Iterative Optimization Technology Combined with Wavelength Selection Based on Excess Absorption for a Process
Analytical Technology Calibration–Minimum Approach

Shojiro Shibayama,¹ Hiromasa Kaneko,¹ and Kimito Funatsu^{*,1}

¹*Department of Chemical System Engineering, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan*

* Corresponding author: Tel.: +81 3 5841 7751; Fax.: +81 3 5841 7771.

E-mail: funatsu@chemsys.t.u-tokyo.ac.jp.

Abstract

Process analytical technology (PAT) plays an important role in the pharmaceutical industry. Calibration-free/minimum methods in PAT are expected to aid in a deeper understanding of processes in the early development stage of new drugs. Iterative optimization technology (IOT), an existing calibration-free method, is not able to predict the compositions of nonideal mixtures because the Beer–Lambert law does not hold in some wavelength regions. In this paper, we propose IOT with wavelength selection based on excess absorption (WLSEA), which is available with at least one calibration sample. Excess absorption (EA) is the residual between the measured and ideal spectra of a mixture, and includes noise and spectral change related to molecular interactions. WLSEA determines a threshold of EA that separates noise and spectral change by

minimizing prediction errors of IOT. Consequently, WLSEA selects a set of regions where predictive accuracy of IOT is high. WLSEA-IOT can be applied to predict compositions of both ideal and nonideal mixtures [that have ideal regions](#). The performance of the proposed IOT is verified by analyses with three types of mixture spectra. The proposed wavelength selection method will [enhance both development of quantitative methods and analyses of molecular interactions with infrared spectroscopy](#).

Keywords

Process analytical technology

Iterative optimization technology

Wavelength selection

Excess absorption

Calibration-minimum method

Infrared spectroscopy

1. Introduction

Process analytical technology (PAT) plays an important role in the pharmaceutical industry.[1-6] One of the objectives of PAT is to effectively monitor pharmaceutical processes with nondestructive methods, such as Raman, ultraviolet, and infrared spectroscopy. Attempts have been made to realize PAT with statistical methods such as partial least squares[7] (PLS) to construct regression models between the properties of products and intermediates and spectra using many training samples and to estimate properties from spectra. Process monitoring with PLS is applied to established drug manufacturing processes.

Calibration-free/minimum approaches are useful in the early development stage of new drugs. One field of calibration-free approach studies is called self-modeling curve resolution (SMCR).[8-10] SMCR includes evolving factor and window factor analyses. Evolving factor and window factor analyses are concentration calculation methods for chromatographic data. Other types of calibration-free methods used for mixture uniformity monitoring[11,12] use changes in the standard deviations of mixture spectra. However, calibration-free methods cannot accurately predict component concentrations, although component concentrations in mixtures provide a deep understanding of processes. To tackle this problem, a calibration-free method called iterative optimization technology (IOT) has been recently reported.[13] IOT is a type of least squares method that is based on the Beer–Lambert law considering some physical constraints. IOT can be applied to ideal mixtures because the Beer–Lambert law only holds for ideal regions in a mixture spectrum. However, IOT cannot predict mole fractions of nonideal mixtures that are affected by spectral changes related to molecular interactions.

If spectral change could be formulated based on molecular interactions, the mole fractions of nonideal mixtures could be predicted. However, formulation with few calibration samples is difficult. Nonlinear IOT has been proposed for prediction of nonideal mixtures.[13] However, the predictive accuracy of nonlinear IOT largely depends on training samples and the wavelength regions used. IOT with virtual molecular interaction spectra (IOT-VIS) has been developed.[14] It can express spectral change as nonlinear terms and requires few training samples. The predictive ability of IOT-VIS has been verified with mixtures but it requires nonideal mixtures as training samples. Indirect hard modeling (IHM)[15,16] is a hybrid method of mathematical and physical approaches that formulates the peaks in mixture and pure-component spectra. However, IHM requires as many spectra of training samples as there are pure components. Although it can be applied to nonideal mixtures, formulation of the spectral change for calibration-free/minimum methods is difficult.

Wavelength selection is another way to deal with nonideal mixtures and improve the predictive ability of IOT. However, existing wavelength selection methods are not appropriate for calibration-free/minimum approaches. For instance, genetic algorithm-based wavelength selection[17] (GAWLS) selects wavelength regions where the predictive accuracy of IOT is high.[18] Nonetheless, GAWLS requires many mixture spectra to construct a robust model. The SMCR method SIMPLISMA[19,20] selects wavelength regions where the intensity is related to the concentration of each pure component, although it uses many spectra to calculate the standard deviation of the intensities at each wavelength.

Defining applicability domains (ADs), which are data domains where statistical models can produce true predictive performance, is still challenging for calibration-free/minimum methods owing to shortage of

training samples. ADs should be defined for all data-driven modeling to assess the predictive accuracy for each query or new spectrum. Nevertheless, it is difficult to define ADs in calibration-free/minimum prediction because ADs are generally based on the similarity between training and test data.[\[21-24\]](#) The spectral similarity of infrared spectra[\[25\]](#) in test and training samples cannot stably assess the predictive accuracy because of the unknown size of noise in the spectra, small number of training samples, and absence of a proper similarity index.

