

## Wavelet Methods in Chemometrics: Quantitative Spectrometric Multicomponent Analysis

U. Depczynski, K. Jetter, J. Stockler  
Universität Hohenheim, Institut für Angewandte Mathematik und Statistik  
D-70593 Stuttgart, Germany  
e-mail: kjetter@uni-hohenheim.de

and

K. Molt, A. Niemöller  
Gerhard-Mercator-Universität Duisburg, FB 6, Instrumentelle Analytik  
D-47048 Duisburg, Germany  
e-mail: molt@lims.uni-duisburg.de

**Keywords:** wavelets, NIR-spectrometry, pretreatment, multicomponent analysis, genetic algorithm

### ABSTRACT

The application of different types of wavelet transforms in spectrometric multicomponent analysis is explored. From the complete set of wavelet coefficients a subset is selected in a two-step procedure. The criterion of selection is an optimal calibration model. The results obtained are comparable to factor analytical methods (like PCR and PLS) which today are used predominantly in the field of chemometrics. One advantage of wavelets is the fact, that in contrast to factor analysis no data preprocessing is necessary.

### INTRODUCTION

The wavelet transform of spectral data can be a useful technique in applied spectroscopy. Recently, such methods were devised for the compression and denoising of spectra [1,2]. The wavelet packet transform of Near Infrared data was used for pattern recognition [3]. In our present work, we use wavelet transforms in connection with the quantitative analysis of chemical mixtures. In example 1 (Figure 1a) we take data from Near Infrared spectra of mixtures of the three xylene-isomers (*o*-, *m*-, and *p*-xylene), and create a calibration model for each component. Example 2 (Figure 1b) deals with synthetic spectra in which two Lorentz profiles with different amplitudes are superimposed. High frequency random noise and various baseline drifts are added. The aim of this study is to develop an automated and fast calibration procedure without the need of extensive and time consuming pretreatment of the spectra. In this short paper we only present the basic principles of our method. A complete description is in preparation and will be published elsewhere.

### Preprocessing of the Spectral Data

In practical quantitative analysis, a proper pretreatment of the spectral data is necessary. Usually, the high frequency noise is accompanied by low frequency components, e.g. drifts of the baseline. Therefore, methods for dealing with noise from different sources are required. For conventional quantitative methods, like e.g. PCR, there is no standard procedure for deciding which pretreatment will give the best results. We propose the use of the wavelet transform as a preprocessor of the data for the following reasons. The vanishing moment property of the wavelet can completely eliminate the drifts of the baseline which are of low polynomial order. Furthermore, the high frequency noise can be reduced by

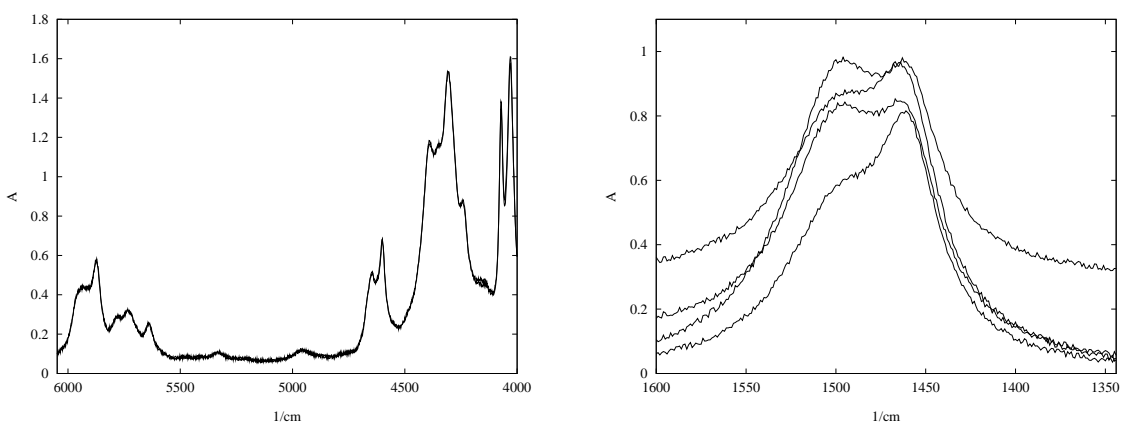


Figure 1. Data sets of example 1 and 2: on the left (a) two exemplary Near Infrared spectra of mixtures of the three xylene isomers, on the right (b) four exemplary spectra of a simulated set with two peaks and different drifts of the baseline.

