Transfer of calibration function in near-infrared spectroscopy

Article *in* Chemometrics and Intelligent Laboratory Systems · February 1995

DOI: 10.1016/0169-7439(95)80023-3

ELSEVIER

Chemometrics and Intelligent Laboratory Systems 27 (1995) 189–203

Chemometrics and intelligent laboratory systems

Transfer of calibration function in near-infrared spectroscopy

M. Forina [a,*], G. Drava [a], C. Armanino [a], R. Boggia [a], S. Lanteri [a], R. Leardi [a], P. Corti [b], P. Conti [c], R. Giangiacomo [d], C. Galliena [d], R. Bigoni [e], I. Quartari [e], C. Serra [e], D. Ferri [f], O. Leoni [g], L. Lazzeri [g]

[a] *Istituto di Analisi e Tecnologie Farmaceutiche ed Alimentari, Via Brigata Salerno (Ponte), I-16147 Genova, Italy*
[b] *Dipartimento Farmaco Chimico Tecnologico, Via Banchi di Sotto 55, I-53100 Siena, Italy*
[c] *Dipartimento di Scienze Chimiche, Via S. Agostino 1, I-62032 Camerino, Italy*
[d] *Istituto Sperimentale Lattiero Caseario, Via A. Lombardo 11, I-20075 Lodi, Italy*
[e] *Eridania Laboratorio Chimico Centrale, Via Argine Ducale 397, I-44100 Ferrara, Italy*
[f] *Istituto Sperimentale Agronomico, Via Ulpiani 5, I-70125 Bari, Italy*
[g] *Istituto Sperimentale per le Colture Industriali, Via di Corticella 133, I-40129 Bologna, Italy*

Received 20 September 1993; accepted 24 March 1994

Abstract

A procedure for the transfer of the regression equation in near-infrared spectroscopy (NIRS), from a first instrument to a second instrument, is presented. The procedure uses partial least squares (PLS) regression twice: in the first step to compute the relationship between the spectra of transfer samples of the two instruments, and in the second step to compute the regression equation (relationship between chemical variables and spectral variables) of the first instrument. These two PLS steps are combined to predict the regression equation of the second instrument. Sometimes the PLS relationship between the two instruments is obtained from the principal components of the spectra of the two instruments. The procedure is applied to a set of 60 samples of soy flour, representative of the Italian soy production. 40 samples were used both as transfer samples and to compute the regression equation. 20 samples were used as evaluation set. Spectra were recorded with four different instruments, in four different laboratories. The result of the transfer procedure were evaluated by means of the standard error of prediction (*SEP*) with the predicted regression equation. Owing also to the great number of samples in the transfer set, and to the noise filtering effect of the twin PLS procedure, *SEP* with the predicted regression equation is not greater than that with the regression equation computed directly from the second instrument. The effect of some parameters, such as the number of PLS latent variables in the two steps, is also studied.

1. Introduction

Near-infrared spectroscopy (NIRS) is more and more used in process analysis and in the analysis of perishable raw materials, because of its speed and because it requires no or little sample treatment. As the number of measured reflectances or absorbances increased, from the first NIR instruments to the instruments of today, the use of suitable regression techniques such as principal component regression (PCR) or partial least

* Corresponding author.

SSDI 0169-7439(94)00028-H

squares regression (PLS) became essential to extract from the spectra measured on a training set (the 'standards' used for calibration) the multivariate regression equation, which computes the chemical quantity as a function of the measured physical quantities.

NIRS requires the standardization of the calibration procedure between different instruments (e.g., from an instrument in a central laboratory to a slave instrument), or for the same instrument, to correct for day-to-day variations: standardization means that the calibration procedure is transportable from an instrument to a second instrument, or from a day to another day.

The transfer of the calibration procedure is needed to avoid repetition of the whole calibration procedure, with the measure of the spectra on the whole calibration set and the computation of the calibration function.

Several procedures have been suggested to perform the transfer of calibration in NIRS. Generally, these procedures are based on the 'transfer of spectra'. This means that from the spectrum obtained by a satellite instrument (or in the day) B, the spectrum of the instrument (or in the day) A (master or reference instrument, or reference day) is computed. So all the spectra are referred to a standard instrument, and the calibration regression function obtained by the standard instrument can be applied, no matter where or when the original spectrum was obtained.

In the procedure of Shenk and Westerhaus [1,2], for each wavelength of the master instrument the wavelength of the satellite instrument with the highest correlation is detected. With this wavelength and the two neighbouring wavelengths of the satellite instrument, a quadratic model for the wavelength of the master instrument is computed. The correlation of the model with the wavelength of the master instrument is computed, and the location corresponding to the maximum correlation is considered as the satellite wavelength that most matches the master wavelength. After this correction for wavelength shifting, a correction for the spectral intensity is applied.

Wang and co-workers [3,4] compared five procedures, the Shenk procedure and four methods developed in the Laboratory of Chemometrics of Seattle. The piecewise direct standardization (PDS) was the method with the best performances, as measured by the standard error of prediction, *SEP*. PDS is based on PCR or PLS regression.

Be $\mathbf{X}^A$ the matrix of spectra obtained by instrument A on the samples used as 'transfer set', and $\mathbf{X}^B$ the matrix of spectra measured by instrument B on the same samples. The transformation matrix $\mathbf{F}$

$$\mathbf{X}^A = \mathbf{X}^B\mathbf{F}$$

is obtained as a banded diagonal matrix, by the regression coefficients $\boldsymbol{b}$ of each wavelength v (reflectance, absorbance or transform) of the instrument A on a limited number of wavelengths of the instrument B, near to v. $\mathbf{F}$ is then used to transfer the spectra from instrument B to instrument A.

Dardenne et al. [5] used the Shenk procedure, with special attention to two relevant problems, the use of the transfer procedure in a network with different instruments, and the set of standardization samples required to compute the transfer rule.

Recently [6,7] we suggested to perform the transfer of calibration by means of the transfer of the regression equation, that means that the regression coefficients $\boldsymbol{c}^B$ of the calibration function that computes the chemical quantity y from the spectrum $\boldsymbol{x}^B$ for the instrument B:

$$y = \boldsymbol{x}^{B\mathrm{T}}\boldsymbol{c}^B$$

(where $\boldsymbol{x}^{B\mathrm{T}}$ is the row vector transpose of $\boldsymbol{x}^B$) are predicted from $\boldsymbol{c}^A$ and, of course, from a transfer matrix $\mathbf{M}$, analogous to $\mathbf{F}$, but obtained by only one step of PLS regression, applied to all the wavelengths in matrices $\mathbf{X}^A$ and $\mathbf{X}^B$.

In this paper the complete theory of the transfer of calibration equation is presented, with the experimental results obtained by using four different instruments in four laboratories. This work has been realized as a collaborative study of the Italian Group for NIRS. It stresses on the transfer algorithm more than on the application to real problems, which requires further work and comparison with other procedures.

The samples for the transfer of spectra (trans-

fer set, matrices **M**) are the same as the ones used to obtain the regression equation (calibration set, vectors c). Only in a limited number of cases the transfer of spectra was obtained with a reduced number of samples. Moreover, the validation procedure is that of the single evaluation set. Work is in progress to apply more reliable evaluation procedures (cross-validation, repeated evaluation set), to use generic standards in the transfer set and to study the minimum number of standards to be used.

2. Experimental

2.1. Samples

60 samples of soy seeds were collected from extensive soy fields in many Italian regions, to be representative of the Italian soy production. Seeds were milled, and subsamples from the 60 samples of soy flour were sent to four laboratories.

The chemical analysis was done with the following methods.

