Chemometrics and Intelligent Laboratory Systems 131 (2014) 31–36

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

# Variable selection based on locally linear embedding mapping for near-infrared spectral analysis

Ruifeng Shan, Wensheng Cai, Xueguang Shao *

*State Key Laboratory of Medicinal Chemical Biology, Research Center for Analytical Sciences, College of Chemistry, Nankai University, Tianjin 300071, PR China*

ARTICLE INFO

*Article history:*
Received 30 May 2013
Received in revised form 5 December 2013
Accepted 14 December 2013
Available online 21 December 2013

*Keywords:*
Variable selection
Locally linear embedding
Monte Carlo cross validation
Partial least squares regression
Forward stepwise selection
Near-infrared spectroscopy

ABSTRACT

Locally linear embedding (LLE) is a nonlinear dimensionality reduction method that can preserve the relationship between samples in the mapping space. The neighbors in high dimensional space will keep their relative position in LLE space. A method based on the effect of the variables on the relative position of the samples in LLE space was proposed for variable selection in NIR spectral analysis. In the method, the spectra are mapped into LLE space with all variables at first, and then the mapping is repeated by removing a variable from the spectra. Therefore, the movement of the samples in LLE space caused by a variable can be used to evaluate the effect of the variable on the spectra. The variables that cause a large movement will be the important ones to affect the relationship of the spectra. For further selection of the informative variables specific to the target component, a forward stepwise selection is applied to the variables selected by LLE method. To validate the performance of the proposed method, it was applied to the partial least squares (PLS) modeling of three NIR spectral datasets of corn, pharmaceutical tablets and tobacco lamina samples. Results show that the proposed method can effectively select the informative variables from the NIR spectra, and build a parsimonious model by using several tens of selected variables.


## 1. Introduction

Multivariate calibration methods have been extensively used in the near-infrared (NIR) spectroscopic quantitative analysis. Much attention has been paid to variable selection in NIR spectral analysis for building accurate and parsimonious models. Methods based on optimization algorithms such as genetic algorithm (GA) [1–3], particle swarm optimization (PSO) [4], and interval partial least squares (iPLS) [5,6] have been applied to search the optimal subset of variables. The results of these methods demonstrated that better prediction can be obtained using the selected variables than the full spectrum. However, optimization algorithms generally need larger number of parameters and are time-consuming. Therefore, simple and efficient methods based on statistics were used for the problem. Uninformative variable elimination (UVE) and its variants [7–9], randomization test (RT) [10], Bayesian variable selection [11], successive projections algorithm (SPA) [12,13], etc. have been proposed. These methods evaluate the variables statistically and then select the variables with higher or lower statistical value. Additionally, stepwise selection method was widely used in NIR spectral analysis due to the simplicity. Competitive adaptive reweighted sampling (CARS) [14,15] selects variables in a stepwise and efficient way. In our recent works, methods based on the detection of the influential variables [16] and latent projective graph (LPG) [17] were proposed. These works proved that satisfactory PLS models can be established using several tens or even several informative variables.

Manifold learning techniques are developed for nonlinear dimensionality reduction, which discover compact representations of high dimensional data by recovering the underlying low dimensional manifold. Manifold learning algorithms such as locally linear embedding (LLE) [18], isometric mapping (Isomap) [19], Hessian LLE [20], and Laplacian Eigenmap [21] have been proposed. These methods are all based on Euclidean distance for exploiting the neighborhood information as same as locally weighted regression (LWR) and soft independent modeling of class analogy (SIMCA). Among these methods, LLE is a local method similar with LWR. The both of them need to select the nearby points and determine the weights. However, the weights in LLE are optimized by minimizing the reconstruction errors, while the weights in LWR are calculated according to the distance between the predicted data and the training data points [22]. Additionally, compared with LWR, LLE preserve the relationship between samples in mapping space and find the embedding in a noniterative way. Therefore, LLE and its extensions have become a promising techniques and used to solve the problem of dimension reduction of high dimensional data, such as face recognition [18,23], NIR spectra [24,25], gene expression [26,27], etc. Furthermore, it was also widely used in the data visualization.