thresholding of the wavelet transform, which was used in [2]. Finally, we can employ fast numerical algorithms with computational complexity  $\mathcal{O}(N \log_2 N)$  where  $N$  is the length of the spectral data.

### Selection of the Coefficients

Our procedure automatically generates a linear model which uses selected coefficients of the transformed data. For this purpose we are given a certain number of sample spectra  $S_i, 1 \leq i \leq n$ . (We use  $n = 30$  in our experiments.) The transformed data  $S_i^*, 1 \leq i \leq n$ , are lists of wavelet coefficients of the spectra. The property of the spectra for which the calibration is performed is expressed by a vector  $(p_i, 1 \leq i \leq n)$ . The most difficult part is the selection of coefficients. This is done in two steps by a well-designed optimization technique. The choice of too many coefficients may lead to *overfitting*. To avoid this a special functional  $f$  is developed and used for the optimization in the second step which includes a genetic algorithm.

## THE WAVELET TRANSFORM

The spectral data are represented in absorbance units at equally spaced wavenumbers. In our examples we apply three different types of wavelet transforms to the non-pretreated data.

### Sturm-Liouville Wavelets

These are wavelets on the interval  $[0, 1]$  which allow the treatment of non-periodic boundary conditions. They are defined in terms of the eigenfunctions of a Sturm-Liouville operator

$$\mathcal{L}[f] := -\frac{1}{\rho} ((p \cdot f')' - q \cdot f)$$

on the weighted Sobolev space  $H_\rho^2([0, 1])$ . Their general construction is described in [4,5]. Here we use the family of wavelets

$$\psi_{j,k}(x) = \frac{2}{\sqrt{2^j + 1}} \sum_{i=2^j+1}^{2^{j+1}} \cos\left(i\pi \frac{2k+1}{2^{j+1}}\right) \cos(i\pi x), \quad 0 \leq k \leq 2^j - 1, j \geq 1.$$

If we include three scaling functions on the lowest scaling level  $j = 1$ , then the whole family is a Riesz basis for the subspace of  $H^2([0, 1])$  with boundary conditions  $f'(0) = f'(1) = 0$ . (Other sorts of boundary conditions could be implemented by different families  $\psi_{j,k}$ .) Their Riesz bounds are  $2/3$  and  $2$ . Furthermore, there exist fast decomposition and reconstruction algorithms which are mainly based

on the DCT-I transform. The computational complexity of this wavelet transform is  $\mathcal{O}(N \log_2 N)$  for each spectrum  $S_i$ ,  $1 \leq i \leq n$ , if  $N = 2^\nu$  denotes the number of data in  $S_i$ .

### Periodic Wavelets and Frames

The use of periodic wavelets is still common for finite datasets. Let  $N = 2^\nu$  as before. In order to reduce the side lobe effects, we add a linear polynomial to the sequence  $S_i$  such that the new sequence vanishes on the boundary. Then we apply the periodic wavelet decomposition algorithm, see e.g. [6]. In our examples we use the symmlet-8 orthonormal wavelet which was proposed for denoising of signals [7]. Since it has 8 vanishing moments, it eliminates polynomial drifts of the baseline up to degree 7 except for regions near the boundary.