In this paper, we focus on wavelength selection for two reasons. First, formulation by IOT-VIS requires training samples that have spectral change, which means that the formulation places a restriction on training samples. IOT with wavelength selection will provide high accuracy without constraints on calibration samples because ideal regions are expected to not significantly change. Thus, IOT with wavelength selection will realize highly accurate prediction of nonideal mixtures. Second, wavelength selection may contribute to definition of ADs. ADs are able to be defined in ideal mixtures without any training samples if the ideal mixtures can be simulated from only pure-component spectra. Thus, ADs may be defined even in nonideal mixtures after ideal regions are selected by wavelength selection. Wavelength selection that is suitable for IOT prediction is expected to simultaneously solve the two problems.

One possible key to realize wavelength selection for calibration–minimum prediction is excess absorption.[\[26-28\]](#) Excess absorption is the residual between the mixture spectrum and the linear combination of the mole fraction and spectrum of each pure component. Excess absorption is useful for analyzing spectral change in mixture spectra because of molecular interactions. However, excess absorption studies do not mention regions where the Beer–Lambert law holds, that is, where IOT prediction provides

high accuracy. Excess absorption is expected to aid wavelength selection because excess absorption contains information about spectral change and noise.

The objective of this study is to develop calibration–minimum concentration prediction of nonideal mixtures. We propose IOT with wavelength selection based on excess absorption (WLSEA) to simultaneously improve the predictive accuracy and define the AD. IOT with WLSEA (WLSEA-IOT) predicts mole fractions after selecting regions where a mixture spectrum has a linear relationship with the pure-component spectra. WLSEA-IOT can be performed with at least one calibration sample, which realizes a calibration–minimum prediction. The performance of the proposed method is verified with three types of mixtures.

2. Methods

2.1. IOT with wavelength selection based on excess absorption

IOT is a calibration-free concentration prediction method; it is explained in detail in Appendix A and the literature.[13] IOT minimizes the excess absorption $\mathbf{x}_{\text{excess}}$ given by the following equation:

$$\mathbf{x}_{\text{excess}} = \mathbf{x}_{\text{mix}} - \sum_{i=1}^c r_i \mathbf{x}_{\text{pure},i}, \quad (1)$$

where $\mathbf{x}_{\text{mix}}$, r_i , $\mathbf{x}_{\text{pure},i}$, and c are the mixture spectra, mole fraction of pure component i , spectrum of pure component i , and number of pure components, respectively. IOT prediction of ideal mixtures succeeds because the small variance of excess absorbance provides accurate prediction results. Similarly, the predictive accuracy of IOT will improve if only regions where excess absorbance is sufficiently small are selected, even for nonideal mixtures.

For a mixture spectrum, we assume the existence of a threshold $e_{\text{threshold}}$ to select regions where excess absorbance $x_{\text{excess},d}$ at the d -th wavelength is sufficiently small. For a training sample t , when $e_{\text{threshold}}$ is given, we remove regions where the absolute value of x_{excess} is above $e_{\text{threshold}}$, as shown in Fig. 1(a). The remaining regions $\mathbf{v}_{\text{linear}}$ (shown as the red lines in Fig. 1(a)) are defined as follows:

$$\mathbf{v}_{\text{linear}} = \{d \mid |x_{\text{excess},d}| < e_{\text{threshold}}, d = 1, 2, \dots, v_{\text{max}}\}, \quad (2)$$

where v_{max} is the number of measurement points in the spectrum. The selected regions $\mathbf{v}_{\text{linear}}$ are called linear or ideal regions because $\mathbf{x}_{\text{mix}}$ on $\mathbf{v}_{\text{linear}}$ has a linear relationship with $\mathbf{x}_{\text{pure},i}$ and the Beer–Lambert law holds.

We can assess the $e_{\text{threshold}}$ performance as the prediction error after IOT prediction with $\mathbf{v}_{\text{linear}}$: the smaller the prediction error, the better the $e_{\text{threshold}}$ performance. The optimized $e_{\text{threshold}}$ for a training sample t is obtained by minimizing a prediction error indicator: root mean square error (RMSE) $\mathbf{r}_{\text{train},t}$, as shown in Fig. 1(b). RMSE $\mathbf{r}_{\text{train},t}$ is given by

$$\text{RMSE } \mathbf{r}_{\text{train},t} = \sqrt{\frac{1}{c} \sum_{i=1}^c |r_{\text{obs},i,t} - r_{\text{pred},i,t}|^2}, \quad (3)$$

where $r_{\text{obs},i,t}$ and $r_{\text{pred},i,t}$ are the observed and predicted mole fractions of pure component i of sample t , respectively. The optimum $e_{\text{threshold}}$ is searched for from 0.1% to 10% of the 0.95 quantile of $\mathbf{x}_{\text{mix}}$ (in increments of 0.1%). RMSE $\mathbf{r}_{\text{train},t}$ is a function of $e_{\text{threshold}}$, which is usually a monomodal or bimodal function and in experiments, is convex downward. Therefore, minimization of RMSE $\mathbf{r}_{\text{train},t}$ probably converges. In general, all one hundred values are searched. However, other optimization methods, such as the golden section search, are also available for rapid and effective searching. The selected region $\mathbf{v}_{\text{linear}}$ with optimum $e_{\text{threshold}}$ is the final region $\mathbf{v}_t$ for training sample t .

The intersection of $\mathbf{v}_t$ ($t = 1, 2, \dots, t_{\text{max}}$, where t_{max} is the number of training samples) is the final set of

regions ν used for prediction. This ν selection is called WLSEA. Note that linear regions are not theoretically linear regions because they are selected for highly accurate IOT prediction. For example, regions where large noise exists are removed.