- Moisture: gravimetric, by drying at 105°C for 4 h; result as the mean of three determinations (inaccuracy as 95% confidence interval: ±1.5%).
- Proteins: determination of nitrogen by the Kjeldahl method, catalyst selenic mixture. Nitrogen percentage is multiplied by 6.25 to obtain the total protein content; result as the mean of three determinations (inaccuracy as 95% confidence interval: ±2.5%).
- Oil: determined by Soxhlet extraction; result as the mean of three determinations (inaccuracy as 95% confidence interval: ±2%).
- Spectra were recorded in four different laboratories, namely: (A) Technicon InfraAlyzer 400R (Bologna), 19 wavelengths; (B) Technicon InfraAlyzer 400R (Bari), 19 wavelengths; (C) Technicon InfraAlyzer 500 (Lodi), 350 wavelengths; (D) NIRSystem 6500 (Ferrara), 700 wavelengths.

Spectra were the average respectively of three (instruments A and B) and of five replicates (instruments C and D).

In the case of instruments C and D only 175 wavelengths were used in data analysis (one every two and one every four, respectively).

Table 1
Characteristics of training and evaluation sets

Variable	Mean	Std. dev.	Minimum	Maximum
Training set				
Moisture	11.880	3.210	5.9	18.4
Proteins	34.705	3.872	29.0	43.4
Oil	18.793	2.434	14.7	22.9
Evaluation set				
Moisture	11.785	3.110	7.5	17.9
Proteins	35.330	2.298	31.7	42.4
Oil	19.100	2.350	15.3	22.5

Samples were divided into a training set of 40 objects and an evaluation set of 20 objects. Subdivision was done in laboratory C, in the usual way of the laboratory, without the use of rules of experimental design to have both optimal training and evaluation sets. Simply, the distributions of the response variables were inspected, and the evaluation samples were selected to obtain about the same distributions for the training and the evaluation sets.

Also for storage of samples and for measurements, the procedures were those of ordinary routine work in the four laboratories.

The training set of 40 objects was used both to compute the calibration coefficients (calibration step, see below) and for the spectrum transfer step. In a limited number of cases the spectrum transfer set was performed with 3–10 objects.

Table 1 shows the main characteristics of the training and evaluation sets.

3. Theory

The following notation will be used for a matrix: ${}_I\mathbf{X}_V$, with the left pedix for the number of rows, and the right pedix for the number of columns. The transpose of ${}_I\mathbf{X}_V$ is denoted by ${}_V\mathbf{X}_I$. Lowercase letters corresponding to the number of rows or columns are used to indicate a scalar, as x_{iv}.

So ${}_I\boldsymbol{x}$ is a column vector of I rows, and $\boldsymbol{x}_V$ is a row vector of V columns.

${}_I\mathbf{X}_V$ indicates the matrix of predictor variables

(original physical quantities, or modified quantities as derivatives, autoscaled data, centered data). ${}_I\mathbf{X}_V^A$ indicates that the physical quantities refer to laboratory A.

We will use I for the number of objects in the training set ($I = 40$), and J for the evaluation set, so that ${}_J\mathbf{X}_V$ indicates the matrix of the predictor variables for the 20 objects of the evaluation set.

${}_I\mathbf{X}_M$, with $M = V + 1$ indicates the augmented matrix, where a column M of all 1 ($x_{iM} = 1$) is added to the original matrix ${}_I\mathbf{X}_V$.

${}_I\mathbf{Y}_3$ is the response matrix, of three response variables (moisture, proteins, oil); ${}_I\mathbf{y}_k$ is the column vector of the kth response variable.

3.1. Principal components (reduction step)

Sometimes, to reduce the number of variables in the predictor matrix, the eigenvectors of ${}_I\mathbf{X}_V$ are computed:

$${}_I\mathbf{X}_V = {}_I\mathbf{S}_E \, {}_E\mathbf{L}_V \tag{1}$$

E components are retained; ${}_I\mathbf{S}_E$, and the related matrix of scores of the evaluation set

$${}_J\mathbf{S}_E = {}_J\mathbf{X}_V \, {}_V\mathbf{L}_E \tag{2}$$

are then processed as ${}_I\mathbf{X}_V$ and ${}_J\mathbf{X}_V$.

3.2. Regression y–X (calibration step)

In each laboratory, PLS regression computes the vector of the regression coefficients, so that:

$${}_I\mathbf{y}_k = {}_I\mathbf{X}_M^A \, {}_M\mathbf{c}_k^A + {}_I\mathbf{e}_k^A \tag{3}$$

where ${}_M\mathbf{c}_k^A$ is the vector of the regression coefficients for the kth response variable computed with data of laboratory A, and ${}_I\mathbf{e}_k^A$ is the vector of residuals. The value in the Mth row of the vector of regression coefficients is the intercept.

Regression coefficients have been computed by means of the algorithm of Marengo and Todeschini [8]; they depend on the number P of latent variables used in the PLS regression.

The values estimated by the regression are given by:

$${}_I\hat{\mathbf{y}}_k = {}_I\mathbf{X}_M^A \, {}_M\mathbf{c}_k^A \tag{4a}$$

and for a generic object:

$$\hat{y}_{ik}^A = {}_i\mathbf{x}_M^A \, {}_M\mathbf{c}_k^A \tag{4b}$$

The before-regression variance of the kth response variable, independent of the laboratory, is given by:

$$s_{bk}^2 = \frac{\sum_i (y_{ik} - \bar{y}_k)^2}{I - 1} \tag{5}$$

for the training set. $\bar{y}_k$ is the mean of the response variable, computed on the data of the training set.

For the evaluation set, Eq. (5) is modified to:

$$s_{bk}^2 = \frac{\sum_j (y_{jk} - \bar{y}_k)^2}{J} \tag{6}$$

$\bar{y}_k$ is the same as in Eq. (5).

The after-regression variance for the training set is:

$$s_{ak}^2 = \frac{\sum_i (y_{ik} - \hat{y}_{ik}^A)^2}{I - L} \tag{7}$$

L is 1 plus the number of latent variables; it is M in the ordinary least-squares regression.

The square root of this variance (residual standard deviation of fitting) is generally called *SEC* (standard error of calibration).

The after-regression variance for the evaluation set is:

$$s_{ak}^2 = \frac{\sum_j (y_{jk} - \hat{y}_{jk}^A)^2}{J} \tag{8}$$

The square root of this variance (residual standard deviation of prediction) is generally called *SEP* (standard error of prediction).

Often, to measure the goodness of fit the following quantity is used:

$$R^2 = 1 - \frac{\sum_i (y_{ik} - \hat{y}_{ik}^A)^2}{\sum_i (y_{ik} - \bar{y}_k)^2} \tag{9}$$

3.3. Regression X–X (spectrum transfer step)

PLS regression between two blocks (training set) of predictor variables (two instruments) gives

two matrices of regression coefficients:

$${}_I\hat{\mathbf{X}}_V^A = {}_I\mathbf{X}_M^B \, {}_M\mathbf{M}_V^{AB} \quad (10)$$

$${}_I\hat{\mathbf{X}}_V^B = {}_I\mathbf{X}_M^A \, {}_M\mathbf{M}_V^{BA} \quad (11)$$

These equations do not require the same number of variables V in the two **X** blocks. Moreover, matrices **X** can be those of original data or of scores (we will specify the various possibilities as type **X–X**, **S–S**, **X–S** of data for the inter-instruments relationship, but otherwise we will use **X–X** independently of the nature of the two matrices).

The last row of a matrix **M** is the column of intercepts.

3.4. Transfer of the regression equation

Let us now use Eq. (4a) in the form:

$${}_I\hat{\mathbf{y}} = {}_I\mathbf{X}_M^A \, {}_M\boldsymbol{c}^A \quad (12)$$

where the index k of the response variable is omitted.