Another type of wavelet transform can be obtained, if one uses more shifts of the mother wavelet on each scaling level. This leads to so-called oversampling frames which were introduced by Chui and Shi [8]. Given a fixed oversampling rate  $r \geq 2$ ,  $r \in \mathbb{N}$ , we let

$$\psi_{j,k/r}(x) := 2^{j/2} \psi(2^j x - k/r), \quad j, k \in \mathbb{Z}.$$

This family defines a *frame* of  $L_2(\mathbb{R})$ , if  $\psi$  is one of the known orthonormal or biorthogonal wavelets. Again we choose  $\psi$  to be the symmlet-8 wavelet in our experiments. The transformation of functions  $f \in L_2(\mathbb{R})$  is defined by the *frame analysis*

$$c_{j,k/r} := \langle f, \psi_{j,k/r} \rangle = \int_{-\infty}^{\infty} f(x) \psi_{j,k/r}(x) dx, \quad j, k \in \mathbb{Z}.$$

For periodic data we use the same periodization technique as above. Fast algorithms for the computation of these inner products are available, if either  $r$  is odd [9] or  $r$  is a power of 2 [10]. The computational complexity has the order  $\mathcal{O}(rN)$  where  $N$  denotes the length of the data. The gain of this transform over the usual wavelet transform (where  $r = 1$ ) depends on the applications. It was shown in [9] that they perform better for echo cancellation or in more general situations, where translation invariance of the transform is important. We observed in our applications, that some of the frame coefficients  $c_{j,k/r}$ ,  $k \not\equiv 0 \pmod{r}$ , have higher correlation with the property of the spectra than the wavelet coefficients. Hence they are likely to be good candidates for generating linear models for calibration.

## CALIBRATION

Let  $S_1^*, \dots, S_n^*$  denote the transformed sample spectra, and  $p_i$  the value of the property of  $S_i$  for which the calibration is needed. We divide them into two subsets, the calibration set  $A = \{S_1^*, \dots, S_m^*\}$  and the validation set  $B$  with the remaining  $n - m$  spectra. The spectra in the validation set are assumed to be independent from those of the calibration set. The validation of a certain calibration is necessary as a check with respect to overfitting.

Let us first describe the criteria used for optimization. The wavelet transformed spectra are treated as unstructured lists; i.e., no preference of any scaling levels is made beforehand. According to certain rules, which are explained below, we choose  $t$  indices of coefficients (where  $t \leq m$ ). Then the best least squares fit for the vector  $(p_i; 1 \leq i \leq m)$ , which contains the properties of the spectra in  $A$ , is determined. This fit only uses the calibration spectra in  $A$ . Let us denote by  $\hat{p}_i$ ,  $1 \leq i \leq n$ , the predicted property of  $S_i^*$  by this linear model. In the absence of a validation set  $B$ , a suitable measure for the standard deviation is

$$\text{SEE} = \left( \sum_{i=1}^m (\hat{p}_i - p_i)^2 / (m - t - 1) \right)^{1/2}.$$

It can be observed that pure minimization of SEE has over fitting as a side-effect. Therefore we compute another quantity from our independent validation spectra in  $B$ . This is called the standard error of analysis in [11] and is defined as

$$\text{SEA} = \left( \sum_{i=m+1}^n (\hat{p}_i - p_i)^2 / (n - m) \right)^{1/2}.$$

The functional  $f$  for optimization is a weighted sum of both SEE and SEA. The most difficult task of the calibration lies in the selection of a small number of coefficients for the linear regression model. In the first selection step a fixed number  $M$  of wavelet coefficients is preselected. In the second selection step a partial minimum of our functional  $f$  is computed by a deterministic procedure. The further optimization can now be viewed as a nonlinear integer programming problem in  $M$  variables which can take only values 0 and 1. A popular class of algorithms for finding near best solutions are genetic algorithms, see [12].

### First selection step

Denote by  $s_{i,k}^*$  the  $k$ -th coefficient in our lists  $S_i^*$ . We write  $r_k$  for the (absolute) correlation of the vector  $(s_{1,k}^*, \dots, s_{m,k}^*)$  with  $(p_1, \dots, p_m)$ . Only the calibration spectra in  $A$  are used here. Then the lists  $S_i^*$ ,  $1 \leq i \leq n$ , are sorted by decreasing value of  $r_k$  and only the first  $M$  coefficients are kept, where  $M$  is a fixed parameter which might depend on the application.

