Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window Partial Least-Squares Regression with Applications to Mid-Infrared and Near-Infrared Spectroscopic Data

Jian-Hui Jiang,[†,‡] R. James Berry,[†] Heinz W. Siesler,[§] and Yukihiro Ozaki*[,†]

Department of Chemistry, School of Science and Technology, Kwansei-Gakuin University, Gakuen, Sanda, Hyogo 669-1337, Japan, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, P. R. China, and Department of Physical Chemistry, University of Essen, Essen D45117, Germany

A new wavelength interval selection procedure, moving window partial least-squares regression (MWPLSR), is proposed for multicomponent spectral analysis. This procedure builds a series of PLS models in a window that moves over the whole spectral region and then locates useful spectral intervals in terms of the least complexity of PLS models reaching a desired error level. Based on a proposed theory demonstrating the necessity of wavelength selection, it is shown that MWPLSR provides a viable approach to eliminate the extra variability generated by non-composition-related factors such as the perturbations in experimental conditions and physical properties of samples. A salient advantage of MWPLSR is that the calibration model is very stable against the interference from non-composition-related factors. Moreover, the selection of spectral intervals in terms of the least model complexity enables the reduction of the size of a calibration sample set in calibration modeling. Two strategies are suggested for coupling the MWPLSR procedure with PLS for multicomponent spectral analysis: One is the inclusion of all selected intervals to develop a PLS calibration model, and the other is the combination of the PLS models built separately in each interval. The combination of multiple PLS models offers a novel potential tool for improving the performance of individual models. The proposed procedures are evaluated using two open-path Fourier transform infrared data sets and one near-infrared data set, each having different noise characteristics. The results reveal that the proposed procedures are very promising for vibrational spectroscopy-based multicomponent analyses and give much better prediction than the full-spectrum PLS modeling.

Multicomponent spectral analysis has come into widespread use in analytical chemistry. A main goal of multicomponent spectral analysis is to construct a calibration model relating the outputs of multivariate spectrometers to the compositions or the properties of analytical samples. In most situations, a linear calibration model is established due to the mathematical simplicity and the physical or chemical interpretability. While the advances in modern spectroscopic instrumentation have brought enhanced resolution and sensitivity as well as easiness in spectral measurements, the expanded amount of data collected and the increased complexity of samples practically involved persist in a need of useful approaches to combat the overdetermined systems inherent in multicomponent spectral analysis and build robust and stable linear calibration models. A variety of linear regression methods have been proposed for multicomponent spectral analysis, among which the most popular are the so-called latent variable (LV) methods[1−8] including principal component regression (PCR),[1] partial least-squares (PLS) regression,[2] and their analogues.[3−7] A theoretical demonstration has been given that, under certain assumptions, the addition of spectral channels always improves the prediction performance.[9] The implication of this proof is that these LV methods may eliminate the necessity of wavelength selection and have a built-in capacity to deal with the overdetermined problem of full-spectrum calibration. However, there is increasing evidence indicating, either theoretically[10,11] or experimentally,[12,13] that wavelength selection can still significantly refine the performance of these full-spectrum calibration techniques. It has been recognized that the ideal assumptions on which the theoretical proof is based may be unrealistic, and the elimination

[1] Geladi, P.; Kolwalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1.
[2] Sjostrom, M.; Wold, S.; Lindberg, W.; Persson J. A.; Martens, H. *Anal. Chim. Acta* **1983**, *150*, 61.
[3] Stone, M.; Brook, R. J. *J. R. Stat. Soc. B* **1990**, *52*, 237.
[4] de Jone, S.; Kiers, H. A. L. *Chemom. Intell. Lab. Syst.* **1992**, *14*, 155.
[5] de Jong, S. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251.
[6] Wentzell, P. D.; Andrews, D. T.; Kowalski, B. R. *Anal. Chem.* **1997**, *69*, 2299.
[7] Kalivas, J. H. *J. Chemom.* **1999**, *13*, 111.
[8] Burnham, A. J.; McGregor, J. F.; Viveros, R. *Chemom. Intell. Lab. Syst.* **1999**, *48*, 167.
[9] Lober, A.; Kowalski, B. R. *J. Chemom.* **1988**, *2*, 67.
[10] Xu, L.; Schechter, I. *Anal. Chem.* **1996**, *68*, 2392.
[11] Spiegelman, C. H.; McShane, M. J.; Goetz, M. J.; Motamedi, M.; Yue, Q. L.; Cote, G. L. *Anal. Chem.* **1998**, *70*, 35.
[12] Kalivas, J. H.; Roberts, N.; Sutter, J. M. *Anal. Chem.* **1989**, *61*, 2024.
[13] Rimbaud, D. J.; Walczak, B.; Massart, D. L.; Last, I. R.; Prebble, K. A. *Anal. Chim. Acta* **1995**, *304*, 185.

* To whom correspondence should be addressed. E-mail: ozaki@kwansei.ac.jp. Fax: +81-795-65-9077.
[†] Kwansei-Gakuin University.
[‡] Hunan University.
[§] University of Essen.

10.1021/ac011177u CCC: $22.00
Published on Web 06/13/2002

of uninformative spectral channels is still of potential importance in the practice of multicomponent spectral analysis even in situations where the LV methods are applied.

Wavelength selection is composed of the decision of a subset of spectral channels with which the established calibration model gives the minimum errors in prediction. The benefit gained from wavelength selection is not only the stability of the model to the collinearity in multivariate spectra but also the interpretability of the relationship between the model and the sample compositions. A number of procedures have been proposed for wavelength selection in multicomponent spectral analysis. These procedures can be distinguished from each other either in the objective criterion for measuring the optimality of wavelength subsets or in the search algorithm to locate the optimal subsets. Typical objective criteria include the spectral signal-to-noise ratio, the condition number or determinant of the calibration matrix, Akaike information criterion (AIC), and Mallows Cp statistics as well as some estimates of the mean squared error in prediction (MSEP),[14] while routine search algorithms comprise the stepwise selection,[15] simplex optimization,[12] branch and bound combinatorial search,[16] simulated annealing,[17] and genetic algorithms (GAs).[18] However, most of the conventional procedures, generally coupled with an inverse least-squares step, are designed to select a few wavelengths such that the overdetermined system of multicomponent spectral analysis can be converted to an exactly determined one. Because only a small number of spectral channels are utilized in calibration modeling and much information in the whole spectra is not exploited, these approaches may be exposed to significant loss of analytical precision as well as analytical accuracy. To remedy the defect of the conventional wavelength selection methods, considerable effort has been directed toward developing new wavelength selection methods that can be effectively coupled with the full-spectrum calibration techniques.

There are different approaches to wavelength selection that can be implemented in conjunction with the LV modeling techniques. Some of the approaches rank the spectral channels based on the uncertainty of the associated regression coefficients. The wavelengths with large uncertainty are taken as uninformative ones and may be eliminated stepwise during the modeling, while the wavelengths with small uncertainty may be included stepwise in the model.[11,19−21] Other procedures directly optimize the wavelength ranges or subsets together with the number of LVs using GA-based search strategy to minimize an estimate of MSEP.[22−24] In contrast, the present study focuses on the selection of wavelength intervals instead of individual spectral points. The philosophy of search for spectral intervals is the continuity of most kinds of spectral responses. For example, vibrational and rotational spectra give Voigt profiles that generally have a full width at half-height at least 4 cm^{-1}, usually 8−20 cm^{-1}. This implies the existence of intrinsic spectral intervals. In addition, the use of spectral intervals rather than individual spectral points not only enable a straightforward coupling of the wavelength selection procedure with the full-spectrum modeling techniques, thereby providing the possibility of improved analytical accuracy, but also make it possible to implement an effective algorithm for ascertaining the intervals beneficial for the modeling.

In the present study, we demonstrate that the prediction error of indirect (or inverse) calibration may be inflated by including nonideal spectral regions, and a common feature of the nonideal spectral regions is the increased complexity in LV models when these regions are used for calibration modeling. A new wavelength interval selection method, moving window partial least-squares regression (MWPLSR), is thus proposed. This method builds a series of PLS models in a window that moves over the spectral direction and then locates useful spectral intervals in terms of the model complexity and the sum of residuals. When multiple spectral intervals are selected, two strategies are suggested for coupling the MWPLSR procedure with PLS for multicomponent spectral analysis: One is the inclusion of all selected intervals to build a PLS calibration model, and the other is the combination of the PLS models built separately in each interval.

