Optimal modeling pattern of variables selection on analog complex using UVE-PLS regression

Qianqian Li 1,2, Yue Huang 1,3, and Kuangda Tian 1

1 China Agricultural University, Beijing, 100083, People’s Republic of China
2 Beijing University of Chinese Medicine, Beijing 100029, People’s Republic of China
3 Author to whom any correspondence should be addressed. E-mail: orange07@126.com

Keywords: uninformative variables elimination, partial least squares, near-infrared, variable selection

Abstract
This study aimed to determine the composition of chemical complex by partial least square (PLS) regression models combined with uninformative variable elimination (UVE). The near-infrared (NIR) spectra of the forty samples were determined and then UVE was used to compress full NIR spectra from 12011 redundant variables to dozens of variables. Finally, 54, 16, 27, 31 and 42 variables were selected by UVE for 2,2,4-Trimethylpentane, Heptane, Cyclohexane, Ethyl formate and Butyl acetate respectively. Selected variables were used as the inputs of PLS model for quantitative analysis which made the prediction of the model more robust and accurate compared with the conventional PLS.

Introduction
Nowadays, near-infrared (NIR) spectroscopy (4000 ~ 12500 cm⁻¹) are widely employed for qualitative and quantitative analysis in agriculture and food quality evaluation [1–3]. NIR spectroscopy can be used in analysis due to the molecular bonds of C–H, N–H, O–H and C=O whose combination tone and the overtones absorption are all in the NIR region [4]. As a rapid and well-established technique for complex analysis, it is nondestructive and requiring minimal sample processing prior to analysis [5] compared to destructive and tedious conventional measurements.

NIR has been proved to be a powerful analytical tool combined with partial least squares (PLS) which is a commonly used multivariate calibration [6, 7]. However, full spectra have much redundant information in the spectral data such as noise, background and overlapping information that would influence the model development and prediction [8]. It is essential and important to compress the large amount of data and to choose the useful and relevant information before executing PLS modeling [9–11].

UVE has been used to solve such problems and improve the quality of models by eliminating the irrelevant information and noise from the data matrix [12–14]. Generally a PLS regression is followed to develop the model with the chosen variables. Therefore, calibration and prediction models obtained by the selecting characteristic information may be better than those obtained by the full-spectrum because the characteristic wavelengths instead of raw spectra are used for developing the model [15, 16].

In this study, the UVE-PLS method was applied to NIR data of 40 samples for analysis consisting of five components including 2,2,4-Trimethylpentane, Heptane, Cyclohexane, Ethyl formate and Butyl acetate. Specially, thirty calibration samples were used to develop PLS regression model and other ten samples was sequentially chosen as the independent prediction set to evaluate the accuracy of the model.

Experimental
Samples and materials
2,2,4-Trimethylpentane, Heptane, Cyclohexane, Ethyl formate, Butyl acetate were mixed with different proportions for preparation of 40 solutions. Hexane, Ethyl acetate, Benzaaldehyde, Toluene, Acetophenone were...
added to each sample so as to make the matrix complicate and to check the feasibility of this method. CCl₄ was employed as the solvent. Each sample was prepared with the weight percentage of 2,2,4-Trimethylpentane, 0.854% ∼ 26.58%, Heptane, 0.944% ∼ 43.54%, Cyclohexane, 0.157% ∼ 5.60%, Ethyl formate, 0.140% ∼ 8.058% and Butyl acetate, 0.344% ∼ 12.70%. To avoid collinearity, the forty samples were not arranged in concentration gradient sequence. Each component of the sample was orderly separated into thirty calibration samples and ten prediction samples.

Spectral analysis
The spectral data were acquired by an NIR spectrophotometer (Spectrum One NTS, Perkin Elmer, USA) over the wavelength range from 12,000 to 4000 cm⁻¹. The spectrum was scanned 32 times with the averaged spectrum acquired, and the resolution was 2 cm⁻¹. A quartz cell with a 1.0 mm path length was employed. CCl₄ solution was used as a blank.

UVE-PLS regression
The UVE-PLS algorithm is actually performs UVE before doing PLS regression. UVE is a variable selection method based on the stability analysis of regression coefficient. Stability value was acquired by hundreds of repeating Monte Carlo cross calculation. PLS was based on the ability to mathematically correlate spectral data to a property matrix of interest while simultaneously accounting for all other significant spectral factors that interfered the spectrum.

Model evaluation
Results of linear modeling of PLS were investigated and compared for quantitative analysis of the compositions in the complex system. Different processes of spectral data were evaluated and then optimized at the stage of calibration. Generally, a good model should have higher correlation coefficient (R²) and residual prediction deviation (RPD) values and simultaneously lower RMSE values.

