TEMPERATURE-ROBUST MULTIVARIATE CALIBRATION

TEMPERATURE-ROBUST MULTIVARIATE CALIBRATION

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. mr. P.F. van der Heijden
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 30 juni 2004, te 10.00 uur

door

FLORIAN WÜLFERT

geboren te Friedrichshafen

Promotor:

prof.dr. A.K. Smilde

Co-promotor:

dr. W.Th. Kok

FACULTEIT DER NATUURWETENSCHAPPEN, WISKUNDE EN INFORMATICA

To my parents

Promotiecommissie:

prof.dr. L.M.C. Buydens (Katholieke Universiteit Nijmegen)

prof.dr. C. Gooijer (Vrije Universiteit, Amsterdam)

prof.dr. P.J. Schoenmakers (Universiteit van Amsterdam)

prof.dr. H.W. Siesler (Universität Duisburg-Essen, Germany)

prof.dr. B. Smit (Universiteit van Amsterdam)

TABLE OF CONTENTS

1. GENERAL INTRODUCTION

In the last years the number and range of chemometric applications has increased considerably due to an increase in both demand and supply.

On the demand side it has been the development, automation and digitization of chemical analytical techniques that has stimulated the need for applied chemometric techniques. Not only is a spectrum or chromatogram hardly ever plotted anymore on paper as primary way of recording, the nowadays common digital storage also opens the possibility of advanced data analyses long after the chemical analysis is finished. Furthermore, the technical improvements of the analytical instruments have led to a much higher flow of data resulting typically in data file sizes in the order of megabytes per experiment. These developments have consequently led to a flow of analytical data that cannot be handled anymore in the traditional ways. Another type of demand is stimulated not so much by the complexity of the data analysis but by the necessary speed: The development of fast, versatile but non-selective spectroscopic process analyzers. These new instruments and methods are optimized for short analysis time, instrumental robustness and linear response over a wide range rather than selectivity and accuracy. Consequently they can often only be used in combination with multivariate calibration methods that are able to compensate for the lack of selectivity and consequent interferences.

On the supply side two major trends make it possible to meet the increased demand described above. A first aspect is that chemometrics, - as all other computationally intensive disciplines -, can take advantage from the fact that computing power and data storage capacities double every 1½ to 2 years. Second, and perhaps as a consequence of the first, chemometrics has created a vast and even still growing collection of algorithms and tools applicable to a wide spectrum of analytical problems. This does not mean that theoretical advancement has become unnecessary. The data analysis of multidimensional measurements is still an evolving field, but the work of

the average chemometrician is shifting from algorithm development towards application and method development.

In this more practical line of work robustness, simplicity and data preprocessing play an equally important role in the model as the multivariate algorithm itself. Generally, while assessing which methods, algorithms or strategies are possible, different aspects have to be addressed. Before evaluating the possible choices, the aim and setting of the model to be built have to be considered: Is the model only descriptive, predictive or both; how many samples can be measured; what is known about the analysis and its environment? These considerations give the boundaries between which a solution must be found. In a next step, the more technical aspects have to be addressed; a few examples are given below:

- Interference: Measurements are influenced by variation other than the one induced by the analyte or entity to be modeled. Noise or interfering species with different gradations of correlation can make modeling more difficult and complex, especially if the interferences are difficult to map or quantify a priori.
- Linearity: The question is whether the problem at hand is fundamentally linear (Lambert-Beer's law) or not (exponential decay). Furthermore, the problem might be linear in principle but not in practice due to a deviation from the ideal case, e.g., saturation effects. The method to tackle the non-linearity, e.g. linearization by preprocessing, approximation with a linear model or use of a non-linear algorithm, will vary and depend on the data.
- Time, stability and long term effects: Depending on the application, a model has to be valid over different time spans. In the case of a "long-life" model changes in environment or aging of apparatus might lead to the necessity of additional standardization strategies.

Considerations and aspects mentioned above are obviously only the most common and can be extended considerably in more specific cases. Therefore, the work presented in the next chapters cannot be comprehensive but will deal with only certain aspects of linearity and interference.

More specifically, the work concentrates on non-linear interferences on spectra hampering the use of linear multivariate models and the possible ways to deal with it. As an illustration, temperature-induced absorption shifts in short-wave Near-Infrared spectra are chosen for the non-linear interference. The shifting absorption bands are fundamentally non-linear because they can be described as response-shifts between information channels: Each wavelength represents an information channel where the information about the concentration and identity are represented by the absorption (response) and wavelength (channel) respectively. An absorption shift over different wavelengths, or in other words a leakage between the channels, can never be described by a linear function. Partial Least Squares (PLS) regression models will stand as example for linear models, since PLS is one of the most often applied multivariate linear calibration models. Due to the non-linear nature of the interference, the multivariate linear model will either identify the shifted bands as “new” components, incorporating them in a more complex model, or significantly loose accuracy in a model with constant complexity. The exact effect of the non-linearity on the model depends on the calibration and preprocessing strategy used.

Therefore, Chapters 2 to 5 will attempt to give an overview of the possibilities to deal with and eventually correct for non-linear interferences. Generally two main strategies can be identified: the either implicit or explicit inclusion of the interference into the model or the correction for the interference previous to the calibration model itself. Chapter 2 describes how the temperature fluctuations affect the short-wave Near-Infrared spectra of the chosen mixtures, leading to a shifting absorption band. Using

a synthetic set of signals (a shifting Gaussian peak) and a simple descriptive multivariate model (Principal Component Analysis) the effects of response shifts on linear models are visualized. Finally, different ways to include the interfering temperature into the regression models are compared to the regression without the non-linear interference. In Chapter 3 methods with a more explicit inclusion of the temperature into the model and a variable elimination method using the same PLS algorithm as for the final regression model are tested and compared to the previous models. Furthermore the (im)possibilities of linear basis transformations are evaluated more fundamentally in order to assess the limits of linear approaches to tackle non-linearities. In Chapter 4 an extension of a preprocessing algorithm for discrete correction situations (Piecewise Direct Standardization correction for e.g. two instruments, PDS) is developed to accommodate the continuous character of temperature fluctuations. This continuous PDS technique results in a non-linear pre-processing without resorting to elimination or selection of variables. In Chapter 5, a variable selection based on a probabilistic algorithm (Simulated Annealing) is examined in order to assess the possibilities of a more sophisticated variable selection technique and in order to compare the results with those of some of the earlier models.

All the above-described models and strategies use inclusion or pre-processing and may require information about and/or additional measurements of the non-linearly interfering temperature. In order to enable a quick overview of the approaches used and information needed, the most important characteristics of the presented methods are summarized in the following table (Table 1-1).

Table 1-1: Models and strategies used

Category and	Model Type	Chapter	Pre-processing	Temp. known for Calibration	Temp. known for Prediction
Implicit inclusion	Global	2	None	✗	✗
Explicit inclusion in calibration model	Local + interpolation	2	None	✓	✓
	Incl. in X	3	None	✓	✓
	Incl. in Y	3	None	✓	✗
Data pre-processing	2-step PLS	3	Linear	✓	✗
	Basis projection	3	Linear	✗	✗
	Var. selection PLS-UVE	3	Non-linear	✓	✗
	CPDS	4	Non-linear	✓	✓
	Var. selection SA	5	Non-linear	✗	✗

✓: Knowledge of temperature is required in order to be able to use the model.

✗: Temperature is not required but in case of calibration it should be possible to assume, by e.g. the size of the dataset, that the temperature variation is well spread in order to be excluded as a confounding factor.

2. INFLUENCE OF TEMPERATURE ON VIBRATIONAL SPECTRA AND CONSEQUENCES FOR THE PREDICTIVE ABILITY OF MULTIVARIATE MODELS.

Abstract

Temperature, pressure, viscosity and other process variables fluctuate during an industrial process. When measuring vibrational spectra on- or in-line for process analytical and control purposes, the fluctuations influence the shape of the spectra in a non-linear manner. The influence of these temperature induced spectral variations on the predictive ability of multivariate calibration models is assessed. Short wave NIR spectra of ethanol/water/2-propanol mixtures are taken at different temperatures and different local and global partial least squares calibration strategies are applied. The resulting prediction errors and sensitivity vectors of a test set are compared. For data with no temperature variation, the local models perform best with high sensitivity but the knowledge of the temperature for prediction measurements cannot aid in the improvement of local model predictions when temperature variation is introduced. The prediction errors of global models are considerably lower when temperature variation is present in the dataset but at the expense of sensitivity. In order to be able to build temperature-stable calibration models with high sensitivity, a way of explicitly modeling the temperature should be found.

Based on: Wülfert, F.; Kok, W.Th.; Smilde, A.K.; *Anal. Chem.* **1998**, *70*, 1761-1767.

Introduction

General

Mid infrared, Near-Infrared (NIR) and short-wave NIR spectroscopic techniques in combination with multivariate calibration are finding an increasing range of applications in process analysis [1, 2, 3, 4, 5, 6, 7]. The spectroscopic analysis can be done in- or on-line and, in contradiction to slower classical off-line techniques, the results can be used for process control purposes.

The high sensitivity and consequently short pathlengths (in the range of a few μm) of mid-IR instrumentation is often not compatible with industrial environments. With the orders of magnitude lower absorbance of the overtones in NIR and short-wave-NIR , much more robust flow cells can be used which are not susceptible to blockage.

By moving the measurement from the well controlled laboratory to the process environment, external process variables such as temperature, pressure, flow turbulence will also affect the measurements. The difficulty to keep these variables constant or even the inevitability to change their value during the process (e.g. temperature programming in batch processes) makes it necessary to study the influence on the spectra and therefore also on the calibration models.

Temperature effects on vibrational spectra

Vibrational spectra from liquid and solid samples do not only show isolated molecular features, such as structure and functional groups, but also inter- or intramolecular features, such as hydrogen bonding. These weaker forces influence the vibrational modes[8, 9, 10, 11, 12, 13, 14, 15, 16, 17] of molecular bonds but are themselves affected by conditions such as temperature and pressure. Therefore the variations in, e.g., temperature translate via the changes in intermolecular forces to modifications of the vibrational spectra.

The influence of the temperature on the O-H stretch band and its overtones has been described in various articles[18, 19, 20, 21, 22]. The hydroxyl group gives rise to two bands for its stretching mode: a sharper band for the “free” O-H groups and a broader one for the stretch mode of hydrogen-bonded O-H groups. The broad band, which can be seen as an overlay of many bands that belong to different cluster sizes formed by hydrogen bonding, is shifted towards lower energies (higher wavelength) relative to the free O-H stretch. Rising the temperature decreases the average cluster size and increases the relative absorbance of free groups[23].

This can be seen most clearly in water spectra where the hydroxyl band shifts to the lower wavelengths and becomes sharper when the temperature is increased. The increase of free O-H groups can also be observed for alcohols, but a combination C-H stretch mode that absorbs in the same region makes the effect less apparent. Similar effects can be observed for spectra of polyamides and polyurethane, where the N-H groups can form hydrogen bonds[24, 25, 26]. The bands originating from N-H stretching modes are influenced by the temperature much in the same way as for hydroxyl groups.

Effects of shifts and peak distortion on multivariate regression

Due to a lack of selectivity NIR applications consist mostly of spectroscopic measurements in combination with multivariate data analysis. Partial Least Squares (PLS) and Principal Component Regression (PCR) are the most common methods. Both methods assume linear additivity. This means that absorption spectra are supposed to increase linearly with the concentration (linearity) and that a mixture of components gives a spectrum that is a linear combination of the pure spectra (additivity). Any deviation from this ideal behavior has to be approached by using more components in the PLS or PCR model.

Spectra that exhibit shifts or other changes in their shape do not conform to the linearity demand and consequently a multivariate model will have to use

more regression factors than is to be expected by the chemical rank (number of components in the mixture).

Scope of this chapter

To study the effect of external variation on the predictive ability of multivariate calibration for spectral data, the temperature has been chosen as the external variable. Short-wave NIR spectra, measured at different temperatures, of mixtures containing ethanol, water and 2-propanol are used as data and PLS regression is employed as data analysis method. Two different types of PLS models are compared: local models that apply to samples of one temperature and global models that can be used for samples at different temperatures. The difference in prediction error for the different models is used to evaluate which calibration strategy can handle temperature-influenced spectra. Explanation of the differences in predictive ability is sought by inspecting the sensitivity vectors for the analytes.

Experimental section

Apparatus

Mixtures of ethanol, water and 2-propanol have been prepared using an analytical balance and kept in airtight sample flasks. Fresh p. a. quality alcohols and sub-boiled water have been used. Closed quartz cells with 1 cm path length have been used in order to prevent dissipation of the alcohols during the measurement. The spectra have been taken on a HP 8453 Spectrophotometer with a thermostatically controlled cell holder and cell stirring module (Hewlett Packard, Palo Alto, CA, USA). The wavelength range used was 580 to 1091 nm with 1 nm resolution and the integration time was 20 s. The collection of the spectra was done on a Hewlett Packard Vectra XM2 PC using the UV-Visible Chemstation software (Rev A.02.04). The temperature of the sample has been regulated using an external Pt-100 sensor immersed in the sample and linked to the controller of a Neslab microprocessor EX-111 circulator bath.

For simulations and the data processing Matlab (ver. 4.2 and 5; The Mathworks Inc.) and the PLS toolbox (ver. 1.4) were used on a Pentium-class computer.

Mixture design

In order to span the concentration variation a mixture design (Figure 2-1) has been set up. The mole fraction levels that obey this design have been mixed and are given in Table 2-1. In order to perform linearity and additivity tests, the spectra of the pure components have also been measured.

The 19 mixtures and the three pure components have been measured at temperatures of 30, 40, 50, 60 and 70°C (± 0.2°C).

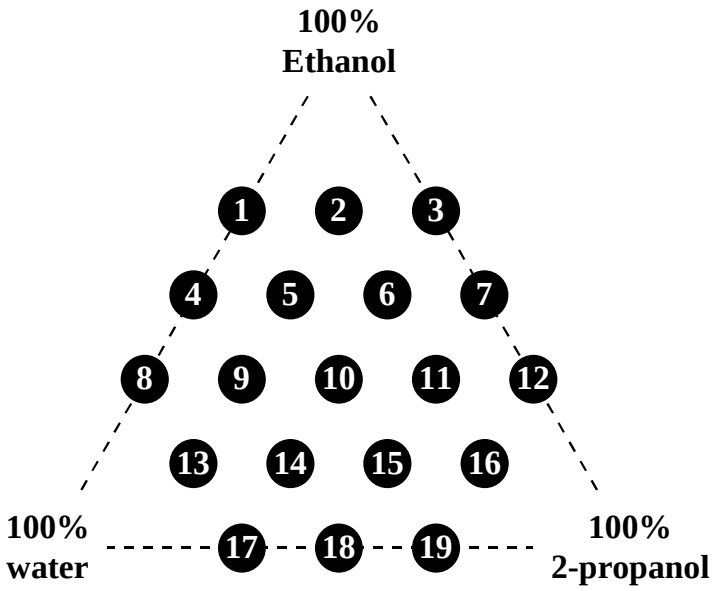


Figure 2-1: Mixture design for ethanol, water and 2-propanol mole fractions.

Table 2-1: Mole-fractions of the samples in %

	ethanol	*water*	*2-prop.*
1	66.4	33.6	0
2	67.2	16.3	16.5
3	66.6	0	33.4
4	50.0	50.0	0
5	50.0	33.3	16.7
6	49.9	16.7	33.3
7	50.0	0	50.0
8	33.3	66.7	0
9	33.2	50.0	16.7
10	33.3	33.4	33.3
11	32.2	16.6	51.2
12	33.5	0	66.5
13	16.6	66.7	16.7
14	16.7	50.0	33.3
15	16.6	33.3	50.1
16	16.2	16.3	67.5
17	0	66.7	33.3
18	0	50.0	50.0
19	0	33.4	66.6

Data analysis

Pretreatment and analysis of experimental data

The measured spectra are pretreated to remove instrumental baseline drift. Straight lines are fitted through the wavelength range 749-849 nm, where no absorbance bands are present, and subtracted from the spectra. The data analysis is performed on the region 850-1049 nm. The absorption at lower wavelengths is too low to be considered significant and absorption above 1050 nm is very noisy due to instrumental effects.

The data analysis consists of PLS1 regressions using the mean-centered pretreated spectra as **X**-block and mean-centered mole fractions for each chemical component separately as **y**-vector. For the different models that will be used the data is always split into a training set for building the respective model and a test set for estimating the predictive quality of that model. When building the model, cross validation techniques are used to estimate the number of latent variables (LV's).

PLS models have been built for each temperature separately (local models) and for the full dataset containing all temperatures (global models). These two cases are fundamentally different when used for prediction of new samples.

Local models

When building small, local models for each temperature it is also necessary to know the temperature of the new samples in the prediction step, otherwise it is not possible to choose one of the local models. If a model and a prediction sample are measured at the same temperature, the mole fraction can directly be predicted (case *a*). Another possibility is that the temperature of the new sample falls in between the model-temperatures (case *b*). In the latter case the estimated concentration of the new sample from one of the models is expected to be biased. In order to achieve a

better prediction, the mole fraction can be estimated by interpolating between the results of the models.

