Improvement of Iterative Optimization Technology (for Process Analytical Technology Calibration-Free/ Minimum Approach) with Dimensionality Reduction and Wavelength Selection of Spectra

Hiromasa Kaneko¹, Koji Muteki², Kimito Funatsu¹

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

²Pfizer Global Supply, Global Technology Services, Eastern Point Rd, Groton, Connecticut 06340, United States

Corresponding author: Kimito Funatsu

Tel: +81-3-5841-7751

Fax: +81-3-5841-7771

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

Abstract

Process analytical technology plays an important role in the pharmaceutical industry. Calibration-free/ minimum approach, iterative optimization technology (IOT), was previously proposed to predict the composition of a mixture while maintaining a similar [prediction ability](#) to calibration models such as a partial least squares. However, for the mixture case which includes similar structured materials, it would be essentially difficult to provide good prediction on mixture component ratio. This study presents a method which can improve the prediction ability of IOT through reducing dimensionality of spectra with optimal selection of wavelength. It involves using a latent variable model approach for dimensionality reduction of spectra in IOT and genetic algorithm-based wavelength selection for optimal wavelength-region selection. Through the analyses of numerical simulation data and real industrial data, it was confirmed that the proposed method achieved higher predictive accuracy compared to the traditional IOT.

Keywords: Process analytical technology, Iterative optimization technology, Dimensionality reduction, Wavelength selection

1 Introduction

Process analytical technology (PAT) [1-4] plays an important role in the pharmaceutical industry. PAT is used extensively in process development, process understanding, and process control [5]. Often, quantitative measurements are desired/required and a calibrated model will have to be developed and implemented. A calibrated model could account for active pharmaceutical ingredient (API) lot-to-lot variability and scale-up operations and was applied to optimization of process conditions in pilot-scale plant and industrial-scale plant of the high shear wet granulation process [6]. García-Muñoz et al. proposed to use a calibrated model for feed-forward process control and for designing quality into process conditions and materials' properties (Quality by Design) [7]. The development, implementation and maintenance of these quantitative models are both resource and time intensive.

Muteki et al [8] previously proposed calibration-free/ minimum approach, iterative optimization technology (IOT), which is used to predict the composition of a mixture while maintaining a similar [prediction ability](#) to calibration models such as a partial least squares (PLS) [9] model. It involves using only pure component spectra and mixture component spectra (without a calibration data set). Linear and non-linear IOT have been successfully applied (by offline/online) to some practical pharmaceutical process (e.g. blending, feed frame, solvent mixture, reaction, etc).

However, a key (remaining) question was how to deal with a dependency (multi-collinear relationship) among pure component spectra during the optimization computation. For the mixture case which includes similar structured materials, it would be essentially difficult to provide good prediction on mixture component ratio, although the degree of collinearity of spectra can be changed by different measurement conditions of spectrometer. This problem has often been an important obstacle when applying IOT to mixture cases having more than several components. In addition, nonlinear relationships between the absorption of pure components and that of mixture components also deteriorate [prediction ability](#) of IOT. Although nonlinear IOT has been developed already [8], this requires to optimize a parameter using training samples and linear models would be more robust than nonlinear ones. Selection of only absorption bands with there exist linear relationships between the absorption of pure components and that of mixture components would be one of the solutions.

This study presents a method which can improve the prediction ability of IOT through reducing dimensionality of spectra with optimal selection of wavelength. It

involves using a latent variable model (LVM) approach [10] such as principal component analysis (PCA) [11] and PLS for dimensionality reduction of spectra in IOT and genetic-algorithm-based wavelength selection (GAWLS) [12] for optimal wavelength-region selection. LVM is used as preprocessing of spectra and collinearity in pure component spectra would be handled. IOT estimation is applied to the calculation of a fitness value in GAWLS and a combination of wavelength-regions with which a predictive IOT model can be optimized would be obtained.

The effectiveness of the proposed methods is demonstrated using simulation data where pure component spectra have strong correlation and real industrial data of multicomponent mixtures. The proposed methods achieve higher predictive accuracy compared to the traditional IOT.

2 Method

2.1 LVM-IOT

Prediction ability of IOT decreases when there is collinearity between pure component spectra and collinearity between wavelengths (see Appendix A). To handle collinearity and guarantee independency of pure component spectra, our approach is dimensionality reduction and we apply latent variable modelling (LVM) as preprocess of IOT. The details of IOT are shown in Appendix A.

