Spectra Calibration Modeling and Statistical Analysis for Cumulative Quality Interpretation and Prediction

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

DOI 10.1002/aic.12592
Published online March 31, 2011 in Wiley Online Library (wileyonlinelibrary.com).

An improved calibration modeling and statistical analysis algorithm is proposed for spectra quality interpretation and prediction is presented here. In a previous work, the frequency-band varying characteristics of the underlying spectra over the entire wavelength were treated by separately analyzing the spectra in each sub-band. Following that, the current major task lies in how to further comprehend and model the cumulative effects of different sub-bands on qualities from the inter-sub-band viewpoint. It reveals that one part of the underlying variation in each sub-band stays invariable over sub-bands, whereas the other part changes with the alternation of sub-bands. The original variation in each sub-band can thus be separated into two different parts, the common and specific ones. They reveal sub-band-similar and dissimilar contributions on quality interpretation, respectively, which are referred to ''repetitive'' and ''complementary'' cumulative effects in this approach. Correspondingly, different calibration modeling and analyses are performed to explore their respective and joint roles in quality interpretation. The feasibility of the proposed calibration analysis algorithm is verified through both simple numerical data and real spectra data. © 2011 American Institute of Chemical Engineers *AIChE J,* 58: 466–479, 2012

Keywords: independent component analysis, sub-band separation, repetitive and complementary cumulative effects, quality interpretation and prediction, sub-band-common and specific variations

Introduction

During the past decades, the use of spectroscopic information[1–20] 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, a calibration model is often constructed to form the quantitative prediction relationship between the mixture spectra and the reference concentration of constituents.[11–20] Besides the common multivariate calibration methods[21–25] used for spectra, independent component analysis (ICA)[26] has been developed considering that the mixture spectra are actually often a linear combination (with coefficients corresponding to the proportions) of the spectra of its constituent species.[11,27–30] It allows for the decomposition of concentration values and pure spectra profiles of the different species from spectral measurement. Previous work[11,27,28,30] have reported the effectiveness of ICA in recovering the

Correspondence concerning this article should be addressed to F. Gao at kefgao@ust.hk.

constituent species of interest as well as determining their effects on the observed mixture spectra. This is called blind source signal separation process.

Independent component regression (ICR) method was first proposed by Chen and Wang[27] to the near-infrared (NIR) spectra data analysis. The authors have pointed out that by comparing the spectra of separated independent components (ICs) with the spectra library of pure substances, it was possible to identify unknown constituent species existing in the mixture. Ideally, if the separated ICs exactly matched the pure substances constituting the mixture, then mixing matrix would agree well with the concentration of the substances existing in mixture. Considering they could not match very well in practice, regression analysis could be conducted between the estimated mixing matrix and the real concentration measurements based on simple least-squares algorithm. Shao et al.[28] further reported besides the completely equivalent quantitative prediction performance compared with principal component regression (PCR), ICs could give more chemical explanations than principal components, which were found to be strongly correlated to the NIR spectra of source components in spectra.

Zhao and Gao[1] reported in their previous research work that 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. Certain source species may be important in some spectra regions and thus can be accurately decomposed by ICA, whereas in other spectra regions, different dominant source species are found and will be estimated better. Single or unified ICA decomposition over the full-range spectra region can not identify all source species sufficiently enough and desirably track the varying chemical characteristics along the wavelength direction. To solve this problem, Zhao and Gao[1] proposed the sub-band separation strategy, in which multiple sub-bands are identified from the whole wavelength region. Without losing generality, over different sub-bands, different ICA decomposition relationships were figured out. Then the most quality-related information from different sub-bands was integrated for quality prediction. Improved prediction performance was reported compared with unified ICR model since source species and the corresponding mixing coefficients could be better decomposed from the mixture spectra after the sub-band separation. On the other hand, these separated multiple sub-bands prepare a good quality-related statistical analysis platform. It is statistically meaningful if one can clearly know how the underlying variation information develops, how they correlate with each other and especially how they influence the concerned qualities. By exploring the predictor information effectively in advance, it can help to enhance one's process understanding and design more efficient calibration analysis strategy for improved quality interpretation. This is a critical and meaningful issue, deserving special attention.

In this work, a spectra calibration modeling and statistical analysis strategy is designed for quality interpretation based on sub-band separation, which focus on analyzing the cumulative effects on qualities over different sub-bands. The process is assumed that different mixing coefficients can be decomposed by ICA estimation in each band, which is directly related with the concentrations. Based on the band-dependent mixing models, the idea of variation division is first presented to distinguish the common predictor information from specific information over sub-bands. It is found that in spite of the inter-sub-band dissimilarity, some of the underlying variations are similar and consistent, which can be well approximated by defining some common-to-sub-band latent variables (LVs). The other underlying information is specific to sub-band, which changes from one sub-band to another and thus cannot be comprehensively matched by a uniform LV structure. It is easy to understand that the sub-band-common LVs will describe the similar quality variations, imposing ''repetitive'' cumulative effects with respect to quality interpretation; whereas the sub-band-specific predictor information will explain different parts of quality variations, presenting ''complementary'' cumulative effects. A more comprehensive predictor information exploration is achieved prior to regression modeling, where common and specific explanatory variations are separated from each other and thus given different calibration analyses. Their different cumulative effects are clearly revealed and finally wisely combined for more effective quality interpretation.

This article is organized as follows. In the next section, the cumulative calibration analysis strategy is described in detail in a three-step modeling fashion. Its underlying principle is explained and knowledge of some related modeling methods is also introduced. The applications to both numerical case and the real spectra case demonstrate the feasibility of the proposed method. Discussion is conducted based on the results, highlighting the suitability of the proposed method and its advantages for spectra calibration analysis and understanding. Finally, conclusions are drawn in the last section.

Methodology

As analyzed before, underlying mixture spectra, there are different dominative source spectra over different wavelength regions. One spectra wavelength region should be further divided into several different sub-bands if the inherent spectra information changes. From the inter-sub-band viewpoint, their contributions to qualities share both similarity and dissimilarity. In the proposed calibration analysis strategy, blocking of spectra variables (i.e., sub-band separation along wavelength direction) and partition of underlying variations (i.e., separation of common and specific variations in each sub-band) are combined to check the spectra information more comprehensively for cumulative quality interpretation. It is implemented in three consecutive steps. First, multiple spectra sub-bands are identified from the original wavelength space. In the second step, sub-band-common and specific variations are distinguished in each sub-band. In the third step, based on the above variation division, different regression models are designed respectively to reveal their different cumulative effects on quality interpretation. The modeling procedure is described in the following subsections.

Spectra sub-band separation

As mentioned before, the underlying sources are not necessarily persistently dominative throughout the whole

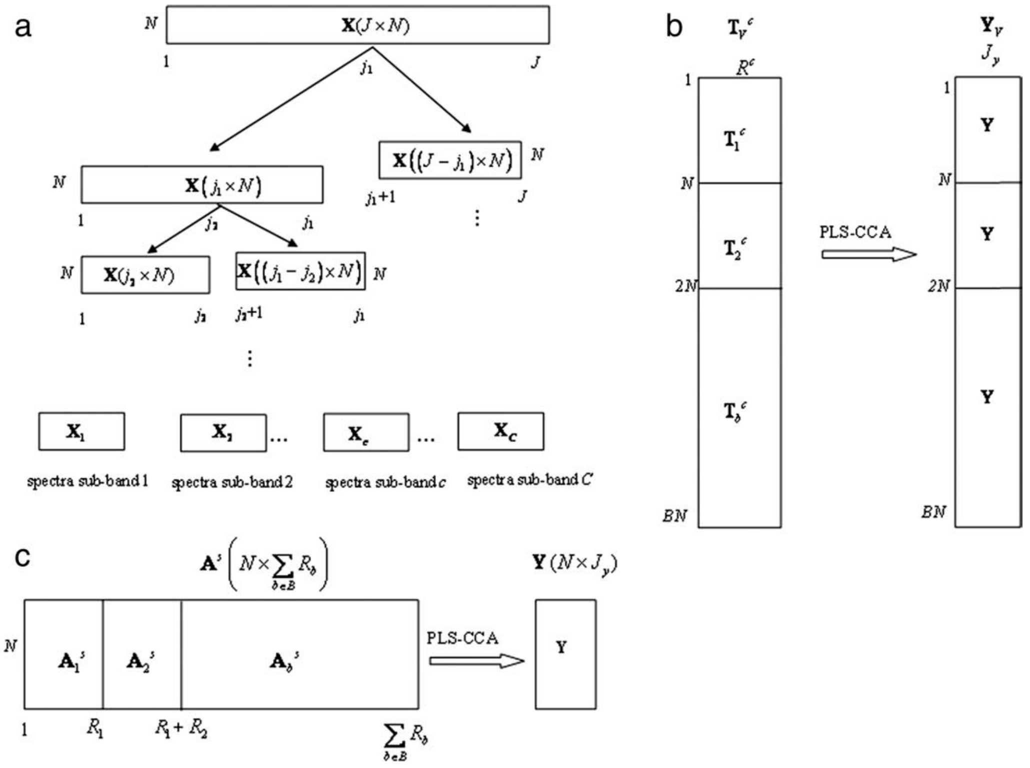

Figure 1. (a) Spectra sub-band separation scheme, (b) regression modeling in common part, and (c) regression modeling in specific part.

wavelength region. The dominant sources in some wavelength regions may be less important in the other regions. Therefore, if the sources are decomposed by unified ICA estimation focusing on the entire wavelength region, they actually reconstruct the mixture spectra from the viewpoint of global optimization. Sources may not be decomposed well enough, resulting in less accurate mixing coefficients. Considering the changes of underlying dominant sources in spectra profile along wavelength, it is necessary to divide a spectral wavelength region into several meaningful sub-bands. Different sub-bands can be identified by evaluating the resulting IC extraction performance and model representability. In our previous work,[1] an automatic sub-band separation algorithm has been designed which is used here for identification of multiple sub-bands. Its goal is to find the segments along wavelength direction which can be well approximated by one ICA model with sufficient representability. The detail can be found in our previous work.[1]