Case *a*: At each temperature models for each chemical compound are built from samples that are on the “edge” of the experimental design (samples 1, 2, 3, 4, 7, 8, 12, 13, 16, 17, 18, 19) and the sample in the “center” (sample 10). The test set is given by the remaining concentration levels (samples 5, 6, 9, 11, 14, 15). As can be seen from the graphical representation in Figure 2-2 A, no extrapolating prediction will be done. The results from local models case *a* can also bee seen as a “best case scenario” considering that temperature does not play any role.

Leave-one-out cross validation is used to establish the number of LV’s in all models, using the prediction error for the left-out samples and visual examination of the loading as criteria.

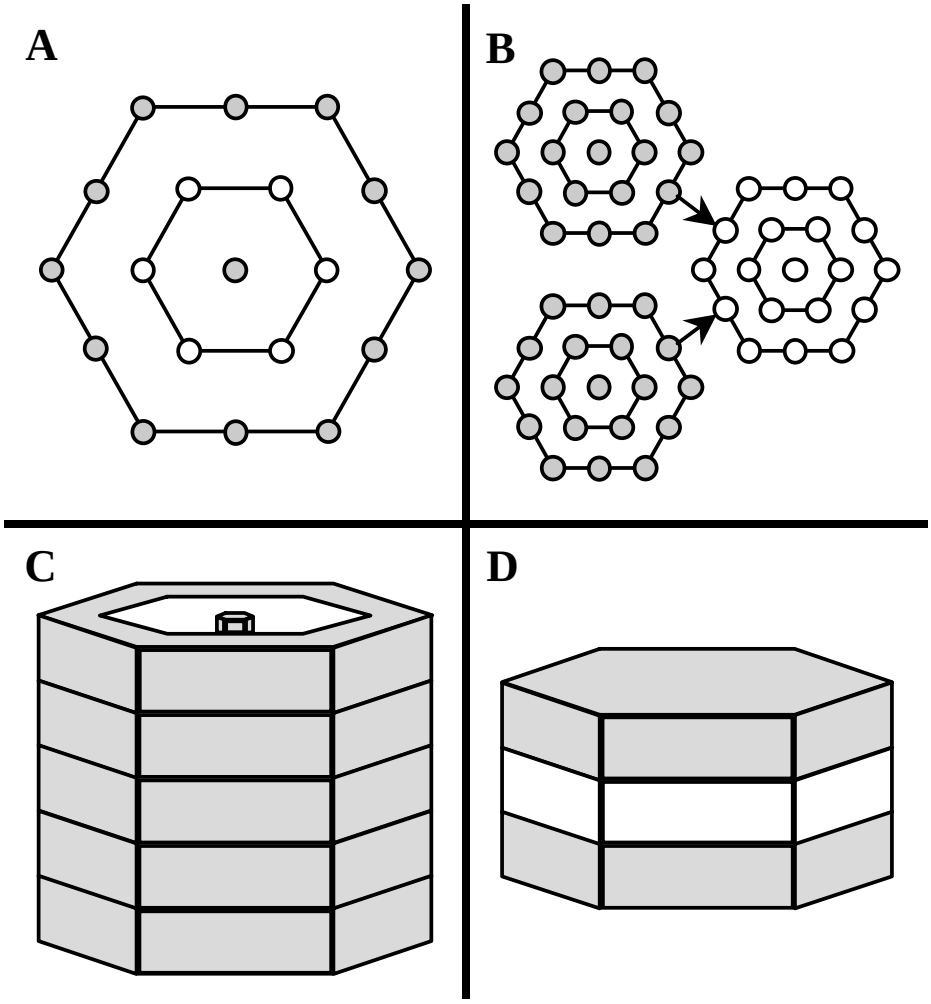


Figure 2-2: Graphical representation of training (gray circles and areas) and test (white circles and areas.) sets. A: Local models case *a*; B: Local models case *b*; C: Global models case *a*; D: Global models case *b*; .

Case *b*: Since the test set consists of samples measured at a different temperature all samples from the experimental design can be used for building the model. Three models are built from the spectra at 30, 50 and 70°C and the prediction samples are the spectra measured at 40 and 60°C (See Figure 2-2 B).

The models are build with the same number of LV's as established for the models in case *a*. The mole fractions of the test set are estimated by averaging the predicted mole fraction resulting from the two models at the nearest temperatures [1].

$$\hat{\mathbf{y}}_{40°C} = \tfrac{1}{2}\left(\hat{\mathbf{y}}_{30°C} + \hat{\mathbf{y}}_{50°C}\right) \quad ; \quad \hat{\mathbf{y}}_{60°C} = \tfrac{1}{2}\left(\hat{\mathbf{y}}_{50°C} + \hat{\mathbf{y}}_{70°C}\right) \quad [1]$$

Global models

With one global model for all temperatures it is neither necessary to know the temperature of a new sample to be predicted nor that of the training set samples. The global model treats temperature as an unknown interferent. PLS uses the covariance between **X** and **y** to establish a regression model that explains the variation in **y** with variation in **X**. If the spectrum of the interferent correlates perfectly with that of the analyte, the PLS algorithm cannot distinguish between analyte and interferent. The weaker the correlation between interferent and analyte becomes, the easier the PLS algorithm can distinguish between them. The spectrum of temperature (if seen as interferent) is strongly non-linear and different from that of the chemical compounds. It may therefore be advantageous but not necessary to know the temperatures of the training samples and to vary temperature independently from the concentrations in order to minimize the covariance between them.

The differences between a prediction sample with a temperature that "fits" into a model (case *a*) or a sample with a temperature that falls in between models (case *b*) does not apply to general models. The temperature is assumed unknown and the cases can therefore not be distinguished.

For comparison of the predictive abilities however, it is useful to build global models that use exactly the same test and training data as the local models.

Case *a*: The same mixtures are used as training and test sets as in the local models. Instead of building 5 models for the 5 temperatures, all training sample measurements are used to build one global model and to predict all measurements of the test set (see Figure 2-2 C)

Leave more out cross validation was performed on the training set leaving one concentration out for all temperatures at each cross validation step. In this way the disturbance of the design by the left out samples is comparable to that during the cross validation used in the local case. Because of the higher number of training samples it is possible to apply additionally a stratified leave out procedure for verification. The difference between stratified and leave-one-concentration-out strategies is, that with stratified five different mixtures (one per temperature) are left out at random, which is repeated until all concentrations have been left out once for each temperature.

Case *b*: Again the same data is used for training and test sets as with the local models. Two models are made: one using all mixtures at 30 and 50°C for building the model and all mixtures at 40°C for prediction. The other model uses all mixtures at 50 and 70°C as training set and all mixtures at 60°C as test set (see Figure 2-2 D).

The number of LV's used is equal to that of the global model case *a*.

Performance measures

Prediction errors: The root mean squared error (RMSE) is used as performance criterion in cross validation (RMSECV), where it is used to estimate the necessary number of LV's, as well as in prediction (RMSEP), where it is used to asses the predictive power of the model. In both cases the RMSE is calculated in the common way as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{n}} \qquad [2]$$

where $\hat{y}_i$ and y_i are respectively the predicted and real values of sample *i* of the *n* samples in either the cross validation or test set.

In order to place the prediction error in a more recognizable setting the mean relative error (MRE) is also used to summarize the results for each type of model. This way an impression can be given on how many percent the prediction is inaccurate.

Sensitivity vectors: In classical first order univariate calibration the sensitivity is an important characterization of a calibration model. It can be calculated as the difference in net analyte signal (response without the offset) of two measurements at different concentrations resulting in the slope of the calibration line. The higher the sensitivity, the better the model performs, since even slight differences in analyte concentration give a distinctively different response.

Recently a method has been proposed to determine the net analyte signal (NAS) and the sensitivity vector not only for classical univariate and multivariate calibration but also for inverse multivariate calibration methods such as PLS[27]. This extension means that no longer all pure spectra and all concentrations have to be known. The method consists of reconstructing the **X** data (response matrix) by its description used in the calibration model (product of x-loading and x-score blocks). By applying rank annihilation it is possible to eliminate the part of the reconstructed response which is contributed by the analyte. The result is an estimation of the response matrix of only the interferents without the analyte. In classical multivariate calibration the pure spectrum is needed for the rank annihilation step.

Lorber et al.[27] show that a linear combination of mixed spectra can also be used, as long as the analyte is present in those spectra. The NAS can then be estimated as the part of a new measurement that is not described by and therefore orthogonal to the interferents-response matrix. The norm of the NAS vector is (for the linear case) proportional to the concentration. Division of the NAS vector by the sample concentration leads to a sensitivity vector for each of the new measurements. Ideally all sensitivity vectors for new samples are the same but in practice they form only estimates of the concentration-normalized pure spectrum.

When applying net analyte signal, its norm and sensitivity as figures of merit, precautions have to be taken in the case of mean centered data. The linear combination of mixed spectra used in the rank annihilation step cannot be the sum of all spectra from the training set, since they sum up to zero. Therefore spectra with the highest analyte concentration (for ethanol: samples 1, 2, 3; for water: samples 8, 13, 17; for 2-propanol: samples 12, 16, 19) have been chosen. Furthermore, prediction samples with an analyte concentration very near to the mean concentration show a sensitivity vector consisting only of amplified measurement noise, since both NAS and concentration will become almost zero. Because of this artifact, only sensitivity vectors of test samples with a different mole fraction than the mean (one third) and common to all test sets are used for interpretation and comparison (for ethanol: samples 5, 6, 14, 15; for water: samples 6, 9, 11, 14; for 2-propanol: samples 5, 9, 11, 15) .

Results and Discussion

Simulations

To assess the influence of spectral shifts and broadening on multivariate models simulations have been carried out. Especially the increase of complexity (the number of principal components needed to describe the data) was estimated.

Three data sets of Gaussian peaks showing either an increase in area, a shift or changing width were generated and Principal Component Analysis (PCA) was applied to these mean centered datasets. The loadings and scores of the datasets (Figure 2-3) show that only variation in area is a linear phenomenon. Variation in the position of the maximum or in the width of the Gaussian peaks lead to a PCA description with more then one principal component (PC).

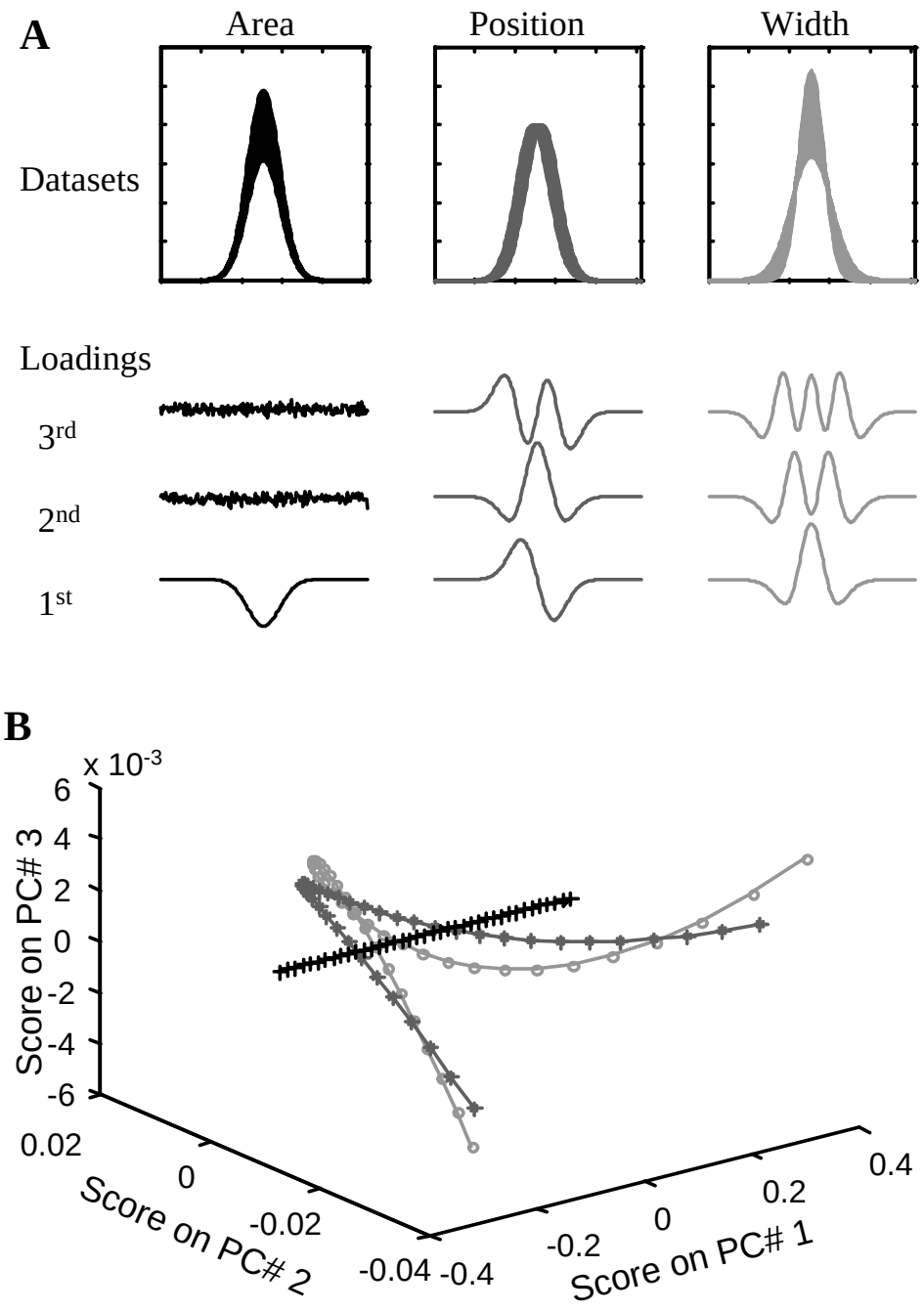


Figure 2-3: Changing area, position and width of a peak and it's effects on multivariate space. *A:* Datasets and loadings. *B:* Score values Area; Position; Width.

It is shown clearly that for the area variation, only the first loading vector has any meaning whilst the second and third merely describe the white noise that has been added. The score plots also show the linearity for the area increase data, since only the first PC contains significant score values.

Contrary, the loadings for the shift and broadening datasets show systematic information even at higher PC's than shown here, until finally noise level is reached. Their respective score plots have a clear 3-D character (corkscrew) since they are non-linear effects and have to be approached by several principal components. An increase in complexity can therefore also be expected for spectra that show shift or broadening of bands.

Qualitative analysis of the data set

Spectra of the pure components have been measured for qualitative evaluation of the temperature effects and testing linearity. Figure 2-4 gives a good impression of the temperature effects on the absorption bands, the band assignments were done using the spectra shown by Bonanno et al.[17]. For water a temperature increase leads to a band shift towards lower wavelengths together with an absorption increase and band narrowing. Rising the temperature decreases the cluster size of hydrogen bonded molecules and increases therefore the fraction of "free" hydroxyls. The alcohols show a very slight decrease of the 3rd C-H overtone, an increase in free O-H and probably some increase in the C-H combination band.

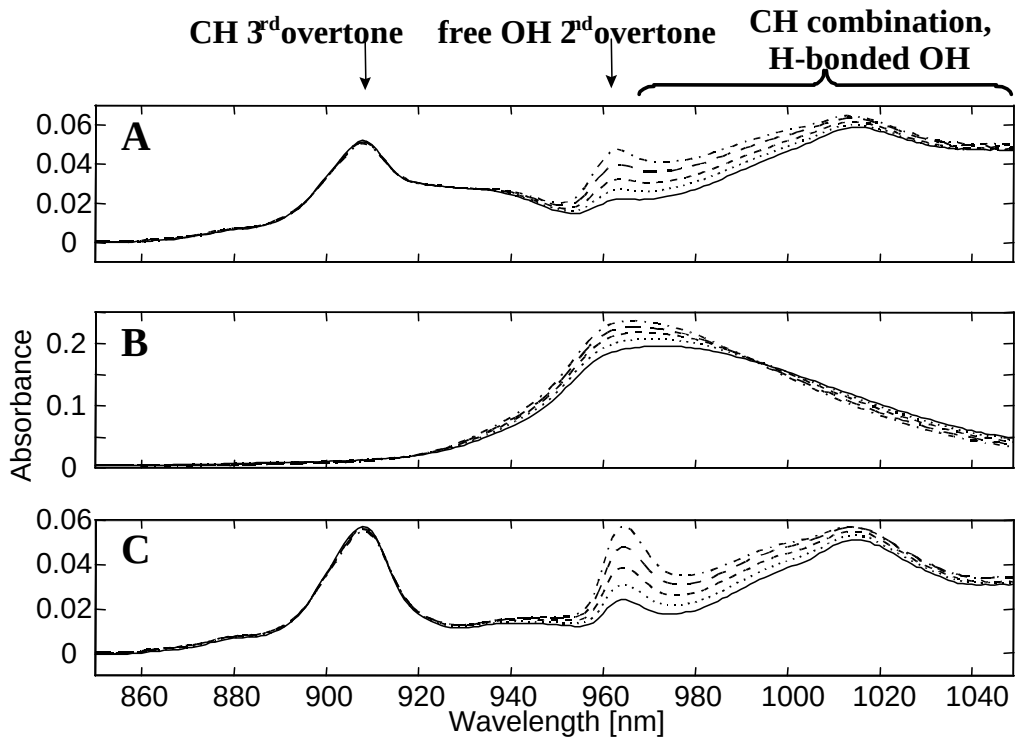


Figure 2-4: Spectra of the pure components at different temperatures (——30°C ····· 40°C - - - - 50°C – – 60°C and - · - · - 70°C); *A* ethanol, *B* water, *C* 2-propanol

In order to test the linearity and additivity synthetic spectra have been composed by addition of the pure component-spectra multiplied with the concentration levels as in Table 2-1. These synthetic spectra were compared with the measured spectra. In Figure 2-5 the differences between some synthetic and real spectra are shown. Deviation from linearity and additivity were especially found with mixtures containing a high fraction of water (sample 13). In comparison, the differences between the real and the synthetic spectra were much smaller for mixtures without water (sample 7).