Results and discussion
Spectrum of the complex
The raw spectra of the forty samples were presented in figure 1. Each spectrum has 12001 data points. Some regions such as 4000 ∼ 4500 cm⁻¹, 4700 cm⁻¹, 5500 ∼ 6000 cm⁻¹, 7000 ∼ 7400 cm⁻¹ and 8100 ∼ 8500 cm⁻¹ were sensitive transmitted, while the remaining region of the spectral curves were flatten at the range of 7000 ∼ 7400 cm⁻¹ and 8100 ∼ 8500 cm⁻¹ which were much weaker than the other peaks because these regions were the second overtones of C–H and certainly with less information. Forty samples with their spectral curves overlapped.

Modelling by UVE-PLS
Figure 2 was the informative variables remained after UVE. The dark and the blue spectrum were respectively representing the NIR spectrum of each pure chemical and the complex. Variables with relative low S/N were
and 5804 cm$^{-1}$ and CH$_2$ groups, respectively. The bands evolving around 8195 cm$^{-1}$ are assigned to the stretching and bending vibration. Peaks from 5820 cm$^{-1}$ almost removed. Region 8500 cm$^{-1}$ with strong signal were overtones and combinations of C–H and O–H. If the chemical did have the independent featured peak then it would be selected by UVE, if not, this algorithm may also choose the peak from the overlapped spectrum.

Figure 2(A) showed the 54 variables out of 12001 after UVE processing of 2,2,4-Trimethylpentane as in vertical red lines. The peak from 4544.5 cm$^{-1}$ to 4550.5 cm$^{-1}$ were assigned to the combination tone of C–H of CH$_2$ stretching and bending vibration. Peaks from 5820 cm$^{-1}$ to 5825 cm$^{-1}$ were assigned to the first overtone region 2 × C–H stretching vibration of CH$_3$ and CH$_2$. The peaks of 8290 cm$^{-1}$ to 8299 cm$^{-1}$, 8358 cm$^{-1}$ to 8365 cm$^{-1}$ were assigned to the second overtone of 3 × C–H stretching vibration of CH$_2$ and CH$_3$. The regions between 5820 cm$^{-1}$ to 5825 cm$^{-1}$ and 8358 cm$^{-1}$ to 8365 cm$^{-1}$ were the characteristic peaks of 2,2,4-Trimethylpentane as shown in the pure spectrum of figure 2(a), which demonstrated the variables were just accurately selected.

Figure 2(B) had only 16 variables remaining after UVE of Heptane. Regions of 5063 cm$^{-1}$ to 5064 cm$^{-1}$ and 5201 cm$^{-1}$ were the combination tone of C–H. 5454 cm$^{-1}$, 5854 cm$^{-1}$ and 5872 cm$^{-1}$ were assigned to the first overtone of C–H stretching vibration. 5454 cm$^{-1}$, 5854 cm$^{-1}$ and 5878 cm$^{-1}$ were the characteristic peak of Heptane. However, 5063 cm$^{-1}$ to 5064 cm$^{-1}$ and 5201 cm$^{-1}$ may come from Cyclohexane, which in return indicated UVE selected not only the analyzing sample but also other samples which possessed the overlapped variables in the complex.

Figure 2(C) demonstrated the selected 27 variables of Cyclohexane in vertical red lines. 5035 cm$^{-1}$ and 5048 cm$^{-1}$ to 5049 cm$^{-1}$ were the combination tone of C–H, 5297 cm$^{-1}$ to 5298 cm$^{-1}$, 5500 cm$^{-1}$ to 5503 cm$^{-1}$ and 5605 cm$^{-1}$ to 5606 cm$^{-1}$ were assigned to the first overtone of C–H stretching vibration. The peaks around 7172 cm$^{-1}$ to 7175 cm$^{-1}$ and 7082 cm$^{-1}$ were caused by the 2C–H stretching and C–H deformation vibration of CH$_3$ and CH$_2$ groups, respectively. The bands evolving around 8195 cm$^{-1}$ to 8196 cm$^{-1}$ were assigned to the second overtone of 3 × C–H stretching vibration of CH$_2$ and CH$_3$. Therefore, as executing UVE, it did not mean selecting the variables with larger absorbance but those with higher regression coefficients.

Figure 2(D) was the retained 31 variables of Ethyl formate after UVE. The region among 4601 cm$^{-1}$ to 4602 cm$^{-1}$, 4635 cm$^{-1}$ to 4636 cm$^{-1}$, 4664 cm$^{-1}$ to 4672 cm$^{-1}$ and 4681 cm$^{-1}$ to 4682 cm$^{-1}$ were the combination tone of C–H and C=O stretching vibrations of HCOOCH$_2$CH$_3$ which was special in Ethyl formate as the independent characteristic peak and can be identified from others and the variables were exactly with the largest absorbance as well as the highest regression coefficients.

