

Modified secured principal component regression for detection of unexpected chromatographic features in herbal fingerprints

Bo-Yan Li,^{ab} Yun Hu,^b Yi-Zeng Liang,^{*a} Pei-Shan Xie^a and Yukihiro Ozaki^{*b}

Received 21st September 2005, Accepted 4th January 2006

First published as an Advance Article on the web 23rd January 2006

DOI: 10.1039/b513365c

Secured principal component regression is modified for the qualitative analysis of chromatographic fingerprint data sets of herbal samples with residual concentrations. After chromatographic shift-correction and autoscaling are performed on the data, this modified secured principal component regression (msPCR) can detect unexpected chromatographic features in various herbal fingerprints. The successful application of msPCR to two real herbal medicines of *Erigeron breviscapus* from different geographical origins and *Ginkgo biloba* from various sources or vendors demonstrates that the proposed method can detect reasonably unexpected features differing from the regulars or not being modeled. From a chemical point of view, the causes have also been explained to corroborate the results. Moreover, it presents a viable approach for the qualitative evaluation of diverse herbal objects with a regular class of chromatographic fingerprints.

Introduction

Despite its existence and continuing use over many centuries, and its popularity and extensive use during the last decade, traditional herbal medicine has not been officially recognized in most countries due not only to the lack of research data to support its safety and efficacy but also to a lack of adequate or accepted research methodology for its evaluation.¹ Because a herbal medicine or formula often consists of many complex phytochemicals, sometimes it becomes imprecise or difficult for conventional approaches to fulfil its practical quality assessment.^{2–4} There are two major trends for quality control of traditional medicine. One quality control mode is originally from chemically synthesized pharmaceuticals, namely, selecting a known active constituent or a marker compound from the herbal drug as the qualitative and quantitative target to assess its authenticity and inherent quality. Several decades have passed with such a quality control mode being applied to herbal medicines with no significant changes except for renewing and advancing the analytical technologies. However, for natural products derived from herbal medicines with the inherent uncertainty feature of their secondary metabolic substances, the drawback of such a quality control mode has surfaced more and more obviously. Therefore, another practical approach, generated from a comprehensive mode, is necessary to control the quality of herbal products, and the complex formulation of herbal medicine in particular. As demanded, fingerprinting procedures have emerged which appear efficient for serial analyses in quality control and

stability tests of herbal medicine.⁵ Chromatography fingerprinting emphasizes an integral formulation of pharmacologically active and phytopharmaceutically characteristic components of samples with similar or different attributions.^{6–11} The quality consistency and stability of herbal extracts or products can be assessed by their integral fingerprint patterns in quantified operation procedures. Therefore, it was formally introduced by WHO in Munich in 1991,¹² and subsequently accepted.

Fingerprinting is a logical result of both the development of herbal medicine and the progress of relevant research for quality assessment,^{1,11} and also a demand of both the compilation of pharmaceutical attributes and the comparison of componential distributions of herbal samples.^{5,9,11,13–16} With respect to its uncertainty and inclusion, the multiple-level and integral information of a chromatographic fingerprint is more interesting for quality control than the information from the conventional mode. Moreover, fingerprinting makes the inherent quality of herbal medicine visible through observing and comparing fingerprint patterns.¹¹

There have been only a few methods to evaluate the fingerprint quality of herbal materials or pharmaceutical products, such as correlative chromatography,^{4,16} comparative analysis,^{4,17} wavelet analysis and artificial neural networks (ANN).^{9,18,19} Either taking advantage of spectral correlation and chromatographic features or using chemometric methods and pattern recognition techniques such as K-nearest neighbors (KNN)^{9,17} and soft independent modeling of class analogy (SIMCA),⁹ these techniques distinguish which are of good quality and which not within chromatographic fingerprint sets.

The relationship of the fingerprints could be commonly analyzed through comparison with a certain reference, in terms of similarity or dissimilarity, presented as correlation coefficient or congruence coefficient *etc.*^{4,17} However, it is apparent

^aCollege of Chemistry and Chemical Engineering, Research Center of Modernization of Chinese Herbal Medicines, Central South University, Changsha, 410083, P. R. China. E-mail: yizeng_liang@263.net; Fax: +86 731 8830831; Tel: +86 731 8830831

^bDepartment of Chemistry and Research Center for Near-Infrared Spectroscopy, School of Science and Technology, Kwansei-Gakuin University, 2-1 Gakuen, Sanda, 669-1337, Japan. E-mail: ozaki@kwansei.ac.jp; Fax: +81 79 565 9077; Tel: +81 79 565 8349

that two problems with this comparison exist: how to achieve a reasonable reference and to what extent the objects are similar to such a reference. Popularly, the reference may be derived either from the extracts of a standard herbal medicine or proportioned mixtures of herbal medicines (e.g. EGb761)^{3,5} or from computation with some mathematical method.⁴ The (dis)similarities between herbal objects and the reference often undertake themselves to a qualified threshold.¹¹ Although such comparison attaches importance to the integral relationship of the fingerprints, sometimes masking and swamping effects might occur either explicitly or implicitly. The masking effect means that an unexpected sample could not be identified due to its high similarity to the reference (e.g. the identification of three species of *Coptis chinensis*, *C. teet-Oides* C. Y. Cheng, and *C. deltoidea* C. Y. Cheng et Hsiao from herb *Rhizoma Coptidis*). The swamping effect encompasses wrongly discriminating a desirable sample illegal on account of its low similar value, which is usually influenced by the diversity of chromatographic compositional distribution (e.g. the determination of herb *Houttuynia cordata* Thunb. from different sources).

In the present study, modified secured principal component regression (msPCR) is introduced. It is flexible to detect unexpected chromatographic features, by means of the systematic fit error compensation within a series of interval moving windows. msPCR is based on but different to secured principal component regression (sPCR).^{20,21} The latter was originally developed to detect and characterize the uncalibrated spectral features in measured spectra. It includes a series of steps, such as two-step denoise, linear or nonlinear residual decision, reflecting-line estimate, and so forth, for the approximate compensation of systematic fit errors and the correction of the disturbances. Therefore, when the investigated objects are diverse chromatographic fingerprints in compositional distribution, it is not friendly to use and the results are also not satisfactory. However, the residual concentrations after the systematic fit error compensation *via* msPCR can provide qualitative information about these features with respect to chromatographic peaks and positions, which is valuable for quality evaluation of herbal samples. This method allows one to avoid the above effects as much as possible.

Two real chromatographic fingerprint data sets of *Erigeron breviscapus* and *Ginkgo biloba* herbal samples are analyzed *via* msPCR with residual discrimination, and the results are demonstrated from a chemical point of view, paralleled with a characteristic trial of the phytochemicals of herbal samples.

Modified secured principal component regression algorithm

Throughout the remainder, matrices will be noted in capital boldface letters and column vectors in lower-case italic boldface letters. Transposed objects are indicated by superscript T. Subscripts cal and meas specify calibration and measurement items, respectively. Elements of matrices or vectors are denoted with lower-case non-bold letters with italic subscript characters.