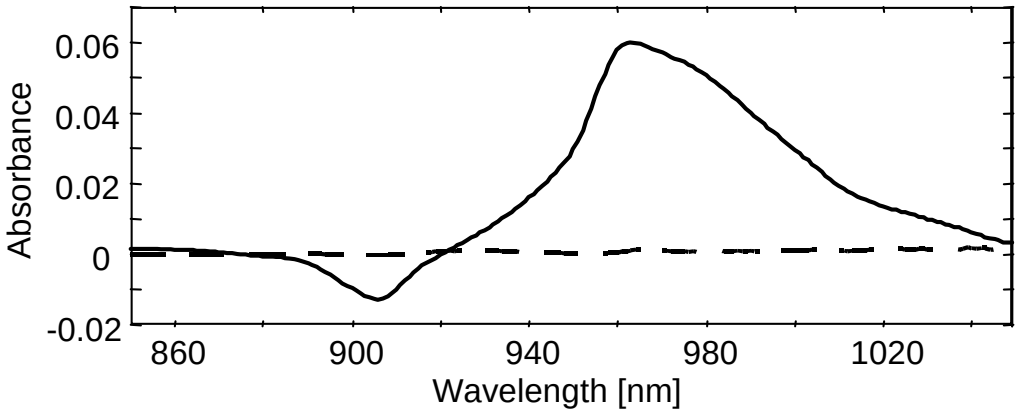


Figure 2-5: Difference between real and synthetic spectra. Solid line sample: 13 ($^1/_6$ ethanol, $^2/_3$ water, $^1/_6$ 2-propanol). Dashed line: sample 7 (½ ethanol / ½ 2-propanol) .

A PLS regression of the spectra on their mole fractions is therefore expected to need more LV's than would be expected by the chemical rank[28].

Local models

Case *a*: Leave one out cross-validation is performed on the training set (see Figure 2-2) for calibration models for each of the three chemical compounds at each of the 5 temperatures.

For all local models the RMSECV does decrease considerably until 4 LV's are included, staying more or less constant for more LV's. The models for water give an about factor three lower RMSECV but show the same behavior. This is due to the fact that water has a higher absorption in the wavelength range studied than the alcohols.

Visual inspection of the loading plots indicates for all models that only the first four loadings show systematic spectral information; higher LV's consist primarily of noise. Therefore, 4 LV's have been used to build the PLS-models for predicting the mole fractions in the test set. Note that in the ideal

case (linearity and additivity) the model would only consist of two LV's since the chemical rank is two (3 components with closure) and the spectra are mean centered [29]. The non-additive behavior of especially water is responsible for the higher number of LV's necessary in practice.

The individual predicted mole fractions for each test sample and temperature did not show any anomalies such as outliers or systematic errors. The results will therefore be summarized by giving the values for the RMSEP and MRE per model only (see Table 2-2). The error for the prediction of water is considerably lower than for the alcohols. On the average a prediction for one of the components at any temperature would be about 3% inaccurate.

Table 2-2: RMSEP ($\cdot 10^{-2}$) and MRE for the different models.

		Temperature [°C] of		ethanol		water		2-propanol	
		Sample	Model	RMSEP	MRE	RMSEP	MRE	RMSEP	MRE
Local	a	30	30	1.77	4.0%	0.92	3.2%	1.24	3.2%
		40	40	1.06	2.1%	0.67	1.3%	0.93	2.4%
		50	50	1.66	4.0%	1.11	2.8%	2.18	7.4%
		60	60	0.98	3.0%	0.43	1.4%	0.83	2.3%
		70	70	1.12	3.4%	0.38	1.3%	1.47	2.5%
			Mean	**1.32**	**3.3%**	**0.70**	**2.0%**	**1.33**	**3.6%**
	b	40	30 & 50	1.81	3.8%	0.51	1.4%	2.74	7.5%
		60	50 & 70	2.77	8.5%	1.13	3.1%	1.92	5.6%
			Mean	**2.29**	**6.1%**	**0.82**	**2.3%**	**2.33**	**6.5%**
Global	a	30	30-70	1.38	4.9%	1.25	3.4%	1.13	3.1%
		40	30-70	1.32	5.0%	0.55	1.9%	1.64	5.1%
		50	30-70	3.77	13.1%	0.79	2.3%	4.08	14.9%
		60	30-70	1.59	5.5%	0.84	3.1%	1.74	4.0%
		70	30-70	1.75	4.9%	0.76	2.1%	1.75	4.6%
			Mean	**1.96**	**6.7%**	**0.84**	**2.6%**	**2.07**	**6.3%**
	b	40	30 & 50	1.17	3.4%	0.95	2.9%	1.03	2.2%
		60	50 & 70	1.24	4.4%	0.93	2.1%	1.30	3.7%
			Mean	**1.21**	**3.9%**	**0.94**	**2.5%**	**1.17**	**3.0%**

The higher signal of water translates to a higher norm of the NAS's and sensitivities for water (Table 2-3). The increase in absorption with the increase in temperature for all three components (Figure 2-4) also gives rise to higher sensitivity at higher temperatures. The sensitivity vectors for all samples (except the samples with mole fraction 1/3 as explained in Performance measures) are very similar, as shown for ethanol in Figure 2-6a.

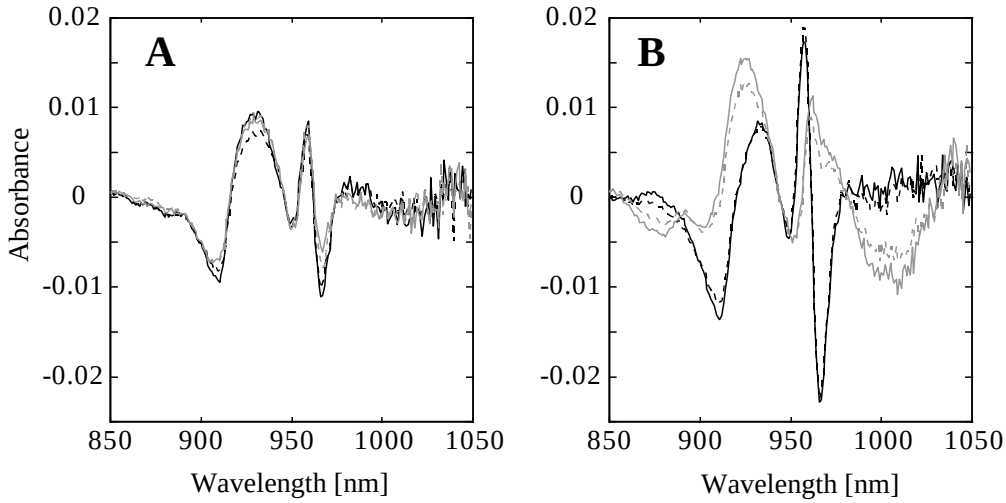


Figure 2-6: Sensitivity vector plots for ethanol prediction of samples ―5, - -6, —14 and - -15 measured at 40°C. *A*: Local model case *a* at 40°C.*B* : Local model case *b*, vectors (model at 30°C).

Table 2-3: Norm of the sensitivities for prediction samples: 5, 6, 14 ,15 for ethanol, 6, 9, 11, 14 for water and 5, 9, 11, 15 for 2-propanol.

		Temperature [°C] **of**		**Sensitivity norm (·10^{-2}) for:**		
		Sample	**Model**	**ethanol**	**water**	**2-prop**
Local	**a**	30°C	30°C	5.60	9.33	6.47
		40°C	40°C	5.40	10.6	6.47
		50°C	50°C	5.77	10.4	6.54
		60°C	60°C	5.81	12.3	7.20
		70°C	70°C	6.56	13.4	8.47
			Mean	**5.83**	**11.2**	**7.03**
	b	40°C	30°C	8.45	12.3	9.61
		40°C	50°C	8.35	11.2	7.62
		60°C	50°C	7.68	13.0	9.55
		60°C	70°C	10.0	14.0	9.64
			Mean	**8.63**	**12.7**	**9.11**
Global	**a**	30°C	30-70°C	3.46	7.32	3.10
		40°C	30-70°C	4.00	7.33	3.37
		50°C	30-70°C	3.77	7.20	3.09
		60°C	30-70°C	3.66	7.54	3.09
		70°C	30-70°C	3.67	7.72	3.35
			Mean	**3.71**	**7.42**	**3.20**
	b	40°C	30 & 50°C	3.97	8.47	3.64
		60°C	50 & 70°C	3.83	7.89	3.25
			Mean	**3.90**	**8.18**	**3.45**

Case *b*: With these models predictions for a test temperature are calculated as the average prediction based on local models built for the two "neighboring" temperatures. The prediction errors found are given in Table 2-2.

The models can obviously not predict with a good accuracy measurements done at a different temperature. Averaging the predicted mole fraction improves the prediction error to approximately half of the prediction error given by the PLS models at the two nearest temperatures. Still, the prediction errors are almost twice as high as in case *a*.

The sensitivities (Table 2-3) are higher than for local models case *a*. This is due to the fact that the NAS does not describe only the analyte but also the temperature difference between the training set and prediction samples. This is shown by comparing the plots in Figure 2-6, revealing the difference between the sensitivities of case *a* and *b*. The same test samples measured at 40°C exhibit very different and irregular sensitivity vectors when predicted by a model at 30°C. The rank annihilation step causes the net analyte signal to describe everything except absorption due to water and 2-propanol at 30°C. The difference between the sensitivities for samples 5,6 and 14,15 shows clearly that the temperature effect, now incorrectly included in the NAS and sensitivities, is dependent on the concentrations.

Global models

Case *a*: The training set for all five temperatures is used to build the model and the mole fractions of the test set at all temperatures are predicted.

For both cross validation strategies the RMSECV steadily decreases with the number of LV's up to seven LV's included in the model when it stops decreasing significantly. The loading plots show that the LV's higher than seven describe mostly noise. Therefore models with 7 latent variables were built form the training sets.

Apparently, the nonlinearity of the temperature effects forces the PLS algorithm to model some systematic information in such high LV's. Roughly, the number of LV's can be rationalized as two LV's necessary to describe the chemical problem, two further to explain the non-additive behavior of water (see local models) and three more LV's for the description of the nonlinearities due to temperature variation.

The prediction errors for the test set at the different temperatures are given in Table 2-2. In absolute terms (RMSEP) the global model performs worse than the local model case a and comparable to case *b*. The high mean relative error compared to the equivalent predictions by the local models is caused by the fact that the model makes a relative high error when predicting lower mole fractions.

The norms of the sensitivity vectors (Table 2-3) are considerably lower than those for the local models. This leads to the conclusion that, due to the variation caused by temperature, the model is forced to use a smaller amount of the spectra for prediction of the analyte.

Case *b*: In this case data at two temperatures (30 and 50°C or 50 and 70°C) are used for building a model and the spectra at the temperature in between (40 or 60°C resp.) are used as prediction set. As it was the case for the local models, the global models case *b* are built with the same number of LV's (7) as in case *a*. Table 2-2 displays the RMSEP and MRE values for the two test sets. When compared to the results of the corresponding local model, the global model predicts more accurate in almost all cases. As a whole, the predictive performance is comparable to that of local models case *a*, being slightly better for the alcohols and slightly worse for water. Considering that the local models are in a way a "best case scenario" it means that the temperature effect on the predictions is reduced to a minimum.

The sensitivity-norms (Table 2-3) are only little higher than for global models case *a*, especially for the test set at 40°C predicted with the spectra at 30 and 50°C. The smaller temperature span and mainly the higher number of calibration samples improves the predictive ability in comparison to case *a*. Still, a considerable part of a spectrum is not used for prediction due to the temperature effects as can be seen from comparing the sensitivity-norms with the local model case *a*.

Conclusions

Global models in which the temperature is modeled as an unknown interferent perform only slightly inferior to local models which are calibrated and used for a specific temperature. Global models, however, have a tendency to become (very) complex. The obtained global models needed seven LV's, three to describe the temperature interference, two for the non-additive behavior of water whilst the chemical system is of rank two. If temperature is treated as an unknown interferent, it is more important to span the variation due to concentration rather than for many temperature levels.

Interpolation between local models, to accommodate temperatures not present in the calibration set, performs poorly.

Further research will aim to describe the temperature effects explicitly; either by preprocessing data before calibration or by inclusion of temperature into a calibration model itself.

References

[1] Blaser, W. W.; Bredeweg, R. A.; Harner, R.S.; LaPack, M.A.; Leugers, A.; Martin, D. P.; Pell, R. J.; Workman, J., Jr.; Wright, L. G. *Anal. Chem.* **1995**, *67*, 47R-70R.

[2] DeThomas, F. A.; Hall, J. W.; Monfre, S.L. *Talanta* **1994**, *41*, 425-431.

[3] Frank, I. E.; Feikema, J.; Constantine, N.; Kowalski, B. R. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 20-24.

[4] Hall, J. W.; McNeil, B.; Rollins, M. J.; Draper, I.; Thompson, B. G.; Macaloney, G.; *Appl. Spectrosc.* **1996**, *50*, 102-108.

[5] Siesler, H.W. *Landbauforschung Völkenrode*, **1989**, 112-118.

[6] Hildrum, K.I.; Isaksson, T.; Næs, T.; Tandberg, A. *Near Infrared Spectroscopy: Bridging the Gap between Data Analysis and NIR Applications*; Ellis Horwood: New York, 1992.

[7] Burns, D. A.; Ciurczak, E.W. *Handbook of Near-Infrared Analysis*, 1st edition; Marcel Dekker Inc.: New York, 1992.

[8] Cho, T.; Kida, I.; Ninomiya, J.; Ikawa, S. *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 103-107.

[9] Czarnecki, M.A.; Czarnecka, M.; Ozaki, Y. *Spectrochim. Acta*, **1994**, *50*, 1521-1528.

[10] Czarnecki, M.A.; Liu, Y.; Ozaki, Y.; Suzuki, M.; Iwahashi, M. *Appl. Spectrosc.* **1993**, *47*, 2162-2168.

[11] Hazen, K. H.; Arnold, M. A.; Small, G. W. *Appl. Spectrosc.* **1994**, *48*, 477-483.

[12] Kamiya, N.; Sekigawa, T.; Ikawa, S. *J. Chem. Soc. Faraday Trans.* **1993**, *89*, 489-493.

[13] Liu, Y.; Czarnecki, M.A.; Ozaki, Y.; Suzuki, M.; Iwahashi, M. *Appl. Spectrosc.* **1993**, *47*, 2169-2171.

[14] de Noord, O.N. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 85-97.

[15] Okuyama, M.; Ikawa, S. *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 3065-3069.

[16] Ozaki, Y.; Liu, Y.; Noda, I. *Appl. Spectrosc.* **1997**, *51*, 526-535.

[17] Bonanno, A.S.; Olinger, J.M.; Griffiths, P.R. *in [6]*.

[18] Libnau, F.O.; Kvalheim, O.M.; Christy, A. A.; Toft, J. *Vib. Spectrosc.* **1994**, *7*, 243-254.

[19] Pegau, W. S.; Zaneveld, J. R. V. *Limnol. Oceanogr.* **1993**, *38*, 188-192.

[20] Finch, J. N.; Lippincott,k E. R. *J. Chem. Phys.*, **1956**, *24*, 908-909.

[21] Finch, J. N.; Lippincott,k E. R. *Phys. Chem.* , **1957**, *61*, 894-902.

[22] Kemeny, G. J. *in [7]*.

[23] Noda, I. ; Liu, Y.; Ozaki, Y.; Czarnecki, M.A. *J. Phys. Chem.* **1995**, *99*, 3068-3073.

[24] Liu, Y. ; Czarnecki, M.A.; Ozaki, Y. *Appl. Spectrosc.* **1994**, *48*, 1095-1101.

[25] Wang, F.C.; Feve, M.; Lam, T. M.; Pascault, J. P. *J. Polym. Sci.: Phys.* **1994**, *32*, 1305-1313.

[26] Eschenauer, U.; Henck, O.; Hühne, M.; Wu, P.; Zegber, I.; Siesler, H. W. *in [6]*.

[27] Lorber A.; Faber, K.; Kowalski, B.R. *Anal. Chem.* **1997**. *69*, 1620-1626.

[28] DiFoggio, R. *Appl. Spectrosc.* **1995**, *49*, 67-75.

[29] Pell, R. J.; Seasholtz, M. B.; Kowalski, B. R. *J. Cemom.* **1992**, *6*, 57-62.

3. LINEAR TECHNIQUES TO CORRECT FOR TEMPERATURE INDUCED SPECTRAL VARIATION IN MULTIVARIATE CALIBRATION.

Abstract

The influence of external physical variation such as temperature fluctuations on NIR spectra and their effect on the predictive power of calibration models such as PLS has been studied. Different methods to correct for the temperature effect by explicitly including the temperature in a calibration model have been tested. The results are compared to the implicit inclusion which takes the temperature into account only through the calibration design. Two data sets are used, one well designed data set measured in the laboratory and one industrial data set consisting of measurements for process samples. For both data sets the explicit inclusion of the temperature in the calibration models did not result in an improvement of the prediction accuracy compared to implicit inclusion.