By the separation algorithm, B different spectral sub-bands are obtained as shown in Figure 1a. Let $\{\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_b, \ldots, \mathbf{X}_B\}$ be the corresponding $(J_b \times N)$-matrices of the separated spectra sub-bands (where subscript b is used to represent one specific sub-band and J_b is the number of wavelength variables in each sub-band). They have the same N samples scanned at different spectral regions, which all point to the same concentration matrix ($\mathbf{Y}$). From a mathematical point of view, by performing ICA decomposition, their underlying source spectra profiles are figured out, respectively, in different spectra sub-bands:

$$\begin{aligned}\hat{\mathbf{X}}_1 &= \mathbf{S}_1\mathbf{A}_1\\ \hat{\mathbf{X}}_2 &= \mathbf{S}_2\mathbf{A}_2\\ &\vdots\\ \hat{\mathbf{X}}_b &= \mathbf{S}_b\mathbf{A}_b\\ &\vdots\\ \hat{\mathbf{X}}_B &= \mathbf{S}_B\mathbf{A}_B\end{aligned} \tag{1}$$

where different source spectra ($\mathbf{S}_b$ ($J_b \times R_b$) and multiple mixing relationships ($\mathbf{A}_b$ ($R_b \times N$)) are decomposed (where R_b is the number of retained ICs, which is related with the number of active/dominant sources in the mix spectra in each sub-band and may be different over different spectra sub-bands). $\mathbf{A}_b$ reveals the different contributions of source spectra profiles to mixture spectra in different sub-bands.

Based on the designed sub-band separation strategy, multiple spectra sub-bands are obtained, enclosing different spectra wavelengths. It provides different analysis platforms for ICA-based source signal extraction and improves the ICA model representability. From the local analysis viewpoint, one can also obtain more details in different spectra sub-bands. Moreover, it is easy to check how they will influence qualities under the supervision of each other by performing the corresponding statistical analysis.

Common and specific variation division

First, the simple theoretical analysis and support with respect to variation division for cumulative quality analysis is given as below. By sub-band separation, multiple data sets are obtained corresponding to different sub-bands. Over different sub-bands, different ICA decomposition results generally result in different mixing relationships and thus are related to different quality interpretability. One quality index (concentration) may be better explained by the mixing coefficient in one sub-band if the uncovered IC can better match the corresponding real source. From another viewpoint, this provides a quality-related multiple-source data analysis platform. One analysis way is to treat each predictor block independently, which, however, isolates the role of each sub-band and overlooks the inter-sub-band interaction. Multiway partial least squares (MPLS)[31] algorithm can analyze them from a global viewpoint but loses the local information. Multiblock[32–35] modeling idea is expected to obtain both local information (block scores) and global information (super scores) simultaneously from the data. Further, Reinikainen and Hoskuldsson[36] reported a priority PLS regression analysis method and its successful application to a multistep industrial process, which gave multiple phases descending priority following operation time sequence. Instead of treating all the blocks in a parallel mode, it modeled the separate predictor blocks serially, where the variations in quality that were not modeled by the previous blocks would be left to be explained by the following blocks. However, from a more general viewpoint, the significance of multiple phases does not always agree with the time sequence of operation process. Here, for spectra analysis, over multiple sub-bands, there is similarity to a certain extent among their variations, which is deemed to be driven by some common LVs here. It is not difficult to understand that these common LVs will explain the same quality variability repetitively. It means no new quality variations can be additionally interpreted sub-band after sub-band if all the underlying LVs are common over sub-bands. Therefore, their cumulative effects are actually pseudo, called repetitive cumulation here. Moreover, besides the similarity, these sub-bands also present dissimilarity, which reveals their specific variations driven by different LVs. It is easy to understand that these different LVs will explain different quality variations complementarily. It means that new quality information can be explored step by step if all the underlying LVs are different over sub-bands. Such cumulative effects are called complementary cumulation here. Clearly, the two types of cumulations have different characteristics and different influential relationships with qualities. From the inter-sub-band viewpoint, each sub-band covers both the above variations. Instead of adopting global modeling or multiple isolated models, it is necessary to distinguish the variations over different sub-bands and thus give them a detailed analysis for cumulative quality interpretation.

Motivated by such cognition, a two-step multiset analysis algorithm, which was proposed in our previous work (Zhao et al., submitted), can be modified here to relate the inherent variation information of multiple data spaces. Different from the original algorithm (Zhao et al., submitted), which focused on extracting the underlying correlations over multiset data, this study pays attention to their underlying variations. For easier readability, a brief description of the two-step LV extraction algorithm is given in Appendix. The variation division is directly based on ICA decomposition result, which is much easier than that based on the other methods (such as PLS) which focus on the lengthy wavelength profiles instead of the estimated mixing relationships. By this method, each original spectral sub-band can be separated into two different parts with different variation information enclosed. One is called the common part, which is driven by the common variation LVs, revealing inter-sub-band similarity. The other is called the uncommon part, which is supported by those sub-band-specific LVs, revealing inter-sub-band dissimilarity.

All the mixing matrices obtained from different spectra sub-bands can be organized as multiple data blocks: $\left\{\mathbf{A}_1^{\mathrm{T}}, \mathbf{A}_2^{\mathrm{T}}, \ldots, \mathbf{A}_b^{\mathrm{T}}, \ldots, \mathbf{A}_B^{\mathrm{T}}\right\}$. Input them to the two-step LV extraction algorithm shown in Appendix. For each sub-band, $\mathbf{A}_b^{\mathrm{T}}$ $(N \times R_b)$ $(b = 1,2,\ldots B)$, which may have different variables, the common LVs are obtained: $\mathbf{T}_g$ $(N \times R^{\mathrm{c}})$, where R^{c} is the retained number. Then, each original data space $(\mathbf{A}_b)$ is separated into two different parts $(\mathbf{A}_b^{\mathrm{c}}$ and $\mathbf{A}_b^{\mathrm{s}})$, one explained by the common LVs in linear combination and the other uninterpretable, which enclose the cross-sub-band similar and dissimilar underlying variation information, respectively:

$$\begin{aligned}
\mathbf{A}_b^{\mathrm{T}} &= \mathbf{A}_b^{\mathrm{cT}} + \mathbf{A}_b^{\mathrm{sT}} \\
\mathbf{A}_b^{\mathrm{cT}} &= \mathbf{T}_b^{\mathrm{c}}\mathbf{P}_b^{\mathrm{cT}} = \mathbf{T}_b^{\mathrm{c}}\left(\mathbf{T}_b^{\mathrm{cT}}\mathbf{T}_b^{\mathrm{c}}\right)^{-1}\mathbf{T}_b^{\mathrm{cT}}\mathbf{A}_b^{\mathrm{T}} \\
\mathbf{A}_b^{\mathrm{sT}} &= \mathbf{A}_b^{\mathrm{T}} - \mathbf{A}_b^{\mathrm{cT}} = \left(\mathbf{I} - \mathbf{T}_b^{\mathrm{c}}\left(\mathbf{T}_b^{\mathrm{cT}}\mathbf{T}_b^{\mathrm{c}}\right)^{-1}\mathbf{T}_b^{\mathrm{cT}}\right)\mathbf{A}_b^{\mathrm{T}}
\end{aligned} \tag{2}$$

where predictor loadings $\mathbf{P}_b^{\mathrm{cT}} = \left(\mathbf{T}_b^{\mathrm{cT}}\mathbf{T}_b^{\mathrm{c}}\right)^{-1}\mathbf{T}_b^{\mathrm{cT}}\mathbf{A}_b^{\mathrm{T}}$ are also the linear combination coefficients corresponding to the common LVs. Actually, $\mathbf{G}_{\mathbf{T}_b^{\mathrm{c}}} = \mathbf{T}_b^{\mathrm{c}}\left(\mathbf{T}_b^{\mathrm{cT}}\mathbf{T}_b^{\mathrm{c}}\right)^{-1}\mathbf{T}_b^{\mathrm{cT}}$ is the orthogonal projector onto the column space of $\mathbf{T}_b^{\mathrm{c}}$, and $\mathbf{H}_{\mathbf{T}_b^{\mathrm{c}}} = \mathbf{I} - \mathbf{G}_{\mathbf{T}_b^{\mathrm{c}}} = \mathbf{I} - \mathbf{T}_b^{\mathrm{c}}\left(\mathbf{T}_b^{\mathrm{cT}}\mathbf{T}_b^{\mathrm{c}}\right)^{-1}\mathbf{T}_b^{\mathrm{cT}}$ is the anti-projector with respect to the column space of $\mathbf{T}_b^{\mathrm{c}}$. Therefore, from another viewpoint, the two sub-bands can also be regarded as the ones obtained by projecting $\mathbf{A}_b$ onto different projectors, $\mathbf{A}_b\mathbf{G}_{\mathbf{T}_b^{\mathrm{c}}}$ and $\mathbf{A}_b\mathbf{H}_{\mathbf{T}_b^{\mathrm{c}}}$. It is clear the two sub-bands are orthogonal with each other as $\mathbf{A}_b^{\mathrm{c}}\left(\mathbf{A}_b^{\mathrm{s}}\right)^{\mathrm{T}} = \mathbf{A}_b\mathbf{G}_{\mathbf{T}_b^{\mathrm{c}}}\left(\mathbf{A}_b\mathbf{H}_{\mathbf{T}_b^{\mathrm{c}}}\right)^{\mathrm{T}} = \mathbf{0}$. Here, it should be noted that $\mathbf{A}_b^{\mathrm{s}}$ cover both sub-band-specific systematic variations and residuals, such as measurement errors and noises, which will be further decomposed in the following regression modeling.

