Available online at www.sciencedirect.com

Journal of Pharmaceutical and Biomedical Analysis 46 (2008) 568–573

JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS

www.elsevier.com/locate/jpba

Short communication

Determination of total polyphenols content in green tea using FT-NIR spectroscopy and different PLS algorithms

Quansheng Chen[a,*], Jiewen Zhao[a], Muhua Liu[b], Jianrong Cai[a], Jianhua Liu[a]

[a] *School of Food & Biological Engineering, Jiangsu University, 212013 Zhenjiang, PR China*
[b] *College of Engineering, Jiangxi Agricultural University, 330045 Nanchang, PR China*

Received 4 August 2007; received in revised form 23 October 2007; accepted 25 October 2007
Available online 1 November 2007

Abstract

This paper attempted the feasibility to determine content total polyphenols content in green tea with near infrared (NIR) spectroscopy coupled with an appropriate multivariate calibration method. Partial least squares (PLS), interval PLS (iPLS) and synergy interval PLS (siPLS) algorithms were performed comparatively to calibrate regression model. The number of PLS components and the number of intervals were optimized according to root mean square error of cross-validation (RMSECV) in calibration set. The performance of the final model was evaluated according to root mean square error of prediction (RMSEP) and correlation coefficient (R) in prediction set. Experimental results showed that the performance of siPLS model is the best in contrast to PLS and iPLS. The optimal model was achieved with $R = 0.9583$ and RMSEP = 0.7327 in prediction set. This study demonstrated that NIR spectroscopy with siPLS algorithm could be used successfully to analysis of total polyphenols content in green tea, and revealed superiority of siPLS algorithm in contrast with other multivariate calibration methods.

Keywords: Near infrared spectroscopy; PLS; iPLS; siPLS; Green tea; Total polyphenols

1. Introduction

Tea polyphenols substance is of great interest due to its beneficial medicinal properties [1]. There is increasing evidence that polyphenols substances found in tea can enhance general health. Recently, many researches have suggested that antioxidants found in polyphenols substances, may have an important role to prevent cardiovascular disease [2], chronic gastritis [3,4] and some cancers [5,6,7]. Additionally, polyphenols compounds are mainly responsible for the characteristic astringent and bitter taste of tea brews [8]. In recent years, many methods of analysis have been employed to determine total polyphenols content in tea, such as colorimetric measurements and titration method with potassium permanganate [9]. However, these methods are all time-consuming. Near infrared reflectance spectroscopy is a fast, accurate and non-destructive technique that can be employed as a replacement of time-consuming chemical method.

Near infrared (NIR) spectroscopy has been proved to be a powerful analytical tool used in the agricultural, nutritional, petrochemical, textile and pharmaceutical industries [10–19]. Since 1990s, attempts have been made to simultaneously predict alkaloids and phenolic substance in green tea leaves using near infrared spectroscopy [20,21]. Some studies on the quantitative analysis of total antioxidant capacity in green tea by NIR are also reported in 2003 and 2004 [22,23]. Recently, Some researchers applied near infrared spectroscopy to analyze simultaneously the content of free amino acids, caffeine, total polyphenols and amylose in green tea [24–27].

For these works mentioned above, near infrared spectral data calibrations are often made with the classical multivariate calibration analysis, for example, partial least squares (PLS) regression and artificial neural net (ANN). Many spectral pretreatment methods have been developed to reduce the effects of variations in the spectral data that are not related to the chemical variations in the samples [24,25]. These methods often improve the calibrations, but they do not take into account that there might be spectral regions that do not contain any information about the chemical variations in the samples [28]. In fact, one of the major problems in multivariate data analysis is to select appropriate

[*] Corresponding author. Tel.: +86 511 8790308; fax: +86 511 8780201.
E-mail address: q.s.chen@hotmail.com (Q. Chen).

0731-7085/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2007.10.031

spectral region in order to achieve the best performance. Some researchers have constructed PLS models in different spectral regions to quantify compositions content in tea, however, these regions were selected manually [27]. Spectral regions selected manually might as well weaken the performance of the calibration model without prior experienced knowledge about NIR spectroscopy.

In recent years, both theoretical and experimental evidence have been published that spectral region selection can significantly improve the performance of these calibration techniques [29,30]. It is so important to select specific regions where contain much information that generate more stable models with superior interpretability, and this will produce the lowest prediction error. Some methods have been recently described in the literature to implement spectral region selection and have used PLS for multivariate calibration in each subset [30].