The selection of spectral intervals has been addressed in several works.[10,25−27] Norris used a manual procedure for selecting the best spectral regions via examination of the correlation coefficients between the spectral derivatives and the analyte concentrations with empirical optimization of the derivative gap size.[25] Xu and Schechter proposed a wavelength selection criterion based on the relative error in the norm of the net analytical signal and then optimized both the window position and the window size in such a way that the window gave the minimum relative error.[10] Their method is distinguished from MWPLSR in that it aims at rectifying the full-spectrum approaches in situations where the norm of the net analytical signal is small and the first-order approximation of prediction error becomes unrealistic, while MWPLSR focuses on the cases where the first-order approximation of prediction error is dominated by nonideal spectral intervals and the full-spectrum approaches can be improved via the elimination of these nonideal spectral intervals. Norgaard and coworkers proposed a spectral interval selection procedure, interval partial least-squares regression (iPLS), based on an idea relatively similar to that behind MWPLSR.[26,27] However, unlike the MWPLSR procedure that explores a window gradually moving along the spectral direction using PLS models of varying dimensionalities, the iPLS method tests a series of adjacent but nonoverlapping windows using PLS regression with the same model dimensionality. This may make the iPLS algorithm rather sensitive to the choice of model dimensionality and increase the risk of missing the optimal window. Moreover, MWPLSR can locate the optimized spectral intervals directly, while iPLS requires a postoptimization step to refine the spectral window initially

[14] Martens, H.; Næs, T. *Multivariate Calibration*; Wiley: New York, 1989.
[15] Brown, P. J. *J. Chemom.* **1992**, *6*, 151.
[16] Liang, Y.-Z.; Xie, Y.-L.; Yu, R.-Q. *Anal. Chim. Acta* **1989**, *222*, 347.
[17] Horchner, U.; Kalivas, J. H. *Anal. Chim. Acta* **1995**, *311*, 1.
[18] Lucasius, C. B.; Kateman, G. *Trends Anal. Chem.* **1991**, *10*, 254.
[19] Rimbaud, D. J.; Massart, D. L.; de Noord, O. E.; de Jong, S. Vandeginste, B. G. M.; Sterna, C. *Anal. Chem.* **1996**, *68*, 3851.
[20] Westad, F.; Martens, H. *J. Near Infrared Spectrosc.* **2000**, *8*, 117.
[21] McShane, M. J.; Cameron, B. D.; Cote, G. L.; Spiegelman, C. H. *Appl. Spectrosc.* **1999**, 53, 1575.
[22] Shaffer, R. E.; Arnold, M. A.; Small, G. W. *Anal. Chem.* **1996**, *68*, 2663.
[23] Bangalore, A.; Shaffer, R. E.; Small, G. W.; Arnold, M. A. *Anal. Chem.* **1996**, *68*, 4200.
[24] Berry, R. J. In *Handbook of Vibrational Spectroscopy*; Chalmers, J., Griffiths, P. R., Eds.; John Wiley & Sons: Chichester, U.K., 2001.
[25] Norris, K. H. In *Proceedings of the 1982 IUFST Symposium*; Martens, H., Eds.; Applied Science Publishers: Oslo, 1983.
[26] Norgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B. *Appl. Spectrosc.* **2000**, *54*, 413.
[27] Munck, L.; Nielsen, J. P.; Moller, B.; Jacobsen, S.; Sondergaard, I.; Engelsen, S. B.; Norgaard, L.; Bro, R. *Anal. Chim. Acta* **2001**, *446*, 171.

ascertained. On the other hand, iPLS estimates the calibration model based on all the wavelengths from the accepted windows, which may deteriorate the performance of the total model by including spectral windows with varying goodness. Instead, the present method constructs the total model through a weighted averaging of the local models, which provides a robust way to exploit the information comprised in the local models and avoids possible accumulation of errors in multiple spectral intervals.

The proposed procedures are evaluated using three vibrational spectroscopic data sets, i.e., two open-path Fourier transform infrared (OP/FT-IR) data sets and one near-infrared (NIR) data set, each having different noise characteristics. The issue of wavelength selection is of particular importance in vibrational spectroscopy-based multicomponent spectral analysis, since IR, Raman, and NIR spectra generally show relatively high sensitivity to small perturbations in the experimental conditions as well as the physical properties of samples in comparison with ultraviolet and visible spectra. These non-composition-related factors may make responses at some local spectral intervals deviate significantly from ideal situations, and elimination of these uninformative intervals can substantially improve the analytical accuracy in modeling the concentrations. The results obtained show that the proposed procedures yield superior performance compared to the full-spectrum PLS modeling.

THEORY

Theoretical Background. Consider the indirect (or inverse) calibration model routinely used in multicomponent spectral analysis

$$c = \mathbf{r}^{\mathrm{T}}\mathbf{b} + e \tag{1}$$

where the dependent variable c is the analyte concentration in a sample, the explanatory variables $\mathbf{r}$ are the spectral responses measured at I wavelengths of the sample, the superscript T denotes the matrix or vector transposition, $\mathbf{b}$ is the coefficient vector to be estimated, and e is a model error. Given data, $\mathbf{c} = (c_1,..., c_N)^{\mathrm{T}}$ being the analyte concentrations in N calibration samples, $\mathbf{R}$ being the $N \times I$ response matrix whose nth row $\mathbf{r}_n^{\mathrm{T}}$ is the spectrum of the nth calibration sample, eq 1 can be rewritten as

$$\mathbf{c} = \mathbf{R}\mathbf{b} + \mathbf{e} \tag{2}$$

with $\mathbf{e} = (e_1,..., e_N)^{\mathrm{T}}$. Without loss of generality, one can assume that both $\mathbf{R}$ and $\mathbf{c}$ are columnwise standardized; i.e., each column has zero mean and unit variance. The goal of indirect calibration is to estimate $\mathbf{b}$ such that using eq 1 one can accurately predict the concentrations of the analyte in new samples. This is essentially a linear regression problem with major concern about prediction.

It has been shown that the true regression coefficient vector is given by[28]

$$\mathbf{b} = \mathrm{nas}/||\mathrm{nas}||^2 \tag{3}$$

where nas designates the net analytical signal (NAS) of the analyte[28] and || . || denotes the Euclidean norm of a vector. Ideally, it is assumed that the errors are independent and identically distributed and then the theoretically achievable minimum MSEP is

$$\mathrm{msep}(c) = \sigma^2\ ||\mathbf{b}||^2 = \sigma^2/||\mathrm{nas}||^2 \tag{4}$$

That is to say, the MSEP decreases with the increase of the length of nas, provided the NAS can be available without errors and the errors are independent and identically distributed. As a matter of fact, this conclusion is the basis for the theoretical proof that the addition of spectral channels always improves the prediction performance.[9] However, spectral practice frequently goes contrary to the ideal assumption that the errors are identically distributed and the NAS cannot always be obtained without any deviations. To estimate the error in prediction of concentrations, the errors in $\mathbf{r}$ and $\mathbf{b}$ both should be taken into account.