Regression modeling for cumulative analysis

Based on the separation of different sub-bands as well as common and specific variations, different regression modeling and statistical analysis strategies can then be designed to reveal their different cumulative effects as shown in Figures 1b, c. In the sub-band-common part, the predictor variables enclose the similar underlying variation information as all of them are the linear combinations of one subset of common LVs and thus reveal similar relationships with qualities. A uniform regression model common to all sub-bands can thus be uncovered. In the left sub-band-specific part, the predictor variations are different and specific to each sub-band, revealing their dissimilarity in quality prediction. Therefore, different regression model structures should be designed in each sub-band.

In the common part, as shown in Figure 1b, the similar LVs $(\mathbf{T}_b^{\mathrm{c}}$ $(N \times R^{\mathrm{c}}))$ over different sub-bands are arranged in variable-unfolding way, forming $\mathbf{T}_v^{\mathrm{c}}$ $(BN \times R^{\mathrm{c}})$. The quality variables are also duplicated correspondingly, forming $\mathbf{Y}_v$ ($BN \times J_y$), which is arranged considering the similar covarying relationship between $\mathbf{T}_b^{\mathrm{c}}$ ($N \times R^{\mathrm{c}}$) and qualities ($\mathbf{Y}$). A unified regression model can be extracted by PLS-canonical correlation analysis (CCA) algorithm[37]:

$$\begin{aligned}
&\mathbf{T}^{\mathrm{c}} = \mathbf{T}_V^{\mathrm{c}}\mathbf{R}^{\mathrm{c}}\\
&\mathbf{P}^{\mathrm{cT}} = \left(\mathbf{T}^{\mathrm{cT}}\mathbf{T}^{\mathrm{c}}\right)^{-1}\mathbf{T}^{\mathrm{cT}}\mathbf{T}_V^{\mathrm{c}} = \Lambda^{\mathrm{c}-1}\mathbf{T}^{\mathrm{cT}}\mathbf{T}_V^{\mathrm{c}}\\
&\mathbf{Q}^{\mathrm{cT}} = \left(\mathbf{T}^{\mathrm{cT}}\mathbf{T}^{\mathrm{c}}\right)^{-1}\mathbf{T}^{\mathrm{cT}}\mathbf{Y}_{VB} = \Lambda^{\mathrm{c}-1}\mathbf{T}^{\mathrm{cT}}\mathbf{Y}_V\\
&\hat{\mathbf{T}}_b^{\mathrm{c}} = \mathbf{T}_b^{\mathrm{c}}\mathbf{P}^{\mathrm{cT}}\\
&\hat{\mathbf{Y}}_{V,b} = \mathbf{T}_b^{\mathrm{c}}\mathbf{Q}^{\mathrm{cT}}
\end{aligned} \tag{3}$$

where, the unified PLS-CCA weights model, $\mathbf{R}^{\mathrm{c}}$ ($R^{\mathrm{c}} \times A^{\mathrm{c}}$) is intrinsically doubly controlled by PLS and CCA weights. $\mathbf{T}^{\mathrm{c}}$ are the extracted PLS-CCA LVs used for quality prediction, from which, the common PLS-CCA LVs for each sub-band can be readily separated, $\mathbf{T}^{b,\mathrm{c}}$ ($N \times A^{\mathrm{c}}$). Λ^{c} is a diagonal matrix with equal element as the LVs $\mathbf{T}^{\mathrm{c}}$ have a unity variance resulting from CCA algorithm. $\mathbf{P}^{\mathrm{c}}$ and $\mathbf{Q}^{\mathrm{c}}$ are weights for predictors and qualities, respectively. The uniform regression model can be regarded as an averaged version over all sub-bands. $\hat{\mathbf{T}}_b^{\mathrm{c}}$ and $\hat{\mathbf{Y}}_{V,b}$ are the modeled predictor and quality variations by the sub-band-common part. PLS-CCA[37] combines PLS and CCA,[38,39] where, as a postprocessing, CCA is implemented on PLS LVs. In this way, it avoids the rank-deficiency problem of single CCA algorithm, gets rid of the quality-uninformative variation in PLS LVs and thus directly maximizes the regression correspondence. Therefore, compared with $\mathbf{T}_b^{\mathrm{c}}$, $\hat{\mathbf{T}}_b^{\mathrm{c}}$ only cover close quality-related systematic variations by excluding those quality-uninformative ones.

Over sub-bands, the similar contributions of the common parts suggest similar quality predictions, which may vary slightly resulting from measurement noises and modeling errors. The average quality prediction of common parts over sub-bands is thus defined as the final prediction $\hat{\mathbf{Y}}^{\mathrm{c}} = \frac{1}{B}\sum_{b\in B}\hat{\mathbf{Y}}_{V,b}$.

In the specific part, the variations of quality can be explained step by step from one sub-band to another as they contribute differently to qualities. This allows quality interpretation in a complementary cumulative fashion. Therefore, their joint contributions to qualities can be explored by putting all specific LVs one by one, composing a joint predictor space, $\mathbf{A}^{\mathrm{s}}\left(N \times \sum_{b\in B} R_b\right) = \left[\mathbf{A}_1^{\mathrm{sT}}, \mathbf{A}_2^{\mathrm{sT}}, \ldots, \mathbf{A}_b^{\mathrm{sT}}, \ldots, \mathbf{A}_B^{\mathrm{sT}}\right]$. By performing PLS-CCA modeling,[37] the end-of-wavelength quality analysis results are obtained in the sub-band-specific part:

$$\begin{aligned}
&\mathbf{T}^{\mathrm{s}} = \mathbf{A}^{\mathrm{s}}\mathbf{R}^{\mathrm{s}}\\
&\mathbf{P}^{\mathrm{sT}} = \left(\mathbf{T}^{\mathrm{sT}}\mathbf{T}^{\mathrm{s}}\right)^{-1}\mathbf{T}^{\mathrm{sT}}\mathbf{A}^{\mathrm{s}} = \Lambda^{\mathrm{s}-1}\mathbf{T}^{\mathrm{sT}}\mathbf{A}^{\mathrm{s}}\\
&\mathbf{Q}^{\mathrm{sT}} = \left(\mathbf{T}^{\mathrm{sT}}\mathbf{T}^{\mathrm{s}}\right)^{-1}\mathbf{T}^{\mathrm{sT}}\mathbf{Y} = \Lambda^{\mathrm{s}-1}\mathbf{T}^{\mathrm{sT}}\mathbf{Y}\\
&\hat{\mathbf{A}}^{\mathrm{s}} = \mathbf{T}^{\mathrm{s}}\mathbf{P}^{\mathrm{sT}}\\
&\hat{\mathbf{Y}}^{\mathrm{s}} = \mathbf{T}^{\mathrm{s}}\mathbf{Q}^{\mathrm{sT}}\\
&\mathbf{E}^{\mathrm{s}} = \mathbf{A}^{\mathrm{s}} - \hat{\mathbf{A}}^{\mathrm{s}}
\end{aligned} \tag{4}$$

where the model parameters in the specific part are denoted similarly to those in the common part shown in Eq. 3. The specific predictor error $\mathbf{E}^{\mathrm{s}}$ may cover both quality-irrelevant systematic variations and the measurement errors. Further, the participated specific variations in each sub-band can be also calculated as $\hat{\mathbf{A}}_b^{\mathrm{sT}} = \mathbf{T}^{\mathrm{s}}\mathbf{P}_b^{\mathrm{sT}} = \mathbf{T}^{\mathrm{s}}\left(\mathbf{T}^{\mathrm{sT}}\mathbf{T}^{\mathrm{s}}\right)^{-1}\mathbf{T}^{\mathrm{sT}}\mathbf{A}_b^{\mathrm{sT}}$ (where $\mathbf{P}_b^{\mathrm{s}}$ is the sub-band-specific loadings, which can be directly split from $\mathbf{P}^{\mathrm{s}}$ corresponding to the *b*th sub-band).

In summary, the underlying variations underlying each sub-band are formulated as below:

$$\mathbf{A}_b^{\mathrm{T}} = \mathbf{A}_b^{\mathrm{cT}} + \mathbf{A}_b^{\mathrm{sT}} = \mathbf{A}_b^{\mathrm{cT}} + \hat{\mathbf{A}}_b^{\mathrm{sT}} + \mathbf{E}_b^{\mathrm{s}} = \mathbf{T}_b^{\mathrm{c}}\mathbf{P}_b^{\mathrm{cT}} + \mathbf{T}^{\mathrm{s}}\mathbf{P}_b^{\mathrm{sT}} + \mathbf{E}_b^{\mathrm{s}} \tag{5}$$

where $\mathbf{T}_b^{\mathrm{c}}$ can be readily calculated in each local sub-band, whereas $\mathbf{T}^{\mathrm{s}}$ can only be obtained by considering all sub-bands together. $\mathbf{T}_b^{\mathrm{c}}\mathbf{P}_b^{\mathrm{cT}}$ model the common variations as $\mathbf{T}^{\mathrm{s}}\mathbf{P}_b^{\mathrm{sT}}$ evaluate the specific variations in each sub-band for cumulative quality interpretation.

The final quality prediction can be readily obtained by combining the predicted qualities from two different parts over all sub-bands using weights derived from PLS-CCA algorithm:

$$\begin{aligned}
&\left[\hat{\mathbf{Y}}^{\mathrm{c}}, \hat{\mathbf{Y}}^{\mathrm{s}}\right] \xrightarrow{\text{PLS}-\text{CCA}} \mathbf{Y}\\
&\hat{\mathbf{Y}} = \left[\hat{\mathbf{Y}}^{\mathrm{c}}, \hat{\mathbf{Y}}^{\mathrm{s}}\right]\Theta
\end{aligned} \tag{6}$$

where $\Theta\left(2J_y \times J_y\right)$ is the regression coefficients corresponding to two different types of quality predictions. In this way, the different cumulative contributions of the common and specific parts are stacked.

