Chemometrics and Intelligent Laboratory Systems 55 (2001) 53–65

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

# Near-infrared spectral data transfer using independent standardization samples: a case study on the trans-alkylation process

Kwang-Su Park [a], Young-Hyun Ko [a], Hyeseon Lee [a], Chi-Hyuck Jun [a,*], Hoeil Chung [b], Min-Sik Ku [b]

[a] *Department of Industrial Engineering, Pohang University of Science and Technology, San 31, Hyoja-dong, Nam-gu, Pohang 790-784, South Korea*
[b] *NIR Project Team, SK Corporation, 110 Kosa-dong, Nam-Gu, Ulsan, South Korea*

Received 5 May 2000; accepted 6 November 2000

**Abstract**

A variety of standardization or transfer methods between near infrared spectrometric instruments are applied for the content prediction of five major constituents of the product at trans-alkylation process with spectra measured on two different instruments. Because process samples are difficult to be stored, we use independent transfer samples by blending some pure materials for the spectrum standardization of the process samples. Using the independent standardization samples, we investigate the transfer performance of well-known piecewise direct standardization combined with several regression methods on the raw spectra. Also, we propose some indirect standardization methods utilizing wavelet transferred scores or factor scores through principal component analysis and partial least squares. The standardization by transferring scores takes only a few transfer coefficients, but it shows similar performance to the spectrum transfer case. In addition, we show the possibility of using a fewer number of stable samples than the original set of samples for the standardization with similar performance. © 2001 Elsevier Science B.V. All rights reserved.

*Keywords:* NIR; Standardization; PDS; Wavelet; Factor transfer

## 1. Introduction

The use of statistical and mathematical techniques (multivariate calibrations) for the analysis of near-infrared (NIR) spectra is increasing in a variety of areas. Multivariate methods are useful in chemistry or environment data analysis to predict material compositions or in the process data analysis to monitor and control a process. A typical procedure in a multivariate calibration for quantitative analysis with the spectroscope is taking a series of samples with known property of interest and constructing the model to predict the characteristics for new samples. Calibra-

[*] Corresponding author. Tel.: +82-54279-2197; fax: +82-54279-2870.
*E-mail address:* chjun@postech.ac.kr (C.-H. Jun).

0169-7439/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0169-7439(00)00115-5

tion models are based on a large number of samples that require considerable time and cost for the preparation and measurement.

As many researchers [1–6] have pointed out, there are three problems in the use of calibration model with the spectroscope. The first occurs when a calibration model developed on one instrument is to be transported to other instruments. Even though the type of these instruments is identical, difference in spectral responses may cause erroneous results on other instruments. Constructing a new calibration model is not simple when the samples are numerous, chemically or physically unstable, hazardous, or when several instruments at separate sites are involved. The second problem is observed when the instrumental responses measured on a single instrument change over a period of time because of temperature fluctuation, electric drift, wavelength or detector intensity instability, instrument part exchange, et al. Finally, the third problem is caused by the difference between samples coming from different production situations such as sample finish, particle size, surface texture, density, et al. Standardization is required to avoid repetition of the whole calibration procedure including the measurements of calibration spectra and response variables and the computation of calibration parameters.

Some methods such as simple slope and bias correction of the predicted values [7], robust wavelength selection [8], orthogonal signal correction [9], filtering based on a finite impulse response [10], and positive matrix factorization (PMF) as a data reconstruction [11] may have been applied for the standardization. Standardization methods using the linear transfer matrix including transfer of spectra or transfer of coefficients of calibration model have been suggested and focused by many researchers [1–6,12,13]. Duponchel et al. [14] investigated the nonlinear mapping from spectrum to spectrum. Forina et al. [4] proposed the transfer of the regression equation from master to slave instrument with two-step partial least square. Walczak et al. [15] introduced a new method based on transferring spectra in the wavelet domain.

The standardization with transfer matrix is often used, which is constructed by mapping the spectrum of the master instrument or reference day from the spectrum obtained by a slave instrument or in the day. In this paper, we use “master” for the original instrument that a calibration model has been based on and ‘slave’ for a new satellite one that a new calibration model should be provided for. The patent method [12] is the correction of each wavelength according to the highest correlation between master and slave instrument followed by the correction of spectral intensity. Wang et al. [6] proposed four transfer methods and compared them with the patent method. Among them, the most well-known and popular methods are direct standardization (DS) and piecewise direct standardization (PDS).

In this work, we use the set of independent standardization samples that is called the generic set. The generic set is obtained by combining pure major constituents of the product at trans-alkylation process and some other materials having similar properties. The generic samples are easily taken under different times and situations by blending pure constituents according to the pre-designed combinations, whereas the process sample is not easy to store and to keep in the original state after a period of time.

After constructing the calibration model for five major constituents of the process, we apply PDS combined with three regression methods such as ordinary least square (OLS), principal component analysis (PCR) and partial least square (PLS) between instruments on raw spectral domain. Also, we propose so called indirect standardization (IDS) methods on score projection domain utilizing wavelet transformation, PCA and PLS. Using the standardization by score transfer, the size of transfer matrix becomes much smaller than using the standardization of raw spectra data. Thus the IDS would have the practical advantage of fast information transferring when standardization of remote satellite instruments is needed, especially in on-line analysis. In each of standardization methods (PDS and IDS), using the pre-determined calibration model for each response variable of the process samples, we estimate transfer regression parameters and search for a suitable size of transfer window and number of factors in PDS or number of transfer levels or factors in IDS. We then compare the performance of the standardization methods with each other. In addition, we reduce the number of samples to be involved in the transfer to search a suitable number of generic transfer samples by eliminating less informative samples from the original set of generic samples. The minimum number of transfer samples would be particularly desirable when a company is operating several laboratory stations at different sites.

