AIChE

A Robust Calibration Modeling Strategy for Analysis of Interference-Subject Spectral Data

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

Fuli Wang
College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning Province, P.R. China

DOI 10.1002/aic.11998
Published online August 26, 2009 in Wiley InterScience (www.interscience.wiley.com)

Preprocessing and correction of mixture spectra have been an important issue with regard to the removal of undesired systematic variation due to variations in environmental, instrumental, or sample conditions. In this article, a new robust calibration modeling strategy is proposed on the basis of independent component analysis (ICA). It aims at separating the interference-subject parasitic subspace from the interference-immune common subspace among all considered cases. The common subspace is further divided into two orthogonal parts according to their relationship with quality: one is quality-irrelevant and the other is quality-informative, in which, only the second part is employed for quality prediction. Focusing on each subspace, it identifies distinct types of underlying source components underlying different spectra subspaces, analyzes their characteristics and roles, and accordingly models them for different applications, respectively. This approach provides a comprehensive insight into the inherent nature of interference-subject mixture spectra. Furthermore, several model statistics are defined to give quantitative indication on the effectiveness of the correction strategy. The feasibility and performance of the proposed method are illustrated with data from laboratory experiments. © 2009 American Institute of Chemical Engineers *AIChE J,* 56: 196–206, 2010

Keywords: independent component analysis (ICA), robust multivariate calibration, quality prediction, interference-subject spectra, preprocessing and correction

Introduction

During the past decades, the use of spectroscopic information[1–4] had received much attention and begun to emerge as an important technique, which is being heavily encouraged and practiced for different purposes. To analyze the substance composition in mixture samples, calibration model is constructed using the mixture spectra together with the reference concentrations of constituents to form a quantitative prediction relationship. Common multivariate calibration methods[5–9] used for spectra analysis include principal component regression (PCR) and partial least-squares (PLS) regression. They are based on a fact that variable collinearity is typical in spectral data, and its presence can result in unsatisfactory prediction performance. To deal with the problem of high data dimensionality and redundancy, PCR and PLS reduce the number of spectral variables by means of feature extraction, so the original spectral data space is shrunk to a subspace of smaller dimension. Those underlying features are used for regression modeling.

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

The calibration models generally suffer from a lack of robustness. A calibration model developed under one specific circumstance usually performs less well under another, because the spectrum of a testing sample may mismatch with that of a calibration sample, due to changes in external parameters G. Variations in environmental, instrumental, and sample conditions are often called ''interference factors'' that are different from the parameter of interest. For example, different temperatures may cause the measured spectrum to change even for the same solution with the identical concentration. Therefore, the calibration model built at a given temperature cannot be applied confidently to the concentration prediction under another temperature. Generally, it is required that the influence of temperature on spectrum should be removed prior to the development of calibration model. A possible solution to calibration transfer problem is to measure every sample under the new circumstance and construct a new model for it, which, would be both costly and time consuming. The problem of robust calibration modeling has been extensively reviewed and chemometric strategies are numerous. One common and basic idea is the preprocessing optimization of calibration dataset. Some adjustment or correction to remove all or a major part of the interferences before commencing regression, would be useful. Preys et al.[10] summarized the chemometric strategies to improve calibration robustness into two main types: (a) ''generic'' methods, which tend to correct spectra according to their direct contribution to quality atrributes without taking into account the considered interference factors G. Among them, variable selection[11–16] and certain orthogonal projection methods, such as orthogonal signal correction (OSC),[17–21] etc., are some usual preprocessing techniques designed to exclude sources that are of little or no predictive value from regressors during calibration modeling. (b) ''specific'' methods, which make full use of the G-relevant information and can eliminate their interference prior to modeling even when the specific values of G are not explicitly known. For example, external parameter orthogonalisation (EPO)[22] and transfer by orthogonal projection (TOP),[23] can identify the induced variability due to G and then remove the influence of such variations on spectra measurement. By means of the above preprocessing, the filtered calibration data only cover the common variations regarding the response of interest among different cases. Therefore, the calibration model can be built automatically to be as insensitive as possible to the external parameter influence and thus may be transferred without adjustment. Hansen[24] introduced a preprocessing method called independent interference reduction (IIR) to model the interference effects by implementing PCA on a large number of spectral samples that have the same level for the response of interest but are designed intentionally to cover all known sources of external variation. This was followed by subtraction of the modeled interference information from the calibration matrix. Roger et al.[22] developed an EPO-PLS method to correct temperature-induced spectra variation. It derived the variation directions along which the temperature-induced perturbation located, constructed the interference-orthogonal projector on the basis of these directions, and then projected each spectra data onto it to obtain the common variation information for calibration. Andrew and Fearn[23] proposed a method called transfer by orthogonal projection (TOP) for deriving calibrations robust to between-instrument variation. The idea was to derive the between-instrument variation directions by carrying out PCA on a specifically designed data matrix in which each sample is the mean of spectra data set scanned on each instrument.

Reviewing the previous correction strategies, it can be seen that the common idea is based on interference modeling and orthogonal projection. They generally relied on PCA to find the G-spread variation information in which, the first several columns of PCA loadings relate to the main interferences, and the left ones correspond to the common spectra variation. To obtain a number of well-defined and representative PCA loadings, it often requires that the spectra data should be large enough.[24] Moreover, based on only second-order statistical information, sometimes the variation directions could not be figured out correctly. Actually, NIR spectra of a mixture are often the linear combination of the spectra of its constituent species. It would be very useful if the component spectra can be recovered from the mixture.[25–27] However, PCA, PCR, and PLS are not designed for such a purpose. Independent component analysis (ICA) algorithm[28] is able to deliver this function. It finds independent components (ICs) that constitute the observed variables. It is distinctive to other methods, because it is aimed at separating the spectra of the constituent components of the mixture as well as determining their concentrations, which is called blind source signal separation process. This is clearly beneficial when all or some components of a mixture are unknown. Chen and Wang[25] have applied ICA on near-infrared spectral data, which successfully proved the effectiveness of ICA for recovering the components of interest from spectra mixture. However, if the mixture spectra are subject to interferences, the constituent separation result will be distorted in which, the estimated ICs may be far from the real constituent substances.

In this article, a new robust calibration modeling strategy is developed using ICA algorithm for spectral data. One of its advantages is that it does not need so many modeling samples as previous methods. The proposed preprocessing approach aims at splitting the original spectral space into different subspaces. The general principle of the proposed method is that the interference-induced spectra variation can be regarded as the linear combination of some underlying unobserved and independent sources. Accordingly, the mixed spectra of interference sources can be separated as a parasitic subspace from the spectral space of $\mathbf{X}$ by ICA. The residual subspace is the common part among different cases independent of interference factors G, which is further partitioned into two different sections quality-related vs quality-orthogonal, by revealing their direct relationships with quality. Then the regression model is created only using the quality-relevant information in the interference-filtered common subspace, resulting in an enhanced causal relationship, and presenting a high potential for the maintenance of prediction performance when it is transferred to another case different from the reference one.

This article is organized as follows. First, the proposed method is introduced and its underlying principle is clarified. Moreover, a quantitative calibration analysis is developed for model validation and performance evaluation. Second, its effects on robust enhancement are evaluated by two real

spectra cases with respect to between-instrument variation and temperature-induced spectra vibration. Results are presented and discussed. Finally, conclusions are drawn in the last section.

Methodology

Theory analysis

For a set of spectra with J wavelengths acquired on N samples, $\mathbf{X}(J \times N)$, and the quality data, concentration matrix, $\mathbf{Y}(N \times J_y)$ (where, J_y is the number of quality variables), a common independent component regression (ICR) model[25] can be formulated as below:

$$\begin{aligned} \mathbf{X} &= \mathbf{SA} + \mathbf{E} \\ \widehat{\mathbf{Y}} &= \mathbf{A}^{\mathrm{T}}\mathbf{B} \end{aligned} \tag{1}$$

where, $\mathbf{S}(J \times R)$ is the estimated independent components from the observed variables, which are the spectra estimation of the pure constituents in the mixture. R is the number of ICs. Ideally, if the estimated ICs exactly match the pure substances constituting the mixture, the mixing matrix, $\mathbf{A}(R \times N)$, will agree well with the concentrations of the substances in mixtures. In practice, they can't match with each other very well, and therefore it cannot be taken for granted that the elements in the matrix $\mathbf{A}(R \times N)$ are concentrations. Therefore, like PCR, regression analysis is performed between the estimated concentrations and real ones to derive the regression relationship $\mathbf{B}$. $\mathbf{E}$ is the residual matrix caused by normal random measurement noises, which are well controlled by routine laboratories. That is, if the same sample is scanned multiple times by the same instrument under the same condition, these spectra will differ slightly and stochastically.