The limitation of WLSEA in principle is not to work on nonideal mixtures that have no ideal regions. The user of this method can determine in the $e_{\text{threshold}}$ optimization step whether WLSEA works properly or not, by checking the minimized RMSE $r_{\text{train},t}$ value. The prediction accuracy of WLSEA-IOT will be bad if RMSE $r_{\text{train},t}$ is not small enough.

(a) Wavelength selection based on $e_{\text{threshold}}$

(b) Optimization of $e_{\text{threshold}}$ by RMSE $r_{\text{train},t}$ minimization

Fig. 1. Concept of wavelength selection based on excess absorption. Wavelength regions where

the absolute excess absorbance is less than $e_{\text{threshold}}$ are selected (red part of the spectrum in (a)). The optimum $e_{\text{threshold}}$ is determined by minimizing a function of $e_{\text{threshold}}$ (RMSE $\mathbf{r}_{\text{train},t}$).

2.2. Wavelength selection refinement

It is possible that a set of selected regions $\mathbf{v}$ has wavelength regions with large noise. Refined regions that help to more easily associate $\mathbf{v}$ with peaks can be obtained by eliminating noise regions. The refinement procedure is shown in Fig. 2. In WLSEA, a set of regions is selected from selected regions $\mathbf{v}$ after determining if each selected region contributes to maintain the IOT predictive ability. Each selected region is evaluated by the length of one region divided by root mean square (RMS) $\mathbf{x}_{\text{excess}}$:

$$\begin{aligned} \text{RMS } \mathbf{x}_{\text{excess}} &= \sqrt{\frac{1}{v_{\text{max}}} \sum_{d'=1}^{v_{\text{max}}} \left| x_{\text{mix},d'} - \sum_{i=1}^c r_{\text{obs},i} x_{\text{pure},i,d'} \right|^2} \\ &= \sqrt{\frac{1}{v_{\text{max}}} \sum_{d'=1}^{v_{\text{max}}} |x_{\text{excess},d'}|^2}, \end{aligned} \quad (4)$$

where v_{max} is the number of measurement points that belong to one region. This indicator can evaluate the contribution of one region to the IOT predictive accuracy. The length of one region and RMS $\mathbf{x}_{\text{excess}}$ correspond to the reliability of the region and noise. For instance, if the length is too short (e.g., 1) the region is obviously noise. Conversely, if RMS $\mathbf{x}_{\text{excess}}$ is too large, the noise of the region is so large that IOT predictive accuracy decreases. Each selected region in $\mathbf{v}$ has an indicator value. Regions are added one by one starting from the one with a highest value of the indicator until RMSE $\mathbf{r}_{\text{train}}$ is below a threshold. RMSE $\mathbf{r}_{\text{train}}$ is given by

$$\text{RMSE } \mathbf{r}_{\text{train}} = \sqrt{\frac{1}{ct} \sum_{t=1}^{t_{\max}} \sum_{i=1}^c |r_{\text{obs},i,t} - r_{\text{pred},i,t}|^2}. \quad (5)$$

In this study, the threshold is set to 0.02.

Fig. 2. Concept of the refinement.

2.3. Evaluation of applicability domains

For any spectrum, the prediction accuracy is not always high in a prediction model. In the case of WLSEA-IOT, linear regions of the test samples may be considerably different from those of the training samples. Therefore, we must first assess the reliability of the prediction results as ADs. The $\text{RMSE } \mathbf{x}_{\text{mix}}$ used for this method is

$$\text{RMSE } \mathbf{x}_{\text{mix}} = \sqrt{\frac{1}{v_{\max} n} \sum_{j=1}^n \sum_{d'=1}^{v_{\max}} |x_{\text{obs},d',j} - \sum_{i=1}^c r_{\text{pred},i,j} x_{\text{pure},d',i}|^2}, \quad (6)$$

where n is the number of samples. Note that $\text{RMSE } \mathbf{x}_{\text{mix}}$ is calculated for $\mathbf{x}_{\text{mix}}$ over the selected $\mathbf{v}$. Thus, $v_{\max}$ is the total length of the selected regions $\mathbf{v}$. $\text{RMSE } \mathbf{x}_{\text{mix}}$ is used to assess $\text{RMSE } \mathbf{r}$, which is expressed as follows:

$$\text{RMSE } \mathbf{r} = \sqrt{\frac{1}{cn} \sum_{i=1}^c \sum_{j=1}^n |r_{\text{obs},i,j} - r_{\text{pred},i,j}|^2}, \quad (7)$$

where $r_{\text{pred},i,j}$ is the predicted mole fraction of pure component i in mixture j . When assessing the reliability of each prediction result, n in eqs 6 and 7 is set to one.

The AD is generally defined from training samples. However, there are only a couple of training samples available in a situation where the proposed method is applied. Therefore, simulated spectra are generated to assess the prediction reliability with RMSE $\mathbf{x}_{\text{mix}}$. Virtual spectra $\mathbf{x}_{\text{mix},\text{sim}}$ are generated using the following equation:

$$\begin{aligned}\mathbf{x}_{\text{mix},\text{sim}} &= \sum_{i=1}^c r_i \mathbf{x}_{\text{pure},i} + \mathbf{e} \text{ at } \mathbf{v}, \\ e_{d'} &\sim N(0, \sigma^2) \quad d' = 1, 2, \dots, v_{\text{max}},\end{aligned}\tag{8}$$

where $\mathbf{e}$, $e_{d'}$, and v_{max} are a noise vector, noise that follows a normal distribution at wavelength d' , and the number of elements in $\mathbf{v}$. $\mathbf{x}_{\text{mix},\text{sim}}$ is assumed to be an ideal mixture spectrum. If $\mathbf{x}_{\text{mix}}$ on $\mathbf{v}$ is completely linear, the predictive accuracy for $\mathbf{x}_{\text{mix}}$ coincides with that for $\mathbf{x}_{\text{mix},\text{sim}}$. Unlike the RMSE r of $\mathbf{x}_{\text{mix}}$, the RMSE r of $\mathbf{x}_{\text{mix},\text{sim}}$ can be calculated because the mole fraction of $\mathbf{x}_{\text{mix},\text{sim}}$ is known.

