Elimination of the Uninformative Calibration Sample Subset in the Modified UVE(Uninformative Variable Elimination)–PLS (Partial Least Squares) Method

Jun KOSHOUBU,* Tetsuo IWATA, and Shigeo MINAMI*****

**JASCO Technical Research Laboratory Corporation, 2963-3, Ishikawa, Hachioji, Tokyo 192–0032, Japan*
***Department of Mechanical Engineering, Faculty of Engineering, University of Tokushima,*
Minami-Jyosanjima-cho-2, Tokushima 770–8506, Japan
****Osaka Electro-Communication University, 18-8, Hatsu-cho, Neyagawa, Osaka 572–8530, Japan*

In order to increase the predictive ability of the PLS (Partial Least Squares) model, we have developed a new algorithm, by which uninformative samples which cannot contribute to the model very much are eliminated from a calibration data set. In the proposed algorithm, uninformative wavelength (or independent) variables are eliminated at the first stage by using the modified UVE (Uninformative Variable Elimination)–PLS method that we reported previously. Then, if the prediction error of the ith ($1 \le i \le n$) sample is larger than 3σ, the corresponding sample is eliminated as uninformative, where n is the total number of calibration samples and σ is the standard deviation calculated from the other n–1 samples. Calculation of σ by the leave-one-out manner enhances the ability to identify the uninformative samples. The final PLS model is constructed precisely because both uninformative wavelength variables and uninformative samples are eliminated. In order to demonstrate the usefulness of the algorithm, we have applied it to two kinds of mid-infrared spectral data sets.

(Received September 4, 2000; Accepted October 27, 2000)

In the previous paper, we reported a modified version of the UVE-PLS (Uninformative Variable Elimination–Partial Least Squares) method[1] originally developed by Centner *et al*.[2] The UVE-PLS method is an algorithm to increase the predictive ability of the standard PLS method, where wavelength (or independent) variables which cannot contribute to the model construction very much are eliminated. The key is to make a comparison between experimental variables and purposely-added noise variables with respect to the degree of contribution to the model. The number of the noise variables is the same as the number of the experimental ones. In the modified version, the overfitting problem in the original UVE-PLS algorithm was solved by introducing the PRESS (Prediction Error Sum of Squares) criterion[3,4] and the calculation time was reduced by a factor of several times.

In practical situations, however, unexpected experimental errors or measurement noise are introduced into concentration (or dependent) variables as well as wavelength variables. They must deteriorate the predictive ability of the PLS model. In some cases, samples which cannot be used as calibration data at all might be introduced accidentally for some reasons. Although many robust modeling techniques[5,6] for coping with such problems have been developed so far, most of them use all of the wavelength variables given. Among the wavelength variables, therefore, uninformative ones which cannot contribute to the model will be included. In order to increase the predictive ability of the model, therefore, such uninformative variables should be eliminated in advance, and thereafter uninformative samples should be eliminated. In other words, uninformative samples should be eliminated by taking into account both wavelength variables and concentration variables.

Basing upon such an idea, we have improved our previous method, that is the modified UVE-PLS (MUVE-PLS) method, so that the uninformative samples (or concentration variables) are eliminated. In the present article, we call such a procedure USE (Uninformative Sample Elimination), and the whole method is called MUVE-USE-PLS. In order to demonstrate the effectiveness of the MUVE-USE-PLS method, we have applied it to two kinds of mid-infrared absorption spectral data sets: water–ethanol mixtures[1] and ethyl acetate–acetonitrile mixtures.

Uninformative Sample Elimination (USE)

Figure 1 shows a procedure of the MUVE-USE-PLS method:

(1) First, the modified UVE (MUVE) that we have reported previously is applied for a calibration data set with the final number of latent variables (LVs) A. At this stage, uninformative wavelength variables are eliminated.

(2) For the ith ($1 \le i \le n$) sample, the value of the prediction error $e(i)$ is calculated. At the same time, the value of RMSEP (Root Mean Squares Error of Prediction) is evaluated.

