Spectra Data Analysis and Calibration Modeling Method Using Spectra Subspace Separation and Multiblock Independent Component Regression Strategy

Chunhui Zhao, Furong Gao and Fuli Wang
Dept. of Chemical and Biomolecular Engineering, The Hong Kong University of Science and Technology,
Clear Water Bay, Kowloon, Hong Kong SAR

DOI 10.1002/aic.12333
Published online July 19, 2010 in Wiley Online Library (wileyonlinelibrary.com).

In this article, a spectra data analysis and calibration modeling approach is proposed for the estimation of the concentration of sources species in chemical mixture. Based on the multiplicity of underlying spectra characteristics, it designs spectra subspace separation and multiblock independent component regression modeling strategy. It is performed in two steps: The first step aims at an automatic partition of the original wavelength space into different spectra subspaces to reveal the changes of underlying spectra information. In different spectra subspaces, each being well fitted by one independent component analysis (ICA) model, it better explores the existing chemical constituent species of interest. In the second step, multiblock regression system is designed for concentration estimation. The advantage is mainly to allow for easier interpretation and enhanced understanding by zooming into different smaller specific segments and thus well tracking the wavelength-varying effects on qualities. It is theoretically and experimentally illustrated that the proposed method can result in better predictive power compared with standard ICR (SICR) modeling focusing on the full-range wavelength. © 2010 American Institute of Chemical Engineers *AIChE J,* 57: 1202–1215, 2011

Keywords: independent component regression, spectra subspace separation, multiblock regression system, source spectra profiles, component concentration estimation

Introduction

During the past decades, the use of spectroscopic information[1–19] has received much attention and begun to emerge as an important technique, which is being heavily encouraged and practiced for different purposes. Predicting a dependent variable from the spectra measurement is a frequently encountered problem in chemometrics. To analyze the substance composition in chemical mixture, calibration model is often constructed using the mixture spectra together with the reference concentration of constituents to form a quantitative prediction relationship.[10–19] Common multivariate calibration methods[20–24] used for spectra analysis include principal component regression (PCR) and partial least-squares (PLS) regression. They are based on such a fact that variable collinearity is typical among spectra wavelengths in which some can be represented as a linear combination of some latent variables (LVs). To deal with the problem of high data dimensionality and redundancy since spectra data often

Correspondence concerning this article should be addressed to F. Gao at kefgao@ust.hk.
Current Address of Fuli Wang: College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning Province, P. R. China.

comprise more variables (wavelengths) than observations (samples), PCR and PLS reduce the number of spectra variables by means of LV extraction, so the original spectra data space is shrunk to a feature subspace of smaller dimension. Those underlying features are then used for regression modeling. Moreover, having realized that not all wavelengths are useful for quality prediction, various variable (feature) selection methods[8,9,12,19] have been designed. In this way, only those most quality-related wavelengths are retained as descriptors for regression modeling, which, however, may lead to information loss compared with the original spectra space, more or less. Actually, spectra measurement of a mixture is often a linear combination (with coefficients corresponding to the proportions) of the spectra of its constituent species.[10,25–28] It would be useful if the component spectra can be recovered from the mixture spectra. None of above mentioned methods can properly identify the unknown species in mixture as well as their concentration.

Independent component analysis (ICA) algorithm[29] is deemed to be able to deliver this function. Given only observations that are assumed to be linear combinations of some source signals, it is designed based on higher-order statistics to extract the constituent species in the form of independent components (ICs) as well as determining their effects on the observed mixture spectra, which is called blind source signal separation process. This is clearly beneficial when all or some components of a mixture are unknown. Previous work[10,25,26,28] have applied ICA on spectra data, which successfully proved their effectiveness in recovering the components of interest from spectra mixture and estimating their concentration. IC regression (ICR) method was first proposed by Chen and Wang[25] to the near-infrared (NIR) spectra data analysis. The authors have pointed out that by comparing the spectra of separated ICs with the spectra library of pure substances, it was possible to identify unknown constituent species existing in the mixture. Shao et al.[26] further reported ICR could be a promising tool to retrieve both quantitative and qualitative information from complex chemical data sets when used to build the regression model between NIR spectra and the routine components of plant samples. Besides the completely equivalent quantitative prediction performance compared with PCR, ICs could give more chemical explanations than principal components (PCs), which were found to be strongly correlated to the NIR spectra of source components in spectra.

Although many investigators have reported that ICs extracted from mixture spectra are able to capture the essential structure and are chemically meaningful, some underlying problems have not been desirably addressed nevertheless. Commonly, evolving along wavelength direction, the mixture spectra show significant fluctuations and dynamics, which actually are alternately dominated by the spectral profiles of different constituent species. Some basic chemical components are mainly influential to limited spectra wavelength regions. It reveals different underlying chemical characteristics and different influential relationships on quality properties. Here it is called "multiplicity" of spectral characteristics. It would be meaningful to have an insight on the local underlying chemical information within one specific spectra wavelength region. Clearly, a unified ICA feature extraction model over the full-range spectra space is not sufficient and desirable enough to recover the real constituent features and track the varying chemical characteristics along wavelength direction. Considering that multiplicity of spectral characteristics, it is necessary to divide the original spectra data space into different subspaces as indicated by the underlying changes, analyze their respective roles as well as their correlations and thus develop the corresponding spectra data analysis and modeling framework. This is of particular interest and should be especially addressed, which, however, has not been brought into focus yet in the foregoing reviewed literatures.

Based on the above analyses, a multibock ICR (MBICR) spectra data analysis and calibration modeling strategy is presented on the basis of spectra subspace separation. Particularly, it will address the changes of underlying chemical characteristics over different wavelength regions and their different roles in concentration estimation. To achieve the above purpose, it is implemented in two steps. First, considering the specific effect of ICA on spectroscopy, it is used to identify multiple spectra subspaces from the original wavelength space. In different subspaces, which cover different spectra wavelengths and enclose different underlying chemical characteristics, both constituent spectra and their mixing relationships are figured out, revealing different dominant source spectra profiles. In the second step, a multiblock regression system is designed on the basis of subspace separation in which their different specific effects on concentration estimation are checked as well as their integrated contributions. The resulting regression system thus reveals more comprehensive underlying information beneficial to quality prediction.

This article is organized as follows. In the next section, the knowledge of ICA is introduced, including its basic theory and model representation. The multiblock spectra data analysis and calibration modeling strategy is then described in detail and its underlying principle is also explained. The application to the real spectra case demonstrates the effectiveness of the proposed method. Discussion is conducted based on the results, highlighting the suitability of the proposed method, also pointing out its future directions and possible improvements. Finally, conclusions are drawn in the last section.

The Conventional ICR Algorithm

As a two-step calibration method, the basic idea of conventional ICR[25,26] is a combination of ICA algorithm and multiple linear regression (MLR), which is conceptually very similar to PCR modeling idea. The only difference is that ICs and mixing matrix obtained by ICA are used for regression instead of PCs and loadings matrix obtained by PCA. In the first-step, ICA feature extraction is performed to estimate both the latent components $\mathbf{s}$ and the demixing relationship $\mathbf{W}$ from the process measurements $\mathbf{x}$ without any related prior knowledge, a process termed blind separation. It is assumed that the J spectra measurement variables $x_1,x_2,\ldots,x_J$ can be described as linear combinations of R ($R \leq J$)ICs $s_1,s_2,\ldots,s_R$. It finds the statistically independent non-Gaussian hidden factors, or as independent as possible in terms of higher-order statistics. Such a representation is reported to be able to recover the true structure of the measured data with rich chemical meanings. Ideally, if the separated ICs exactly match the pure substances constituting the mixture, then mixing matrix will agree well with the concentrations of the substances existing in mixture. In practice, however, they cannot match very well, and therefore it cannot be taken for granted that the elements in the mixing matrix are concentration. Then in the second step, like PCR, regression analysis can be readily performed between the estimated mixing matrix and the real concentration measurements based on simple least-squares algorithm.

In the first step, for a set of spectra with J wavelengths acquired on N samples, $\mathbf{X}(J \times N)$, a common ICA model can be formulated as below:

$$\begin{aligned} \mathbf{S} &= \mathbf{XW} \\ \mathbf{A} &= \left(\mathbf{S}^{\mathrm{T}}\mathbf{S}\right)^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{X} \\ \mathbf{X} &= \mathbf{SA} + \mathbf{E} \end{aligned} \tag{1}$$

where, $\mathbf{S}(J \times R)$ is the estimated ICs from the observed spectra, which actually are the spectra estimation of the pure constituents in the mixture. $\mathbf{W}(N \times R)$ is the demixing matrix by which the ICs can be directly calculated from the mixture spectra. The mixing matrix, $\mathbf{A}(R \times N)$, actually indicates the effects of the substances on mixture spectra. The implicit reasonable hypothesis of the above model is thus that the mixture spectra are a linear combination (with mixing coefficients corresponding to the proportions) of the component spectra. $\mathbf{R}(J \times N)$ is the unexplained residual. R is the number of retained ICs.