Matrix **X** and vector $\boldsymbol{c}$ can be partitioned to separate the column of 1 and the intercept:

$${}_I\hat{\mathbf{y}} = \left({}_I\mathbf{X}_V^A \quad {}_I\mathbf{1}\right)\begin{pmatrix} {}_V\boldsymbol{c}^A \\ c^A \end{pmatrix}$$

$$= \left({}_I\mathbf{X}_V^A \, {}_V\boldsymbol{c}^A \quad {}_I\mathbf{1}\, c^A\right) \quad (13a)$$

where vector ${}_I\hat{\boldsymbol{y}}$ is regarded as the sum of two vectors, the second, ${}_I\mathbf{1}\, c^A$, being a column of constant terms, the intercepts.

In the same way, for the estimate of $\boldsymbol{y}$ from spectra measured by instrument B:

$${}_I\hat{\boldsymbol{y}} = \left({}_I\mathbf{X}_V^B \, {}_V\mathbf{c}^B \quad {}_I\mathbf{1}\, c^B\right) \quad (13b)$$

Now, substituting in Eq. (13a) to ${}_I\mathbf{X}_V^A$ its estimate given by Eq. (10):

$${}_I\hat{\boldsymbol{y}} = \left({}_I\hat{\mathbf{X}}_V^A \, {}_V\boldsymbol{c}^A \quad {}_I\mathbf{1}\, c^A\right)$$

$$= \left({}_I\mathbf{X}_M^B \, {}_M\mathbf{M}_V^{AB} \, {}_V\boldsymbol{c}^A \quad {}_I\mathbf{1}c^A\right) \quad (14)$$

By partitioning ${}_I\mathbf{X}_M^B$ as

$$\left({}_I\mathbf{X}_V^B \quad {}_I\mathbf{1}\right)$$

and ${}_M\mathbf{M}_V^{AB}$ as

$$\begin{pmatrix} {}_V\mathbf{M}_V^{AB} \\ \boldsymbol{m}_V^{AB} \end{pmatrix}$$

(the Mth row of ${}_M\mathbf{M}_V^{AB}$ is the row of the V intercepts $\boldsymbol{m}_V^{AB}$):

$${}_I\hat{\mathbf{X}}_V^A = \left({}_I\mathbf{X}_V^B \quad {}_I\mathbf{1}\right)\begin{pmatrix} {}_V\mathbf{M}_V^{AB} \\ \boldsymbol{m}_V^{AB} \end{pmatrix}$$

$$= \left({}_I\mathbf{X}_V^B \, {}_V\mathbf{M}_V^{AB} \quad {}_I\mathbf{1}\, \boldsymbol{m}_V^{AB}\right)$$

and

$${}_I\hat{\mathbf{X}}_M^A = \left({}_I\mathbf{X}_V^B \, {}_V\mathbf{M}_V^{AB} \quad {}_I\mathbf{1}\, \boldsymbol{m}_V^{AB} \quad {}_I\mathbf{1}\right)$$

is partitioned into two matrices and a column vector.

So Eq. (14) becomes:

$${}_I\hat{\boldsymbol{y}} = \left({}_I\mathbf{X}_V^B \, {}_V\mathbf{M}_V^{AB} \quad {}_I\mathbf{1}\, \boldsymbol{m}_V^{AB} \quad {}_I\mathbf{1}\right)\begin{pmatrix} {}_V\boldsymbol{c}^A \\ c^A \end{pmatrix}$$

$$= \left({}_I\mathbf{X}_V^B \, {}_V\mathbf{M}_V^{AB} \, {}_V\boldsymbol{c}^A \quad {}_I\mathbf{1}\, \boldsymbol{m}_V^{AB} \, {}_V\boldsymbol{c}^A \quad {}_I\mathbf{1}\, c^A\right) \quad (15)$$

By comparing Eq. (15) with Eq. (13b)

$${}_I\hat{\boldsymbol{y}} = \left({}_I\mathbf{X}_V^B \, {}_V\boldsymbol{c}^B \quad {}_I\mathbf{1}\, c^B\right)$$

giving the estimate ${}_I\hat{\boldsymbol{y}}$ directly from the spectra ${}_I\mathbf{X}_V^B$ measured from instrument B, we can see that the slopes of the regression Eq. (13b) can be estimated by:

$${}_V\hat{\boldsymbol{c}}^{BA} = {}_V\mathbf{M}_V^{AB} \, {}_V\boldsymbol{c}^A \quad (16)$$

and the intercept c^B by:

$$\hat{c}^{BA} = \boldsymbol{m}_V^{AB} \, {}_V\boldsymbol{c}^A + c^A \quad (17)$$

so that the final equation for the estimate of the regression coefficients of instrument B from the regression coefficients computed for instrument A and from the inter-**X** blocks regression matrix **M** is given by:

$${}_M\hat{\boldsymbol{c}}^{BA} = \begin{pmatrix} {}_V\mathbf{M}_V^{AB} \, {}_V\boldsymbol{c}^A \\ \boldsymbol{m}_V^{AB} \, {}_V\boldsymbol{c}^A + c^A \end{pmatrix} \quad (18)$$

This equation can be applied also when the number of variables in the two **X** blocks is different, V_A and V_B. In this case, it changes into:

$$ {}_M\hat{\boldsymbol{c}}^{BA} = \begin{pmatrix} {}_{V_B}\mathbf{M}^{AB}_{V_A}\ {}_{V_A}\boldsymbol{c}^{A} \\ \boldsymbol{m}^{AB}_{V_A}\ {}_{V_A}\boldsymbol{c}^{A} + c^{A} \end{pmatrix} \quad (19) $$

As a result of the transfer of coefficients, two estimates ${}_I\hat{\boldsymbol{y}}$ are available for the instrument B: the direct estimate (with local or 'computed' regression coefficients $\boldsymbol{c}$):

$$ {}_I\hat{\boldsymbol{y}} = {}_I\mathbf{X}^{B}_{M}\ {}_M\boldsymbol{c}^{B} \quad (20) $$

and the estimate with the predicted regression coefficients $\hat{\boldsymbol{c}}$:

$$ {}_I\hat{\boldsymbol{y}} = {}_I\mathbf{X}^{B}_{M}\ {}_M\hat{\boldsymbol{c}}^{BA} \quad (21) $$

Consequently, for each predicted instrument B and for each predictor instrument A, we obtain four parameters describing the goodness of the results:

SEC^B: standard error of calibration with locally computed regression coefficients $\boldsymbol{c}$.

SEP^B: standard error of prediction with locally computed regression coefficients $\boldsymbol{c}$.

SEC^{BA}: standard error of calibration with predicted regression coefficients $\hat{\boldsymbol{c}}^{BA}$.

SEP^{BA}: standard error of prediction with predicted regression coefficients $\hat{\boldsymbol{c}}^{BA}$.

When, in Eq. (10), ${}_I\hat{\mathbf{X}}^{A}_{V} = {}_I\mathbf{X}^{B}_{M}\ {}_M\mathbf{M}^{AB}_{V}$ the scores ${}_I\mathbf{S}^{B}_{E}$ have been used instead of ${}_I\mathbf{X}^{B}_{V}$, so that $M = E + 1$, E being the number of retained components, from the coefficients estimated for the scores by Eq. (18) or (19), the coefficients of the regression function for the original variables can be estimated using the matrix of loadings **L**:

$$ {}_I\hat{\boldsymbol{y}} = {}_I\mathbf{S}^{B}_{M}\ {}_M\hat{\boldsymbol{c}}^{BA} = \left({}_I\mathbf{S}^{B}_{E}\ {}_E\hat{\boldsymbol{c}}^{BA} \quad {}_I\mathbf{1}\ \hat{c}^{BA}\right) $$

$$ = \left({}_I\mathbf{X}^{B}_{V}\ {}_V\mathbf{L}_E\ {}_E\hat{\boldsymbol{c}}^{BA} \quad {}_I\mathbf{1}\ \hat{c}^{BA}\right) $$

so that the V slopes in the regression equation with predictors **X** can be obtained by the E

Table 2
Effect of the number of components of global-centered data retained in the block **S**, E, on the performances (*SEC* and *SEP*) of PLS regression y–**S**