Principal component regression

Principal component analysis (PCA) and principal component regression (PCR)^{22–24} are popular techniques for multivariate data analysis and evaluation of unknown measurements. In this study, with known regular samples PCR first builds an appropriate calibration model for herbal fingerprints and then uses this model to evaluate unknown measurements and detect unexpected fingerprints *via* their residual concentrations in the resulting scores from their projection onto significant principal components (PCs), which can be summarized as follows.

Suppose a calibration matrix \mathbf{X}_{cal} consists of k -column regular fingerprints and m chromatographic absorptions measured at regular time intervals. Then, PCA is conducted by singular value decomposition (SVD) of \mathbf{X}_{cal} .^{25,26}

$$\mathbf{X}_{\text{cal}(m \times k)} = [\mathbf{x}_1 \dots \mathbf{x}_k] = \mathbf{U}_{(m \times k)} \mathbf{S}_{(k \times k)} \mathbf{V}_{(k \times k)}^T = \mathbf{P}_{(m \times k)} \mathbf{T}_{(k \times k)}^T \quad (1)$$

where $\mathbf{P} = [\mathbf{p}_1 \dots \mathbf{p}_k]$ is defined as the score matrix, $\mathbf{T} = [\mathbf{t}_1 \dots \mathbf{t}_k]$ the loading matrix. Due to noise in the actual system, the rank r of \mathbf{X}_{cal} is decided as the number of relevant PCs different from noise, usually $r < k$. Hence, eqn (1) is expressed with two blocks,

$$\mathbf{X}_{\text{cal}(m \times k)} = [\mathbf{X}_{\text{cal}(m \times r)} \mathbf{X}_{\text{noise}(m \times (k-r))}] = [\mathbf{P}_{\text{cal}(m \times r)} \mathbf{P}_{\text{noise}(m \times (k-r))}] \mathbf{T}_{(k \times k)}^T \quad (2)$$

where $\mathbf{P}_{\text{cal}(m \times r)} = [\mathbf{p}_1 \dots \mathbf{p}_r]$ involves information of the r relevant PCs. This will be used for regression calibration. $\mathbf{P}_{\text{noise}(m \times (k-r))} = [\mathbf{p}_{r+1} \dots \mathbf{p}_k]$ encompasses noise. Then, $\mathbf{X}_{\text{meas}(m \times n)} = [\mathbf{x}_{\text{meas}_1} \dots \mathbf{x}_{\text{meas}_j} \dots \mathbf{x}_{\text{meas}_n}]$ consisting of n -column measured fingerprints can be analyzed by multivariate least squares fit to obtain measurement \mathbf{T}_{meas} ,

$$\mathbf{X}_{\text{meas}(m \times n)} = \mathbf{P}_{\text{cal}(m \times r)} \mathbf{T}_{\text{meas}(n \times r)}^T + \mathbf{E} \quad (3)$$

$$\mathbf{T}_{\text{meas}(n \times r)}^T = (\mathbf{P}_{\text{cal}(m \times r)}^T \mathbf{P}_{\text{cal}(m \times r)})^{-1} \mathbf{P}_{\text{cal}(m \times r)}^T \mathbf{X}_{\text{meas}(m \times n)} \quad (4)$$

In view of the orthogonality of the PCs, namely $\mathbf{P}_{\text{cal}(m \times r)}^T \mathbf{P}_{\text{cal}(m \times r)} = \mathbf{I}$, we rewrite eqn (4),

$$\mathbf{T}_{\text{meas}(n \times r)} = \mathbf{X}_{\text{meas}(m \times n)}^T \mathbf{P}_{\text{cal}(m \times r)} \quad (5)$$

Then, the residuals of each measurement are derived by use of calibration $\mathbf{P}_{\text{cal}(m \times r)}$ and measurement $\mathbf{T}_{\text{meas}(n \times r)}$,

$$\mathbf{E} = \mathbf{X}_{\text{meas}(m \times n)} - \mathbf{P}_{\text{cal}(m \times r)} \mathbf{T}_{\text{meas}(n \times r)}^T = \mathbf{X}_{\text{meas}(m \times n)} - \mathbf{P}_{\text{cal}(m \times r)} \mathbf{P}_{\text{cal}(m \times r)}^T \mathbf{X}_{\text{meas}(m \times n)} = (\mathbf{I}_{(m \times m)} - \mathbf{P}_{\text{cal}(m \times r)} \mathbf{P}_{\text{cal}(m \times r)}^T) \mathbf{X}_{\text{meas}(m \times n)} \quad (6)$$

The residual matrix $\mathbf{E} = [\mathbf{e}_{\text{meas}_1} \dots \mathbf{e}_{\text{meas}_j} \dots \mathbf{e}_{\text{meas}_n}]$ contains errors and the part of the data set \mathbf{X}_{meas} not explained by r PCs, \mathbf{I} denotes a unit matrix of size $(m \times m)$. If there are no unexpected chromatographic features presenting in the measured herbal samples, all the residual concentrations of \mathbf{E} will merely consist of noise. Otherwise, high concentrations of

some residuals will feature additional chromatographic profiles of unexpected samples.

Secured compensation for systematic fitting errors

Although this conventional PCR is able to elucidate implicitly the chromatographic occurrences in the calibration data,²⁷ its procedure cannot handle the unexpected features not being contained in calibrations but emerging during measurements. Moreover, this unexpected emergence indicates the quality problem of diverse herbal objects, and these unexpected features need to be taken care of and detected. Generally, the residuals \mathbf{E} are disturbed by real unexpected measurements, unavoidable noise and systematic fit errors from the regression model. However, factually, it is impossible to split up of the residuals into individual disturbances just as described in the literature.^{20,21} In order to correctly detect unexpected profiles and securely evaluate residual concentrations, a compensation for systematic fit errors is necessarily performed on \mathbf{E} . Vogt and Mizaikoff developed an sPCR method for this compensation, and applied it to detect and correct uncalibrated spectral features. In contrast to spectra, herbal fingerprints comprise a multitude of ingredients and their relative contents are diverse. Therefore, such diversity being considered, sPCR is modified for detection of unexpected features and evaluation of chromatographic measurements in our practice. The residual profiles within a certain narrow interval chromatographic window are compared with the parts of the r PCs located inside the same chromatographic window. If the residuals are found to be linearly dependent on the parts of the r PCs inside the same window, it can be concluded that the residuals are systematic fit errors, for which approximate compensations are conducted as described in the following section. If the residuals are independent of the parts of the r PCs, high residuals indicate unexpected chromatographic features.