A graphically oriented local modeling procedure called interval partial least squares (iPLS) was presented for usage on NIR spectral data [30,31]. It has shown that selective optimum interval in the spectral data could give precision prediction models. Norgaard et al. (2000) also proposed a method called Synergy Interval PLS (siPLS) to select several intervals spectra data which could split the data set into a number of intervals (variable-wise) and calculates all possible PLS model combinations of two, three or four intervals.

In this research, we investigate and compare the results provided by PLS, iPLS and siPLS procedures for NIR quantitative analysis of total polyphenols content in green tea. We systematically studied the different steps that have to be gone through in model calibration. The number of PLS factor and the number of regions intervals were optimized according to the root mean square error of cross-validation (RMSECV) in calibration set. The performance of the final model was evaluated according to the root mean square error of prediction (RMSEP) and the correlation coefficient (R) in prediction set.

2. Materials and methods

2.1. Sample preparation

All tea samples came from different provinces in China, and they were all already on stock within 4 months period. Taking into consideration the heterogeneity of tea samples, the samples would be ground before analysis. For the grinding, the whole tealeaves were put into a small electric coffee mill and ground during 10 s. After this procedure, the powders are sieved with a mesh width 500 μm and these sieved powders are used for the further analysis.

2.2. Chemical analysis

Total polyphenols content were reference measured by a photometric Folin-Ciocalteu assay according to a proposed international standard method [9]. Absorbance (E) at 540 nm of the reaction solution is determined in a 1 cm light-path cell by a Lengguang-752 spectrophotometer (Lengguang Optical Instrument Ltd. Co., Shanghai, China). The calibration standard is gallic acid.

2.3. Spectra collection

The NIR spectra were collected in the reflectance mode using the Antaris. Near infrared spectrophotometer (Thermo Electron Co., USA) with an integrating sphere. Each spectrum was the average spectrum of 32 scans. The range of spectra is from 10,000 to 4000 cm^{-1}, and the data were measured in 3.856 cm^{-1} intervals, which resulted in 1557 variables.

The standard sample accessory holder was used for performing the tea spectra collection. The sample accessory holder is sample cup specifically designed by Thermo Electron Co. For each tea sample, 10 ± 0.1 g of dry tealeaves were filled into the sample cup in the standard procedure depending upon the bulk density of materials. The corresponding amount of dry tea powders was densely packed into the sample cup and then compressed by closing it. When spectra collecting, tea sample was collected one times every rotating the cup 120° angle, thus, each sample was collected three times. The average of the three spectra, which were collected from the same tea sample, was used in the next analysis. The temperature was kept around 25 °C and the humidity was kept at a steady level in the laboratory.

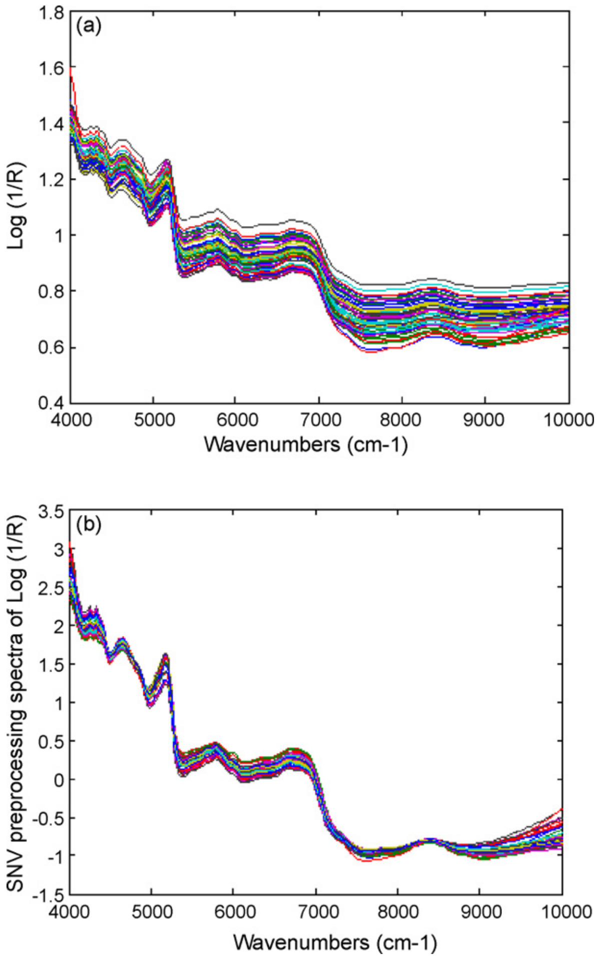

Fig. 1. Spectra of tea obtained from (a) raw data and (b) SNV preprocessing data.