Assuming that the errors are independent, one can reach the first-order approximation of MSEP from eq 1 as follows:

$$\mathrm{msep}(c) = \sum_{i=1}^{I} b_i^{\,2} d^2(r_i) + \sum_{i=1}^{I} r_i^{\,2} d^2(b_i) \tag{5}$$

where b_i and r_i are the ith elements of $\mathbf{b}$ and $\mathbf{r}$, respectively, and $d^2(\ )$ denotes the squared error of a variable. Notice that the MSEP comprises two parts; one is the estimation errors in the regression coefficients, and the other is the errors in the spectra measured. If J spectral channels are added in calibration modeling, the MSEP is given by

$$\mathrm{msep}(c) = \sum_{i=1}^{I+J} b'^{\,2}_i d^2(r_i) + \sum_{i=1}^{I+J} r_i^{\,2} d^2(b'_i) \tag{6}$$

Then, the variation in MSEP is

$$\Delta\mathrm{msep}(c) = \sum_{i=1}^{I} (b'^{\,2}_i - b_i^{\,2}) d^2(r_i) + \sum_{i=I+1}^{I+J} b'^{\,2}_i d^2(r_i) + \sum_{i=I+1}^{I+J} r_i^{\,2} d^2(b'_i) + \sum_{i=1}^{I} r_i^{\,2} \{d^2(b'_i) - d^2(b_i)\} \tag{7}$$

One can assume that the estimation errors in regression coefficients in the originally selected spectral channels are not significantly affected by the addition of spectral channels; then variation in MSEP can be approximately represented as

$$\Delta\mathrm{msep}(c) = \sum_{i=1}^{I} (b'^{\,2}_i - b_i^{\,2}) d^2(r_i) + \sum_{i=I+1}^{I+J} b'^{\,2}_i d^2(r_i) + \sum_{i=I+1}^{I+J} r_i^{\,2} d^2(b'_i) \tag{8}$$

Now, it is clear that the addition of spectral channels has two kinds of effect on the MSEP. On one hand, the magnitude of b_i gets smaller in the originally selected regions ($b'^{\,2}_i \le b_i^{\,2}$, $i = 1, 2, ...,$

[28] Lorber, A. *Anal. Chem.* **1986**, *58*, 1167.

I), so the first term at the right side of eq 8 takes a negative value, which makes the MSEP decrease. On the other hand, the magnitude of b_i and the errors in regression coefficients associated with the added spectral channels varies from 0 to nonzero value ($b'^2_i \geq 0$, $r_i^2 d^2(b'_i) \geq 0$, $i = I + 1, 2, ..., I + J$), and then the second and the third terms at the right side of eq 18 have a positive value, which causes an increase in the MSEP. As a consequence, if the errors in the spectra obtained at added wavelengths or the estimation errors in the regression coefficients at added spectral channels are too large, the MSEP may be inflated by the inclusion of these spectral channels. In other words, the selection of suitable spectral channels that have good signal-to-noise ratio and provide an accurate estimate for the regression coefficients is capable of improving the accuracy of multicomponent spectral analysis techniques.

Thus far, it has been demonstrated that the spectral channels having poor signal-to-noise ratio and giving large uncertainty in the estimate of regression coefficients will induce an inflated error in the prediction of concentrations. The uncertainty in the estimate of regression coefficients is generated by two factors. One is the error in the concentrations, and the other is the uncertainty in the spectra. In LV modeling, the regression coefficient vector is a certain linear combination of the spectra of calibration samples with combination weights relevant to the concentrations. The errors in the concentrations are propagated into the regression coefficient vector of the whole spectral region through the combination weights, which has a similar effect at different spectral channels. In contrast, the errors in the spectra can be propagated into regression coefficients through the basis vectors in the linear combination, which has different contributions at varying spectral channels. If there is large uncertainty in certain spectral channels, the uncertainty in regression coefficients at the corresponding spectral channels will also be relatively large. Because the effect of errors in the concentrations cannot be reduced via wavelength selection, the objective of wavelength selection can only be set to improving the prediction accuracy by eliminating the spectral channels with large uncertainty or including merely the spectral channels with small uncertainty.

Before the description of the proposed wavelength selection procedure, the characteristic of spectral channels with large uncertainty needs to be addressed first. Actually, the spectral channels with large uncertainty mean that the responses at these channels are severely contaminated by the factors that cannot be modeled using the calibration samples. Such factors include large random errors, nonlinearity, and drifts created by the changes in instrumental parameters, experimental conditions, or physical properties (non-composition-related properties) of samples. Since these non-composition-related factors introduce additional variability in the responses at these spectral channels, if these spectral channels are used for calibration modeling based on an LV method such as PLS, an increased number of LVs has to be constructed to account for extra variability generated by the non-composition-related factors. This results in increased complexity or model dimensionality in the LV model. That is to say, the spectral channels with much uncertainty can be characterized by the significantly increased model dimensionality (the number of LVs) of the LV-based calibration model built using these spectral channels. Conversely, the spectral channels with little uncertainty can be identified as those giving the least model dimensionality.

Wavelength Interval Selection by Moving Window Partial Least-Squares Regression. The motivation for selecting spectral intervals is the continuity of spectral responses. That is, if there is a wavelength informative for the modeling, there must be a spectral interval around the wavelength that contains useful information for the model building. Analogously, if a spectral channel is contaminated by non-composition-related factors, the wavelength interval around the channel will also be interfered with these factors. Based on the conclusion of the preceding section, it is clear that the spectral intervals with large uncertainty can be identified by the significantly increased model dimensionality of the LV-based calibration model built using the spectral interval, while the spectral intervals with small uncertainty can be ascertained as those with the least model dimensionality. This is the basis for the proposed wavelength interval selection procedure, moving window partial least-squares regression.

In MWPLSR, a spectral window that starts at the ith spectral channel and ends at the ($i + H - 1$)th spectral channel is constructed. For simplicity, the window position is used for denoting the starting position of the window. The spectra obtained in the spectral window is a submatrix $\mathbf{R}_i$ ($N \times H$ matrix) containing the ith to the ($i + H - 1$)th columns of the calibration matrix $\mathbf{R}$. The PLS models with different numbers of LVs can then be built to relate the spectra in the window to the concentrations of the analyte. That is,

$$\mathbf{c} = \mathbf{R}_i\mathbf{b}_{i,k} + \mathbf{e}_{i,k} \tag{9}$$

where $\mathbf{b}_{i,k}$ ($H \times 1$ vector) is the regression coefficient vector estimated using PLS with k PLS components and $\mathbf{e}_{i,k}$ is the residue vector obtained with a k-component PLS model. The window is moved over the whole spectral region. At each position, the PLS models with varying PLS component number is built for the calibration samples, and the sums of squared residues (SSR), i.e., the squared norms of the residue vectors, are calculated with these PLS models and plotted as a function of the position of the window. This yields a number of residue lines, with each line associated with the SSR for a certain model dimensionality in the corresponding window position. Obviously, the SSR of the PLS models at a window position will decrease with the increase of the PLS components. Furthermore, based on the aforementioned conclusion, provided the window is positioned in a spectral interval comprising useful information for the modeling and the window size is suitably defined, the SSR is expected to reach an acceptable error level with a relatively small number of PLS components. On the contrary, if the window is located in a spectral interval contaminated significantly by uncertain factors, the SSR cannot approach the desirable error level with a small number of PLS components, and the desired PLS model dimensionality has to be substantially increased such that much more PLS components can be exploited to reduce the SSR. Therefore, by analyzing the desired PLS model dimensionality as the function of the window position, the spectral intervals containing information beneficial for calibration modeling as well as the spectral interval comprising significant uncertainty can be ascertained. Then these informative spectral intervals are selected and utilized for building the calibration model based on PLS.

It is important to note that the PLS method with one dependent variable (PLS1) is employed in the present study for modeling in the spectral window. The major advantage with the use of PLS1 is that each component can be examined independently, since the optimal conditions for determining each component are different in most situations.

Once multiple spectral intervals are selected, two strategies are suggested for PLS modeling of the calibration equation using the selected spectral intervals. One is to include all the selected spectral intervals and develop a PLS model using the selected intervals. The other is to separately build individual PLS models in each interval and construct a linear combination of all the separate PLS models for prediction. The second strategy is described in the subsequent section.