L	*	E									
		19	14	12	10	8	7	6	5	4	3
SEC											
1	1.3149	0.9022	0.9464	0.9949	1.0212	1.0658	1.0658	1.0977	1.0979	1.1194	1.1190
2	1.0527	0.7233	0.7988	0.8769	0.9189	0.9861	0.9861	1.0421	1.0428	1.0730	1.0731
3	1.0132	0.7041	0.7274	0.7522	0.7607	0.8128	0.8135	0.8540	0.9492	0.9861	
4	0.9645	0.6981	0.7310	0.7592	0.7697	0.8232	0.8239	0.8658	0.9623		
5	0.9204	0.7080	0.7414	0.7700	0.7806	0.8348	0.8356	0.8781			
6	0.8413	0.7183	0.7522	0.7813	0.7920	0.8470	0.8478				
7	0.8019	0.7291	0.7635	0.7930	0.8039	0.8598					
8	0.7824	0.7404	0.7753	0.8053	0.8163						
9	0.7768	0.7523	0.7878	0.8182	0.8294						
10	0.7775	0.7647	0.8008	0.8317							
SEP											
1	1.4310	1.2644	1.2948	1.2566	1.2789	1.2959	1.2957	1.2732	1.2893	1.3173	1.3286
2	1.3826	1.3602	1.3396	1.2670	1.3208	1.3606	1.3605	1.2933	1.3053	1.3615	1.3635
3	1.4677	1.2753	1.3159	1.3971	1.4598	1.4950	1.4708	1.3784	1.3874	1.4739	
4	1.4498	1.3092	1.3519	1.4241	1.4756	1.5039	1.4798	1.3787	1.3880		
5	1.3266	1.3092	1.3520	1.4242	1.4756	1.5039	1.4798	1.3787			
6	1.4404	1.3092	1.3520	1.4242	1.4756	1.5039	1.4798				
7	1.5502	1.3092	1.3520	1.4242	1.4756	1.5039					
8	1.5644	1.3092	1.3520	1.4242	1.4756						
9	1.4126	1.3092	1.3520	1.4242	1.4756						
10	1.3861	1.3092	1.3521	1.4240							

Instrument: A; response: moisture; *: original variables (regression y–**X**); L: number of latent variables in PLS regression.

slopes in the regression equation with predictors **S** as

$$_{V}\hat{\boldsymbol{c}}^{BA} = {}_{V}\mathbf{L}_{E\ E}\hat{\boldsymbol{c}}^{BA} \quad (22)$$

(the intercept unchanged).

4. Results and discussion

The results of the described procedure depend on factors controlling the two (*y*–**X** and **X**–**X**) or three (**X**–**S**, *y*–**S** and **S**–**S**) steps: (a) pretreatment of data matrix $_{I}\mathbf{X}_{V}$; (b) number of principal components computed in **X**–**S** reduction step; (c) number of PLS latent variables in *y*–**X** or *y*–**S** calibration step; (d) number of PLS latent variables in **X**–**X** or **S**–**S** transfer step.

With four instruments and three response variables *y*, each instrument can be predicted from three instruments, so that 36 predictions were studied as a function of the above factors.

Only some results are reported here, generally in the form of figures, corresponding to the different cases of the spectrum transfer step (**X**–**X**, **X**–**S**, and **S**–**S**) and to the three different response variables. These results are representative of all the results obtained. Other results are available upon request to the authors.

When used, scores were those of the global-centered data (only a limited study was performed on data pretreatment for step **X**–**S**: global centering gave the best results, both with regard to double cross-validation of the number of significant eigenvectors and in predictive ability in the step *y*–**S**).

The number of components was selected to obtain in the calibration step *y*–**S** about the same predictive ability (*SEP*) as that was obtained with the original data in the calibration *y*–**X**. Table 2 shows an example of the results.

Figs. 1 and 2 refer to B–A, type **X**–**X**, response variable moisture.

Fig. 3 shows the regression coefficients for instrument B, $\boldsymbol{c}^{B}$, computed by OLS (ordinary least-squares regression) and by PLS, and $\boldsymbol{c}^{BA}$, predicted from instrument A.

Figs. 4 and 5 refer to C–D, type **S**–**S**, response variable proteins.

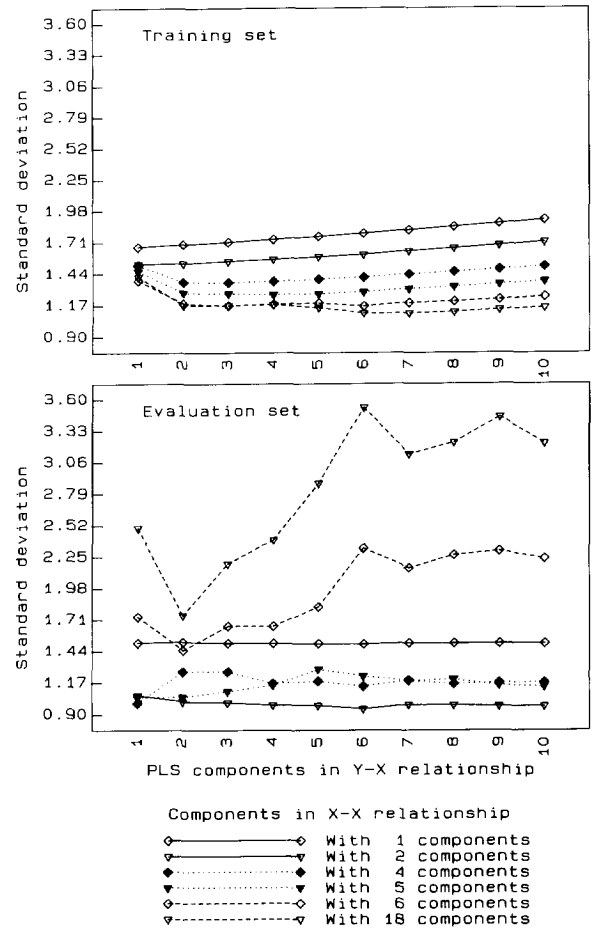

Fig. 1. Standard error of calibration *SEC* (top) and standard error of prediction *SEP* (bottom) with predicted regression coefficients. Response variable: moisture. Prediction of instrument B from A, type **X**–**X**.

Figs. 6 and 7 refer to A–C, type **X**–**S**, response variable oil. These figures refer to the case where the number of variables in the two blocks of predictors is different (19 for **X** and 15 for **S**), so that Eq. (19) is applied.

Fig. 8 shows the regression coefficients for instrument C, $\boldsymbol{c}^{C}$, as computed when the original variables **X** are used in the *y*–**X** step.

Fig. 9 shows the regression coefficients $\boldsymbol{c}^{CA}$, predicted by using the interblock relationship C–A type **S**–**X**, and then Eq. (22) to obtain the regression coefficients of the original variables from the regression coefficients of scores.

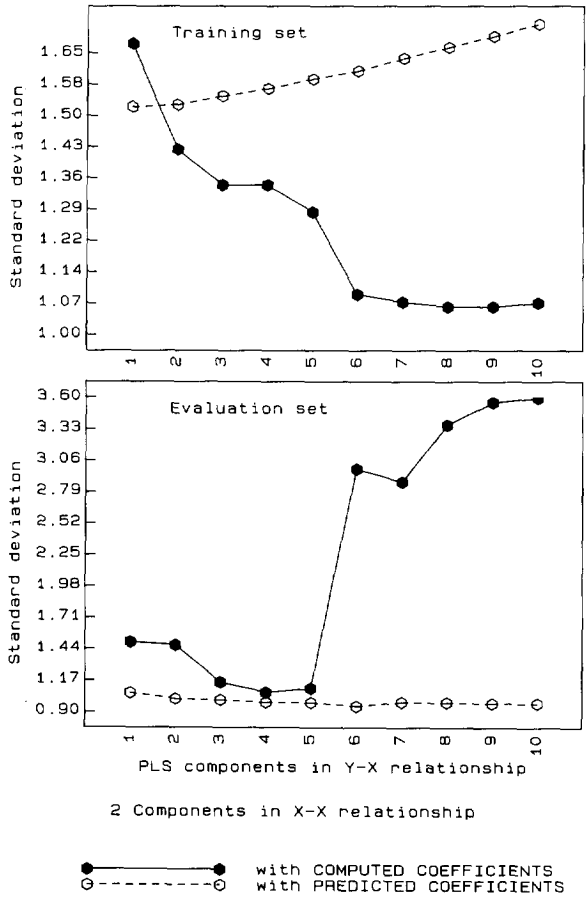

Fig. 2. Standard error of calibration *SEC* (top) and standard error of prediction *SEP* (bottom) with computed and predicted regression coefficients. Response variable: moisture. Prediction of instrument B from A, type **X–X**.