In this study, σ^2 in eq 8 is equal to the standard deviation of the excess absorbance $\mathbf{x}_{\text{excess},t}$ of training sample t multiplied by 1 to 20. r_i in eq 8 is set to the mole fraction of training sample t . One hundred $\mathbf{x}_{\text{mix},\text{sim}}$ are generated for each set of r_i and σ^2 . For example, if two training samples are used, 4000 $\mathbf{x}_{\text{mix},\text{sim}}$ spectra per set of training samples will be obtained. Note that the number of times $\mathbf{x}_{\text{mix},\text{sim}}$ is generated and the number of σ^2 candidates are arbitrary. Therefore, these numbers can be changed for further analyses.

2.4. Outline of WLSEA-IOT

The WLSEA-IOT procedure is shown in Fig. 3. The product set of $\mathbf{v}_t$ is selected regions $\mathbf{v}$, which can be used for IOT prediction either without or after the refinement step. The error in the IOT prediction is estimated at the end of the procedure.

Fig. 3. Outline of WLSEA-IOT.

3. Materials

3.1. Simulation data

To validate the performance and investigate the limitations of the proposed method, five-component mixtures were simulated. The simulated mixtures were generated from convolution of Gaussian and Lorentzian functions, which is known in curve fitting studies as the pseudo-Voigt function.^[29,30] The pseudo-Voigt function V is expressed as follows:

$$V(\nu, \omega_p, \gamma_p, \alpha_p, \beta_p) = \beta_p \alpha_p \exp\left(-4 \ln 2 \frac{(\nu - \omega_p)^2}{\gamma_p^2}\right) + (1 - \beta_p) \alpha_p \frac{\gamma_p^2}{(\nu - \omega_p)^2 + \gamma_p^2}, \quad (9)$$

where ν , ω_p , γ_p , α_p , and β_p are the wavelength, peak location, peak width, peak height, and Gaussian–Lorentzian weight of peak p , respectively. Five pure-component spectra were generated with the parameters shown in Table 1. Each spectrum has two separate peaks. The generated pure-component spectra are shown in Fig. 4. Subsequently, seven mixture spectra were generated based on $\mathbf{x}_{\text{pure},i}$. Three pseudo-Voigt functions were added as spectral change because of molecular interactions, as shown in Table 2. The sum of the terms in Table 2 is $\mathbf{x}_{\text{mix}}$. The peak height of the spectral change is correlated with the mole fraction of the

first, second, and fourth pure components. The mole fractions of the first to fourth pure components were randomly generated. Each of the mole fractions has a normal distribution with mean and variance of 0.125 and 0.041^2 , respectively. The mole fraction of the fifth pure component was the remainder (sum of mole fractions is one). Shapes of two simulated mixture spectra are shown in Fig. 5.

Table 1. Equations of the simulated pure-component spectra.

component	equation
1 st component	$x_{\text{pure},1} = V(\nu, 800, 10, 1, 0.5) + V(\nu, 1800, 10, 1, 0.5)$
2 nd component	$x_{\text{pure},2} = V(\nu, 1000, 10, 1, 0.5) + V(\nu, 2000, 10, 1, 0.5)$
3 rd component	$x_{\text{pure},3} = V(\nu, 1200, 10, 1, 0.5) + V(\nu, 2200, 10, 1, 0.5)$
4 th component	$x_{\text{pure},4} = V(\nu, 1400, 10, 1, 0.5) + V(\nu, 2400, 10, 1, 0.5)$
5 th component	$x_{\text{pure},5} = V(\nu, 1600, 10, 1, 0.5) + V(\nu, 2600, 10, 1, 0.5)$

Fig. 4. Shapes of the simulated pure-component spectra.

Table 2. Terms of the partial simulated mixture spectra.

component	term
1 st component	$r_1 \mathbf{x}_{\text{pure},1} + V(\nu, 800, 10, r_1, 0.5)$
2 nd component	$r_2 \mathbf{x}_{\text{pure},2} + V(\nu, 1000, 10, r_2^2, 0.5)$
3 rd component	$r_3 \mathbf{x}_{\text{pure},3}$
4 th component	$r_4 \mathbf{x}_{\text{pure},4} + V(\nu, 1400, 10, r_4, 0.5)$
5 th component	$r_5 \mathbf{x}_{\text{pure},5}$

Fig. 5. Shapes of two simulated mixture spectra.

3.2. Actual ideal binary mixture

One of the weaknesses of the proposed method is noise and baseline shift sensitivity because $e_{\text{threshold}}$ and ν_t are determined based on excess absorbance, which is affected by noise. In actual mixture cases, a Savitzky–Golay (SG) filter[31] is applied to reduce noise and baseline shift effects. The SG filter reduces noise and baseline shift effects by smoothing and differentiation. Moreover, differentiation by the SG filter does not disturb the wavelength selection because the Beer–Lambert law is not affected by ν differentiation.