In this work, the external interferences G are taken into account, and the spectra are measured with varying interference levels. Under the influence of G, the spectra will vary significantly even for the same sample. Moreover, the resulting variation information, which is not useful to the estimation of constituent concentrations, can result in high prediction error. The interference variations, here called structured/systematic noises, are significantly different from the normal random noises in nature. They are deemed to follow specific characteristics and can cause a systematic alteration of the spectra. One attraction of developing robust calibrations is that it should in principle be possible to transfer the calibration to further, as yet unseen, interferences of the same types. Therefore, the underlying characteristics and rule that the interferences G follow should be exploited and modeled, which will form a parasitic subspace. Then, their effects embedded into the mixture spectra can be eliminated by subtraction. Calibration analysis can thus be performed within the left common subspace in which, further considering the direct relationship with quality, the quality-irrelevant, and quality-relevant information can be distinguished from each other. Taking into account the specific effectiveness of ICA for spectra analysis, in this section, our work will try to figure out these different subspaces from the viewpoint of blind source signal separation.

From a mathematical point of view, the objective is to try to formulate the underlying characteristics of mixture spectra using ICA algorithm[29] and thus construct the calibration model as below:

$$\begin{aligned} \mathbf{X} &= \mathbf{S}_n\mathbf{A}_n + \left(\mathbf{S}_o\mathbf{A}_o + \mathbf{S}_q\mathbf{A}_q\right) + \mathbf{E} \\ \mathbf{Y} &= \mathbf{A}_q^{\mathrm{T}}\mathbf{B} + \mathbf{F} \end{aligned} \tag{2}$$

where, $\mathbf{S}_n(J \times R_n)$ denotes R_n unknown interference source signals and $\mathbf{A}_n(R_n \times N)$ is their mixing relationship, i.e., their contributions to the spectra vibration of the same samples scanned at different levels of G, which actually reveals how those interference sources $\mathbf{S}_n$ influence mixture spectra. Here, it is reasonable to simply regard the mixing coefficient as the contribution magnitude of each IC to spectra since the ICs are mutually independent. $(\mathbf{S}_o\mathbf{A}_o + \mathbf{S}_q\mathbf{A}_q)$ reveals the underlying common subspace among those considered interference-induced cases. $\mathbf{S}_o(J \times R_o)$ reveals R_o quality-irrelevant components split from those quality-related components $\mathbf{S}_q(J \times R_q)$, which both belong to the common subspace; $\mathbf{A}_o(R_o \times N)$ and $\mathbf{A}_q(R_q \times N)$ are respectively their mixing relationships. $\mathbf{E}(J \times N)$ and $\mathbf{F}(N \times J_y)$ are residual matrices, revealing random noises and errors. In this way, the original spectral space is split into three meaningful portions, covering two systematic noise subspaces: one is interference-induced, $\mathbf{S}_n\mathbf{A}_n$, and the other is quality-orthogonal, $\mathbf{S}_o\mathbf{A}_o$, which are both quality-uninformative. Only the quality-related information in common part is used for regression modeling and quality prediction, resulting in the enhanced causal relationship $\mathbf{B}(R_q \times N)$.

ICA-based interference correction

In the following, based on Eq. 2, the correction solution is formulated. Let $\{\mathbf{X}^1, \mathbf{X}^2,\ldots,\mathbf{X}^k\}$ be k $(J \times N)$-matrices of spectral data with the same N samples scanned at k different levels of G, which all point to the same concentrations of constituents in mixtures, $\mathbf{Y}$. Assuming that the common part among the k spectral matrices has been figured out, each spectral data set can be formulated respectively:

$$\begin{aligned} \mathbf{X}^1 &= \left(\mathbf{S}_o\mathbf{A}_o + \mathbf{S}_q\mathbf{A}_q\right) + \mathbf{S}_n\mathbf{A}_n^1 + \mathbf{E}^1 \\ \mathbf{X}^2 &= \left(\mathbf{S}_o\mathbf{A}_o + \mathbf{S}_q\mathbf{A}_q\right) + \mathbf{S}_n\mathbf{A}_n^2 + \mathbf{E}^2 \\ &\vdots \\ \mathbf{X}^k &= \left(\mathbf{S}_o\mathbf{A}_o + \mathbf{S}_q\mathbf{A}_q\right) + \mathbf{S}_n\mathbf{A}_n^k + \mathbf{E}^k \end{aligned} \tag{3}$$

Thus, the difference of spectra at varying G levels is reflected by the linear combinations, $\mathbf{S}_n\mathbf{A}_n^k$ in which, $\mathbf{A}_n^k$ reveals the different contributions of interference sources to spectra for different cases.

Arbitrarily choose one spectral matrix as the master one (here $\mathbf{X}^1$ is selected), and then define the "difference spectra" by subtracting the master spectra from the other slave ones:

$$\begin{aligned} \mathbf{X}^2 - \mathbf{X}^1 &= \mathbf{S}_n\left(\mathbf{A}_n^2 - \mathbf{A}_n^1\right) + \mathbf{E}^2 - \mathbf{E}^1 \\ \mathbf{X}^3 - \mathbf{X}^1 &= \mathbf{S}_n\left(\mathbf{A}_n^3 - \mathbf{A}_n^1\right) + \mathbf{E}^3 - \mathbf{E}^1 \\ &\vdots \\ \mathbf{X}^k - \mathbf{X}^1 &= \mathbf{S}_n\left(\mathbf{A}_n^k - \mathbf{A}_n^1\right) + \mathbf{E}^k + \mathbf{E}^1 \end{aligned} \tag{4}$$

From the dimensionality point of view, the difference operation is equivalent to a simple data preprocessing. Differently, the reference values, i.e., the chosen master spectra, $\mathbf{X}^1$, are not fixed but vary along wavelengths and over samples. It can transform the absolute spectral profile driven by both the inherent constituents and interferences G into relative spectral changes dominated by G. Therefore, such a preprocessing allows us to view clearly the underlying characteristics of relative spectra changes.

To get a more synthetical view, all the difference values between different spectra sets are put side by side: $\mathbf{X}^d(J \times (k-1)N) = [\mathbf{X}^2 - \mathbf{X}^1, \mathbf{X}^3 - \mathbf{X}^1 \ldots, \mathbf{X}^k - \mathbf{X}^1]$, from which, the interference sources $\mathbf{S}_n$ can be extracted:

$$\mathbf{X}^d = \mathbf{S}_n \mathbf{A}_n^d + \mathbf{E}_n^d \tag{5}$$

Once $\mathbf{S}_n$ are identified, their mixing matrix, $\mathbf{A}_n^i (i = 1,2,\ldots,k)$, underlying each original spectral dataset, can be derived using simple least-squares algebra and their influence imposed for each case, i.e., the parasitic subspace $\widetilde{\mathbf{X}}^i$, is then identified from the mixture spectra:

$$\begin{aligned} \mathbf{A}_n^i &= \left(\mathbf{S}_n^{\mathrm{T}}\mathbf{S}_n\right)^{-1}\mathbf{S}_n^{\mathrm{T}}\mathbf{X}^i \\ \widetilde{\mathbf{X}}^i &= \mathbf{S}_n\mathbf{A}_n^i = \mathbf{S}_n\left(\mathbf{S}_n^{\mathrm{T}}\mathbf{S}_n\right)^{-1}\mathbf{S}_n^{\mathrm{T}}\mathbf{X}^i \end{aligned} \tag{6}$$

Here, it readily solves the collinearity problem of typical MLR calculation by guaranteeing an invertible matrix $\mathbf{S}_n^{\mathrm{T}}\mathbf{S}_n$ because of the mutual orthonormality of the ICs. Actually, $\mathbf{S}_n^{\mathrm{T}}\mathbf{S}_n$ is a diagonal matrix with identical diagonal elements.