Based on: Wülfert, F.; Kok, W.Th.; de Noord, O.E.; Smilde, A.K.; *Chemom. Intell. Lab. Syst.* **2000**, 51, 189-200.

Introduction

Near-Infrared (NIR) spectroscopy in combination with multivariate calibration models has an increasing application range in process analysis[1, 2, 3, 4, 5, 6, 7, 8]. The fast spectroscopic methods make in- or on-line analysis attractive in industrial applications. However, as the measurements are not done under well-controlled laboratory circumstances, they will also reflect variations in physical variables such as temperature, pressure and viscosity [9, 10]. These external variations give rise to changes in band shapes by changing the weaker inter- and intramolecular forces. The strongest effects can be observed for bands of functional groups with H-bonding, such as hydroxyl groups, due to the fact that the intermolecular forces have much stronger effects on them[11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29].

The resulting shifting and broadening of bands are non-linear phenomena which complicate the application of linear multivariate models. As shown in a previous study[30], it is still possible to build global multivariate regression models by including the temperature as an interferent in the calibration design (implicit inclusion). The resulting global models can then approximate the non-linear temperature effects in the **X** data matrix by including more latent variables than expected from the chemistry of the problem. As the temperature is neither used explicitly as predicting (**X**) or dependent (**y**) variable the accuracy is not expected be optimal.

In order to achieve a better handling of the temperature influences, explicit inclusion of the temperature into the model is often expected to improve the accuracy. In this study explicit inclusion of the temperature is done along different lines: by direct inclusion in the calibration models, by preprocessing of the spectra and by expression of the spectra on a different basis. Two very different data sets are used: the first (data set *A*) contains spectra of ternary mixtures of ethanol, water and 2-propanol, following a calibration design, with the mole fractions as the variables to be predicted. The second (data set *B*) originates from industrial samples, containing spectra of heavy

oil fractions with the density as predicted variable. The prediction accuracies obtained with the different explicit models are compared with those obtained with the implicit models described previously[30].

Theory

In the following the calibration models with explicit inclusion of temperature variation and the data preprocessing steps that have been used are described and their choice is shortly justified. All models are PLS[31, 32, 33] regression models, which are inverse calibration models of the type:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F}$$

where **Y** represents the *k* dependent variable(s), e.g., concentrations of *k* chemical components, **X** the independent variables, e.g., *n* spectra with *m* wavelengths, **B** the regression coefficients and **F** the residuals.

Method 1, Temperature as X variable: In most cases temperature measurements are readily available at the same time that spectra are measured. Since the temperature is therefore a known quantity it can be appended to the spectra and used as independent variable in order to improve prediction. The **X** block contains then the spectra and the appended temperature while the **y** vector contains either the mole fractions for one component, or the density (data set *A* or *B*, respectively.).

The collinearity between temperature affected regions of the spectrum and the appended temperature variable could lead to regression models that recognize the temperature effects more easily and either are able to correct for it or give less weight to these regions. Since the **X** block does now contain incomparable variables (temperature and absorptions) the data has to be scaled, i.e. all variables are scaled to unit variance (auto scaling) or the variance of the temperature variable is scaled to match the variance of the spectra (block scaling). Through these scalings the temperature can be

given the same weight as either only one wavelength or a complete spectrum.

Method 2, T as Y variable in PLS2: Adding the temperature as a predicted variable results in an **X** block containing only the spectra and an **Y** block containing the temperature and one of the mole fractions (*A*) or the density (*B*).

The simultaneous prediction of the **y** variable and the temperature is seen as a way to enable the model to identify the spectral regions which are temperature dependent. This is in line with inverse calibration, where the underlying variable(s) causing the variation in the spectra are collected in the **Y** block. In this case the temperature is also causing variation in the spectra. Note that the temperature of the unknown sample does not have to be known; it will be predicted from the spectrum. The calibration method used is PLS2, where the suffix '2' indicates that there is more than one variable in the **Y** block. PLS2 uses the fact that there is correlation in the **Y** block. In this special case, such a correlation is not present due to the design of the data. Hence, PLS2 might give poor results in this case. Yet it is worthwhile to examine this approach.

Method 3, T in Y for two step PLS: Instead of simultaneous prediction with the dependent variable, a temperature model can also be built prior to the final calibration model.

In step one, a PLS model is built between all calibration spectra (**X**-block) and the temperature (**y**-block). This model is calculated with one component and describes the covariation between the spectra and the temperature. In step two the **X**-residuals from the first model are used to build the calibration model for predicting **y**. The temperature induced variation is supposed to be removed by the preprocessing model but the success of this model will depend on how independent temperature and concentration manifest themselves in the spectra.

Method 4, Robust variable selection: Another approach to eliminate the temperature effects on the prediction is to select only the **X**-variables (wavelengths) which are insensitive to temperature. Contrary to the previous method, rather than removing the temperature induced variation from whole spectra, a variable selection method tries to exclude the temperature effects by eliminating the variables that carry the temperature variation.

In order to assess which variables are informative for the prediction of the **y** variable but do not reflect temperature induced variation, the uninformative-variable-elimination (UVE) method[34] is adapted. The UVE method uses a comparison between the spectral variables and appended artificial noise in order to estimate whether a spectral variable has predictive power or not. Several calibration models are built (using a jackknife leave-one-out method) and the resulting regression coefficients of the spectral and artificial noise variables are compared with each other. This is done by calculating a reliability coefficient from the mean and standard deviation over the several models for each regression coefficient. Spectral variables that are not considerably more reliable than the artificial noise variables are eliminated.

The UVE method is extended for this application by applying it for both the **y**-variable and temperature prediction. This allows the building of the final calibration model from only those wavelengths that are considered informative for **y** but not for the temperature.

Method 5, Basis projection: By expressing spectra on a different basis, a separation of the temperature effects from the concentration information is sought. Ideally a spectrum projected on the new basis would result in an expression of temperature and concentration effects on different coefficients. Such a new basis can either be formed by the data itself or by mathematical functions.

A data-driven basis can be formed from the spectra at different temperatures that represent the extreme points of a calibration design (for instance pure component spectra). Ideally a spectrum measured at a certain temperature would result in high concentration related coefficients on the extreme spectra which are measured at the same temperature.

For a mathematical basis, Wavelet Packets (WP) represent a possibility as they are localized in both frequency and location (compact support). WP Transform is therefore ideal to describe and filter local effects and has found chemical applications in the denoising and compression of signals[35, 36, 37, 38].

The need for well designed data limits the data-driven approach to the first data set (*A*). The Wavelet Packet transform will only be applied in a tentative study on a small simulated data set due to the intensive calculations needed for selecting different WP-bases. A more general analysis of the possibilities of transformations between orthogonal bases is given in the Appendix.

Experimental section

Apparatus and experimental details

Data set *A*: Using an analytical balance, mixtures of ethanol, water and 2-propanol have been prepared from p. a. grade alcohols and subboiled water. Airtight sample flasks and a closed 1 cm quartz cell have been used in order to prevent evaporation and consequent concentration changes during storage and measurements. A HP 8453 Spectrophotometer linked to a HP Vectra XM2 PC (Hewlett Packard, Palo Alto, CA, USA) was used to take the spectra in the wavelength range 580 to 1091 nm (1 nm resolution, 20 s integration time). The wavelength range from 850 to 1049 was used for building the calibration models. The sample temperature during the measurements has been controlled using a thermostatically controlled cell holder and cell stirring module with an accuracy of 0.2°C.

Data set *B*: Spectra of heavy oil products were measured in a temperature controlled flow cell on a Bomem MB 160 FTNIR spectrometer in the spectral range between 6206 and 3971 cm^{-1}. Baseline correction was applied by subtracting the average absorbance in the range 4810-4800 cm^{-1} and the last 400 variables (4740-3971 cm^{-1}) were used for the calibration models. The quality variable to be predicted is the density at 15°C, which was measured according to ASTM D4052. This method has a long term standard deviation of 0.0015 g/ml for the current type of product.

Small simulated data set: Two vectors of 16 data points were made using the Gaussian function with a width (standard deviation) of 1 data point . The shift between the two Gaussians peaks has been set at 1 data point (see Figure 3-8 in Appendix).

Simulations and data processing were done on a Pentium-class computer using Matlab (ver. 4.2 and 5; The Mathworks Inc.) and the PLS toolbox (ver. 1.4, Eigenvector Research).

Calibration design

Data set *A*: Samples from a mixture design which include all possible secondary and ternary mixtures with mole fraction levels of 1/6, 1/3, 1/2 and 2/3 (see Figure 3-1) have been measured at temperatures of 30, 40, 50, 60 and 70°C (± 0.2°C). Models are built using a training set consisting of the samples that are on the edge of the experimental design (samples 1, 2, 3, 4, 7, 8, 12, 13, 16, 17, 18, 19) and the sample in the "center" (sample 10) measured at all temperatures. The other samples are used as independent test set.

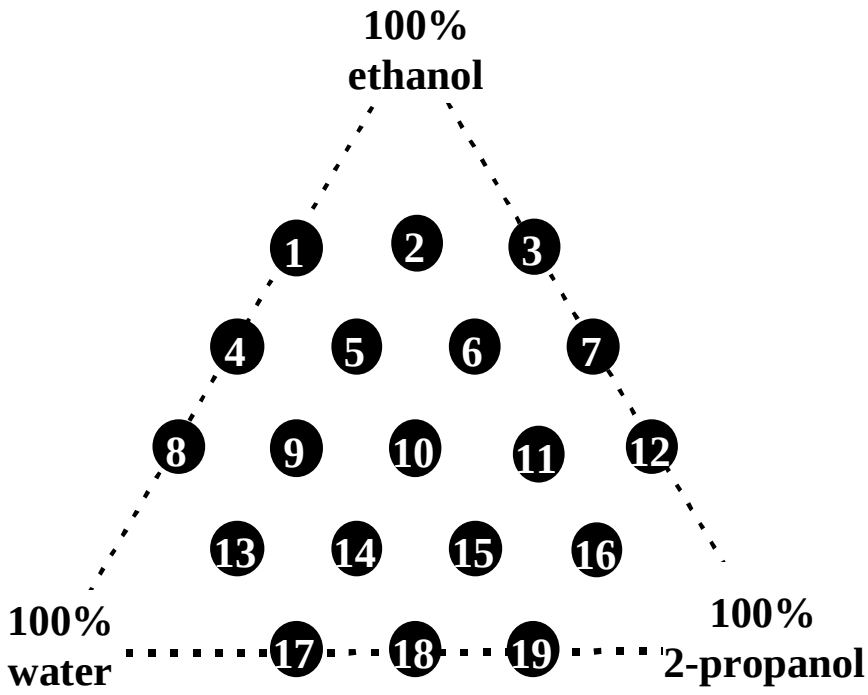


Figure 3-1: Graphical representation of mixture design for data set A.

Data set *B*: The data set consists of 232 spectra taken from 64 heavy oil product samples measured at 95, 100 and 105°C. On the 168 spectra of the 56 samples that are neither duplicate measurements nor quality control

samples a PCA model is applied and the score values of the first two principal components are plotted against each other (see Figure 3-2). From this score plot 14 samples are selected evenly from all regions to be added to all duplicate and quality control sample measurements to form the test set. The training set consists therefore of 42 samples measured at all temperatures (126 spectra, 42 densities). The test set consists of 14 unique samples, 7 duplicates and the quality control sample (total of 106 spectra, 22 densities).

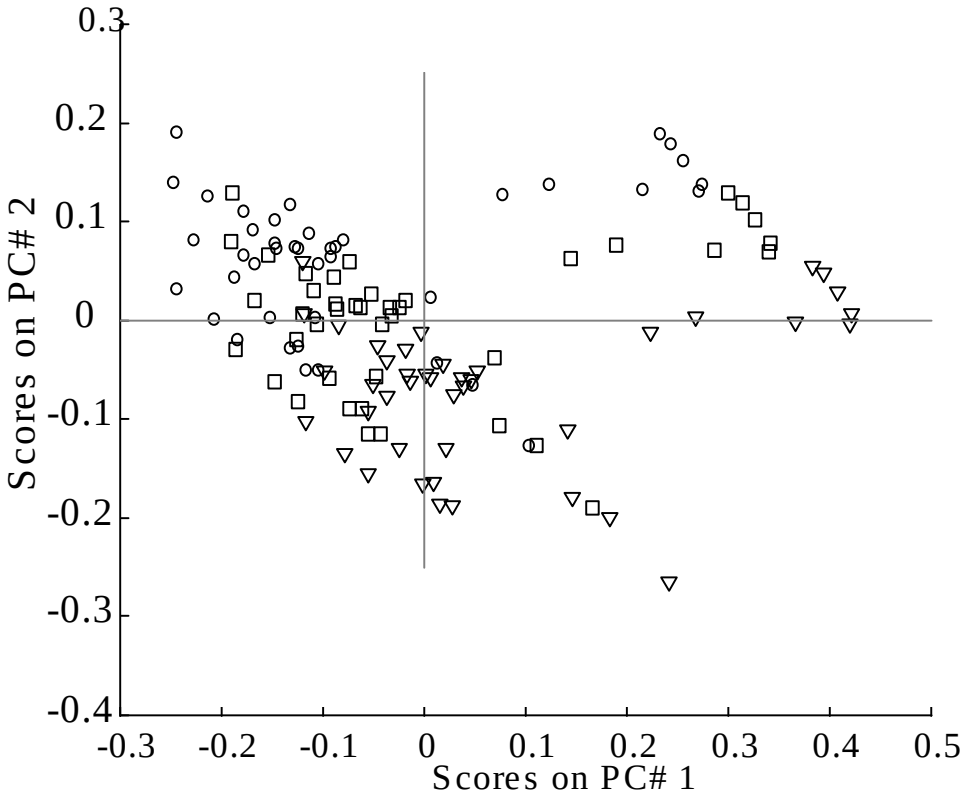


Figure 3-2: PCA on data set B, the temperature effect can clearly be seen; circles: measurements at 105°C, squares: measurements at 100°C and triangles: measurements at 95°C.

Results and Discussion

Figure 3-3 shows some exemplary spectra for data set *A* and the temperature effect for one of the mixtures.

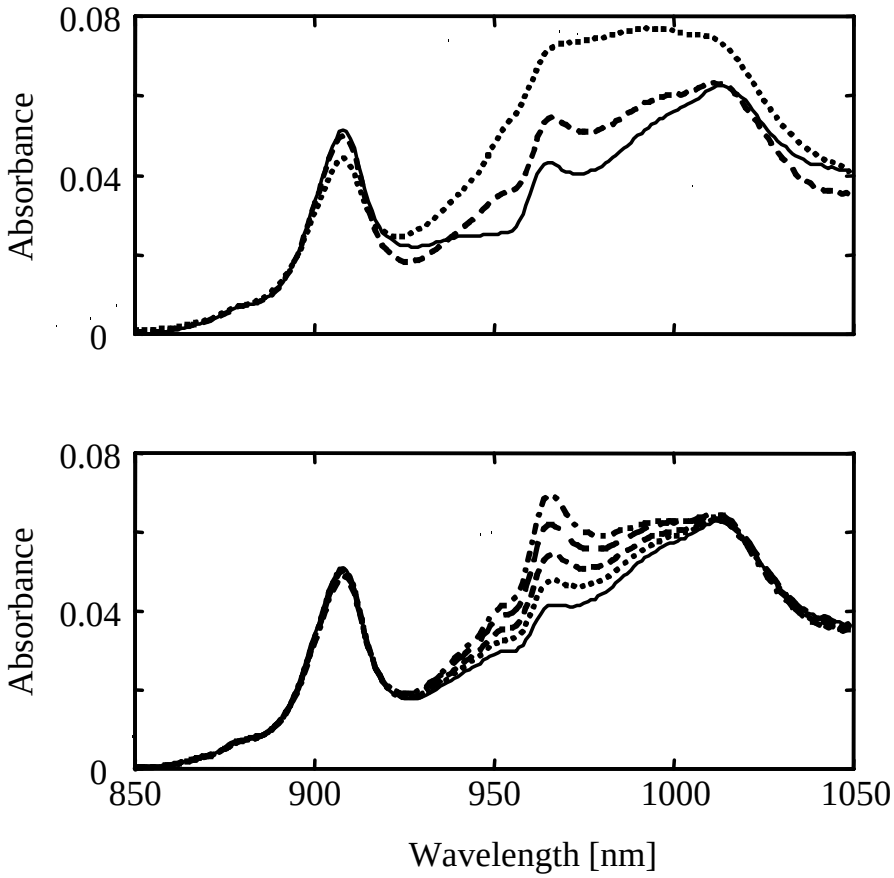


Figure 3-3: Spectra of different ternary mixtures taken at 50°C (top), temperature effect on one mixture, spectra taken at 30, 40, 50, 60 and 70°C.

The figure shows that a distinction between spectral regions containing composition information and regions with temperature interference can not be made intuitively.

The same conclusion can be drawn for Figure 3-4, showing a similar set of spectra obtained with heavy oil samples (data set *B*).

