A spectral similarity measure using Bayesian statistics

Article *in* Analytica chimica acta · April 2009

DOI: 10.1016/j.aca.2009.01.024 · Source: PubMed

3 authors, including:

Feng Gan
Sun Yat-Sen University
26 PUBLICATIONS **267** CITATIONS
SEE PROFILE

Philip K Hopke
Clarkson University
973 PUBLICATIONS **25,261** CITATIONS
SEE PROFILE

Analytica Chimica Acta 635 (2009) 157–161

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

A spectral similarity measure using Bayesian statistics

Feng Gan[a,*], Philip K. Hopke[b], Jiajun Wang[c]

[a] School of Chemistry and Chemical Engineering, Sun Yat-sen University, Guangzhou 510275, China
[b] Center of Air Resources Engineering and Science, Clarkson University, Potsdam, NY 13699, USA
[c] Honghe Cigarette General Factory, Yunnan 652300, China

ARTICLE INFO

Article history:
Received 18 December 2008
Accepted 8 January 2009
Available online 19 January 2009

Keywords:
Spectral similarity measure
Bayesian statistics
Near-infrared spectrum
Tobacco

ABSTRACT

A spectral similarity measure was developed that can differentiate subtle differences between two spectra. The spectra are digitalized into a vector. The difference between the two spectra is defined by a difference vector, which is one spectrum minus the other. The spectral similarity measure is transformed into a hypothesis test of the similarities and differences between the two spectra. The scalar mean of the difference vector is used as the statistical variable for the hypothesis test. A threshold for the hypothesis that the spectra are different was proposed. The Bayesian prior odds ratio was estimated from multiple spectra of the same sample. The posterior odds ratio was used to quantity the spectral similarity measure of the two spectra. Diffuse reflectance near-infrared spectra of tobacco samples of two formulations were used to demonstrate this method. The results show that this new method can detect subtle differences between the spectra.

1. Introduction

Many types of spectra have been widely used in modern analytical chemistry. Spectral similarity measures have been applied to solve a variety of problems. Library searches [1,2] of mass spectra are an example of the application of spectral similarity measures. A target mass spectrum of a substance is compared with the standard spectra in a mass spectral database (for example, the SpecInfo [3]). A similarity index (SI) is calculated to produce a list of candidate compounds and the one with the largest SI value is the most possible candidate. However, the experience of experts is generally needed to make final conclusions from the library search results. So far, spectral similarity measures have been extended to other kinds of spectral data such as fluorescence spectra, infrared spectra and chromatography profiles [4,5]. The fields of application of spectral similarity measures have been extended to source determination [6–11] of a substance and chemical process control [12]. Such spectral comparisons are routinely made in everyday work. Thus, there is a need to develop widely applicable methods for estimating indicators of spectral similarity.

To date, many spectral similarity measures have been proposed. The earliest method of determining spectral similarity measure was to make visual comparisons. Two spectra are plotted together, and if they overlap well in a visual comparison, they are assumed to be the same (or similar). Otherwise they are dissimilar. Although eyes have the ability to differentiate even rather subtle differences between two spectra, this ability varies among individuals so it cannot be an objective scientific approach. Although the method is still used today, calculation of some type of SI [13] for two spectra has become the primary approach in chemistry. Efforts to develop new SI (or dissimilarity index) have been recently published [14,15]. In a sense, the currently methods to calculate SIs are applicable to a variety of problems comparing the similarities and differences between two spectra. However, the definition of a quantitative threshold for the SI that clearly delineates when spectral differences can be ascertained with a known level of certainty is still a problem. There is no theoretical basis to establish an objective threshold. In previous work [10], a method for a spectral similarity measure based on Bayesian statistics [16] was presented. The difference between our method and others is that we transformed spectral similarity measure into a hypothesis test. We found that hypothesis testing provides an objective solution to the threshold problem.