The preprocessing will then transform each original spectral data $\mathbf{X}^i$ into $\breve{\mathbf{X}}$ by removing those interferences:

$$\breve{\mathbf{X}}^i = \mathbf{X}^i - \widetilde{\mathbf{X}}^i = \left(\mathbf{I} - \mathbf{S}_n\left(\mathbf{S}_n^{\mathrm{T}}\mathbf{S}_n\right)^{-1}\mathbf{S}_n^{\mathrm{T}}\right)\mathbf{X}^i \tag{7}$$

The corrected spectra, $\breve{\mathbf{X}}^i$, are interference-insensitive and thus should be similar among different case since they are from the same sample. In contrast, without preprocessing, the spectra of the same sample scanned at different values of G might mismatch more than the spectra of two different samples measured at the same G value.

O-ICR calibration modeling

After the above correction, the left spectra data represent the common part among different cases. It does not mean that the corrected spectra are all useful for quality prediction. The conventional ICR, as an exclusive two-step implementation algorithm, has the risk similar to PCR. That is, those ICs decomposed from the observed mixed signals are not ensured close causal relationship with quality properties. Besides the systematic and structured interference factors, spectra observations often contain another kind of systematic variations that are of little or no predictive feature. Therefore, it may be observed that the extracted first several ICs capture most systematic information, but not necessarily, explain the quality properties. Generally, more ICs have to be employed for comprehensive quality description. However, it increases the model complexity although increasing descriptor information in a regression model will improve the fitting to the training data. In particular, it will often cause a substantial reduction in the generalized predictive ability of the model. Therefore, it is a good practice to do a proper filtering before embarking on calibration modeling so that the process information that shows the highest correlation with or contributes most to the concerned quality variations can be paid enough attention to. Moreover, removal of superfluous descriptor information can predigest calibration modeling and improve model interpretation.

Up to now, there have been various feature selection methods, which can be used to preprocess those input data, and thus simplify the model structure. Preys et al.[10] have ranged them into the class of "generic" methods. Among them, variable selection[11–16] can directly reduce the model dimension by removing those input variables, which are irrelevant to qualities. In contrast, OSC,[17–21,30,31] as a signal correction technique, tries to remove quality-irrelevant components instead of directly extracting quality-related components. It has been widely used prior to calibration modeling, such as PLS or PCR. After OSC preprocessing, the relationships between the rest PLS latent variables (LVs) and qualities are enhanced, which thus compresses the model size by reducing the number of required LVs in calibration model. Here, OSC algorithm is employed to further filter the spectra. From the mathematical viewpoint, it is performed in the similar way to the ordinary PLS algorithm except for the desired objective and the meaning of the extracted latent components. Instead of the maximization of covariance between $\mathbf{X}$ and $\mathbf{Y}$ in PLS, it minimizes this covariance, i.e., to get some scores from $\mathbf{X}$, which have no relationship with qualities. This is calculated by means of orthogonal projection operation to achieve as close to orthogonality between the OSC components and $\mathbf{Y}$ as possible. The O-ICR method proposed here is a modification to the original ICR algorithm. It can be regarded as an integration of the regular ICR modeling with OSC preprocessing. Moreover, it should be noted that in spectra data analysis, the mixing matrix directly corresponds to the concentration information ($\mathbf{Y}$). Therefore, the proper object to be preprocessed by OSC is the mixing vector instead of each IC. The details of the proposed O-ICR method are shown in Appendix. In this work, O-ICR provides a way to further remove quality-irrelevant systematic variation from the G-corrected spectra data; in other words, to remove variability in $\mathbf{X}$ that is orthogonal to $\mathbf{Y}$ with the additional benefit that such a kind of variation itself can be analyzed individually.

Using the proposed O-ICR algorithm shown in Appendix, the underlying source components are retrieved and the mixing relationships can be explored as follows for the chosen master mixture $\breve{\mathbf{X}}^1$:

$$\begin{aligned} \mathbf{S}_o &= \breve{\mathbf{X}}^1\mathbf{W}_o^1 \\ \mathbf{A}_o^1 &= \left(\mathbf{S}_o^{\mathrm{T}}\mathbf{S}_o\right)^{-1}\mathbf{S}_o^{\mathrm{T}}\breve{\mathbf{X}}^1 \\ \overline{\breve{\mathbf{X}}}^1 &= \breve{\mathbf{X}}^1 - \mathbf{S}_o\mathbf{A}_o^1 = \left(\mathbf{I} - \mathbf{S}_o\left(\mathbf{S}_o^{\mathrm{T}}\mathbf{S}_o\right)^{-1}\mathbf{S}_o^{\mathrm{T}}\right)\breve{\mathbf{X}}^1 \\ \mathbf{S}_q &= \overline{\breve{\mathbf{X}}}^1\mathbf{W}_q^1 \\ \mathbf{A}_q^1 &= \left(\mathbf{S}_q^{\mathrm{T}}\mathbf{S}_q\right)^{-1}\mathbf{S}_q^{\mathrm{T}}\overline{\breve{\mathbf{X}}}^1 \\ \breve{\mathbf{X}}^1 &= \mathbf{S}_o\mathbf{A}_o^1 + \mathbf{S}_q\mathbf{A}_q^1 + \mathbf{E}^1 \end{aligned} \tag{8}$$

where, the mixing relationship, $\mathbf{A}_q^1$, denotes the concentrations of estimated constituents in mixture.

The calibration model is then built only using the quality-related information:

$$\widehat{\mathbf{Y}}^1 = \mathbf{A}_q^{1\mathrm{T}}\mathbf{B} \tag{9}$$

Moreover, based on Eq. 8, it is easy to derive the direct regression relationship from the interference-corrected master spectra to concentrations:

$$\begin{aligned}\widehat{\mathbf{Y}} &= \overline{\breve{\mathbf{X}}}^{1\mathrm{T}}\mathbf{S}_q\left(\mathbf{S}_q^{\mathrm{T}}\mathbf{S}_q\right)^{-1}\mathbf{B} \\ &= \breve{\mathbf{X}}^{1\mathrm{T}}\left(\mathbf{I}-\mathbf{S}_o\left(\mathbf{S}_o^{\mathrm{T}}\mathbf{S}_o\right)^{-1}\mathbf{S}_o^{\mathrm{T}}\right)\mathbf{S}_q\left(\mathbf{S}_q^{\mathrm{T}}\mathbf{S}_q\right)^{-1}\mathbf{B} \\ &= \breve{\mathbf{X}}^{1\mathrm{T}}\Theta\end{aligned} \tag{10}$$

where, $\Theta = \left(\mathbf{I}-\mathbf{S}_o\left(\mathbf{S}_o^{\mathrm{T}}\mathbf{S}_o\right)^{-1}\mathbf{S}_o^{\mathrm{T}}\right)\mathbf{S}_q\left(\mathbf{S}_q^{\mathrm{T}}\mathbf{S}_q\right)^{-1}\mathbf{B}$.

In the similar way, the mixing relationship specific to each case, $\mathbf{A}_q^i$, can be figured out respectively and the calibration relationship can be readily transferred to them:

$$\begin{aligned}&\mathbf{A}_o^i = \left(\mathbf{S}_o^{\mathrm{T}}\mathbf{S}_o\right)^{-1}\mathbf{S}_o^{\mathrm{T}}\breve{\mathbf{X}}^i \\ &\overline{\breve{\mathbf{X}}}^i = \breve{\mathbf{X}}^i - \mathbf{S}_o\mathbf{A}_o^i = \left(\mathbf{I}-\mathbf{S}_o\left(\mathbf{S}_o^{\mathrm{T}}\mathbf{S}_o\right)^{-1}\mathbf{S}_o^{\mathrm{T}}\right)\breve{\mathbf{X}}^i \\ &\mathbf{A}_q^i = \left(\mathbf{S}_q^{\mathrm{T}}\mathbf{S}_q\right)^{-1}\mathbf{S}_q^{\mathrm{T}}\breve{\mathbf{X}}^i \\ &\widehat{\mathbf{Y}}^i = \mathbf{A}_q^{i\mathrm{T}}\mathbf{B} = \breve{\mathbf{X}}^{i\mathrm{T}}\Theta\end{aligned} \tag{11}$$

From the above equation, it can be seen that the transferability of calibration for quality prediction actually depends upon the performance of both interference correction and O-ICR modeling. That is, the validity of this approach depends on the successful identification of both the important interference sources and pure constituent species.