One of the most famous LVM methods is PCA, the details of which are described in Appendix B. PCA can reduce the dimension of $\mathbf{X}$ and extract uncorrelated PCs which are input variables for IOT.

PCA can be used in different ways. Figure 1 shows images of two ways to employ PCA. First, principal components (PCs) are extracted from pure component spectra data matrix $\mathbf{X}_{\text{pure}} \in R^{I \times N}$ (I represents the number of pure components and N represents the number of wavelengths). By using a loading matrix $\mathbf{P} \in R^{N \times q}$ (q represents the number of PCs), score vectors $\mathbf{T}_{\text{mix}} \in R^{J \times q}$ (J represents the number of mixture component spectra) can be obtained from mixture component spectra data matrix $\mathbf{X}_{\text{mix}} \in R^{J \times N}$ as follows:

$$\mathbf{T}_{\text{mix}} = \mathbf{X}_{\text{mix}} \mathbf{P}. \quad (1)$$

IOT can estimate mole fractions $\mathbf{R} \in R^{I \times J}$ using $\mathbf{T}_{\text{pure}} \in R^{I \times q}$ instead of $\mathbf{X}_{\text{pure}}$ and $\mathbf{T}_{\text{mix}}$ instead of $\mathbf{X}_{\text{mix}}$ in eqs. (A.7) and (A.8). This method is named PCA-IOT1. PCA-IOT1 can handle correlation between wavelengths.

Second, PCs are extracted from pure component spectra data matrix $\mathbf{X}_{\text{pure}}^T \in R^{N \times I}$ after transposition (lower part in Figure 1). IOT can estimate $\mathbf{R}$ for $\mathbf{T}_{\text{pure}} \in R^{N \times q}$, $\mathbf{R}_{\text{Tpure}} \in R^{q \times J}$, using $\mathbf{T}_{\text{pure}}^T$ instead of $\mathbf{X}_{\text{pure}}$ and $\mathbf{X}_{\text{mix}}$ in eqs. (A.7) and (A.8). $\mathbf{R}_{\text{Tpure}}$ is transformed to $\mathbf{R}$ using $\mathbf{P} \in R^{I \times q}$ as follows:

$$\mathbf{R} = \mathbf{P} \mathbf{R}_{\text{Tpure}}. \quad (2)$$

This method is named PCA-IOT2. PCA-IOT2 can handle correlation between pure spectra.

For PCA-IOT1, PCA can be applied also to the data matrix combined with $\mathbf{X}_{\text{pure}}$ and $\mathbf{X}_{\text{mix}}$, and $\mathbf{T}_{\text{pure}}$ and $\mathbf{T}_{\text{mix}}$ are extracted simultaneously. This is named PCA-IOT3. PLS instead of PCA can be performed as LVM. PLS requires objective variables $\mathbf{Y} \in R^{J \times I}$. The mole fraction of each pure component is 1 for $\mathbf{Y}$ -variables, which is also used when Amigo *et al.* applied PLS to classical least squares [13]. PLS is applied as is the case with PCA-IOT1. This is named PLS-IOT.

2.2 GAWLS-IOT

Except an LVM approach, selection of the suitable absorption bands is another solution to guarantee the independency of pure component spectra. In addition, absorption bands having nonlinearity between pure component spectra and mixture component spectra, which come from molecular interactions such as hydrogen bonding, deteriorate prediction ability of IOT. Unnecessary wavelengths must be removed from pure component spectra and mixture component spectra.

Our second approach is wavelength selection or variable selection. Many variable selection methods such as uninformative variable elimination [14], searching combination moving window PLS [15] and model population analysis [16] have been developed and have produced good results. Not a single absorption band but a region of absorption bands would be reasonable to represent the absorbance of a representative band of pure components and mixture components. Thus, we apply genetic-algorithm-based wavelength selection (GAWLS) method [17] which is a wavelength-regions selection method to IOT. GAWLS is one of the methods used to select combinations of important wavelengths using regions as a unit of measurement. A genetic algorithm (GA) [18] is applied to select wavelength-regions. A GA is an optimization method that is used in biology to model principles of natural evolution. Species having a high level of fitness under certain environmental conditions can

prevail in the next generation, and the best species may be reproduced by crossover together with the random mutation of chromosomes in those species that survive. The solution space around superior individuals is searched for preferentially, which leads to the discovery of a solution that is close to the optimum. IOT is performed to calculate a fitness value. This method is named GAWLS-IOT.