There is another challenge for spectral similarity measures beyond the threshold problem. The challenge comes from subtle differences between two spectra. These subtle differences mean that the two spectra may overlap quite well in general except for some local regions. Even in these local regions, the differences may not be significant to the naked eye. Such subtle differences will cause problems for many methods. For example, many methods such as correlation coefficient or the vector cosines [4] will give a SI value of approximately 1.0 (or 100%) for these kinds of spectral pairs. Our previous method also failed when these spectral types are encountered [10].

* Corresponding author. Tel.: +86 20 31725310; fax: +86 20 84112245.
E-mail address: cesgf@mail.sysu.edu.cn (F. Gan).

0003-2670/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2009.01.024

In this paper, a new spectral similarity measure has been developed based on our previous work. Theoretically, our previous method is a widely applicable method because no assumptions were made with respect to the type of spectral data. Further improvements to the method are needed to enable it to cope with the challenge from subtle differences between the two spectra. The problems to be solved are defining the boundary for the alternative hypothesis and the prior odds ratio. We expected that subtle differences between two spectra can be differentiated if these problems are resolved. Our improved method was applied to spectra of tobacco samples of two formulations and it did successfully differentiate the subtle differences among the spectra.

2. Theory

A spectrum is a collection of responses of an instrument to a sample at different analytical channels. For the convenience of mathematical analysis, the spectrum is written in vector form as $\mathbf{r}(r_1, r_2, \ldots, r_n)$, where $r_i(m)(i = 1, 2, \ldots, n)$ is the response at ith analytical channel for the m th sample, n is the number of analytical channels. As the existent of measurement errors, a spectrum will has following expression:

$$\mathbf{r}(m) = \mathbf{r}_0(m) + \mathbf{e}(m) + \mathbf{s}(m) \tag{1}$$

where $\mathbf{r}_0(m)$ is the true spectrum of the sample m; $\mathbf{e}(m)$ is the vector of errors from random noise in the measurement of sample m; $\mathbf{s}(m)$ is the vector of the systematic errors in analytical channels.

In an ideal situation that there are no any measurement errors, two spectra (j and k) from same sample will be the same. It means

$$\mathbf{r}_0(j) - \mathbf{r}_0(k) = \mathbf{0}. \tag{2}$$

In practical analytical measurement, errors exist. So, one obtains

$$\mathbf{r}(j) - \mathbf{r}(k) \neq \mathbf{0}. \tag{3}$$

Under reproducible conditions, the systematic errors are usually assumed to be the same. Thus, the controlling errors are the random noise. The differences between two spectra will be

$$\Delta\mathbf{r} = \mathbf{r}(1) - \mathbf{r}(2) = \mathbf{e}(1) - \mathbf{e}(2) \tag{4}$$

Since $\mathbf{e}(1)$ usually does not equal $\mathbf{e}(2)$, the measurement errors from the random noise are the main problem in practical measurements of spectral similarity. Since the random noise cannot be eliminated, we have to accept some given level of random noise. Eq. (4) shows that the so-called similarity measure can be based on statistical analysis of the differences between the two spectra. Based on this consideration, we establish following hypothesis test for determining the similarity of two spectra.

$$\begin{cases} \mathrm{H}_0: & \Delta\mathbf{r} = \mathbf{r}(1) - \mathbf{r}(2) = 0 \\ \mathrm{H}_1: & \Delta\mathbf{r} = \mathbf{r}(1) - \mathbf{r}(2) \neq 0 \end{cases} \tag{5}$$

where H_0 is the null hypothesis and H_1 the alternative hypothesis. When H_0 holds, we can conclude that the two spectra are the statistically indistinguishable. Therefore, the two spectra must also be the same chemically.