In this work, a method for variable selection is proposed based on the effect of a variable on the mapping distance in LLE space. The spectra

* Corresponding author at: College of Chemistry, Nankai University, Tianjin 300071, PR China. Tel.: +86 22 23503430; fax: +86 22 23502458.
*E-mail address:* xshao@nankai.edu.cn (X. Shao).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.12.002

are mapped into the low dimensional space by LLE operation. According to the principle of LLE, the relationship, i.e., the relative position, between the samples in the spectral space will not change in the mapping space of LLE. However, if a variable that significantly affect the spectra is removed, the relative position of the samples in LLE space may change accordingly. Therefore, the change of the position, i.e., the movement of the samples, in LLE space caused by removal of a variable can be used to evaluate the effect of the variable on the spectra. Taking the movement calculated by the average Euclidean distance as a criterion, the variables that cause a large movement will be the important ones to the spectra.

## 2. Theory and calculations

LLE is a nonlinear mapping method that computes low-dimensional, neighborhood-preserving embeddings of high-dimensional inputs [18,23]. The basic idea of LLE is to approximate each data point by a linear combination of its neighbors and to find a low dimensional configuration of data points. In LLE algorithm, each data point and its neighbors are assumed to lie on or close to a locally linear patch of a manifold. Therefore, a data point can be approximated as a linear combination of its neighbors based on the assumption of local linearity. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a set of $N$ points in a high $D$ dimensional data space $R^D$. The corresponding set of $N$ points in a low $d$ dimensional data space $R^d$ is denoted as $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_N\}$. For each data point $\mathbf{x}_i$, find its $K$ neighbors by using Euclidean distance at first, and then the reconstruction weights $\mathbf{w}_i$ that best reconstruct $\mathbf{x}_i$ linearly by its $K$ nearest neighbors can be optimized by minimizing the following cost function:

$$\varepsilon(\mathbf{W}) = \sum_{i=1}^{N} \left| \mathbf{x}_i - \sum_{j=1}^{K} w_{ij}\mathbf{x}_j \right|^2 \tag{1}$$

under the constraints that each vector of reconstruction weights $\mathbf{W}$ sums to unity. It should be noted that the constrained weights are invariant to rotations, rescalings, and translations. Therefore, the optimized reconstruction weights can characterize intrinsic geometric properties of each neighborhood in the high dimensional space.

In order to preserve the local geometry of the data in low dimensional space, the embedding $\mathbf{Y}$ of $\mathbf{X}$ can be reconstructed with the weights by minimizing the embedding cost function:

$$\phi(\mathbf{Y}) = \sum_{i=1}^{N} \left| \mathbf{y}_i - \sum_{j=1}^{K} w_{ij}\mathbf{y}_j \right|^2 \tag{2}$$

and subjecting to the following constraints:

$$\begin{cases} \sum_{i=1}^{N} \mathbf{y}_i = \mathbf{0} \\ \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i{}^{\mathrm{T}}\mathbf{y}_i = \mathbf{I} \end{cases} \tag{3}$$

In the calculations, a new sparse symmetric and positive semi-definite matrix $\mathbf{M}$ is constructed based on the matrix $\mathbf{W}$ [18]. Then, the constrained minimization problem can be converted to solving eigenvalue problem of the matrix $\mathbf{M}$ as calculated by

$$\mathbf{M} = (\mathbf{I} - \mathbf{W})^{\mathrm{T}}(\mathbf{I} - \mathbf{W}) \tag{4}$$

The eigenvectors of $\mathbf{M}$ associated with the smallest $d$ nonzero eigenvalues constitute the low embedding outputs $\mathbf{Y}$.

The mapping quality is rather sensitive to the number $K$ of the nearest neighbors, as indicated in Eq. (1). In this paper, therefore, the residual variance, defined by $1-\rho^2_{\mathbf{D_X D_Y}}$, is employed as a quantitative measure of the embedding results when the low dimensionality $d$ is 3, and the optimal value for $K$ is determined by [18]:

$$K_{opt} = \arg\min\left(1-\rho^2_{\mathbf{D_X D_Y}}\right) \tag{5}$$

where $\mathbf{D_X}$ and $\mathbf{D_Y}$ are the matrices of Euclidean distances (between pairs of points) in the input data matrix $\mathbf{X}$ and the output data matrix $\mathbf{Y}$, respectively, and $\rho$ is the standard linear correlation coefficient of $\mathbf{D_X}$ and $\mathbf{D_Y}$.