In GAWLS-IOT, two actual values of a chromosome represent one region of wavelengths. Figure 2 shows the basic concept of GAWLS-IOT. Hence, GAWLS-IOT can select important wavelengths using regions as a unit of measurement. In Figure 2, the numbers of selected wavelengths are 1364 through 1381 and 1723 through 1744, and the number of windows (NW) is 2. Only a set of selected wavelength-regions is used in IOT modelling. A negative value of the root mean squared error (RMSE) which is an index of estimation error for training data is used as a fitness value of the chromosome. In this way, a set of an optimum set of wavelength-regions, with which highly predictive model (the RMSE value is low) can be constructed, is obtained.

The number of chromosomes, the number of generations, the maximum width of window and NW must be set beforehand. The minimum width of window is zero.

2.3 GAWLS-LVM-IOT

To handle collinearity of pure component spectra in wavelength-regions selected with GAWLS-IOT, not IOT but LVM-IOT such as PCA-IOT1 and PCA-IOT2 can be used in the calculation of a fitness value in GA. This is named GAWLS-LVM-IOT such as GAWLS-PCA-IOT1 and GAWLS-PCA-IOT2.

3 Results and Discussion

To verify the effectiveness of the proposed methods, we analyzed two sets of numerical simulation data including 10 mixture materials and two sets of real industrial data including multiple mixture materials.

3.1 Numerical simulation data where pure component spectra have strong correlation

In this case study, we analyzed 10 virtual pure component spectra having several degrees of collinearity and mixture component spectra represented by linear combinations of the pure component spectra. 10 virtual pure component spectra are generated first. A spectrum including two normal distribution functions is assumed a pure component spectrum as shown in Figure 3. The averages of the normal distribution functions are 50 and 200. The second one whose average is 200 is doubled. 10 virtual pure component spectra differ in the standard deviations of the normal distribution

functions. Intervals of the standard deviations relate to collinearity in virtual pure component spectra, i.e. smaller intervals mean stronger correlation in virtual pure component spectra. Intervals changes; 0.1, 0.5, 1, 1.5, 2, 3, 4, 5, 6, 7, 8, 9 and 10. Then, for each set of pure component spectra, we changed each composition randomly under the conditions of averages of 0.1 and the sum of the compositions of 1. Mixture component spectra are linear combinations of pure component spectra. Noise following the normal distribution of average of 0 and standard deviation of 0.01 is added to pure component spectra and mixture component spectra. The number of samples is 100. Table 1 shows average of correlation coefficients between pure spectra for each interval, which means the degree of overlap of pure spectra. When intervals are small, pure spectra are correlated and the degree of overlap is high.

IOT, PCA-IOT1, PCA-IOT2, PCA-IOT3 and PLS-IOT were used to estimate compositions of 10 components for each sample. RMSE values are calculated using all composition values of 100 samples. Figure 4 shows RMSE values for each interval of the standard deviations of the normal distribution functions. By decreasing the interval and increasing correlation in pure component spectra, the RMSE values increased for normal IOT. By using PCA-IOTs and PLS-IOT, the RMSE values decreased for each interval, i.e. each degree of correlation. PCA and PLS could handle collinearity in pure component spectra in IOT modeling and could improve [prediction ability](#) of IOT successfully. The difference in PCA-IOTs and PLS-IOT is not large in this study.

3.2 Numerical simulation data where spectra have unnecessary wavelengths

Then, we analyzed 10 virtual pure component spectra mixture component spectra represented by linear combinations of the pure component spectra, but the pure component spectra and mixture spectra include unnecessary wavelengths. A spectrum including two normal distribution functions is assumed a pure component spectrum as shown in Figure 5. The averages of the normal distributions are 20 and 50 and the standard deviations are both 5. The first one whose average is 20 is tripled. 10 virtual pure component spectra differ in the average of the second normal distribution function. The interval is 20 and the averages are 50, 70, 90, 110, 130, 150, 170, 190, 210 and 230. There exists little correlation in virtual pure component spectra. Then, for each set of pure component spectra, we changed each composition randomly under the conditions of averages of 0.1 and the sum of the compositions of 1. Mixture component spectra are linear combinations of pure component spectra. However, only in the calculation of mixture component spectra, virtual intensity of pure component spectra for the first normal distribution function is zero. Thus, as shown in Figure 6, virtual intensity in the

corresponding region is also zero. This region is unnecessary wavelengths (variables), which is clear by comparing Figures 5 and 6. Noise following the normal distribution of average of 0 and standard deviation of 0.01 is added to pure component spectra and mixture component spectra. The number of samples is 100.