The parameters of the SG filter are the order of differentiation, window size, and order of the polynomial. The latter two parameters are related to the noise reduction process. These three parameters are adjusted by minimization of RMSE r_{train} . Ideal binary mixtures consisting of isopropanol acetate (IPA) and heptane were used to analyze the performance of the SG filter and parameter fitting for the SG filter. They were also used to analyze the performance of the proposed method for ideal mixtures. Seventeen samples with different mole fractions were provided by Pfizer Inc. All of the regions were used in this study, while the 1500–1620 and 1820–2040 nm regions were used in previous research.[13] Shapes of pure-component and two-component-mixture spectra are shown in Fig. 6.

Fig. 6. Pure-component and two-component-mixture spectra of an ideal binary mixture.

3.3. Actual nonideal binary mixture

To show that the proposed method can be applied to nonideal mixtures and that the refinement and AD evaluation perform well, we analyzed nonideal binary mixtures consisting of water and IPA. These spectra were also measured by Pfizer Inc.[13] The pure component and the mixture spectra are shown in Fig. 7. Eight

mixtures with different mole fractions were available.

(a) Pure-component spectra

(b) Mixture spectra of the eight samples

Fig. 7. Spectra of nonideal binary mixtures.

4. Results and Discussion

4.1. Prediction results for the simulated data

In this section, we will discuss the performance of WLSEA and a requirement of WLSEA-IOT. Two

predictions were performed with simulated data: prediction of spectra over the complete wavenumber range ($650\text{--}2750\text{ cm}^{-1}$) and prediction of the spectra over half the wavenumber range ($650\text{--}1700\text{ cm}^{-1}$). In the latter analysis, each pure component spectrum contained only one peak, meaning that the spectra of the first, second, and fourth pure components each have only one nonideal peak. To predict the mole fraction, $e_{\text{threshold}}$ candidates were generated from 0.1% to 10% of the maximum value of $\mathbf{x}_{\text{mix}}$ (in increments of 0.1%) rather than the 0.95-quantile of $\mathbf{x}_{\text{mix}}$ because the simulated data does not have noise. Note that $e_{\text{threshold}}$ candidates are generally generated from 0.1% to 10% in increments 0.1% of the 0.95-quantile of $\mathbf{x}_{\text{mix}}$ to reduce the noise effect. The number of training samples was two. For the complete analysis, the prediction results, selected regions $\mathbf{v}_t$, and $\mathbf{x}_{\text{excess}}$ are shown in Fig. 8(a), Fig. 9(a), and Fig. 10(a), respectively. The three figures show that the proposed method succeeded in producing a highly accurate prediction for the simulated data.

(a) Complete wavenumber range

(b) Wavenumber range 650–1700 cm^{-1}

Fig. 8. Prediction results for five simulated component mixtures. The mole fractions of all of the pure components are plotted. Red asterisks and blue circles are the results of WLSEA-IOT and IOT, respectively.

(a) Selected regions v_i for the complete wavenumber range analysis

(b) Selected regions ν_i for the half wavenumber range analysis

Fig. 9. Selected regions ν_i when using the (a) complete wavenumber range and (b) 650–1700 cm^{-1} range.

(a) Complete wavenumber range

(b) Wavenumber range 650–1700 cm^{-1}

Fig. 10. x_{excess} using the (a) complete wavenumber range and (b) the 650–1700 cm^{-1} range.

Fig. 8(b), Fig. 9(b), and Fig. 10(b) show the results of WLSEA-IOT with half the wavenumber range. The predictive accuracy with only the 650–1700 cm^{-1} range (Fig. 8(b)) is poorer than that with all wavenumbers (Fig. 8(a)). This causes bad prediction results, which is confirmed in Fig. 11. Fig. 11 is the result with one set of training samples using half the wavenumber range. The mole fractions of the first and fourth pure components were not predicted well because their peaks are not selected, as shown in Fig. 9(b) and Fig. 10(b). It is concluded that the proposed wavelength selection works well for the simulated data and that a requirement of the proposed method is that the selected regions $\mathbf{v}$ must contain at least one peak per pure component spectrum.

Fig. 11. Prediction results with one set of training samples using the $650\text{--}1700\text{ cm}^{-1}$ regions.

4.2. Results for the ideal binary data

We performed a set of analyses on ideal binary mixtures to show that the parameter fitting for the SG filter and the SG filter itself perform well. The parameter candidates for this ideal mixture case are shown in Table

3.

Table 3. Candidates for the SG filter parameters

parameter	values
order of differentiation	0, 1, 2
order of polynomial	1, 2, 3
window size	5, 10, 15, 20

Note that the order of the polynomial is greater than or equal to the order of differentiation. In addition, the order of the polynomial was not 0 in the analyses because the results with order of polynomial 0 are the same as those without the SG filter.

(a) All results

(b) Only 0–0.015 RMSE r_{all} and 0–0.01 RMSE r_{train}

Fig. 12. SG filter parameter fitting results. D denotes the order of differentiation.