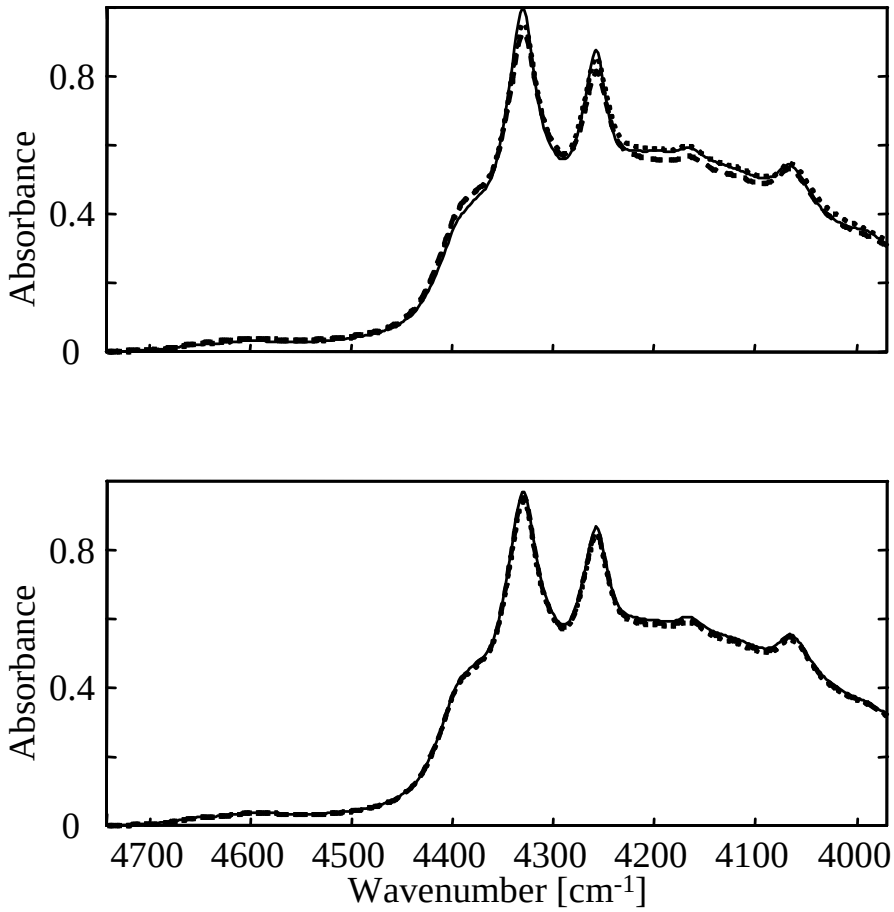


Figure 3-4: Spectra of different heavy oil products taken at 100°C (top), temperature effect on one sample, spectra taken at 95, 100,105°C.

In order to assess model complexity, for all methods leave-one-out cross-validation has been applied on the training set, leaving the data of one sample at all temperatures out for each cross-validation step.

The model complexity and prediction errors (RMSEP's) are given in Table 3-1 and Table 3-2 for data sets *A* and *B* respectively. In the following only deviating behavior and differences between the models will be discussed.

Table 3-1: Results for data set A.

	No. of lv's	RMSEP Ethanol	RMSEP water	RMSEP 2-prop	RMSEP Mean
Reference	7	0.0196	0.0084	0.0207	0.0162
Reference auto scaled	7	0.0276	0.0094	0.0279	0.0216
Method 1.	7	0.0201	0.0117	0.0232	0.0183
Method 2.	7	0.0189	0.0116	0.0202	0.0169
Method 3.	1 ; 6*	0.0581	0.1323	0.0816	0.0907
Method 4.	4	0.0231	0.0127	0.0265	0.0208
Method 5.	7	0.0224	0.0102	0.0294	0.0207

* 1st value (1) for preprocessing model, 2nd value (6) for calibration model.

Table 3-2: Results for data set B.

	No. of lv's	RMSEP Density
Reference	6	0.00324
Reference auto scaled	6	0.00356
Method 1.	6	0.00363
Method 2.	6	0.00410
Method 3.	1 ; 5*	0.01303
Method 4.	6	0.00414

* 1st value (1) for preprocessing model, 2nd value (5) for calibration model.

Reference method, Global model:. In a previous article[30] it was shown that the implicit inclusion of the temperature through the calibration design results in reasonable predictions for data set *A* but at the cost of an increased model complexity by 3 latent variables when compared to models without temperature effects. The results of the global model will be used as a reference and are visualized in the parity plot in Figure 3-5.

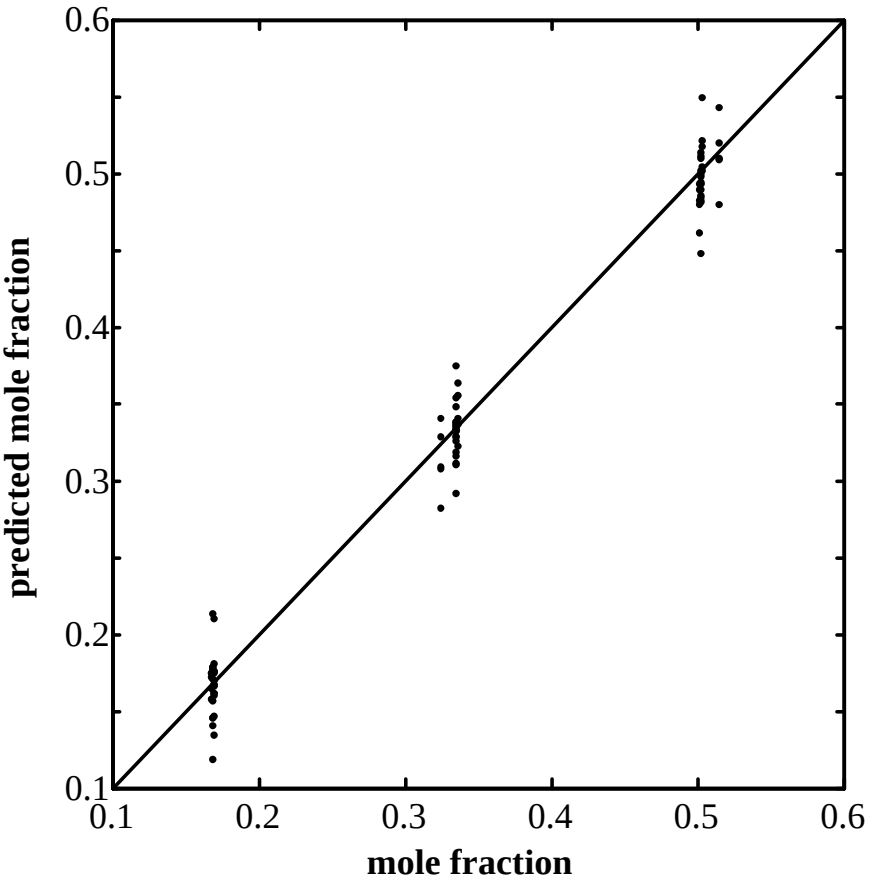


Figure 3-5: Predicted versus real mole fraction for all three components for the reference method applied on data set A.

Data set *B*: This data set has not been used with the reference method earlier and the results will therefore be given in short: Cross-validation

results in a model with 6 latent variables (only one higher than for a model on data at one temperature) as the RMSECV decreases steeply until the 6th latent variable where it stabilizes, only decreasing very slightly for more complex models. The good agreement between prediction and off-line measurement is shown in the parity plot Figure 3-6.

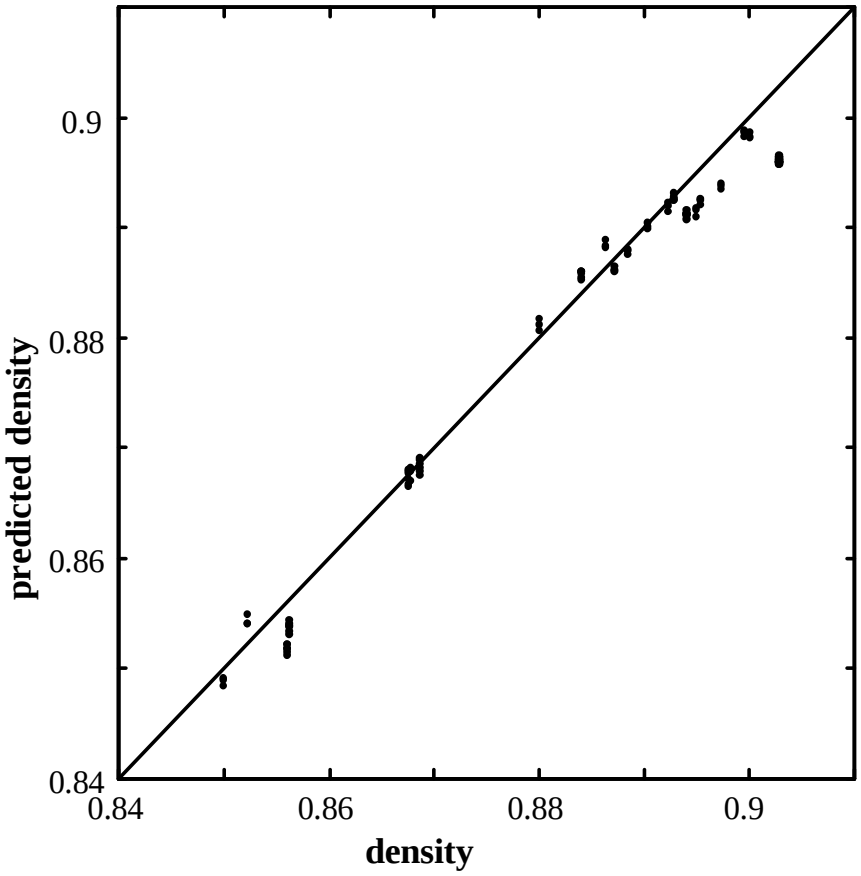


Figure 3-6: Predicted versus real density for the reference method applied on data set B.

Method 1, T as X variable: For both data sets, the cross-validation shows no significantly different behavior for auto scaled or block scaled data and therefore only the results for auto scaling are given. The similar behavior for

the different scaling methods can be sought in the fact that the PLS algorithm compensates with different values in the weight vectors. The results for both data sets indicate that no significant gain can be achieved by including the temperature in the **X** block. Whether the lack of improvement can be attributed to the necessary scaling (see also the results for an auto scaled reference method) or whether the PLS model was unable to identify the temperature affected regions could not be determined.

Method 2, T as Y variable in PLS2: For this method the use of a **Y** matrix instead of a vector leads to the necessity to auto scale the **Y**-block and use of the iterative NIPALS PLS2 algorithm. Convergence of the NIPALS algorithm was slow for the prediction of 2-propanol for data set *A* and even slower for data set *B*, needing more than 100 iterations.

The difference between the results for data sets *A* and *B* (reasonable results for *A* and the much worse convergence problems for *B*) can be explained by their different structure: data set *A* contains 5 temperature levels while data set *B* contains only 3. While it was possible to predict the temperature simultaneously with the **y**-variable (an error of 2°C on a range of 40°C for data set *A* and an error of 1°C on a range of 10°C for data set *B*), no gain in the accuracy in interest could be observed. This is most probably due to the fact that PLS2 needs correlated **Y**-variables for an improved performance over “normal” PLS1 models.

Method 3, T as Y for two step PLS: Although for both models cross-validation resulted in final calibration models with one latent variable less than the reference method, these models are not simpler as one additional latent variable has been used for the first temperature correcting PLS model.

For both data sets the prediction errors are much larger than for the reference method. The preprocessing correction model seems to be successful in modeling the temperature by using much of the variance

present in the spectra, even though only one latent variable was used. Since 95% of the variance of **X** for data set *A* and 60% of the variance for data set *B* have been used up, the residual matrices used for the prediction models do not contain enough variance for modeling the mole fractions or density. This can be explained by the fact that the concentration and temperature induced effects on the spectra are not independent of each other; both parameters have influence on hydrogen bonding and other intermolecular forces. Consequently, they manifest themselves on the spectra in a very similar way and a separation of these effects with PLS based models is thus not achievable.

Method 4, Robust variable selection: After elimination of the uninformative and temperature sensitive spectral variables a considerable data reduction is achieved: Out of the 200 spectral variables for data set *A*, only 32, 44 and 45 variables were retained for water, ethanol and 2-propanol prediction respectively. For data set *B*, 243 out of 400 wavelength were considered as robust by this method.

For both data sets the resulting calibration models were not more accurate than the reference method, but for data set *A* the calibration models were more parsimonious. The better data reduction and conciseness for data set *A* is a consequence of the larger temperature range and the higher number of temperature levels present in that data set, enabling the UVE method to identify the variables to be left out.

Interpretation of the variable selection is possible, especially for data set A where also pure components have been measured. In Figure 3-7 the pure spectra of water measured at two temperatures are shown to illustrate this. The selection of informative variables for the water content in mixtures (a), the temperature (b) and the final selection of robust variables (c) are shown in the plots.

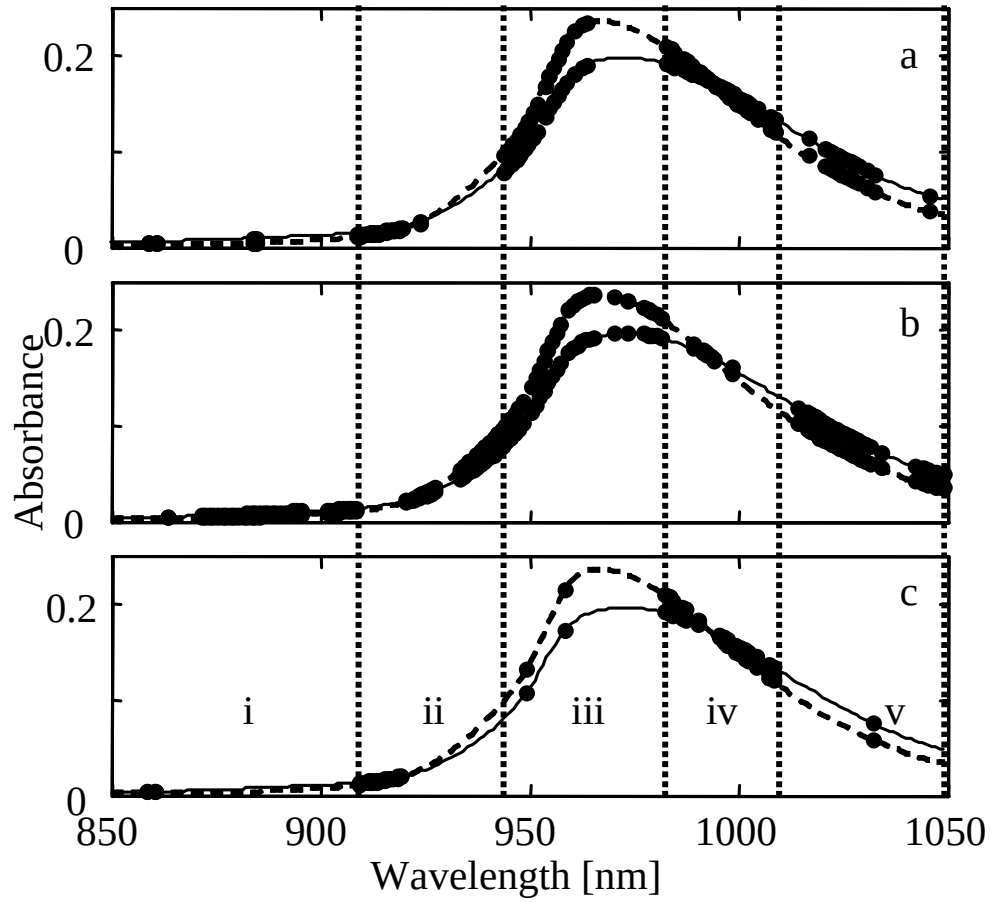


Figure 3-7: Pure spectra of water at 30°C (dashed line) and 70°C (solid line). Circles represent the wavelengths selected by UVE for prediction of water content (*a*), temperature (*b*) and variables only selected for water but not for temperature prediction (*c*).

The spectra have been divided in 5 regions for easier explanation. In the first region (i) the absorbances are low, the UVE method for water rejects almost all wavelengths but selects most for temperature since some temperature induced variation is present. In the second region (ii), there are higher water absorbances present but in the upper half of this region the temperature effects and the CH-band of the alcohols interfere. Only some wavelengths in the lower half are therefore selected. The highest

absorbances but also the largest temperature effects are found in the third region (iii) where also the free OH-band of the alcohols adds to the interference. Almost no wavelengths are considered informative for water and uninformative for temperature. In the next region (iv) most of the wavelengths can be selected: water absorption is high and temperature effects as well as alcohol interference are relatively low. The last region (v) again results in almost no selected wavelengths due to the interference of the temperature and alcohols.

Method 5, Base projection: For a data-driven basis the data set *A* has been used. As new basis vectors the ternary mixtures at the border of the calibration design have been used (samples #2, 13 and 16). This leads to 15 data-driven base vectors at 5 different temperatures on which the data is projected. The projection did not result in a few high coefficients belonging to base spectra measured at the same temperature as the projected spectrum. Moreover no systematic distribution of the signal over the new coefficients could be discovered. Prediction errors are therefore expectedly higher than those obtained with global models.

For the tentative study on the small simulated data set different wavelet families have been used. For each family 677 complete bases are possible to describe the 16-points data space and a complete search over all bases was performed. The basis which distinguished the best between temperature effects (represented by a shift of one wavelength) and concentration effects was selected. For details see the Appendix.