Combination of Multiple PLS Models in Spectral Interval Selection. If two or more spectral intervals are selected by MWPLSR, multiple PLS calibration models can be obtained by developing a model in each interval. Suppose that there are *J* spectral intervals selected and *J* PLS models are established as follows:

$$\mathbf{c} = \mathbf{R}_j\mathbf{b}_j + \mathbf{e}_j \qquad j = 1, 2, ..., J \tag{10}$$

where $\mathbf{b}_j$ is the estimate of regression coefficients with suitable PLS components and $\mathbf{e}_j$ is the model error for the PLS model in the *j*th spectral interval. In model combinations of these PLS models for prediction of concentrations in unknown samples, the calibration model is computed as a certain linear combination of the *J* PLS models. That is,

$$\mathbf{y} = \sum_{j=1}^{J} w_j\mathbf{R}_j\mathbf{b}_j \tag{11}$$

where w_j is the combination weight ($j = 1, 2, ..., J$). Apparently, the combination weights can be determined directly using least-squares regression by minimizing $||\mathbf{c} - \mathbf{y}||^2$. However, as each PLS model is built as the estimate of the concentrations **c**, the sum of the combination weights tends to approach 1. If some weights have negative values, the other weights may have a value larger than 1. On the other hand, since each PLS model is obtained using different spectral intervals, it is expected that the errors in $\mathbf{R}_j\mathbf{b}_j$ ($j = 1, 2, ..., J$) are independent of each other. Then, the error in the combination model is the weighted sum of the errors in each model with the weights being the squares of the combination weights. Therefore, if there are combination weights larger than 1, the error in the corresponding model will be inflated. This is an undesirable property for the model combination. To circumvent the problem, it is necessary to put a certain constraint on the combination weights. A straightforward constraint is that the weights are restricted to the domain [0, 1]. However, since the sum of the combination weights tends to approach 1, it is enough to restrict the combination weights to being nonnegative. Therefore, the combination problem of multiple PLS models can be formulated as minimizing the sum of squared deviations between the actual concentrations and the combined model subjected to the constraint that the combination weights are nonnegative. That is,

$$\min \quad ||\mathbf{c} - \sum_{j=1}^{J} w_j\mathbf{R}_j\mathbf{b}_j||^2 \tag{12}$$

$$\text{subject to} \quad w_j \geq 0 \ (j = 1, 2, ..., J)$$

The solution of the combination weights can be easily approached using a nonnegative least-squares algorithm.

It is noteworthy that the aforementioned procedure to determine the combination weights is based on the fact that the models to be combined are all constructed properly; that is, each model is built to yield optimized prediction. In cases where some models are not properly built, it is better to utilize an MSEP-based loss function for the determination of combination weights. A common resort is the cross-validation PRESS (sum of squared residues in prediction) or the MSEP on an additional validation set, and it can be immediately implemented by replacing **c** and $\mathbf{R}_j$ ($j = 1, 2, ..., J$) by their counterparts in the validation set. In the present study, the PLS models in different spectral intervals are all properly constructed based on a representative calibration set, so the aforementioned procedure can be employed directly.

EXPERIMENTAL SECTION

OP/FT-IR Data. These data have been reported in detail previously.[24] Two OP/FT-IR data sets were synthesized by adding reference spectra to experimentally measured open-path background spectra. The open-path background spectra were measured in the range from 700 to 3000 cm^{-1} with a Bomem MB-104 spectrometer over several weeks under a variety of conditions (temperature, humidity, path lengths). In each measurement, two single-beam spectra were obtained with the same nominal path length and the ratios determined and converted to absorbance. In data A, the interferograms for the background were measured at a 1-cm^{-1} spectral resolution and processed using medium Norton−Beer apodization, while in data B, the background spectra resolution were obtained at an 8 cm^{-1}. All background spectra were corrected according to the procedures outlined in U.S. EPA Method TO-16 that addresses OP/FT-IR measurements.[29]

The reference spectra of 100 samples were generated using the pure spectra measured for five pure compounds, i.e., methanol, ethanol, 1-propanol, and 1-butanol, and 2-propanol. The strongest absorption peak for each pure spectrum was first scaled to a concentration of 0.3 absorbance unit and assigned a value of 1 arbitrary concentration unit (ACU).[30] The concentrations of the 5 components in 100 samples were created using random numbers ranging from 0 to 1. Then the reference spectra of these samples were generated exactly in terms of Beer's law using the scaled pure spectra and the concentrations matrix. The final response spectra were then synthesized by adding the reference absorbance spectra to the real open-path background absorbance spectra. The reference spectra were manipulated as interferograms using medium Norton−Beer apodization and truncated to 1-cm^{-1} resolution in data A and to 8-cm^{-1} resolution in data B.

The 100-sample data were split into calibration and prediction sets for modeling the concentration of ethanol in the mixtures.

[29] Compendium Method TO-16 Long-Path Open-path Fourier Transform Infrared Monitoring of Atmospheric Gases, EPA/635/R-96/010b, U.S. Environmental Protection Agency, Research Triangle Park, NC, 1999.

[30] Anderson, R. L.; Griffiths, P. R. *Anal. Chem.* **1975**, *47*, 2339.

The calibration set that was composed of 75 samples was used to build the models for prediction using full-spectrum PLS or the proposed method. The prediction set that comprises the remaining 25 samples was used to evaluate the behavior of the models.

NIR Data. The samples were prepared gravimetrically and composed of four components including ethanol (EtOH, p.a. 99.8%, Roth), 1-butanol (BuOH, p.a. 99.5%, Roth), ethyl acetate (EtOAc, p. a. 99.5%, Baker) and *n*-butyl acetate (BuOAc, technical grade, 98%, Riedel-de-Haen). The concentration range of the samples was chosen according to the situation of monitoring the reaction

$$\mathrm{EtOAc} + \mathrm{BuOH} \xrightarrow{t\text{−BuOK}} \mathrm{BuOAc} + \mathrm{EtOH}$$

where the catalyst, potassium *tert*-butylate, was not added such that stable mixtures could be obtained.

The NIR spectra were recorded on a Foss 6500 spectrometer using the transflection mode. The sample was positioned in a quartz vessel with a gold-coated reflector (0.5-mm layer thickness corresponding to ~1-mm sample thickness) on the quartz window that is illuminated from below. The radiation transflected from the sample was collected by four PbS detectors positioned under the quartz window with an inclination angle of 45°. Each sample was measured in duplicate by rotating the quartz vessel by 45° between the recording of spectra, and 32 scans were accumulated for one spectrum in the wavelength range from 1100 to 2498 nm with an interval of 2 nm. The spectral resolution was 10 nm at 1600 nm. A ceramic plate was used as reference, and NIR spectra of the sample were measured at 22 °C.

The 37-sample data were split into calibration and prediction sets for modeling the concentration of EtOAc in the mixtures. The calibration set was composed of 19 samples randomly chosen from the whole set, while the prediction set comprised the remaining 18 samples.

Throughout the present study, the window size for MWPLSR is set to 20 spectral points. It was found that the window size had no significant effect on the residue lines obtained, provided the window size was larger than the desired model dimensionality and smaller than the spectral intervals to be sought for. For simplicity of comparison, the dimensionality for the PLS model constructed using certain spectral regions was determined to be the number where the SSR value begins to decrease insignificantly with the increase of model dimensionality. It was also checked in the present study that, with a representative calibration set, this procedure gave model dimensionalities consistent with those determined by the validation methods.

RESULTS AND DISCUSSION

OP/FT-IR Data A. The spectra of the OP/FT-IR data A are shown in Figure 1. The measurement errors in the data mainly arise from the instrumental noise in the background spectra. One can see a number of "spikes" over the whole spectral region. These spikes are due to detector error and strong background absorption. Over the long path length, the background components, water vapor and CO_2 in the atmosphere, have very strong absorption in the spectral ranges of 1200−2000 and 2250−2400 cm^{-1}, and the absorbances in these spectral ranges are extremely large. As a result, when the ratios of these spectra (that have very strong intensities) are determined, two very small detector errors will be greatly amplified in these spectral ranges. These spikes are unavoidable at the 1-cm^{-1} resolution. A decrease in the spectral resolution may eliminate most of the spikes, but it may also exclude some analytical information as well. For OP/FT-IR field operation and calibration, the U.S. EPA currently specifies a 1-cm^{-1} resolution for common use.[29]

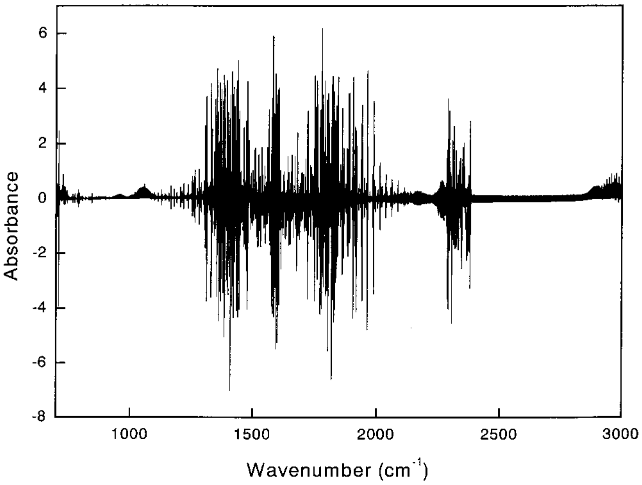

Figure 1. OP/FT-IR spectra obtained at 1-cm^{-1} resolution in the range of 700−3000 cm^{-1} for 100 mixture samples of methanol, ethanol, 1-propanol, 1-butanol, and 2-propanol.