After first-step analysis, the mixing matrix ($\mathbf{A}(R \times N)$) is prepared as regressors for the second-step calibration analysis. A linear regression can be readily calculated to relate it with the quality, concentration matrix, $\mathbf{Y}(N \times J_Y)$ (where, J_y is the number of quality indices) through a simple least-squares calculation:

$$\begin{aligned} \Theta &= \left(\mathbf{AA}^{\mathrm{T}}\right)^{-1}\mathbf{AY} \\ \hat{\mathbf{Y}} &= \mathbf{A}^{\mathrm{T}}\Theta \end{aligned} \tag{2}$$

Here, ICR readily solves the typical collinearity problem of MLR calculations by guaranteeing an invertible matrix $\mathbf{AA}^{\mathrm{T}}$.

Proposed MBICR Algorithm

As analyzed before, for spectra data, the wavelength variables are of great amount and over different wavelength regions, the underlying chemical characteristics are different with different dominative source spectra. A unified ICA model focusing on full-range wavelength may not probe the varying underlying information comprehensively enough when different underlying spectra information is mixed together. The resulting mixing matrix, which is directly related to the estimation of constituent concentration, can result in high prediction error when the recovered ICs poorly match the real pure substances existing in the chemical mixture. Therefore, one spectra data space should be further divided into several different subspaces if the inherent spectra information changes. Without losing generality, over different subspaces, different underlying chemical characteristics and different effects on the final products are hinted. To further investigate their specifics, multiblock modeling idea is deemed to be proper. Multiblock strategy[30–34] was developed to improve the interpretability of multivariate methods and has been widely applied to practical industrial processes for fault detection and diagnosis. It was reported to be able to catch the abnormal event earlier and better reveal the subsection within which the operation fault had occurred. Here, in the designed MBICR approach, the key point is how to automatically partition the spectra wavelengths into conceptually meaningful blocks. Blocking of spectra variables can help to check the spectra space more comprehensively, revealing both the local detail of different spectra wavelength regions and their global information, and then employ them for quality prediction. The modeling detail is described in the following subsections.

Spectra subspace separation

To improve the spectra model representability and the resulting IC extraction performance, one should properly identify different segments along wavelength direction. Each of them should be able to be well approximated by one ICA model with sufficient representability. The similar principle was once adopted by Camacho and Pico[35] to perform automatic phase division. Here inspired by its core idea, a complete automatic spectra subspace separation algorithm is designed. It performs a greedy search based on ICA model performance evaluation. A division performance index, the mean squared errors (MSE), is set up, which is related to the unexplained squared error of each wavelength region after ICA model fitting. It is calculated as MSE $= \frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}\left(x_{i,j} - \hat{x}_{i,j}\right)^2$, where $x_{i,j}$ is the measured spectra variable ($j = 1,2,\ldots,J$) in each sample ($i = 1,2,\ldots,N$) and $\hat{x}_{i,j}$ is its reconstructed value by ICA. During the separation process, every possible wavelength interval will be tried and the best spectra separation points will be figured out in terms of reconstruction power. Any possible spectra subdivision is evaluated with the percentage of reduction of MSE, which is used to check whether the subdivision can improve the representation performance of ICA model. The one with the highest reduction value of MSE is chosen and the procedure recursively proceeds with the resulting segments (corresponding submatrices of spectra data), as shown in Figure 1.

The input of the division algorithm involves:

(a) the whole spectra wavelength space

(b) the minimum length of the separated wavelength region (minlen)

(c) the improvement threshold of MSE to accept a subdivision

(d) the maximum number of separated spectra subspaces (maxnum)

The output is the spectra subspace separation result. The recursive division procedure is implemented as below:

(1) Set the initial number of separated spectra regions *numreg* = 1.

(2) Input the prepared spectra data set, perform the modified ICA algorithm[36] on the data and get the initial ICA model. Calculate the resulting MSE_0 index value to evaluate the model representation performance.

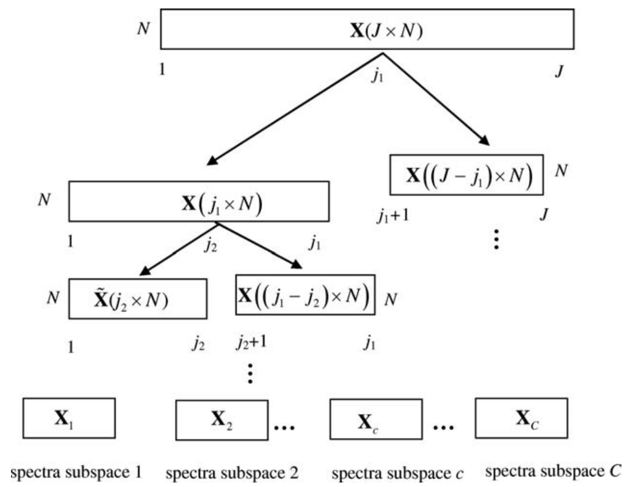

Figure 1. Spectra subspace identification scheme.

(3) For each wavelength interval (j) within the current spectra data set, if the subdivision in j generates two segments with length higher than the predefined minimum region length (minlen), perform ICA on the two segments respectively; calculate the resulting new MSE values for both segments, denoted as MSE_1^{j-1} and MSE_1^{j-2}.

(4) Find the sampling wavelength interval ($j^\bullet$) at which the average of MSE_1^{j-1} and MSE_1^{j-2} is lowest. Comparing both MSE_1^{j-1} and MSE_1^{j-2} with MSE_0, if the improvement for either segment, $\frac{\mathrm{MSE}_0-\mathrm{MSE}_1^{j-1}}{\mathrm{MSE}_0}$ and $\frac{\mathrm{MSE}_0-\mathrm{MSE}_1^{j-2}}{\mathrm{MSE}_0}$, does not reach the predefined improvement threshold, then stop this branch.

(5) Otherwise, accept subdivision and update the number of separated wavelength regions numreg := numreg + 1. If the current numreg is no longer less than the predefined maxnum, stop the iteration procedure, output the division result.

(6) Otherwise, recursively repeat steps (2)–(5) for either of the two resulting segments respectively, each now employed as the new spectra input data space in step (2).

Based on the above spectra subspace separation strategy, multiple specific regions are obtained, enclosing different spectra wavelengths. It provides different analysis scenes for ICA-based source signal extraction and improves the ICA model representability. The advantage of the subspace separation is that in addition to a global view for the whole spectral region, one also obtains local analysis angles for different spectra wavelength regions. It is easy to explore how they will act under the influence of each other. It is deemed that their respective underlying characteristics as well as contributions to quality properties tend to be hidden by each other to a certain extent when the whole spectra wavelengths are not separated. By zooming into separate spectra subspaces, it provides potential for capturing the key factors in different subspaces for quality interpretation and improvement.

Moreover, it should be noted in the separation strategy, one question remains to be discussed in the context of the ICA projection: the choice of the number of ICs, that is, the model order. Because of the intrinsic formulation of the ICA problem, the IC number cannot exceed the sample number of observed spectra. Here, first, focusing on the entire wavelength region, the number of retained ICs for modeling is determined by cross-validation so that it can lead to a small reconstruction error (MSE). This is implemented and evaluated by considering both training and testing data sets. Then the same number of ICs is retained for modeling the resulting subspaces during the whole subspace separation procedure. It is done by considering that the determined IC number is mainly used to aid in evaluating whether the separation could improve the ICA model representability with the same model order. After the final subspace separation result is obtained, in each subspace, the new proper model order will be determined properly again for multiblock ICR modeling.

Analysis and discussion

From the above algorithm, the identification of spectra subspaces have a close relationship with the choice of three input parameters, including the minimum length of the separated wavelength region (minlen), the improvement threshold of MSE to accept a subdivision and the maximum number of separated spectra subspaces (maxnum); and the searching procedure itself. Here, focusing on these factors, their effects and roles will be analyzed and discussed.