While a solution had been found, applying this solution on Gaussians with a 2 points shift instead of one point resulted in totally different signals. Apparently, the solution is not applicable to other shifts indicating that no general solution to filter out shifts can be formulated by means of WPT. This is in accordance with the findings described in the Appendix, which prove that transforms from one orthogonal basis to another (like WPT) cannot isolate non-linear effects on a few dedicated coefficients.

Furthermore an extensive basis search on real data would gain unmanageable proportions (for data with 256 points: 1.9×10^{45} possible bases per family). Because of the unsatisfactory results of the small simulation, the computational power needed and more principally the theoretical considerations explained in the Appendix, a further study on spectroscopic data was not considered.

Conclusions

Summarizing, it can be noted that none of the different methods for explicit inclusion of the temperature into the calibration models leads to an improvement when compared to the more basic idea of implicit inclusion. For none of the considered methods the predictive ability improved (lower RMSEP) and in general the models also did not become simpler (i.e., lower number of latent variables). The consistency of these results for two completely different data sets (simple mixtures of known composition and complex oil fractions) indicates that non-linearities such as the temperature effects cannot be corrected for nor modeled further with linear techniques than already done by the implicit inclusion through a good calibration design.

Furthermore, it can be concluded from the Appendix that non-linear effects (such as the temperature induced band shifts and broadening) cannot be filtered out or resolved by an orthogonal basis transformation. The effect can only be linearized to different degrees of efficiency by different linear transformations. A full description and inclusion of non-linear effects into a calibration model is therefore only possible by using non-linear transformations. Further research must consequently be focused on non-linear approaches.

Appendix

As described in the theory chapter, the idea is that application of WPT[35, 36, 38] on spectra with temperature induced variation should concentrate the temperature effects on certain WP coefficients. A tentative study on a small artificial data set (see Figure 3-8) has been performed to assess the possibilities of such a transform.

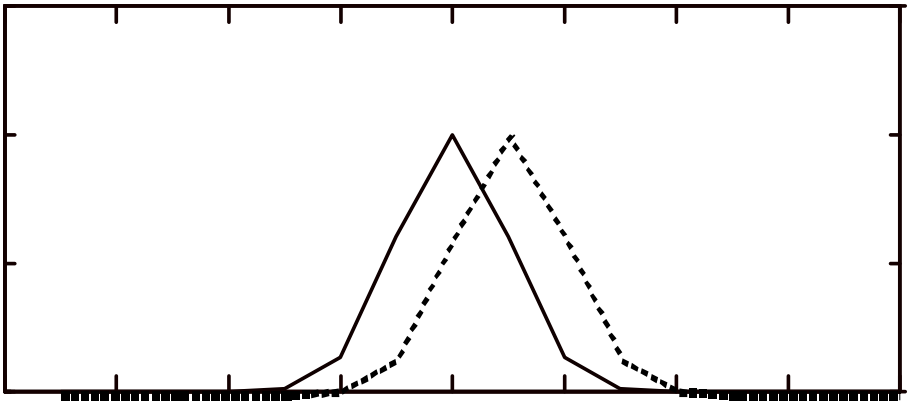


Figure 3-8: Two data vectors with shift used for study on WPT.

Different wavelet basis functions (families) have been used (Haar, Beylkin, Coiflet 1-5, Daubechies 4-20 even, and Symmlet 4-10). For each family, 667 different complete basis sets are possible to describe the data space with 16 variables, which can be calculated recursively by:

$$N_j = N_{j-1}^2 + 1$$

resulting in 2, 5, 26, 677, basis sets for signals with length of 2, 4, 8, 16, data points.

The selection of one basis is often described as a path in consecutive filtering operations with low and high pass filters resulting in the detail and

approximation of the signal. The path, where only the approximation is filtered again in the following step, is called the Wavelet Transform and is therefore only one of the possible Wavelet Packet Transforms (see Figure 3-9).

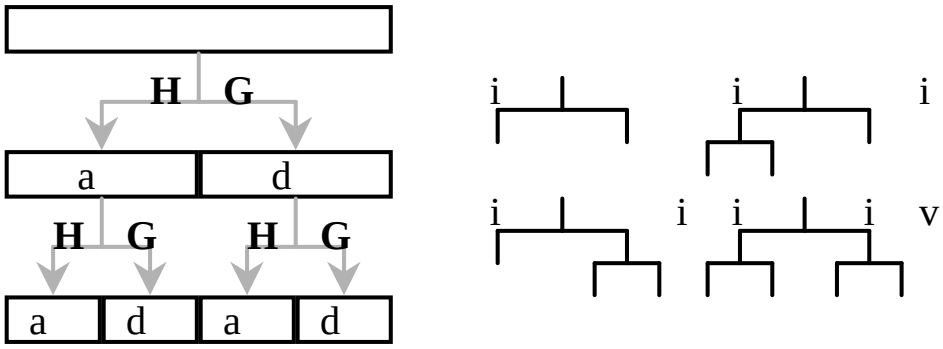


Figure 3-9: All possible filter operations *H* and *G* resulting in approximation *a* and detail *d* (left). Representation of the 4 paths (right), selecting linearly independent and complete bases; the fifth path, the non-transform is obviously not represented.

All bases are orthogonal (orthonormal basis vectors) like the Dirac basis where the signal is expressed originally in. Therefore, the transformation consists of the multiplication of the signal vector with an orthogonal matrix composed of the chosen wavelet packet functions.

The following ranking criteria were applied to ensure that transforms would score high when effectively removing the shift without too much loss of signal and shape:

Select a complete basis.

Project the signal and the shifted version on this basis.

Compare coefficients of both signals on the wavelet basis.

Count the number of basis functions for which:

both signals give high coefficients (>1% of the total power of the signal) and

both signals give equal coefficients (difference < 1% of the value of the coefficients).

Zero the coefficients for all other basis functions.

Transform signals back and calculate the power (norm) of the reconstructed signals.

The extensive search and ranking resulted in a number of good transforms, the best solution being a basis set of the Haar family which removed effectively the difference but retained most of the signal power (96%) and peak shape (see Figure 3-10). However, when applying this solution on vectors with a 2 points shift, the resulting signals were totally different of each other. Apparently, the solution found for a 1 point shift is not applicable to other shifts. In the following, this result will be discussed in two manners, first a more intuitive, geometric approach and later a more formalistic mathematical explanation.

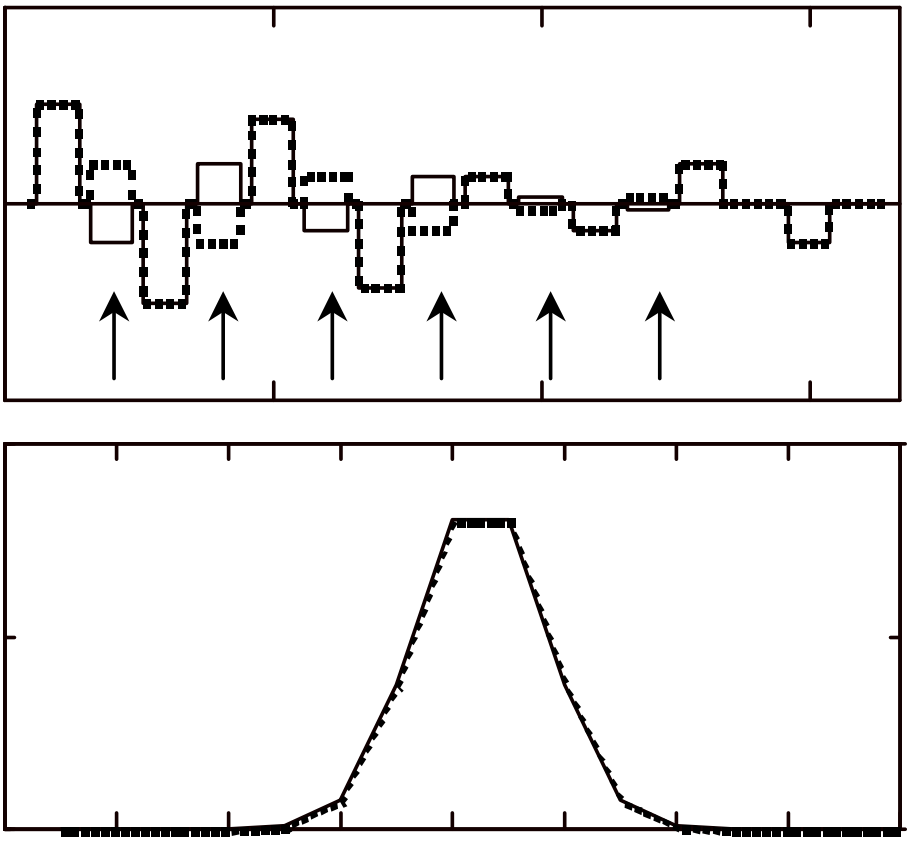


Figure 3-10: WP coefficients for best basis (top), coefficients marked with arrows were zeroed before the back transform, resulting in almost identical data vectors (bottom).

For comprehensibility a signal with 3 data points is taken as an example. A linear effect like a pure intensity change will manifest itself in the three dimensional data space as a straight line. It is possible to rotate (perform a transform with an orthogonal matrix) the basis vectors in such a way that the line lies on one of the rotated basis vectors. On the new rotated basis only one dimension is needed to describe the intensity change.

On the contrary, a shape changing effect like a shift does not manifest itself in the data space as a line but rather as a more complex shape, e.g. a spiral

(in the case of a signal with more than 3 data points this is still possible to see by projecting into fewer dimensions with e.g. PCA[30]). Intuitively it can be understood that by rotating the basis vectors whilst preserving their orthogonality it will not be possible to find a position where the spiral becomes a sinus or even a straight line. Even when dropping the orthogonality demand, new basis vectors will not be able to bring the spiral into a two or one dimensional form, the shift will therefore also influence all three coefficients on the new basis. The 1-point shift solution found in the small WPT study is nothing more than the solution that finds the straight line between 2 points on the spiral. A third situation (2 point shift) does not lie on this line.

For the more formalistic approach, consider two signals $A(x)$ and $B(x)$ of four data points (where x stands for the position in the vector so x=1,2,3,4) shifted by a positive integer Δ. In general a signal is expressed on the Dirac basis with the orthonormal basis vectors $\delta_i(x)$ and coefficients a_i and b_i. The basis vectors $\delta_i(x)$ are the Kronecker delta functions which are one for x=i and zero for $x \neq$i:

$$A(x) = a_1\delta_1(x) + a_2\delta_2(x) + a_3\delta_3(x) + a_4\delta_4(x) \qquad [1]$$
$$B(x) = b_1\delta_1(x) + b_2\delta_2(x) + b_3\delta_3(x) + b_4\delta_4(x)$$
$$B(x) = A(x - \Delta)$$

The goal of the WP transform is to concentrate the shift on a limited number of basis functions, so that the signals on the new WP basis $f_i(x)$ would have the following form:

$$A(x) = c_1 f_1(x) + c_2 f_2(x) + c_3 f_3(x) + c_4 f_4(x) \qquad [2]$$
$$B(x) = c_1 f_1(x) + c_2 f_2(x) + d_3 f_3(x) + d_4 f_4(x)$$

Where the shift is concentrated on the basis functions f_3 and f_4 (with different coefficients) and the common features of the two signals are expressed on the basis functions f_1 and f_2 (having equal coefficients c_1 and c_2). Furthermore the sought solution is needed to be general, which means that for all possible coefficients a_i and b_i some coefficients on the new basis should always be identical (e.g. c_1 and c_2).

Since both bases (Dirac and Wavelet Packets) are orthonormal the following holds:

$$\int_{-\infty}^{\infty} f_n f_m dx = 0 \quad ; \qquad \int_{-\infty}^{\infty} \delta_n \delta_m dx = 0 \quad \text{for} \qquad n \neq m \qquad [3]$$

$$\int_{-\infty}^{\infty} f_n f_m dx = 1 \quad ; \qquad \int_{-\infty}^{\infty} \delta_n f_m \delta x = 1 \quad \text{for} \qquad n = m$$

The following integral:

$$\int_{-\infty}^{\infty} B(x) f_1(x) dx \qquad [4]$$

$$= c_1 \underbrace{\int_{-\infty}^{\infty} f_1(x) f_1(x) dx}_{1} + c_2 \underbrace{\int_{-\infty}^{\infty} f_2(x) f_1(x) dx}_{0}$$

$$+ d_3 \underbrace{\int_{-\infty}^{\infty} f_3(x) f_1(x) dx}_{0} + d_4 \underbrace{\int_{-\infty}^{\infty} f_4(x) f_1(x) dx}_{0}$$

$$= c_1$$

can also be rewritten since $B(x)=A(x-\Delta)$:

$$\int_{-\infty}^{\infty} B(x) f_1(x) dx = c_1 = \int_{-\infty}^{\infty} A(x-\Delta) f_1(x) dx = \qquad [5].$$

$$c_1 \int_{-\infty}^{\infty} f_1(x-\Delta) f_1(x) dx + c_2 \int_{-\infty}^{\infty} f_2(x-\Delta) f_1(x) dx +$$

$$c_3 \int_{-\infty}^{\infty} f_3(x-\Delta) f_1(x) dx + c_4 \int_{-\infty}^{\infty} f_4(x-\Delta) f_1(x) dx$$

This can only be equal to c_1 for all $c_i \in$ IR when:

$$\int_{-\infty}^{\infty} f_1(x-\Delta)\, f_1(x)dx = 1 \; ; \; \int_{-\infty}^{\infty} f_2(x-\Delta)\, f_1(x)dx = 0 \qquad [6]$$

$$\int_{-\infty}^{\infty} f_3(x-\Delta)\, f_1(x)dx = 0 \; ; \; \int_{-\infty}^{\infty} f_4(x-\Delta)\, f_1(x)dx = 0$$

Combining equations 3 and 6 results in the following equality:

$$\int_{-\infty}^{\infty} f_1(x)\, f_1(x)dx = \int_{-\infty}^{\infty} f_1(x-\Delta)\, f_1(x)dx = 1 \qquad [7]$$

which cannot be true.

Therefore, it can be concluded that it is not possible to find a transform between orthonormal bases (WPT, FT) which gives a general solution for concentrating a shift on a few coefficients.

References

[1] Blaser, W. W.; Bredeweg, R. A.; Harner, R.S.; LaPack, M.A.; Leugers, A.; Martin, D. P.; Pell, R. J.; Workman, J., Jr.; Wright, L. G., Process analytical chemistry, *Anal. Chem.* **1995**, *67*, 47R-70R.

[2] DeThomas, F. A.; Hall, J. W.; Monfre, S.L., Real-time monitoring of polyurethan production using near-infra-red spectroscopy, *Talanta* **1994**, *41*, 425-431.

[3] Frank, I. E.; Feikema, J.; Constantine, N.; Kowalski, B. R., Prediction of Profuct Quality from Spectral Data Using the Partial Least-Squares Method, *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 20-24.

[4] Hall, J. W.; McNeil, B.; Rollins, M. J.; Draper, I.; Thompson, B. G.; Macaloney, Near-infra-red spectroscopic determination of acetate, ammonium, biomass, and glycerol in an industrial Escherichia coli fermentation , *Appl. Spectrosc.* **1996**, *50*, 102-108.

[5] Siesler, H.W., Near-Infrared Spectroscopy in the Industry: Acceptance, New Developments and New Results, *Landbauforschung Völkenrode*, **1989**, *107*, 112-118.

[6] Cooper, J.B., Wise, K.L., Welch, W.T., Sumner, M.B., Wilt, B.K., Bledsoe, R.R., Comparison of near-IR, Raman, and mid-IR spectroscopies for the determination of BTEX in petroleum fuels, *Appl. Spectrosc.* 1997, *51*, 1613-1620.

[7] Hildrum, K.I.; Isaksson, T.; Næs, T.; Tandberg, A. *Near Infrared Spectroscopy: Bridging the Gap between Data Analysis and NIR Applications*; Ellis Horwood: New York, 1992.

[8] Burns, D. A.; Ciurczak, E.W. *Handbook of Near-Infrared Analysis*, 1st edition; Marcel Dekker Inc.: New York, 1992.

[9] Yalvac, E.D.; Seasholtz, M.B.; Beach, M.A.; Crouch, S.R., Real-time analysis of light alkenes at elevated temperatures and pressures by fibre-optic near-infra-red spectroscopy, *Appl. Spectrosc.* 1997, *51*, 1565-1572

[10] DeBraekeleer, K.; Cuesta Sánchez, F.; Hailey, P.A.; Sharp, D.C.A.; Pettman, A.J.; Massart, D.L., Influence and correction of temperature perturbations on NIR spectra during the monitoring of a polymorph conversion process prior to self-modelling mixture analysis, *J. Pharmac. Biomed. Analysis* **1998**, *17*, 141-152.

[11] Cho, T.; Kida, I.; Ninomiya, J.; Ikawa, S. , Intramolecular Hydrogen Bond and Molecular Conformation. Part 2, *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 103-107.

[12] Czarnecki, M.A.; Czarnecka, M.; Ozaki, Y.; Iwahashi, M., Are the integrated absorption coefficients temperature dependent? FT-NIR study of the first overtone of the OH stretching mode of octanoic acid, *Spectrochim. Acta*, **1994**, *50A*, 1521-1528.