Because the OP/FT-IR spectra in the ranges of 1200−2000 and 2250−2400 cm^{-1} are dominated by the measurement errors and the samples themselves do not have absorption in the range of 2000−2250 cm^{-1}, one may well utilize the spectral regions of 700−1200 and 2400−3000 cm^{-1} for the quantification of ethanol in the mixtures. As a matter of fact, it was found that, even though the whole spectral range was included in the spectral interval selection, the proposed approach is still capable of indicating that the spectral range of 1200−2400 cm^{-1} is uninformative and should be excluded in the modeling. To keep consistency with the implementation of OP/FT-IR analysis, we will only focus on the treatment of the spectral ranges of 700−1200 and 2400−3000 cm^{-1}.

The first 20 residue lines obtained by MWPLSR for OP/FT-IR data A in the spectral ranges of 700−1200 and 2400−3000 cm^{-1} are depicted in Figure 2a and d, respectively. As excessive components (20 components) are used for the PLS model, the residue lines indicate the achievable error level (represented by the SSR value) for the prediction of ethanol is ~10^{-3}. In Figure 2a, there are two spectral regions with which the built PLS models reach the error level. The residue lines in these two regions are replotted at an amplified scale in Figure 2b and c. One can observe that the window-based PLS models attain to the error level with five components when the window is positioned in the spectral interval of 719−738 cm^{-1} (Note that the right boundary of the interval can be extended backward by 19 spectral points, the size of the window minus 1.), and introduction of more components does not improve significantly the fitness of the model. This indicates that the spectral interval of 719−738 cm^{-1} may be informative for modeling the concentration of ethanol. Likewise, one can ascertain from Figure 2c that the spectral interval of 1013−1029 cm^{-1} may be useful for the modeling. In Figure 2c, there are two spectral intervals around 1039 and 1050 cm^{-1} with

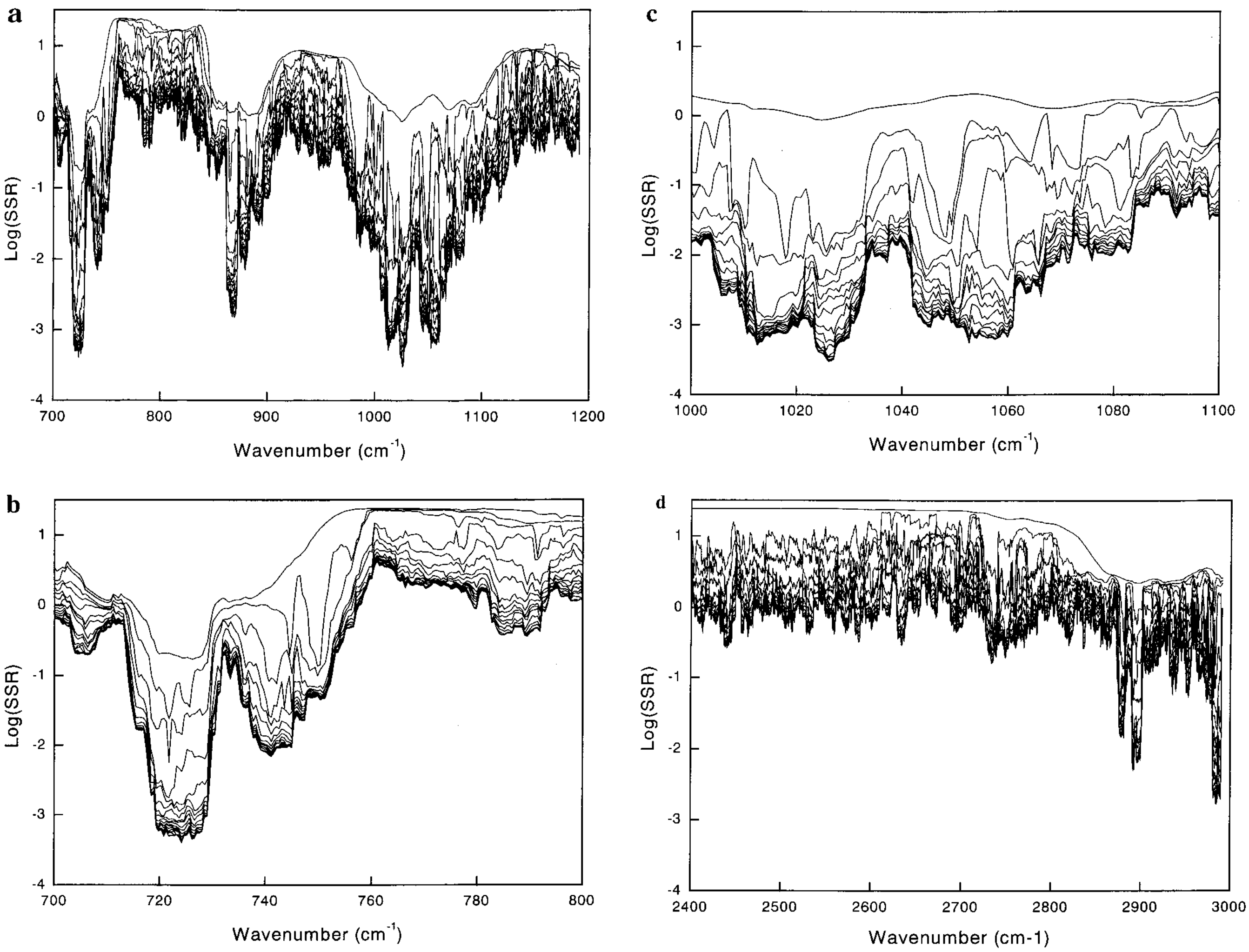

Figure 2. Residue lines obtained by MWPLSR of the OP/FT-IR spectra A for the calibration samples. The residue lines in the range of (a) 700−1200, (b) 700−800, (c) 1000−1100, and (d) 2400−3000 cm^{-1}.

which the built PLS models can also achieve the error level. However, these two intervals are excluded due to the fact that the residue lines continue to lower down significantly until more than 10 components were exploited in the PLS models, indicative of extra uncertainty in these spectral intervals. In Figure 2d, one sees that when the window is in the region of 2400−3000cm^{-1}, the residue lines for the 20-component PLS models are still significantly higher than the achievable error level. This implies that the spectra of the samples in the spectral region are severely contaminated by the instrumental noise and the resulting spectra show substantial deviation from the ideal linear response. To model the concentration of ethanol using such noise-distorted spectra, PLS has to exploit more components to compensate for the model deviations, and the fitness of the PLS model to the data cannot reach a desirable level, provided the spectral window is not large enough. Therefore, it was ascertained that the spectral region of 2400−3000 cm^{-1} was uninformative and could be eliminated in the modeling. Based on the above findings, it is clear that for the OP/FT-IR data A only two small spectral intervals, 719−738 and 1013−1029 cm^{-1}, are not contaminated substantially by the instrumental noise and can be useful for modeling of the concentration of ethanol. These two spectral intervals were then selected in subsequent calibration and prediction.