For a new sample, $\mathbf{x}_{\mathrm{new}}$ ($J \times 1$), the preprocessing is carried out and its concentration can be estimated as follows:

$$\begin{aligned}&\breve{\mathbf{x}}_{\mathrm{new}} = \left(\mathbf{I}-\mathbf{S}_n\left(\mathbf{S}_n^{\mathrm{T}}\mathbf{S}_n\right)^{-1}\mathbf{S}_n^{\mathrm{T}}\right)\mathbf{x}_{\mathrm{new}} \\ &\mathbf{a}_{o\,\mathrm{new}} = \left(\mathbf{S}_o^{\mathrm{T}}\mathbf{S}_o\right)^{-1}\mathbf{S}_o^{\mathrm{T}}\breve{\mathbf{x}}_{\mathrm{new}} \\ &\overline{\breve{\mathbf{x}}}_{\mathrm{new}} = \breve{\mathbf{x}}_{\mathrm{new}} - \mathbf{S}_o\mathbf{a}_{o\,\mathrm{new}} \\ &\mathbf{a}_{q\,\mathrm{new}} = \left(\mathbf{S}_q^{\mathrm{T}}\mathbf{S}_q\right)^{-1}\mathbf{S}_q^{\mathrm{T}}\overline{\breve{\mathbf{x}}}_{\mathrm{new}} \\ &\widehat{\mathbf{y}}_{\mathrm{new}}^{\mathrm{T}} = \mathbf{a}_{q\,\mathrm{new}}^{\mathrm{T}}\mathbf{B} = \breve{\mathbf{x}}_{\mathrm{new}}^{\mathrm{T}}\Theta\end{aligned} \tag{12}$$

Orthogonalization property

Based on two-step preprocessing, involving both interference correction and quality-orthogonal information removal, the original spectral space is separated into three different subspaces according to whether they are influenced by G and related to quality, which are $\mathbf{S}_n\mathbf{A}_n^i$, $\mathbf{S}_o\mathbf{A}_o^i$, and $\mathbf{S}_q\mathbf{A}_q^i$. Some orthogonalization relationships exist between the three subspaces, which are clarified as below.

(a) $\breve{\mathbf{X}}^{i\mathrm{T}}\widetilde{\mathbf{X}}^i = \mathbf{0}$

For Eq. 7, $\mathbf{S}_n(\mathbf{S}_n^{\mathrm{T}}\mathbf{S}_n)^{-1}\mathbf{S}_n^{\mathrm{T}}$ is actually the orthogonal projector onto the column space of $\mathbf{S}_n$, and $\mathbf{I}-\mathbf{S}_n(\mathbf{S}_n^{\mathrm{T}}\mathbf{S}_n)^{-1}\mathbf{S}_n^{\mathrm{T}}$ is the antiprojector with respect to $\mathbf{S}_n$-space. Accordingly, it can be found out that the parasitic subspace is orthogonal against the corrected spectral space: $\breve{\mathbf{X}}^{i\mathrm{T}}\widetilde{\mathbf{X}}^i = \mathbf{0}$, which means that each sample in original spectral space is split into two mutually orthogonal parts.

(b) $\mathbf{S}_o^{\mathrm{T}}\mathbf{S}_n = \mathbf{0}$

Resulting from the orthogonality claimed in (a), it is readily to know that the underlying source components are orthogonal with each other: $\mathbf{S}_o^{\mathrm{T}}\mathbf{S}_n = \mathbf{0}$ due to $\mathbf{W}_o^{1\mathrm{T}}\breve{\mathbf{X}}^{1\mathrm{T}}\widetilde{\mathbf{X}}^1\mathbf{W}_n^1 = \mathbf{0}$.

(c) $(\mathbf{S}_o\mathbf{A}_o^i)^{\mathrm{T}}\overline{\breve{\mathbf{X}}}^i = \mathbf{0}$ and $\mathbf{S}_o^{\mathrm{T}}\mathbf{S}_q = \mathbf{0}$

The two subspaces separated from $\breve{\mathbf{X}}^i$, $\mathbf{S}_o\mathbf{A}_o^i$, and $\overline{\breve{\mathbf{X}}}^i$ are also orthogonal based on the similar reason with that of (a), as well as the source ICs, $\mathbf{S}_o$ and $\mathbf{S}_q$.

Therefore, by means of orthogonalization treatment, the three subspaces reveal different underlying information, which are guaranteed to have no overlap.

Total number of components

Possible loss of information and ambiguity regarding discarded variation are issues that complicate pretreatment of spectral data. There is also a risk that excessive preprocessing may result in an overfitting problem. It is known that OSC makes use of the rigor orthogonality constraint between components and qualities to extract and remove quality-irrelevant features from the original spectra measurement. From the mathematical viewpoint, quality-orthogonal means completely quality-irrelevant, which, however, may result in such a risk that the spectra of calibration data may be overcorrected if overmany OSC components are allowed. Particularly, it may lose generality when referring to new samples. In the proposed algorithm, the selection of the number of ICs, respectively referring to the interference sources, the quality-orthogonal components and constituent sources, is of importance. On the one hand, from the difference spectra, it is possible to calculate at the most N interference ICs. Then this preprocessing can be applied with a number of ICs R_n varying from 1 up to N, the size of which directly determines the extent of interference removal. On the other hand, both number of quality-orthogonal ICs and quality-informative ICs, R_o and R_q, also vary from 1 up to N, the sizes of which directly decide the regression relationship as shown in Eq. 10. Therefore, the number of all components should be determined with caution to obtain the desired transferable calibration and quality prediction. Here, two evaluation indices are defined to determine their respective number.

Fixing the number of quality-orthogonal and quality-predictive components prior, a similarity index is characterized as follows to check the effects of different number of interference ICs:

$$\begin{aligned}\mathrm{Simi}_{R_n} = &\frac{1}{R_o(k-1)}\sum_{i=2}^{k}\sum_{r=1}^{R_o}\frac{\left|\cos\left(\mathbf{a}_o^{i,r},\mathbf{a}_o^{1,r}\right)\right|}{1+\left|\left|\mathbf{a}_o^{i,r}\right|-\left|\mathbf{a}_o^{1,r}\right|\right|} \\ &+\frac{1}{R_q(k-1)}\sum_{i=2}^{k}\sum_{r=1}^{R_q}\frac{\left|\cos\left(\mathbf{a}_q^{i,r},\mathbf{a}_q^{1,r}\right)\right|}{1+\left|\left|\mathbf{a}_q^{i,r}\right|-\left|\mathbf{a}_q^{1,r}\right|\right|}\end{aligned} \tag{13}$$

where, ($N \times 1$)-dimension $\mathbf{a}_o^{i,r}$ and $\mathbf{a}_q^{i,r}$ are, respectively, the rth row vector of $\mathbf{A}_o^i$ and $\mathbf{A}_q^i$. Operator cos() stands for the cosine of angle between two vectors. $|\mathbf{a}_o^{i,r}|$ stands for the model of $\mathbf{a}_o^{i,r}$.

In this way, the similarity index simultaneously considers the information of the space angles and their modular differences. It is apparent that the index ranges from 0 up to 1. The larger the similarity value, the more similar the mixing relationship. $\mathrm{Simi}_{R_n} = 0$ if $\mathbf{A}_s^i \perp \mathbf{A}_s^1$, while $\mathrm{Simi}_{R_n} = 1$ if and only if $\mathbf{A}_s^i = \mathbf{A}_s^1$ for any one of the $k - 1$ mixing spaces.