[13] Czarnecki, M.A.; Liu, Y.; Ozaki, Y.; Suzuki, M.; Iwahashi, M. , Potential of Fourier Transform Near-Infrared Spectroscopy in Studies of the Dissociation of Fatty Acids in the Liquid Phase, *Appl. Spectrosc.* **1993**, *47*, 2162-2168.

[14] Hazen, K. H.; Arnold, M. A.; Small, G. W. , Measurement of glucose in water with first-overtone near-infra-red spectra, *Appl. Spectrosc.* **1994**, *48*, 477-483.

[15] Kamiya, N.; Sekigawa, T.; Ikawa, S., Intramolecular Hydrogen Bond and Molecular Conformation. Part 1, *J. Chem. Soc. Faraday Trans.* **1993**, *89*, 489-493.

[16] Liu, Y.; Czarnecki, M.A.; Ozaki, Y.; Suzuki, M.; Iwahashi, M. *Appl. Spectrosc.* **1993**, *47*, 2169-2171.

[17] de Noord, O.E., Multivariate calibration standardization, *Chemom. Intell. Lab. Syst.* **1994**, *25*, 85-97.

[18] Okuyama, M.; Ikawa, S., Intramolecular Hydrogen Bond and Molecular Conformation. Part 3, *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 3065-3069.

[19] Ozaki, Y.; Liu, Y.; Noda, I., Two-Dimensional Infrared and Near-Infrared Correlation Spectroscopy: Applications to Studies of Temperature-Dependent Spectral Variations of Self-Associated Molecules, *Appl. Spectrosc.* **1997**, *51*, 526-535.

[20] Bonanno, A.S.; Olinger, J.M.; Griffiths, P.R. *in [7]*.

[21] Libnau, F.O.; Kvalheim, O.M.; Christy, A. A.; Toft, J., Spectra of water in the near- and mid-infrared region, *Vib. Spectrosc.* **1994**, *7*, 243-254.

[22] Pegau, W. S.; Zaneveld, J. R. V., Temperature-dependent absorption of water in the red and near-infrared portions of the spectrum, *Limnol. Oceanogr.* **1993**, *38*, 188-192.

[23] Finch, J. N.; Lippincott, E. R., Hydrogen Bond Systems: Temperature Dependence of OH Frequency Shifts and OH Band Intensities, *J. Chem. Phys.* **1956**, *24*, 908-909.

[24] Finch, J. N.; Lippincott, E. R., Hydrogen Bond Systems - Temperature Dependence of OH Frequency Shifts and OH Band Intensities, *Phys. Chem.* **1957**, *61*, 894-902.

[25] Kemeny, G. J. *in [7]*.

[26] Noda, I.; Liu, Y.; Ozaki, Y.; Czarnecki, M.A. *J. Phys. Chem.* **1995**, *99*, 3068-3073.

[27] Liu, Y.; Czarnecki, M.A.; Ozaki, Y., Fourier Transform Near-Infrared Spectra of *N*-methylacetamide: Dissociation and Thermodynamic Properties in Pure Liquid Form and in CCl_4 Solutions, *Appl. Spectrosc.* **1994**, *48*, 1095-1101.

[28] Wang, F.C.; Feve, M.; Lam, T. M.; Pascault, J. P., FTIR Analysis of Hydrogen Bonding in Amorphous Linear Polyurethanes. I. Influence of Temperature, *J. Polym. Sci.: Phys.* **1994**, *32*, 1305-1313.

[29] Eschenauer, U.; Henck, O.; Hühne, M.; Wu, P.; Zegber, I.; Siesler, H. W. *in [**6**]*.

[30] Wülfert F.; Kok, W.Th.; Smilde, A.K., Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models, *Anal. Chem.* **1998**, *70*, 1761-1767.

[31] Geladi, P.; Kowalski, B.R., Partial Least-Squares Regression: A Tutorial, *Anal. Chim. Acta* **1986**, *185*, 1-17.

[32] Höskuldsson, A., PLS Regression Methods, *J. Chemom.* **1988**, *2*, 211-228.

[33] Martens H.; Næs, T. *Multivariate Calibration*; John Wiley & Sons Ltd.: Chichester, 1989.

[34] Centner, V.; Massart, D.L.; de Noord, O.E.; de Jong, S.; Vandeginste, B.M.; Sterna, C., Elimination of uninformative variables for multivariate calibration, *Anal. Chem.* **1996**, *68*, 3851-3858.

[35] Bos, M.; Vrielink, J.A.M., Wavelet transform for pre-processing IR spectra in the identification of monoand di-substituted benzenes, *Chemom. Intell. Lab. Syst.* **1994**, *23*, 115-122.

[36] Chau, F.T.; Shih, T.M.; Gao, C.K.; Chan, C.K., Application of the fast wavelet transform method to compress ultra-violet-visible spectra, *Appl. Spectrosc..* **1996**, *50*, 339-348.

[37] Walczak, B.; Massart, D.L., Wavelet packet transform applied to a set of signals: a new approach to the best-basis selection, *Chemom. Intell. Lab. Syst.* **1997**, *38*, 39-50.

[38] Jouan-Rimbaud, D.; Walczak, B.; Poppi, R.J.; de Noord, O.E.; Massart, D.L., Application of wavelet transform to extract the relevant component from spectral data for multivariate calibration, *Anal. Chem.* **1997**, *69*, 4317-4323.

4. CORRECTION OF TEMPERATURE INDUCED SPECTRAL VARIATION BY CONTINUOUS PIECE-WISE DIRECT STANDARDIZATION

Abstract

In process analytical applications it is not always possible to keep the measurement conditions constant. However, fluctuations in external variables such as temperature can have a strong influence on measurement results. For example, non-linear temperature effects on Near-infrared (NIR) spectra may lead to a strongly biased prediction result from multivariate calibration models such as PLS. A new method, called Continuous Piece-wise Direct Standardization (CPDS) has been developed for the correction of such external influences. It represents a generalization of the discrete PDS calibration transfer method and is able to adjust for continuous non-linear influences such as the temperature effects on spectra. It was applied to short-wave NIR spectra of ethanol/water/ 2-propanol mixtures measured at different temperatures in the range 30 - 70 °C. The method was able to remove almost completely the temperature effects on the spectra and prediction of the mole fractions of the chemical components was close to the results obtained at constant temperature.

Based on: Wülfert, F.; Kok, W.Th.; de Noord, O.E.; Smilde, A.K.; *Anal. Chem.* **2000**, 72, 1639-1644.

Introduction

Fast spectroscopic techniques such as near-infrared (NIR) spectroscopy play a prominent role in process analysis[1, 2, 3, 4, 5, 6, 7]. The possibility to measure in-line or on-line and the short analysis time (in the order of milliseconds to seconds) make NIR spectroscopy an interesting alternative to classical process analytical methods such as chromatography, since speed is obviously of paramount importance in process control and monitoring. Since the selectivity of NIR spectroscopy is low, multivariate methods such as partial least squares (PLS)[8, 9, 10] are often necessary for calibration. However, in-line and on-line NIR measurements in an industrial environment can be influenced by fluctuations of external variables such as temperature or pressure[11, 12]. A multivariate calibration model is easily disturbed by such fluctuations and will perform poorly when the effect of the external variation on the NIR spectra is not taken into account.

One possibility to address external fluctuations is by implicit modeling through the inclusion of temperature into the calibration design. In previous work the effect of a fluctuating temperature on the predictive ability of a calibration model for NIR spectra was studied[13]. It was found that with the implicit inclusion of the temperature in the model the predictive ability was still satisfactory; however, the complexity of the calibration model was strongly increased.

Alternative approaches are the explicit inclusion of the fluctuating external variable into the calibration model (as an additional variable) or linear preprocessing methods. However, it has been shown[14] that such methods do not work to correct for temperature effects on NIR spectra and do not lead to better results than implicit modeling. It appears that it is not possible to correct for temperature influences with linear techniques because of the non-linear character of the effects (shape changes caused by an influence on the long range intermolecular and intramolecular forces such as hydrogen bonding[15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]).

For the correction of complex non-linear spectral effects the technique of piece-wise direct standardization (PDS)[32] has been developed, which consists of the multiplication of the spectra with a banded transformation matrix. It has widespread use in situations for calibration transfer between measurements performed on two different instruments or under two different sets of conditions[33, 34, 35, 36, 37, 38, 39]. Although PDS is a linear operation, it can find a linear solution to correct for complex and non-linear differences between two discrete situations, comparable to a straight line that can always be drawn between two points, even if the underlying function is curved.

Since temperature is not a discrete variable, the original PDS method is not suitable for correction of temperature fluctuations. To transform spectra measured at many different (discrete) temperatures, numerous individual PDS models would have to be built, based on an impractical large number of standardization measurements. Moreover, it is not possible to use PDS models for intermediate temperatures (see Figure 1) and the utilization of the experimental data is not optimal, since for each discrete PDS correction model only part of the standardization measurements is used (measurements at two temperatures).

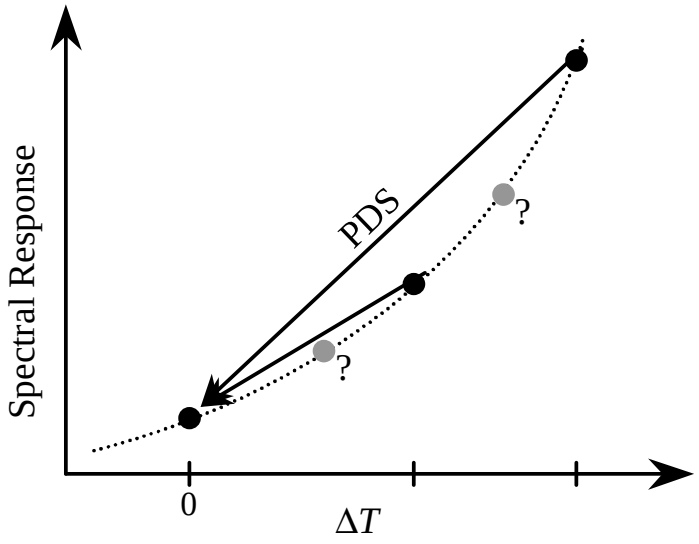


Figure 4-1: Illustration of how the linear PDS correction can deal with discrete temperature differences (black dots) but does not give a solution for measurements at other levels (gray dots).

In this paper a newly developed technique is presented for the correction of external fluctuations on spectral data: continuous piece-wise direct standardization (CPDS). This CPDS technique can be regarded as a generalization of PDS to continuous variables. The method has been evaluated with a set of NIR data measured on ternary mixtures of water, ethanol and 2-propanol at different temperatures. The results are compared with implicit inclusion of the temperature in the calibration model.

Theory

For ease of understanding, the data set which will be used further on as an application is also used as an example in this section. The data set consists of ternary mixtures of ethanol, water and 2-propanol, prepared according to a mixture design[13] and SW-NIR spectra taken at 5 different temperatures (30, 40, 50, 60 and 70°C). Goal of the procedure is to enable the usage of a calibration model built from training spectra and mole fractions ($\mathbf{X}_{train}$, $\mathbf{y}_{train}$) at a certain reference temperature for prediction of samples measured at different temperatures without loosing precision.

Table 4-1: Steps representing the construction and application of CPDS

Description	Parameters	Data used	Result
Build calibration model :	cal. temperature; # of latent variables	Cal. Spectra $\mathbf{X}_{train}$; mole fractions $\mathbf{y}_{train}$	regr. vector $\mathbf{b}_{pls}$
Build PDS transfer matrices:	Window size*; # of lv's*	Standard spectra $\mathbf{X}_{stand}$; $\mathbf{b}_{pls}$	discr. transf. matrices $\mathbf{P}_{\Delta T}$
Build CPDS model:	Degree of polynomial*	ΔT discr. Transf. matr. $\mathbf{P}_{\Delta T}$	cont. transf. model $\hat{\mathbf{P}}(\Delta T)$
Application and Validation:		$\mathbf{X}_{test}$; $\mathbf{y}_{test}$; $\mathbf{b}_{pls}$ T; $\hat{\mathbf{P}}(\Delta T)$	pred. mole fract. $\mathbf{y}_{pred}$; RMSEP

***** Estimation of window size, number of latent variables and degree of polynomial is performed in one cross-validation procedure.

To achieve this a correction model is built from standardization spectra ($\mathbf{X}_{stand}$) taken at different temperatures which will be valid for a whole temperature range. In the following, the method will be explained in detail, a short scheme is given in Table 4-1.

A Partial Least Squares (PLS) calibration model between spectra and mole fractions is built at the calibration temperature from the training set. For the selection of the calibration temperature two possibilities are considered, the

lowest and the midpoint temperature are logical choices. The choice of the lowest calibration temperature might be advantageous from practical and instrumental viewpoint. An example for choosing the lowest temperature is the monitoring or control of a batch process that follows or induces a temperature gradient or program. The calibration temperature would then be chosen as the temperature at which the batch process is started. A midpoint calibration temperature would be preferred for a continuous process application, where the temperature fluctuates around this midpoint.

Next, discrete calibration transform solutions are found by PDS models, which form the foundation of CPDS. The solutions consist of banded transformation matrices **P** that can correct for the variation in spectra between two distinct measurement situations A and B:

$$\mathbf{X}_{\mathrm{A}} = \mathbf{X}_{\mathrm{B}} \cdot \mathbf{P} \qquad \text{Equation 1}$$

where $\mathbf{X}_{\mathrm{A}}$ and $\mathbf{X}_{\mathrm{B}}$ are matrices of dimensions $I \times J$ representing the spectra of I samples taken over a spectral range of J wavelengths and **P** represents the banded transformation matrix of dimensions $J \times J$.

The transformation matrix **P** is obtained from the standardization spectra $\mathbf{X}_{\mathrm{stand}}$ by regressing the absorbance values $x_{\mathrm{A},j}$ (wavelength j, situation A) on a window $x_{\mathrm{B},j-k}$ to $x_{\mathrm{B},j+k}$ (wavelengths $\pm k$ around j, situation B). The regression vectors $\mathbf{b}_j$ (column vector with length $2 \cdot k+1$) are calculated using PLS and are given by:

$$x_{\mathrm{A},j} = \begin{bmatrix} x_{\mathrm{B},j-k} & x_{\mathrm{B},j-k+1} & \cdots & x_{\mathrm{B},j+k-1} & x_{\mathrm{B},j+k} \end{bmatrix} \cdot \mathbf{b}_j + e_j \qquad \text{Equation 2}$$

After calculation of all regression vectors $\mathbf{b}_j$, these form the diagonal band of the transformation matrix **P** by placing $\mathbf{b}_j$ on the *j*-th column, ranging from row *j-k* until *j+k* (see Figure 4-2).

Multiplying spectra measured under condition B with the found **P** matrix will then transform the spectra into a form as if they were measured under condition A (Equation 1). Calculating the **P** matrices for transforming the spectra at the different temperatures (e.g. 40, 50, 60 and 70°C) back to the calibration temperature (e.g. 30°C) gives 4 discrete solutions, that is: $\mathbf{P}_{\Delta T=10}$, $\mathbf{P}_{\Delta T=20}$, $\mathbf{P}_{\Delta T=30}$ and $\mathbf{P}_{\Delta T=40}$. Each $\mathbf{P}_{\Delta T}$ gives the discrete transformation from a temperature which is ΔT higher but interpolation for other temperature differences is not possible.

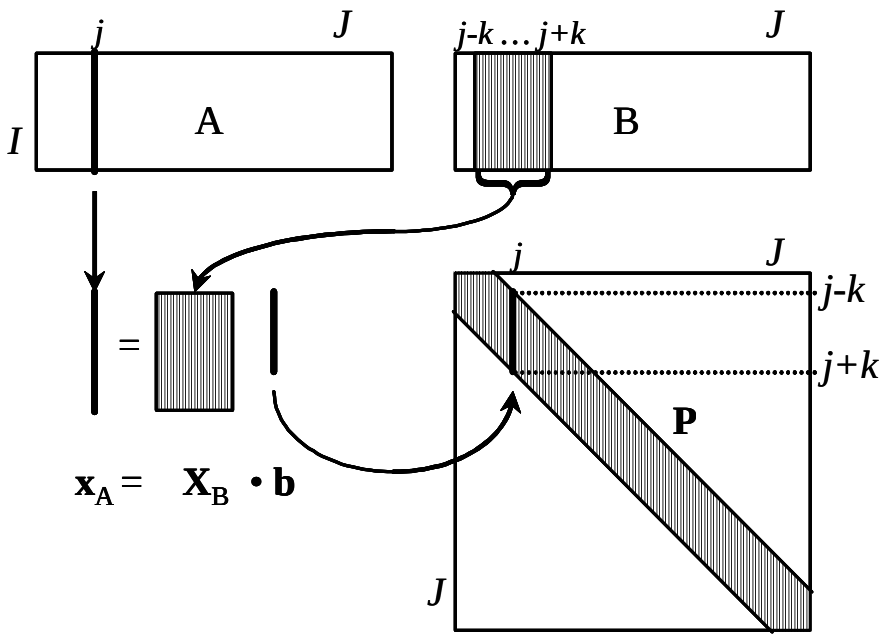


Figure 4-2: Schematic representation of one step of the PDS algorithm, the absorption values at one wavelength (**j**) under situation ***A*** are regressed on the absorption values in a window (**j-k…j+k**) of wavelengths measured under situation ***B***. The resulting regression vector forms the **j**-th column on the band of the transformation matrix.