The cumulative quality analysis procedure is simply summarized as follows:

1. By sub-band separation algorithm, B different spectral sub-bands are obtained;
2. common and specific variation are distinguished by two-step LV extraction algorithm;
3. regression modeling variable-unfolding in common part over sub-bands by arranging predictors as $\mathbf{T}_v^{\mathrm{c}}$ ($BN \times R^{\mathrm{c}}$);
4. regression modeling in specific part over sub-bands by arranging predictors as $\mathbf{A}^{\mathrm{s}}\left(N \times \sum_{b\in B} R_b\right)$; and
5. final quality prediction by combining the predicted qualities from both sub-band-common and specific parts.

Simulations and Discussions

Case study 1

The numerical example will simply illustrate the influence of separation of sub-bands on ICA decomposition result and the resulting mixing coefficients as well as the extraction of common variations over sub-bands. Three 200-wavelength sources ($\mathbf{s}_i$) are considered, whose profiles are shown in Figure 2a with the original mixing relationships (a_i) shown in Figure 2b. In this case, two sub-bands are artificially set from the variation profiles of sources. That is, Source 1 dominates throughout the wavelength region, whereas Source 2 only in Sub-band 1 (wavelength 1–100) with random variation in Sub-band 2 and Source 3 only in Sub-band 2

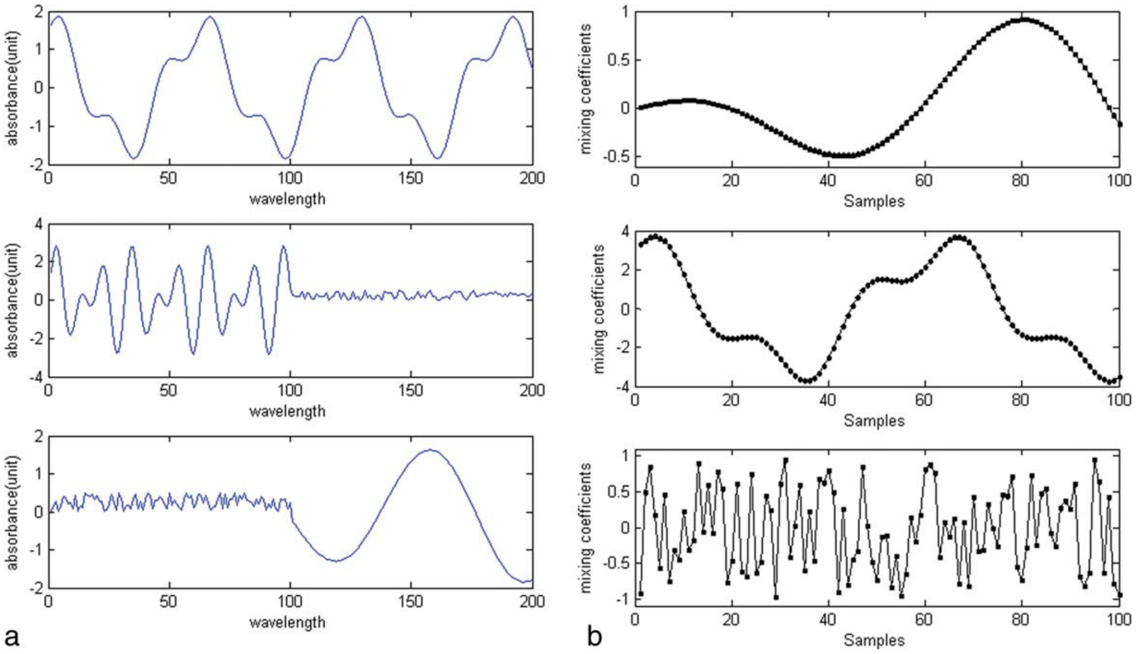

Figure 2. (a) Original source profile and (b) the mixing relationships for the composition of mixture spectra.

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

(wavelength 101–200). By linear combination, we can get the mixture spectra profile of each observation, $\mathbf{x} = \sum_{i=1}^{3} \mathbf{s}_i a_i$, where the real mixing coefficients (a_i) are quality indices. Moreover, each observation profile is added by normally distributed random noises which are $\sim$1% of the source variations. In all, 100 observations are generated and used for analysis, in which, 60 samples are used for model training, and the other 40 samples are used for validation.

For training data, in the two different sub-bands, respectively, the decomposed ICA mixing coefficients are shown in Figure 3a taking example for the first 3. Comparatively, without sub-band separation, the mixing coefficients by single ICA (SICA) decomposition over the entire wavelength region are shown in Figure 3b. For testing data, the decomposed mixing relationships are shown in Figure 3c by two-sub-band ICA estimation in comparison with that by SICA shown in Figure 3d. When compared with the real mixing coefficients shown in Figure 2b, we can see that SICA results maybe only better approximate one part of all the mixing coefficients. By two-sub-band ICA decomposition, the mixing coefficients resulting from different sub-bands may be more comprehensive for the interpretation of real mixing relationships (i.e., the qualities). For example, from training data, the third IC in Sub-band 2 may be more like the first real mixing coefficient although Sub-band 1 may not decompose it well. The decomposition results in two different sub-bands can then be combined with each other for the interpretation of qualities. Clearly, if SICA decomposition is performed on the entire wavelength region, although some of the real mixing coefficients can be decomposed well; however, there are still some ones which can not be uncovered, which thus loses the interpretability of the corresponding quality indices. For example, it is interesting to find that the first real mixing coefficient is not decomposed well enough by SICA for both training and testing data although it is regarded that the profile of the first source is persistently important throughout the entire wavelength region. Therefore, we can conclude that the ICA decomposition performance may be influenced by different factors so that ICA may not decompose the real sources accurately enough. By sub-band separation, it increases the chance for real mixing relationships to be estimated. More information may be provided as in different sub-bands different mixing coefficients can be well decomposed and thus complement each other for quality interpretation.

Based on the two-sub-band ICA decomposition result, the common variations in mixing relationships over two sub-bands are extracted as shown in Figure 4, which are more like the first mixing relationship shown in Figure 2b. Moreover, they match well with each other, revealing the same variations over the two sub-bands.

To sum up, in this simple numerical example, analysis results illustrate how sub-band separation influences the mixing relationships decomposed by ICA and the common variations over sub-bands can be figured out.

Case study 2

The used data set consists of spectra from 80 samples of corn with wavelength ranging 1100–2498 nm at 2 nm intervals (700 channels), which are scanned on m5 NIR spectrometer. Therefore, we can collect a 700 × 80 spectral matrix. 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

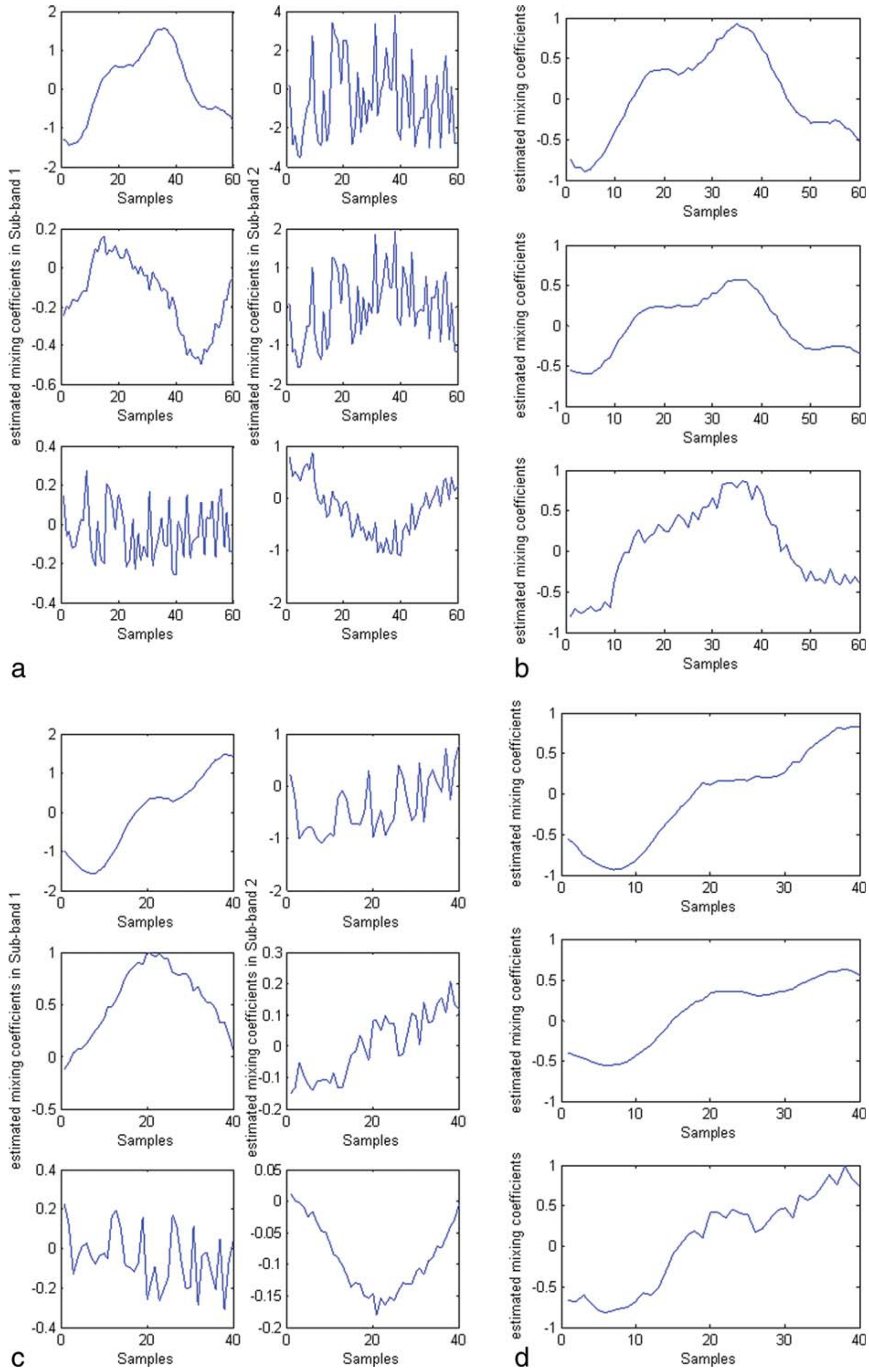

Figure 3. Mixing coefficient decomposition results for training data by (a) two-sub-band ICA and (b) SICA as well as for testing data by (c) two-sub-band ICA and (d) SICA.