To implement the hypothesis shown in Eq. (5), we pretreat the spectra. Since the random errors have a mean of zero, the elements of $\mathbf{e}$ have a distribution of $e_i \sim N(0, \sigma_i^2)$, where σ_i is the standard deviation in analytical channel i. Thus, in the absence of bias, the intensity at analytical channel i has the distribution of $N(r_{o,i}, \sigma_i^2)$, where $r_{o,i}$ is the true value of the intensity in analytical channel i. $r_{o,i}$ and σ_i can be estimated by repeated measurement of the same sample under reproducible conditions. The distribution above can be transformed into $N(r'_{o,i}, 1)$ by unit-variance scaling $r'_{o,i} = r_{o,i}/\sigma_i$.

After this transformation, the scaled elements of $\Delta\mathbf{r}'_0(j, k)$, each has a distribution of $N(0, 2)$. Hereafter, the vectors are scaled in this way. We cannot directly estimate $\Delta\mathbf{r}'_0(j, k)$ since we can only collect data with its embedded noise. However, the ordered difference vector, $\Delta\mathbf{r}'(j, k)$ can be calculated for the pair of the observed spectra j and k. The distributions of $\Delta\mathbf{r}'(j, k)$ will then be approximately $N(0, 2)$. It means that in the absence of errors, the summation of the elements of the $\Delta\mathbf{r}'(j, k)$ will be zero.

$$T = \frac{1}{n}(\Delta r'_1 + \Delta r'_2 + \cdots + \Delta r'_n) = \frac{1}{n}\left[\sum_{j=1}^{n} r(j) - \sum_{k=1}^{n} r(k)\right] \tag{6}$$

If the two spectra are identical, the expectation value of T is zero ($E(T) = 0$) and the variance of T, $V(T)$ is $2/n$. We then consider the following hypothesis instead of that given in Eq. (5).

$$\begin{cases} \mathrm{H}_0: & E(T) = 0 \\ \mathrm{H}_1: & E(T) \neq 0 \end{cases} \tag{7}$$

When the T value is given, the posterior probabilities of H_0 and H_1 are as follows:

$$p(\mathrm{H}_0|T) = \frac{p(T|\mathrm{H}_0)p(\mathrm{H}_0)}{p(T|\mathrm{H}_0)p(\mathrm{H}_0) + p(T|\mathrm{H}_1)p(\mathrm{H}_1)} \tag{8}$$

$$p(\mathrm{H}_1|T) = \frac{p(T|\mathrm{H}_1)p(\mathrm{H}_1)}{p(T|\mathrm{H}_0)p(\mathrm{H}_0) + p(T|\mathrm{H}_1)p(\mathrm{H}_1)} \tag{9}$$

where $p(\mathrm{H}_0|T)$ and $p(\mathrm{H}_1|T)$ are the posterior probabilities for the null hypothesis and the alternative hypothesis, respectively. Both $p(\mathrm{H}_0)$ and $p(\mathrm{H}_1)$ are the prior probabilities. The posterior odds ratio will be

$$\frac{p(\mathrm{H}_0|T)}{p(\mathrm{H}_1|T)} = \left[\frac{p(\mathrm{H}_0)}{p(\mathrm{H}_1)}\right]\left[\frac{p(T|\mathrm{H}_0)}{p(T|\mathrm{H}_1)}\right] \tag{10}$$

The second term in Eq. (10) is the prior odds ratio. When there is no information about the prior probabilities, a reasonable assumption is $p(\mathrm{H}_0) = p(\mathrm{H}_1) = 0.5$. However, in practical applications, we may have some information on the measured spectra so a value for this ratio can be provided as discussed later in this section. For the convenience, we define:

$$\alpha = \frac{p(\mathrm{H}_0)}{p(\mathrm{H}_1)} \tag{11}$$

However, the alternative hypothesis $\mathrm{H}_1: T \neq 0$ is too broad to calculate the probability in practical applications. The previous setting of $T = 1$ for H_1 may not be sufficiently stringent [10]. This lack of stringency might explain the failure of the method in differentiating subtle differences between two spectra. Since the variance of T equals $2/n$, the standard deviation is $\sigma = \sqrt{2/n}$. Since 3σ represents a 99.7% confidence bound, a T of $3\sqrt{2/n}$ for H_1 should provide an appropriate criterion value.