When G-sensitive ICs are determined, the number of quality-orthogonal and predictive ICs should be selected to get better quality prediction results. The mean prediction square error (MSE) is formulated based on validation dataset as follows:

$$\mathrm{MSE} = \frac{1}{kN_{\mathrm{te}}J_y}\sum_{i=1}^{k}\sum_{m=1}^{N_{\mathrm{te}}}\sum_{j=1}^{J_y}(y_{m,j}^i - \hat{y}_{m,j}^i)^2 \tag{14}$$

where, subscripts m and j denote the validation sample and quality variable, respectively; $\hat{y}$ is the quality prediction and y is the real quality measurement. These MSE values from all k cases are then integrated to quantitatively evaluate the current prediction performance.

Combining the above two indices, the respective number of various ICs can be properly selected by cross-validation or trial and error. Moreover, it should be pointed out that due to the mutual independency of ICs, it is appealing that there is no need to re-estimate the mixing model whenever excluding one IC. It just needs to remove the row coefficient vector in $\mathbf{A}_s^i$ ($R_s \times J$) corresponding to the removed IC. This can greatly ease the calculation of model validation.

Quantitative model statistics

In this subsection, six quantitative model statistics are present to analyze the effects of the proposed pretreatment method on spectral data, which respectively check the amount of different systematic information located in different subspaces. These model statistics will be further quantitatively clarified in the simulation section.

(1) Modeled variation of $\mathbf{X}$

$$R^2\mathbf{X} = 1 - \frac{\sum \mathbf{E}^2}{\sum \mathbf{X}^2} \tag{15}$$

(2) Modeled interference-induced variation of $\mathbf{X}$

$$R^2\widetilde{\mathbf{X}} = \frac{\sum (\mathbf{S}_n\mathbf{A}_n)^2}{\sum \mathbf{X}^2} \tag{16}$$

(3) Quality-predictive variation of $\mathbf{X}$

$$R^2\breve{\mathbf{X}}_q = \frac{\sum (\mathbf{S}_q\mathbf{A}_q)^2}{\sum \mathbf{X}^2} \tag{17}$$

(4) Modeled quality-orthogonal variation of $\mathbf{X}$

$$R^2\breve{\mathbf{X}}_o = \frac{\sum (\mathbf{S}_o\mathbf{A}_o)^2}{\sum \mathbf{X}^2} \tag{18}$$

(5) Predicted variation of $\mathbf{Y}$

$$R^2\widehat{\mathbf{Y}} = \frac{\sum \widehat{\mathbf{Y}}^2}{\sum \mathbf{Y}^2} \tag{19}$$

(6) Quality-predictive explanations:

$$\begin{aligned}\mathbf{X} &= \mathbf{SA} + \mathbf{E}\\ R^2\mathbf{XY}_{\mathrm{corr}} &= \frac{\mathrm{CCA}(\mathbf{A}_q, \mathbf{Y})}{\mathrm{CCA}(\mathbf{A}, \mathbf{Y})}\end{aligned} \tag{20}$$

where, CCA() stands for those accountable variations in $\mathbf{Y}$ by the identified canonical variates from mixing matrix using canonical correlation analysis (CCA) algorithm.[32–34] The numerator denotes the underlying quality-predictive information after spectra correction as calculated in Eq. 9, whereas the denominator reveals the original one without any preprocessing. In this work, we have made such an important assumption that the interference factors G have no relation with quality properties, which means that the underlying quality-explicable information covered in spectral data should remain invariable no matter whether interferences are corrected. By Eq. 20, one can quantitatively check the scenario, e.g., whether some quality-predictive information is lost during the preprocessing and how much.

In conclusion, by ICA algorithm, the proposed method can provide more chemical meanings, which are useful for further understanding and analysis.

Simulations and Discussions

In this section, the new algorithm is applied to two study cases, in which, its performance is evaluated based on two types of interferences, between-instrument and temperature-varying variations.

Experiment datasets

Case Study 1: Robust Calibration to Between-Instrument Spectral Variation. The first data set consists of spectra from 80 samples of corn with wavelength ranging 1100–2498 nm at 2 nm intervals (700 channels), in which, each spectra sample are scanned on three different NIR spectrometers (m5, mp5, and mp6). Therefore, we can collect 80 × 3 (samples × instruments) spectral observations in all. They all correspond to the same concentrations, which are involved in the response matrix $\mathbf{Y}$(80 × 4), referring to four constituents, moisture, oil, protein, and starch. The corn data are available at the Eigenvector Research homepage: http://www.eigenvector.com/DATA/Corn.

Case Study 2: Robust Calibration to Temperature-Induced Spectral Variation. The spectral data of case study 2, obtained from the literature[35] are ternary mixtures of water, ethanol and 2-propanol recorded in a 1 cm cuvette with the wavelength range 580–1091 nm. The short-wave NIR spectra of 19 samples are taken at different temperatures (30, 40, 50, 60, 70°C), i.e., 19 × 5 (samples × temperatures) spectral observations in all, and the temperature of the samples is controlled (~0.2° variation). Besides the spectral measurement $\mathbf{X}$, the quality matrix $\mathbf{Y}$(19 × 3) describes different contents of ethanol, water, and isopropanol.

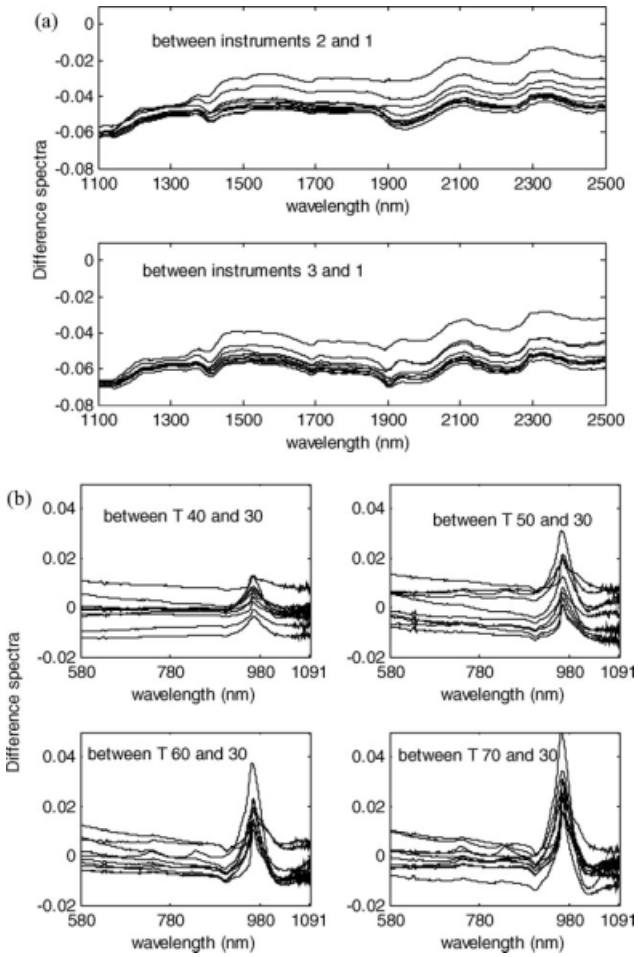

Figure 1. Difference spectra before spectral correction for (a) corn data and (b) ternary mixture.

Simulation methodology and results

In order not to interfere with the correction strategy presented here, no preprocessing has been carried out on the spectra. In the first case, three data sets $\{\mathbf{X}^1,\mathbf{X}^2,\mathbf{X}^3\}$ from three different instruments are used, and in the second case, five data sets $\{\mathbf{X}^1,\mathbf{X}^2,\mathbf{X}^3,\mathbf{X}^4,\mathbf{X}^5\}$ with different temperatures are employed, both of which cover only ten samples in each training data set. The rest data are then used as testing set. In each case, due to the influences of interference factors, the spectral profiles are different even they are from the same samples covering the same constituents with the same concentrations. In both case studies, master spectra are arbitrarily chosen and used to set up the calibration model since their choice would not impose significant influence on the implementation performance of the proposed correction strategy. Some idea of the effect of G on the spectra may be obtained from the difference spectra shown in Figure 1. For case 1, the two spectra are respectively the difference values between instruments 2 and 1 and between instruments 3 and 1 before removal of between-instrument variation, in which, samples from instrument 1 is selected as master ones. For case 2, the four spectra show the differences of spectra measurements influenced by different measurement temperatures, in which, the spectra taken at 30°C are used as the master ones. Both cases show that spectral data does not agree well with each other. On the one hand, the difference magnitude is determined by the size of interference levels. For example, for case study 2 shown in Figure 1b, the greater the difference of temperature, the larger the magnitude of difference spectra. Moreover, spectra on different wavelengths are distorted and affected diversely. On the other hand, compare the difference spectra shown in Figures 1a and b, and it can be found that different types of interference factors may distort the spectra in different ways.