2.4. Spectral preprocessing [32]

Fig. 1(a) presents the raw spectral profile of tea, and raw spectral data need be conducted on spectral preprocessing. In this study, three spectral preprocessing methods were applied comparatively, and they were standard normal variate transformation (SNV), mean centering (MC) and multiplicative scatter correction (MSC). SNV is a mathematical transformation method of the log(1/*R*) spectra used to remove slope variation and to correct for scatter effects. Each spectrum is corrected individually by first centering the spectral values, and then the centered spectrum is scaled by the standard deviation calculated from the individual spectral values. Mean centering is to calculate the average spectrum of the data set and subtract the average from each spectrum. MSC is another important procedure for the correction of scatter light, on the basis of different particle sizes, and the technique is also used to correct for additive and multiplicative effects in the spectra.

To compare results obtained by three preprocessing methods, SNV preprocessing method is as good as MSC and much better than MC. This is because dry tealeaves are particle solids, which bring to scatter light easily; while, SNV and MSC spectral preprocessing methods can remove slope variation and correct light scatter due to different particle sizes. Therefore, SNV spectral preprocessing method was applied in this research, and the spectra after SNV preprocessing are presented in Fig. 1(b).

2.5. Software

All algorithms were implemented in Matlab V7.0 (Mathworks, USA) under Windows XP. Result Software (Antaris System, Thermo Electron Co., USA) was used in NIR spectral data acquisition. The iPLS and siPLS algorithms used in this work were downloaded from http://www. models.kvl.dk/.

3. Results and discussion

3.1. Spectra investigation

Fig. 1(a) shows the spectra for the original data. Seen from Fig. 1(a), the intensive spectral peaks are mainly in the region of 4000–8500 cm^{-1}. In additionally, some spectral regions exhibiting a high noise level (e.g. 10000–9000 and 5000–4000 cm^{-1}) should be excluded in data processing. According to the investigation of spectra, the spectral region of 5002.44–9002.08 cm^{-1} was selected in next analysis.

3.2. Calibration of models

All 71 samples are divided into two subsets. One of subset is called calibration set, which is used to build model, and other is called prediction set, which is used to test the robustness of model. To avoid bias in subset division, this division is made as follows: all samples had been sorted according to their respective *y*-value (viz. the reference measurement value of total polyphenols content). In order to come to a 3/2 division of calibration/prediction spectra, the two spectra of every five samples are divided into the prediction set, so that finally the calibration set contains 43 spectra, the remaining 28 spectra constitute the prediction set. Seen from Table 1, the range of *y*-value in calibration set covers the range in the prediction set, therefore the distribution of the samples is appropriate in calibration and prediction set.

The performance of the final PLS model is evaluated according to the root mean square error of prediction (RMSEP) and the correlation coefficient (*R*) in prediction set. For RMSECV, a leave-one-sample-out cross-validation is performed: the spectrum of one sample of the calibration set is deleted from this set and a PLS model is built with the remaining spectra of the calibration set. The left-out sample is predicted with this model and the procedure is repeated with leaving out each of the samples of the calibration set. The RMSECV is calculated as follows:

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_{\setminus i} - y_i)^2}{n}} \qquad (1)$$

where n is the number of samples in the calibration set, y_i is the reference measurement result for sample i and $\hat{y}_{\setminus i}$ is the estimated result for sample i when the model is constructed with sample i removed. The number of PLS factors included in the model is chosen according to the lowest RMSECV. This procedure is repeated for each of the preprocessed spectra. For the test set, the root mean square error of prediction (RMSEP) is calculated as follows:

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (2)$$

where n is the number of samples in the test set, y_i is the reference measurement result for test set sample i and $\hat{y}_i$ is the estimated result of the model for test sample i.

Finally the model with the overall lowest RMSECV will be selected as final model. Correlation coefficients between the predicted and the measured value are calculated for both the calibration and the test set, which are calculated as follows Eq. (3), where $\bar{y}$ is the mean of the reference measurement results for all samples in the train and test sets.