Modification of secured principal component regression

sPCR compensates approximately the systematic fit errors in the residuals by means of many so-called reflecting-line estimates in terms of three necessary requirements, as documented in ref. 20 and 21. Moreover, this compensation involves a prior two-step denoise and as well decision about the residuals being linearly dependent on or independent of the PCs, based on the largest singular value ratio of two submatrices. Therefore, it is not easy to use when there are a lot of diverse objects in compositional distribution. Analogous to this compensation, we develop a flexible method. First, we split the data set $\mathbf{X}_{\text{meas}(m \times n)}$ in the chromatographic direction into a series of narrow windows with a certain width, say $\mathbf{X}_{i_w\text{in}\text{size}}$ of size ($w\text{in}\text{size} \times n$). In the subscript, $w\text{in}\text{size}$ indicates the width of the chromatographic window, i be the i^{th} narrow window, ranging from 1 to q , and q is the last of all the windows. Moreover, these narrow windows are equi-interval and consecutive. Thus, the residual profiles are recalculated as a second matrix $\mathbf{D} = [d_{\text{meas}_1} \dots d_{\text{meas}_j} \dots d_{\text{meas}_n}]$ by adding appropriate compensation of the systematic fit errors to the original residuals $\mathbf{E} = [e_{\text{meas}_1} \dots e_{\text{meas}_j} \dots e_{\text{meas}_n}]$ in a piecewise manner,

$$\mathbf{D}_{i_w\text{in}\text{size}} = (\mathbf{I} - \mathbf{P}_{i_w\text{in}\text{size}}\mathbf{P}_{i_w\text{in}\text{size}}^T)\mathbf{E}_{i_w\text{in}\text{size}} (i = 1, 2, \dots, q) \quad (7)$$

In eqn (7), the matrix $\mathbf{E}_{i_w\text{in}\text{size}}$ denotes the part of original residuals \mathbf{E} inside the i^{th} narrow window, $\mathbf{P}_{i_w\text{in}\text{size}}$ contains the parts of the PCs inside the same window, \mathbf{I} is a unit matrix of size ($w\text{in}\text{size} \times w\text{in}\text{size}$), $\mathbf{D}_{i_w\text{in}\text{size}}$ is the recalculated residuals corresponding to the residuals $\mathbf{E}_{i_w\text{in}\text{size}}$ after compensation for the systematic fit errors. Compared with sPCR, this modified algorithm is friendly to use, and the results are more reasonable. Of sPCR, there is a limitation for detecting uncalibrated absorptions, *i.e.* it does not work when some uncalibrated measurements fall completely into the calibration model. msPCR does not have this limitation, on the contrary, it would rather concern unexpected chromatographic features in a set data of herbal fingerprints for quality control.

Assessment of residuals with variance discriminant

When \mathbf{X}_{meas} is regressed into the calibration model, the residual profiles \mathbf{D} have to be evaluated as to whether they contain only noise or some unexpected chromatographic features. For this purpose, a decision threshold is necessarily defined for the noise level of data. On the assumptions that the experimental noise of herbal fingerprints is subject to a statistical distribution and the noise level remains stable over time under the same experimental conditions and allowing for the $(k - r)$ noise PCs of the calibration data, namely $\mathbf{P}_{\text{noise}(m \times (k-r))} = [p_{r+1} \dots p_k]$, we may obtain the noise $\mathbf{X}_{\text{noise}(m \times n)}$ with $\mathbf{X}_{\text{cal}(m \times k)}$ and $\mathbf{P}_{\text{noise}(m \times (k-r))}$,

$$\mathbf{X}_{\text{noise}(m \times k)} = \mathbf{P}_{\text{noise}(m \times (k-r))}\mathbf{P}_{\text{noise}(m \times (k-r))}^T\mathbf{X}_{\text{cal}(m \times k)} \quad (8)$$

From this, the noise level is deduced as the average of standard deviation of the noise $\mathbf{X}_{\text{noise}(m \times k)}$,

$$\text{std}^{\text{noise}} = \frac{1}{k} \sum (\text{var}(\mathbf{X}_{\text{noise}(m \times k)}))^{1/2} \quad (9)$$

Based on this noise level, the decision threshold is defined as,

$$\text{thre} = 3\text{std}^{\text{noise}} = \frac{3}{k} \sum (\text{var}(\mathbf{X}_{\text{noise}(m \times k)}))^{1/2} \quad (10)$$

In the same way, the standard deviations of the measurement residuals $\mathbf{D} = [d_{\text{meas}_1} \dots d_{\text{meas}_j} \dots d_{\text{meas}_n}]$ can also be calculated,

$$\text{std}^{\mathbf{D}} = (\text{var}(\mathbf{D}_{(m \times n)}))^{1/2} = [\text{std}_1^{\mathbf{D}} \dots \text{std}_j^{\mathbf{D}} \dots \text{std}_n^{\mathbf{D}}] \quad (11)$$

Concerning the detection of unexpected features, $\text{std}_j^{\mathbf{D}}$ is compared with the decision thre , respectively. If $\text{std}_j^{\mathbf{D}} \leq \text{thre}$, it is concluded that the column residual d_{meas_j} consists only of noise. Otherwise, there is more than noise in the residual d_{meas_j} , which indicates unexpected chromatographic features. Hence, msPCR accomplishes all its procedures consequently.

Experimental^{3,4,28}

Plant materials and reagents

Erigeron breviscapus materials were collected from different geographical origins or producing fields in Yunnan province, P. R. China. *Ginkgo biloba* samples were purchased from several pharmaceutical stores, vendors/companies and

collected from various producing areas in the mainland, P. R. China. All of these samples were identified by one of authors (advanced pharmacist) from Guangzhou Institute for Drug Control, P. R. China. Standard extract EGb761 was kindly donated to one of authors from Guangzhou Institute for Drug Control, P. R. China by the Beaufour-Ipsen Company in France.

Analytical grade methanol and phosphoric acid were purchased from chemical reagents factory of Guangzhou, Guangdong, P. R. China. The other reagents used were also of analytical grade. Ultrapure water (18.2 M Ω) was obtained by means of a Milli-Q apparatus from the Millipore Corporation (France) and was used for mobile phase preparation. The mobile phase was vacuum filtered through a filter of pore size 0.45 μm .

Sample preparation

Extraction of raw materials of *Erigeron breviscapus*. For analytical purposes, each plant sample was ground to fine powder by a pulverizer. A 0.50 g amount of the fine powder was extracted with 30 ml methanol in a water bath, under reflux at 80 $^{\circ}\text{C}$ for 30 min, and with a second aliquot of 20 ml methanol for 15 min. The two extracts were blended, filtered and taken to dryness in a water bath. The residue was dissolved into 5 ml of 60% methanol and was filtered through a film of pore size 0.45 μm . 10 μl filtrate was analyzed by HPLC-DAD.

Extraction of extracts and products of *Ginkgo biloba*. Pre-weighted extracts of *Ginkgo biloba* were dissolved with 5 ml methanol. A 1 ml volume of this solution was filtered through a Millipore filtration unit type HV 0.45 μm . 10 μl filtrate was injected into the HPLC system.

Pre-weighted *Ginkgo biloba* products were extracted using CQ250 ultrasonic cleaner with 20 ml methanol for 15 min, and the extract was kept still for a moment at room temperature. The solution was then filtered through a glass filter covered with a filter paper. Next, the solution evaporated in vacuum to about 1 ml, and was diluted into methanol in a 5 ml volumetric flask. Then, a 1 ml volume of this solution was filtered through a Millipore filtration unit type HV 0.45 μm . 10 μl filtrate was injected into the HPLC system.