By performing ICA on those difference spectra, two underlying source components, i.e., interference source signals, are figured out and shown in Figure 2 a, taking example for case 2. From their spectra, they explain and agree well with the larger variation of original spectra around the wavelength of 980 nm. Moreover, checking the mixing coefficients shown in Figure 2b, the two interference components indicate different contribution values and influence effects. Generally speaking, the larger the temperature differs, the larger the magnitude of mixing coefficients is, i.e., the G-induced effects are more significant on the spectra, which agrees well the real situation. Moreover, as indicated by the mixing coefficients, the interference ICs impose influences on spectra differently in different samples.

On the basis of the extraction of interference ICs, their effects are removed and the differences of corrected spectra for both cases are shown in Figure 3. Compared with those in Figure 1, the magnitude of difference spectra has been greatly reduced, revealing that after correction, the spectra are transformed to match with each other better. Taking example for case study 2, one can clearly see that the large difference of original spectra around the wavelength of 980 nm has been removed to some extent. The results give one a good impression of the elimination of the temperature effects on the spectra changes, which directly and effectively demonstrate the function of the proposed correction strategy. Calibration analysis is then performed based on the corrected master spectra and quality matrix for each case. Respectively using the proposed O-ICR algorithm and the original ICR algorithm, the regression coefficients between the estimated mixing matrix and the real concentrations are calculated. Taking example for case study 1, they are shown as below:

$$\begin{bmatrix} -0.2403 & -0.1303 & -0.0045 & -1.7069 \\ -0.3750 & 0.0898 & 2.1178 & -0.2757 \\ 1.8870 & -0.2286 & -0.1428 & 0.1168 \\ 0.0285 & 1.9052 & -0.2435 & 0.1518 \end{bmatrix}$$

$$\text{and} \begin{bmatrix} 0.3560 & -0.0644 & -0.4363 & 0.4075 \\ 0.2596 & -0.0632 & -0.2154 & 0.1731 \\ -0.1519 & 0.0609 & -0.0252 & 0.0816 \\ 0.1841 & 0.0555 & -0.7875 & 0.8896 \end{bmatrix}$$

In the two matrices, each row represents the contribution of each estimated constituent component for four concentrations in mixture and each column indicates the contribution of all estimated ICs for each concentration. Checking those coefficients obtained from O-ICR in detail, it is found that moisture is mainly determined by IC3, oil by IC4, protein by IC2, and starch by IC1. Therefore, it can be deduced that

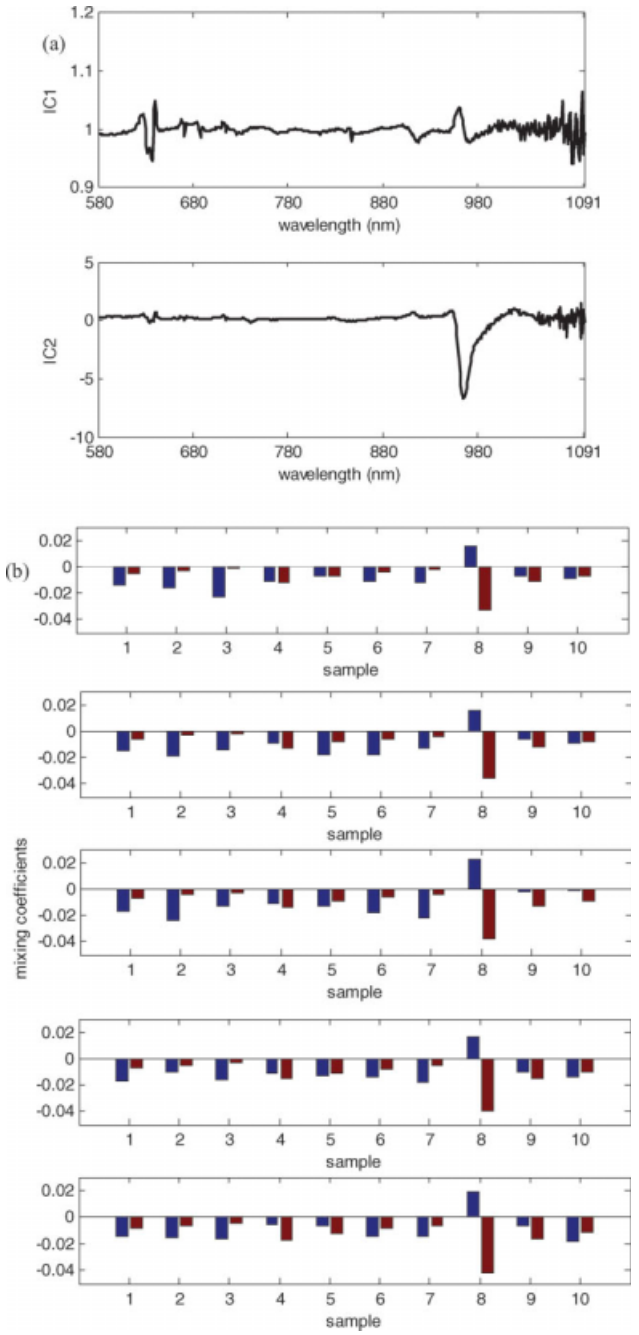

Figure 2. Interference analysis in ternary mixture (a) spectra profile of the estimated interference ICs and (b) mixing coefficients for five temperature-different cases (T = 30, 40, 50, 60, and 70).

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

these ICs respectively correspond to the four pure species in mixture. Therefore, using O-ICR algorithm, the mixing coefficients are more close to the real concentration, i.e., the estimated ICs more approach the real constituents in mixture. In contrast, using conventional ICR, the extracted ICs are more likely to be the linear combination of real compositions.

Because the effects of interference factors have been removed to some extent, it should in principle be possible to transfer the calibration to other instruments or other temperature-induced situations different from the master one. Table 1 shows the simulation results for both cases using the proposed method compared with those obtained from untreated original spectra. The root mean square error of calibration (RMSEC) is calculated for the master spectra (those of instrument 1 in case study 1 and $T = 30$ in case study 2). And the root mean square error of prediction (RMSEP) is used to evaluate the model transferability when the calibration model built based on master spectra is applied to the slave ones for quality prediction. Both of the two statistical indices are calculated by combining the quality prediction results from training and testing data. The results demonstrate that with no correction strategy, the direct application of the primary model to data acquired under the different situations may lead to poor predictions. Such a problem is obviously more apparent when the differences of spectra are large. As expected, the proposed correction strategy improves the model transferability in both examples only based on a small quantity of calibration modeling samples. Moreover, it can be seen that the preprocessing has different effects on the improvement of calibration robustness