Since the distributions of $p(T|\mathrm{H}_0)$ and $p(T|\mathrm{H}_1)$ are $N(0, 2/n)$ and $N(3\sqrt{2/n}, 2/n)$, respectively, the likelihood ratio (LR, the third item in Eq. (10)) has the form of

$$LR = \frac{p(T|\mathrm{H}_0)}{p(T|\mathrm{H}_1)} = \frac{(n/4\pi)^{1/2}e^{-(n/4)T^2}}{(n/4\pi)^{1/2}e^{-(n/4)(T-3\sqrt{2/n})^2}} = e^{-(n/4)(6\sqrt{2/n}T-(18/n))} \tag{12}$$

So, the posterior odds ratio (POR) shown in Eq. (10) will have the form:

$$POR = \frac{p(\mathrm{H}_0|T)}{p(\mathrm{H}_1|T)} = \alpha e^{-(n/4)(6\sqrt{2/n}T-(18/n))} \tag{13}$$

If $POR > 1$, we can accept H_0; otherwise, we will reject H_0 and accept H_1. The spectral similarity measure is thereby transformed into a hypothesis test.

In practical analytical measurement, the parameter T is calculated from the μ'_1 values that can be positive or negative. Thus, T

Table 1
Information of the NIR spectra.

Date	No.	Numbers	Percent of pass
November 2005	1–10	10	100
December 2005	11–48	38	71.1
January 2006	49–90	42	95.2
February 2006	91–93	3	100.0
March 2006	94–120	27	85.2
April 2006	121–150	30	0.0
May 2006	151–154	4	0.0
July 2006	155–170	16	0.0

will be positive or negative. If T is positive, we can implement right side testing and POR is calculated by Eq. (13); if T is negative, we can implement left side testing, and the T value for H_1 will then be $-3\sqrt{2/n}$. However, a simple calculation based on Eq. (12) shows that we can still use Eq. (13) to calculate the POR if we use the absolute value of T. The prior odds ratio α is the parameter that is uncertain. If there is no prior information, we have to assume that $\alpha = 1$. However, when we have additional information, α can be set to values greater than 1. Suppose that two spectra come from samples of the same specific substance, and thus, we believe that these two spectra should be identical except for the noise. Thus, if the noise is such that the calculated $POR < 1$, we know that we need to set the α value sufficiently large to make $POR > 1$ to correspond to our knowledge of the nature of the truly similar samples being measured.

3. Experimental section

The spectral data used in this paper are historical results from the Honghe cigarette general factory of Honghe tobacco group, Yunnan Province, China. Tobacco samples were collected at the third class formula production line (a batch process) in the factory. Under normal processing conditions, one kilogram of cut tobacco is gathered at the production site. The gathered cut tobacco was uniformly mixed and 30 g was used to make a powder with a 0.25-mm grain size. The tobacco powders were kept in a sealed valve bags for future measurements.

A PerkinElmer Spectrum One NTS spectrometer equipped with an integrating sphere for diffuse reflectance analyses, a tungsten halogen lamp and an InGaAs detector were used to measure the NIR spectra of the tobacco powders. The spectral resolution was 8 cm^{-1}. The tobacco powders were placed in a quartz base sample dish and the dish is positioned on integrating sphere. Each NIR sample spectrum was the average of 64 scans.

Pretreatment of the data was performed to eliminate the baseline drift in the diffuse reflectance near- infrared spectroscopy and the differential scattering by particles of different diameters. A differential operation based on the Savitzky–Golay method [17] was used to remove the baseline drift. A multiple scattering correction [18] was used to eliminate the influence of scattering. All of these pretreatments were performed using Prirouette v3.11 (Infometrix, Inc.) with default settings. All data calculations of our method were performed using Matlab 7.0.