### Second selection step

After this we minimize the functional  $f$  over all subsets of the leading  $K$  coefficients, where  $1 \leq K \leq M$ . This amounts to the solution of  $M$  linear models of small size. In our applications we found that  $M \leq 16$  gives satisfactory results. This simple and rather abbreviated method, however, was not regarded as optimal. To improve the selectivity and to allow for all combinations of wavelet coefficients to be taken into account for the optimization we developed a genetic algorithm for coefficient selection.

The genetic algorithm (GA) uses chromosomes which are bitstrings of length  $M$  in our case. If the  $\mu$ -th bit (called a gene) is on, then the coefficient at index  $\mu$  is selected for the linear model. The so-called fitness function of the GA is the functional  $-f$ , hence finding a chromosome of maximal fitness solves our minimization problem. The genetic algorithm produces subsequent generations (which are collections of chromosomes) by certain rules which involve random parameters. As a starting population we take randomly generated chromosomes and include the best chromosome which was found at the end of the deterministic procedure (so-called *hybrid GA*). Several operators from [12] for creating the child generation are implemented: uniform crossover, mutation and invaders. The procedure starts with high probability for uniform crossover and low probability for mutation and invaders and ends with the opposite distribution of probabilities. Our experiments show that this technique often improves the optimal value for  $f$  by a factor 1.5 upto 4.

## NUMERICAL RESULTS

For our first example we use the three Near Infrared spectra of the pure isomers *o*-xylene, *m*-xylene, and *p*-xylene, measured on a Perkin Elmer 1700X NIR-FTIR-spectrometer using 16 scans with optical resolution of  $4 \text{ cm}^{-1}$ , while digital resolution was set to  $2 \text{ cm}^{-1}$ . Then 30 sample mixtures  $S_i$  are generated by superposition of the three spectra. The reason for our use of synthesized rather than measured spectra of mixtures lies in the precise control over noise parameters, which we can introduce into the calibration. In this example we add random noise with an amplitude of 0.01 absorbance units (AU) to each  $S_i$ . All mixtures contain *m*-xylene as main component (80{93 %) and *o*- and *p*-xylene (3{10 %) as minor components. Two examples, the graphs of  $S_5$  and  $S_{10}$ , are shown in Figure 1a. Visible differences only occur in small regions of the spectrum, e.g. near  $4100 \text{ cm}^{-1}$ . We use 20 spectra for the calibration set  $A$  and the remaining 10 spectra for validation, as described in the previous section.

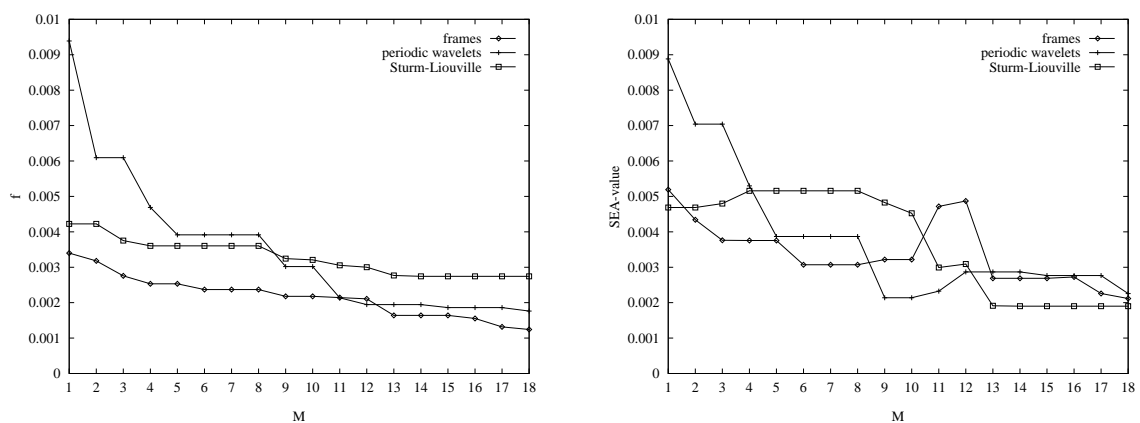


Figure 2. Calibration of the component *o*-xylene: the *x*-axis refers to the number  $M$  of preselected coefficients; on the left (a) the optimal values of  $f$ , on the right (b) the SEA-values for the prediction of 10 independent spectra.