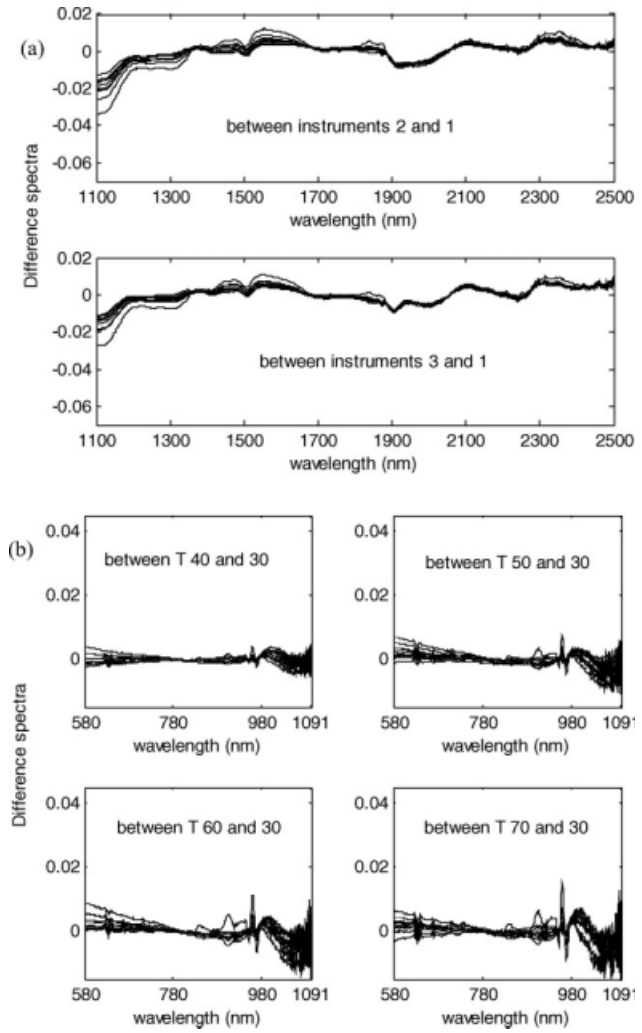

Figure 3. Difference spectra after spectral correction for (a) corn data and (b) ternary mixture.

Table 1. Root Mean Square Errors of Calibration and Prediction for Corn Data and Ternary Mixtures of Water, Ethanol, and 2-Propanol

	Index							
	RMSEC	RMSEP		RMSEC	RMSEP			
Methods	Instrument 1	Instrument 2	Instrument 3	$T = 30$	$T = 40$	$T = 50$	$T = 60$	$T = 70$
No interference correction	0.5496	0.7509	1.1627	0.375	0.9588	1.662	3.1791	4.9397
	0.8501	0.8384	0.8272	0.1792	0.2675	0.3137	0.4465	0.469
	0.7049	0.896	2.4081	0.1786	1.1412	1.713	1.3048	1.611
	0.6043	0.8023	2.3657					
After interference correction	0.1239	0.3774	0.413	0.1436	0.2623	0.5378	0.718	1.5547
	0.6414	0.6101	0.6177	0.1318	0.1283	0.2886	0.306	0.3924
	0.115	0.3837	0.4335	0.1779	0.2245	0.2363	0.2682	0.2182
	0.2237	0.2385	0.4192					

referring to different quality attributes. Further, taking example for the master spectra, a summary of calibration modeling results are given in Table 2, which show the amount of various kinds of modeled variation information. For spectra data **X** in study case 2, the structured information amounts to 95.847% of the total variation; 34.345% of the variation in original spectra is modeled as interference sensitive. The left variations are interference independent and will serve for calibration analysis, which, however, does not indicate that they are all correlated with quality, i.e., they are not all useful for quality prediction. Based on the proposed O-ICR algorithm, the quality-orthogonal variations are modeled amounting to 20.161%, which can not describe quality property and does not take part in any quality prediction. Only 42.494% of **X** is used to explain quality variation; 97.724% of **Y** is modeled by **X** with good predictive ability, as indicated by the statistic $R^2\widehat{\mathbf{Y}}$. The statistical index, $\boldsymbol{R}^2\mathbf{XY}_{\mathrm{corr}}$, provides important clues to the verification of preprocessing effect on quality prediction. Ideally, the spectra correction should not influence the underlying quality-predictive competency. In the simulation, the amount of quality-explicable information in **X** after correction is slightly smaller than that before preprocessing, which reveals that the interference variation may have some relationship with quality more or less. Ideally, if interferences are strictly irrelevant to quality properties, i.e., orthogonal to quality from a mathematical point of view, the quality-predictive explanations underlying the spectral data would exactly match and the predictive ability remains constant, no matter whether interference correction is implemented or not.

The simulations conclude with illustrations of how the proposed calibration modeling strategy performs on real cases, revealing the desirable improvement in model transferability and quality prediction. It also provides the basis and potential for future work. For example, an important issue that can be further addressed may refer to the problem of wavelength selection, which has been mentioned by previous work.[13,14,16,36–38] It directly picks out and thus pays more attention to those important ranges of wavelengths, which contribute in an effective way to quality prediction. Moreover, further analyses can also be conducted focusing on the comprehensive comparisons between the proposed method and all the other methods to reveal their respective advantages and disadvantages. Maybe a better robust calibration modeling strategy can be obtained if their merits are combined. It is a meaningful issue and deserves further investigation.

Conclusion

In this article, a robust calibration modeling strategy is presented, which demonstrates that it is possible to identify sources of variation affecting spectra measurement and calibration performance and observe their effects by ICA. Its underlying idea is to separate different subspaces from the original spectral data by two-step correction. The assumption is that they can be described as a linear combination of different source components. This has been verified in this literature. The first-step preprocessing plays an important role in correcting the interference-induced spectral variations and thus makes the calibration model transferable among different cases. The second-step correction excludes the quality-orthogonal information from regression analysis, which thus enhances the causal relationship with qualities and facilitates parsimonious calibration model structure. Some quantitative model statistics are also defined to verify the modeling results and evaluate the modeling performance. Simulation examples have shown the effectiveness of the proposed method with smaller prediction error and less sensitive to the environmental and instrumental changes. The proposed algorithm provides a tool for more informative statistical explanations and better chemically interpretable characteristics for the analysis of interference-subject spectra. It should be generally applicable to a broad range of multivariate analysis applications aiding in the optimization of the calibration model.

Table 2. Summary of Robust Calibration Modeling Results for Corn Data and Ternary Mixtures

Number of Features				Model Statistics					
$\mathbf{S}_n$	$\mathbf{S}_o$	$\mathbf{S}_q$	LVs	$R^2\mathbf{X}$	$R^2\widetilde{\mathbf{X}}$	$R^2\breve{\mathbf{X}}_o$	$R^2\breve{\mathbf{X}}_q$	$R^2\widehat{\mathbf{Y}}$	$R^2\mathbf{XY}_{\mathrm{corr}}$
1	1	4	2	0.87556	0.29461	0.20731	0.49808	0.96877	0.97950
2	1	4	2	0.95847	0.34345	0.20161	0.42494	0.97724	0.97103

Acknowledgments

The work was supported by the China National 973 program (2009CB320603) and the National Natural Science Foundation of China (No. 60774068).