## 2. Experimental

### 2.1. Notations

The notation of subscript and/or superscript of a matrix is the following: $\mathbf{X}_{IJ}^{(\mathrm{M})}$ for spectra measured on master instrument and $\mathbf{X}_{IJ}^{(\mathrm{S})}$ for spectra on slave instrument. The superscript (M or S) will be omitted when representing both. The first subscript $(I)$ represents the number of rows (the number of samples) in the matrix and the second subscript $(J)$ denotes the number of columns (the number of wavelengths) in the matrix, which are often omitted. Lowercase letter indices are used to denote a part of $\mathbf{X}_{IJ}$, as $\mathbf{X}_{Ij}$ for $j$th column vector and $X_{ij}$ for a scalar, that is, $(i,j)$th element of $\mathbf{X}_{IJ}$. Furthermore, the transpose matrix of $\mathbf{X}_{IJ}$ is represented by the $\mathbf{X}_{JI}$. A variety of data sets (matrices) are defined as follows:

| | |
|---|---|
| $\mathbf{Y}$ | matrix of constituent contents of the process samples for the set of training or test samples |
| $\mathbf{X}$ | spectra matrix for the set of training or test samples |
| $\mathbf{G}$ | spectra matrix for the set of generic samples |
| $\mathbf{S}$ | spectra matrix for the standardization, a subset of $\mathbf{G}$ or $\mathbf{X}$ |
| $\boldsymbol{W}(j)$ | $j$th portion of spectra matrix for a given range of wavelengths |
| $\boldsymbol{D}$ | high-frequency components (details) of spectra $\mathbf{X}$ in the discrete wavelet transform |
| $\mathbf{F}$ | transfer matrix from spectra to spectra (called $\mathbf{X}$–$\mathbf{X}$ type), or from details to details (called $\mathbf{D}$–$\mathbf{D}$ type) |
| $\mathbf{X}^{(\mathrm{SM})}$ | transferred spectra matrix from $\mathbf{X}^{(\mathrm{S})}$, $\mathbf{X}^{(\mathrm{SM})}=\mathbf{X}^{(\mathrm{S})}\mathbf{F}$ |
| $\mathbf{D}^{(\mathrm{SM})}$ | transferred details matrix from $\mathbf{D}^{(\mathrm{S})}$, $\mathbf{D}^{(\mathrm{SM})}=\mathbf{D}^{(\mathrm{S})}\mathbf{F}$ |
| $\mathbf{P}$ | loading matrix of training spectra $\mathbf{X}$ |
| $\mathbf{Z}$ | score matrix of training spectra from the model of $\mathbf{X}=\mathbf{ZP}+\mathbf{E}$ |
| $\mathbf{R}$ | weighted loading matrix of training spectra in PLS to obtain scores $\mathbf{Z}=\mathbf{XR}$ |
| $\mathbf{Q}$ | loading matrix of standardization spectra $\mathbf{S}$ |
| $\mathbf{T}$ | score matrix of standardization spectra from the model of $\mathbf{S}=\mathbf{TQ}+\mathbf{E}$ |
| $\mathbf{E}$ | residual matrix of spectra matrix ($\mathbf{X}$ or $\mathbf{S}$) |
| $\mathbf{V}$ | transfer matrix from score to score (called $\mathbf{Z}$–$\mathbf{Z}$ type) |
| $\mathbf{Z}^{(\mathrm{SM})}$ | transferred score matrix from $\mathbf{Z}^{(\mathrm{S})}$, $\mathbf{Z}^{(\mathrm{SM})}=\mathbf{Z}^{(\mathrm{S})}\mathbf{V}$ |
| $\mathbf{U}$ | residual matrix of score matrix $\mathbf{Z}$ or $\mathbf{T}$ |

In addition, $L$ represents the number of principal components or latent variables. To reduce the number of variables for the calibration or for the transfer, the predictor matrix $(\mathbf{X}_{IJ})$ may be decomposed as $\mathbf{X}_{IJ}=\mathbf{T}_{IL}\mathbf{P}_{LJ}+\mathbf{E}_{IJ}$ where $L$ components are retained. Through this projection procedure, the number of the predictor variables is decreased from $J$ to $L$ and all of new score vectors are orthogonal to each other.

### 2.2. Training and test sets for the calibration and the standardization

The process data set consists of 88 samples and each sample has five main constituents (Benzene, Toluene, Xylene, Trimethylbenzene, and Ethylmethylbenzene). TMB and EMB denote the last two constituents, respectively. The contents of five constituents in weight percentage are predicted and the root mean square (RMS) error of each constituent is evaluated to see the predictive ability of calibration model and the effectiveness of spectra transfer from slave to master instrument.

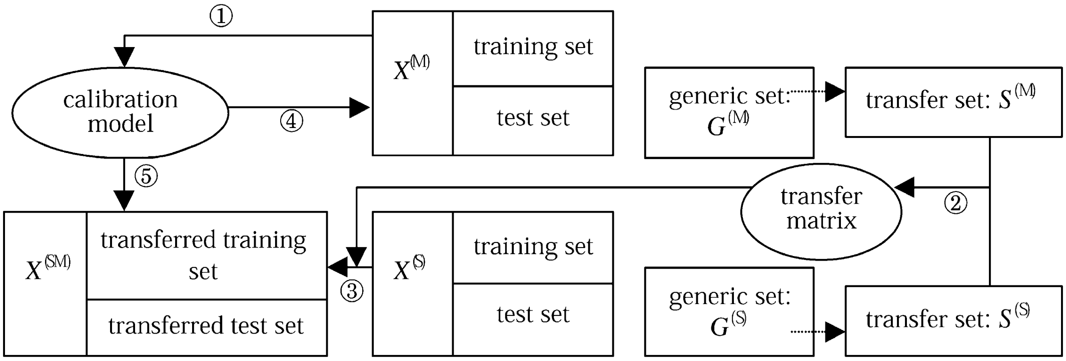
The 88 process samples are divided into a training set of 44 samples and a test set of 44 samples. Subdivision is done through the Duplex algorithm [16] applied to the master spectra data to produce independent but similar sample distributions in the training and the test set. The Duplex method is modification of the Kennard–Stone (KS) [17] algorithm to overcome the drawback of the KS by dividing samples into two groups alternately. The KS algorithm sequentially selects a sample to maximize the minimal Euclidean distances between already selected samples and the remaining samples. This procedure is repeated until the required samples are obtained, whereas the Duplex method selects representative calibration and test data sets of equal size based on the KS algorithm applied to each subset alternately. The KS or Duplex algorithm should be taken after excluding outliers, if any, since the outmost sample is likely selected first.

The calibration model is obtained through the cross validation with leave-one-out method using the training set measured on the master instrument. The analysis procedure including the model calibration and the standardization is plotted in Fig. 1. The prediction model is obtained from the spectra treated by first differencing (that is, simply taking difference in spectral intensity between $i$th wavelength and ($i+1$)th wavelength) and by mean centering. For the standardization, only first differencing is applied to the spectra. Programs for the preprocessing and modeling are written in MATLAB® (The Language of Technical Computing, Version 5.2.0, the MathWorks, 1998).