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

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. The corn spectra data are available at the Eigenvector Research homepage: http://www.eigenvector.com/DATA/Corn.

Simulation Methodology and Results

First, the mixture spectra are shown in Figure 5 taking example for the training samples. Along wavelength direction, the spectral profiles show great fluctuation. As analyzed before, different basic chemical components dominate over

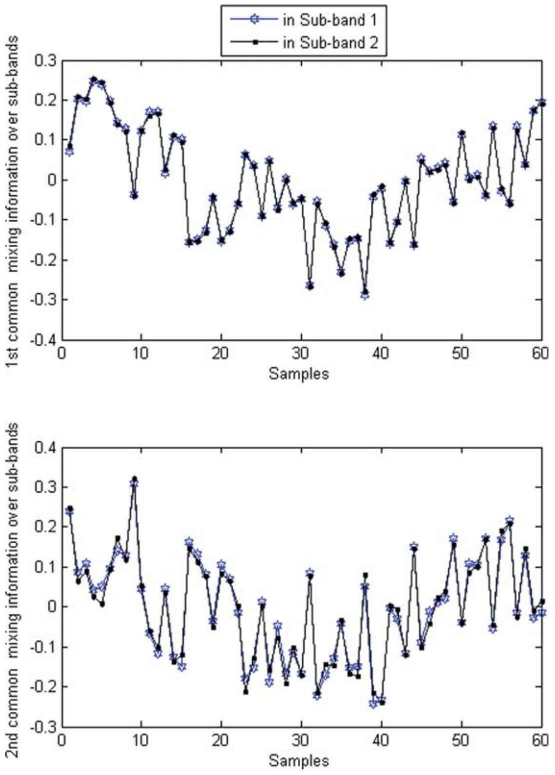

Figure 4. Common variations in mixing relationships over two sub-bands.

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

different spectra ranges. By performing the proposed sub-band separation, five different spectra sub-bands are obtained, covering different wavelengths, 1100–1364, 1366–1468, 1470–1894, 1896–2392, 2394–2498 nm, respectively, as shown in Figure 5. Over different sub-bands, the source spectra are extracted as well as their mixing coefficients. By cross-validation, different number of ICs is retained in each sub-band to recover the mixture spectra as shown in Table 1. The coefficient of determination (R^2), the commonly used statistical metric, is used here to describe how much of the variability in the data is captured by the prediction model. The modeled spectral variations are quantified by $R^2\hat{\mathbf{X}}_b = \frac{\sum_{b=1}^{B}\hat{\mathbf{x}}_b^2}{\sum_{b=1}^{B}\mathbf{x}_b^2} \times 100\%$, telling the model reconstruction competency in each sub-band. Taking example for the first 2 source components, their spectra profiles are illustrated in Figure 6a over five different sub-bands, respectively. They are quite different from those decomposed by SICA modeling results (where the entire wavelength band is considered together) shown in Figure 6b. For clear comparison, the ICs decomposed by SICA are also displayed separately in five sub-bands. Moreover, the mixing parameters corresponding to the different decomposed active ICs over sub-bands are shown in Figure 6c using the proposed method. It is clear that over different spectra wavelength ranges, the dominant source spectra may be different. Correspondingly, their contributions to the mixture spectra may be described with different accuracy. The separated sub-bands 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 sub-band.

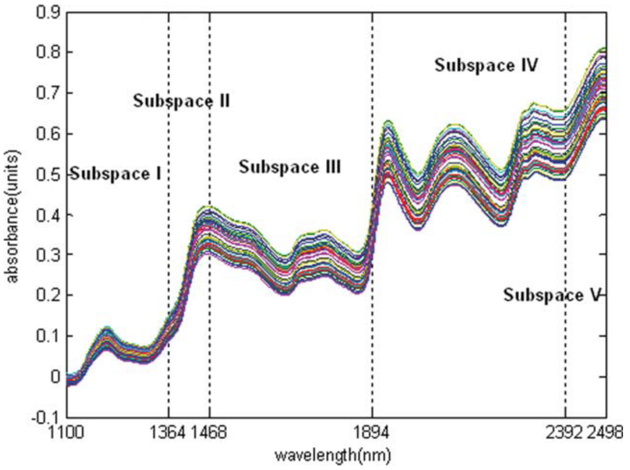

Figure 5. Mixture spectra trajectory and the sub-band separation result for corn data.

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

Multiple mixing matrices form a new predictor space, preparing a desirable statistical analysis platform to explore their cumulative effects on qualities. In our previous work,[1] the improvement of quality prediction performance using sub-band separation compared with SICR has been reported. In the current case illustration, the study of their different cumulative effects will be focused on. First, in each sub-band, variation division is performed to decompose the predictor information according to their different manners in cumulative quality interpretation. The common LVs are distinguished from those specific LVs, forming common and specific parts, respectively. The division result is summarized in Table 2, where only one common LV is figured out in each sub-band in this study. They can form a representative common LV vector over all sub-bands, $\mathbf{t}_v^c$ ($BN \times 1$) (where $B = 5$ and $N = 50$), corresponding to the rearranged concentration matrix $\mathbf{Y}_v$ ($BN \times J_R$). The contribution of common part to quality interpretation can be directly calculated by lease squares estimation as only one predictor variable is available. For the specific part, the variation information over all sub-bands can be collected, forming a new predictor space,

Table 1. ICA Decomposition Result over Different Spectra Sub-Bands

	Spectra Sub-Bands				
Statistics	I	II	III	IV	V
Model order	6	5	6	5	4
ICA reconstruction $R^2\hat{\mathbf{X}}_b$ (%)	88.58	98.80	99.52	99.86	99.91

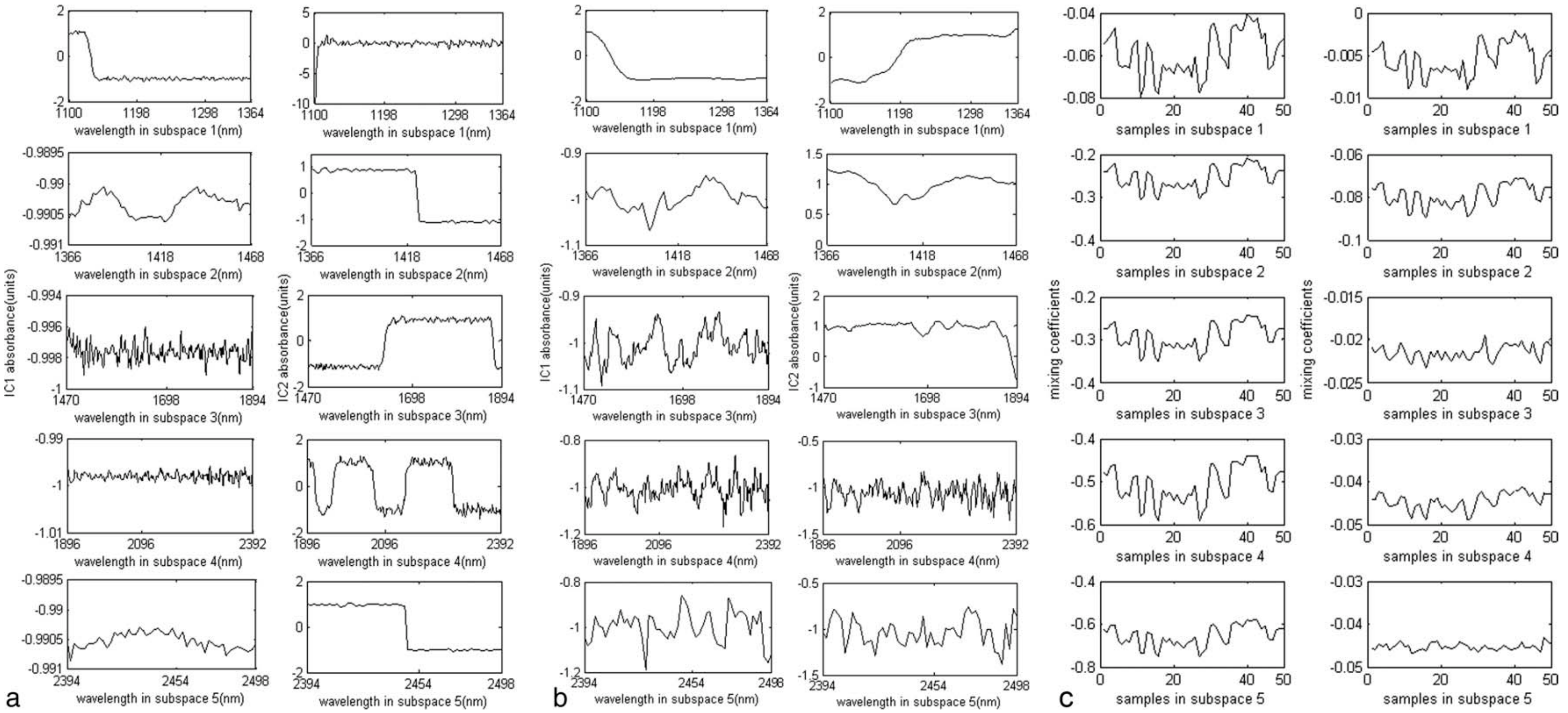

Figure 6. Spectra profile of the first 2 ICs over five spectra sub-bands: (a) Using the proposed method, (b) using single ICA decomposition, and (c) mixing coefficients for the first 2 ICs over five spectra sub-bands using the proposed method.