Based on the property of LLE mapping, a method for variable selection in modeling of NIR spectra is proposed. Clearly, the relative position of the samples in high dimensional space can be kept by LLE transformation. If a variable that significantly affect the spectra is removed from the spectra, the relative position of the samples would be affected in LLE space. The change of the position in LLE space caused by removal of a variable can be used to evaluate the effect of the variable on the spectra. Therefore, the method maps the full spectra into LLE space at first, and then the mapping is repeated by removing a variable from the spectra. With the data in LLE space, the movement of the samples caused by the removal of each variable can be calculated by averaging the movement (Euclidean distance) of the samples. If the movements are ranked in a descending order, a sequence indicating the significance of the variables to the spectra can be obtained.

Only the influence of the variables on the spectra is involved in the method. For building an efficient model of a component, however, the variables specific to the component are more effective. Therefore, a forward stepwise selection (FSS) is applied to the selected variables by LLE. In the calculations, PLS models with an increasing number of the variables along the ranked sequence are evaluated with the root mean squared error of cross-validation (RMSECV). When a variable makes the RMSECV smaller, the variable is accepted, otherwise, rejected. The accepted variables are taken as the final selected variables to build the PLS model of a target component. LLE-FSS-PLS is named for the method. Additionally, Monte Carlo cross-validation (MCCV) is adopted in this study. In the calculation, half of the samples in the calibration set are randomly sampled to building the model and the remaining half is used for validation. 1000 repetitions are used for calculating the RMSECV.

## 3. Descriptions of the datasets

### 3.1. Dataset 1

The dataset was downloaded from http://software.eigenvector.com/Data/Corn/index.html, which consists of NIR spectra, measured with three spectrometers, and the moisture, oil, protein and starch values of 80 corn samples. The spectra measured on mp5 NIR spectrometer and the moisture values are used in this study. Each spectrum was recorded in the wavelength range 2498–1100 nm (4003–9091 $cm^{-1}$) with the digitization interval 2 nm. 56 spectra were selected, by using Kennard–Stone (KS) algorithm, as calibration set and the other 24 spectra were taken as prediction set.

### 3.2. Dataset 2

The dataset was downloaded from the website of international diffuse reflectance conference (IDRC), http://www.idrc-chambersburg.org/shootout2002.html. It contains the spectra of 655 pharmaceutical tablets from two spectrometers (Foss NIRSystems and Multitab Spectrometers) and the spectra of each instrument were split into a calibration set with 155 spectra, a validation set with 40 spectra and a test set with 460 spectra. Transmittance mode was used and the spectral region is from 600 to 1898 nm with 2 nm increments. The assay values of the active pharmaceutical ingredient (API) were included. In this work, the

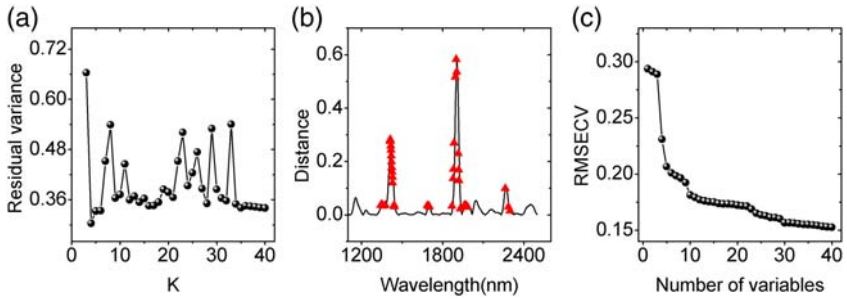


**Fig. 1.** Parameter optimization and variable selection of the dataset 1. (a) Variation of residual variance with the number of the nearest neighbors. (b) Average mapping distance caused by removal of each variable. The triangles indicate the finally selected variables by LLE-FSS. (c) Variation of RMSECVs with the number of selected variables.

spectra of the second instrument were used and, as suggested, the 155 and 460 spectra were used as calibration and prediction sets, respectively. Furthermore, the variables from 600 to 1622 nm were used and the possible outliers (Nos. 19, 122, 126, 127 in calibration set and Nos. 11, 145, 267, 295, 294, 342, 313, 341, and 343 in prediction set) [28,29] were removed before calculation.