Parameter Choice. Generally, for the above-mentioned three parameters, their relaxation or tightness may lead to the variety of subspace separation result to a certain extent, indirectly impose effects on the accuracy and sensitivity of subspace model representation, and finally influence the quality prediction performance. However, up to now, the selection of these parameters is not something for which specific rules can easily be given. It is also hard to define definite criterion or uniform standard to strictly quantify them. Instead, it greatly depends on the specific characteristics of each practical case. For different cases, maybe different parameter values should be chosen. Sometimes, prior expertise can provide a reference standard for the general idea of parameter choice. When no prior information exists, as a rule of thumb, the parameters can be determined by trial and error. Therefore, their determination is inevitably affected more or less by artificial subjectivity factors. Here focusing on their different effects on separation result and thus modeling performance, it is simply discussed as below:

The minimum length of the separated wavelength region (minlength) can be combined with the maximum number of subspaces (maxnum) in order not to derive too many subspaces with undesirably small size. The maxnum parameter equal to 1 and minlength equal to J mean no subspace separation can be performed, resulting global ICA modeling result; while smaller minlength and larger maxnum mean more subspaces can be obtained. It should be noted that there is not such a definite relationship: $\mathrm{maxnum} = \frac{J}{\mathrm{minlength}}$. They, actually, can be deemed to be two-fold constraints. It is possible that at one iteration step, the obtained subspace length no longer satisfies the requirement of minlength, but the resulting subspace number may still stay well below the upper limit of maxnum. Generally, the minlength should not be too small in order to avoid obtaining subspaces with insufficient wavelength intervals, which will not provide stable and sufficient statistical information. Moreover, it should not be too large in order to avoid subspaces with overlong wavelengths which may cover different underlying chemical characteristics and thus lose the sensitiveness to their changes.

The improvement threshold directly determines whether a division can be accepted or not based on evaluating whether the resulting model representation can better reconstruct the current underlying variation information. It is fluctuant from 0 to 1. A threshold value equal to 0 means the separation can be accepted even if it only presents a very small improvement. A threshold value equal to 100% means no subdivision will be justified, resulting in a global ICA representation. A value between 0 and 1 results in a compromise between the two above extreme cases.

From the above analysis, these parameters actually present a compromise between model stability and sensitivity. Generally speaking, smaller improvement threshold and min-length value and larger maxnum parameter result in more spectra subspaces with shorter wavelength each of which can be easily fitted with a simple ICA model but the model stability may be compromised to a certain extent. On the contrary, fewer subspaces are obtained with longer wavelength where each can provide more modeling information but the model sensitivity may be sacrificed more or less since different characteristics are mixed together. By setting different parameter values step by step, one may generally interrogate their specific effect.

Searching Algorithm. During the separation iteration, it performs a greedy search based on ICA model performance evaluation. However, greedy algorithms mostly (but not always) fail to find the globally optimal solution, because they usually do not operate exhaustively. They may make commitments to certain choices too early, which prevent them from finding the best overall solution later. Nevertheless, they are useful because they are quick to think up and often give good approximations to the optimum. Alternatively, there are also many other well-developed algorithms,[37–39] which are reported to find global optimal point, such as evolutionary programming (EP),[38] branch and bound,[39] and so on. Evolutionary programming, originally conceived by Lawrence, is a stochastic optimization strategy, which places emphasis on the behavioral linkage between parents and their offspring. Branch and bound is the popular approach for solving a given NP-hard (nondeterministic polynomial-time hard) discrete optimization problem for the best solution in which a tree is set up with the same starting and ending nodes and a tree search algorithm is basically used. They guarantee to get an optimal solution, but the only drawback for them is that they generally take exponential number of repetition of steps. For example, during solving the well-known traveling salesman problem (TSP),[39] branch and bound algorithm takes care of all possible tours and neglects only those about which it is sure they could not produce optimal tour, resulting in complexity in exponentials. Comparatively, the principal advantage of greedy selection is that it is cheap, both in space and time. The main drawback, however, is the potential convergence to a suboptimal solution. This is not to say that all greedy algorithms fail to find an optimum consistently. Greedy search may return a global optimum if each locally selected optimum is part of this global optimum, where the problem is said to have ''optimal substructure.'' In general, they often provide a successful tradeoff between solution optimality of the problem and the speed of computing the problem.

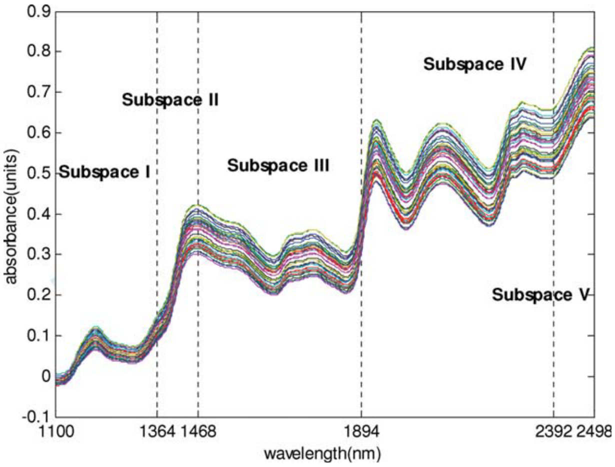

Figure 2. Mixture spectra trajectory and the subspace separation result.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Multiblock ICR modeling

Over different spectra regions, the underlying spectra information may vary significantly. Located in different spectral regions, the influences on ICA feature extraction and regression performance are also different. For certain source species, it can be accurately extracted in some spectra subspace, whereas in other spectra subspaces, it cannot be estimated reliably enough or even cannot be figured out since it does not dominate the current mixture spectra. Different IC decomposition results lead to different effects on quality prediction, so that information from multiple subspaces should be combined advisably for comprehensive and reliable source species estimation. Multiblock modeling idea is expected to be able to wisely achieve this purpose. Besides the detailed comprehension of local spectra regions and a global view when all spectra wavelengths are considered simultaneously, another attraction is that it is possible to develop a robust calibration model since different wavelength regions are stacked smartly. Moreover, the importance of extracted ICs can also be evaluated by checking the different roles of their associated mixing coefficients in quality prediction.

In the following, based on separation of different spectra regions, that is, blocking of spectra wavelength variables, the multiblock regression system is formulated.

Let $\{\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_c,\ldots,\mathbf{X}_C\}$ be C ($J_c \times N$)-matrices of spectra measurements with the same N samples scanned at different spectral regions (where J_c is the number of wavelength variables in each subspace), which all point to the same concentration ($\mathbf{Y}$) of constituents in mixture. From a mathematical point of view, by performing ICA decomposition,[36] their underlying source spectra profiles are figured out respectively located in different spectra subspaces:

$$\begin{aligned} \hat{\mathbf{X}}_1 &= \mathbf{S}_1\mathbf{A}_1 \\ \hat{\mathbf{X}}_2 &= \mathbf{S}_2\mathbf{A}_2 \\ &\vdots \\ \hat{\mathbf{X}}_C &= \mathbf{S}_C\mathbf{A}_C \end{aligned} \tag{3}$$

Table 1. ICA Decomposition Result Over Different Spectra Subspaces

		Spectra subspaces				
		I	II	III	IV	V
Model order		6	5	6	5	4
ICA reconstruction $R^2\hat{\mathbf{X}}_c$ (%)	Training samples	88.5828	98.8009	99.5211	99.8590	99.9089
	Testing samples	88.3905	98.8029	99.5253	99.8545	99.9054

where, different source spectra ($\mathbf{S}_c(J_c \times R_c)$) and multiple mixing relationships ($\mathbf{A}_c(R_c \times N)$) are decomposed (where R_c is the number of retained ICs, which may be different over different spectra subspaces). $\mathbf{A}_c$ reveals the different contributions of source spectra profiles to mixture spectra in different subspaces/blocks. Here it should be noted that the ICA model order (R_c) may be different from that used in spectra subspace separation. In each spectra subspace, the proper IC number is chosen by cross-validation focusing on both training and testing data sets so that it can lead to a small reconstruction error by $\mathbf{X}_c - \hat{\mathbf{X}}_c = \mathbf{X}_c - \mathbf{S}_c\mathbf{A}_c$.

All the mixing matrices obtained from different spectra subspaces can thus be organized as multiple blocks: $\mathbf{A} = [\mathbf{A}_1^{\mathrm{T}}, \mathbf{A}_2^{\mathrm{T}},\ldots,\mathbf{A}_c^{\mathrm{T}},\ldots,\mathbf{A}_C^{\mathrm{T}}]$, which reveal underlying information for quality description. Moreover, it should be noted that these block mixing relationships may correlated with each other and