Chromatography system and procedure

Determination of extracts of *Erigeron breviscapus*. The analysis of samples from *Erigeron breviscapus* was achieved on an analytical HPLC unit (Hewlett-Packard series 1100, Agilent Inc.), using a reversed-phase Alltima C₁₈ (Alltech Inc., 5 μm particle size; 250 \times 4.6 mm I.D.) column. The mobile phases were composed of A (CH₃OH)–B (0.1% H₃PO₄) (25 : 75, v/v) at the start, and then linearly changed to A–B (75 : 25, v/v) at 50 min. The flow rate was 0.8 ml min⁻¹. The column temperature was kept at 25 $^{\circ}\text{C}$. The detection was accomplished with a diode-array detector and chromatograms were recorded in the range of 200–400 nm at 1 nm step⁻¹.

Determination of extracts and of products of *Ginkgo biloba*. The analysis of samples from *Ginkgo biloba* was performed on

a high performance liquid chromatography (HP series 1100, Agilent Inc.) equipped with a diode-array detector and an intelligent quaternary pump. A Spherisorb ODS2 C₁₈ column (4.0 \times 250 mm, 5 μm particle size, Agilent Inc.) was used with the following mobile phase A₁ consisting of water–acetonitrile–isopropanol–citric acid (1000 : 200 : 30 : 4.92 g), and B₁ (1000 : 470 : 50 : 6.08 g) after being filtered through a film of pore size 0.45 μm . At flow rate of 1.0 ml min⁻¹, a gradient elution program was set from 100% A₁ to 100% B₁ over a period of 25 min. The column compartment was kept at 25 $^{\circ}\text{C}$. The detection was carried out using a diode-array UV detector in the range of 200–400 nm.

Data analysis

Data analysis was performed on a Pentium 1.7 G processor. All involved programs were coded in MATLAB 5.3.

Results and discussion

Correction of chromatographic shifts

In chromatography it is commonly encountered that disturbances, particularly chromatographic shifts introduced by the measurement device, may trouble the data processing and analysis. Therefore, the chromatographic shift of original fingerprints should be corrected as necessary with local least squares procedures in advance of performing msPCR. The local least squares procedures have been well elaborated in ref. 4 and 28 and are not expanded on here.

Autoscaling of chromatographic data

Autoscaling^{29–33} is one of the data pretreatments. In a chromatographic fingerprint set, it may be written in matrix notation as,

$$\mathbf{Y} = (\mathbf{X} - (\mathbf{I} \cdot \mathbf{I}^T/m)\mathbf{X})\mathbf{W} = (\mathbf{I} - \mathbf{I} \cdot \mathbf{I}^T/m)\mathbf{X}\mathbf{W} \quad (12)$$

where \mathbf{I} is a one-vector, matrix \mathbf{X} includes all the objects with the length m after chromatographic shifts have been corrected, *i.e.* both the corrected calibration fingerprints and the corrected measurements are subjected to autoscaling, and \mathbf{Y} is the matrix holding autoscaled fingerprint data. \mathbf{W} is a diagonal matrix with the scaling parameter for the i^{th} column fingerprint x_i of matrix \mathbf{X} on its i^{th} diagonal element w_i . \mathbf{I} denotes a unit matrix of size ($m \times m$). Examining eqn (12), autoscaling mainly consists of two parts: centering across the column fingerprints, the left-half multiplication with projection $(\mathbf{I} - \mathbf{I} \cdot \mathbf{I}^T/m)$, and scaling within the column fingerprints, the right-half multiplication with weighted matrix \mathbf{W} . In our study, the weight of the i^{th} column fingerprint x_i is chosen to be the inverse of the standard deviation of it, say, $w_i = \text{var}(x_i)^{-1}$.

Centering across every column fingerprint of data sets should be performed, as stated by Bro and Smilde,³³ if there is an average constant part across this vector or if modeling such an average can provide an approximately reasonable model for the data. Simply speaking, centering is able to make a difference to the fingerprint data by respectively removing individual averages of all the column fingerprints of matrix \mathbf{X} , sensibly reducing the rank of the PCR model for calibration,

and helpfully avoiding the numerical problem of rank estimation of matrix \mathbf{X}_{cal} , significantly increasing fit to both the calibration objects and the measurements. Generally, a proper centering operation removes the averages from the column data but does not change the structural model of the data.

Scaling within the column fingerprints with the inverse of the standard deviation of the residual variance of each corresponding fingerprint is used to adjust scale differences of individual fingerprints. This performance makes the variance of each chromatographic object be initially identical in the model. Thus, how the measurement fingerprints are subjected to the msPCR model can be described with systematic fitting variation as much as possible, particularly in the case that some fingerprint measurements have different chromatographic responses (e.g. two fingerprint segments in Fig. 1). The autoscaled fingerprints have the same shape as the original fingerprints corrected for chromatographic shift.

Real *Erigeron breviscapus* data set of chromatographic profiles

The task of the calibration model is to examine three components of the regression triplet in terms of data, model and method.³⁴ For this consideration, we herein illustrate the procedure of msPCR with respect to its performance ability by first analysing a data set of thirty-three chromatographic fingerprints from different herbal *Erigeron breviscapus* samples measured through a liquid chromatographic column. Seen from their appearances, most of these fingerprints seem similar, but there exist chromatographic shifts and some quantitative variations of the phytochemical compositions, which make immediate comparison and evaluation of these fingerprints difficult and also even unreasonable. Fig. 1 depicts two amplified original chromatographic segments of regular samples at the wavelength of 280 nm. Therefore, with a local least squares procedure, such shift disturbance of every original fingerprint has to be prior resolved for subsequent modelling calibration and detection.

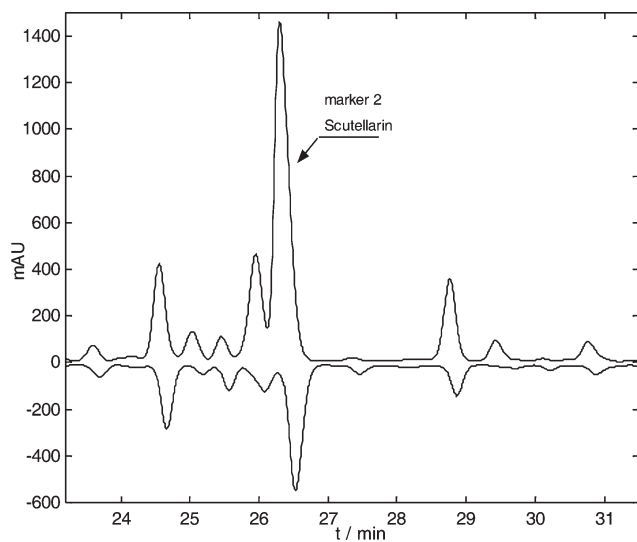


Fig. 1 An amplified segment of two representative chromatographic fingerprints of *Erigeron breviscapus* herbal samples measured at the wavelength of 280 nm by HPLC-DAD.