Figure 2(E) indicated the 42 chosen variables of Butyl acetate. 4700 cm$^{-1}$ and 4766 cm$^{-1}$ to 4767 cm$^{-1}$ were the combination tone of C–H and C=O stretching vibrations of CH$_3$COOCH$_2$CH$_2$CH$_3$, 5784 cm$^{-1}$ to 5785 cm$^{-1}$ and 5804 cm$^{-1}$ to 5805 cm$^{-1}$ were the characteristic variables of Butyl acetate and UVE also selected some overlapped variables among 5784 cm$^{-1}$ to 5785 cm$^{-1}$ and 5804 cm$^{-1}$ to 5805 cm$^{-1}$.

**Evaluation of the modelling**

To improve the predictability of the model, the screening of wavelength range was of particular importance in the regression modeling of spectral data. So the calibration and prediction models of the 40 samples were almost removed.

---

**Figure 2.** The retained variables after UVE selection. (A) 2,2,4-Trimethylpentane; (B) Heptane; (C) Cyclohexane; (D) Ethyl formate; (E) Butyl acetate.
built from the retained variables of 2,2,4-Trimethylpentane, Heptane, Cyclohexane, Ethyl formate and Butyl acetate with only 54, 16, 27, 31 and 42 of the initial 12001 with the variables no more information than noise eliminated.

Table 1. PLS model parameters with or without UVE for the five components in complex.

| Methods  | Components                | LVs | $R^2_c$ (%) | $R^2$ (%) | RMSEC | RMSEP | RPD  | RSD (%) |
|----------|----------------------------|-----|-------------|-----------|-------|-------|------|--------|
| UVE-PLS  | 2,2,4-Trimethyl-pentane    | 6   | 96.13       | 99.96     | 0.0122 | 0.029 | 11.91| 0.05   |
|          | Heptane                    | 6   | 94.67       | 99.10     | 0.0978 | 0.179 | 3.22 | 0.21   |
|          | Cyclohexane                | 4   | 95.58       | 99.57     | 0.0578 | 0.038 | 13.50| 0.05   |
|          | Ethyl formate              | 4   | 95.81       | 99.94     | 0.0168 | 0.010 | 37.03| 0.02   |
|          | Butyl acetate              | 6   | 92.64       | 95.13     | 0.0537 | 0.013 | 3.16 | 0.05   |
| PLS      | 2,2,4-Trimethyl-pentane    | 6   | 93.62       | 98.51     | 0.0439 | 0.131 | 2.61 | 0.24   |
|          | Heptane                    | 6   | 72.74       | 77.60     | 0.2720 | 0.674 | 0.85 | 0.78   |
|          | Cyclohexane                | 4   | 94.23       | 99.06     | 0.0478 | 0.086 | 6.04 | 0.11   |
|          | Ethyl formate              | 4   | 96.72       | 99.91     | 0.0118 | 0.013 | 27.92| 0.03   |
|          | Butyl acetate              | 6   | 77.96       | 82.90     | 0.0560 | 0.201 | 0.918| 0.82   |

Figure 3. UVE-PLS regressions of calibration and prediction sets. (A) 2,2,4-Trimethylpentane; (B) Heptane; (C) Cyclohexane; (D) Ethyl formate; (E) Butyl acetate.
The retained variables were used to establish PLS regression by data central-normalization before modeling. Regression analysis did not generate any outliers. Table 1 lists the LVs of 2,2,4-Trimethylpentane, Heptane, Cyclohexane, Ethyl formate and Butyl acetate, respectively, namely, the best UVE-PLS calibration and prediction model was composed of 6, 6, 4, 4 and 6 factors.

The results of calibration and prediction sets of the five chemicals after variables selecting were presented in figure 3. As shown, the predicted values associated to UVE-PLS generally correlate the measured. According to figure 3(A) and table 1, 2, 2, 4-Trimethylpentane showed a good linearity with the determination coefficient \((R^2)\) of 99.96% and RMSEC of 0.0122. While the model established for the real test depended on its ability to predict external samples, the RMSEP, RSD (%) and the RPD were 0.029, 0.05 and 11.91, respectively. High \(R^2\) of 99.96% (shown in table 1) and the RPD was much over limit of 3, which meant the predictability of 2, 2, 4-Trimethylpentane was acceptable. Similarly, figures 3(B)–(E) showed the regressions of calibration and prediction sets of Heptane, Cyclohexane, Ethyl formate and Butyl acetate, respectively.

Results of UVE-PLS models of the other four components were listed in table 1. From table 1 it was not only found that the four calibration models of Heptane, Cyclohexane, Ethyl formate and Butyl acetate were all well established with \(R^2\) (99.10%, 99.57%, 99.94%, 95.13%) and RMSEC(0.0978, 0.0578, 0.0168, 0.0537) but also the models performed good predictability with the RMSEP (0.179, 0.038, 0.010, 0.013) and RPD (3.22, 13.50, 37.03, 3.16). RSD% were respectively 0.21%, 0.05%, 0.02% and 0.05% which indicate the precision was all acceptable and dozens of selected variables were enough to develop the model in theory and in results above.