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (3)$$

Table 1
The reference measurements and sample numbers in calibration and prediction set

Set	Units (%)	S.N.[a]	Range	Mean	S.D.[b]
Calibration set	g/g	43	14.93–25.46	19.61	2.71
Prediction set	g/g	28	15.84–24.39	19.69	2.61

[a] S.N., sample number.
[b] S.D., standard deviation.

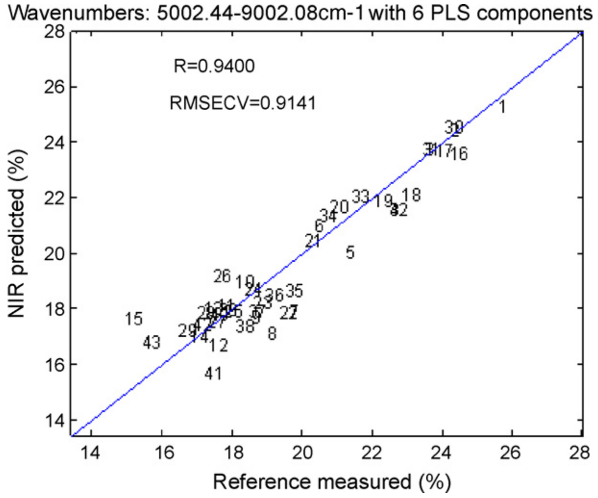

Fig. 2. Reference measured vs. NIR predicted by PLS in calibration set.

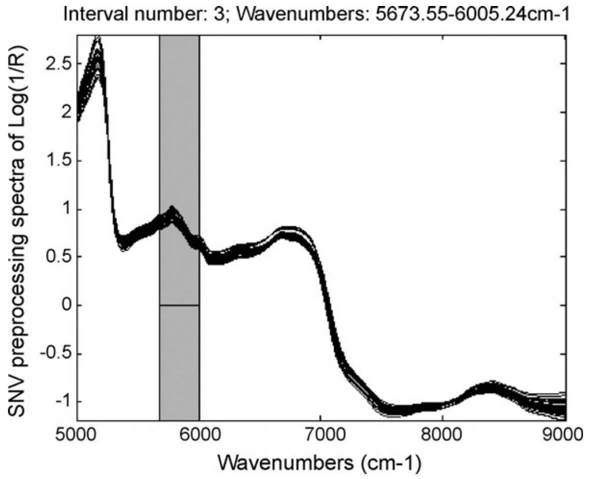

Fig. 3. Optimal spectral region selected by iPLS with wavenumbers 5673.55–6005.24 cm^{-1}.

3.2.1. Results of PLS model

In the application of PLS algorithm, it is generally known that the number of PLS components is a critical parameter in calibrating model. The optimum number of PLS components is determined by the lowest root mean square error cross-validation (RMSECV). The lowest RMSECV is 0.9141 when 6 PLS components are included in calibration model. Therefore the optimal number of PLS components is 6.

Fig. 2 is the scatter plot showing a correlation between reference measured and NIR predicted in calibration set by PLS model. Here, the value of root mean square error of cross-validation (RMSECV) is 0.9141, and correlation coefficient (*R*) is 0.9400 in calibration set. When the performance of PLS model is evaluated by the samples in prediction set, the root mean square error of prediction (RMSEP) is 1.0719 and correlation coefficient (*R*) is 0.9085 in prediction set.

3.2.2. Results of iPLS model

The development of spectral interval selection was first accomplished by iPLS. The interval PLS (iPLS) algorithm [30,31] used here was developed by Norgaard et al. (2000). The principle of this algorithm is to split the spectra into some smaller equidistant regions, next, to develop PLS regression models for each of the sub-intervals. Thereafter, root mean square error of cross-validation (RMSECV) is calculated for every sub-interval. The region with the lowest RMSECV is chosen.

It is serious effect on the performance of iPLS model when split the spectra into different number of intervals, therefore, the number of intervals should been optimized according RMSECV. Results show that the optimal iPLS model is obtained with 12 intervals and 6 PLS components, and the lowest RMSECV is 0.9348 when the optimal interval selected is number 3, corresponding to wavenumbers in the range 5673.55–6005.24 cm^{-1}, which is shown in Fig. 3.

