10.5	10.07
241	9.27	9.13

[a] Y_{ref} are the values obtained with the reference method for each outlier and y (outlier) are the estimated values.

For the LM method, the important PCs are selected as in PCR, and the first seven important PCs are the 5th, 7th, 10th, 12th, 4th, 9th and 2nd PC. For this data set, there are no outliers in two dimensions but 3 (1.01% of the prediction objects), 18 (6%), 34 (11.45%), 47 (15.82%) and 69 (23.23%) outliers are detected in 3–7 dimensions, respectively.

In order to compare the values, the outliers detected in the LM method in different dimensions are removed from the test set and then PCR and PLS are applied. Table 8 shows that the results from the LM method are comparable with the results from the PCR and PLS method when the outliers are removed in PCR and PLS. As in the previous case, PCR with selection of PCs gives better results than PLS.

Table 10
Corn data: comparison of the RMSEP value using the LM method with the value from PCR and PLS in 3–7 dimensions for all objects (outliers + points within the convex hull)

Dimensions	PLS	PCR	LM
3	1.0167	0.8152	1.0381
4	0.9315	0.7222	0.7999
5	0.8216	0.6552	0.6868
6	0.7227	0.6091	0.6462
7	0.6118	0.5697	0.6616

The response values for the outliers in these dimensions are well predicted. Table 9 shows the results for three dimensions. As is shown in the Table 10 the RMSEP for all objects is of the same order as the value using only the points within the lattice. The results for LM in the case of 4–6 dimensions are better than the results for PLS and similar to those for PCR. The result in seven dimensions for LM is also comparable with the results for PLS and PCR.

4. Conclusions

The Law of Mixtures method was presented here as an alternative to PCR/PLS for multivariate calibration. In all cases studied, the LM method gives at least similar results as PCR or PLS. A drawback of the method is the large number of outliers that is found when more than two dimensions are used. This high number of outliers is due to the small number of objects in multiple dimensions, and in part to the manner in which the data was split. If the duplex method was used to split the data, in some cases the number of outliers can reach 50% of the objects (for instance, the hydrogen data) when only four dimensions are applied. In the case of the alfalfa data, 29% of the objects were detected as prediction outliers. As in the case of the corn data for the same number of dimensions, only 6% of the objects were detected as prediction outliers. Here, the corn data set was randomly split and the influence of the amount of data on the number of outliers is clear.

With the three examples presented here it is shown that LM is a very good method to predict the concentration of the objects inside the lattice. It is shown also that acceptable prediction is possible for the points considered as outliers in prediction. However, other ways of treating such outliers can be imagined. It seems necessary to separate true outliers from outliers that are not true outliers, but merely objects situated on the border of the data set that became outliers through data splitting. Alternative methods for creating lattice, such as Delaunay triangulation will also be investigated.

References

[1] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics (Part B), Elsevier, Amsterdam, 1998, pp. 223–225.

[2] A. Espinosa, M. Sanchez, S. Osta, C. Boniface, J. Gil, A. Martens, B. Descales, D. Lambert, M. Valleur, Oil Gas J. 17 (1994) 49–56.

[3] J.A. Fernández Pierna, F. Wahl, O.E. de Noord, D.L. Massart, Chemom. Intell. Lab. Syst. 63 (2002) 27–39.

[4] F.P. Preparata, M.I. Shamos, Computational Geometry: An Introduction, Springer, Berlin, 1985.

[5] H. Martens, T. Naes, Multivariate Calibration, Wiley, New York, 1989.

[6] J.A. Fernández Pierna, L. Jin, F. Wahl, D.L. Massart, in: Proceedings of the 2nd International Symposium on PLS and Related Methods, Capri, Italy, October 2001.

[7] R. Carlson, Design and Optimization in Organic Synthesis, Data handling in Science and Technology, vol. 8, Elsevier, Amsterdam, 1992.

[8] I. Ruisánchez, F.X. Rius, S. Maspoch, J. Coello, T. Azzouz, R. Tauler, L. Sarabia, M.C. Ortiz, J.A. Fernández, D.L. Massart, A. Puigdomènech, C. García, Chemom. Intell. Lab. Syst. 63 (2002) 93–105.

[9] R.D. Snee, Technometrics 19 (1977) 415–428.

[10] H. van der Voet, Chemom. Intell. Lab. Syst. 25 (1994) 313–323;
H. van der Voet, Chemom. Intell. Lab. Syst. 28 (1995) 315.

[11] J. Shao, J. Am. Stat. Assoc. 88 (1993) 486–494.

[12] R.R. Picard, R.D. Cook, J. Am. Stat. Assoc. 79 (1984) 575–583.

[13] Q. Xu, Y. Liang, Chem. Intell. Lab. Syst. 56 (2001) 1–11.

[14] S. Gourvénec, J.A. Fernández Pierna, D.L. Massart, D.N. Rutledge, in: Proceedings of the 2nd International Symposium on PLS and Related Methods, Capri, Italy, October 2001.