Figs. 10 and 11 refer to B–A type **S–S**, by using eight principal components in the **X–S** step. The response variable is moisture, to compare results with those in Figs. 1 and 2.

In Fig. 12 the results obtained when only five objects are used in the spectrum transfer step **X–X** (instruments B–A, moisture) are reported. The five objects were selected among the 40 objects of the training set. The two first principal components of the **X** block were computed, and the choice was made to obtain a D-optimal design.

4.1. Residual standard deviation on the evaluation set (SEP)

Generally, the residual standard deviation on the evaluation set, *SEP*, with the predicted coefficients $\hat{c}^{AB}$ is less than that with the computed coefficients c^{A}. This is the result of greatest importance: it demonstrates that the transfer of

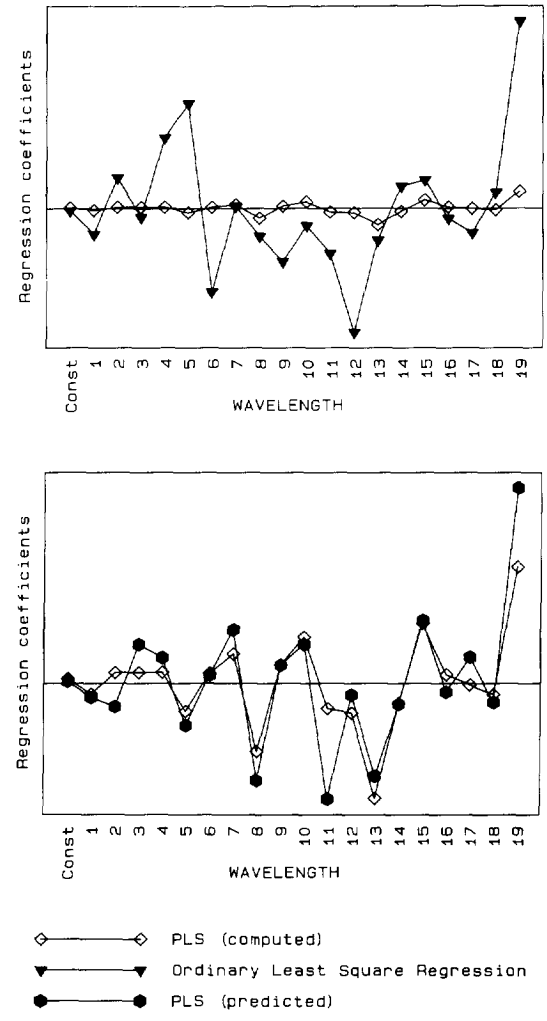

Fig. 3. (Top) Regression coefficients for instrument B, c^{B}, computed by OLS (ordinary least-squares regression) and by PLS (direct calibration). (Bottom) Regression coefficients computed by PLS (direct calibration) and predicted from instrument A (magnified scale).

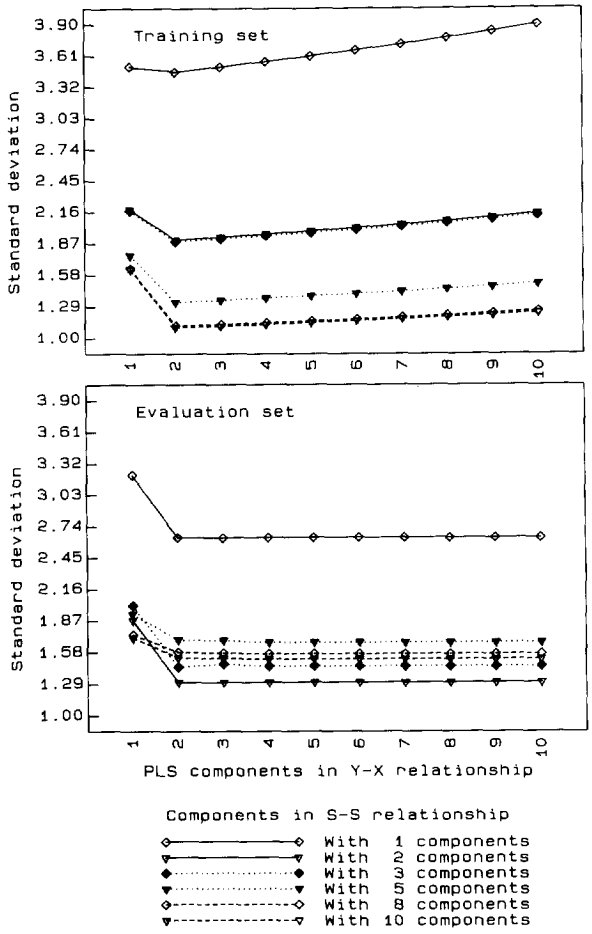

Fig. 4. Standard error of calibration *SEC* (top) and standard error of prediction *SEP* (bottom) with predicted regression coefficients. Response variable: proteins. Prediction of instrument C from D, type S–S, 15 principal components in S blocks.

the calibration equation is not only feasible, but generally favourable.

PLS regression is a biased regression technique (in comparison with OLS, under the hypothesis that the predictor variables are all and only those relevant to the relationship with the response variable). It can introduce a (low) systematic error but reduces the random error. OLS can use a lot of noise to obtain the maximum fitting to data in the training set. In the case of instrument B for the response moisture, with OLS R^2 is 91.63%, corresponding to a residual standard deviation of 2.19. The residual standard deviation on the evaluation set is 3.69, very high. OLS diagnostics show very high inflation factors (up to 300000). The confidence intervals for each regression coefficient are also higher than the value of the coefficients itself, so that this value is not significant.

OLS minimizes the sum of squares: in Eq. (7), the sum of squares is divided by the number of degrees of freedom to obtain the residual standard deviation. The optimum fitting of OLS is generally achieved with great absolute values of the regression coefficients, both positive and neg-

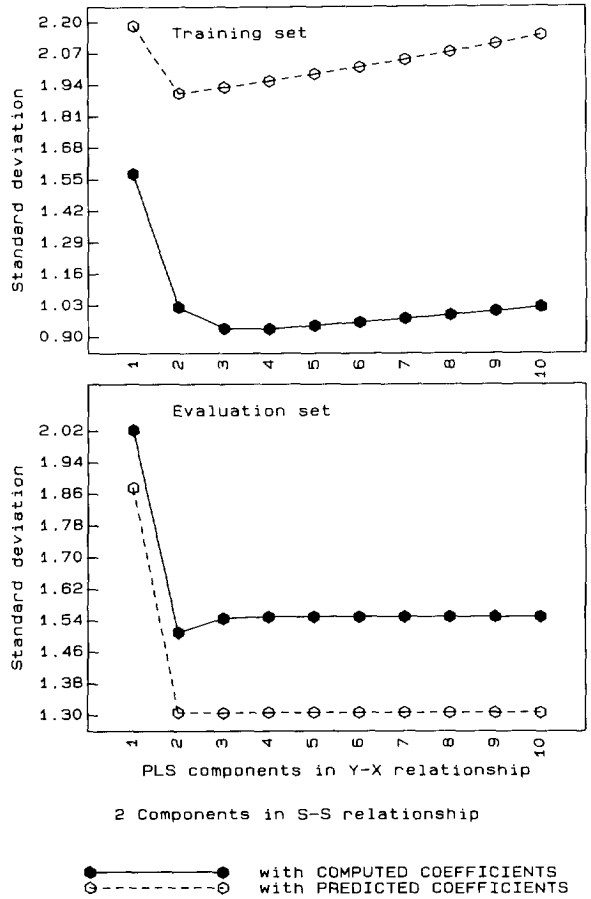

Fig. 5. Standard error of calibration *SEC* (top) and standard error of prediction *SEP* (bottom) with computed and predicted regression coefficients. Response variable: proteins. Prediction of instrument C from D, type S–S, 15 principal components in S blocks.