IOT and GAWLS-IOT were used to estimate compositions of 10 components for each sample. 30 samples were set as training data for GAWLS-IOT and estimation accuracy for 70 test samples was compared in IOT and GAWLS-IOT. Figure 7 shows the plot of simulated and estimated all 10 compositions of 70 samples for IOT and the RMSE value including all 10 components was 0.0337. In this case, the results of PCA-IOTs and PLS-IOT are almost the same as that of IOT, which would be because correlation in pure component spectra is little.

The number of chromosomes was set to 300 and the number of generations was set to 500 for GAWLS-IOT. NW was set as 10 because there are 10 pure components and each pure component spectrum would have its own important wavelength-region. We performed GAWLS-IOT 10 times, i.e. 10 runs because the final solution of GA depends on initial chromosomes. The average value of the 10 RMSE values was 0.0097 and greatly decreased, compared to that of IOT. Additionally, as shown in the plot of simulated and estimated compositions in Figure 7, the [prediction ability](#) dramatically improved by using GAWLS-IOT. This was because unnecessary wavelengths corresponding to the first normal distribution function in Figures 5 and 6 could be removed successfully as shown in the wavelength selection results for 10 calculations in Figure 8.

3.3 Real industrial data including five mixture materials

Five mixture materials consisting of an Active Pharmaceutical Ingredient (API), Microcrystalline Cellulose (MCC), Fast Flo Lactose 316 (FFL), Sodium Starch Glycolate (SSG) and Stearic Acid (SA) were used with a rotary tablet press employing continuous mass feeder and continuous mixing [8]. The tablet press feed frame was modified to incorporate an NIR probe to directly measure the blend composition circulating within the feed frame just prior to tableting. An ePAT 611 NIR was mounted to the feed frame from the top of paddle wheel. The NIR spectral region collected was 1100 nm to 2075 nm with a 4 nm spectral resolution.

We prepared 100 samples including 30 samples as training data for GAWLS-IOT and GAWLS-PCA-IOTs. As a preprocessing method, the Savitzky–Golay (SG) filter [19] was employed. The SG filter can differentiate and smooth spectra simultaneously and the details of the SG method should be referred to as the reference

[20,21]. The width is 11, the order of the polynomial is 3 and the order of the derivative is 1. These hyper-parameters were determined through a trial and error process.

Table 2 shows RMSE values for test data in IOT, PCA-IOT1, PCA-IOT2, PCA-IOT3 and PLS-IOT modeling. Each RMSE value and a total RMSE value are large for IOT. RMSE values of SSG and SA, which are low materials, must be lower particularly since RMSE values exceed theoretical values. Although RMSE values for PCA-IOT1 and PLS-IOT are the same as those for IOT, RMSE values decreased by using PCA-IOT2 and PCA-IOT3 and the RMSE value for SSG became low for PCA-IOT3. The collinearity between pure component spectra would be responsible rather than the collinearity between wavelengths and important PCs could be extracted by including mixture component spectra to pure component spectra. However, total RMSE values are still large and the RMSE value for SA whose theoretical value is 1% must decrease more.

Then, we applied GAWLS-IOT, GAWLS-PCA-IOT1 and GAWLS-PCA-IOT2 to improve predictive ability. The number of chromosomes was set to 300 and the number of generations was set to 500. NW changed to 2, 3, 4 and 5. Total RMSE values for each NW are shown in Figure 9. Although there is some variability in RMSE values, all RMSE values are lower than the total RMSE values of IOT, PCA-IOT1, PCA-IOT2, PLS-IOT3 and PLS-IOT in Table 2. The GAWLS-based IOT could select appropriate wavelength-regions and [prediction ability](#) of IOT improved. As NW increases, RMSE values decrease and the RMSE value of GAWLS-PCA-IOT1 is minimum when NW is five. However, it would not be reasonable to increase NW furthermore because there are only five pure components and selection results of wavelength-regions becomes complicated as NW increases. By using not only GAWLS but also PCA, RMSE values are low when NW is small in particular. Small NW means the small number of input variables, which is prone to produce high correlation between pure component spectra. Collinearity in pure component spectra would be handled well with PCA and [prediction ability](#) of IOT improved.