As indicated in Fig. 1, the standardization will be performed using the generic set instead of using the process samples. The generic set contains 21 samples and weight percentages of each constituent are listed in Table 1. A generic sample is composed of pure material and/or mixture of pure materials to play a representative role of an independent sample of the process.

Table 1
Pre-determined contents (in wt.%) of 21 generic samples

| No. | Benzene | Toluene | Ethylbenzene | *p*-Xylene | PDEB |
|---|---|---|---|---|---|
| 1 | | 100.00 | | | |
| 2 | | | 100.00 | | |
| 3 | | | | 100.00 | |
| 4 | | | | | 100.00 |
| 5 | 19.97 | 20.01 | 20.00 | 20.00 | 19.99 |
| 6 | 39.97 | 9.99 | 10.01 | 29.99 | 9.99 |
| 7 | 9.99 | 40.00 | 10.00 | 9.99 | 30.00 |
| 8 | 30.00 | 9.99 | 40.00 | 10.00 | 10.01 |
| 9 | 9.98 | 29.99 | 9.99 | 40.00 | 10.00 |
| 10 | 10.00 | 10.00 | 30.00 | 10.00 | 40.00 |
| 11 | 60.00 | 10.03 | 10.00 | 9.99 | 9.99 |
| 12 | 10.01 | 59.99 | 9.99 | 10.00 | 9.99 |
| 13 | 10.00 | 10.01 | 59.99 | 10.01 | 9.99 |
| 14 | 10.01 | 10.01 | 10.00 | 59.99 | 10.00 |
| 15 | 9.99 | 10.00 | 9.99 | 10.01 | 60.00 |
| 16 | 29.99 | 5.00 | 29.99 | 5.00 | 30.00 |
| 17 | 30.00 | 30.00 | 4.99 | 29.99 | 5.00 |
| 18 | 5.00 | 29.99 | 29.99 | 31.20 | 3.80 |
| 19 | 4.99 | 5.03 | 30.00 | 29.99 | 30.00 |
| 20 | 100% 2,2-Dimethylbutane | | | | |
| 21 | 100% 2,3,4-Trimethylpentane (*iso*-Octane) | | | | |



① : Make calibration model by PCR

② : Make transfer matrices for $X$-$X$ type, $D$-$D$ type or $Z$-$Z$ type standardization

③ : Apply transfer matrix to data set measured on slave instrument
to generate the transferred spectra

④ : Prediction for the test set measured on master instrument

⑤ : Prediction for the transferred test spectra set

Fig. 1. Calibration and standardization procedures with independent transfer sets.
]

### *2.3. NIR instruments with sample spectrum and response measurement*

Four NIR spectra sets (two sets of process samples and two sets of generic samples) are obtained on two instruments (we call one master and the other slave) that are the same type (NIRSystems on-line model 5000 spectrometer) made by the same manufacturer (Foss NIRSystems, Silver Spring, MD). These are equipped with a quartz halogen lamp, PbS detector, and a fiber optic interactance probe. The resolution of collected spectra is 10 nm with 2-nm data point intervals from 1100 to 2500 nm. The fiber optic probe consists of concentric rings of illuminating fibers, receiving fibers, and a reflecting mirror. The size of fiber optic is 2.54 cm in outer diameter and 15.2 cm in length. The distance between the optical fibers and the reflecting mirror is 1 cm, resulting in an actual pathlength of 2 cm.

The NIR spectra scanned on two instruments of 88 process samples and 21 generic samples are plotted in Fig. 2a and c, respectively. Fig. 2b and d show spectral differences of process and generic samples between the two instruments, respectively.

The content in weight percentage of each of five constituents in the process samples is measured through the gas chromatography. The averages and standard deviations of these contents in weight percentage are shown in Table 2 for the training and test sets.

Since the process and the generic samples are highly complex mixtures of a variety of carbon chain length hydrocarbons and oxygenates, the resulting NIR spectrum is the summation and overlapped spectra of these components. As shown in Fig. 2a and c, the spectra are dominated by overtone and combination bands of CH, $CH_2$, and $CH_3$ of various hydrocarbons. The wavelength bands from 1100 to 1270 nm and 1830 to 2100 nm correspond to the second overtone and combination bands, respectively. Most useful spectral information is included in the 1100 to 1640 nm and 1810 to 2100 nm wavelength ranges and we use these bands for the calibration. There seem to be significant spectral variations in the rest of ranges (1640–1810 and 2100–2500 nm). However, the spectral differences in these ranges are not purely from the difference in sample composition, but from the differences in instrumental performance such as detector, optical fiber, and alignment. The big spectral differences in 1640–1810 and 2100–2500 nm range are mostly from differences in detector linearity and optical fiber, respectively. Additionally, there is no useful spectral information in these ranges because of the strong absorption of NIR radiation by the relatively long optical pathlength. Consequently, these two ranges were excluded for further study. Since calibration is made on these selected ranges of spectrum, we consider only these ranges for the standardization.

### *2.4. Performance measures*

To measure the performance of standardization, the transfer error (TE) is computed from the transferred spectra and the spectra collected on the master instrument, which is given by TE $= \sqrt{\sum_{i=i}^{I}\sum_{j=1}^{J}\left(X_{ij}^{(\mathrm{M})} - X_{ij}^{(\mathrm{SM})}\right)^2/(IJ)}$ . TET is defined to be the TE for the transferred training set and TEV is for the transferred test (or validation) set. Using the calibration models developed on the master instrument for the $k$th constituent, we obtained the root mean square error of calibration ($\mathrm{RMSEC}_k$) for training spectra and the root mean square error of validation ($\mathrm{RMSEV}_k$) for test spectra collected on the master instrument. Furthermore, to examine the performance of standardization, the root mean square error of transfer ($\mathrm{RMSET}_k$) for training spectra and the root mean square error of prediction ($\mathrm{RMSEP}_k$) for test spectra transferred from the spectra measured on the slave instrument are obtained. We obtain these four performance measures based on the formula of $\sqrt{\sum_{i=1}^{I}\left(Y_{ik} - \hat{Y}_{ik}\right)^2/I}$, where $Y_{ik}$ represents the content of $k$th constituent measured by the reference method and the predicted value $\hat{Y}_{\mathrm{ik}}$ is obtained through the calibration model developed on the master instrument. In addition, to measure the goodness of fit for the $k$th constituent, the quantity of $R_k^2 = 1 - \sum_{i=1}^{I}(Y_{ik} - \hat{Y}_k)^2/\sum_{i=1}^{I}(Y_{ik} - \bar{Y}_k)^2$ is calculated where $\bar{Y}_k$ represents the average content of $k$th constituent.