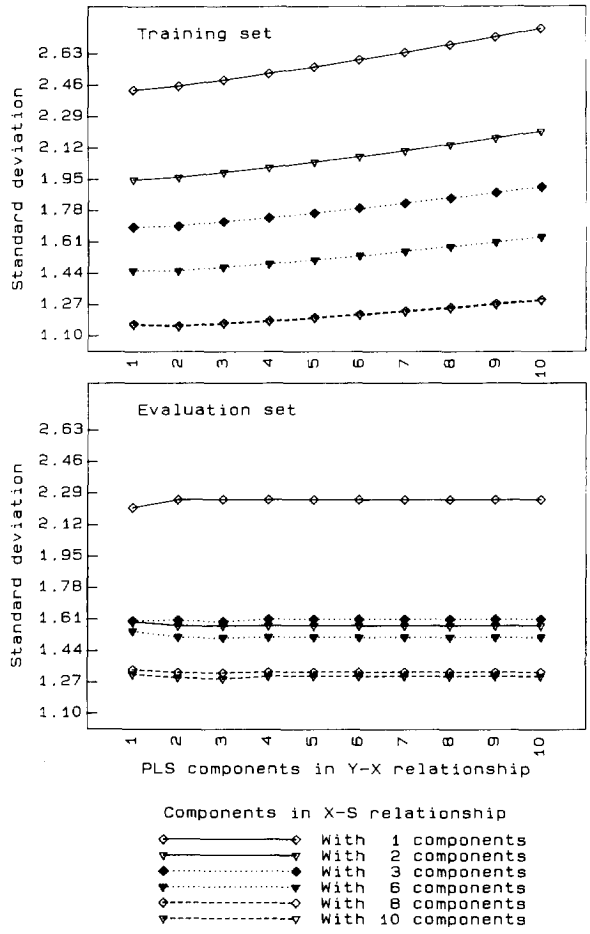

Fig. 6. Standard error of calibration *SEC* (top) and standard error of prediction *SEP* (bottom) with predicted regression coefficients. Response variable: oil. Prediction of instrument A from C, type **X–S**, 13 principal components in **S** blocks.

ative (as shown in Fig. 3). The estimate of the response variable is obtained as the sum of great positive and negative contributions, and a relatively small error on one predictor variable causes, consequently, a great error on the response.

PLS reaches an R^2 value as high as that of OLS only with a great number of latent variables (PLS tends to the model of OLS as the number of latent variables tends to the number of predictor variables): however, when the number of latent variables is low, the residual standard deviation in fitting can be lower than that of OLS, 1.06 with eight latent variables.

The regression coefficients in the PLS model are very low, as compared with the OLS coefficients. The estimate of the response variable results as the sum of two relatively small positive and negative contributions. An error on one of the predictor variables causes a relatively small error on the response variable.

More important is the residual standard deviation on the evaluation set: for PLS it is 1.07 with four latent variables (see Fig. 2). When the predicted coefficients are used, the residual standard deviation on the evaluation set is 0.95 (two latent variables in the **X–X** step, five to seven latent

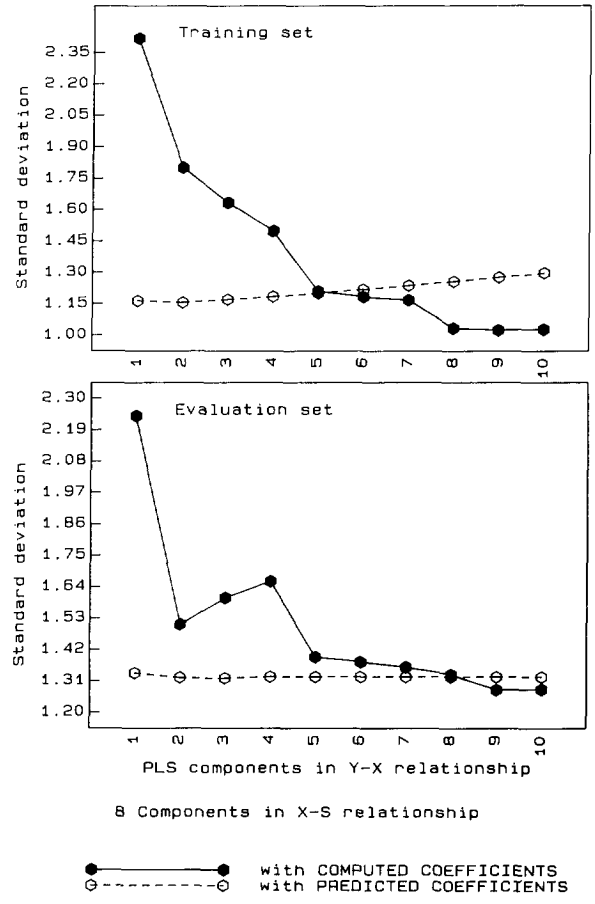

Fig. 7. Standard error of calibration *SEC* (top) and standard error of prediction *SEP* (bottom) with computed and predicted regression coefficients. Response variable: oil. Prediction of instrument A from C, type **X–S**, 13 principal components in **S** blocks.

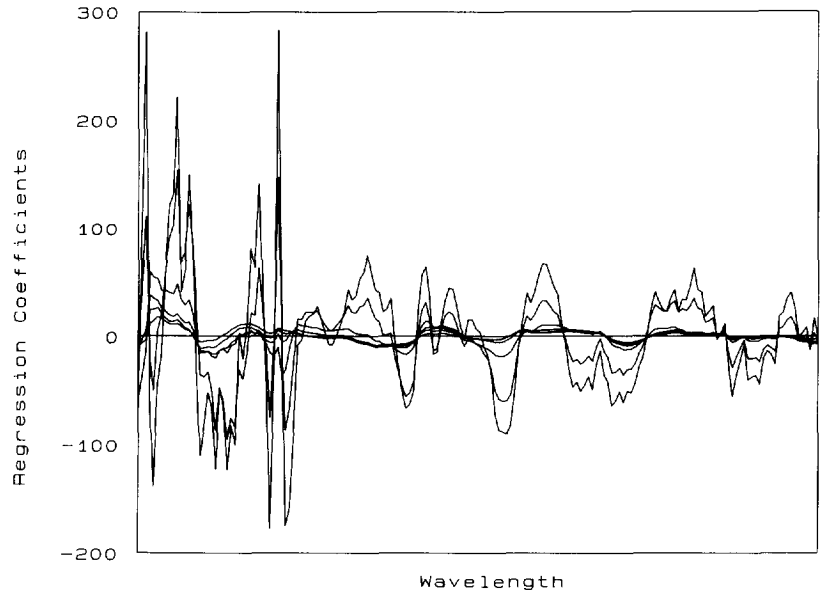

Fig. 8. Regression coefficients of the original variables, instrument C, direct calibration. The coefficients are reported for five to ten latent variables used in the PLS regression: the increase of the number of latent variables corresponds to regression coefficients higher in absolute value.