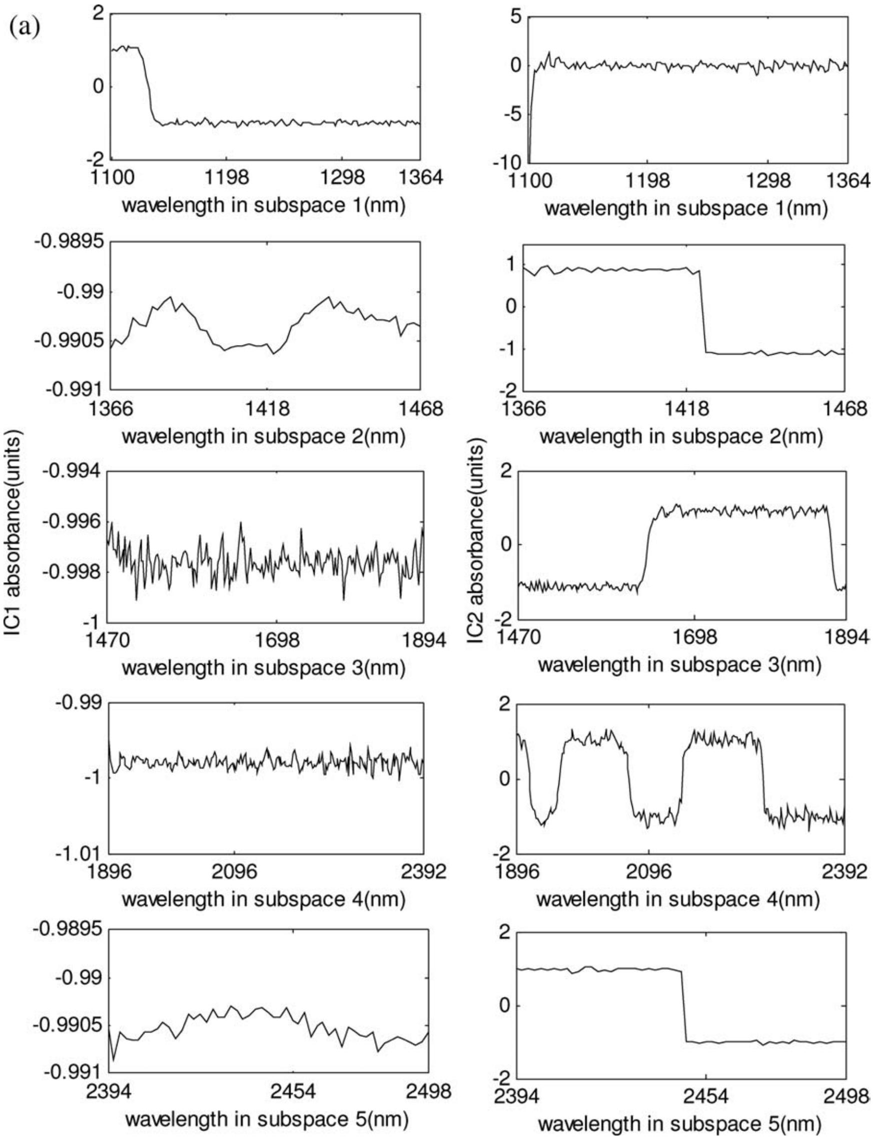

Figure 3. (a) Spectra profile of the first two ICs over five spectra subspaces; (b) mixing coefficients for the first two ICs over five spectra subspaces.

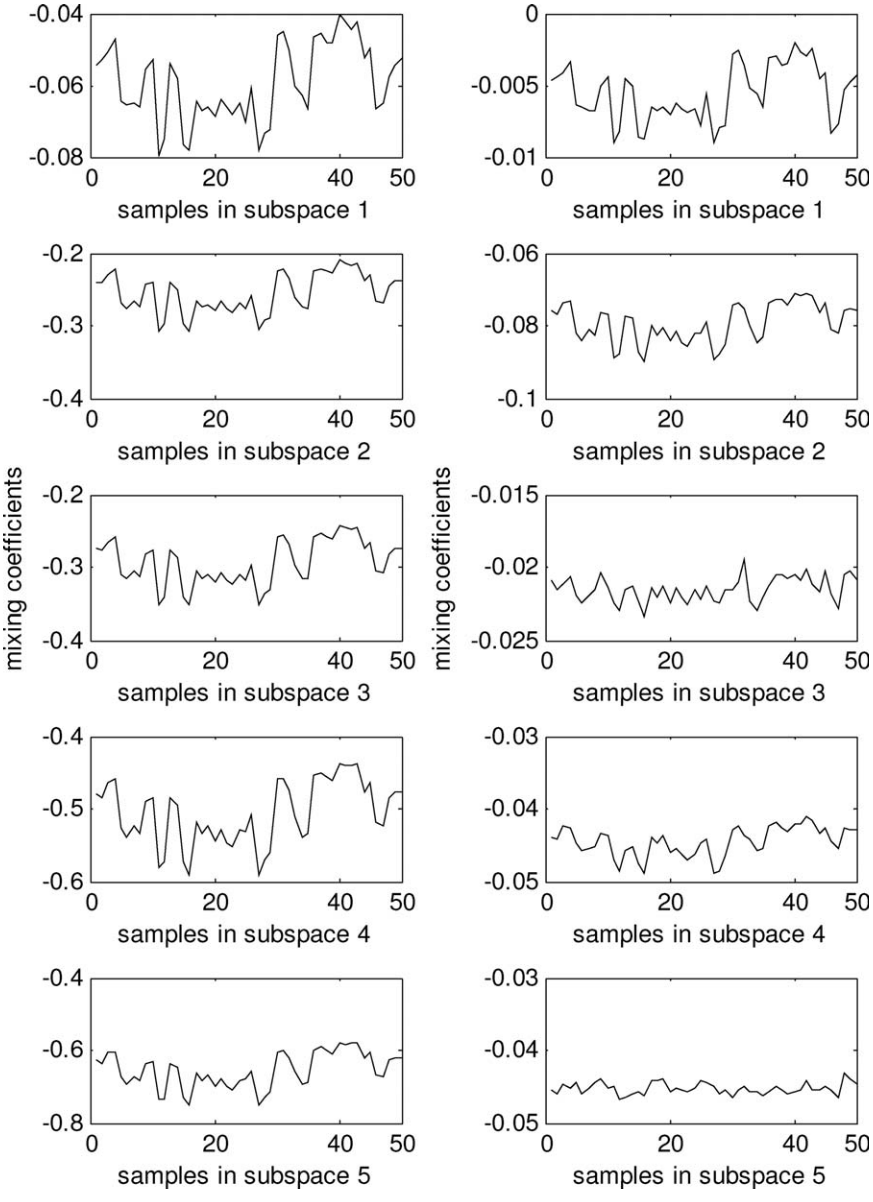

Figure 3. (Continued.)

may also cover some quality-irrelevant information more or less. The existence of quality-irrelevant variations in descriptor space may weaken the regression correspondence and thus lead to complex regression model structures. Therefore, some strategy should be adopted so that the really quality-related underlying information can be focused on. Instead of the simple least-squares calculation as commonly done in ICR,[25,26] a two-step regression modeling strategy proposed by Yu and MacGregor,[40] which combined PLS and canonical correlation analysis (CCA),[41,42] is especially meaningful. It was called PLS-CCA, where, as a post-processing, CCA was implemented on PLS LVs. In this way, it avoided the rank-deficiency problem of single CCA algorithm, got rid of the quality-uninformative variation in PLS LVs, and thus directly maximized the regression correspondence. A parsimonious regression model was thus obtained with the same prediction ability as the standard PLS model. Here, it will be adopted as the basic regression modeling algorithm from which multiblock PLS-CCA (MBPLS-CCA) method is designed as shown in Appendix A. It is proved valuable in subspace-wise calibration modeling in which how different spectra regions covary with qualities under the influence of each other are of special interest.

Based on the MBPLS-CCA modeling algorithm, input the prepared ICA mixing matrices ($\mathbf{A} = [\mathbf{A}_1^{\mathrm{T}}, \mathbf{A}_2^{\mathrm{T}},\ldots,\mathbf{A}_c^{\mathrm{T}},\ldots,\mathbf{A}_C^{\mathrm{T}}]$) and the concentration measurements ($\mathbf{Y}$). For each data table, the means of each column are subtracted to approximately eliminate the main nonlinearity. And each variable is scaled to unit variance to handle different measurement units, thus giving each equal weight. The regression system is then set up from different standpoints:

From the local eye located in each spectra subspace:

$$\begin{aligned}\mathbf{T}_c &= \mathbf{A}_c\mathbf{R}_c\\ \hat{\mathbf{A}}_c &= \mathbf{T}_c\mathbf{P}_c^T\\ \hat{\mathbf{A}}_c^t &= \mathbf{T}\mathbf{P}_c^{tT}\\ \hat{\mathbf{Y}}_c &= \mathbf{T}_c\mathbf{Q}_c^T\end{aligned} \tag{4}$$

where, $\mathbf{R}_c$ is block weights matrix used to directly calculate the block LVs ($\mathbf{T}_c$); $\mathbf{P}_c$ and $\mathbf{P}_c^t$ are the loadings matrices corresponding to local LVs ($\mathbf{T}_c$) and global LVs ($\mathbf{T}$) to describe each spectra block from which two types of block variations ($\hat{\mathbf{A}}_c$ and $\hat{\mathbf{A}}_c^t$) are modeled, revealing the underlying local information. $\hat{\mathbf{Y}}_c$ is the local quality prediction result obtained by each spectra block.