### 3.3. Dataset 3

The dataset consists of diffuse reflectance NIR spectra of 307 tobacco lamina samples and their total sugar contents. The spectra were measured on an MPA FT-NIR spectrometer (Bruker, Germany). Each NIR spectrum was recorded in the wavenumber range of 3999.7–11995.3 $cm^{-1}$ with the digitization interval ca. 4 $cm^{-1}$. The contents of total sugar were measured on an Auto Analyzer III (Bran + Luebbe, Germany) following the industrial standard procedures. In the calculations, the variables from 3999.7 to 9002.5 $cm^{-1}$ (1298 data points) were used, and the calibration set was selected by using KS method. 204 spectra were used as calibration set and the other 103 spectra were taken as prediction set.

## 4. Results and discussion

### 4.1. Variable selection for dataset 1

As described above, the number $K$ of the nearest neighbors in LLE needs to be determined before the mapping. If $K$ is set too small, the mapping does not reflect any global property in the high dimensional space, in contrast, if $K$ is too large, small-scale structures of the high dimension data may be eliminated, making the mapping lose its nonlinear character. In order to investigate the effect of $K$ on the mapping, the residual variance is calculated with different $K$ values ranging from 3 to 40 using the calibration set of the dataset. Fig. 1(a) shows the variation of the residual variance with parameter $K$. It can be seen that the variances fluctuate significantly with the increase of the nearest neighbors. Therefore, the optimal number, 4, with the minimum of the variance is used for this dataset.

In order to investigate the effect of the variables on the spectra, the average mapping distance in LLE space is calculated. Fig. 1(b) shows the distribution of the average mapping distance caused by removal of each variable. It can be seen that the variables with the larger mapping distance are located around 1400 nm and 1900 nm.

To further selection of the informative variables for building the model of predicting moisture, FSS is performed with the variable sequence ordered with the distance values. Fig. 1(c) shows the variation of the RMSECVs of the calibration set with the increase of the accepted variable number along the variable sequence. As mentioned above, only the variables that make the RMSECV smaller are accepted. Therefore, only the accepted variables are shown in the figure. It is clear that, for this dataset, 40 variables are finally selected as the significant ones. Furthermore, the rationality of the variable sequence can be demonstrated by the variation trend of the curve, i.e. the effect of the variable on the RMSECV value becomes smaller and smaller along the sequence.

The triangles in Fig. 1(b) display the 40 finally selected variables. It can be seen that most of the selected variables are concentrated around 1400 nm and 1910 nm. The two bands clearly correspond to the water absorption [30,31] and the first overtones of O-H stretching [32]. However, well resolved peaks in the figure are not selected, the reason may