### *2.5. Spectra transfer*

The relation between two sets of independent standardization samples measured on the master and the slave instruments gives a spectral transfer matrix

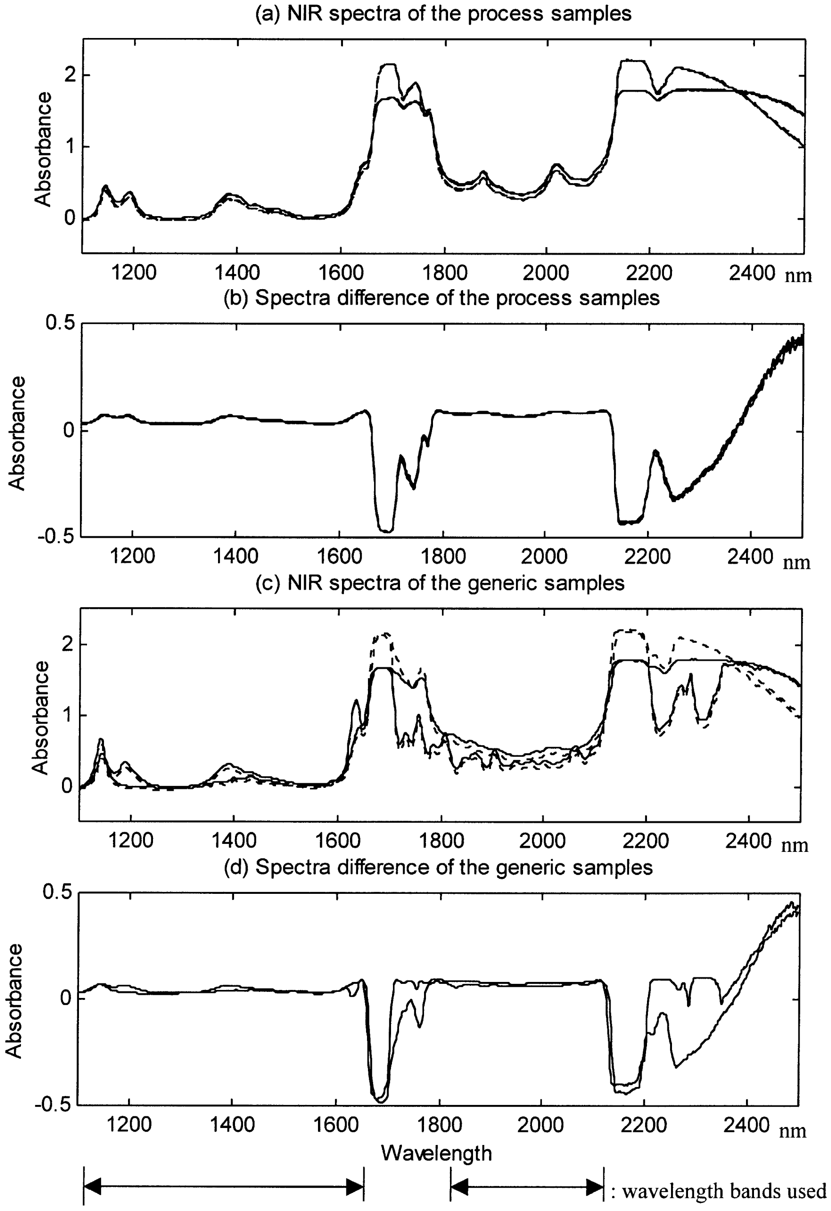


Fig. 2. NIR spectra and spectral differences between master (solid line) and slave (dashed line) instruments.

**F**. With this transfer matrix **F**, the transferred spectra matrix is estimated. In this work, we investigate the usability of inter-instrumental relationship such as **X**–**X**, **D**–**D** and **Z**–**Z** type transfers.

### *2.5.1. Piecewise direct standardization (PDS)*

The main drawback of direct standardization (DS) is that the whole range of spectrum collected in the prediction step should be used to reconstruct all spectrum intensity values of the transferred spectrum, which may lead to over-fitting and eliminate the variance of inter-spectra. As in our case, DS is not effective when the generic set is used for the standardization. In order to overcome those drawbacks, an improved version of DS, called PDS, has been developed. In the research of the Wang and Kowalski

Table 2
Content distributions (in wt.%) of five constituents in training and test set

| Constituent | Benzene | Toluene | Xylene | TMB | EMB |
|---|---|---|---|---|---|
| *Training set* | | | | | |
| Average | 6.5993 | 28.5772 | 34.2924 | 20.2211 | 0.7522 |
| Standard deviation | 2.5340 | 5.7324 | 4.0665 | 5.4577 | 0.5216 |
| Range | 11.0523 | 22.5301 | 16.2294 | 22.1379 | 2.0945 |
| *Test set* | | | | | |
| Average | 6.8041 | 29.1237 | 35.0555 | 19.5608 | 0.6760 |
| Standard deviation | 2.4388 | 5.1528 | 3.2467 | 4.8130 | 0.4050 |
| Range | 10.6145 | 20.6708 | 13.1122 | 20.2000 | 1.7087 |

[5] and the Wang et al. [6], they have indicated that PDS may be better than DS for the spectral standardization. The main difference between PDS and DS is in the range of spectra taken for the transfer. PDS is based on the fact that the spectral information contained in a certain wavelength on the master instrument is highly correlated to the spectra of neighbor wavelengths on the slave instrument. In PDS the $j$th wavelength of the transferred spectrum is reconstructed from wavelengths contained in the $j$th spectral window $\boldsymbol{W}(j)_{I\vartheta} = [\mathbf{S}^{(S)}_{I,j-g}, \mathbf{S}^{(S)}_{I,j-g+1}, \ldots, \mathbf{S}^{(S)}_{Ij}, \ldots, \mathbf{S}^{(S)}_{I,j+h-1}, \mathbf{S}^{(S)}_{I,j+h}]$, where $\vartheta = g + h + 1$ is a window size.