From the global eye dwelling upon all block wavelengths:

$$\begin{aligned}\mathbf{T} &= \mathbf{A}\mathbf{R}\\ \hat{\mathbf{Y}} &= \mathbf{T}\mathbf{Q}^{\mathrm{T}}\end{aligned} \tag{5}$$

where, by weights matrix $\mathbf{R}$, the super LVs ($\mathbf{T}$) are calculated taking into consideration the whole spectra space. $\mathbf{Q}$ is loadings matrix for qualities and $\hat{\mathbf{Y}}_c$ is the final quality prediction.

It can be seen that the subspace separation result provides a rational regression modeling platform, where various model statistics can be calculated to quantitatively evaluate the different roles of different spectra subspaces for an improved model interpretation and spectra understanding. For example, how many of spectra variations are systematic information and how many of them take part in quality prediction, how many of quality variations can be explained by the current subspace, and which subspace is of comparative importance to quality interpretation and prediction. They will be further quantitatively clarified in the Simulation section.

Simulations and Discussions

In this section, the performance of calibration modeling and quality prediction using the proposed method is illustrated through real experiment spectra data. Quantitative and qualitative analyses are performed with respect to the effects of the spectra subspace separation as well as their respective effects on quality prediction in the multiblock modeling strategy. It well illustrates the improvement of the modeling performance and chemical interpretation resulting from the use of spectra subspace separation and the multiblock modeling strategy.

Experiment dataset

The used data set consists of spectra from 80 samples of corn with wavelength ranging 1100-2498nm at 2 nm intervals (700 channels), which are scanned on m5 NIR spectrometer. Therefore, we can collect 700 × 80 spectral observations in all. The corresponding concentration values are involved in the response matrix $\mathbf{Y}(80 \times 4)$, referring to four constituents, moisture, oil, protein, and starch. Most of the data should be used for training, and a smaller portion of the data is used for testing. Here, they are simply randomly partitioned into two sets, 50 samples used for model training and the other 31 used as testing data, which is approximately 2:1 between the samples of two data sets. The corn spectra data are available at the Eigenvector Research homepage: http://software.eigenvector.com/Data/Corn/.

Simulation methodology and results

First, the mixture spectra are shown in Figure 2 taking example for the training samples. Along wavelength direction, the spectral profiles show great fluctuation. As analyzed before, different basic chemical components dominate over different spectra ranges. By performing the proposed subspace separation, five different spectra blocks are obtained, covering different wavelengths, 1100–1364, 1366–1468, 1470–1894, 1896–2392, 2394–2498 nm, respectively, as shown in Figure 2. Over different subspaces, the source spectra are extracted as well as their mixing coefficients. By cross-validation, different number of ICs is retained in each subspace to recover the important source components as shown in Table 1. Moreover, the modeled spectral variations are also calculated, telling the model reconstruction competency in each subspace. Taking example for the first two source components, their spectra profiles are illustrated in Figure 3a over five different subspaces respectively. Moreover, the corresponding ICA mixing parameters are also shown in Figure 3b. It is clear that over different spectra wavelength ranges, both source spectra and their contributions to describe the mixture spectra are different, which demonstrates their different underlying chemical characteristics. The separated subspaces provide different statistical analysis platforms, which may reveal different impacts of different substances on mixture spectra. Especially, if the spectra library of pure substances are known, one just needs to compare the uncovered ICs with them, and then it is easy to identify the unknown constituent species existing in the mixture and know which chemical components dominate in each wavelength subspace. Even the spectra library is not known, we can check this information from another viewpoint. As mentioned before, the better the uncovered ICs match the real pure substances, the better the corresponding coefficients in the mixing matrix approximate the real concentrations. Therefore, here we will focus on analyzing the modeled mixing matrix in each subspace and relate them to the concentration measurement, respectively. It should be noted in this analysis each subspace is handled solely and the block-wise correlations are not taken into consideration to avoid their impacts on the checking of dominate local information. First, the correlations between mixing vectors and the real concentrations are calculated over five different subspaces and shown in Table 2 where for simplicity, their correlation relationships are uniformly expressed in terms of absolute values. In each subspace, for each IC, the first two largest correlation values are marked bold for clarity. In Subspace I, clearly, the decomposed ICs should be correlated with moisture and oil closer than protein and starch since they give higher correlation coefficients. In Subspace II, the ICs also share a certain relationship with the third source substance (protein). Especially, in Subspace V, the latter three ICs generally relate to protein and starch better than moisture and oil. From the results, it is clear that over different subspaces, the ICA result may approximate different concentration indices with different reliabilities. Moreover, in each subspace, the modeled concentration variations are

Table 2. Subspace-Based Statistical Analysis Results

Subspace No.	$\mathbf{Y}_1$	$\mathbf{Y}_2$	$\mathbf{Y}_3$	$\mathbf{Y}_4$
Correlation analysis between mixing coefficients and concentrations				
I	**0.7888**	**0.5970**	0.4405	0.0534
	0.8042	**0.5729**	0.4232	0.0791
	0.7966	**0.5990**	0.4432	0.0579
	0.7613	**0.5654**	0.4228	0.0592
	0.8191	**0.5994**	0.4173	0.0885
	0.7886	**0.5985**	0.4271	0.0704
II	**0.7711**	0.4216	**0.5799**	0.0677
	0.7742	**0.5649**	0.3871	0.1038
	0.7048	0.3843	**0.5493**	0.0835
	0.6830	**0.6075**	0.5373	0.1062
	0.7479	**0.5541**	0.3895	0.0918
III	**0.7773**	0.4319	**0.5806**	0.0608
	0.6480	**0.3469**	0.1175	0.2523
	0.5597	**0.5339**	0.4270	0.0214
	0.0343	**0.3065**	**0.3632**	0.2053
	0.7478	**0.6314**	0.2867	0.2154
	0.6914	**0.5705**	0.3230	0.1300
IV	**0.7814**	**0.5864**	0.4417	0.0520
	0.8213	**0.6591**	0.2964	0.2053
	0.8317	0.3605	**0.4690**	0.0130
	0.3739	**0.5605**	**0.3889**	0.1099
	0.8872	0.5226	**0.5995**	0.0988
V	**0.7881**	**0.5959**	0.4263	0.0693
	0.1422	0.0216	**0.4723**	**0.5085**
	0.3623	0.3133	**0.7316**	**0.4780**
	0.0450	0.0612	**0.5721**	**0.5623**
Modeled concentration variations (%)				
I	75.0847	57.7553	43.9528	27.9708
II	89.7976	43.7199	47.5727	32.5448
III	80.9594	50.4350	58.7748	57.5218
IV	87.2299	56.3211	89.9269	79.9212
V	63.8754	36.1122	58.3791	35.3719

also quantified as shown in Table 2. Generally, the underlying characteristics over five subspaces are all informative for the estimation of the first concentration index (moisture) while different subspaces may be differently informative for the other concentration indices.

The multiple subspaces prepare a desirable regression modeling and quality prediction platform. Considering that these spectra subspaces may correlate and influence with each other, the regression system can be explored by multiblock strategy (MBPLS-CCA). Collecting the five mixing matrices $\mathbf{A} = [\mathbf{A}_1^{\mathrm{T}}, \mathbf{A}_2^{\mathrm{T}}, \mathbf{A}_3^{\mathrm{T}}, \mathbf{A}_4^{\mathrm{T}}, \mathbf{A}_5^{\mathrm{T}}]$, multiblock regression system is set up to relate them with the real concentration measurements ($\mathbf{Y}$) in which the number of retained LVs is set to be 3 by cross-validation. For the four concentration variables, the quality prediction results for training and testing samples are shown in Figure 4a,b, respectively. Visually, they well-capture the general quality variations and yield a generally satisfying prediction trend, which demonstrates both the fitness ability and generalization adaptability for quality estimation. Comparatively, the prediction results by the standard ICR algorithm (here simply notated as SICR) are shown in Figure 5. It is designed as follows: first, a unified ICA decomposition is performed to decompose the entire wavelength range, and then single PLS-CCA is used to fit the regression relationship. In this way, the comparison will focus on revealing the effect of spectra subspace separation. Combining the results shown in Figures 4 and 5, clearly, SICR yields worse prediction performance than MBICR. This demonstrates that resulting from the varying underlying spectra characteristics, a unified ICA decomposition model can not recover source components well enough over the full-range wavelengths. Therefore, the undesirable extraction result, as the descriptors, will directly influence the following regression modeling performance. A summary of multiblock regression modeling results is given in Table 3 based on the training samples in which both modeled descriptor variations and quality variations are taken into account. The associated local variations in each subspace, including two types, one modeled by local LVs ($\mathbf{T}_c$) and the other by global LVs ($\mathbf{T}$), are calculated. From the result, $R^2\hat{\mathbf{A}}_c$ modeled by $\mathbf{T}_c$ is greatly larger than $R^2\hat{\mathbf{A}}_c^t$ modeled by $\mathbf{T}$, which agrees well with what has been mentioned in Appendix A. The predicted quality variations ($R^2\hat{\mathbf{Y}}_c$) are also accounted with respect to the four quality indices respectively, revealing different prediction power of each subspace under the supervision of the other ones. Moreover, it should be noted that they are different from the results shown in Table 2 [modeled concentration variations (%)], where each subspace is isolated from the others without considering their correlations. Here, based on the use of multiblock modeling strategy, it takes the mutual influences of different subspaces into consideration. From the result shown in Table 3, for different quality indices, different subspaces have different contributions. Normally, the critical-to-quality spectra subspace is more accurate and reliable for the prediction of quality variable, so it should have larger $R^2\hat{\mathbf{Y}}_c$. For example, for the