According to eqn (12), all the chromatographic fingerprints are autoscaled for msPCR. As for the calibration data, the fingerprints for a model should be selected on the relevant principle that they are able to feature as fundamental chromatographic information about as many regular herbal samples as possible. Since the investigated herbal samples were derived and influenced from different collected seasons, various geographical origins and culture locations and so on, two chromatographic profiles from any different samples are not necessarily consistent in composition with each other or proportional at all despite the shift correction. On account of this reason, seventeen regular fingerprints are selected with SIMCA³⁵ into matrix \mathbf{X}_{cal} serving as calibration objects. Consequently, the first five PCs (\mathbf{P}_{cal}) are determined to be sufficient to model \mathbf{X}_{cal} , i.e. $r = 5$. The estimate of the relevant PCs is thoroughly reviewed elsewhere.³⁶ Here, it is not elaborated. All the remaining PCs ($\mathbf{P}_{\text{noise}}$) of the calibration data are utilized to achieve the noise $\mathbf{X}_{\text{noise}}$ and the decision threshold (thre).

When the msPCR algorithm runs its procedure, the width of the narrow chromatographic windows for systematic fitting error compensation, as mentioned above, has to be determined, which is at least equal to or larger than the number of PCs, say $\text{winsize} \geq r$. So as to secure appropriate compensation of systematic fitting errors and correctly detect and evaluate the measurements of herbal samples, a good winsize should not exceed r significantly. Practically, the changes of winsize do not much influence the results, when its magnitude approaches the number of PCs. In our two instances, the window width is determined empirically, $\text{winsize} = 8$ for *Erigeron breviscapus*, $\text{winsize} = 6$ for *Ginkgo biloba*.

Sixteen measurement fingerprints other than those calibration objects are detected *via* the msPCR algorithm, and their residual profiles are evaluated whether they contain only noise or unexpected chromatographic features. Table 1 lists the results of detection and evaluation of these measurements. As can be seen from Table 1, the residual deviations (std^{D}) of measured samples 1, 3, 4 and 8 are obviously larger than the decision threshold (thre). On the contrary, the residual

Table 1 Results of msPCR and sPCR modeling of sixteen measurement fingerprints from *Erigeron breviscapus* samples

| Series no. | msPCR | sPCR |
|------------|---------------------------------------|---------------------------------------|
| | $\text{std}^{\text{D}} - \text{thre}$ | $\text{std}^{\text{D}} - \text{thre}$ |
| 1 | 0.2803 | -0.0674 |
| 2 | -0.0359 | -0.3082 |
| 3 | 0.3861 | 0.0928 |
| 4 | 0.6830 | 0.3984 |
| 5 | -0.1247 | -0.3886 |
| 6 | -0.1016 | -0.3744 |
| 7 | -0.1879 | -0.4419 |
| 8 | 0.6244 | 0.3313 |
| 9 | -0.1973 | -0.4457 |
| 10 | -0.1620 | -0.4177 |
| 11 | -0.2155 | -0.4634 |
| 12 | -0.1497 | -0.4128 |
| 13 | 0.0660 | -0.2445 |
| 14 | -0.1306 | -0.3852 |
| 15 | -0.0301 | -0.2948 |
| 16 | -0.0216 | -0.3042 |

deviations (std^D) of measured samples 2, 5, 6, 7, 9, 10, 11, 12, 14, 15 and 16 are less than the decision threshold (thre), and the residual deviation (std^D) of measured sample 13 is just above the decision threshold. This suggests that measurements 1, 3, 4 and 8 feature unexpected chromatographic profiles, while only measurement noise contributes to the residual concentrations of measurements 2, 5, 6, 7, 9, 10, 11, 12, 14, 15 and 16 after their projection onto the r PCs and compensation for systematic fitting errors.

In order to clarify this outcome we have to take herbal samples, particularly those containing unexpected chromatographic features, into account since the chromatographic profiles relate to individual samples. The fingerprints of measured samples 4, 8 and one regular sample in the calibration are shown together in Fig. 2. By visual comparison, their phytochemical ingredients are extremely different in spite of being obtained under the same experimental extract and chromatographic conditions. The botanical materials of measured sample 4 took on three pieces of capitulate inflorescence, unlike regular samples of *Erigeron breviscapus* with just single inflorescence, while the material of measured sample 8 held two pieces of inflorescence. Therefore, measured samples 4 and 8 are excluded from herb *Erigeron breviscapus*. The eluting sequences of phytochemicals in the fingerprints of measured samples 1 and 3 were nearly consistent with that of the regular sample (see Fig. 3), but many compositions are very diverse, especially active component scutellarin (marker 2) and characteristic component pyromeconic acid (marker 1). Viewed from the characteristics of medicinal materials, the leaves of samples 1 and 3 appeared filemot. It is possible that different breed and harvest result in their inferior quality.

The methanol extracts of various parts of herbal materials have also been analyzed by HPLC-DAD, and the results

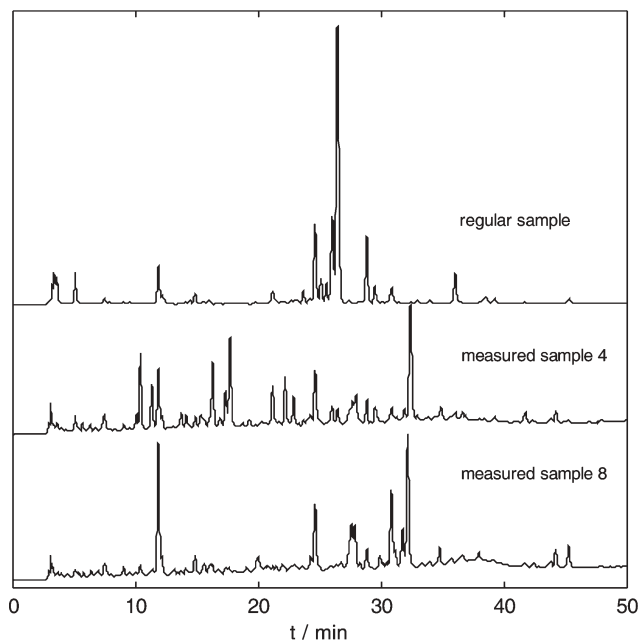


Fig. 2 The fingerprint chromatograms of measured samples 4 (middle), 8 (bottom) with one regular sample (top) from *Erigeron breviscapus* at a wavelength of 280 nm.