Comparison of conventional PLS

To compare two methods of UVE-PLS and PLS on the selected variables and raw spectrum, the same pretreatment method and LVs were chosen in developing UVE-PLS models. Performance of selected variables by UVE was compared with PLS models using full spectrum and the results were presented in table 1. Poor predictive ability in table 1 may result from the low S/N redundant variables which had little contribution on developing the model.

\(R^2\) and RPD of 2,2,4-Trimethylpentane, Heptane, Cyclohexane, Ethyl formate and Butyl acetate with selected variables by UVE were all higher than that with the raw spectrum by PLS. Meanwhile, the RMSEC, RMSEP with selected variables by UVE of 2,2,4-Trimethylpentane, Heptane, Cyclohexane, Ethyl formate and Butyl acetate were all lower than that with the raw spectrum by PLS. The predictability index of RMSEP and RPD were mostly used to evaluate whether the models performed well or not. Selected variables of the five components were with far lower RMSEP values (0.028, 0.179, 0.038, 0.010, 0.013 versus 0.131, 0.674, 0.086, 0.013, 0.201) and higher RPD values (11.91, 3.22, 13.50, 37.03, 3.16 versus 2.61, 0.852, 6.04, 27.92, 0.918) than that with the full spectrum. And RSD% (0.05, 0.21, 0.05, 0.02, 0.05 versus 0.24, 0.78, 0.11, 0.03, 0.82) of single PLS processing also increased a lot, revealing the poor predictability by the raw spectra.

It was observed from the table 1 that, the predictability of models by UVE-PLS with selected variables apparently outperformed that by single PLS with the full spectrum by removing the influence of low S/N variations and with featured variables.

Conclusion

In this study, the 40 samples of 2,2,4-Trimethylpentane, Heptane, Cyclohexane, Ethyl formate and Butyl acetate were surveyed by NIR spectroscopy. UVE was utilized to eliminate uninformative variables from 12001 to 54, 16, 27, 31 and 42 respectively. The selected variables included the combination tone and the overtones absorption of C–H, O–H and C=O. The values of appraisal index, namely \(R^2\), RMSEC and RMSEP, indicated that the prediction precision of UVE-PLS regression were superior to PLS method. Compared with PLS models, the RMSEC and RMSEP of UVE-PLS models were generally much smaller. Results of PLS after UVE indicated UVE was a powerful variable selection approach in chemometrics and can be combined with NIR or other spectroscopy to determine the chemical composition in complex precisely and rapidly.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID iDs

Yue Huang @ https://orcid.org/0000-0001-8563-232X
References

[1] Baranwal Y, Román-Ospino A D, Keyvan G and Ramachandran R 2019 Int. J. Pharm. 565 419–36
[2] Roza-Delgado B, Garrido-Varo A, Soldado A and Arrojo A G 2017 Food Control 76 74–81
[3] Liu F, He Y and Wang L 2008 Anal. Chim. Acta 610 196–204
[4] Luypaert J, Massart D L and Vander-Heyden Y 2007 Talanta 72 865–83
[5] Sjöholm K, Tenhunen J, Tammisola J, Pietila K and Horne S 2018 J. Am. Soc. Brew. Chem. 54 135–40
[6] Mabood F, Hussain J, Jabeen F and Abbas J 2018 Food Additives and Contaminants: Part A 35 1052–1060
[7] Yang J X, Lou X, Yang Q, Yang H and Liu C Y 2019 Anal. Lett. 52 2188–2202
[8] Su W H and Sun D W 2018 Comprehensive Reviews In Food Science And Food Safety 17 220–39
[9] Chen Q, Jiang P and Zhao J 2010 Spectrochimica Acta Part A 76 50–5
[10] Milanez K D, Nóbrega T C, Nascimento D S, Galvão R K and Pontes M J 2017 Anal. Chim. Acta 984 76–85
[11] Giglio C and Brown S D 2018 J. Chemom. 32 e3034
[12] Andries J P, Heyden Y V and Buydens L 2017 Anal. Chim. Acta 982 57–67
[13] Yan H, Han B, Wu Q, Jiang M and Gui Z 2011 Spectrochimica Acta Part A 79 174–84
[14] Center V, Massart S, de Noord O E, de Jong S, Vandeginste B M and Sterna C 1996 Anal. Chem. 68 3851–8
[15] Liu X W, Cui X Y, Yu X M and Shao X G 2017 Chin. Chem. Lett. 28 1447–52
[16] Kuligowski J, Carrión D, Quintás G, Garrigues S and Guardia M 2012 Food Chem. 131 353–9