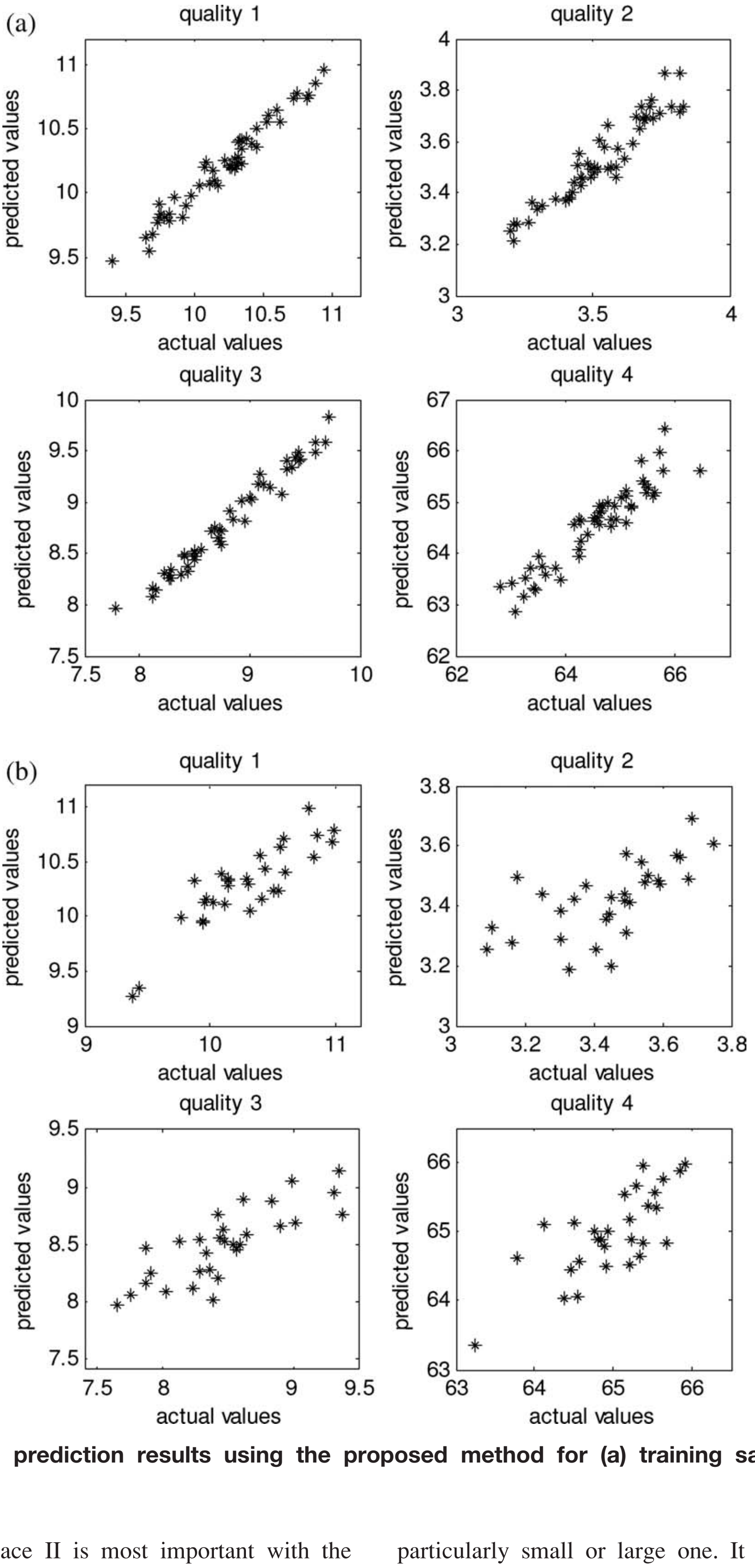

Figure 4. Concentration prediction results using the proposed method for (a) training samples and (b) testing samples.

first quality index, Subspace II is most important with the largest $R^2\hat{\mathbf{Y}}_c$, whereas for the third and fourth quality indices, Subspace V is more critical. Moreover, it is found that each subspace can account for one part of quality variations more or less and the contributions are comparative with no particularly small or large one. It means each spectra subspace is necessary and should be combined to give a comprehensive description of the quality properties. Moreover, the final quality prediction results are also evaluated as indicated by the value of $R^2\hat{\mathbf{Y}}_c$. Clearly, by stacking the

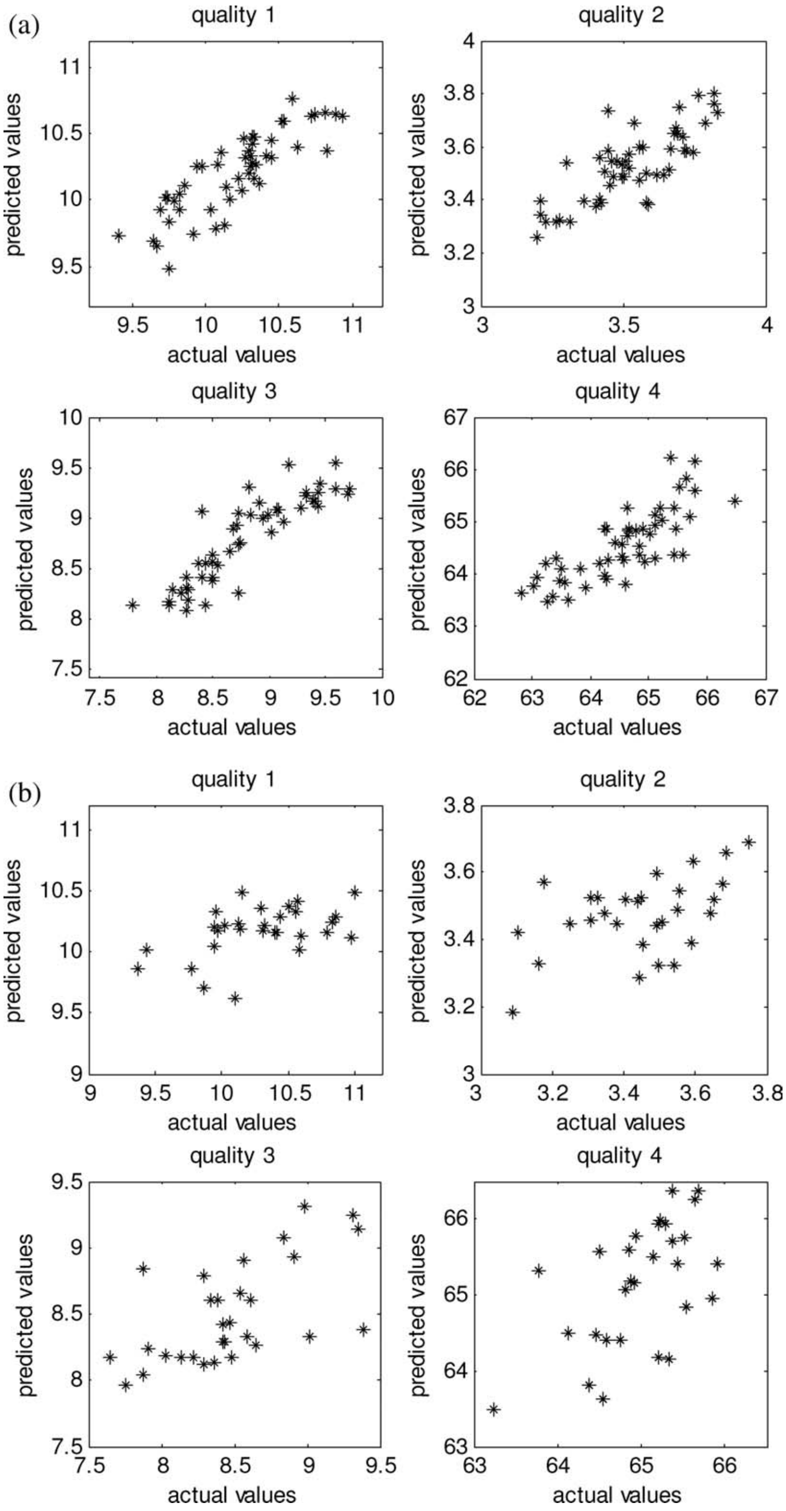

Figure 5. Concentration prediction results using SICR method for (a) training samples and (b) testing samples.

contributions of all blocks, more quality variations are captured. Based on the above quantitative evaluations, we can further realize the meaning of subspace separation in improving statistical interpretation.