For the $j$th wavelength, regression coefficients $(\boldsymbol{f}_j)$ relating to the spectral intensity values on the master instrument at the $j$th wavelength and those on the slave instrument in the corresponding spectral window are computed by ordinary least squares (OLS), PCR or PLS using the equation: $\mathbf{S}^{(M)}_{Ij} = \boldsymbol{W}(j)_{I\vartheta}\, \boldsymbol{f}_j + \boldsymbol{e}_j$ where $\boldsymbol{e}_j$ is a random error vector. The use of this moving spectral window leads to obtain a banded diagonal transfer matrix $\mathbf{F} = \text{diag}(\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_{J-1}, \boldsymbol{f}_J)$. In the prediction step, a transferred spectra set is obtained by $\mathbf{X}^{(SM)} = \mathbf{X}^{(S)}\mathbf{F}$.

### 2.5.2. Indirect standardization through wavelet transfer

According to the Daubechies [18] and Chui [19], the wavelet transformation enables analysis of signal at different levels of resolution. The discrete wavelet transformation (DWT) is the wavelet transformation in the time domain (in our case, wavelength domain) and the result is therefore time-scale (in our case, wavelength–absorbance) domain analysis of the signal. A discrete signal of the length $N = 2^H$ can be decomposed into its low frequency components (called approximations) and high-frequency components (called details denoted by $\mathbf{D}$) on the $H$ different scales (levels) of resolution, recursively. At each stage of the DWT, the sampled signal is passed through a low-pass filter (scaling filter) and a high-pass filter (wavelet filter) satisfying the orthogonal conditions.

In the work of Walczak et al. [15], they present a standardization method based on transferring spectra in wavelet domain. It contains the decomposition of the slave spectra $(\mathbf{X}^{(S)})$ into the wavelet coefficients (details, $\mathbf{D}^{(S)}$), transferring it to the wavelet coefficients of master $(\mathbf{D}^{(SM)})$ and finally reconstructing it as master spectra $(\mathbf{X}^{(SM)})$. We call the transfer from slave to master as **D**–**D** type standardization. In this analysis, the transfer matrix from $\mathbf{D}^{(S)}_{I2h}$ to $\mathbf{D}^{(SM)}_{I2h}$ is obtained from the generic samples by the PDS with window size one where $h = 1, \ldots H$. The transferring from slave to master is done only up to a suitable level $(h)$. That is, the wavelet coefficients under the level are unchanged after transferring. The transfer from spectra to the wavelet coefficients and the reconstruction are based on the Daubechies wavelets. The linear-padding is applied to the spectra range omitted between 1640 and 1810 nm and after the 2100 nm to make $N = 512$ variables with $H = 9$.

### 2.5.3. Indirect standardization by factor transfer

The transfer in PDS is based on the relation between spectra measured on the master instrument and those measured on the slave instrument. In this work, we also consider the relation between the scores calculated from the spectra on the master instrument and those on the slave instrument, which is called a **Z**–**Z** type transfer.

Using the decomposition algorithm for principal component analysis, spectra measured on the master instrument are decomposed by $\mathbf{X}^{(M)} = \mathbf{Z}^{(M)}\mathbf{P}^{(M)} + \mathbf{E}$ and the projected scores of slave spectra to the latent structures of master is obtained by $\mathbf{Z}^{(S)} = \mathbf{X}^{(S)}\mathbf{P}^{(M)}$. For the standardization sets on both instruments, the same operation is possible such as $\mathbf{T}^{(M)} = \mathbf{S}^{(M)}\mathbf{P}^{(M)}$ and $\mathbf{T}^{(S)} = \mathbf{S}^{(S)}\mathbf{P}^{(M)}$. Using the equation: $\mathbf{T}^{(M)}_{IL} =$

$\mathbf{T}_{IL}^{(S)}\mathbf{V}_{LL} + \mathbf{U}_{IL}$, the factor transfer matrix $\mathbf{V}_{LL} = (\mathbf{T}_{LI}^{(S)}\mathbf{T}_{IL}^{(S)})^{-1}\mathbf{T}_{LI}^{(S)}\mathbf{T}_{IL}^{(M)}$ is obtained and new transferred scores are computed by $\mathbf{Z}^{(SM)} = \mathbf{Z}^{(S)}\mathbf{V}$. Using these transferred scores, a new spectra set is finally estimated by $\mathbf{X}^{(SM)} = \mathbf{X}^{(S)} - \mathbf{Z}^{(S)}\mathbf{P}^{(M)} + \mathbf{Z}^{(SM)}\mathbf{P}^{(M)} = \mathbf{E}^{(S)} + \mathbf{Z}^{(SM)}\mathbf{P}^{(M)}$. In this case, differently from the DS or the PDS, we do not disregard the residual matrix $\mathbf{E}^{(S)}$ in the standardization.

The scores obtained by the PLS are available since we use the loading matrix of training data measured on the master instrument in which the values of response variable are known to us. In this case, the loading matrix $\mathbf{P}^{(M)}$ in equations $\mathbf{Z}^{(S)} = \mathbf{X}^{(S)}\mathbf{P}^{(M)}$, $\mathbf{T}^{(M)} = \mathbf{S}^{(M)}\mathbf{P}^{(M)}$ and $\mathbf{T}^{(S)} = \mathbf{S}^{(S)}\mathbf{P}^{(M)}$ should be replaced by the weighted loading matrix $\mathbf{R}^{(M)} = \mathbf{W}^{(M)}(\mathbf{P}^{(M)}\mathbf{W}^{(M)})^{-1}$ where $\mathbf{W}$ is the weight matrix and $\mathbf{P}$ is the loading matrix in PLS analysis.

When using the factor transfer, there are some advantages. The factor transfer shrinks the size of transfer matrix since it uses $\mathbf{V}$ ($L \times L$ square matrix) instead of $\mathbf{F}$ ($J \times J$ square matrix) and usually $L \ll J$. Moreover, this procedure increases the stability of transfer because the variability of spectra is integrated on a few latent structures. In our experience, to apply factor transfer for standardization, similar number of factors to the one for calibration is sufficient.