Fig. 4 is the scatter plot showing a correlation between reference measured and NIR predicted in calibration set by iPLS model. Here, the value of root mean square error of cross-validation (RMSECV) is 0.9348, and correlation coefficient (*R*) is 0.9374 in calibration set. When the performance of iPLS model is evaluated by the samples in prediction set, the root mean square error of prediction (RMSEP) is 1.3307 and correlation coefficient (*R*) is 0.8550 in prediction set.

3.2.3. Results of siPLS model

Synergy interval PLS (siPLS) algorithm used here was also developed by Norgaard et al. (2000) [30]. The basic principle of this algorithm is same as iPLS. First, it is to split the data set into a number of intervals (variable-wise), next, to develop PLS regression models for all possible combinations of two, three or four intervals. Thereafter, RMSECV is calculated for every combination of intervals. The combination of intervals with the lowest RMSECV is chosen.

The number of intervals was also optimized according RMSECV in siPLS model calibration. Table 2 shows the results of siPLS model calibration when split the spectra into different number of intervals. The optimal siPLS model was obtained with 23 intervals and 5 PLS components, because the lowest RMSECV is 0.7514, which is prominent with the bold in

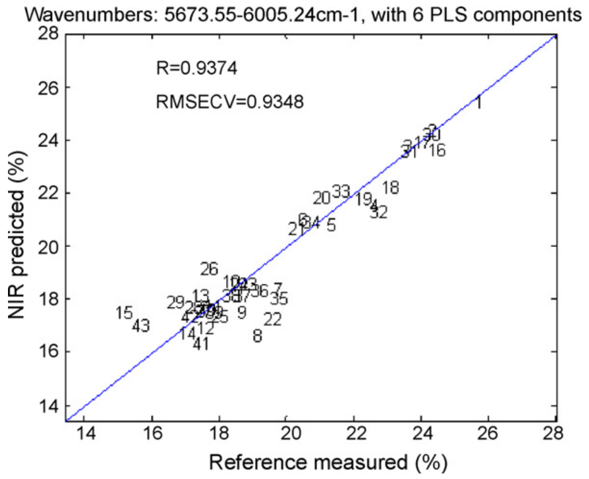

Fig. 4. Reference measured versus NIR predicted by iPLS in calibration set.

Table 2
Results of siPLS calibration model selected different spectral regions

Number of intervals	PLS components	Selected intervals	RMSECV
10	5	[8 10]	0.8415
11	6	[6 7 8 10]	0.8631
12	5	[9 12]	0.8404
13	5	[10 13]	0.8275
14	5	[11 14]	0.8343
15	9	[3 4]	0.8124
16	9	[3 4]	0.7957
17	5	[10 13 17]	0.8225
18	5	[11 14 18]	0.8131
19	5	[14 18 19]	0.8139
20	5	[12 14 16 20]	0.8224
21	5	[15 16 20]	0.8199
22	5	[16 17 21]	0.7965
23	**5**	**[17 18 22]**	**0.7514**
24	9	[4 5 6]	0.7620
25	6	[9 10 16 22]	0.8172

Table 2. The optimal combinations of intervals selected are number 17, 18 and 22. It is corresponding to 7791.01–7960.71, 7964.57–8134.27 and 8658.82–8828.52 cm^{-1} in the spectral regions, which is shown in Fig. 5.

Fig. 6 is the scatter plot showing a correlation between reference measured and NIR predicted in calibration set by iPLS model. Here, the value of root mean square error of cross-validation (RMSECV) is 0.7514, and correlation coefficient (R) is 0.9597 in calibration set. When the performance of siPLS model was evaluated by the samples in prediction set, the root mean square error of prediction (RMSEP) is 0.7372 and correlation coefficient (R) is 0.9583 in prediction set.

3.2.4. Discussion of the results

Comparing three results from PLS, iPLS and siPLS models, siPLS is the best, next to PLS, while, the performance of iPLS is the worst. Such phenomena are explained by the following reasons. (1) PLS is performed on full spectral region (5002.44–9002.08 cm^{-1}) to calibrate global model, so some noisy spectral information inevitably weaken the performance of model. (2) iPLS actually gives an overview spectral data to select the interesting spectral region and remove some noisy regions, but only one interval was selected to calibrate PLS model, so that useful spectral information was removed, therefore, the performance of model inevitably decline. (3) In contrast with iPLS, siPLS shows its incomparable superiority. siPLS not only possesses same advantages as iPLS, but also overcome the disadvantages of iPLS, because siPLS combines with two, three or four intervals to calibrate PLS model, so as not to lose much useful information in calibrating model. In conclusion, the siPLS is superior to iPLS and full spectrum PLS.