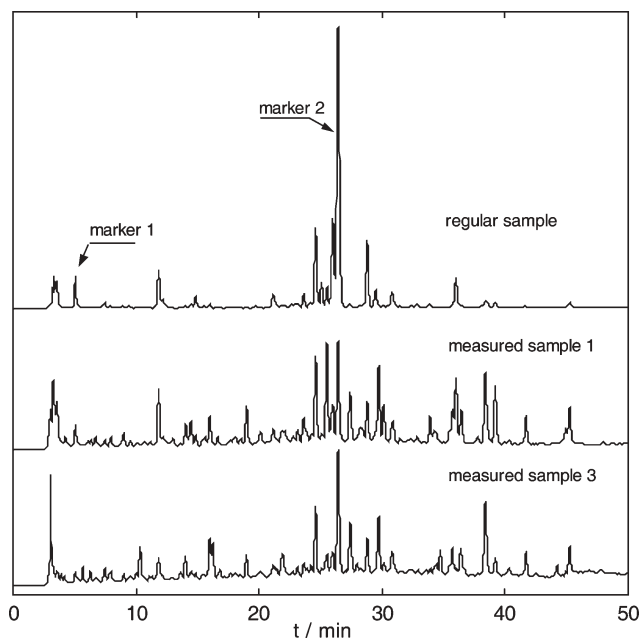


Fig. 3 The fingerprint chromatograms of measured samples 1 (middle), 3 (bottom) with one regular sample (top) from *Erigeron breviscapus* at a wavelength of 280 nm.

suggest that different parts contain different phytochemicals at low concentrations. For example, leaf, stem and flower with no root together constitute samples 14, 15 and 16, which make their chromatographic profiles in compositional distribution a little different from the major regular fingerprints. The measured sample 13 is composed of only leaf. Hence, it could not be qualified into the samples of good quality. However, the measured fingerprints 2, 5, 6, 7, 9, 10, 11 and 12 are generally in accordance with the profiles of regular samples even though there are insignificant concentration variations of some compositions.

The qualitative evaluation of the measurements may be observed using the first two PCs, say PC1 and PC2, as shown in Fig. 4. The fingerprints are moderately classified based on their chromatographic features. The calibrations are denoted with black circle “•”, and the measurements with open circle “○”. Seventeen regular samples cluster together (black circles in Fig. 4), accordingly defined as a regular class. Measured samples 4 and 8 are clearly differentiated, farthest apart from the regular samples, and samples 1 and 3 locate in the middle of the score plot, kept from the regular class. While samples 2, 5, 6, 7, 9, 10, 11, 12, 14, 15 and 16 are scattered in or near the regular class, as shown in Fig. 4. The projection point of sample 13 is a little way from them. It proves that the msPCR algorithm is capable of detecting unexpected chromatographic features and evaluating herbal samples qualitatively when enough regular samples are given to the calibration model and these regular samples selected as calibration objects in a clear class should specify a herbal plant species in terms of *Erigeron breviscapus* without controversy.

Now we compare the results of msPCR and sPCR modeling of sixteen measurement fingerprints from *Erigeron breviscapus* samples, as listed in Table 1. One can observe that the (std^D – thre) values of these samples calculated from msPCR are

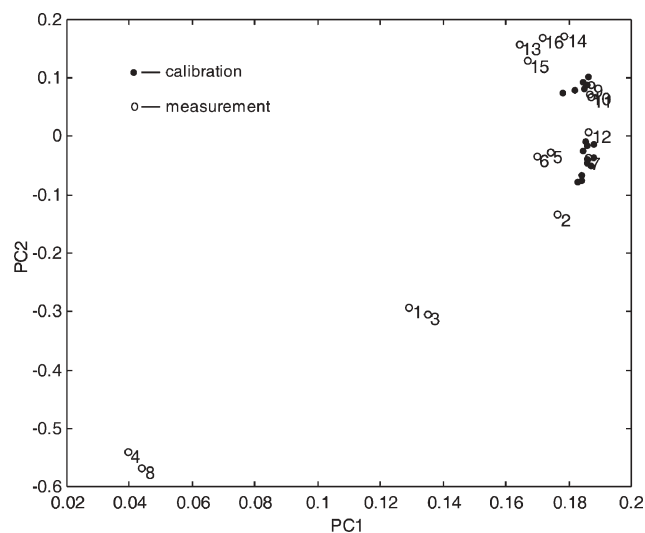


Fig. 4 The PCA score plot for all the chromatographic fingerprints from *Erigeron breviscapus*.

acceptable, though they are larger by almost 0.3 than the corresponding ones from sPCR. This indicates msPCR is able to improve reliably the detection of unexpected chromatographic features in herbal fingerprints.

Regarding the calibration objects as measurements, we tested msPCR with completely satisfactory results, which are listed in Table 2.

Real *Ginkgo biloba* data set of chromatographic fingerprints

Flavonoids are some of the pharmacologically active and chemically characteristic ingredients from *Ginkgo biloba*. EGb761 consisting mainly of thirty-three flavonoids^{5,37,38} is well defined as a standard extract for the quality control of beneficial extracts from *Ginkgo biloba*, and recognized widely in the world. Fig. 5 shows the chromatographic fingerprint of EGb761 measured at a wavelength of 360 nm. In this figure, twenty-one peaks are marked with letters “a’”, “b’”, “c’”, etc. Of them, marker components “h’”, “q’”, “r’” represent

Table 2 Results from performing msPCR on seventeen calibration objects from *Erigeron breviscapus* samples

| Series no. | std ^D – thre |
|------------|-------------------------|
| 1 | -0.1630 |
| 2 | -0.1773 |
| 3 | -0.1819 |
| 4 | -0.1662 |
| 5 | -0.1895 |
| 6 | -0.1981 |
| 7 | -0.1947 |
| 8 | -0.1112 |
| 9 | -0.1864 |
| 10 | -0.2083 |
| 11 | -0.1584 |
| 12 | -0.1879 |
| 13 | -0.2216 |
| 14 | -0.1824 |
| 15 | -0.1989 |
| 16 | -0.1788 |
| 17 | -0.1516 |

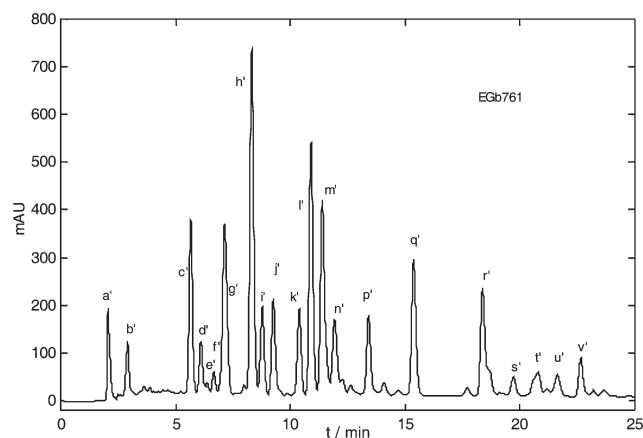


Fig. 5 Chromatographic fingerprint of all the flavonoids in standard extract EGb761 from herbal *Ginkgo biloba* measured at a wavelength of 360 nm by HPLC-DAD.

active or characteristic luteolin, 3-*O*-[2-*O*-[6-*O*-(*p*-hydroxy-*trans*-cinnamoyl)- β -D-glucosyl]- α -L-rhamnosyl]quercetin, and 3-*O*-[2-*O*-[6-*O*-(*p*-hydroxy-*trans*-cinnamoyl)- β -D-glucosyl]- α -L-rhamnosyl]kaempferol respectively. On the consistency principle, when the chromatographic fingerprints of all the beneficial extracts from *Ginkgo biloba* are very similar to that of EGb761, they are of good quality. However, in our application of msPCR, the quality evaluation of the beneficial extracts is not based on simple similarity of their chromatographic fingerprints with that of EGb761 but on their residual concentrations regressed by a clear class of regular samples that are determined to be good, which is achieved still with SIMCA.