RMSE values for test data in GAWLS-IOT, GAWLS-PCA-IOT1 and GAWLS-PCA-IOT2 are shown in Table 2 when NW is five, which is because there are five pure components and each pure component spectrum would have its own important wavelength-region. RMSE values for all components decreased and RMSE values of low components became sufficiently low by using the proposed methods. Figure 10 shows estimation results for IOT, GAWLS-IOT, GAWLS-PCA-IOT1 and GAWLS-PCA-IOT2. Although the estimation errors are large for every component in IOT modeling, the estimation performance improved dramatically by using

GAWLS-IOT and selecting only important wavelength-regions (see Figures 10(a) and 10(b)). However, the estimation errors for SA are still large and there is some bias in estimated values. By combining GAWLS and PCA with IOT, the estimation errors of SA decreased and the bias also became small although the estimation results of the other components did not change much as shown in Figures 10(c) and 10(d). This would be because collinearity in pure component spectra could be dealt with using PCA and compositions of even low materials could be optimized correctly in IOT modeling. We confirmed that [prediction ability](#) of IOT could improve by using the proposed methods.

As another way to use GAWLS-IOT, GAWLS-PCA-IOT1 and GAWLS-PCA-IOT2, we simulated a situation that high estimation accuracy of only API was required. In this case, a fitness value in GAWLS can be modified, which is a negative value of RMSE of only API composition. By modifying a fitness value, the RMSE value of API could be reduced to about 0.6, which is much lower than those in Table 2. Of course, estimation accuracy of only low materials can improve by set a fitness value as a negative value of RMSE of only low materials. Our proposed methods can be customized according to the objective of IOT estimation.

3.4 Real industrial data including four mixture materials

Four mixture materials consisting of an Active Pharmaceutical Ingredient (API), Microcrystalline Cellulose (MCC), Fast Flo Lactose 316 (FFL) and Magnesium Stearate (MgSt) were used with continuous mixer employing continuous mass feeder. An ePAT 611 NIR was mounted to the continuous mixer. The NIR probe directly measures the blend composition at the discharge of continuous mixer. The NIR spectral region collected was 1100 nm to 2075 nm with a 4 nm spectral resolution. At the steady state condition of MCC, FFL and MgSt mixtures, a small amount of API was invested to investigate the process dynamic response.

This data set was used as an application of our proposed methods. The SG filter was employed first as is the case with the section 3.3. The hyper-parameters were the same as those in the section 3.3. As shown in Figure 11, API composition values estimated with IOT is not in agreement with the theoretical value of API composition soon after an API spike, which is because materials are not mixed well first. To estimate ideal API composition at an API spike, we used spline interpolation [22]. First API composition values are estimated in IOT modeling at several points after the materials are mixed to some extent (see Figure 11). Then, API composition values at the other points can be estimated using a cubic spline interpolation based on the values estimated with IOT. The pink or gray line in Figure 11 is the interpolation result. The interpolated

value at an API spike is set as the estimated value of API composition.

Figure 12(a) shows the plots of theoretical and estimated API composition values in four batches in IOT modeling. Although estimated API values are in good agreement with the theoretical API values at API spikes, IOT cannot estimate API values of zero accurately after API drains some time later.

GAWLS-IOT and GAWLS-PCA-IOTs were used to improve predictive accuracy for not only API values at API spikes but also API values of zero some time later. The number of chromosomes was set to 300 and the number of generations was set to 500. NW was set as four because there are four pure components and each pure component spectrum would have its own important wavelength-region. The estimation results are shown in Figure 12. The data are very close to the diagonals and prediction accuracy increased greatly for all GAWLS-IOT, GAWLS-PCA-IOT1 and GAWLS-PCA-IOT2. As shown in time plots of theoretical and estimated API composition values for IOT and GAWLS-PCA-IOT1 as an example of the proposed methods, estimated API values converge to zero and appropriate time change of API is verified by using GAWLS-PCA-IOT1 (see Figure 13). We confirmed that [prediction ability](#) of IOT improved by using the proposed methods.

4 Conclusions

For the improvement of [prediction ability](#) of IOT, we combine LVM and GAWLS with IOT to reduce dimensionality in spectra and select important wavelength-regions, respectively. Different ways to use LVM by changing spectra data matrix exist as shown in the section 2.2. By using LVM-IOT in the calculation of a fitness value in GAWLS, the problems of collinearity in pure component spectra and unnecessary wavelengths can be solved simultaneously.

Through case studies of numerical simulation data sets including collinearity in pure component spectra and unnecessary wavelengths, it was demonstrated that PCA-IOTs and PLS-IOT as LVM-IOTs could handle collinearity in pure component spectra appropriately and GAWLS-IOT could select only important wavelength-regions. In addition, the proposed GAWLS-PCA-IOTs which are the combination of IOT, PCA and GAWLS showed the much improvement of [prediction ability](#) of IOT and achieved the highly predictive ability in applications using real industrial data sets of multicomponent mixtures.