## 3. Analysis results and discussion

### 3.1. *Development and test of the multivariate calibration models*

When the PCR and the PLS are applied to the spectra measured on the slave instrument without standardization, in our case, the PLS is more sensitive to out-ranged spectra and shows less effective result than the PCR. Thus, we adopt the PCR for the calibration modeling. Table 3 shows the result of calibration including the explained proportion of **X** and **Y** variance, RMSEC and RMSEV. For the content prediction of five constituents, 3 to 10 factors are used. The numbers of factors are determined as the one having the smallest cumulative values of predicted residual error sum of squares ($\mathrm{PRESS}_{l,k}$) for the $k$th constituent. It is obtained by $\sum_{i=1}^{I}(\hat{Y}_{(l,i)ik} - Y_{ik})^2$, where $\hat{Y}_{(l,i)ik}$ is the $i$th predicted value using $l$ principal components calibration model except $i$th observation from the calibration set because we use leave-one-out cross validation.

From the values of RMSEC and RMSEV in Table 3, we observe some overfit for EMB and underfit for TMB, but we assume these are pertinent since we are more interested in the standardization than the calibration. Large values of RMSEP using the test set measured on the slave instrument before standardization indicate that the spectra measured on the slave instrument needs to be standardized before applying the calibration model.

### 3.2. *Piecewise direct standardization from spectra to spectra*

For PDS, we use OLS, PCR and PLS to regress between spectral intensities measured on the slave instrument and those on the master instrument. Using the transfer matrix obtained with the 21 generic samples through three regression methods, the values of RMSET are computed according to the size of window and the number of transfer factors with the training samples transferred from the slave instru-

Table 3
The result of calibration with PCR models for five constituents

| Constituent | NCF[a] | RMSECV | Explained<br>$X$ variance | Explained<br>$Y$ variance | RMSEC | RMSEV | RMSEP |
|---|---|---|---|---|---|---|---|
| Benzene | 7 | 0.1111 | 99.88 | 99.88 | 0.0874 | 0.0898 | 0.1695 |
| Toluene | 9 | 0.2791 | 99.90 | 99.86 | 0.2083 | 0.2813 | 2.1331 |
| Xylene | 10 | 0.4152 | 99.91 | 99.45 | 0.2994 | 0.2924 | 0.9178 |
| TMB | 3 | 0.4124 | 98.95 | 99.50 | 0.3819 | 0.3481 | 1.3860 |
| EMB | 9 | 0.0576 | 99.90 | 99.35 | 0.0417 | 0.0544 | 0.5600 |

[a] NCF = number of calibration factors.

ment. When the OLS is used to obtain transfer matrix by regression coefficients between spectra of generic samples, the window size should be small enough to prevent from overfitting because the spectral intensities are highly correlated with each other.

The transfer matrix is taken at the window size and the number of transfer factors which yield the minimum RMSET and it is applied to the test spectra measured on the slave instrument. After applying the calibration models to the transferred spectra, we tabulate the results for five constituents, respectively, in Table 4.

RMSEPs by PDS through three regression methods in Table 4 are similarly improved as compared to the RMSEPs in Table 3. Among three regression methods for PDS, the result through PCR is best except for Xylene. Still, OLS, the simplest, seems to be quite comparable. It is interesting to see that better spectrum fittings (smaller TET and TEV) of PDS do not always lead to the better prediction performances (smaller RMSET and RMSEP).

### 3.3. Indirect standardization through wavelet transfer

To implement IDS by transferring wavelet coefficients, we determine the maximum level of DWT (1 $\le h \le H$) and make the transfer matrix (**F**) for the wavelet coefficient transfer up to the determined level where **F** is $2^h \times 2^h$ size matrix. We compute RMSET after the standardization of training samples according to the number of DWT level (from one to nine) and take the number of DWT levels to minimize RMSET for each constituent. In this case, we use PDS for transferring of wavelet coefficients from slave to master with window size 1.

Using the number of factors for the calibration and the DWT level for the standardization, we obtain standardization results in Table 5. The effectiveness of IDS by wavelet coefficient transfer is similar to those of PDS in Table 4. It is shown in Table 5 that the content prediction of each constituent can be improved even though the levels of DWT are smaller than nine except for Xylene. Though spectrum fittings (TET and TEV) of IDS by transferring wavelet coefficients are worse than those of PDS, prediction performance is similar to the PDS.

### 3.4. Indirect standardization by factor transfer

For the IDS by PCR score transfer, we determine the number of principal components for the transfer matrix (**V**). We compute the RMSET after the standardization of training samples according to the

Table 4
The result of standardization by PDS

| Regression method | Constituent | STW[a] | NTFP[b] | TET | RMSET | TEV | RMSEP |
|---|---|---|---|---|---|---|---|
| OLS | Benzene | 1 | | 0.00024 | 0.15241 | 0.00024 | 0.1361 |
| | Toluene | 3 | | 0.00026 | 0.39117 | 0.00025 | 0.42638 |
| | Xylene | 1 | | 0.00024 | 0.41253 | 0.00024 | 0.3509 |
| | TMB | 1 | | 0.00024 | 0.39113 | 0.00024 | 0.37786 |
| | EMB | 1 | | 0.00024 | 0.0673 | 0.00024 | 0.07515 |
| PCR | Benzene | 11 | 3 | 0.00056 | 0.11048 | 0.00055 | 0.10506 |
| | Toluene | 19 | 7 | 0.00036 | 0.41683 | 0.00035 | 0.40072 |
| | Xylene | 11 | 11 | 0.00026 | 0.41203 | 0.00025 | 0.52619 |
| | TMB | 19 | 5 | 0.00059 | 0.38586 | 0.00058 | 0.3727 |
| | EMB | 9 | 5 | 0.00026 | 0.054 | 0.00025 | 0.06162 |
| PLS | Benzene | 9 | 3 | 0.00032 | 0.12478 | 0.00032 | 0.11412 |
| | Toluene | 13 | 5 | 0.00029 | 0.41022 | 0.00029 | 0.44479 |
| | Xylene | 15 | 9 | 0.00035 | 0.39776 | 0.00035 | 0.48687 |
| | TMB | 15 | 3 | 0.00064 | 0.39944 | 0.00062 | 0.38301 |
| | EMB | 15 | 9 | 0.00035 | 0.05358 | 0.00035 | 0.06316 |

[a] STW = size of transfer window.
[b] NTFP = number of transfer factors for PDS.

Table 5
The result of standardization by wavelet scores according to the transfer level

| Constituent | NWL[a] | TET | RMSET | TEV | RMSEP |
|---|---|---|---|---|---|
| Benzene | 6 | 0.00046 | 0.1472 | 0.00046 | 0.13201 |
| Toluene | 5 | 0.00064 | 0.35521 | 0.00065 | 0.3605 |
| Xylene | 9 | 0.00026 | 0.47843 | 0.00025 | 0.41048 |
| TMB | 8 | 0.00032 | 0.43235 | 0.00031 | 0.41974 |
| EMB | 6 | 0.00046 | 0.05249 | 0.00046 | 0.06013 |

[a] NWL = number of DWT level for the transfer matrix.

number of transfer factors and obtain the number of transfer factors that is minimizing RMSET for each constituent.