We first checked the result of parameter fitting with a set of training samples (Fig. 12). The number of training samples was two. $e_{\text{threshold}}$ was determined by a golden section search. The lower and upper bounds were 0.1% and 50% of the 0.95-quantile of $\mathbf{x}_{\text{mix}}$ when the order of differentiation was 1 or 2, while they were 0.1% and 5% of the 0.95-quantile of $\mathbf{x}_{\text{mix}}$ when the order of differentiation was 0. Fig. 12 shows the correlation between RMSE $\mathbf{r}$ in eq 7 and RMSE $\mathbf{r}_{\text{train}}$ of all 17 samples. RMSE $\mathbf{r}_{\text{all}}$ and RMSE $\mathbf{r}_{\text{train}}$ of all of the predictions have positive correlation, which indicates that minimization of RMSE $\mathbf{r}_{\text{train}}$ results in minimum RMSE $\mathbf{r}_{\text{all}}$. It is concluded that parameter fitting performs successfully.

The minimum RMSE $\mathbf{r}_{\text{train}}$ in Fig. 12 is when the order of differentiation is 1. This is because baseline shift, which is observed in Fig. 13 and Fig. 14, greatly affects the prediction accuracy. Fig. 13 and Fig. 14 show selected regions $\mathbf{v}$ and smoothed $\mathbf{x}_{\text{excess}}$ of five of the 17 samples. Without differentiation, the shape of the mixtures is affected by baseline shift (Fig. 14(a)) while the shape is not affected by baseline shift with first-order differentiation (Fig. 14(b)). Considering the results in Fig. 12-Fig. 14, one is the best order of differentiation. The prediction error increased with order of differentiation two because a large differentiation order amplifies the noise effect.

Fig. 13. Selected regions ν for the zeroth and first derivative.

(a) With SG filter and without differentiation

(b) With SG filter and first-order differentiation

Fig. 14. x_{excess} of five samples with SG filter and either zeroth-order or first-order differentiation.

We limited the order of differentiation to one to reduce computation time. For the golden section search, 0.1% and 50% of the 0.95-quantile of x_{mix} were the lower and upper bounds of the search section. The prediction results are shown in Fig. 15. The red asterisks show the results of WLSEA-IOT for all of the 17 samples with all sets of training samples, while the blue circles are the results of IOT for the 17 samples. IOT results were predicted with all regions. The dotted line denotes the range where the prediction error is 0.01. Figure 15 shows that the predictive accuracy of the proposed method competes with that of IOT, although some prediction errors of the results are above 0.01. In conclusion, the SG filter and wavelength selection provide a good prediction for ideal mixtures.

Fig. 15. Prediction results with SG filter and first-order of differentiation.

4.3. Results for nonideal binary data

Nonideal binary mixtures were used to show that the proposed method can be applied to nonideal mixture systems. First, the SG filter parameters were adjusted. The values of the SG filter parameters are shown in Table 4. The number of training samples was two. The order of differentiation was fixed to 1 because RMSE r_{train} was high with an order of differentiation of 0 or 2.

Table 4. Candidates for the SG filter parameters.

parameter	values
order of differentiation	1
order of polynomial	1, 2, 3
window size	5, 10, 15, 20

The prediction results and selected regions are shown in Fig. 16 and Fig. 17. The results of IOT were predicted with all regions. In Fig. 16(a), most of the results are within the dotted lines that denote predicted results differing from observed values by 0.01. The prediction results of WLSEA-IOT are higher than those of IOT. Fig. 16(b) shows the prediction results after refinement. The prediction results in Fig. 16(b) are poorer than those in Fig. 16(a). Selected regions from the first to the $_{8C_2}$ -th (=28) set of training samples are shown in Fig. 17. Although refinement makes the predictive accuracy slightly poorer, unnecessary regions, mainly where the wavenumber is greater than 1700 cm^{-1} , were successfully removed. Fig. 18 shows $\mathbf{x}_{\text{excess}}$, and indicates the selected $\mathbf{v}$ in Fig. 17 contains noise regions whose wavenumber is greater than 1700 cm^{-1} .

(a) Before refinement

(b) After refinement

Fig. 16. Prediction results. The red asterisks consist of 28 sets of eight samples.

(a) Before refinement

(b) After refinement

Fig. 17. Selected wavelength regions with the 28 sets of training samples.

Fig. 18. x_{excess} of the smoothed nonideal binary mixtures. The region where the wavenumber is greater than 1700 cm^{-1} only shows noise.

After refinement, the ADs were evaluated. The results with four sets of training samples are shown in Fig. 19. The red asterisks and blue circles denote the prediction and simulation results, respectively. The results indicate that the method successfully evaluates the prediction error for each sample at any RMSE x_{mix} . Thus, WLSEA-IOT takes estimated prediction errors into account when using prediction results for new validation spectra. It is concluded that samples in the selected regions ν are completely linear and their predictive accuracy follows that of simulated spectra based on $x_{\text{pure},i}$.

(a) First set of training samples

(b) Second set of training samples

(c) Tenth set of training samples

(d) Twenty-second set of training samples

Fig. 19. ADs based on $\text{RMSE } x_{\text{mix}}$. The red asterisks and the blue circles show the prediction results for actual mixtures and randomly generated spectra, respectively.

5. Conclusions

In this paper, we propose a novel calibration–minimum prediction method WLSEA-IOT that can be applied even to nonideal mixtures [that have ideal regions](#). WLSEA is generally performed after a SG filter to reduce the effects of baseline shift and noise. The three parameters of the SG filter are adjusted by minimization of the RMSE r_{train} . Refinement of the regions and an AD evaluation method are also proposed. The refinement method obtains more reliable regions. ADs are defined using simulated spectra because only a limited number of training samples are available.