Optimization of the number of PCs in PCA-IOTs and PLS-IOT and Optimization of NW in GAWLS-IOT and GAWLS-PCA-IOTs remain to be issues in the future. We believe that by applying our proposed method to process control and

achieving the highly accurate prediction, pharmaceutical processes will be operated effectively and stably.

Acknowledgements

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 Daniel O. Blackwood, Yang A. Liu, Sonja S. Sekulic, Mojgan Moshgbar, David Lauri Pla in Pfizer Inc for providing the process data and useful discussion.

References

- [1] J.A. Lopesa, 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.
- [2] Y. Roggo, P. Chalus, L. Maurer, C.L. Martinez, A. Edmond, N. Jent, A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *J. Pharmaceut. Biomed.* 44 (2007) 683–700.
- [3] D. Bu, B. Wan, G. McGeorge, A discussion on the use of prediction uncertainty estimation of NIR data in partial least squares for quantitative pharmaceutical tablet assay methods. *Chemom. Intell. Lab. Syst.* 120 (2013) 84–91.
- [4] E.M. Hansuld, L. Briens, A review of monitoring methods for pharmaceutical wet granulation. *Int. J. Pharm.* 472 (2014) 192–201.
- [5] G.L. Reid, H.W. Ward II, A.S. Palm, K. Muteki, Process analytical technology (PAT) in pharmaceutical development. *Amer. Pharm. Rev.* 15 (2012) 49–55.
- [6] K. Muteki, K. Yamamoto, G.L. Reid, M. Krishnan, De-risking scale-up of a high shear wet granulation process using latent variable modeling and near-infrared spectroscopy, *J. Pharm. Innov.* 6 (2011) 142–156.
- [7] S. García-Muñoz, S. Dolph, H.W. Ward II, Handling uncertainty in the establishment of a design space for the manufacture of a pharmaceutical product, *Comput. Chem. Eng.*, 37 (2010) 1098–1107.
- [8] K. Muteki, D.O. Blackwood, B. Maranzano, Y. Zhou, Y.A. Liu, K.R. Leeman, G.L. Reid, Mixture component prediction using iterative optimization technology (calibration-free/minimum approach), *Ind. Eng. Chem. Res.* 52 (2013) 12258–12268.

- [9] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemom. Intell. Lab. Syst.* 58 (2001) 109–130.
- [10] C.M. Bishop, M.E. Tipping, A hierarchical latent variable model for data visualization, *IEEE T. Pattern Anal.* 20 (1998) 281–293.
- [11] S. Wold, Principal component analysis, *Chemom. Intell. Lab. Syst.* 2 (1987) 37–52.
- [12] M. Arakawa, Y. Yamashita, K. Funatsu, Genetic algorithm-based wavelength selection method for spectral calibration, *J. Chemom.* 25 (2011) 10–19.
- [13] J.M. Amigo, C. Ravn, N.B. Gallagher, R. Bro, A comparison of a common approach to partial least squares-discriminant analysis and classical least squares in hyperspectral imaging. *Int. J. Pharm.* 373 (2009) 179–182.
- [14] W. Cai, Y. Li, X. Shao, A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra, *Chemom. Intell. Lab. Syst.* 90 (2008) 188–194.
- [15] W. Ni, S.D. Brown, R. Man, Stacked partial least squares regression analysis for spectral calibration and prediction, *J. Chemom.* 23 (2009) 505–517.
- [16] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, Model population analysis for variable selection, *J. Chemom.* 24 (2010) 418–423.
- [17] H. Kaneko, K. Funatsu, A new process variable and dynamics selection method based on a genetic algorithm-based wavelength selection method. *AIChE J.* 58 (2012) 1829–1840.
- [18] D. Whitley, A genetic algorithm tutorial, *Stat. Comput.* 4 (1994) 65–85.
- [19] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36 (1964) 1627–1639.
- [20] J. Luo, K. Ying, J. Bai, Savitzky–Golay smoothing and differentiation filter for even number data. *Signal Process.* 85 (2005) 1429–1434.
- [21] N. Yoshimura, M. Takayanagi, Chemometrics calculations with Microsoft Excel (5), *J. Comput. Chem., Jpn.* 11 (2012) 149–158. (in Japanese)
- [22] http://en.wikipedia.org/wiki/Spline_interpolation
- [23] H.K. Hughes, Suggested nomenclature in applied spectroscopy. *Anal. Chem.* 24 (1952) 1349 – 1355.
- [24] Q. Li, N. Wang, Q. Zhou, S. Sun, Z. Yu, Excess infrared absorption spectroscopy and its application in the studies of hydrogen bonds in alcohol-containing binary mixtures. *Appl. Spectrosc.* 62 (2008) 166–170.
- [25] 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. *J. Phys. Chem. B.* 113 (2009) 11928–11935.