4. Results and discussion

In this study, 170 NIR spectra of tobacco samples of two formulations were examined. The spectra were from independent samples from the production process. The details of these spectra are shown in Table 1 (first three columns). On November 2005, a specific formulation of tobacco was placed into production and this formulation was kept unchanged for four successive months. From April 2006, a new formulation replaced the previous one. In the production process, fluctuations in the working conditions were often encountered that could have an adverse influence on the final quality of cigarette. It was anticipated that our method could detect the fluctuations by measuring the changes in the spectra of the tobacco samples from these production processes.

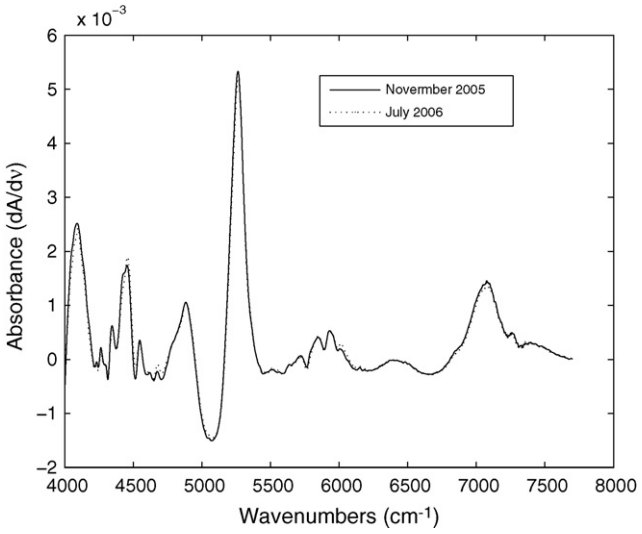

Fig. 1. First derivative spectra of the NIR spectra of the tobacco samples of two formulas.

Fig. 1 shows two first derivative spectra of tobacco samples of the two formulations. The influence of the baseline drift and the scattering were removed so the spectra appear to overlap well. However, subtle differences between them can still be clearly seen in some local areas. These differences can be attributed to the fluctuation in the production process and the differences in the formulations. The calculated SI of the spectra is greater than 0.99 if we use methods such as the correlation coefficient or the vector cosines [4]. Our previous method [10] also gives very large posterior odds ratios for these spectra as being similar. These results support the conclusion that there are no differences between the spectra.

Fig. 2 shows a normal possibility plot of the data at 6398 cm^{-1}. This result shows that an approximately normal distribution was observed in these data. Similar results can be found at various other wave numbers. These results confirm our previous assumption that a Gaussian distribution can be used to describe the distribution of the spectral responses. If the distribution does not follow a Gaussian distribution, a modification of Eq. (12) would be needed depending on the actual distribution.

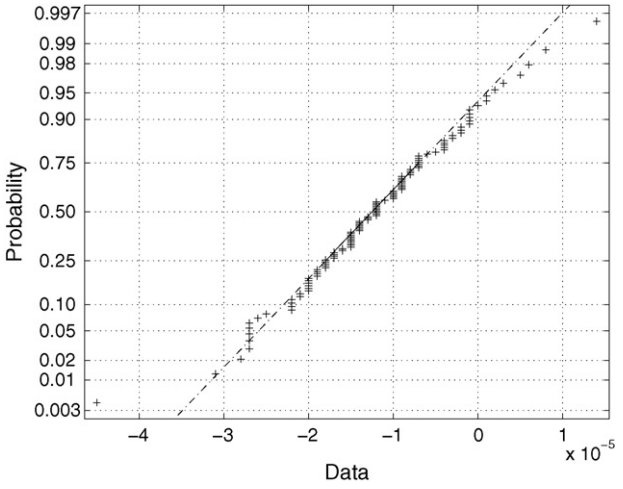

Fig. 2. Normal possibility plot of the data at 6398 cm^{-1}.