Table 2. Summary of Variation Division and Quality-Related Cumulative Analysis Result

Statistics		Spectra Sub-Bands				
		I	II	III	IV	V
	Common Part	1				
Model order	Specific Part	3	3	3	3	2
$R^2\hat{\mathbf{A}}_b^{\mathrm{c}}$ (%)		97.34	96.78	75.98	75.71	29.45
$R^2\hat{\mathbf{A}}_b^{\mathrm{s}}$ (%)		0.82	1.02	11.05	14.06	33.53
$R^2\hat{\mathbf{Y}}_b^{\mathrm{c}}$ (%)		57.76	59.11	58.79	59.45	60.05
		33.91	34.08	33.33	33.49	35.31
		18.92	18.52	18.77	19.39	20.02
		0.23	0.33	0.32	0.26	0.21
$R^2\hat{\mathbf{Y}}_b^{\mathrm{s}}$ (%)		0.04	19.54	4.30	0.73	0.04
		0.37	0.61	23.55	0.96	0.37
		41.08	2.85	42.62	64.65	41.08
		41.15	11.24	66.40	80.32	41.15
$R^2\hat{\mathbf{Y}}^{\mathrm{c}}$ (%)		58.93				
		33.96				
		19.09				
		0.27				
$R^2\hat{\mathbf{Y}}^{\mathrm{s}}$ (%)		22.08				
		25.73				
		73.42				
		87.68				
$R^2\hat{\mathbf{Y}}$ (%)		81.06				
		60.16				
		92.53				
		87.95				

$\mathbf{A}^{\mathrm{s}}\left(N \times \sum_{b=1}^{5} R_b\right) = \left[\mathbf{A}_1^{\mathrm{sT}}, \mathbf{A}_2^{\mathrm{sT}}, \mathbf{A}_3^{\mathrm{sT}}, \mathbf{A}_4^{\mathrm{sT}}, \mathbf{A}_5^{\mathrm{sT}}\right]$. Relating it with the concentration $\mathbf{Y}$ ($N \times J_y$) by PLS-CCA, the systematic variation information really closely related to qualities is extracted, which may only account for a portion of the original variations of the specific part. The residual may cover some systematic variations as well as measurement errors which are both quality-uninformative. In each sub-band, the participated common descriptor variations ($R^2\hat{\mathbf{A}}_b^{\mathrm{c}}$) for the explanation of all quality indices and predicted quality variations ($R^2\hat{\mathbf{Y}}_b^{\mathrm{c}}$) in common part are evaluated quantitatively in the similar way to $R^2\hat{\mathbf{X}}_b$ in Table 1. The results shown in Table 2 obviously indicate that in different sub-bands, the common predictor information is counted to different portions. Moreover, for different quality indices (four quality indices in all), the common part shows different interpretation competency. For example, in Sub-band V, the common predictor variations are counted least; the first quality index can be interpreted best, whereas the fourth quality variable is explained worst. Moreover, for each quality index, the $R^2\hat{\mathbf{Y}}_b^{\mathrm{c}}$ values stay similar over five sub-bands, revealing the similar contributions of common part to qualities although they cannot explain all quality variations. Their average quality interpretation level ($R^2\hat{\mathbf{Y}}^{\mathrm{c}}$) is thus figured out to evaluate the cumulative effects of the common part. Moreover, the predictor variation ($R^2\hat{\mathbf{A}}_b^{\mathrm{s}}$) for the explanation of all quality indices and quality variation ($R^2\hat{\mathbf{Y}}_b^{\mathrm{s}}$) in specific part are also quantified. Comparatively, $R^2\hat{\mathbf{A}}_b^{\mathrm{s}}$ are smaller than $R^2\hat{\mathbf{A}}_b^{\mathrm{c}}$ over all sub-bands except Sub-band V, telling that the common variations are generally more significant in each sub-band in this case. The $R^2\hat{\mathbf{Y}}_b^{\mathrm{s}}$ values vary greatly over five sub-bands, revealing the quite different contributions of specific part to qualities. Moreover, $R^2\hat{\mathbf{Y}}^{\mathrm{s}}$ quantifies the cumulative quality variations explained by specific part over all sub-bands. The final cumulative quality interpretation ($R^2\hat{\mathbf{Y}}$) is then obtained by combining the effects of both common and specific parts over sub-bands. Generally, it satisfies $R^2\hat{\mathbf{Y}} \approx R^2\hat{\mathbf{Y}}^{\mathrm{c}} + R^2\hat{\mathbf{Y}}^{\mathrm{s}}$, which agrees well with the real situation that the concerned quality variations should be complementary as the common part and the specific part cover two different types of descriptor variations, which are orthogonal with each other. Comparatively, without variation division, one can only evaluate the overall effects of each sub-band from a general viewpoint. However, we cannot know how the underlying information behaves in detail as what has been shown by the proposed method. The proposed method enables the separation of different variations and a better representation of underlying information so that predictor information in each data space could be made better use of.

In detail, to compare their different cumulative effects of common and specific parts, the common LVs are also collected in the similar way to what has been done in specific part, forming a joint input space, $\mathbf{T}^{\mathrm{c}}(N \times \mathbf{5}) = \left[\mathbf{t}_1^{\mathrm{c}}, \mathbf{t}_2^{\mathrm{c}}, \mathbf{t}_3^{\mathrm{c}}, \mathbf{t}_4^{\mathrm{c}}, \mathbf{t}_5^{\mathrm{c}}\right]$. Then, PLS regression analysis is performed between $\mathbf{T}^{\mathrm{c}}$ and qualities. The weight coefficients attached to predictors with respect to the first PLS LV are derived and shown in Figure 7a in comparison with the weight coefficients in specific part shown in Figure 7b. It is observed that within the common part the weights generally stay invariable over sub-bands, which means these common LVs act similarly in cumulative quality interpretation when they are put together. This indicates their repetitive cumulative contributions to qualities. In contrast, the weights are different over sub-bands in the specific part,

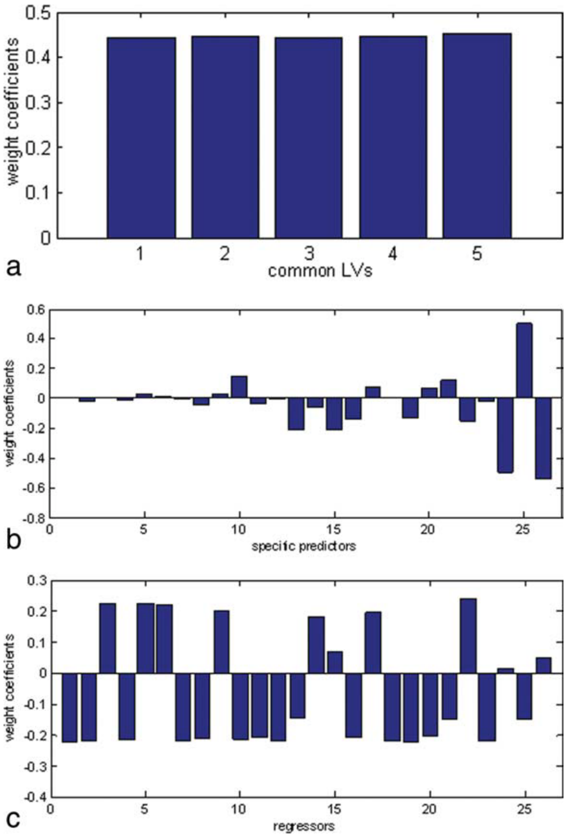

Figure 7. Sub-band-wise predictor weight coefficients with respect to the first PLS LV: (a) For the common part, (b) for the specific part, and (c) for the whole sub-band with no variation division.

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

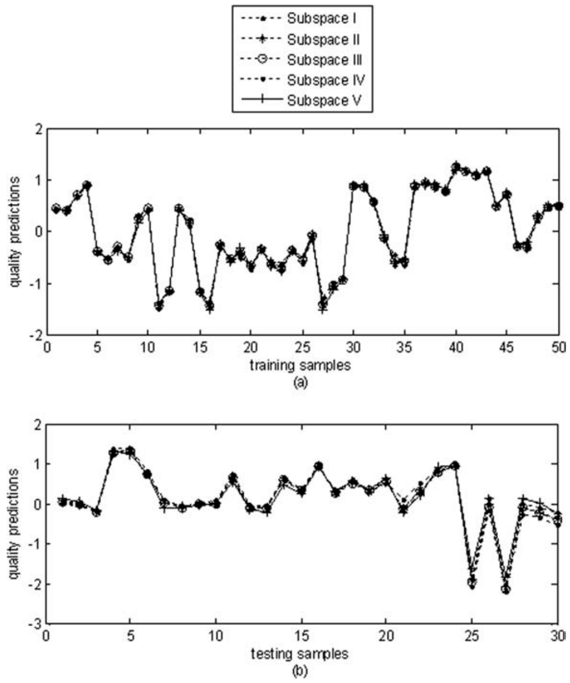

Figure 8. Predicted quality profiles in common part over sub-bands taking example for the first quality index for: (a) Training data and (b) testing data.

revealing the predictors play quite differently from one sub-band to another. Especially, how the operation patterns in the two parts act in quality analysis respectively under the influence of inter-sub-band correlations can be exposed. Generally, the more serious the complementary cumulative effects in specific part are, the more different the weights attached to different sub-bands are. Moreover, comparing the weights in specific part with those obtained without variation division in each sub-band as shown in Figure 7c, it is clear that after the exclusion of common LVs, the weights are more different because only different contributions to qualities are focused on in the specific part.