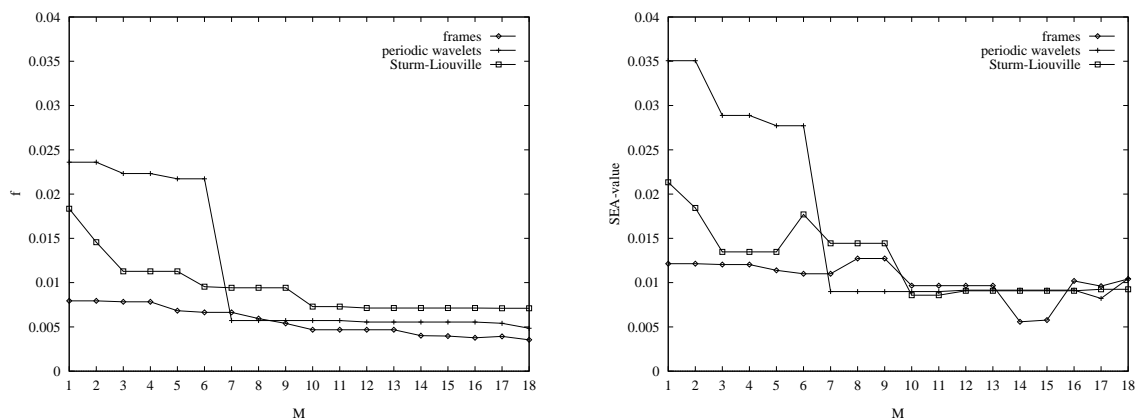


Figure 3. Calibration of the amplitude of the second peak of synthesized spectra from superposition of two Lorentz profiles, with high and low frequency noise added; on the left (a) the optimal values of  $f$ , on the right (b) the SEA-values for the prediction of 10 independent spectra.

Figure 2a shows the results obtained for the calibration of the component *o*-xylene in the mixtures. The values on the *x*-axis denote the number  $M$  of coefficients preselected in the deterministic procedure. The optimal value of the functional  $f$ , which is found by the genetic algorithm with at most 40 generations of maximal size 200, is given on the *y*-axis. CPU-times for the calibration with a fixed  $M = 16$  are around 17 sec on an IBM 3BT Workstation with Power2-processor. The different curves in Figure 2a represent the calibration results for the three types of wavelet transforms which are applied to the datasets  $S_i$ .

In order to test these calibration models, we produced an independent set of 10 spectra  $T_j$ ,  $1 \leq j \leq 10$ , by the same principle as above. The SEA-values for the component of *o*-xylene in these mixtures is shown in Figure 2b. It turns out that the precision of the calibration with SEA-values between 0.002 and 0.003 is satisfactory for all three wavelet types. These results agree well with a calibration by classical *Principal Component Regression* (PCR), a common method in chemometrics [13]. The best calibration model based on PCR using smoothed data gives an SEA-value of 0.002 for the spectra  $T_j$ . We also should mention that in the calibration step the oversampled frame performs better than the other wavelet types. This advantage of frames however disappears in the prediction of independent spectra  $T_j$ . These observations are also true for the other two components of the mixtures, which are not shown here.

Our second example uses synthetic data which consists of two peaks of Lorentz profile

$$A(W) = A_0 \frac{H^2}{H^2 + 4(W - W_m)^2}, \quad 1600 \text{ cm}^{-1} \geq W \geq 1344 \text{ cm}^{-1},$$

with different types of noise added to the spectrum. Here  $W_m$  denotes the peak position on the wave-number-axis ( $W$ ),  $A_0$  is the amplitude and  $H$  denotes the half-band width of the profile. We generated 30 synthetic spectra  $S_i$  with the same location and half-band widths of the peaks ( $W_1 = 1500 \text{ cm}^{-1}$ ,  $H_1 = 70 \text{ cm}^{-1}$ , and  $W_2 = 1460 \text{ cm}^{-1}$ ,  $H_2 = 40 \text{ cm}^{-1}$ ). There is a large variation of the amplitudes of the peaks from 0.44 to 1.02 (peak at  $1500 \text{ cm}^{-1}$ ) and from 0.303 to 0.7 (peak at  $1460 \text{ cm}^{-1}$ ). High frequency random noise of amplitude 0.01 is added to each spectrum. Various drifts of the baseline were used for the different spectra, as can be seen in Figure 1b.