Table 2
Likelyhood ratios of the spectra of November 2005.

Spectrum No.	LR
1	0.0374
2	24.7
3	0.0348
4	3.34×10^{-4}
5	1.33×10^{-7}
6	55.9
7	0.163
8	1.56×10^{-4}
9	3.14
10	8.64×10^{-7}

A standard spectrum is needed before we can implement our method. However, it is not possible to establish a standard spectrum for these samples because they are historical data. The standard spectrum should be developed by repeated measurement of a single same sample under reproducible conditions. In this paper, we use another approach to generate a pseudo-standard spectrum. Table 2 shows the calculated likelihood ratios of the spectra of November 2005. In the calculation, the mean spectrum of the ten spectra was taken as the pseudo-standard spectrum. The sample standard deviations at each wave number of the ten spectra were calculated and used to scale the 10 spectra and the mean spectrum. The likelihood ratio between each of the spectra and the mean spectrum was calculated using Eq. (12). Since there is no prior information about the spectra, a reasonable assumption is $\alpha = 1$. The likelihood ratios show similarities and differences among the spectra with the mean spectrum. It can be seen that the likelihood ratios vary with the sigma values but without any coherent pattern among them. In this paper, we assume that a three sigma interval defines a 99% confidence bound.

Table 2 poses the question of how to treat the spectra whose likelihood ratios are smaller than 1.0. If we treat the spectra in isolation, the differences are significant. However, as the spectra are from the tobacco samples of same formulation, there is reason to believe that the spectra should be the same. The differences among the spectra have been attributed to the fluctuations of the processing conditions. It seems that it is a reasonable compromise to accept that differences arise because we cannot wholly eliminate the fluctuations. On the other hand, as the spectra are from tobacco samples of same formula in the first month. Therefore, these fluctuations in the measurements provides an objective standard to set the prior odds ration so we can evaluate successive production processes. From Eq. (13), one finds that a prior odds ratio $\alpha = 1/1.33 \times 10^{-7}$ ensures that none of the *POR* values for any of the spectra in Table 2 will be less than 1.0. This strategy offers an easy way to determine the prior odds ratio when coping with the practical problem of using Bayesian statistics. This kind of treatment is based on information on the nature of the samples and the measurements rather than from a mathematical analysis.

Although there is no strict theoretical basis, this kind of strategy makes sense in real production process because we can establish parameters to detect the fluctuation in the production process. In this paper, the mean spectrum of the ten spectra is taken as pseudo standard spectrum. The sample standard deviations at each wave number are calculated from the spectra; and the prior odds ratio is $1/1.33 \times 10^{-7}$. These parameters are used in successive calculation of *POR* values of the other spectra shown in Table 1.

Fig. 3 shows the stem plots of the logarithms values of the *POR* for the spectra collected between November 2005 and July 2006. When the logarithm of *POR* is less than 0.0, it provides an indication that the production conditions might have changed. Fig. 3 shows that most of the production processes were operating under normal conditions. The fourth column in Table 1 gives the percent of the

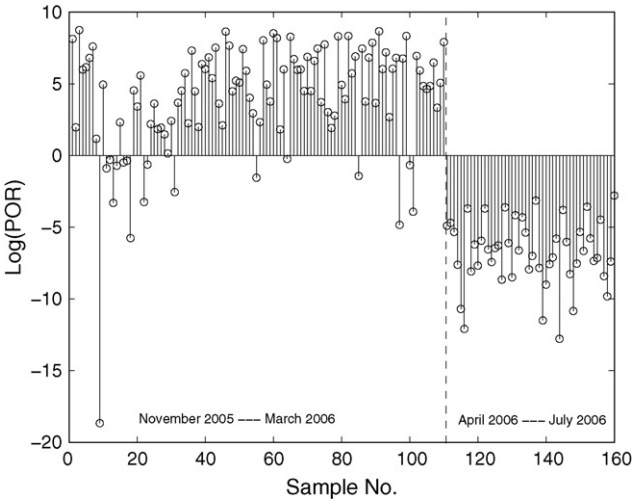

Fig. 3. Logarithm values of the posterior odds ratios for the spectra from November 2005 to July 2006.

spectra that passed the hypothesis test. One can also see from Fig. 3 that the logarithm of the *POR* are all negative from April 2006. This result suggests that the differences in tobacco formulations were the main source of the difference seen in Fig. 2.