Further, in the common part, the same contributions to qualities are revealed, i.e., the same quality prediction competency. This can be clearly seen from Figure 8, where over different sub-bands, the predicted quality profiles match well taking example for the first quality index with respect to both training and testing data. Moreover, as shown in Table 3, the quality predictions by the common part are compared between any two sub-bands, which are quantitatively evaluated by correlation analysis. All show high correlation coefficients, revealing that they are quite similar with each other and thus suggest similar contributions of the common part to qualities over sub-bands. Moreover, it should be noted that the common variation in each sub-band can only explain one part of quality information as indicated by $\boldsymbol{R}^2\hat{\mathbf{Y}}_b^{\mathrm{c}}$ shown in Table 2.

In conclusion, the above illustrations and discussions have illustrated those theoretical analyses mentioned in "Methodology" Section concerning different chemical characteristics over wavelength regions, different variation information within each sub-band and their different cumulative roles in quality prediction and interpretation. By sub-band separation and variation division, one can expect to further check the local characteristics of different variations in different sub-bands and get a more comprehensive understanding before regression modeling. A quantitative statistical analysis can reveal their different roles in cumulative quality interpretation. The result of this study has constituted a step forward towards quality-related cumulative calibration analysis for spectra data. Also, it provides the basis for further work and improvement.

Conclusions

In this article, a spectra calibration analysis method is presented using sub-band separation and variation division to analyze the cumulative effects on quality interpretation along

Table 3. Between-Sub-band Comparison of Quality Predictions in Common Part by Correlation Analysis

Spectra Sub-Bands	I	II	III	IV	V
I	1.0000	0.9984	0.9980	0.9963	0.9945
II	0.9984	1.0000	0.9996	0.9990	0.9974
III	0.9980	0.9996	1.0000	0.9993	0.9982
IV	0.9963	0.9990	0.9993	1.0000	0.9979
V	0.9945	0.9974	0.9982	0.9979	1.0000

wavelength direction. Different spectra sub-bands are separated from the original wavelength region based on the changes of underlying chemical characteristics. Then, focusing on each sub-band, the common and specific variations are distinguished, which reveal different types of cumulative effects, repetitive and complementary cumulations. In this way, we can find more interesting and reliable model representation, better identify the underlying spectra information and get enhanced statistical understanding for spectra quality interpretation. Further, to get better modeling performance, on the one hand, we can improve the sub-bands separation algorithm as well as the division of common and specific variations so as to make better use of them for quality interpretation and prediction. On the other hand, maybe a global model can be designed by adopting proper algorithm instead of the band-varying multiple models to fit the mixture spectra. Although this work only illustrates the effectiveness of the proposed method in spectra analysis, it should be generally applicable to a broad range of practical cases, such as multiphase chemical batch processes. These issues are meaningful and deserve further efforts in future work.

Acknowledgments

This work is supported by China's National 973 program (2009CB320603).

Notation

$\mathbf{X}_b$ ($J_b \times N$) = spectral data in the *b*th sub-band

$\mathbf{Y}$ ($N \times J_y$) = concentration matrix

N = the number of samples in the *b*th sub-band

J_b = the number of wavelength variables in the *b*th sub-band

J_y = the number of concentration indices

$\mathbf{S}_b$ = source spectra in the *b*th sub-band

$\mathbf{A}_b$ = mixing matrix in the *b*th sub-band

R_b = the number of retained ICs in the *b*th sub-band

$\mathbf{T}_g$ ($N \times R^{\mathrm{c}}$) = the global LVs over sub-bands

R^{c} = the retained number of common LVs

$\mathbf{T}_b^{\mathrm{c}}$ = common LVs in the *b*th sub-band

$\mathbf{P}_b^{\mathrm{c}}$ = predictor loadings for common part in the *b*th sub-band

$\mathbf{A}_b^{\mathrm{c}}$ = separated sub-band-common systematic variations in the *b*th sub-band

$\mathbf{A}_b^{\mathrm{s}}$ = separated sub-band-specific variations in the *b*th sub-band

$\mathbf{G}_{\mathbf{T}_b^{\mathrm{c}}}$ = the orthogonal projector onto the column space of $\mathbf{T}_b^{\mathrm{c}}$

$\mathbf{H}_{\mathbf{T}_b^{\mathrm{c}}}$ = the anti-projector with respect to the column space of $\mathbf{T}_b^{\mathrm{c}}$

Cumulative analysis in the common part over sub-bands

$\mathbf{T}_v^{\mathrm{c}}$ = variable-unfolding common LVs

$\mathbf{Y}_v$ = variable-unfolding quality

$\mathbf{R}^{\mathrm{c}}$ ($R^{\mathrm{c}} \times A^{\mathrm{c}}$) = common PLS-CCA model weights

A^{c} = the number of common PLS-CCA LVs

$\mathbf{P}^{\mathrm{c}}$ = predictor weights in common part

$\mathbf{Q}^{\mathrm{c}}$ = quality weights in common part

$\mathbf{T}^{\mathrm{c}}$ = common PLS-CCA LVs

Λ^{c} = diagonal covariance matrix of LVs $\mathbf{T}^{\mathrm{c}}$

$\mathbf{T}^{b,\mathrm{c}}$ ($N \times \Lambda^{\mathrm{c}}$) = PLS-CCA LVs separated from $\mathbf{T}^{\mathrm{c}}$ in the *b*th sub-band

$\hat{\mathbf{T}}_b^{\mathrm{c}}$ = the participated common variations in the *b*th sub-band

$\hat{\mathbf{Y}}_{V,b}$ = predicted quality variations by the common part in the *b*th sub-band

$\hat{\mathbf{Y}}^{\mathrm{c}}$ = predicted quality variations by the common part over all sub-bands

Cumulative analysis in the specific part over sub-bands

$\mathbf{A}^{\mathrm{s}}\left(N \times \sum_{b \in B} R_b\right)$ = joint predictor space by specific variations over all sub-bands

$\mathbf{R}^{\mathrm{s}}$ = specific PLS-CCA model weights

A^{s} = the number of specific PLS-CCA LVs

$\mathbf{P}^{\mathrm{s}}$ = predictor weights in specific part

$\mathbf{Q}^{\mathrm{s}}$ = quality weights in specific part

$\mathbf{T}^{\mathrm{s}}$ = specific PLS-CCA LVs

Λ^{s} = diagonal covariance matrix of LVs $\mathbf{T}^{\mathrm{s}}$

$\hat{\mathbf{A}}^{\mathrm{s}}$ = reconstructed predictor variations in specific part

$\mathbf{E}^{\mathrm{s}}$ = the specific predictor errors

$\mathbf{P}^{\mathrm{s}}$ = predictor weights split from $\mathbf{P}^{\mathrm{s}}$ in the *b*th sub-band

$\hat{\mathbf{A}}_b^{\mathrm{s}}$ = the participated specific variations in the $\mathbf{S}_b$ sub-band

$\hat{\mathbf{Y}}^{\mathrm{s}}$ = predicted quality variations by the specific part over all sub-bands

Final regression modeling

$\hat{\mathbf{Y}}$ = the finally predicted quality

$\Theta\left(2J_y \times J_y\right)$ = the final regression coefficients to combine the common and specific parts

Literature Cited

1. Zhao CH, Gao FR. New spectra data analysis and calibration modeling method using spectra subspace separation and multiblock independent component regression strategy. *AIChE J.* In press.
2. Wold S, Antti H, Lindgren F, Öhman J. Orthogonal signal correction of near-infrared spectra. *Chemom Intell Lab Syst.* 1998;44:175–185.
3. 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.
4. Westerhuis JA, Gurden SP, Smilde AK. Spectroscopic monitoring of batch reactions for on-line fault detection and diagnosis. *Anal Chem.* 2000;72:5322–5330.
5. Gurden SP, Westerhuis JA, Smilde AK. Monitoring of batch processes using spectroscopy. *AIChE J.* 2002;48:2283–2297.
6. 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.
7. 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.
8. 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.
9. Ye SF, Wang D, Min SG. Successive projections algorithm combined with uninformative variable elimination for spectra variable selection. *Chemom Intell Lab Syst.* 2008;91:194–199.
10. Xu H, Liu Z, Cai W, Shao X. A wavelength selection method based on randomization test for near-infrared spectral analysis. *Chemom Intell Lab Syst.* 2009;97:189–193.
11. 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.
12. Sjöblom J, Svensson O, Josefson M, Kullberg H, Wold S. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemom Intell Lab Syst.* 1998;44:229–244.
13. 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 Chemom.* 2003;17:174–185.
14. Trygg J. Prediction and spectral profile estimation in multivariate calibration. *J Chemom.* 2004;18:166–172.
15. Andrew A, Fearn T. Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation. *Chemom Intell Lab Syst.* 2004;72:51–56.
16. Chen T, Morris J, Martin E. Gaussian process regression for multivariate spectroscopic calibration. *Chemom Intell Lab Syst.* 2007;87:59–67.
17. 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. *Chemom Intell Lab Syst.* 2008;91:28–33.