- [26] K. Muteki, J.F. MacGregor, T. Ueda, Rapid development of new polymer blends: the optimal selection of materials and blend ratios. *Ind. Eng. Chem. Res.* 13 (2006) 4653–4660.
- [27] K. Muteki, J.F. MacGregor, Optimal purchasing of raw materials: a data-driven approach. *AIChE J.* 54 (2008) 1554–1559.

Appendix A: IOT [8]

The ideal molar absorption coefficients of mixture components $\varepsilon_{\text{ideal}}$ is given as follows:

$$\varepsilon_{\text{ideal}} = \sum_{i=1}^M r_i \varepsilon_{\text{pure},i} , \quad (\text{A.1})$$

where $\varepsilon_{\text{pure},i}$ and ε_{mix} are the molar absorption coefficients of pure and mixture components, M is the total number of pure components and r_i is a mole fraction of mixture components. The excess molar absorption is expressed using the molar absorption coefficient of mixture components ε_{mix} as follows:

$$\begin{aligned} \varepsilon^{\text{E}} &= \varepsilon_{\text{mix}} - \varepsilon_{\text{ideal}} \\ &= \varepsilon_{\text{mix}} - \sum_{i=1}^M r_i \varepsilon_{\text{pure},i} , \end{aligned} \quad (\text{A.2})$$

which means a deviation from Beers–Lambert law [23-25]. The excess absorbance e can then be rewritten as follows:

$$e = x_{\text{mix}} - \sum_{i=1}^M r_i x_{\text{pure}} . \quad (\text{A.3})$$

where $x_{\text{pure},i}$ and x_{mix} are the absorbance of pure and mixture components. When the number of wavelengths or bands is N , a vector of the excess absorbance $\mathbf{e} \in R^N$ is represented as follows:

$$\mathbf{e} = \mathbf{x}_{\text{mix}} - \sum_{i=1}^M r_i \mathbf{x}_{\text{pure},i} , \quad (\text{A.4})$$

where $\mathbf{x}_{\text{mix}} \in R^N$ is a vector of the absorption of mixture components and $\mathbf{x}_{\text{pure},i} \in R^N$ is a vector of the absorption of pure components.

IOT can be expressed as an optimization problem:

$$\sum_{j=1}^N \left(x_{\text{mix}}(j) - \sum_{i=1}^M r_i x_{\text{pure},i}(j) \right)^2 \quad (\text{A.5})$$

subject to

$$\alpha \leq r_i \leq \beta, \quad \sum_{i=1}^M r_i = 1 \quad (\text{A.6})$$

where α and β are the lower and upper limits, respectively, and $\alpha=0$ and $\beta=1$, basically. r_i is iteratively computed at every new sampling data.

Equations (A.5) and (A.6) show a weighted composition form based on lambert-beers law and common mixture constraints, respectively, and does not involve any model coefficients in advance for the mixture component prediction. Mixture component ratios are then predicted so as to minimize a lambert-beers law deviation on every new sampling data. In common classical least squares, the model coefficient should be first calculated/obtained with calibration samples prior to predicting mixture component ratios on every new sampling data. Both approach looks having similar mathematical algebra form, but differently works.

An assumption in IOT is that the standard spectra are relatively independent to each other [8]. Collinearity in pure component spectra and in wavelengths cause the multi-collinearity problem during the optimization computation. The mole fractions of materials that are spectrally similar will not be able to be predicted correctly. It is important to reduce dimensionality in spectra and select the suitable absorption bands while accounting for the independency of pure component spectra.

Solutions to the optimization problems [26,27] were obtained using MATLAB®.

Figures captions

Figure 1. [Basic concept](#) of two ways to apply PCA.