The calibration is based on fifteen fingerprints of regular extracts from *Ginkgo biloba*, including that from standard EGb761, obtained at a wavelength of 360 nm. The first six PCs are selected for the msPCR model. The measurements consist of fourteen objects. By means of the threshold (thre) derived from X_{cal} , all the measurements are detected, and the results are given in Table 3. Seen from it, the values of their residual deviations (std^D) minus the threshold are obvious: only samples 1, 2, 4 and 9 are determined to be unexpected. The residual profiles of these samples indicate different chromatographic features and concentration distributions of some

Table 3 Results of msPCR modeling of fourteen measurement fingerprints from *Ginkgo biloba* samples

| Series no. | std ^D – thre |
|------------|-------------------------|
| 1 | 0.1015 |
| 2 | 0.0770 |
| 3 | -0.0485 |
| 4 | 0.1032 |
| 5 | -0.2377 |
| 6 | -0.0355 |
| 7 | -0.0011 |
| 8 | -0.0727 |
| 9 | 0.1294 |
| 10 | -0.0824 |
| 11 | -0.0585 |
| 12 | -0.1066 |
| 13 | -0.2317 |
| 14 | -0.2611 |

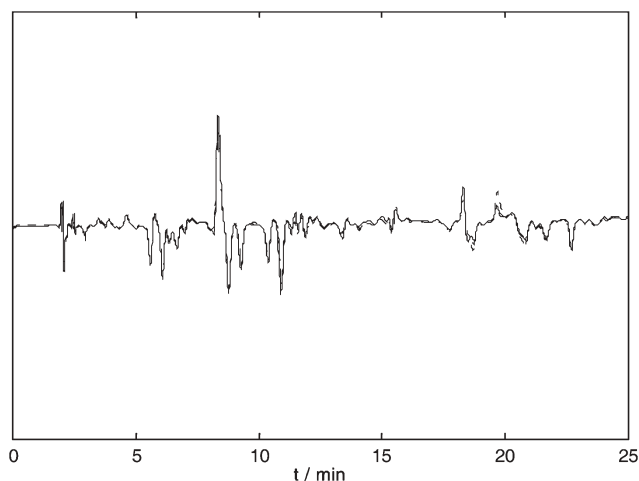


Fig. 6 Residual concentrations of measured samples 1, 2 and 4 tested via msPCR.

chemical ingredients. Fig. 6 and 7 display three residual profiles and chromatographic fingerprints of samples 1, 2, 4, respectively. It is very clear that the presence of phytochemical luteolin in high relative content and the absence of marker component b' with low concentrations of the remainders result chiefly in their irregular residual profiles, and hence, they are determined to be unexpected. Thereby, it is doubtful whether the vendors/companies added cheap chemical luteolin into their products acting as commercial medicines. Sample 9 appears to have a similar compositional distribution to that of EGb761 except for some varying chromatographic features, especially marker components a', h', q', r', etc., which makes sample 9 discriminate from regular samples. In all, it is due to the concentration variations of some relevant ingredients of the chromatographic fingerprints of samples that some herbal extracts or products are successfully detected and discriminated from regular ones, and quality control is ultimately implemented.

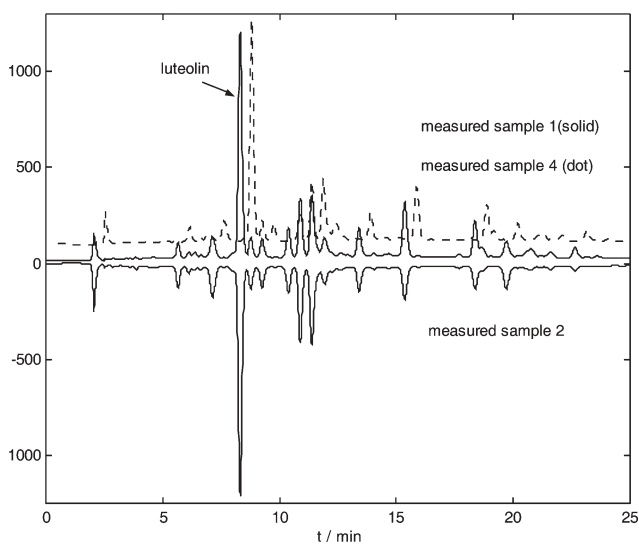


Fig. 7 Chromatographic fingerprints of measured samples 1 (top solid curve), 2 (bottom) and 4 (top dot curve) from herbal *Ginkgo biloba* at a wavelength of 360 nm.

Table 4 Results from performing msPCR on fifteen calibration objects from *Ginkgo biloba* samples

| Series no. | std ^D – thre |
|------------|-------------------------|
| 1 | -0.2701 |
| 2 | -0.2488 |
| 3 | -0.2453 |
| 4 | -0.2296 |
| 5 | -0.2552 |
| 6 | -0.2637 |
| 7 | -0.2315 |
| 8 | -0.2425 |
| 9 | -0.2395 |
| 10 | -0.2367 |
| 11 | -0.2461 |
| 12 | -0.2409 |
| 13 | -0.2543 |
| 14 | -0.2534 |
| 15 | -0.2625 |

Fifteen calibration objects from *Ginkgo biloba* are also tested with msPCR. Table 4 specifies their satisfactory results. All of the chromatographic objects of fifteen calibrations and fourteen measurements are used for PCA. Fig. 8 depicts the three-dimensional (3D) score plot. These first three PCs (PC1, PC2, PC3), which describe the most chromatographic feature variations related to different extracts or products from *Ginkgo biloba*, are used to make differentiation clearer. As before, the calibrations are denoted with black circles “•”, and the measurements with open circles “○”. From Fig. 8, the scores are classified clearly into two groups: one regular group embodies fifteen calibrations and ten measured objects, and the other unexpected group covering samples 1, 2, 4 and 9, a little way from the regular samples.