Table 4 delivers a comparison of quality prediction results by the proposed method and SICR modeling algorithm. Two indices, mean square error of calibration and mean square error of prediction, which are calculated based on training samples and testing ones respectively, are used to evaluate the prediction performance. The results demonstrate that with no subspace separation, the direct application of single ICA decomposition to the full-wavelength spectra may damage the prediction accuracy. Moreover, with the same number of ICs, it shows worse reconstruction performance to fit the spectra variations than multiblock ICA as indicated by smaller $\boldsymbol{R}^2\hat{\mathbf{X}}$. Such a problem may be more apparent when the difference of underlying chemical characteristics in different spectra subspaces is larger. It can be understood from the fact that the more different the spectra subspaces, the larger their spectra variations so that the less representative and more difficult for a unified ICA model to fit the full-wavelength at the same time. Comparatively, it can be expected that spectra variations can be better fitted by dividing the original spectra space into different subspaces according to the changes of underlying spectra characteristics and then modeling them respectively using multiple specific ICA models. From the comparison result shown in Table 4, it also demonstrates that the proposed multiblock modeling strategy improves the model representability and thus directly benefits the following regression modeling and quality prediction.

Summary and Discussion

This report has illustrated how the new statistical analysis and calibration modeling strategy performs on spectra measurement, and how it compares with standard ICR algorithm. The authors believe, it is the first time that the idea of spectra subspace separation is proposed and an automatic partition is achieved. The simulations conclude with illustrations of how the proposed method performs on one real case, revealing the desirable improvement in model representation and quality prediction. It also allows for the interpretation of the spectra data space to be more comprehensive and meaningful.

There may be still many issues to be investigated in future, but the results of this study constitute a step forward toward spectra data analysis and calibration modeling for practical applications. It is hoped that this report can provide the basis for further work and improvement. Future research might profitably take the following directions:

(a) How to further improve the spectra subspace separation algorithm to obtain more proper model representation? As mentioned before, better searching methods, such as those global optimizers, may be used and even further improved for the current study so that the separated multiple subspaces can present more statistical meanings. As one key issue and critical preparation for the following regression modeling, it is of great meaning and also deserves special attention in our future research work.

Table 3. Summary of Modeled Descriptor and Quality Variations

	Spectra subspaces				
(%)	I	II	III	IV	V
$R^2\hat{\mathbf{A}}_c$	99.6423	98.1785	93.5256	95.3799	95.2582
$R^2\hat{\mathbf{A}}_c^t$	66.0583	59.3548	50.1515	68.1070	50.4200
$R^2\hat{\mathbf{Y}}_c$	80.5751	88.6957	69.7438	83.1602	63.6393
	38.0390	36.6112	34.5367	51.2142	36.3350
	33.3824	20.9817	28.9835	37.5851	61.8515
	24.8990	6.4621	17.2549	10.4636	41.6104
$R^2\hat{\mathbf{Y}}$			96.0712		
			89.7510		
			97.3154		
			86.9521		

Table 4. Modeling and Quality Prediction Result Comparison

			ICA reconstruction $R^2\hat{\mathbf{X}}(\%)$	
Index Method	MSEC	MSEP	Training samples	Testing samples
SICR algorithm	0.2851	1.0801	94.2236	94.2176
	0.3739	0.8699		
	0.2108	0.5587		
	0.3711	0.9413		
MBICR algorithm	0.0385	0.3027	99.7157	99.7108
	0.1004	0.8086		
	0.0263	0.2966		
	0.1279	0.3318		

(b) The separated multiple spectra subspaces provide a potential statistical analysis platform. Focusing on them, how to better reveal their underlying information under the supervision of each other? Whether it is possible to distinguish between their common and unique information so as to further improve spectra understanding and calibration performance? This work may be especially fruitful.

(c) Considering that nonlinearity is not uncommon in spectra data, maybe a nonlinear version of spectra subspace separation and LV extraction for calibration modeling is required so as to better decompose the underlying nonlinear structure.

Conclusions

In this article, it presents a new spectra analysis and calibration method by spectra subspace separation and multiblock modeling strategy. Different spectra subspaces are separated from the original wavelength space according to their different underlying chemical characteristics. This provides a multiblock regression analysis platform from which both local information and global information can be readily obtained. In this way, it can find more interesting and reliable model representation and improve quality prediction performance. Simulation examples have shown the effectiveness of the proposed method. The proposed method is especially recommended when the underlying spectra information is neither isolated to a single, small wavelength region nor spread uniformly. It thus should be generally applicable to a broad range of spectra analysis applications aiding in the optimization of calibration model.

Acknowledgments

This work is supported by China's National 973 program (2009CB320603) and the National Natural Science Foundation of China (No. 60774068).

Literature Cited

1. Wold S, Antti H, Lindgren F, Öhman J. Orthogonal signal correction of near-infrared spectra. *Chemometrics Intellig Lab Syst*. 1998;44: 175–185.
2. Bijlsma S, Louwerse DJ, Smilde AK. Rapid estimation of rate constants of batch processes using on-line SW-NIR. *AIChE J*. 1998;44: 2713–2723.