(3) For the ith sample, the standard deviation $\sigma(i)$ of the prediction error is calculated by the leave-one-out manner. It means that $\sigma(i)$ is calculated from the other (n–1) $e(j)$ values except $e(i)$ according to the following equations:

$$e(i) = y_i - \hat{y}_i, \tag{1}$$

$$\mathrm{RMSEP} = \sqrt{\sum_{i=1}^{n} \{e(i)\}^2 \Big/ n}, \tag{2}$$

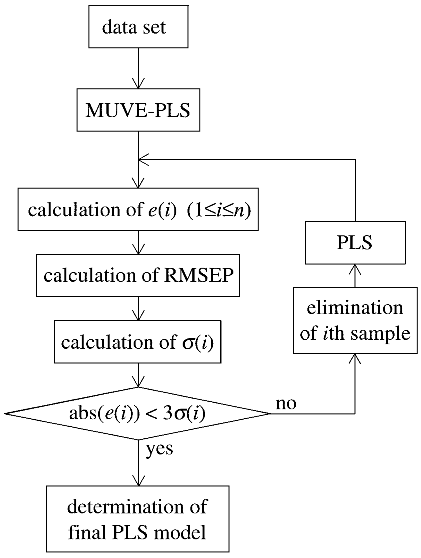
data set
MUVE-PLS
calculation of e(i) (1≤i≤n)
calculation of RMSEP
calculation of σ(i)
abs(e(i)) < 3σ(i)
no
yes
PLS
elimination of ith sample
determination of final PLS model

Fig. 1 Procedure of MUVE-USE-PLS method.

Table 1 Partial mole fractions of ethanol (χ_{eth}) for various water–ethanol mixtures

No.	χ_{eth}	No.	χ_{eth}	No.	χ_{eth}
1	1.000	11	0.317	21	0.072
2	0.881	12	0.281	22	0.058
3	0.788	13	0.248	23	0.041
4	0.695	14	0.224	24	0.036
5	0.621	15	0.198	25	0.024
6	0.553	16	0.171	26	0.019
7	0.493	17	0.150	27	0.012
8	0.441	18	0.124	28	0.006
9	0.399	19	0.110→0.080	29	0.002
10	0.358	20	0.090	30	0.000

Taken and modified by permission of the Society for Applied Spectroscopy from J. Koshoubu, T. Iwata, and S. Minami, *Appl. Spectrosc.*, **2000**, *54*, 148.

$$\overline{e(i)} = \frac{1}{n-1}\sum_{j=1}^{n} e(j) \quad (i \neq j), \tag{3}$$

$$\sigma(i) = \sqrt{\sum_{j=1}^{n}\{e(j) - \overline{e(i)}\}^2 \Big/ (n-1)} \quad (i \neq j), \tag{4}$$

where y_i is the measured value of the ith sample and $\hat{y}_i$ is its predicted value.

(4) We make a comparison of which is larger among the absolute values of $e(i)$, abs$\{e(i)\}$, and $3\sigma(i)$ for $i = 1,2,\ldots,n$.

(5) If abs$\{e(i)\}>3\sigma(i)$, the ith sample is eliminated as an uninformative sample and the PLS model is built from the retained calibration data set with A LVs. Then, we return to (2) again.

(6) If abs$\{e(i)\}\leq 3\sigma(i)$, the final PLS model is built by using the retained data set.

In the proposed method, the ability to discriminate extraordinary samples against ordinary ones is enhanced because of the calculation of the $\sigma(i)$ values by the leave-one-out manner. The MUVE-USE-PLS method can be carried out by a slight change in our previous MUVE-PLS program.