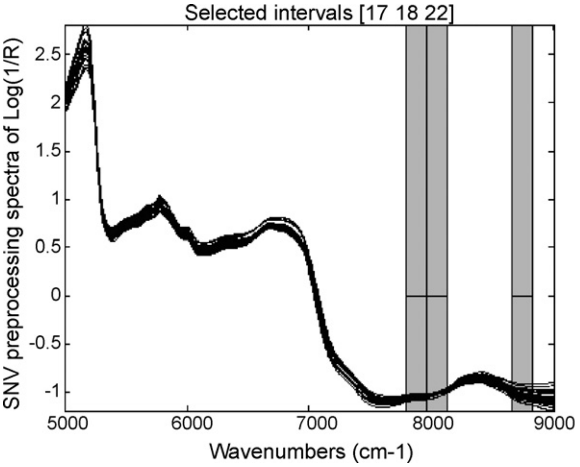

Fig. 5. Optimal spectral region selected by siPLS with wavenumbers 7791.01–7960.71, 7964.57–8134.27 and 8658.82–8828.52 cm^{-1}.

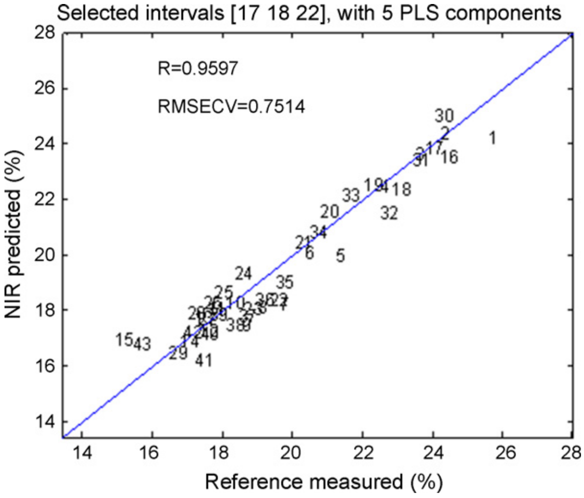

Fig. 6. Reference measured versus NIR predicted by siPLS in calibration set.

4. Conclusion

The overall results sufficiently demonstrate that total polyphenols content in green tealeaves can be determined by NIR spectroscopy coupled an appropriate multivariate calibration method. Compared with PLS, iPLS and siPLS algorithms, the performance of siPLS model is the best. The optimal calibration model was achieved with $R = 0.9583$ and RMSEP = 0.7327 in prediction set. This study demonstrated that NIR spectroscopy with siPLS algorithm could be applied to determine the content total polyphenols in green tea, and siPLS revealed its superiority in contrast with other multivariate calibration methods. It can be concluded that many valid components in tea can be analyzed fast and simultaneously by NIR spectroscopy coupled with siPLS algorithm, and this real-time, at-site measurement will significantly improve the efficiency of quality control and assurance.

Acknowledgements

This work has been financially supported by the Key Natural Science Foundation of Jiangsu Province (Grant No. BK2006707-1) and the National High Technology Research and Development Program of China (863 Project, No. 2006AA10Z263). We are grateful to the website http://www.models.kvl.dk/ where we can download software for iPLS and siPLS free of charge. We also wish to thankful many of their academic colleagues for many stimulating discussion in this field.

References

[1] C.S. Yang, P. Maliakal, X. Meng, Annu. Rev. Pharmacol. Toxicol. 42 (2002) 25–54.

[2] K. Nakachi, S. Matsuyama, S. Miyake, M. Suganuma, K. Imai, Biofactors 13 (2000) 49–54.

[3] V.W. Setiawan, Z.F. Zhang, G.P. Yu, Q.Y. Lu, Y.L. Li, M.L. Lu, M.R. Wang, C.H. Guo, S.Z. Yu, R.C. Kurtz, C.C. Hsieh, Int. J. Cancer 92 (2001) 600–604.

[4] K. Shibata, M. Moriyama, T. Fukushima, A. Kaetsu, M. Miyazaki, H. Une, J. Epidemiol. 10 (2000) 310–316.