Using the number of factors for the calibration and the standardization, we obtain the results for five constituents as in Table 6. The content prediction of each constituent after the IDS by PCR score transfer is worse than that of PDS in Table 4 and that of IDS by wavelet coefficient transfer in Table 5. However, considering its small size of transfer matrix and little calculation effort, the content prediction of constituents seems to be acceptable.

For the IDS by transfer of PLS scores, we determine the number of factors (latent variables) for the transfer matrix (**V**). Similarly to the case of IDS by PCR score transferring, we determine the number of transfer factors that is minimizing RMSET for each constituent. Also, we obtain the results for five constituents in Table 7. We see that the performance of IDS by PLS score transfer is a little better than that of PCR score and similar to the PDS through PCR.

From Tables 6 and 7, we see that the number of transfer factors in most cases is smaller than the number of calibration factors. It means that the difference between instruments seems to be related to small number of factors and the adjustment of those factors through score transfer may be sufficient for each constituent.

Table 6
The result of IDS by PCR score transfer

| Constituent | NTFI[a] | TET | RMSET | TEV | RMSEP |
|---|---|---|---|---|---|
| Benzene | 3 | 0.00048 | 0.13122 | 0.00047 | 0.12208 |
| Toluene | 8 | 0.00058 | 0.62541 | 0.0006 | 0.70087 |
| Xylene | 7 | 0.00053 | 0.58991 | 0.00055 | 0.73705 |
| TMB | 4 | 0.00047 | 0.41854 | 0.00047 | 0.41094 |
| EMB | 4 | 0.00047 | 0.1404 | 0.00047 | 0.14414 |

[a] NTFI = number of transfer factors for IDS.

Table 7
The result of IDS by PLS score transfer

| Constituent | NTFI | TET | RMSET | TEV | RMSEP |
|---|---|---|---|---|---|
| Benzene | 5 | 0.00045 | 0.13241 | 0.00044 | 0.12555 |
| Toluene | 7 | 0.00046 | 0.30777 | 0.00046 | 0.31907 |
| Xylene | 5 | 0.00045 | 0.54048 | 0.00044 | 0.41277 |
| TMB | 6 | 0.00045 | 0.43048 | 0.00045 | 0.38918 |
| EMB | 3 | 0.00046 | 0.07772 | 0.00046 | 0.08184 |

### *3.5. Goodness of fit according to standardization methods*

We obtain value of $R^2$ for each constituent to see the goodness of fit according to standardization methods: none (without standardization), the spectra domain transfer by PDS through three regression methods, and three methods of projection domain transfer (IDS) with the generic standardization samples. After applying the standardization methods and calibration models to the process test samples measured on the slave instrument, we obtain the results as in Table 8.

It can be seen in Table 8 that there is a certain improvement in the prediction of five constituents with the generic standardization samples. Among them, spectra domain transfer by PDS through PCR and projection domain transfer by IDS through PLS show relatively good performance compared with other methods. It is notable that the IDS by PLS score transfer shows a little better performance for the prediction of simple materials (Toluene and Xylene) than the PDS through PCR. However, this is not the case for the complex structured constituent (EMB) because the IDS by PLS score transfer considers a smaller number of factors.

### *3.6. Reduction of generic transfer samples*

To be able to use calibration models developed by the samples on master instruments over time or location, it is recommended to check the stability of the instrumental response and the measurement conditions over stable samples to quantify the spectral dif-

Table 8
Goodness of fit ($R^2$) according to the standardization methods

| Constituent | None | Spectra domain transfer by PDS | | | Projection domain transfer by IDS | | |
|---|---|---|---|---|---|---|---|
| | | OLS | PCR | PLS | Wavelet | PCR | PLS |
| Benzene | 0.9951 | 0.9968 | 0.9981 | 0.9978 | 0.9970 | 0.9974 | 0.9973 |
| Toluene | 0.8246 | 0.9930 | 0.9938 | 0.9924 | 0.9950 | 0.9811 | 0.9961 |
| Xylene | 0.9182 | 0.9880 | 0.9731 | 0.9770 | 0.9836 | 0.9473 | 0.9835 |
| TMB | 0.9151 | 0.9937 | 0.9939 | 0.9935 | 0.9922 | 0.9925 | 0.9933 |
| EMB | N/A | 0.9648 | 0.9763 | 0.9751 | 0.9774 | 0.8704 | 0.9582 |

ferences existing between the master and the slave [20]. In this work, we use the generic standardization samples as the stable samples and wish to reduce the number of generic samples. SK Corporation, the largest refinery in Korea, has 12 different terminals throughout Korea and they want to install NIR instruments at each station. Therefore, calibration transfer from the main lab to each station is really important. Each station is physically remote each other and intellectual level of actual operator is not high enough, so SK Corporation is strongly urged to make overall transfer system as simple as possible. That is why we try to use as minimal number of samples as possible.

Using the same sizes of transfer window in the PDS with 21 generic samples, we may remove samples that give less information for the standardization. To implement the reduction of generic transfer samples, we use the OLS regression method for PDS because the reduction is possible under window size of at most three and still it has good performance for the standardization. We propose the backward elimination method, which removes one sample each time sequentially.

The backward elimination is processed as follows. Starting with the initial set of $G$ samples, we prepare $G$ sets of $(G-1)$ samples by eliminating one sample in turn. We obtain the RMSET for each of five constituents and select the reduced set of $(G-1)$ samples having the minimum in the sum of five RMSETs. Then, we do the same thing to select the next reduced set of $(G-2)$ samples from the reduced set. This procedure may be repeated until the finally reduced set consists of the desirable number of samples required for the standardization.