Conclusions

This study has attempted to describe, detect and evaluate a herbal medicine (extract or product) more with a definite class

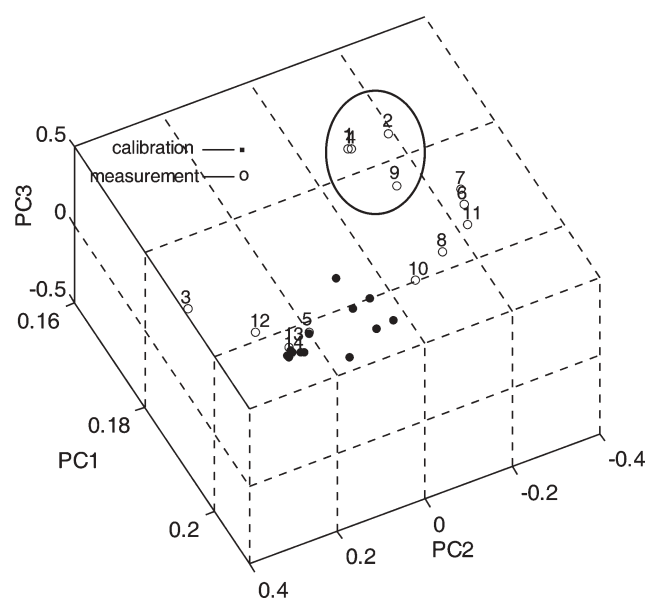


Fig. 8 Three-dimensional score plot from twenty-nine *Ginkgo biloba* chromatographic fingerprints.

of many regular objects than a certain standard reference since careful consideration should be given to intrinsic diversities of phytochemicals of herbal medicine. It is possible and feasible to use msPCR to detect unexpected samples and control the quality of herbal medicines, qualitatively. In addition, when it comes to the use of msPCR for detecting unexpected chromatographic features and evaluating the fingerprint quality, it is very critical to carefully select which fingerprints are used for the calibration model. These calibration fingerprints should contain as much regular chromatographic information as possible that is able to feature the majority of herbal samples. Moreover, the various sources and different influences from growth conditions are also taken into account for herbal samples.

Acknowledgements

This work was supported by research grants from the National Natural Science Foundation of China (No. 20235020) and the Cultivation Fund of the Key Scientific and Technical Innovation Project, Ministry of Education of China (No. 704036). The authors also thank Professor Yi-Ping Du (College of Chemical Engineering, Shandong University of Technology, Zibo, P. R. China), and Professor Qing-Song Xu (ChemoAC, FABI, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium) for helpful discussions.

References

- 1 World Health Organization, *General Guidelines for Methodologies on Research and Evaluation of Traditional Medicines*, Geneva, 2000.
- 2 R. Bauer, *Drug Inf. J.*, 1998, **32**, 101–110.
- 3 F. Gong, Y. Z. Liang, P. S. Xie and F. T. Chau, *J. Chromatogr., A*, 2003, **1002**, 25–40.
- 4 B. Y. Li, Y. Hu and Y. Z. Liang, *J. Sep. Sci.*, 2004, **27**, 581–588.
- 5 A. Hasler and O. Sticher, *J. Chromatogr.*, 1992, **605**, 41–48.
- 6 T. P. Layloff, *Pharm. Technol.*, 1991, **15**, 146–148.
- 7 R. D. Kirchhoefer, *J. AOAC Int.*, 1992, **75**, 577–580.
- 8 Anon, *Gold Sheet*, 1994, **28**, 1–10.
- 9 W. J. Welsh, W. Lin, S. H. Tersigni, E. Collantes, R. Duta and M. S. Carey, *Anal. Chem.*, 1996, **68**, 3473–3482.
- 10 P. Valentão, P. B. Andrade, F. Areias, F. Ferreres and R. M. Seabra, *J. Agric. Food Chem.*, 1999, **47**, 4579–4582.
- 11 P. S. Xie, *Trad. Chin. Drug Res. Clin. Pharmacol.*, 2001, **12**, 141–151.
- 12 World Health Organization, *Guidelines for the Assessment of Herbal Medicines*, Geneva, 1991.
- 13 World Health Organization, *Quality control methods for medicinal plant materials*, Geneva, 1998.
- 14 L. Z. Lin, X. G. He, M. Lindenmaier, G. Nolan, J. Yang, M. Cleary, S. X. Qiu and G. A. Cordell, *J. Chromatogr., A*, 2000, **876**, 87–95.
- 15 F. Gong, Y. Z. Liang, H. Cui, F. T. Chau and B. T. P. Chan, *J. Chromatogr., A*, 2001, **909**, 237–247.
- 16 Y. Hu, Y. Z. Liang, B. Y. Li, C. J. Xu and Z. D. Zeng, *Acta Chim. Sin.*, 2003, **61**, 1466–1470.
- 17 Y. Y. Cheng, M. J. Chen and W. D. Tong, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1068–1076.
- 18 E. R. Collantes, R. Duta, W. J. Welsh, W. L. Zielinski and J. Brower, *Anal. Chem.*, 1997, **69**, 1392–1397.
- 19 Y. Y. Cheng, M. J. Chen and W. J. Welsh, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1959–1965.
- 20 F. Vogt and B. Mizaikoff, *J. Chemom.*, 2003, **17**, 225–236.
- 21 F. Vogt and B. Mizaikoff, *Anal. Chem.*, 2003, **75**, 3050–3058.
- 22 S. Wold, K. Esbensen and P. Geladi, *Chemom. Intell. Lab. Syst.*, 1987, **2**, 37–52.
- 23 H. Martens and T. Naes, *Multivariate Calibration*, John Wiley & Sons, New York, 2nd edn, 1991.
- 24 M. H. Zhang, Q. S. Xu and D. L. Massart, *Chemom. Intell. Lab. Syst.*, 2003, **67**, 175–185.
- 25 G. Golub and C. V. Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 2nd edn, 1989.
- 26 W. Press, S. Teukolsky, W. Vetterling and B. Flannery, *Numerical Recipes in C*, Cambridge University Press, New York, 2nd edn, 1992.
- 27 F. Vogt, U. Klocke, K. Rebstock, G. Schmidtke, V. Wander and M. Tacke, *Appl. Spectrosc.*, 1999, **53**, 1352–1360.
- 28 B. Y. Li, Y. Hu, Y. Z. Liang, P. S. Xie and Y. P. Du, *Anal. Chim. Acta*, 2004, **514**, 69–77.
- 29 R. J. Barnes, M. S. Dhanoa and S. J. Lister, *Appl. Spectrosc.*, 1989, **43**, 772–777.
- 30 R. J. Barnes, M. S. Dhanoa and S. J. Lister, *J. Near Infrared Spectrosc.*, 1993, **1**, 185–186.
- 31 M. S. Dhanoa, S. J. Lister, R. Sanderson and R. J. Barnes, *J. Near Infrared Spectrosc.*, 1994, **2**, 43–47.
- 32 Q. Guo, W. Wu and D. L. Massart, *Anal. Chim. Acta*, 1999, **382**, 87–103.
- 33 R. Bro and A. K. Smilde, *J. Chemom.*, 2003, **17**, 16–33.
- 34 M. Meloun, J. Militký, K. Kupka and R. G. Brereton, *Talanta*, 2002, **57**, 721–740.
- 35 S. Wold and M. Sjostrom, in *Chemometrics: Theory and Application*, ed. B. R. Kowalski, ACS Symposium Series, American Chemical Society, Washington, DC, 1977, vol. 52, pp. 243–282.
- 36 E. N. Malinowski, *Factor Analysis in Chemistry*, John Wiley & Sons, New York, 3rd edn, 2002.
- 37 O. Sticher, *Planta Med.*, 1993, **59**, 2–11.
- 38 World Health Organization, *WHO monographs on selected medicinal plants*, Geneva, 1999, pp. 154–167.