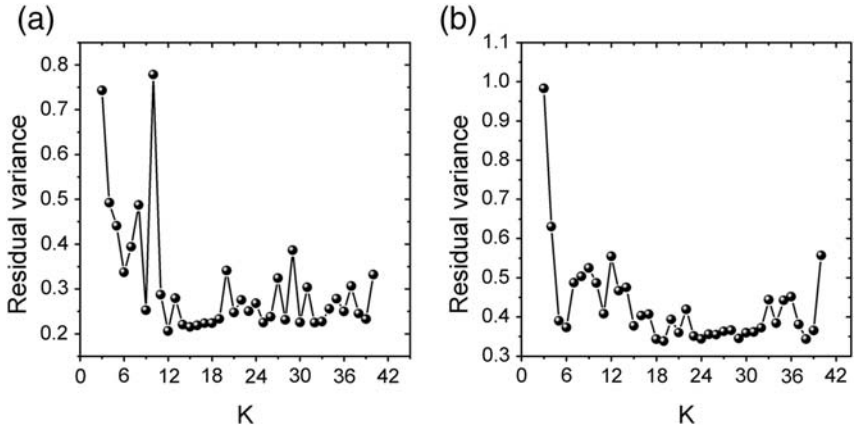


**Fig. 2.** Variation of residual variance with the number of the nearest neighbors for dataset 2 (a) and dataset 3 (b).

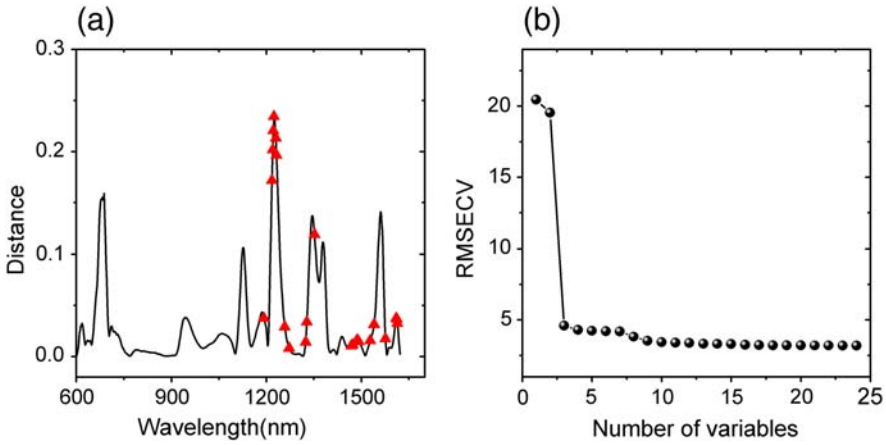


**Fig. 3.** Distribution of average mapping distance caused by removal of each variable (a) and variation of the RMSECVs with the number of selected variables (b) for dataset 2. The triangles in (a) indicate the finally selected variables by LLE-FSS.

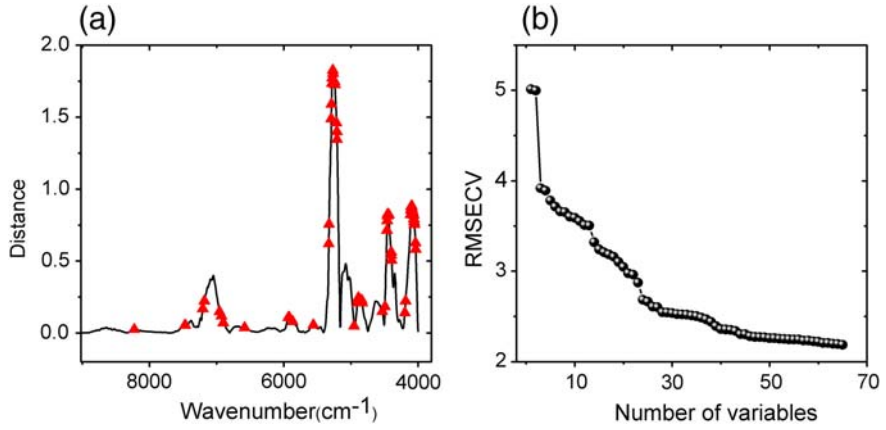


**Fig. 4.** Distribution of average mapping distance caused by removal of each variable (a) and variation of the RMSECVs with the number of selected variables (b) for dataset 3. The triangles in (a) indicate the finally selected variables by LLE-FSS.

be that the informative variables are those reflecting the difference between the spectra. Well resolved peaks may be the common feature of the spectra, instead of the variables specific to the target component. Therefore, the results appear to be reasonable for the model to predict the moisture content.

### 4.2. Variable selection for dataset 2 and 3

Similarly as did for dataset 1, Fig. 2 shows the variation of the residual variance of the calibration set obtained with different number of the nearest neighbors for dataset 2 and 3. It is clear that the optimal $K$ is 12 and 19, respectively.

Fig. 3(a) shows the average mapping distance caused by removal of each variable for dataset 2, and Fig. 3(b) shows the variation of the RMSECVs obtained by FSS with the increase of the number of variables. The former figure suggests that the variables with large distance are concentrated in the wavelength regions of 1100–1622 nm, and the latter figure indicates that only 24 variables are selected for the model. From Fig. 3(a), it can be seen that most of the variables labeled with the triangles are the variables with large distance. This further proves the rationality of the variable sequence ordered by the distance value. However, it is difficult to explain the selected variables for the dataset, because detailed information of the active pharmaceutical ingredient is not contained. On the other hand, comparing Fig. 3(b) with Fig. 1(c), there is a very sharp