The results of PLS modeling for the OP/FT-IR data A using full-spectrum or selected spectral intervals are shown in Table 1. It can be seen that the best spectral interval located by the iPLS algorithm was very close to one of the informative spectral windows given by the proposed method. Moreover, in terms of the root-mean-squared error in prediction (RMSEP), the PLS models based on the selected spectral intervals all give much better performance than the full-spectrum-based PLS models. This confirms the conclusion that the performance of PLS can still be substantially improved by selecting proper spectral regions. Since the absorption bands arising from the C−H stretching mode of ethanol are very similar to those of other alcohols in the mixtures, the spectral region of 2400−3000 cm^{-1} contains less unique information concerning ethanol relative to the spectral region of 700−1200 cm^{-1}, where the band due to the O−H deformation mode is expected. That is to say, the length of NAS of ethanol in the region of 2400−3000 cm^{-1} is smaller than that in the region of 700−1200 cm^{-1}. Then, the theoretically achievable prediction error for the model built on the region of 2400−3000 cm^{-1} is larger than that in the region of 700−1200 cm^{-1}. As a consequence, the full-spectrum PLS model constructed using the spectral region of 2400−3000 cm^{-1} gives the worst prediction in all the models. Moreover, as the weak signal in this region is severely distorted

Table 1. Results of PLS Modeling of OP/IR Data A for Ethanol Using Given Spectral Regions[a]

	PLS	PLS	PLS	PLS^w	PLS^w	PLS^w	PLS$_{com}$	iPLS
spectral region (cm^{-1})	700–1200 2400–3000	700–1200	2400–3000	719–738	1013–1030	719–738 1013–1030	719–738 1013–1030	1012–1030 (1019–1029)
model dimensionality	9	9	9	5	5	6		5
RMSEP	0.0101	0.0092	0.0118	0.0052	0.0043	0.0050	0.0038	0.0043 (0.0046)

[a] PLS, full-spectrum PLS modeling with the given spectral region; PLS^w, PLS modeling with the selected spectral intervals; PLS$_{com}$, combination of PLS models in the selected spectral intervals. The combination weights computed are 0.5315 and 0.4685, respectively, for models in the spectral intervals of 719–738 and 1013–1030 cm^{-1}; iPLS, the spectral interval and the corresponding RMSEP given by the iPLS algorithm. The values in parentheses are those obtained directly by equidistant iPLS, and the values not in parentheses are those obtained after optimizing the selected interval from equidistant iPLS.

by the instrumental noise, the PLS model including both this region and the region of 700–1200 cm^{-1} also yields a slightly deteriorated performance in comparison to the PLS model built only on the spectral region of 700–1200 cm^{-1}.

One can also see that, to construct a proper model for prediction, the full-spectrum PLS calibration requires four more components than the PLS calibration using the selected intervals. This implies that the inclusion of uninformative spectral regions will cause additional variability in the model. Because in practice such extra variability is always generated by complicated baselines or nonlinearity, it cannot be completely accounted for by the calibration samples, thereby introducing increased uncertainty into the model. Therefore, the full-spectrum-based PLS models are less stable than the PLS models based on the selected spectral intervals. Surprisingly, the PLS model using both spectral intervals of 719–738 and 1013–1030 cm^{-1} shows a little inferior performance compared to the PLS model based on the spectral interval of 1013–1030 cm^{-1}, while the combination of two PLS models separately built in two selected intervals give the best performance in the prediction. This might be due to the fact that the noise characteristic in the region of 719–738 cm^{-1} is different from that in the spectral interval of 1013–1030 cm^{-1}, and the noises from these two selected intervals are accumulated. As a matter of fact, the accumulation of noise in these two regions is indicated by the fact that, in the construction of the PLS model, one more component is exploited to account for the noise. The above finding also indicates that, in such situations, a better choice may be the combination of the PLS models separately built on the selected spectral intervals, since it avoids the accumulation of noise via modeling on individual spectral intervals and the difference in model errors can be technically handled using varying combination weights.

OP/FT-IR Data B. The OP/FT-IR data B are shown in Figure 3. With decreased resolution, the instrumental errors of spikes merely appear in the absorption regions of water and CO_2, while the spikes are nearly eliminated in the spectral ranges of 700–1200 and 2400–3000 cm^{-1}. Two absorption bands are observed clearly around 1100 and 2900 cm^{-1}. This suggests an improved signal-to-noise ratio for the analyte, ethanol.

The residue lines obtained by MWPLSR for the OP/FT-IR data in the spectral ranges of 700–1200 and 2400–3000 cm^{-1} are depicted in Figure 4a and b, respectively. These plots indicate that the achievable error level (represented by the SSR value) for the PLS model is approaching 10^{-4}. It is observed in Figure 2a that when the window is located in the spectra interval of 975–1130 cm^{-1}, the PLS models built in the window attain to the error level. Then the spectral interval of 975–1130 cm^{-1} was selected and included in the calibration. One also sees the spectral interval of 805–959 cm^{-1} in which the window-based PLS models give an error level of 10^{-3}. As the error level is usually desirable in multicomponent spectral analysis, the spectral interval was also included in calibration such that the effect of spectral intervals of different error levels could be examined. It is also observed that, to reach the error level, increased model dimensionality has been used in the modeling with these two spectral intervals. Based on the above findings, one can conclude that, with lower wavenumber resolution, better signal-to-noise ratios are expected to be achievable and the useful spectral intervals may be extended; however, the deviation of the model from the ideal linearity is also inclined to increase. Similarly, it can be identified from Figure 4b that the spectral interval of 2826–2965 cm^{-1} is an informative region for the calibration. This spectral interval was then selected for the PLS calibration.

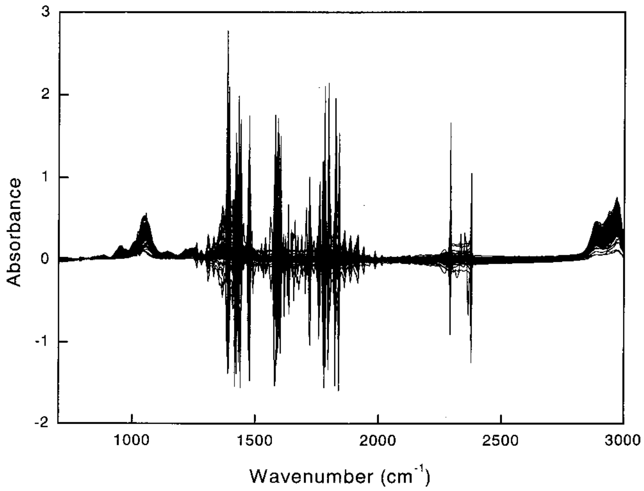

Figure 3. OP/FT-IR spectra obtained at 8-cm^{-1} resolution in the range of 700–3000 cm^{-1} for 100 mixture samples of methanol, ethanol, 1-propanol, 1-butanol, and 2-propanol.

The results of PLS calibration for the OP/FT-IR data using full-spectrum or selected spectral intervals are shown in Table 2. One can see that the model dimensionalities for all the optimal PLS models are larger than the number of species present in the samples. This is also due to the fact that the spectral responses at relatively low spectral resolution may exhibit a slight discrepancy from the ideal linear model. It is also observed that the best prediction is achieved for PLS modeling using two selected spectral intervals, 975–1130 and 2826–2965 cm^{-1}, which verifies

Table 2. Results of PLS Modeling of OP/IR Data B for Ethanol Using Given Spectral Regions

	PLS	PLS	PLS	PLS^w	PLS^w	PLS^w	PLS^w	PLS^w	PLS$_{com}$[a]	PLS$_{com}$[b]	iPLS
spectral region (cm^{-1})	700−1200 2400−3000	700−1200	2400−3000	975−1130	2826−2965	805−959	975−1130 2826−2965	975−1130 2826−2965 805−959	975−1130 2826−2965	975−1130 2826−2965 805−959	1006−1103 (1049−1083)
model dimensionality	10	10	10	8	10	9	10	10			8
RMSEP	0.0047	0.0083	0.0023	0.0022	0.0019	0.0085	0.0012	0.0036	0.0015	0.0015	0.0029 (0.0043)

[a] The combination weights computed are 0.8557 and 0.1443, respectively, for models in the spectral intervals of 975−1130 and 2826−2965 cm^{-1}.
[b] The combination weights computed are 0.8436, 0.1397, and 0.0167, respectively, for models in the spectral intervals of 975−1130, 2826−2965, and 805−959 cm^{-1}.