The performance of the proposed method was verified by three case studies. A case study with simulated data shows that WLSEA performs well and has a requirement. The second case study with ideal mixtures shows that parameter fitting for the SG filter succeeded and the SG filter contributed to the high accuracy. The third case study with nonideal mixtures indicates that WLSEA-IOT can be successfully applied to nonideal mixtures, and that refinement and AD evaluation perform well. [However, the proposed method cannot be applied to nonideal mixtures that have no ideal regions. This is still a problem in some of more complex mixtures than the mixtures used in the case studies.](#)

The proposed wavelength selection method that separates mixture infrared spectra into ideal and nonideal regions will [enhance more precise analyses of molecular interactions with infrared spectroscopy](#).

Appendix A

IOT

IOT is a calibration-free method used to predict the mole fractions of pure components in mixtures. The following optimization problem is solved when predicting the mole fractions:

$$\min_{r_i} \left\| \mathbf{x}_{\text{mix}} - \sum_{i=1}^c r_i \mathbf{x}_{\text{pure},i} \right\|^2,$$

such that

$$\begin{aligned} 0 \leq r_i \leq 1, \sum_{i=1}^c r_i &= 1, \\ \alpha_i \leq r_i \leq \beta_i, \end{aligned} \tag{10}$$

where $\mathbf{x}_{\text{mix}}$, $\mathbf{x}_{\text{pure},i}$, i , c , α_i , and β_i are the mixture spectrum, spectrum of pure component i , pure component index, number of pure components, and lower and upper bounds of the pure component mole fraction r_i , respectively. The predictive accuracy of IOT competes with that of PLS, although IOT requires no training sample.[\[13\]](#) However, this method is only suitable for ideal mixture systems. The prediction accuracy decreases if $\mathbf{x}_{\text{mix}}$ has a nonlinear relationship with $\mathbf{x}_{\text{pure},i}$ because of molecular interactions such as hydrogen bonding and van der Waals forces.

IOT is implemented with the `fmincon` function in the Optimization Toolbox of Matlab R2015a.[\[32\]](#)

Acknowledgment

The authors acknowledge the support of the Core Research for Evolutionary Science and Technology (CREST) project “Development of a knowledge-generating platform driven by big data in drug discovery

through production processes” of the Japan Science and Technology Agency (JST). The authors wish to thank K. Muteki at Pfizer Inc. for providing the actual binary mixture data.

REFERENCES

- [1] D.C. Hinz, Process analytical technologies in the pharmaceutical industry: The FDA’s PAT initiative, *Anal. Bioanal. Chem.* 384 (2006) 1036–1042. doi:10.1007/s00216-005-3394-y.
- [2] J.A. Lopes, P.F. Costa, T.P. Alves, J.C. Menezes, Chemometrics in bioprocess engineering: Process analytical technology (PAT) applications, *Chemom. Intell. Lab. Syst.* 74 (2004) 269–275. doi:10.1016/j.chemolab.2004.07.006.
- [3] A.L. Pomerantsev, O.Y. Rodionova, Process analytical technology: A critical view of the chemometricians, *J. Chemom.* 26 (2012) 299–310. doi:10.1002/cem.2445.
- [4] T.R.M. De Beer, C. Bodson, B. Dejaegher, B. Walczak, P. Vercruysse, a. Burggraeve, et al., Raman spectroscopy as a process analytical technology (PAT) tool for the in-line monitoring and understanding of a powder blending process, *J. Pharm. Biomed. Anal.* 48 (2008) 772–779. doi:10.1016/j.jpba.2008.07.023.
- [5] J. Mantanus, E. Rozet, K. Van Butsele, C. De Bleye, A. Ceccato, B. Evrard, et al., Near infrared and Raman spectroscopy as Process Analytical Technology tools for the manufacturing of silicone-based drug reservoirs, *Anal Chim Acta.* 699 (2011) 96–106. doi:10.1016/j.aca.2011.05.006.
- [6] R.P. Cogdill, C.A. Anderson, M. Delgado-Lopez, D. Molseed, R. Chisholm, R. Bolton, et al., Process analytical technology case study part I: feasibility studies for quantitative near-infrared method

- development., AAPS PharmSciTech. 6 (2005) E262–E272. doi:10.1208/pt060237.
- [7] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemom. Intell. Lab. Syst.* 58 (2001) 109–130. doi:10.1016/S0169-7439(01)00155-1.
- [8] J.-H. Jiang, Y. Liang, Y. Ozaki, Principles and methodologies in self-modeling curve resolution, *Chemom. Intell. Lab. Syst.* 71 (2004) 1–12. doi:10.1016/j.chemolab.2003.07.002.
- [9] M. Maeder, Evolving Factor Analysis for the Resolution of Overlapping Chromatographic Peaks, *Anal. Chem.* 59 (1987) 527–530. doi:Doi 10.1021/Ac00130a035.
- [10] E.R. Malinowski, Window factor analysis: Theoretical derivation and application to flow injection analysis data, *J. Chemom.* 6 (1992) 29–40. doi:10.1002/cem.1180060104.
- [11] P.A. Hailey, P. Doherty, P. Tapsell, T. Oliver, P.K. Aldridge, Automated system for the on-line monitoring of powder blending processes using near-infrared spectroscopy Part I. System development and control, *J. Pharm. Biomed. Anal.* 14 (1996) 551–559. doi:10.1016/0731-7085(95)01674-0.
- [12] S.S. Sekulic, J. Wakeman, D. Phil, P. Hailey A., Automated system for the on-line monitoring of powder blending processes using near-infrared spectroscopy. Part II. Qualitative approaches to blend evaluation, *J. Pharm. Biomed. Anal.* 17 (1998) 1285–1309. doi:10.1016/S0731-7085(98)00025-9.
- [13] K. Muteki, D.O. Blackwood, B. Maranzano, Y. Zhou, Y. A. Liu, K.R. Leeman, et al., Mixture component prediction using iterative optimization technology (Calibration-Free/Minimum Approach), *Ind. Eng. Chem. Res.* 52 (2013) 12258–12268. doi:10.1021/ie3034587.
- [14] Shibayama, S.; Kaneko, H.; Funatsu, K., A Novel Minimal-Calibration Method for Prediction of Mole Fraction in Non-Ideal Mixtures. *AAPS PharmSciTech.* 2016, [accepted](#).