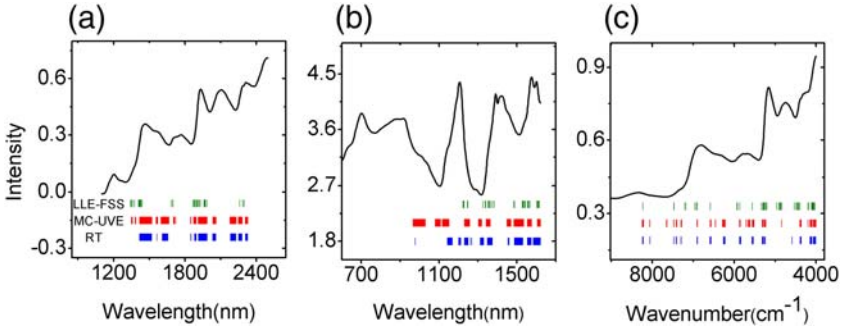


**Fig. 5.** Mean spectrum and the selected variables by LLE-FSS, MC-UVE and RT methods for dataset 1 (a), dataset 2 (b), and dataset 3 (c). The selected variables are labeled with vertical short bars.

decrease in the latter when the number of variables changes from 2 to 3, and changes slightly from 9 to 24. The reason may be that a few variables with larger distance of the selected sequence are more significant to the target component. In fact, the top 7 variables in the sequence correspond to the ones in Fig. 1(a) with very large distance values.

Fig. 4(a) and (b) show, respectively, the average mapping distance caused by removing each variable and the variation of the RMSECVs with the number of selected variables for dataset 3. Unlike the results for dataset 1 and 2, more variables are selected and the distribution of the selected variables is more dispersive. This maybe explained by the prediction target of the model. The sugar content is a total amount of different sugars, such as glucose, levulose, sucrose and maltose etc. Nevertheless, most of the selected variables are around 4400 $cm^{-1}$ and 6900 $cm^{-1}$, which can be assigned to the combinations of O – H/C – O vibrations and the first overtone of O – H stretching [33]. Another difference between the results and that of dataset 1 and 2 is that a few variables with small distance value are selected, as shown in Fig. 4(b). This implies that, for this dataset, the distance parameter is not sufficient enough for evaluating the significance of the variables. In such cases, FSS is necessary to compensate the deficiency.

### 4.3. Evaluation of the selected variables

Fig. 5 shows the mean spectrum and a comparison of the selected variables for the three datasets by LLE-FSS, MC-UVE and RT method. The latter two methods were proposed in our previous works [9,10]. It can be seen that, for all the three datasets, the number of selected variables is in a descending order of MC-UVE, RT and LLE-FSS, although the difference for dataset 3 is not obvious. However, most of the selected variables by LLE-FSS are covered by that by the latter two methods. This can be another proof of the rationality of the selected variables.

To further investigate the selected variables, RMSEPs and correlation coefficients of the prediction set were summarized in Table 1. In the table, the results of the three datasets obtained by PLS with full spectrum without and with continuous wavelet transform (CWT) preprocessing were included as references. CWT is used as a preprocessing technique for removing the variant background in the spectra [34]. In the calculation of CWT, Haar wavelet with a scale parameter 20 was used. Clearly, CWT can improve the models and produce better results. For further comparison of the variable selection methods, the number of variable selected by LLE-FSS is much less than that selected by MC-UVE and RT. For the RMSEP, however, slightly worse results are obtained by LLE-FSS-PLS models for dataset 1 and 3 except that the result of dataset 2 is significantly improved. Similarly, the correlation coefficients obtained by LLE-FSS-PLS are close to that obtained by the other two methods for dataset 1 and 3, but the result of dataset 2 is better than that of the other methods. Although the performance of the models for dataset 1 and 3 is not improved, parsimonious models are obtained using only a small part of the informative variables. Therefore, LLE-FSS method should be efficient technique for selecting informative variables to make the PLS model more parsimonious. In the case of predicting the active pharmaceutical ingredient, only several variables can be used for building an efficient model.