The results of the calibration for the amplitudes of the second peak are shown in Figure 3a. As before, the three curves refer to the results obtained with the three different wavelet types. We used 20 spectra for calibration and 10 spectra for validation. The values of SEA depend on the number  $M$  of preselected coefficients. The SEA-values for the amplitude of the second peak of an independent set of 10 spectra  $T_j$ ,  $1 \leq j \leq 10$ , are shown in Figure 3b. These spectra have the same location and half-band width as before, but different drifts of the baseline and higher noise levels (0.02 for three spectra and 0.3 for one spectrum) were included here. We can draw the same conclusion as for example 1. The genetic algorithm finds smaller values for  $f$ , if the wavelet transform uses the oversampled frames. The test with independent spectra  $T_j$ , however, leads to comparable precision for all three wavelet transforms.

## ACKNOWLEDGMENTS

This research was supported by the Deutsche Forschungsgemeinschaft, project Je 148/7-1.

## REFERENCES

- [1] V.J. Barclay, R.F. Bonner, I.P. Hamilton, *Application of wavelet transforms to experimental spectra: smoothing, denoising, and data set compression*, Anal. Chem. 69, p. 78 (1997).
- [2] F.T. Chau, T.M. Shih, J.B. Gao, C.K. Chan, *Application of the fast wavelet transform method to compress ultraviolet-visible spectra*, Appl. Spectroscopy 50, p. 339 (1996).
- [3] B. Walczak, B. van den Bogaert, D.L. Massart, *Application of wavelet packet transform in pattern recognition of near-IR data*, Anal. Chem. 68, p. 1742 (1996).
- [4] U. Depczynski, *Konstruktion von waveletartigen Zerlegungen auf kompakten Intervallen mit Hilfe der Eigenlosungen Sturm{Liouvillescher Randwertprobleme*, PhD Thesis, University of Duisburg (1995).
- [5] U. Depczynski, K. Jetter, *Multiscale decompositions on the interval based on DCT{I transforms*, in: Advanced Topics in Multivariate Approximation, F. Fontanella, K. Jetter, P.J. Laurent (eds.), World Scientific, Singapore, p.57 (1996).
- [6] J. Buckheit, S. Chen, D. Donoho, I. Johnstone, J. Scargle, *WaveLab Reference Manual*, Stanford University (1995).
- [7] R.R. Coifman, D.L. Donoho, *Translation-invariant de-noising*, in: Wavelets and Statistics, A. Antoniadis, G. Oppenheim (eds.), Springer-Verlag, New York, p. 125 (1995).
- [8] C.K. Chui, X. Shi, *Bessel sequences and a new frames*, Appl. and Comp. Harmonic Analysis 1, p. 29 (1993).
- [9] C.K. Chui, J.C. Goswami, A.K. Chan, *Fast integral wavelet transform on a dense set of time-scale domain*, Numer. Math. 70, p. 283 (1995).

- [10] J. Stockler, *Preconditioning of the frame algorithm*, Proceedings of 1<sup>st</sup> Workshop on Large Scale Scientific Computations, Varna, to appear.
- [11] A. Niemoller, *Quantitative NIR-Spektrometrie wa riger Losungen starker Sauren und Basen*, Diploma Thesis, University of Duisburg (1995).
- [12] L. Davis, *Handbook of Genetic Algorithms*, Van Nostrand, New York (1991).
- [13] H. Martens, T. N s, *Multivariate Calibration*, John Wiley and Sons, New York (1989).