Literature Cited

1. 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.
2. Gurden SP, Westerhuis JA, Smilde AK. Monitoring of batch processes using spectroscopy. *AIChE J*. 2002;48:2283–2297.
3. 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.
4. 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.
5. Geladi P, Kowalski BR. Partial least-squares regression—a tutorial. *Anal Chim Acta*. 1986;185:1–17.
6. Brereton RG. Introduction to multivariate calibration in analytical chemistry. *Analyst*. 2000;125:2125–2154.
7. Kleinbaum DG, Kleinbaum DG. *Applied Regression Analysis and Other Multivariable Methods*, 4th ed. Australia: Thomson Brooks/Cole, 2008:906.
8. Kutner MH, Nachtsheim C, Neter J. *Applied Linear Regression Models*, 4th ed. Boston: McGraw-Hill/Irwin, 2004:701.
9. Ergon R. Reduced PCR/PLSR models by subspace projections. *Chemometr Intell Lab Syst*. 2006;81:68–73.
10. 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. *Chemometr Intell Lab Syst*. 2008;91:28–33.
11. Walmsley AD. Improved variable selection procedure for multivariate linear regression. *Anal Chim Acta*. 1997;354:225–232.
12. Xu L, Zhang WJ. Comparison of different methods for variable selection. *Anal Chim Acta*. 2001;446:477–483.
13. Abrahamsson C, Johansson J, Sparen A, Lindgren F. Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets. *Chemometr Intell Lab Syst*. 2003;69:3–12.
14. 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 Chemometr*. 2003;17:174–185.
15. Galvao RKH, Araujo MCU, Fragoso WD, Silva EC, Jose GE, Soares SFC, Paiva HM. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. *Chemometr Intell Lab Syst*. 2008;92:83–91.
16. Ye SF, Wang D, Min SG. Successive projections algorithm combined with uninformative variable elimination for spectral variable selection. *Chemometr Intell Lab Syst*. 2008;91:194–199.
17. Wold S, Antti H, Lindgren F, Öhman J. Orthogonal signal correction of near-infrared spectra. *Chemometr Intell Lab Syst*. 1998;44:175–185, 1998.
18. Sjöblom J, Svensson O, Josefson M, Kullberg H, Wold S. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemometr Intell Lab Syst*. 1998;44:229–244.
19. Fearn T. On orthogonal signal correction. *Chemometr Intell Lab Syst*. 2000;50:47–52.
20. Andersson CA. Direct orthogonalization. *Chemometr Intell Lab Syst*. 1999;47:51–63.
21. Westerhuis JA, De Jong S, Smilde AK. Direct orthogonal signal correction. *Chemometr Intell Lab Syst*. 2001;56:13–25.
22. Roger JM, Chauchard F, Bellon-Maurel V. EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. *Chemometr Intell Lab Syst*. 2003;66:191–204.
23. Andrew A, Fearn T. Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation. *Chemometr Intell Lab Syst*. 2004;72:51–56.
24. Hansen PW. Pre-processing method minimizing the need for reference analyses. *J Chemometr*. 2001;15:123–131.
25. Chen J, Wang XZ. A new approach to near-infrared spectral data analysis using independent component analysis. *J Chem Inf Comput Sci*. 2001;41:992–1001.
26. Westad F. Independent component analysis and regression applied on sensory data. *J Chemometr*. 2005;19:171–179.
27. Shao XG, Wang W, Hou ZY, Cai WS. A new regression method based on independent component analysis. *Talanta*. 2006;69:676–680.
28. Hyvarinen A, Oja E. Independent component analysis: algorithms and applications. *Neural Netw*. 2000;13:411–430.
29. Lee J, Qin SJ, Lee I. Fault detection and diagnosis based on modified independent component analysis. *AIChE J*. 2006;52:3501–3514.
30. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J Chemometr*. 2002;16:119–128.
31. Yu HL, MacGregor JF. Post processing methods (PLS-CCA): simple alternatives to preprocessing methods (OSC-PLS). *Chemometr Intell Lab Syst*. 2004;73:199–205.
32. Johnson RA, Wichern DW. *Applied Multivariate Statistical Analysis*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1988:607.
33. Anderson TW. *An Introduction to Multivariate Statistical Analysis*, 2nd ed. New York: Wiley, 1984:675.
34. Burnham AJ, Viveros R, MacGregor JF. Frameworks for latent variable multivariate regression. *J Chemometr*. 1996;10:31–45.
35. Wulfert F, Kok WT, Smilde AK. Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models. *Anal Chem*. 1998;70:1761–1767.
36. Krier C, Rossi F, Franois D, Verleysen M. A data-driven functional projection approach for the selection of feature ranges in spectra with ICA or cluster analysis. *Chemometr Intell Lab Syst*. 2008; 91:43–53.
37. Cai WS, Li YK, Shao XG. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemometr Intell Lab Syst*. 2008;90:188–194.
38. Benoudjit N, Melgani F, Bouzgou H. Multiple regression systems for spectrophotometric data analysis. *Chemometr Intell Lab Syst*. 2009;95:144–149.

Appendix: O-ICR Calibration Algorithm

Here, the O-ICR modeling procedure is comprehensively described on the basis of ICA algorithm presented by Lee et al.[29] to derive the quantitative regression relationship.

First, some useful results about ICA should be clearly clarified. In ICA operation, there always satisfies $\mathbf{AW} = \mathbf{I}$ resulting from the following calculation:

$$\begin{aligned} \mathbf{S} &= \mathbf{XW} \\ \mathbf{A} &= (\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{X} \end{aligned} \tag{A1}$$

Generally, $\mathbf{X}$ is whitened to be $\mathbf{Z}$ prior to ICA by $\mathbf{Z} = \mathbf{XQ}$ (where, $\mathbf{Q}$ is whitening matrix), the components of which are uncorrelated and their variances equal unity. In other words, the covariance matrix of $\mathbf{Z}$ equals the identity matrix: $E\{\mathbf{zz}^{\mathrm{T}}\} = \mathbf{I}$. Therefore, it can be further deduced based on Eq. A1 that

$$\begin{aligned} \mathbf{W}_z &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{S} \\ \mathbf{A}_z &= (\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{Z} \end{aligned} \tag{A2}$$

Consequently, we can obtain $\mathbf{A}_z^{\mathrm{T}} = \mathbf{W}_z$, i.e., $\mathbf{a}_z(N\times 1) = \mathbf{w}_z(N\times 1)$, which means that they both correspond to the same information. Thus, vector $\mathbf{w}_z$ can be used in the calculation process as a substitute for vector $\mathbf{a}_z$.

Denoting $\mathbf{w}_z^i$ as the ith row of $\mathbf{W}_z$, the proposed O-ICR algorithm tries to separate the quality-relevant information from quality-orthogonal one in the ith IC and model them respectively, which is actually realized by handling the estimated mixing/demixing relationship. The detailed procedure is given below:

(a) Determine the number of independent components R_z, and set the counter $p = 1$.

(b) Perform ICA[29] on spectra data $\mathbf{Z}$ so as to get one IC and the corresponding demixing vector $\mathbf{w}_z$.

(c) The quality-orthogonal information in demixing relationship $\mathbf{w}_z$ is separated by $\mathbf{w}_{zo} = (\mathbf{I}\text{-}\mathbf{YY}^+)\mathbf{w}_z$ and then normalized by $\mathbf{w}_{zo} = \mathbf{w}_{zo}/\|\mathbf{w}_{zo}\|$. Here, $\mathbf{YY}^+$ is defined as the orthogonal projector onto the orthogonal column space of $\mathbf{Y}$, and $\mathbf{I}-\mathbf{YY}^+$ as the anti-projector with respect to $\mathbf{Y}$-space. In case $\mathbf{Y}$ is not of full column rank, $\mathbf{Y}^+$ is the Moore-Penrose inverse.

(d) Set $p = p + 1$. If $p \leq R_z$, return to step (b) until all desired ICs are calculated and their corresponding quality-orthogonal information is split. Output the quality-orthogonal demixing matrix $\mathbf{W}_{zo}$ $(N \times R_z)$.

(e) The quality-orthogonal demixing matrix with respect to $\mathbf{X}$, $\mathbf{W}_o(N \times R_z)$, can be readily retrieved by anti-whitening:

$$\begin{aligned} \mathbf{S}_o &= \mathbf{ZW}_{zo} = \mathbf{XQW}_{zo} = \mathbf{XW}_o \\ \mathbf{W}_o &= \mathbf{QW}_{zo} \end{aligned} \tag{A3}$$

and calculate the corresponding quality-orthogonal mixing matrix by:

$$\mathbf{A}_o = (\mathbf{S}_o^{\mathrm{T}}\mathbf{S}_o)^{-1}\mathbf{S}_o^{\mathrm{T}}\mathbf{X} \tag{A4}$$

(f) Remove the quality-orthogonal systematic variation from $\mathbf{X}$ and perform ICA on the residual space again to get the quality-predictive ICA model:

$$\begin{aligned} \mathbf{X}_q &= \mathbf{X} - \mathbf{S}_o\mathbf{A}_o \\ \mathbf{S}_q &= \mathbf{X}_q\mathbf{W}_q \\ \mathbf{A}_q &= \left(\mathbf{S}_q^{\mathrm{T}}\mathbf{S}_q\right)^{-1}\mathbf{S}_q^T\mathbf{X}_q \end{aligned} \tag{A5}$$

In the general formulation, the feature extraction of O-ICR can be considered a variant of projection pursuit. It is developed in statistics for finding more "interesting" projection directions, in which the non-Gaussianity and quality-orthogonal are both called projection pursuit "indices". It directly analyzes the quality-irrelevant variation in each mixing vector corresponding to each regular IC. In this way, the corresponding mixing relationship $\mathbf{A}_q$ is closer related with quality so that the regression relationship between it and quality can be better derived using PLS algorithm.

Manuscript received Mar. 15, 2009, and and revision received June 1, 2009.