5. Conclusion

A further study of a spectral similarity measure using Bayesian statistics was presented. The study shows that the transforming spectral similarity measure into Bayesian hypothesis testing is successful. A strategy for determining the prior odds ratio problem in Bayesian statistics is also proposed that is based on the prior spectral information. This strategy makes a fully use of the information of the random errors in the measurements. So it is helpful to make objective similarity measure of the spectra.

Acknowledgments

We thank Professor Qingsong Xu of Institute of Probability and Statistics, Central South University (Changsha) for his suggestion in the revising of this paper. We also thank Dr. Weishi Zheng and Professor Jianhuang Lai of School of Information Science and Technology of Sun Yat-sen University (Guangzhou). They gave instructive talks on Bayesian statistics. This research work was financially supported by National Natural Science Foundation of China (Grants No. 20475067).

References

[1] S.R. Heller, Computer-Supported Spectral Database, John Wiley & Sons, New York, 1986.
[2] P. Willett, J.M. Barnard, G.M. Downs, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.
[3] http://www.cas.org/ONLINE/DBSS/specinfoss.html.
[4] M.J. Chen, Y.Y. Cheng, R.C. Lin, Chin. Trad. Pat. Med. 24 (2002) 905–908.
[5] B.Y. Li, Y. Hu, Y.Z. Liang, P.S. Xie, Y.P. Du, Anal. Chim. Acta 514 (2004) 69–77.
[6] D.L. Duewer, B.R. Kowalski, T.F. Schatzki, Anal. Chem. 47 (1975) 1573–1583.
[7] W.J. Welsh, W. Lin, S.H. Tersigni, E. Collantes, R. Duta, M.S. Carey, W.L. Zielinski, J. Brower, J.A. Spencer, T.P. Layloff, Anal. Chem. 68 (1996) 3473–3482.
[8] J. Li, S. Fuller, J. Cattle, C. PangWay, D.B. Hibbert, Anal. Chim. Acta 514 (2004) 51–56.
[9] J. Li, D.B. Hibbert, S. Fuller, J. Cattle, C.P. Way, Anal. Chem. 77 (2005) 639–644.
[10] F. Gan, R.Y. Ye, J. Chromatogr. A 1104 (2006) 100–105.
[11] A. Nevin, L. Osticioli, D. Anglos, A. Burnstock, S. Cather, E. Castellucci, Anal. Chem. 79 (2007) 6143–6151.
[12] R.D. Maesschalck, F.C. Sanchez, D.L. Massart, P. Doherty, P. Hailey, Appl. Spectrosc. 52 (1998) 725–731.

[13] J. Gasteiger, T. Engel, Chemoinformatics, WILEY-VCH GmbH & Co. KGaA, 2003.
[14] J.F. Li, D.B. Hibbert, S. Fuller, G. Vaughn, Chemom. Intell. Lab. Syst. 82 (2005) 50–58.
[15] Y. Xu, F. Gong, S.J. Dixon, R.G. Brereton, Anal. Chem. 79 (2007) 5633–5641.
[16] S.J. Press, Bayesian Statistics: Principles, Models and Applications, Wiley & Sons, Inc., 2002.
[17] A. Savitzky, M.J.E. Golay, Anal. Chem. 36 (1964) 1627–1639.
[18] P. Geladi, D. MacDougall, H. Martens, Appl. Spectrosc. 39 (1985) 491–500.