Figure 2. Basic concept of GAWLS-IOT. [A chromosome means selected regions of wavelengths and two actual values of a chromosome represent one region of wavelengths. In this figure, the number of windows is set as 2 and the numbers of selected wavelengths are 1364 through 1381 and 1723 through 1744. Only a set of selected wavelength-regions is used in IOT modelling to calculate a fitness value.](#)

Figure 3. Example of a virtual pure component spectrum in the section 3.1. The averages of the normal distribution functions are 50 and 200 and the standard deviations are both 2. The second one whose average is 200 is doubled.

Figure 4. RMSE values for each interval of the standard deviation (std) of the normal distribution functions in the section 3.1.

Figure 5. Example of a virtual pure component spectra in the section 3.2. The averages of the normal distribution functions are 20 and 50 and the standard deviations are both 5. The first one whose average is 20 is tripled.

Figure 6. Example of a mixture component spectrum in the section 3.2.

Figure 7. Plots of simulated and estimated compositions for IOT and GAWLS-IOT in the section 3.2.

Figure 8. Wavelength selection results for 10 calculations of GAWLS-IOT in the section 3.2.

Figure 9. Total RMSE values for each NW in the section 3.3.

Figure 10. Estimation results in the section 3.3. Points mean test data and circles mean training data for GAWLS-IOT, GAWLS-PCA-IOT1 and GAWLS-PCA-IOT2.

Figure 11. Example of IOT prediction from an API spike.

Figure 12. Plots of theoretical and estimated API composition values in the section 3.4.

Figure 13. Time plots of theoretical and estimated API composition values in the section 3.4.

Table 1. Average of correlation coefficients between pure spectra for each interval in the section 3.1.

Interval	Average
0.1	0.992
0.5	0.933
1.0	0.881
1.5	0.848
2.0	0.824
3.0	0.794
4.0	0.775
5.0	0.753
6.0	0.721
7.0	0.683
8.0	0.647
9.0	0.619
10.0	0.599

Table 2. RMSE values [%] for test data in the section 3.3. NW is five for GAWLS-IOT, GAWLS-PCA-IOT1 and GAWLS-PCA-IOT2.

	API	MCC	FFL	SSG	SA	Total
IOT	3.599	8.339	7.603	9.953	2.555	7.012
PCA-IOT1	3.599	8.339	7.603	9.953	2.555	7.012
PCA-IOT2	3.519	5.522	7.544	7.226	2.349	5.613
PCA-IOT3	3.191	3.442	7.740	0.723	2.231	4.182
PLS-IOT	3.599	8.339	7.603	9.953	2.555	7.012
GAWLS-IOT	1.468	1.766	1.246	0.236	0.609	1.204
GAWLS-PCA-IOT1	1.667	1.376	1.283	0.650	0.311	1.170
GAWLS-PCA-IOT2	1.406	1.477	1.033	0.570	0.359	1.066

Figure 1. [Basic concept](#) of two ways to apply PCA.

Figure 2. Basic concept of GAWLS-IOT. A chromosome means selected regions of wavelengths and two actual values of a chromosome represent one region of wavelengths. In this figure, the number of windows is set as 2 and the numbers of selected wavelengths are 1364 through 1381 and 1723 through 1744. Only a set of selected wavelength-regions is used in IOT modelling to calculate a fitness value.

Figure 3. Example of a virtual pure component spectrum in the section 3.1. The averages of the normal distribution functions are 50 and 200 and the standard deviations are both 2. The second one whose average is 200 is doubled.

Figure 4. RMSE values for each interval of the standard deviation (std) of the normal distribution functions in the section 3.1.

Figure 5. Example of a virtual pure component spectra in the section 3.2. The averages of the normal distribution functions are 20 and 50 and the standard deviations are both 5. The first one whose average is 20 is tripled.

Figure 6. Example of a mixture component spectrum in the section 3.2.

(a) IOT

(b) GAWLS-IOT

Figure 7. Plots of simulated and estimated compositions for IOT and GAWLS-IOT in the section 3.2.

Figure 8. Wavelength selection results for 10 calculations of GAWLS-IOT in the section 3.2.

Figure 9. Total RMSE values for each NW in the section 3.3.

(a) IOT

(b) GAWLW-IOT

Figure 10. Estimation results in the section 3.3. Points mean test data and circles mean training data for GAWLS-IOT, GAWLS-PCA-IOT1 and GAWLS-PCA-IOT2.

(c) GAWLS-PCA-IOT1

(d) GAWLS-PCA-IOT2

Figure 10. (Continue)

Figure 11. Example of IOT prediction from an API spike.

Figure 12. Plots of theoretical and estimated API composition values in the section 3.4.

Figure 13. Time plots of theoretical and estimated API composition values in the section 3.4.