3. Westerhuis JA, Gurden SP, Smilde AK. Spectroscopic monitoring of batch reactions for on-line fault detection and diagnosis. *Anal Chem.* 2000;72:5322–5330.
4. Gurden SP, Westerhuis JA, Smilde AK. Monitoring of batch processes using spectroscopy. *AIChE J.* 2002;48:2283–2297.
5. Othman NS, Fevotte G, Peycelon D, Egraz JB, Suau JM. Control of polymer molecular weight using near infrared spectroscopy. *AIChE J.* 2004;50:654–664.
6. Gabrielsson J, Jonsson H, Trygg J, Airiau C, Schmidt B, Escott R. Combining process and spectroscopic data to improve batch modeling. *AIChE J.* 2006;52:3164–3172.
7. Reis MM, Araújo PHH, Sayer C, Giudici R. Spectroscopic on-line monitoring of reactions in dispersed medium: chemometric challenges. *Anal Chim Acta.* 2007;595:257–265.
8. Ye SF, Wang D, Min SG. Successive projections algorithm combined with uninformative variable elimination for spectra variable selection. *Chemometrics Intellig Lab Syst.* 2008;91:194–199.
9. Xu H, Liu Z, Cai W, Shao X. A wavelength selection method based on randomization test for near-infrared spectral analysis. *Chemometrics Intellig Lab Syst.* 2009;97:189–193.
10. Zhao C, Gao F, Wang F. Phase-based joint modeling and spectroscopy analysis for batch processes monitoring. *Ind Eng Chem Res.* 2010;49:669–681.
11. Sjöblom J, Svensson O, Josefson M, Kullberg H, Wold S. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemometrics Intellig Lab Syst.* 1998;44: 229–244.
12. Gusnanto A, Pawitan Y, Huang J, Lane B. Variable selection in random calibration of near-infrared instruments: ridge regression and partial least squares regression settings. *J Chemometrics*. 2003;17: 174–185.
13. Trygg J. Prediction and spectral profile estimation in multivariate calibration. *J Chemometrics*. 2004;18:166–172.
14. Andrew A, Fearn T. Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation. *Chemometrics Intellig Lab Syst.* 2004;72:51–56.
15. Chen T, Morris J, Martin E. Gaussian process regression for multivariate spectroscopic calibration. *Chemometrics Intellig Lab Syst.* 2007;87:59–67.
16. Preys S, Roger JM, BoUlet JC. Robust calibration using orthogonal projection and experimental design. Application to the correction of the light scattering effect on turbid NIR spectra. *Chemometrics Intellig Lab Syst*. 2008;91:28–33.
17. Benoudjit N, Melgani F, Bouzgou H. Multiple regression systems for spectrophotometric data analysis. *Chemometrics Intellig Lab Syst.* 2009;95:144–149.
18. Alciaturi CE, Quevedo G. Bayesian regularization: application to calibration in NIR spectroscopy. *J Chemom.* 2009;23:562–568.
19. Chen T, Martin E. Bayesian linear regression and variable selection for spectroscopic calibration. *Anal Chim Acta.* 2009;631:13–21.
20. Geladi P, Kowalski BR. Partial least-squares regression-a tutorial. *Anal Chim Acta.* 1986;185:1–17.
21. Brereton RG. Introduction to multivariate calibration in analytical chemistry. *Analyst.* 2000;125:2125–2154.
22. Kleinbaum DG, Kleinbaum DG. *Applied Regression Analysis and Other Multivariable Methods*, 4th ed. Australia: Thomson Brooks/ Cole, 2008.
23. Kutner MH, Nachtsheim C, Neter J. *Applied Linear Regression Models*, 4th ed. Boston: McGraw-Hill/Irwin, 2004.
24. Ergon R. Reduced PCR/PLSR models by subspace projections. *Chemometrics Intellig Lab Syst.* 2006;81:68–73.
25. Chen J, Wang XZ. A new approach to near-infrared spectra data analysis using independent component analysis. *J Chem Inf Comput Sci.* 2001;41:992–1001.
26. Shao XG, Wang W, Hou ZY, Cai WS. A new regression method based on independent component analysis. *Talanta.* 2006;69:676–680.
27. Navea S, Tauler R, Juan AD. Monitoring and modeling of protein processes using mass spectrometry, circular dichroism, and multivariate curve resolution methods. *Anal Chem.* 2006;78:4768–4778.
28. Zhao C, Gao F, Yao Y, Wang F. A robust calibration modeling strategy for analysis of interference-subject spectra data. *AIChE J.* 2010;56:196–206.
29. Hyvarinen A, Oja E. Independent component analysis: Algorithms and applications. *Neural Networks.* 2000;13:411–430.
30. Macgregor JF, Jaeckle C, Kiparissides C, Koutoudi M. Process monitoring and diagnosis by multiblock PLS methods. *AIChE J.* 1994; 40:826–838.
31. Kourti T, Nomikos P, Macgregor JF. Analysis, monitoring and fault-diagnosis of batch processes using multiblock and multiway PLS. *J Process Control.* 1995;5:277–284.
32. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *J Chemom.* 1998;12:301–321.
33. Qin SJ, Valle S, Piovoso MJ. On unifying multiblock analysis with application to decentralized process monitoring. *J Chemom.* 2001; 15:715–742.
34. Lopes JA, Menezes JC, Westerhuis JA, Smilde AK. Multiblock PLS analysis of an industrial pharmaceutical process. *Biotechnol Bioeng.* 2002;80:419–427.
35. Camacho J, Pico J. Online monitoring of batch processes using multi-phase principal component analysis. *J Process Control*. 2006; 16:1021–1035.
36. Lee J, Qin SJ, Lee I. Fault detection and diagnosis based on modified independent component analysis. *AIChE J.* 2006;52:3501–3514.
37. Horst R, Tuy H. *Global optimization*, 3rd ed. Berlin: Springer-Verlag, 1996.
38. Fogel LJ, Owens AJ, Walsh MJ. *Artificial Intelligence through Simulated Evolution*. New York: John Wiley, 1966.
39. Azam SM, Rehman M, Bhatti AK, Daudpota N. Parallel branch and bound model using logarithmic sampling (PBLS) for symmetric traveling salesman problem. *World Acad Sci Eng Technol.* 2005;6:66–69.
40. Yu HL, MacGregor JF. Post processing methods (PLS-CCA): simple alternatives to preprocessing methods (OSC-PLS). *Chemometrics Intellig Lab Syst.* 2004;73:199–205.
41. Anderson TW. *An Introduction to Multivariate Statistical Analysis*, 2nd ed. New York: Wiley, 1984.
42. Burnham AJ, Viveros R, MacGregor JF. Frameworks for latent variable multivariate regression. *J Chemom.* 1996;10:31–45.
43. Qin SJ, Valle S, Piovoso MJ. On unifying multiblock analysis with application to decentralized process monitoring. *J Chemom.* 2001; 15:715–742.
44. Westerhuis JA, Smilde AK. Deflation in multiblock PLS. *J Chemom.* 2001;15:485–493.
45. Dayal BS, MacGregor JF. Improved PLS Algorithms. *J Chemom.* 1997;11:73–85.

Appendix

MBPLS-CCA Algorithm

Input data $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_C]$ and $\mathbf{Y}$

Step 1,

Perform regular PLS-CCA on $\mathbf{X}$ and $\mathbf{Y}$ to obtain a pair of super scores $\mathbf{t}$ and $\mathbf{u}$, as well as the weights $\mathbf{w}$ and $\mathbf{r}$ and loadings $\mathbf{p}$ and $\mathbf{q}$.

Step 2,

$\mathbf{r}_b = \mathbf{r}(b)/\|\mathbf{r}(b)\|^2$ where $\mathbf{r}(b)$ is separated from the super weight vector $\mathbf{r}$ that corresponds to block $\mathbf{X}_b$.

$\mathbf{t}_b = \mathbf{X}_b \mathbf{r}_b$

Step 3,

Deflation:

$$\mathbf{q} = \mathbf{Y}_b^{\mathrm{T}} \mathbf{t}_b / \mathbf{t}_b^{\mathrm{T}} \mathbf{t}_b$$

$$\mathbf{F} = \mathbf{Y} - \mathbf{t}\mathbf{q}^{\mathrm{T}}$$

For additional LVs, set $\mathbf{Y} = \mathbf{F}$ and go back to step 1.

Output results:

Super LVs $\mathbf{T}$, super weights $\mathbf{R}$, super loadings $\mathbf{P}$ for descriptor description and super loadings $\mathbf{Q}$ for quality prediction.

Block LVs, $\mathbf{T}_b$, block weights $\mathbf{R}_b$, and block loadings $\mathbf{Q}_b$ for quality prediction. Moreover, two different block loadings $\mathbf{P}_b$ and $\mathbf{P}^t_b$ for block information description can also be calculated by $\mathbf{P}_b = \mathbf{X}_b^{\mathrm{T}}\mathbf{T}_b\ (\mathbf{T}_b^{\mathrm{T}}\mathbf{T}_b)^{-1}$ and $\mathbf{P}^t_b = \mathbf{X}_b^{\mathrm{T}}\mathbf{T}\ (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}$ based on block LVs and super LVs respectively.

Here some important points should be noted:

(a) As analyzed and demonstrated by Qin et al.,[43] the super LVs are equal to the LVs obtained by keeping all blocks together and performing standard regression modeling. In the present MBPLS-CCA algorithm, this conclusion is inherited as shown in Step 1.

(b) $\mathbf{T}_b$ reveal the LV information directly used to construct super LVs ($\mathbf{T}$), which, however, may not be guaranteed to be all informative for quality interpretation. Since the super LVs $\mathbf{T}$ are the sum of $\mathbf{T}_b$, some part in $\mathbf{T}_b$ will be hidden more or less by the other block scores and thus will not exist in $\mathbf{T}$ for quality prediction. Based on block LVs ($\mathbf{T}_b$) and super LVs ($\mathbf{T}$), correspondingly, modeled variations by $\mathbf{T}_b\mathbf{P}_b^{\mathrm{T}}$ and $\mathbf{T}\mathbf{P}_b^{t\mathrm{T}}$ may also reveal different block information respectively. This will be further illustrated in Simulation section.

(c) As analyzed and discussed by Westerhuis and Smilde,[44] some deflation problems exist in multiblock regression method. Since the super LVs ($\mathbf{T}$) summarize the information contained in all blocks, deflation of $\mathbf{X}_b$ using $\mathbf{T}$ gives the same predictions as standard regression model with all variables in one large $\mathbf{X}$-block, but the information of the separate blocks gets mixed up. This leads to block LVs ($\mathbf{T}_b$) that describe not only the information from their specific block but also information from other blocks, resulting in interpretation problems. However, deflation of $\mathbf{X}_b$ using $\mathbf{T}_b$ leads to inferior prediction of $\mathbf{Y}$, since some variation in $\mathbf{t}_b$ is not used for prediction of $\mathbf{Y}$ but is removed from $\mathbf{X}_b$. This information cannot be attained from block $\mathbf{X}_b$ for future components. If the direction is not available in the other blocks either, then this may lead to inferior quality prediction. It has been recommended that if only $\mathbf{Y}$ is deflated using the super scores instead and $\mathbf{X}_b$ is used repeatedly, these problems disappear. No information from one block goes into another block, as was the case with the super LV deflation of $\mathbf{X}_b$ and predictions remain equal to the standard regression model with no blocking. This is based on the theoretical analysis by Dayal and MacGregor[45] that for standard PLS model, instead of deflating both $\mathbf{X}$ and $\mathbf{Y}$, it is possible to only deflate $\mathbf{Y}$. This can be easily extended to the present MBPLS-CCA method as shown in Step 3. Deflation of $\mathbf{Y}$ using super LVs ($\mathbf{T}$) makes $\mathbf{T}$ orthogonal with each other, but the block LVs ($\mathbf{T}_b$) are not guaranteed to be orthogonal. $\mathbf{T}_b$ can give information of the specific block on the relation with $\mathbf{Y}$ in the presence of the other blocks, which makes interpretation much easier.

Manuscript received Mar. 23, 2010, and revision received May 24, 2010.