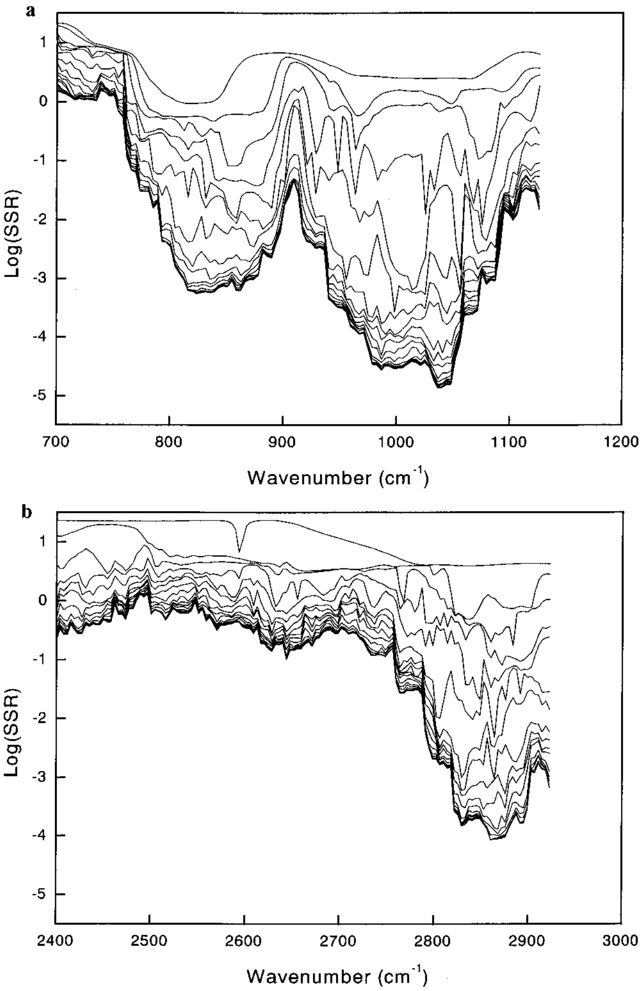

Figure 4. Residue lines obtained by MWPLSR of the OP/FT-IR spectra B for the calibration samples. The residue lines in the range of (a) 700−1200 and (b) 2400−3000 cm^{-1}.

again the fact that the performance of PLS can still benefit from spectral interval selection. The spectra in the range of 700−1200 cm^{-1} are contaminated by the instrumental noise, as is indicated by the fact that the PLS model based on the selected spectral interval, 975−1130 cm^{-1}, exhibits much better performance than the full-spectrum model in the range and the PLS model built on the spectral interval of 805−959 cm^{-1} gives the worst prediction among all the models. Then, it is clear that the spectral interval of 805−959 cm^{-1} contains the interference from non-composition-related factors. Inclusion of this interval in calibration may introduce undesired uncertainty into the model. Therefore, the two full-spectrum PLS models involving this spectral interval and the model based on three selected intervals show slightly deteriorated performance compared to those excluding this spectral interval. The only exception is the model based on the combination of the PLS models built separately on these three spectral intervals of 975−1130, 2826−2965, and 805−959 cm^{-1}. One can see that the model obtained by the combination of PLS models based on three selected spectral intervals gives the prediction as well as the model resulting from the combination of PLS model built on two useful spectral intervals. In fact, according to the combination weights calculated, these two models obtained by combinations are very close to each other, since the combination weight for the model constructed using the uninformative spectral interval, 805−959 cm^{-1}, is very small and the combination weights for the two models based on the other two selected intervals are approximately the same. This implies that the procedure of model combination is capable of identifying the goodness of each model and determining appropriate combination weights for the models in light of the goodness of the model. As a result, the proposed procedure for combining multiple models may be a robust approach to exploit the information comprised in individual models, even though some individual models are constructed improperly. Interestingly, it is observed that iPLS also gives an informative spectral window. However, this spectral window is narrower than the corresponding one located by the MWPLSR procedure. This is an indication that the postoptimization of the spectral window from equidistant iPLS may get trapped into local optimums, suggesting the MWPLSR algorithm is more robust to local optimums than iPLS in ascertaining the best spectral intervals. In addition, one can see that the spectral interval found by iPLS shows slightly worse prediction performance than the spectral region of 2826−2965 cm^{-1}. This is due to the fact that iPLS uses the same model dimensionality to explore the best spectral intervals. As a result, the algorithm may be rather sensitive to the choice of model dimensionality, and then increasing the risk of missing the optimal spectral window associated with slightly increased model complexity and yielding a suboptimal spectral interval.

It is noted that the prediction errors of the PLS model based on the full spectral range of 2400−3000 cm^{-1} is desirably small in this case, which is comparable with that of the model based on the selected interval, 2826−2965 cm^{-1}. This is due to the fact that uninformative spectral regions in the range (2400−2826 and

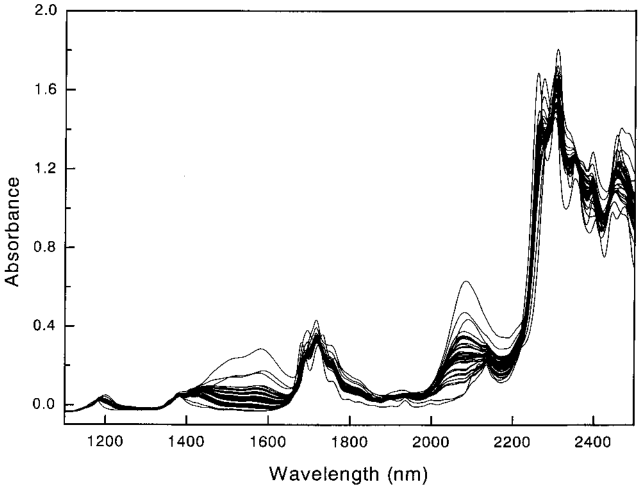

Figure 5. NIR spectra obtained in the range of 1100−2500 nm for 37 mixture samples of ethanol, ethyl acetate, 1-butanol, and *n*-butyl acetate.

2965−3000 cm^{-1}) have only very weak absorption or show strong correlation to the informative region.

NIR Data. The spectra of the NIR data are shown in Figure 5. The NIR spectra of the samples are mainly attributed to the overtone or combination bands of O−H and C−H groups, which are very sensitive to the compositional variations in the samples. With varying compositions of the liquid samples, the interaction between different groups may slightly change and the NIR spectra may deviate to a certain degree from the ideal linearity, which constitutes the major source of errors in the measurements. Due to the possible deviation of the NIR spectra from the ideal linearity, it is expected that a model dimensionality more than the number of components will be used in the PLS modeling.

The residue lines obtained by MWPLSR for the NIR data in the whole spectral range of 1100−2498 nm are depicted in Figure 6a. One can see that the achievable error level (represented by the SSR value) for the window-based PLS models is ~10^{-2}, and there are three spectral intervals in which the window-based PLS models approach the error level. The residue lines in these three regions are replotted in Figure 6b−d separately to explore the fine details. It is ascertained in Figure 6b that the spectral interval between 1100 and 1218 nm is a useful region for the calibration, and six latent variables are needed to build a PLS model reaching the desired error level. This band arises from the second overtone of the C−H stretching vibration. Figure 6c reveals that the spectral interval of 1626−1858 nm is an informative region for modeling the concentration of EtOAc, and the PLS model that gives the desired error level has a model dimensionality of 7. As the bands in the 1100−1218- and 1626−1858-nm regions originate from the