18. Benoudjit N, Melgani F, Bouzgou H. Multiple regression systems for spectrophotometric data analysis. *Chemom Intell Lab Syst.* 2009;95:144–149.
19. Alciaturi CE, Quevedo G. Bayesian regularization: application to calibration in NIR spectroscopy. *J Chemom.* 2009;23:562–568.
20. Chen T, Martin E. Bayesian linear regression and variable selection for spectroscopic calibration. *Anal Chim Acta.* 2009;631:13–21.
21. Geladi P, Kowalski BR. Partial least-squares regression-a tutorial. *Anal Chim Acta.* 1986;185:1–17.
22. Brereton RG. Introduction to multivariate calibration in analytical chemistry. *Analyst.* 2000;125:2125–2154.
23. Kleinbaum DG, Kleinbaum DG. *Applied Regression Analysis and Other Multivariable Methods*, 4th ed. Australia: Thomson Brooks/Cole, 2008.
24. Kutner MH, Nachtsheim C, Neter J. *Applied Linear Regression Models*, 4th ed. Boston: McGraw-Hill/Irwin, 2004.
25. Ergon R. Reduced PCR/PLSR models by subspace projections. *Chemom Intell Lab Syst.* 2006;81:68–73.
26. Hyvarinen A, Oja E. Independent component analysis: algorithms and applications. *Neural Networks.* 2000;13:411–430.
27. 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.
28. Shao XG, Wang W, Hou ZY, Cai WS. A new regression method based on independent component analysis. *Talanta.* 2006;69:676–680.
29. 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.
30. 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.
31. Nomikos P, MacGregor JF. Multi-way partial least squares in monitoring batch processes. *Chemom Intell Lab Syst.* 1995;30:97–108.
32. Macgregor JF, Jaeckle C, Kiparissides C, Koutoudi M. Process monitoring and diagnosis by multiblock PLS methods. *AIChE J.* 1994;40:826–838.
33. 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.
34. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *J Chemom.* 1998;12:301–321.
35. Lopes JA, Menezes JC, Westerhuis JA, Smilde AK. Multiblock PLS analysis of an industrial pharmaceutical process. *Biotechnol Bioeng.* 2002;80:419–427.
36. Reinikainen S, Hoskuldsson A. Multivariate statistical analysis of a multi-step industrial processes. *Anal Chim Acta.* 2007;595:248–256.
37. Yu HL, MacGregor JF. Post processing methods (PLS-CCA): simple alternatives to preprocessing methods (OSC-PLS). *Chemom Intell Lab Syst.* 2004;73:199–205.
38. Anderson TW. *An Introduction to Multivariate Statistical Analysis*, 2nd ed. New York: Wiley, 1984.
39. Burnham AJ, Viveros R, MacGregor JF. Frameworks for latent variable multivariate regression. *J Chemom.* 1996;10:31–45.

Appendix

Two-step common LV extraction algorithm

In case of multiset measurement data, $\mathbf{X}^i$ ($N \times J_i$) ($i = 1,2,\ldots C$), their similar interrelation points to the common LV structures. As any subLVs in each dataset space, $\mathbf{t}_\ell^i$ ($\ell = 1,2,\ldots,J_i$) (where subscript ℓ denotes the number of LVs), must lie in the span of the input variables, there exists linear combination coefficients $\mathbf{a}_j^i = [a_{1.j}^i, a_{2.j}^i,\ldots,a_{n.j}^i]$, such that

$$\mathbf{t}_\ell^i = \sum_{j=1}^{J_i} a_{\ell,j}^i \mathbf{x}_j^i = \mathbf{X}^i \mathbf{a}_\ell^i \tag{A1}$$

That is, each subLV $\mathbf{t}_\ell^i$ is actually a linear function of the original variables in each dataset.

The degree of similarity of subLV should be measured in terms of "how close with each other over sets." However, it would be complicated if all set-to-set interrelationships are simultaneously and directly evaluated. In our method, the cross set similarity assessment can be achieved through the introduction of a global and common basis LV vector, $\mathbf{t}_g$. It can be regarded as the supplementary and pseudo ($C + 1$)st subLV and should approximate all C subLVs as close as possible. That is, these real subLVs should be able to be comprehensively described and even substituted by the global LV as they are correlated with each other as close as possible, or speaking more exactly, as common as possible over sets.

To figure out the common LVs, a two-step extraction procedure is designed. In the first step, the common LVs are preparatorily computed from the original measurement data, and then in the second step, they can be further condensed and refined by enhancing their correlations. Different optimization solutions and constraints are used in the two steps, which both come down to the simple analytic solutions of constrained optimization problems.

The first-step LV extraction

During the first-step LV extraction, we try to find a N-dimensional global basis LV ($\mathbf{t}_g$) together with different linear combinations ($\mathbf{a}^g$) of the variables comprising each of C collective data sets with the cost function and certain constraints as below:

$$\begin{aligned} &\max R^2 = \max \sum_{i=1}^{C} \left(\mathbf{t}_g^{\mathrm{T}} \mathbf{X}^i \mathbf{a}^i\right)^2 \\ &\text{s.t.} \begin{cases} \mathbf{t}_g^{\mathrm{T}} \mathbf{t}_g = 1 \\ \mathbf{a}^{i\mathrm{T}} \mathbf{a}^i = 1 \end{cases} \end{aligned} \tag{A2}$$

$(\mathbf{t}_g{}^{\mathrm{T}} \mathbf{X}^i \mathbf{a}^i)^2$ models the covariance information between subLV ($\mathbf{X}^{\mathrm{i}} \mathbf{a}^i$) and global LV ($\mathbf{t}_g$). It is thus acknowledged that the objective function undesirably involves the module of subLV rather than the pure correlation analysis.

Using a Lagrange operator, the optimization problem finally leads to a simple analytical solution:

$$\begin{aligned} &\sum_{i=1}^{C} \left(\mathbf{X}^i \mathbf{X}^{i\mathrm{T}}\right) \mathbf{t}_g = \lambda_g \mathbf{t}_g \\ &\mathbf{Q} \mathbf{t}_g = \lambda_g \mathbf{t}_g \end{aligned} \tag{A3}$$

This is a standard algebra problem. At the request of the maximal objective function value, i.e., the largest λ_g, analytically, the solution leads to the eigenvalue decomposition on the sum of subset covariances, $\mathbf{Q} = \sum_{i=1}^{c} (\mathbf{X}^i \mathbf{X}^{i\mathrm{T}})$.

The subLV can be calculated by:

$$\mathbf{t}_i = \mathbf{X}^i \mathbf{a}^i = \sqrt{\frac{1}{\lambda_i}} \mathbf{X}^i \mathbf{X}^{i\mathrm{T}} \mathbf{t}_g \tag{A4}$$

where the sub-optimal objective parameter λ_i can be calculated by $\mathbf{t}_g^{\mathrm{T}} \mathbf{X}^i \mathbf{X}^{i\mathrm{T}} \mathbf{t}_g = \lambda_i$.

Orderly, $\bar{R}$ number of global LVs can be derived by Eq. A3 in accord with the descending λ_g, resulting in the same number of subLV vectors in each dataset. This decomposition summarizes and compresses the underlying cross-set covarying information into a new sub-band spanned by $\overline{R}$ subLVs within each dataset, $\bar{T}^i(N \times \bar{R})$. However, the maximization of covariance information, as shown in Eq. A2, may not necessarily imply strong correlations. It is possible that a higher covariance merely results from the larger modules of subLV vectors.

The second-step LV extraction

To get the cross-set common sub-bases which are really close correlated, correlation analysis index should be used instead of the covariance index. Unlike the first-step LV extraction, the second-step LV extraction algorithm inherently excludes the effect of module length of each subLV vector and directly maximizes the sum of their correlations. It is implemented on the basis of the first-step analysis result ($\bar{T}^i(N \times \bar{R})$) and the aim is to maximize the mean square correlations by using the same computation trick, that is, a global/common LV vector is introduced as the third-party one. A constrainted optimization problem is formulated:

$$\begin{aligned} &\max R^2 = \max \sum_{i=1}^{C} \left(\mathbf{t}_g^{\mathrm{T}} \bar{\mathbf{T}}^i \mathbf{a}^i \right)^2 \\ &\text{s.t.} \begin{cases} \mathbf{t}_g^{\mathrm{T}} \mathbf{t}_g = 1 \\ \mathbf{a}^{i\mathrm{T}} \bar{\mathbf{T}}^{i\mathrm{T}} \bar{\mathbf{T}}^i \mathbf{a}^i = 1 \end{cases} \end{aligned} \tag{A5}$$

Using a Lagrange operator, its solution comes down to the eigenvector decomposition of $\mathbf{S} = \sum_{i=1}^{c} (\bar{\mathbf{T}}^i (\bar{\mathbf{T}}^{i\mathrm{T}} \bar{\mathbf{T}}^i)^{-1} \bar{\mathbf{T}}^{i\mathrm{T}})$:

$$\begin{aligned} &\sum_{i=1}^{C} \left(\bar{\mathbf{T}}^i \left(\bar{\mathbf{T}}^{i\mathrm{T}} \bar{\mathbf{T}}^i \right)^{-1} \bar{\mathbf{T}}^{i\mathrm{T}} \right) \mathbf{t}_g = \lambda_g \mathbf{t}_g \\ &\mathbf{S}\mathbf{t}_g = \lambda_g \mathbf{t}_g \end{aligned} \tag{A6}$$

The subLV can be calculated by:

$$\mathbf{t}^i = \bar{\mathbf{T}}^i \mathbf{a}^i = \frac{1}{\sqrt{\lambda_i}} \bar{\mathbf{T}}^i \left(\bar{\mathbf{T}}^{i\mathrm{T}} \bar{\mathbf{T}}^i \right)^{-1} \bar{\mathbf{T}}^{i\mathrm{T}} \mathbf{t}_g \tag{A7}$$

where the sub-optimal objective parameter λ_i can be calculated by $\mathbf{t}_g{}^{\mathrm{T}} \bar{\mathbf{T}}^i (\bar{\mathbf{T}}^{i\mathrm{T}} \bar{\mathbf{T}}^i)^{-1} \bar{\mathbf{T}}^{i\mathrm{T}} \mathbf{t}_g = \lambda_i$.

In turn, one can derive as many $\mathbf{t}_g$ as the rank of $\mathbf{S}$, which is deemed to be less than the sum of the rank of $\bar{\mathbf{T}}^i (\bar{\mathbf{T}}^{i\mathrm{T}} \bar{\mathbf{T}}^i)^{-1} \bar{\mathbf{T}}^{i\mathrm{T}}$. Finally, the retained R number of LV vectors can construct a global LV sub-band $\mathbf{T}_g$ ($N \times R$). Correspondingly, C subLV sub-bands, $\mathbf{T}^i$ ($N \times R$), are also derived, which are actually the projected ones from $\mathbf{T}_g$ ($N \times R$) onto $\bar{\mathbf{T}}^i$ space.

Manuscript received Jun. 19, 2010, revision received Dec. 23, 2010, and final revision received Feb. 9, 2011.