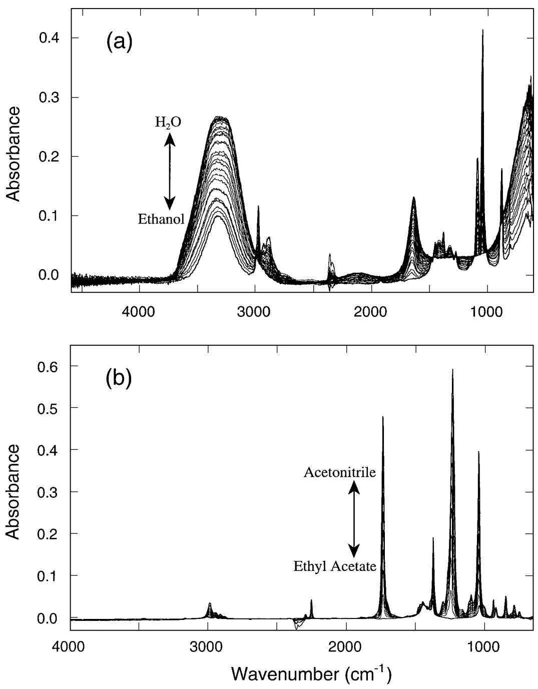

Fig. 2 (a) Mid-infrared absorption spectra of water–ethanol mixtures with various molar fraction ratios listed in Table 1. Taken by permission of the Society for Applied Spectroscopy from J. Koshoubu, T. Iwata, and S. Minami, *Appl. Spectrosc.*, **2000**, *54*, 148. (b) Mid-infrared absorption spectra of ethyl acetate–acetonitrile mixtures with various volume ratios listed in Table 2.

Experimental

Spectral data sets

We have applied the algorithm to two kinds of mid-infrared spectral data sets: data set I and data set II. The data set I is the same as that we previously reported,[1] which consisted of thirty mid-infrared absorption spectra of water–ethanol mixtures with various molar fraction ratios χ_{eth}. In order to demonstrate the sample elimination ability of the USE algorithm in the present article, we purposely varied the ethanol molar fraction of the 19th sample from a true value (χ_{eth} = 0.11) to a wrong one (χ_{eth} = 0.08). The molar fraction ratios of the mixtures are listed in Table 1 and the corresponding spectra are shown in Fig. 2(a) (see in details in Fig. 2 in Ref. 1).

The data set II consists of eleven mid-infrared spectral of ethyl acetate–acetonitrile mixtures with various volume rations. The spectra were measured by using a Fourier transform infrared spectrometer (FT/IR-420, JASCO Co.) with a single reflection attenuated total reflection (ATR) attachment (PIKE Technologies Inc.). Each spectrum consists of 870 variables, corresponding to the absorbance values at wavenumbers from 4000 to 650 cm^{-1}. Volume ratios of ethyl acetate χ_{ea} for the eleven mixtures are listed in Table 2. We purposely varied the ethyl acetate volume ratio of the 9th sample from a true value (χ_{ea} = 80.0) to a wrong one (χ_{ea} = 81.0) again. Figure 2(b) shows the eleven spectra of the mixtures. Ethyl acetate and acetonitrile were obtained from reagent grades (Wako Pure Chemical Co.).

Calibration method

We have applied five kinds of modeling methods for the two calibration data sets. Relations among the five methods are

Table 2 Volume ratios of ethyl acetate (χ_{ea}) for various ethyl acetate–acetonitrile mixtures

No.	χ_{ea}
1	0.0
2	10.0
3	20.0
4	30.0
5	40.0
6	50.0
7	60.0
8	70.0
9	80.0→81.0
10	90.0
11	100.0

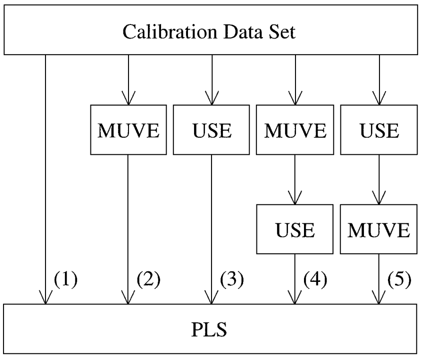

Fig. 3 Five calibration methods: (1) standard PLS method, (2) modified UVE (MUVE) method, (3) USE-PLS method, (4) MUVE-USE-PLS method, (5) USE-MUVE-PLS method.