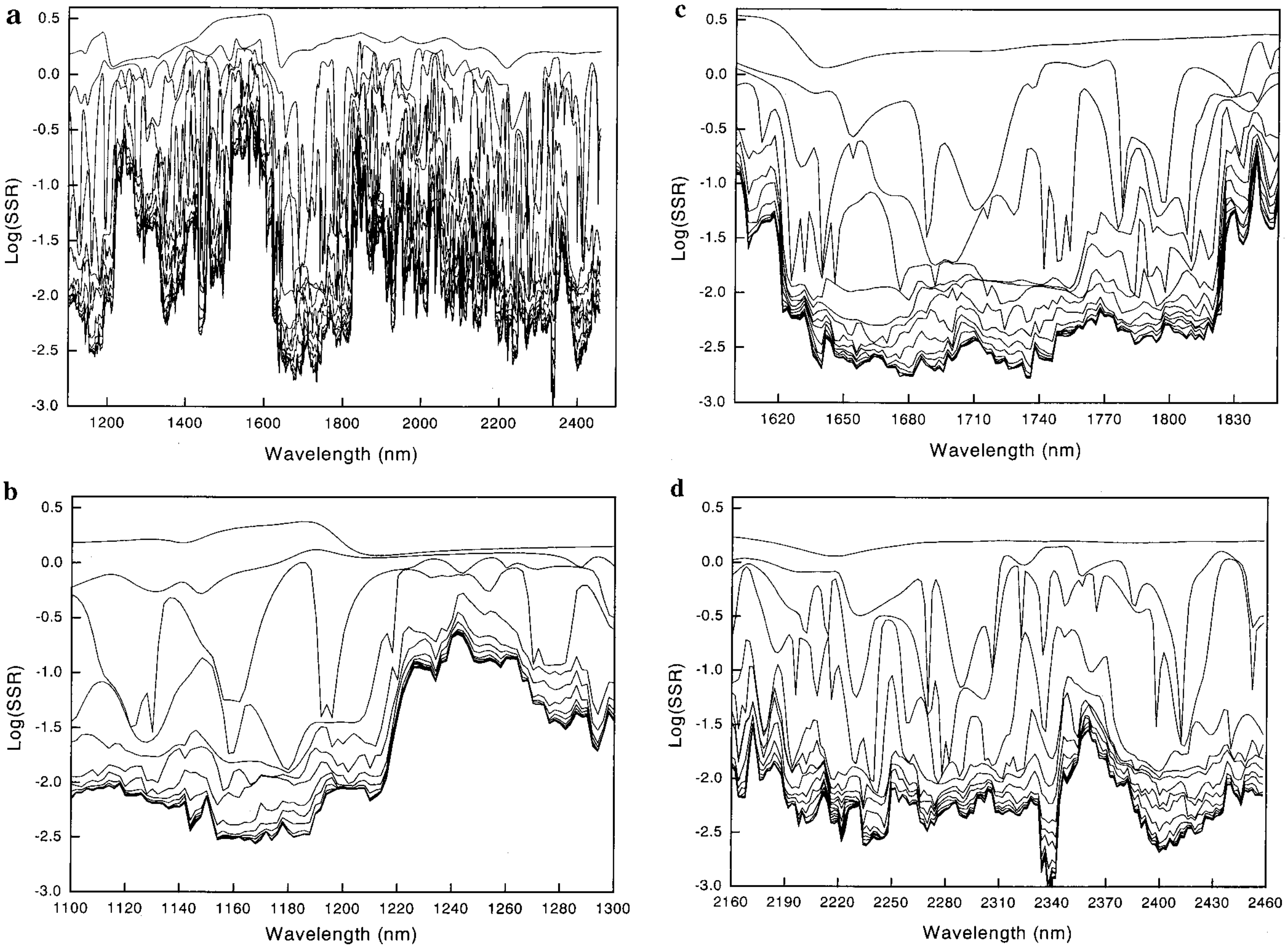

Figure 6. Residue lines obtained by MWPLSR of the NIR spectra for the calibration samples. The residue lines in the range of (a) 1100−2500, (b) 1600−1850, and (c) 2160−2460 nm.

Table 3. Results of PLS Modeling of NIR Data for EtOAc Using Given Spectral Regions

	PLS	PLS^{w}	PLS^{w}	PLS^{w}	PLS^{w}	PLS^{w}	PLS$_{com}$[a]	iPLS
spectral region (nm)	1100−2498	1100−1218	1626−1858	2216−2376	2388−2498	1100−1218 1626−1858 2216−2376 2388−2498	1100−1218 1626−1858 2216−2376 2388−2498	1656−1750 (1660−1698)
model dimensionality	7	6	6	7	4	7		6
RMSEP	0.0394	0.0212	0.0179	0.0165	0.0295	0.0172	0.0157	0.0190 (0.0232)

[a] The combination weights computed are 0, 0.1277, 0.8725, and 0, respectively, for models in the spectral intervals of 1100−1218, 1626−1858, 2216−2376, and 2388−2498 cm^{-1}.

second and the first overtones, respectively, of the C−H stretching mode, the above findings may be the indication that the C−H stretching mode is susceptible to the effect of interaction between the constituents in the samples, and consequently, the C−H stretching bands may deviate from the ideal linearity. Then additional components have to be exploited in the PLS model to address the compositional variations in the samples. From Figure 6d, two spectral intervals, 2216−2376 and 2388−2498 nm, can be identified as the informative regions. The bands in the 2216−2376-nm region are attributed to the combinations of C−H vibrations. One can see that in the spectral region the residue lines continue to lower down significantly until more than four components are used for the PLS model. This is also due to the fact that the band due to C−H combination bands is relatively prone to being affected by the compositional changes in the samples. The band of 2388−2498 nm may arise from the second overtone of the C−H deformation mode. It is noticed that, when the window is positioned in the spectral range, the window-based PLS models approach the desired error level with only four components. This may be an indication that the bands due to the C−H deformation mode are relative stable to the compositional variations in the samples. Based on the above finding, four spectral intervals were selected and PLS models were then constructed using the selected intervals for the quantification of EtOAc in the samples.

It is noteworthy that two meaningful regions, 1400−1600 and 2000−2150 nm, where bands due to the first overtone of the O−H stretching mode and those assigned to the O−H combination mode are expected to appear, respectively, are identified as uninformative regions by MWPLSR. This observation is self-evident, since the analyte, EtOAc, does not have the O−H group.

The results of various PLS models built on the full spectra or the selected spectral intervals are summarized in Table 3. A direct PLS modeling on all the selected spectral intervals gives a RMSEP of 0.0172, which is much better than the full-spectrum-based PLS model. The performance of the PLS models built separately on four selected spectral intervals varies according to the spectral interval used. Better prediction is achieved for the PLS models based on the spectral intervals of 1626−1858 and 2216−2376 nm, while the PLS models constructed using the other two intervals, 1100−1218 and 2388−2498 nm, give inferior performance. This suggests that the interference of random noise in the spectral intervals of 1100−1218 and 2388−2498 nm may be relatively large. Actually, in the combination of the PLS models built in the four selected intervals, the combination weights for the models in the regions of 1100−1218 and 2388−2498 nm are zero, which implies that the two spectral intervals of 1100−1218 and 2388−2498 nm, compared to the other selected regions, are less informative. In addition, this demonstrates the built-in capacity of the proposed approach of model combination to ascertain the goodness of an individual model and handle improper models using small combination weights. One also observes that the informative spectral window given by iPLS is still narrower than the corresponding one generated by MWPLSR, indicating that the post-optimization step is trapped into a local optimum. Moreover, this spectral window yields a slightly worse prediction than the spectral region of 2216−2376 nm. This confirms the conclusion that iPLS may miss some better spectral intervals associated with a slightly increased model dimensionality.

CONCLUSIONS

The present study has demonstrated theoretically that deteriorated performance may be induced by the inclusion of uninformative spectral regions in multicomponent spectral analysis. The uninformative regions are typically characterized by the increased model complexity in the PLS model that is built on these regions. A new spectral interval selection method, MWPLSR, has been proposed for the selection of informative spectral intervals. Once multiple spectral intervals are selected, a novel modeling approach, combination of multiple PLS models, has been developed. The results show that, with the elimination of uninformative spectral regions using the proposed spectral interval selection method, the performance of PLS calibration can still be significantly improved. It is also disclosed that the model combination approach provides a robust way to exploit the information comprised in individual models, as it can avoid possible accumulation of errors in multiple spectral intervals and can handle the models of varying goodness via appropriate combination weights. The model combination approach holds immense potential in chemometric modeling and is expected to find novel implementations in other studies.

Received for review November 14, 2001. Accepted April 10, 2002.

AC011177U