variables in the y–**X** step). This further decrease in *SEP* is due to the filtering effect of the **X**–**X** step, where some noise is cancelled, and the PLS model obtained with predicted coefficients uses a smaller amount of noise. In some way the transfer of the regression coefficients has the same

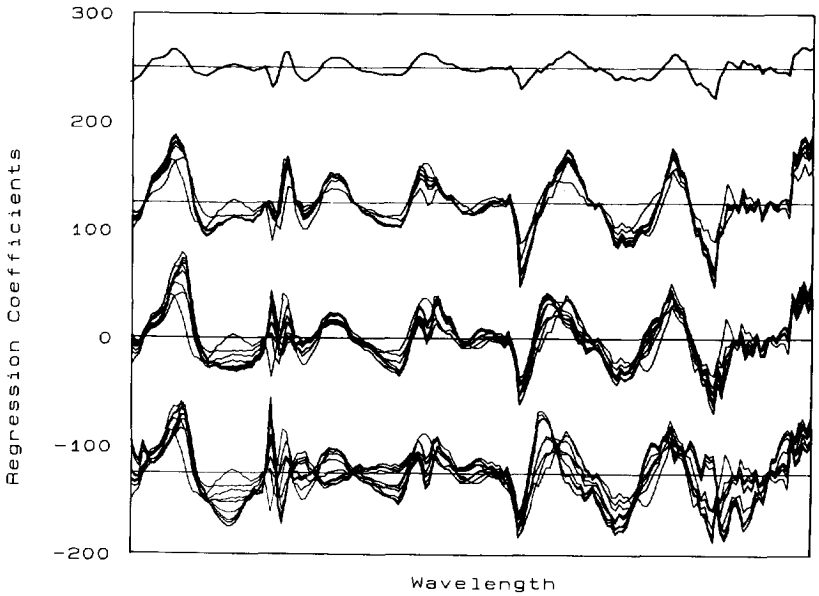

Fig. 9. Regression coefficients of the original variables, instrument C, predicted by instrument D with the use of scores, and from one to ten latent variables in PLS regression. From above: 1, 4, 7 and 10 principal components in the block of scores. The scale of the ordinate refers to the case of seven principal components.

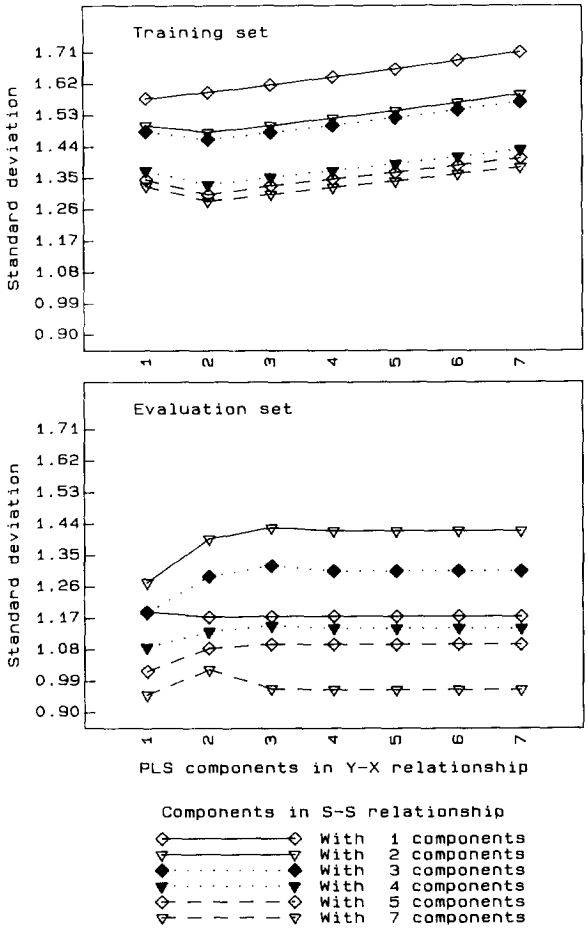

Fig. 10. Standard error of calibration *SEC* (top) and standard error of prediction *SEP* (bottom) with predicted regression coefficients. Response variable: moisture. Prediction of instrument B from A, type S–S, eight principal components in S blocks.

effect as the repetition of a measurement with the same instrument.

The predicted regression coefficients are about the same size as the computed coefficients (see Fig. 3); the differences are not very large, taking into account also the very high correlation between the predictor variables.

4.2. Residual standard deviation on the training set (SEC)

Generally, the residual standard deviation on the training set, *SEC*, with the predicted coefficients $\hat{c}^{AB}$ is greater than that with the computed coefficients c^{A}.

The PLS model with the predicted coefficients is less effective in fitting than the PLS model with the computed (local) coefficients.

What can be surprising is that sometimes *SEC* with the predicted coefficients is higher than *SEP*. This anomaly is reduced when the number of latent variables used in the step **X–X** increases, with the consequent increase of the fitting performance. The observed anomaly seems due to the unicity of the choice of objects in the training and in the evaluation sets. One or a few abnormal

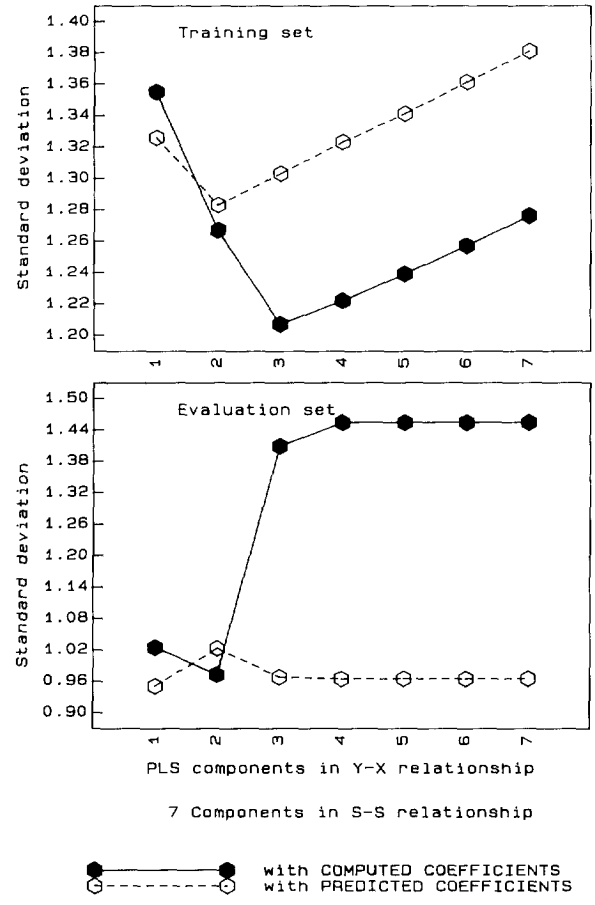

Fig. 11. Standard error of calibration *SEC* (top) and standard error of prediction *SEP* (bottom) with computed and predicted regression coefficients. Response variable: moisture. Prediction of instrument B from A, type S–S, eight principal components in S blocks.