[5] L. Jian, L.P. Xie, A.H. Lee, C.W. Binns, Int. J. Cancer 108 (2004) 130–135.

[6] H. Fujiki, M. Suganuma, S. Okabe, E. Sueoka, N. Sueoka, N. Fujimoto, Y. Goto, S. Matsuyama, K. Imai, K. Nakachi, Mutat. Res. 480–481 (2001) 299–304.

[7] M. Inoue, K. Tajima, M. Mizutani, H. Iwata, T. Iwase, S. Miura, K. Hirose, N. Hamajima, S. Tominaga, Cancer Lett. 167 (2001) 175–182.

[8] D. Zhang, S. Kuhr, U.H. Engelhardt, Z. Lebensm. Unters. Forsch. 195 (1992) 108–111.

[9] ISO (International Standard Organization). Determination of Individual Catechins and Total Polyphenols in Tea, ISO TC 34/SC 8 N 444, 1994.

[10] J.M. Esteban-Diez, C. Gonzalez-saiz, Pizarro, Anal. Chim. Acta 514 (2004) 57–67.

[11] K. Kachrimanis, V. Karamyan, S. Malamataris, Int. J. Pharm. 250 (2003) 13–23.

[12] I. Esteban-Diez, J.M. Gonzalez-saiz, C. Pizarro, Anal. Chim. Acta 525 (2004) 171–182.

[13] Y.A. Woo, H.J. Kim, K.R. Ze, H. Chung, J. Pharmaceut. Biomed. Anal. 36 (2005) 955–959.

[14] Y. Dou, Y. Sun, Y.Q. Ren, P. Ju, Y.L. Ren, J. Pharmaceut. Biomed. Anal. 37 (2005) 543–549.

[15] C.W. Huck, W. Guggenbichler, G.K. Bonn, J. Pharmaceut. Biomed. Anal. 538 (2005) 195–203.

[16] R. De Maesschalck, T.V. Kerkhof, J. Pharmaceut. Biomed. Anal. 37 (2005) 109–114.

[17] C.J. Clark, V.A. McGlone, R.B. Jordan, Postharvest Biol. Technol. 28 (2005) 65–71.

[18] Y.A. Woo, H.R. Lim, H.J. Kim, H. Chung, J. Pharmaceut. Biomed. Anal. 33 (2003) 1049–1057.

[19] V.A. McGlone, R.B. Jordan, R. Seelye, P.J. Martinsen, Postharvest Biol. Technol. 26 (2002) 191–198.

[20] M.N. Hall, A. Robertson, C.N.G. Scotter, Food Chem. 27 (1988) 61–75.

[21] H. Schulz, U.H. Engelhardt, A. Wengent, H.H. Drews, S. Lapczynski, J. Agric. Food Chem. 475 (1999) 5064–5067.

[22] J. Luypaert, M.H. Zhang, D.L. Massart, Anal. Chim. Acta 487 (2003) 303–312.

[23] M.H. Zhang, J. Luypaert, Q.S. Xu, D.L. Massart, Talanta 62 (2004) 25–35.

[24] Q.S. Chen, J.W. Zhao, H.D. Zhang, X.Y. Wang, Anal. Chim. Acta 572 (2006) 77–84.

[25] Q.S. Chen, J.W. Zhao, X.Y. Huang, H.D. Zhang, M.H. Liu, Microchem. J. 83 (2006) 42–47.

[26] Y.F. Luo, Z.F. Guo, Z.Y. Zhu, C.P. Wang, H.Y. Jiang, B.Y. Han, Spectrosc. Spectral Anal. 25 (2005) 1230–1233.

[27] Y.G. Sun, M. Lin, J. Lv, L.H. Xu, Chin. J. Spectrosc. Lab. 21 (2004) 940–943.

[28] C. Abrahamsson, J. Johansson, A. Sparén, F. Lindgren, Chemometr. Intell. Lab. Syst. 69 (2003) 3–12.

[29] O. Kleynen, V. Leemans, M.F. Destain, Postharvest Biol. Technol. 30 (2003) 221–232.

[30] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413–419.

[31] R. Leardi, L. Nørgaard, J. Chemometr. 18 (2004) 486–497.

[32] X.L. Chu, H.F. Yuan, W.Z. Lu, Progr. Chem. 16 (2004) 528–542.