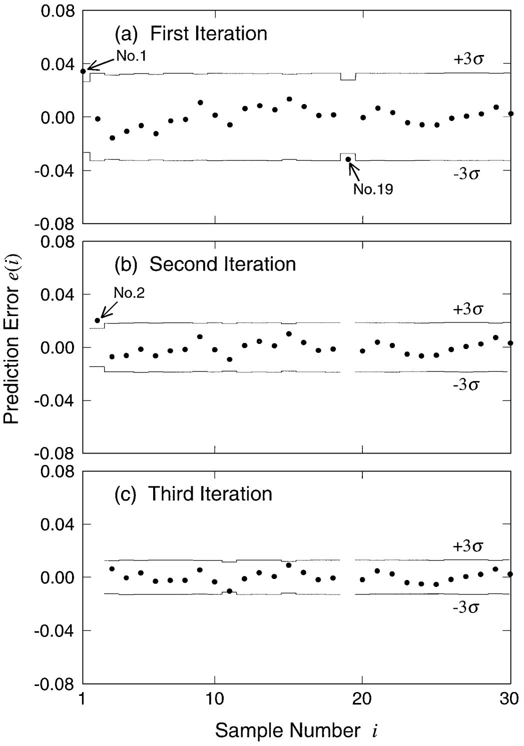

Fig. 4 Application result of the MUVE-USE-PLS method for the calibration data set of water–ethanol mixtures listed in Table 1: (a) plot of prediction error $e(i)$ *vs.* sample number i obtained from the first iteration loop, (b) that obtained from the second iteration loop, and (c) that obtained from the third iteration loop. Two stepwise lines in each figure indicate $\pm 3\sigma(i)$ values as the cutoff levels.

shown in Fig. 3

(1) PLS: Application of the standard PLS method as the criterion minimum RMSEP for the calibration data sets given.

(2) MUVE-PLS: Application of MUVE-PLS method for the calibration data sets. This method is one that we have reported previously.[1]

(3) USE-PLS: Application of the USE algorithm for the calibration data sets given. After the USE, the standard PLS is carried out without the MUVE procedure.

(4) MUVE-USE-PLS: Application of the USE algorithm for the calibration data sets which are processed by the MUVE method. Thereafter, the standard PLS is carried out. This method is one that we propose in this article.

(5) USE-MUVE-PLS: Application of the USE algorithm at first for the calibration data sets given. After the USE, the MUVE-PLS is carried out. This method is the same as the MUVE-USE-PLS but the order of the MUVE and the USE is reversed.

Results and Discussion

Figure 4 shows the application results of the MUVE-USE-PLS method for the spectral data set of thirty kinds of ethanol–water mixtures listed in Table 1. Figure 4(a) shows a plot of the prediction error $e(i)$ as a function of sample number i, which is obtained from the first iteration loop. Two stepwise lines in the figure indicate $\pm 3\sigma(i)$ values, which are used as the criterion for elimination of uninformative samples. From the first iteration, two samples, No.1 and No.19, were eliminated. The sample No.19, the concentration value of which was purposely changed, was reasonably eliminated. Figure 4(b) shows a result obtained from the second iteration loop. At this time, sample No.2 was eliminated. Figure 4(c) shows a result obtained from the third iteration loop and indicates that no sample was eliminated: Individual prediction errors were plotted within $\pm 3\sigma(i)$ values. Among the thirty calibration data, two samples (No.1 and No.2) which were not modified at all were eliminated as uninformative. The reasons for the elimination might be (1) the nonlinearity of spectral intensity and (2) the sparse data density of χ_{eth} in the high concentration region. In the MUVE-USE-PLS algorithm, after all, the final PLS model is built by using the retained 27 samples.

Optimal prediction results obtained from the five different calibration methods are summarized in Table 3. We can find that the MUVE-USE-PLS method proposed in this article brings about a smaller RMSEP value than that of our previous MUVE-PLS method. In other words, the elimination of the two uninformative samples has improved the PLS model by a factor of more than two in the RMSEP sense. On the one hand, the USE-MUVE-PLS method was not able to give a better result than the MUVE-USE-PLS method. This result indicates that elimination of uninformative wavelength variables before that of uninformative samples is important. This is because the number (1038) of wavelength variables are usually much larger than that (30) of concentration variables.

Figure 5 shows the application results of the MUVE-USE-PLS method for the spectral data set of eleven kinds of ethyl acetate–acetonitrile mixtures listed in Table 2. Figure 5(a) shows a plot of the prediction error $e(i)$ as a function of sample

Table 3 Optimal prediction results of experimental data set of water–ethanol mixtures obtained from different calibration methods

Calibration method	RMSEP^{a}	Number of LVs	Number of retained variables	Number of retained samples
(1) PLS	1757	21	1038	30
(2) MUVE-PLS	1053	4	43	30
(3) USE-PLS	1521	15	1038	29
(4) MUVE-USE-PLS	442	4	43	27
(5) USE-MUVE-PLS	794	6	59	29

a. ×10^{-5}.