**Table 1**
Prediction results of the three datasets by different PLS models.

| Methods | Variable number[a] | LV number[b] | R | RMSEP(σ) |
|---|---|---|---|---|
| Corn dataset | | | | |
| PLS | 700 | 5 | 0.8313 | 0.188 |
| CWT + PLS | 700 | 5 | 0.8639 | 0.173 |
| CWT + MC-UVE + PLS | 231 (185–260) | 5 | 0.8952 | 0.146 (0.0011) |
| CWT + RT + PLS | 184 (70–240) | 5 | 0.9027 | 0.149 (0.0047) |
| CWT + LLE-FSS + PLS | 40 | 5 | 0.8944 | 0.156 |
| Pharmaceutical tablet dataset | | | | |
| PLS | 512 | 4 | 0.9520 | 4.896 |
| CWT + PLS | 512 | 4 | 0.9654 | 4.186 |
| CWT + MC-UVE + PLS | 156 (130–225) | 4 | 0.9691 | 3.937 (0.0257) |
| CWT + RT + PLS | 109 (75–195) | 4 | 0.9688 | 3.703 (0.0683) |
| CWT + LLE-FSS + PLS | 24 | 4 | 0.9833 | 2.604 |
| Tobacco lamina dataset | | | | |
| PLS | 1298 | 10 | 0.8813 | 2.561 |
| CWT + PLS | 1298 | 10 | 0.8899 | 2.450 |
| CWT + MC-UVE + PLS | 158 (75–205) | 10 | 0.8878 | 2.460 (0.0284) |
| CWT + RT + PLS | 109 (40–235) | 10 | 0.8902 | 2.440 (0.0507) |
| CWT + LLE-FSS + PLS | 65 | 10 | 0.8876 | 2.503 |

[a] For MC-UVE and RT method, the average of 100 independent runs is used. The numbers in the parenthesis in the column of variable number are the minimal and the maximum number, and the value in the parenthesis in the column of LV number is the standard deviation of the 100 values.
[b] LV number means latent variable number.

## 5. Conclusions

A new method named as the LLE-FSS for variable selection in NIR spectral analysis was proposed based on the effect of the variables on the relative position of the samples in LLE space. With the proposed method, the importance of the variables to the spectra can be evaluated by the movement of the samples in LLE space caused by removal of a variable. With three different NIR datasets of corn, pharmaceutical tablets and tobacco lamina samples, the method was proved very efficient for variable selection to make the PLS model more parsimonious. Compared with full-spectral PLS models, better prediction results can be obtained. Therefore, the method may be a good tool for variable selection in NIR spectral analysis due to its simplicity and efficiency.

## Acknowledgements

This study was supported by National Natural Science Foundation of China (No. 21175074).

## References

[1] K. Hasegawa, Y. Miyashita, K. Funatsu, GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists, J. Chem. Inf. Comput. Sci. 37 (1997) 306–310.
[2] R. Leardi, A.L. Gonzalez, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.
[3] R. Leardi, M. Seasholtz, R. Pell, Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data, Anal. Chim. Acta. 461 (2002) 189–200.
[4] Q. Shen, J.H. Jiang, C.X. Jiao, S.Y. Huan, G.L. Shen, R.Q. Yu, Optimized partition of minimum spanning Tree for piecewise modeling by particle swarm algorithm. QSAR Studies of antagonism of angiotensin II antagonists, J. Chem. Inf. Model. 44 (2004) 2027–2031.
[5] F. Lindgren, P. Geladi, S. Rannar, S. Wold, Interactive variable selection (IVS) for PLS. Part 1: theory and algorithms, J. Chemom. 8 (1994) 349–363.
[6] F. Lindgren, P. Geladi, A. Berglund, M. Sjostrom, S. Wold, Interactive variable selection (IVS) for PLS. Part II: chemical applications, J. Chemom. 9 (1995) 331–342.
[7] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, Anal. Chem. 68 (1996) 3851–3858.
[8] S.F. Ye, D. Wang, S.G. Min, Successive projections algorithm combined with uninformative variable elimination for spectral variable selection, Chemom. Intell. Lab. Syst. 91 (2008) 194–199.
[9] W.S. Cai, Y.K. Li, X.G. Shao, A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra, Chemom. Intell. Lab. Syst. 90 (2008) 188–194.
[10] H. Xu, Z.C. Liu, W.S. Cai, X.G. Shao, A wavelength selection method based on randomization test for near-infrared spectral analysis, Chemom. Intell. Lab. Syst. 97 (2009) 189–193.
[11] T. Chen, E. Martin, Bayesian linear regression and variable selection for spectroscopic calibration, Anal. Chim. Acta. 631 (2009) 13–21.
[12] M.C.U. Araujo, T.C.B. Saldanha, R.K.H. Galvao, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemom. Intell. Lab. Syst. 57 (2001) 65–73.
[13] R.K.H. Galvao, M.C.U. Araujo, W.D. Fragoso, E.C. Silva, G.E. Jose, S.F.C. Soares, H.M. Paiva, A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm, Chemom. Intell. Lab. Syst. 92 (2008) 83–91.
[14] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Anal. Chim. Acta. 648 (2009) 77–84.