After repetition of this procedure we obtain five RMSETs as in Fig. 3 which lead to the minimum sum of RMSETs according to the number of transfer

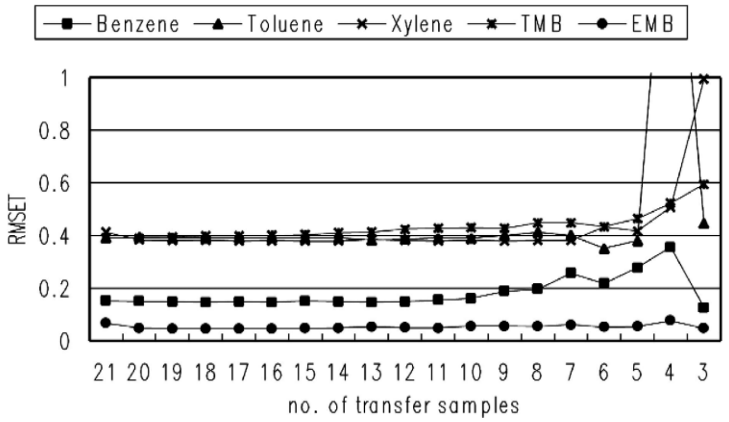


Fig. 3. RMSET according to the number of transfer samples used in the PDS.

Table 9
The result of standardization by reduced 11 generic samples when using the PDS

| Constituent | TET | RMSET | TEV | RMSEP |
|---|---|---|---|---|
| Benzene | 0.00026 | 0.15467 | 0.00026 | 0.14023 |
| Toluene | 0.00028 | 0.39029 | 0.00028 | 0.4215 |
| Xylene | 0.00026 | 0.37842 | 0.00026 | 0.36405 |
| TMB | 0.00026 | 0.42775 | 0.00026 | 0.41536 |
| EMB | 0.00026 | 0.04847 | 0.00026 | 0.05829 |

samples included. The minimum number of samples whose sum of five RMSETs becomes first less than that of total samples is 11 (sample numbers from Table 1 are 1, 2, 3, 5, 6, 7, 9, 11, 12, 20 and 21). Using the reduced 11 standardization samples, we obtain standardization result as shown in Table 9.

Even with the reduced 11 standardization samples, the performance of spectrum transfer and content prediction of five constituents is similar to the case of 21 total samples as in Table 9. For PDS through OLS regression, 11 samples seem to be sufficient for the future use of stable generic transfer samples. These are composed of pure samples (1, 2, 3, 20, 21 in Table 1) and mixtures (5, 6, 7, 9, 11, 12) of several pure materials. Though these reduced samples may be differently selected for different standardization methods, it is observed that five standardization samples (1, 3, 5, 11, 20) are commonly included. For PDS through OLS regression with these five standardization samples, the result is also acceptable as shown in Fig. 3. It seems that these five samples contain the most important information for the standardization of the process samples. That is, in our case, the independent standardization samples including these five samples and some others that may be the mixtures of pure materials are sufficient for the standardization of the process samples. Practically, we want to reduce standardization samples as small as possible. Further study is still ongoing in our laboratory to achieve more simplicity of standardization and transfer samples.

## 4. Conclusions and remarks

Since the process samples may not be easy to be preserved, a set of independent standardization samples is recommended for the standardization of trans-alkylation process samples. It is because the generic set is always possible to be reproducible whenever and wherever needed. If we compare the residual plots of five constituents (though they are not included in this paper) for the calibration set measured on the master and the test set measured on the slave instrument, we may conclude that the transfer of spectra from slave to master can be successful with the generic and that the content prediction of five constituents is applicable to the process samples after the standardization. PDS and IDS perform enough to be acceptable for the standardization of the process samples with the generic samples. However, it should be careful when to apply standardization for some constituents that are quite different from the generic samples.

In this paper, we consider five constituents at the same time to obtain the reduced but common generic samples. However, if only one constituent is concerned at a time, the standardization with the generic samples is expected to show better performance and the number of required generic samples can be significantly reduced although generic samples may be different among constituents.

## References

[1] C.-S. Chen, C.W. Brown, S.-C. Lo, Appl. Spectrosc. 51 (1997) 744–754.

[2] E. Bouveresse, D.L. Massart, Vib. Spectrosc. 11 (1996) 3–15.

[3] E. Bouveresse, D.L. Massart, Chemom. Intell. Lab. Syst. 32 (1996) 201–213.

[4] M. Forina, G. Drava, C. Armanino, R. Boggia, S. Lanteri, R. Leardi, P. Corti, R. Giangiacomo, C. Galliena, R. Bigoni, I. Quartari, C. Serra, D. Ferri, O. Leoni, L. Lazzeri, Chemom. Intell. Lab. Syst. 27 (1995) 189–203.

[5] Y. Wang, B.R. Kowalski, Appl. Spectrosc. 46 (1992) 764–771.

[6] Y. Wang, D.J. Veltkamp, B.R. Kowalski, Anal. Chem. 63 (1991) 2750–2756.

[7] E. Bouveresse, C. Sterna, J.L. Linossier, D.L. Massart, Analysis 24 (1996) 394–397.

[8] H. Swierenga, P.J. de Groot, A.P. de Weijer, M.W.J. Derksen, L.M.C. Buydens, Chemom. Intell. Lab. Syst. 41 (1998) 237–248.

[9] J. Sjöblom, O. Svensson, M. Josefson, H. Kullberg, S. Wold, Chemom. Intell. Lab. Syst. 44 (1998) 229–244.

[10] S.T. Sum, S.D. Brown, Appl. Spectrosc. 52 (1998) 869–877.

[11] Y. Xie, P.K. Hopke, Anal. Chim. Acta 384 (1999) 193–205.

[12] J.S. Shenk, M.O. Westerhaus, Crop Sci. 31 (1991) 1694–1696.

[13] A. Puigdomènech, R. Tauler, E. Casassas, M. Aragay, Anal. Chim. Acta 355 (1997) 181–193.

[14] L. Duponchel, C. Ruckebusch, J.P. Huvenne, P. Legrand, J. Mol. Struct. 480–481 (1999) 551–556.

[15] B. Walczak, E. Bouveresse, D.L. Massart, Chemom. Intell. Lab. Syst. 36 (1997) 41–51.

[16] R.D. Snee, Technometrics 19 (1977) 415–428.

[17] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137–149.

[18] I. Daubechies, Ten Lectures on Wavelets, CBMS-NSF Regional Conference Series in Applied Mathematics 61, SIAM, Philadelphia, 1992.

[19] K. Chui, Wavelets: A Mathematical Tool for Signal Analysis, SIAM, Philadelphia, 1997.

[20] E. Bouveresse, S.C. Rutan, Y. Vander Heyden, W. Penninckx, D.L. Massart, Anal. Chim. Acta 348 (1997) 283–301.