Table 4 Optimal prediction results of experimental data set of ethyl acetate–acetonitrile mixtures obtained from different calibration methods

Calibration method	RMSEP^{a}	Number of LVs	Number of retained variables	Number of retained samples
(1) PLS	60170	4	870	11
(2) MUVE-PLS	54151	4	30	11
(4) MUVE-USE-PLS	34926	4	30	10

a. ×10^{-5}.

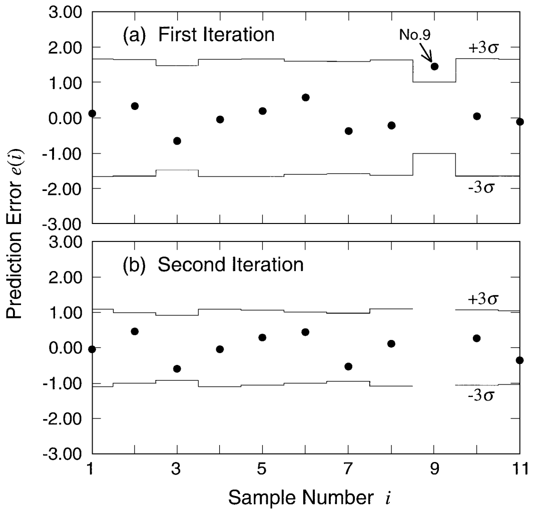

Fig. 5 Application result of the MUVE-USE-PLS method for the calibration data set of ethyl acetate–acetonitrile mixtures listed in Table 2: (a) plot of prediction error $e(i)$ *vs*. sample number i obtained from the first iteration loop, and (b) that obtained from the second iteration loop. Two stepwise lines in each figure indicate ±3$\sigma(i)$ values as the cutoff levels.

number i, which is obtained from the first iteration loop. Two stepwise lines in the figure indicate ±3$\sigma(i)$ values, which are used as the criterion for elimination of uninformative samples as before. From the first iteration, the sample No.9 was eliminated again. The sample No.9, the concentration value of which was purposely changed, was reasonably eliminated. Figure 5(b) shows a result obtained from the second iteration loop and indicates that no sample was eliminated: Individual prediction errors were plotted within ±3$\sigma(i)$ values. Among the eleven calibration data, only No.9 was eliminated as uninformative. In the MUVE-USE-PLS algorithm, after all, the final PLS model is built by using the retained 10 samples. Optimal prediction results obtained from the three different calibration methods are summarized in Table 4. The USE-PLS method and the USE-MUVE-PLS method were not able to give a better RMSEP value than the PLS method. For the spectral data set of ethyl acetate–acetonitrile mixtures, the MUVE-USE-PLS method gave the smallest RMSEP value again among the five different calibration methods.

Conclusion

In order to improve the prediction ability of the standard PLS model, we have proposed a new algorithm where uninformative samples are eliminated from the calibration data set. We called it the MUVE-USE-PLS method, which was an extended version of the MUVE-PLS method that we had reported before.[1] As the criterion for the sample elimination, the value of 3σ was compared with the individual prediction errors, where the value of 3σ was calculated by a leave-one-out manner. Thanks to the procedure, extraordinary data or nonlinearity in the calibration model were able to be eliminated sensitively. This technique might be useful in a practical analysis when a precise model is required.

Abbreviations and Acronisms

PLS: Partial Least Squares
UVE: Uninformative Variable Elimination
MUVE: Modified UVE
USE: Uninformative Sample Elimination
PRESS: Prediction Error Sum of Squares
RMSEP: Root Mean Squares Error of Prediction
LV: Latent Variable

References

1. J. Koshoubu, T. Iwata, and S. Minami, *Appl. Spectrosc.*, **2000**, *54*, 148.
2. V. Centner, D.-L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste, and C. Sterna, *Anal. Chem.*, **1996**, *68*, 3851.
3. D. M. Haaland and E. V. Thomas, *Anal. Chem.*, **1988**, *60*, 1193.
4. T. Iwata and J. Koshoubu, *Bunseki Kagaku*, **1996**, *45*, 85.
5. J. Ferre and F. X. Rius, *Anal. Chem.*, **1996**, *68*, 1565.
6. P. J. Rousseeuw and A. M. Leroy, "*Robust Regression and Outlier Detection*", **1987**, John Wiley and Sons, New York.