[15] K.Y. Zheng, Q.Q. Li, J.Z. Wang, J.P. Geng, P. Cao, T. Sui, X. Wang, Y.P. Du, Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra, Chemom. Intell. Lab. Syst. 112 (2012) 48–54.
[16] X.G. Shao, M. Zhang, W.S. Cai, Multivariate calibration of near-infrared spectra by using influential variables, Anal. Methods 4 (2012) 467–473.
[17] X.G. Shao, G.R. Du, M. Jing, W.S. Cai, Application of Latent projective graph in variable selection for near infrared spectral analysis, Chemom. Intell. Lab. Syst. 114 (2012) 44–49.
[18] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[19] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[20] D.L. Donoho, C. Grimes, Hessian eigenmaps: locally linear embedding techniques for high-dimensional data, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 5591–5596.
[21] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (2003) 1373–1396.
[22] W.S. Cleveland, S.J. Devlin, Locally weighted regression: an approach to regression analysis by local fitting, J. Am. Stat. Assoc. 83 (1988) 596–610.
[23] L.K. Saul, S.T. Roweis, Think globally, fit locally: unsupervised learning of low dimensional manifolds, J. Mach. Learn. Res. 4 (2003) 119–155.
[24] N. Qi, Z.Y. Zhang, Y.H. Xiang, P.D.B. Harrington, Locally linear embedding method for dimensionality reduction of tissue sections of endometrial carcinoma by near infrared spectroscopy, Anal. Chim. Acta. 724 (2012) 12–19.
[25] J. Jacques, C. Bouveyron, S. Girard, O. Devos, L. Duponchel, C. Ruckebusch, Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data, J. Chemom. 24 (2010) 719–727.
[26] M. Pillati, C. Viroli, Supervised locally linear embedding for classification: an application to gene expression data analysis, Proceedings of 29th Annual Conference of the German Classification Society (GfKI 2005), 2005, pp. 15–18.
[27] B. Li, C.H. Zheng, D.S. Huang, L. Zhang, K. Han, Gene expression data classification using locally linear discriminant embedding, Comput. Biol. Med. 40 (2010) 802–810.
[28] K.H. Norris, G.E. Ritchie, Assuring specificity for a multivariate near-infrared (NIR) calibration: the example of the Chambersburg Shoot-out 2002 data set, J. Pharm. Biomed. 48 (2008) 1037–1041.
[29] Z.P. Chen, L.M. Li, R.Q. Yu, D. Littlejohn, A. Nordon, J. Morris, A.S. Dann, P.A. Jeffkins, M.D. Richardson, S.L. Stimpson, Systematic prediction error correction: a noval strategy for maintaining the predictive abilities of multivariate calibration models, Analyst 136 (2011) 98–106.
[30] D. Jouan-Rimbaud, D.L. Massart, R. Leardi, O.E. De Noord, Genetic algorithms as a tool for wavelength selection in multivariate calibration, Anal. Chem. 67 (1995) 4295–4301.
[31] H. Buning-Pfaue, Analysis of water in food by near infrared spectroscopy, Food Chem. 82 (2003) 107–115.
[32] J.H. Jiang, R.J. Berry, H.W. Siesler, Y. Ozaki, Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with application to mid-infrared and near-infrared spectroscopic data, Anal. Chem. 74 (2002) 3555–3565.
[33] A.M.C. Davies, C.E. Miller, Tentative assignment of the 1440 nm absorption band in the near infrared spectrum of crystalline sucrose, Appl. Spectrosc. 42 (1988) 703–704.
[34] X.G. Shao, A.K.M. Leung, F.T. Chau, Wavelet: a new trend in chemistry, Acc. Chem. Res. 36 (2003) 276–283.