- [15] E. Kriesten, F. Alsmeyer, A. Bardow, W. Marquardt, Fully automated indirect hard modeling of mixture spectra, *Chemom. Intell. Lab. Syst.* 91 (2008) 181–193. doi:10.1016/j.chemolab.2007.11.004.
- [16] E. Kriesten, D. Mayer, F. Alsmeyer, C.B. Minnich, L. Greiner, W. Marquardt, Identification of unknown pure component spectra by indirect hard modeling, *Chemom. Intell. Lab. Syst.* 93 (2008) 108–119. doi:10.1016/j.chemolab.2008.05.002.
- [17] M. Arakawa, Y. Yamashita, K. Funatsu, Genetic algorithm-based wavelength selection method for spectral calibration, *J. Chemom.* 25 (2011) 10–19. doi:10.1002/cem.1339.
- [18] H. Kaneko, K. Muteki, K. Funatsu, Improvement of iterative optimization technology (for process analytical technology calibration-free/minimum approach) with dimensionality reduction and wavelength selection of spectra, *Chemom. Intell. Lab. Syst.* 147 (2015) 176–184. doi:10.1016/j.chemolab.2015.08.017.
- [19] W. Windig, J. Guilment, Interactive Self-Modeling Mixture Analysis, *Anal. Chem.* 63 (1991) 1425–1432. doi:10.1021/ac00014a016.
- [20] W. Windig, N.B. Gallagher, J.M. Shaver, B.M. Wise, A new approach for interactive self-modeling mixture analysis, *Chemom. Intell. Lab. Syst.* 77 (2005) 85–96. doi:10.1016/j.chemolab.2004.06.009.
- [21] S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G. Patlewicz, J. Niemela, et al., A stepwise approach for defining the applicability domain of SAR and QSAR models, *J. Chem. Inf. Model.* 45 (2005) 839–849. doi:10.1021/ci0500381.
- [22] H. Kaneko, K. Funatsu, Applicability Domain Based on Ensemble Learning in Classification and Regression Analyses., *J. Chem. Inf. Model.* 54 (2014) 2469–2482. doi:10.1021/ci500364e.

- [23] M.S. Escobar, H. Kaneko, K. Funatsu, Flour concentration prediction using GAPLS and GAWLS focused on data sampling issues and applicability domain, *Chemom. Intell. Lab. Syst.* 137 (2014) 33–46. doi:10.1016/j.chemolab.2014.06.005.
- [24] H. Kaneko, M. Arakawa, K. Funatsu, Applicability Domains and Accuracy of Prediction of Soft Sensor Models, *AIChE J.* 57 (2011) 1506–1513. doi:10.1002/aic.12351.
- [25] V. K., K. M., D. W., Spectral similarity versus structural similarity: infrared spectroscopy, *Anal. Chim. Acta.* 490 (2003) 313–324. doi:10.1016/j.aca.2004.04.014.
- [26] Q. Li, N. Wang, Q. Zhou, S. Sun, Z. Yu, Excess infrared absorption spectroscopy and its applications in the studies of hydrogen bonds in alcohol-containing binary mixtures., *Appl. Spectrosc.* 62 (2008) 166–170. doi:10.1366/000370208783575663.
- [27] Y. Koga, F. Sebe, T. Minami, K. Otake, K. Saitow, K. Nishikawa, Spectrum of excess partial molar absorptivity. I. Near Infrared Spectroscopic Study of Aqueous Acetonitrile and Acetone, *Phys. Chem. Chem. Phys.* 113 (2009) 11928–11935. doi:10.1021/jp901934c.
- [28] F. Sebe, K. Nishikawa, Y. Koga, Spectrum of excess partial molar absorptivity. Part II: a near infrared spectroscopic study of aqueous Na-halides, *Phys. Chem. Chem. Phys.* 14 (2012) 4433. doi:10.1039/c2cp23255c.
- [29] S.D. Bruce, J. Higinbotham, I. Marshall, P.H. Beswick, An analytical derivation of a popular approximation of the Voigt function for quantification of NMR spectra., *J. Magn. Reson.* 142 (2000) 57–63. doi:10.1006/jmre.1999.1911.
- [30] F. Alsmeyer, W. Marquardt, Automatic Generation of Peak-Shaped Models, 58 (2004) 986–994.

- [31]A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal. Chem. 36 (1964) 1627–1639. doi:10.1021/ac60214a047.
- [32]Matlab R2015a, 3 Apple Hill Drive Natick, MA 01760-2098 (2015) <http://www.mathworks.com/>.