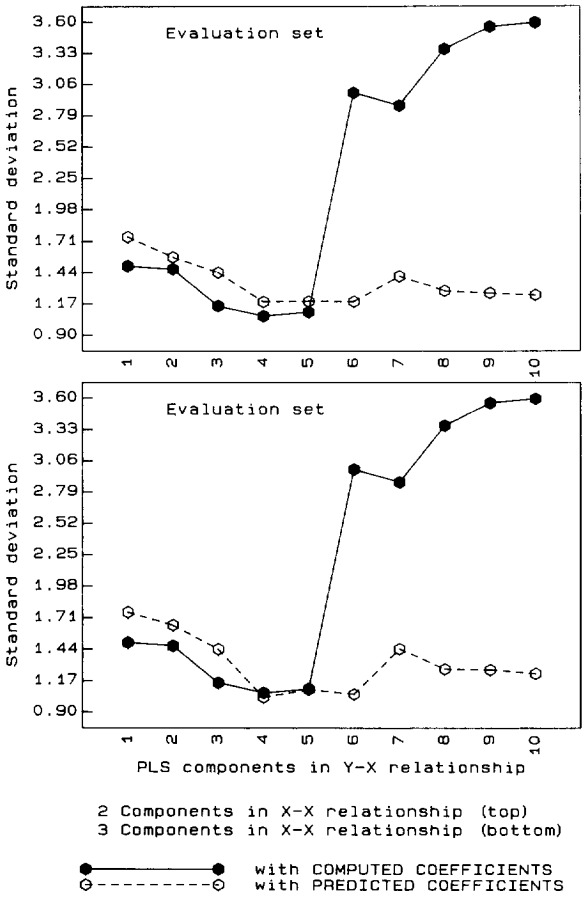

Fig. 12. Standard error of prediction *SEP* with computed and predicted regression coefficients. Five objects used in the training set for the transfer step **X–X**. Response variable: moisture. Prediction of instrument B from A, type **X–X**.

samples can cause the anomaly, as well as irregularities in the dependence of standard deviations on the number of latent variables, as in the case of SEP^B in Fig. 2, from the fifth to the sixth latent variable. A wide systematic study is in progress, where, instead of the single evaluation set, the validation of the whole transfer procedure is performed with the method of cross-validation (with several choices of the number of cancellation groups, from leave-one-out to three cancellation groups) and with the method of the repeated evaluation set (where for each selected percent of the objects in the evaluation set a lot of possible combinations are checked). Results obtained so far indicate that both kinds of anomaly disappear.

4.3. Stability of models

The number of latent variables in both steps, **X–X** and *y*–**X**, is not critical. For example, Fig. 1 shows that about the same *SEP* is obtained with two to five components in the **X–X** step and with one to ten latent variables in the *y*–**X** step. On the contrary the number of latent variables in direct calibration with only a *y*–**X** step appears frequently critical (see for example Fig. 2 as a limiting case, where *SEP* increases from 1.10 with five latent variables to 2.98 with six latent variables). Although this behaviour has been ascribed to the use of the single evaluation set, comparison between the results with computed and predicted coefficients shows that the transfer of calibration causes generally an increase in regularities and in the stability of the regression model with the number of PLS components.

4.4. Use of principal components (S)

When scores of the original predictor variables are used both in direct and in predicted calibration, regularities and stability further increase. The performance (measured by *SEP* with predicted coefficients) is very stable with the number of components retained in the **X–S** step. Fig. 11 compared with Figs. 1 and 2 shows that seven principal components retain the predictive information in the **X** block of 19 predictor variables. The minimum number of principal components to be used in this case without a significant increase in *SEP* was four.

The effect of the increased regularity can be seen also on the regression coefficients computed for the original variables, obtained from the regression coefficients computed or predicted for the scores, using Eq. (22), that rotates the regression coefficients from the space of the principal components to that of the original variables using the matrix of loadings.

Fig. 8 shows that the local regression coefficients $\boldsymbol{c}^C$ computed by PLS in the *y*–**X** step (directly on the original variables) increase in

general very much with the increase in the number of latent variables. Fig. 9 shows that the same coefficients, predicted from instrument D with the use of the **S–S** step, and a different number of principal components, undergo only a limited variation both with the number of latent variables in the ***y*–S** step and with the number of principal components retained in the **X–S** step. The predicted coefficients retain almost the same 'design' of the local coefficients, in spite of the very high correlation among the predictor variables (in the case of sets C and D, of 175 predictor variables, a generic variable has correlation coefficients greater than 0.9999 with about ten contiguous variables).

4.5. Number of samples in the transfer set

Fig. 12 shows an example of the results obtained with the use of only five samples in the transfer step where matrix ${}_{M}\mathbf{M}_{V}^{AB}$ is computed; two or three PLS latent variables were used to obtain **M**. The increase of *SEP* (computed on the same 20 samples of the evaluation set) is very limited. PLS requires at least three samples, generally two samples more than the number of computed latent variables. The number of significant latent variables depends on the complexity of the relationship between the two **X** blocks. However, also when the relationship is simple, so that the transfer requires only one latent variable and three standards, the noise in the measurements limits the goodness of the transfer. PLS regression needs, as all the regression techniques, both an optimal design and enough samples to reduce the effect of the experimental errors.

5. Conclusions

The transfer of the calibration equation in NIRS can be performed without loss of predictive performances with the use of the twin-PLS procedure described here. The use of principal components of the original predictors as variables for the inter-predictor relationship **X–X** makes the procedure also possible in the case of a great number of predictor variables, generally speeds up the procedure, and increases the regularity and stability of the computed and predicted models, without loss of predictive ability.

Some irregularities in the fitting and prediction parameters (residual standard deviations) as a function of the number of PLS components, suggest the use of full validation [9] procedures.

The practical application in a network can suggest the transfer of the spectrum instead of that of the calibration equation: the spectrum transfer procedure suggested here, based on PLS applied to **X–X** or (better) **S–S** blocks, is an alternative to Wang's PDS [3,4] or to Shenk's [1,2] method. The comparison with these two techniques, as well as their possible use in the transfer of the regression equation, requires further work, with the contemporary study of the effect of the number and of the quality of standards used in the transfer step.

Acknowledgments

This research was supported by the Italian Ministry for University and Research (MURST) (40% grant 'Chemiometria' and 60% grant), and by the National Council of Research (CNR, National Committee 'Scienza e Tecnologia dell'Informazione').

References

[1] J.S. Shenk and M.O. Westerhaus, New standardization and calibration procedures for NIRS analytical systems, *Crop Science*, 31 (1991) 1694–1696.

[2] J.S. Shenk, Standardizing NIRS instruments, in R. Biston and N. Bartiaux-Thill (Editors), *Proceedings of the 3rd International Conference on Near Infrared Spectroscopy, Brussel 1990*, Agriculture Research Center, Gembloux, Belgium, 1991, pp. 649–654.

[3] Y. Wang and B.R. Kowalski, Calibration transfer and measurement stability of near-infrared spectrometers, *Applied Spectroscopy*, 46 (1992) 764–771.

[4] Y. Wang, D.J. Veltkamp and B.R. Kowalski, Multivariate instrument standardization, *Analytical Chemistry*, 63 (1991) 2750–2756.

[5] P. Dardenne, R. Biston and G. Simmaeve, Calibration transferability across NIR instruments, in K.I. Hildrum, T. Isaksson, T. Næs and Tandberg (Editors), *Near Infra-Red Spectroscopy*, Ellis Horwood, Chichester, 1992, pp. 453–458.

[6] M. Forina, G. Drava, C. Armanino, R. Boggia, S. Lanteri, R. Leardi, P. Corti, P. Conti, R. Giangiacomo, C. Galliena, R. Bigoni, L. Quartari, C. Serra, D. Ferri, O. Leoni and L. Lazzeri, Il trasferimento della equazione di calibrazione tra differenti strumenti in NIRS, *Atti del Convegno Nazionale di Chimica Analitica, Pavia, 22–25 September 1992*, Società Chimica Italiana Ed., 1992, pp. 44–47.

[7] M. Forina, C. Armanino, R. Giangiacomo, A case study on the transfer of the calibration equation in NIRS, in K.I. Hildrum, T. Isaksson, T. Næs and Tandberg (Editors), *Near Infra-Red Spectroscopy*, Ellis Horwood, Chichester, 1992, pp. 91–96.

[8] E. Marengo and R. Todeschini, A fast method for the calculation of partial least squares coefficients, *Chemometrics and Intelligent Laboratory Systems*, 12 (1992) 117–120.

[9] M. Forina, S. Lanteri, R. Boggia and E. Bertran, Double cross full validation, *Quimica Analitica*, 12 (1993) 128–135.