In order to overcome this restriction a CPDS correction model has to be applied which consists of polynomial fits of the discrete PDS transformation matrices. For each position (*m*,*n*) on the band of the matrices $\mathbf{P}_{\Delta T}$, a polynomial regression is done for the values $p_{m,n}$ on the band against the temperature difference ΔT.

$$p_{m,n}(\Delta T) = a_{m,n}\Delta T^2 + b_{m,n}\Delta T + c_{m,n} + e_{m,n}$$
$$\hat{p}_{m,n}(\Delta T) = a_{m,n}\Delta T^2 + b_{m,n}\Delta T + c_{m,n}$$

Equation 3

This results in estimated transformation matrices $\hat{\mathbf{P}}(\Delta T)$ for all temperature differences that lie in the standardization range. Choosing a first order (straight line) or, as given in Equation 3, a second order (parabolic curve) polynomial gives furthermore the possibility to either describe the temperature dependency of the transformation matrices with a linear or non-linear model. An example of such a fit for one position at the transformation matrices is given in Figure 4-3.

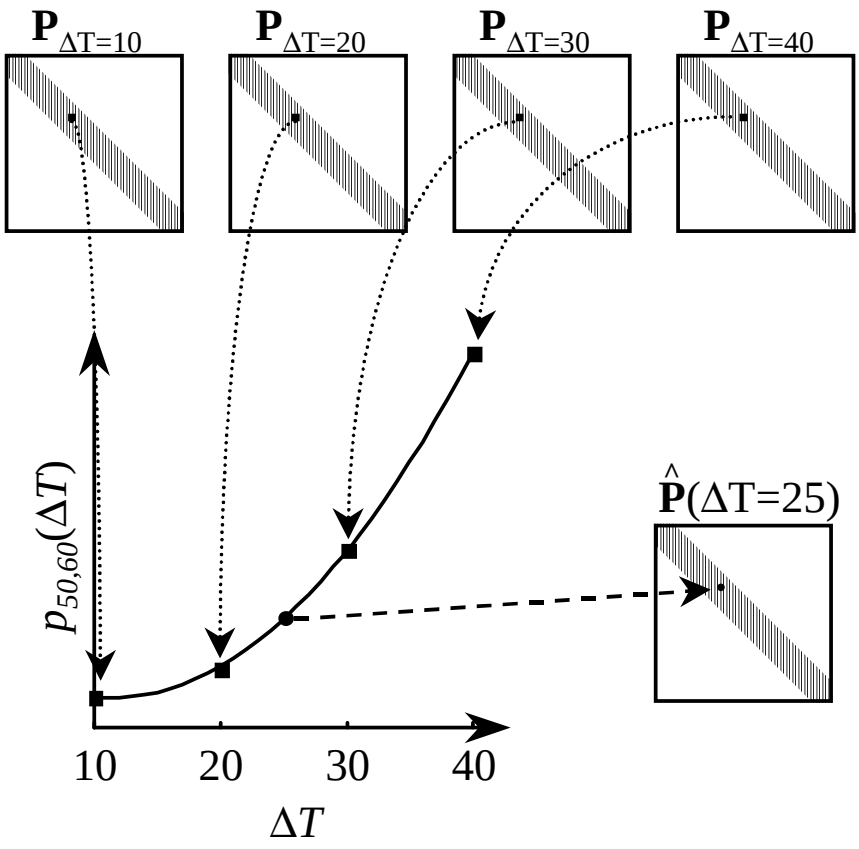


Figure 4-3: Graphical representation of the estimation of the transformation matrix: for each position (e.g. **m=50, n=60**) the $\mathbf{p_{m,n}}$ values from the PDS matrices are fitted with a 2nd degree polynomial. Through the polynomial, an estimated transformation matrix can be found for every $\Delta\mathbf{T}$.

For each new spectrum **x** measured at a certain known temperature ($T=T_{cal}+\Delta T$), the right transformation matrix $\hat{\mathbf{P}}(\Delta T)$ can be built from Equation 3. The temperature influences are consequently removed and the spectrum brought to its temperature corrected version ($\hat{\mathbf{x}}_{Tcorr}$) by:

$$\hat{\mathbf{x}}'_{Tcorr} = \mathbf{x}' \cdot \hat{\mathbf{P}}(\Delta T)$$ Equation 4

The method can now be used for correction of spectra prior to prediction of the mole fractions with the PLS calibration model. This is also done for the spectra of the independent test set $\mathbf{X}_{test}$, where the known concentrations $\mathbf{y}_{test}$ are compared to the predicted $\hat{\mathbf{y}}_{test}$ to estimate the model accuracy. A Root Mean Square Error of Prediction (RMSEP) is calculated according to:

$$RMSEP = \sqrt{\sum_{i=1}^{N_{test}} \frac{\left(y_{i,test} - \hat{y}_{i,test}\right)^2}{N_{test}}}$$ Equation 5

Three parameters have to be estimated to build the CPDS correction method: The number of latent variables and the width of the band (window size 2*k*+1) for the discrete PDS and the degree of the polynomial to generalize these solutions into the CPDS method. The best values for the three parameters are assessed with a leave-one-out cross-validation, although information about e.g. the non-linearity of the external effect (the temperature in this case) can lead to an *a priori* choice of the polynomial degree.

Two measures can be chosen in order to assess the quality of the temperature correction for each cross-validation step of the CPDS models. First, the difference between a temperature corrected spectrum and a spectrum of the same sample measured at the calibration temperature can be considered. This error (E_{corr}) is a measure of the uncorrected temperature effects and artifacts of the correction model in the spectral (or **X**-) space:

$$E_{corr} = \sum_{j=1}^{J} \frac{\left(\hat{x}_{j,T\,corr.} - x_{j,Tcal}\right)^2}{J}$$ Equation 6

where $\hat{x}_{j,Tcorr}$ and $x_{j,Tcal}$ represent the absorbance values of the temperature corrected and at calibration temperature measured spectra with lengths (J) for one sample.

Second, a measure (E_{mod}) for how the spectral differences translate into a difference in prediction of mole fractions (**y**-space) can also be used as cross-validation criterion. For this, the difference between the temperature corrected ($\hat{\mathbf{x}}_{Tcorr}$) and at calibration temperature measured spectrum ($\mathbf{x}_{Tcal}$) is multiplied with the regression vector from the calibration model ($\mathbf{b}_{PLS}$):

$$E_{\mathrm{mod}} = \left(\left(\hat{\mathbf{x}}_{Tcorr.} - \mathbf{x}_{Tcal} \right)' \cdot \mathbf{b}_{PLS} \right)^2 = \left(\hat{y}_{Tcorr.} - \hat{y}_{Tcal} \right)^2 \qquad \text{Equation 7}$$

This is not the same as a prediction error used for cross-validation or assessing the final model quality (see Equation 5), since it does not compare a predicted and a real but two predicted mole fractions. Therefore, the **y**-values for the standardization samples do not have to be known, which can be advantageous. But more importantly, the error only due to correction is minimized without mixing it with an calibration error, which should be minimized with the proper calibration model.

Considering that the performance of the CPDS method can vary for different spectral ranges, E_{mod} does not have to follow the same trend as E_{corr}. By using E_{mod} the spectral difference is weighed with the regression vector of the calibration model. Remaining temperature effects or introduced correction artifacts are thus allowed for wavelengths where the values in $\mathbf{b}_{PLS}$ are near zero.

Experimental Section

The ternary mixtures of ethanol, water and 2-propanol were prepared using an analytical balance. Short-wave NIR spectra of the mixtures were taken from 580 to 1091 nm, the spectral range between 749-849 nm was used for slope and offset correction and the range from 850 to 1049 nm was used for data analysis. The measurements were done using an HP 8454 spectrophotometer equipped with a thermostatically controlled cell holder with stirring module. The temperature was measured and controlled in the closed quartz cells using a Pt-100 sensor linked to a Neslab EX-111 circulator bath. The data analysis was done on a Pentium class computer using Matlab ver.5.2 (Mathworks Inc.) and the PLS toolbox ver.1.5.3b (Eigenvector Research Inc.). Training and standardization sets were chosen from samples representing the edge and center of the mixture design while the samples in between were used as test set. Training and standardization samples do not necessarily have to be the same, usually the standardization samples form a subset of the training set. For this article, however, the training and standardization samples were chosen to be the same to make a fair comparison possible with results obtained with local and global models[13]. The local models were built separately for each temperature and only used for prediction at the same temperature. As they are free of temperature influences, their prediction error can be considered a lower limit. Global models were built from the training samples measured at all 5 temperatures and also used for prediction at all temperatures. They implicitly include temperature effects by using more latent variables (7) than the local models (4). The predictive ability of global models has been shown to be a good indication for how well linear techniques are able to cope with temperature effects[14].

Results and Discussion

Mean centered spectra and mole fractions of ethanol, water and 2-propanol are used as predicting (**X**) and predicted (**y**) variables respectively. For each chemical component one calibration model is built from the training samples using PLS1. As described in an earlier article, leave-one-out cross-validation leads to calibration models requiring 4 latent variables[13]. The study is performed with both 30°C and 50°C as calibration temperature, spectra from the test set measured at the calibration temperature can be predicted directly without correction.

The samples that are used as training set for the calibration model are also used as standardization set but no mean centering is performed. For both calibration temperatures (30, 50°C) the leave-one-out cross-validation was done varying the following factors: the number of latent variables (1 to 5) and the window size (number of neighboring wavelengths *k*) used to build the $\mathbf{P}_{\Delta T}$ matrices (2 to 20), and the degree of the polynomials (1 to 2) used to build the estimated transformation matrices $\hat{\mathbf{P}}(\Delta T)$. For both errors, the spectral difference (E_{corr}; Equation 6) and the prediction difference (E_{mod}; Equation 7), the average square sum over the cross-validation samples is minimized.

The spectral difference (E_{corr}) decreases steeply with increasing window size until it stabilizes around *k*=10 to 15 (window sizes 21 to 31) giving the best results for 3 latent variables used for PDS and a 2nd order polynomial fitting in order to estimate $\hat{\mathbf{P}}(\Delta T)$. The model error due to the uncorrected difference (E_{mod}) shows generally the same trend up to *k*=12 (window size 25), from where on a increase in prediction difference can be noticed. Therefore, a temperature correction model with window size of 25 (*k*=12), 3 latent variables for calculating **P** and 2nd degree polynomials to estimate $\hat{\mathbf{P}}(\Delta T)$ was chosen to build the final CPDS correction models. Note that the 2nd degree polynomial estimation gave not only significantly better results

than an 1st order estimation, which can be expected from the non-linearity of the temperature effect, but was in most cases even better than direct correction with the discrete PDS solutions ($\mathbf{P}_{\Delta T}$). This indicates that, through the polynomial fit, using the information at all measured temperature levels smoothes the transformation matrices, which is advantageous. As an example for the cross-validation results the prediction difference of water using a calibration temperature of 30°C is shown in Figure 4-4.

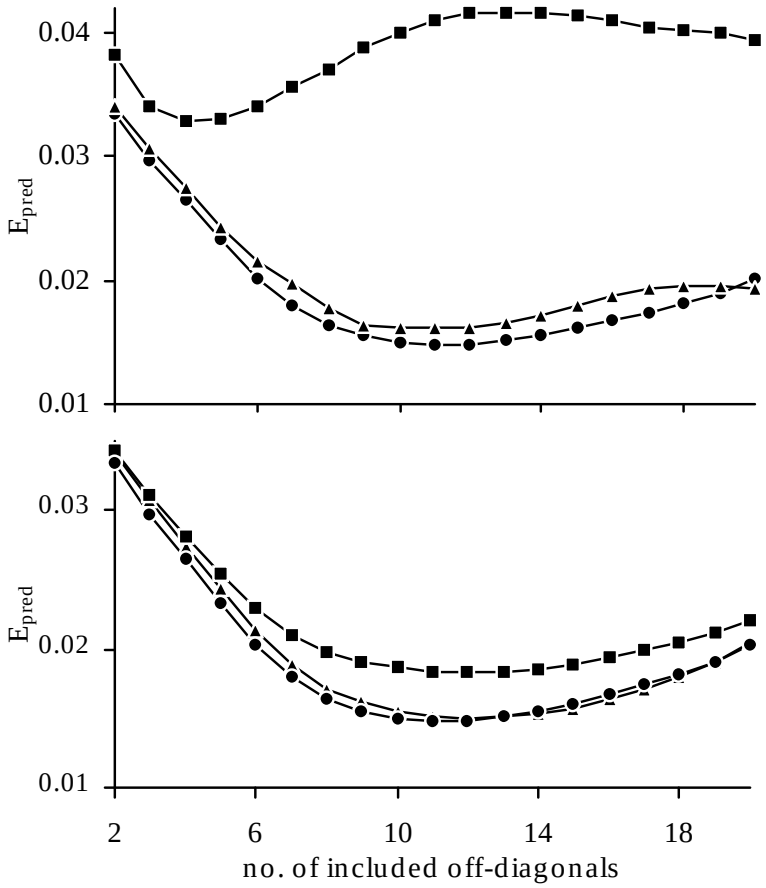


Figure 4-4: Prediction difference for water plotted against the window size. Top: PDS with 2 (squares), 3 (circles) and 4 (triangles) latent variables. Bottom: correction with original (triangles), with 1st degree polynomials estimated (squares) and 2nd degree polynomials (circles) estimated transformation matrices.

For finally assessing the prediction error, the spectra of the test set samples were corrected with estimated transformation matrices. Figure 4-5 shows how the temperature effect is effectively removed from the spectra. The corrected test set spectra are then used with the calibration model and the prediction error of the resulting mole fractions is estimated with the root mean squared error (*RMSEP,* Equation 5),

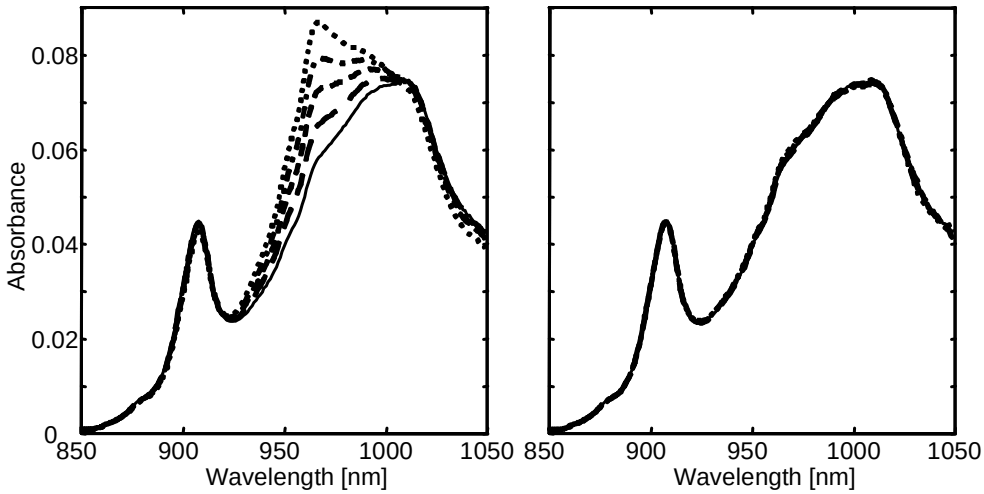


Figure 4-5: Spectra of a test sample before (left) and after correction (right) to lowest temperature. Temperatures of the sample were 30°C (solid line), 40°C (long dashed line), 50°C (dashed line), 60°C (dash-dotted line), 70°C (dotted line).

The errors are given in Table 4-2 for both considered calibration temperatures. Choosing a midpoint calibration temperature leads to better results as the temperature correction does only need to correct for a maximum difference of 20°C and does not need to extrapolate from the higher temperatures to the lowest temperature. Choosing a calibration temperature on the extremes of the temperature range is therefore only to be considered when the application leaves no other option (e.g. following a reaction in temperature programmed batch process) and the results of the midpoint temperature will be considered in the following.

Table 4-2: Prediction errors (RMSEP) of test set after application of CPDS correction model.

Temp. of Calibration→ ↓Prediction	Ethanol		Water		2-propanol	
	30°C	50°C	30°C	50°C	30°C	50°C
30°C	0.0177	0.0152	0.0092	0.0094	0.0124	0.0107
40°C	0.0119	0.0111	0.0058	0.0074	0.0148	0.0128
50°C	0.0149	0.0166	0.0187	0.0111	0.0218	0.0218
60°C	0.0156	0.0091	0.0088	0.0059	0.0118	0.0048
70°C	0.0156	0.0154	0.0131	0.0070	0.0080	0.0136
mean:	**0.0152**	**0.0135**	**0.0111**	**0.0082**	**0.0138**	**0.0127**

For comparison the performance of local and global models are given in Table 4-3. It is evident that the CPDS temperature correcting model performs considerably better than a linear method for prediction of the alcohols and comparable for prediction of water. It also leads to simpler calibration models since no extra latent variables are needed for describing the temperature effects. But an extra effort is needed for building and validating the temperature correction model.

Table 4-3: Comparison of average prediction errors for local, global and CPDS-corrected calibration models.

Model	Temperature (°C)		RMSEP		
	Calibration	Prediction	Ethanol	Water	2-propanol
Local*	30,40,…,70	30,40,…,70	0.013	0.0070	0.013
Global	30-70	30-70	0.020	0.0084	0.021
CPDS	30	30-70	0.015	0.0111	0.014
CPDS	50	30-70	0.014	0.0082	0.013

***** 5 separate models, calibration and predictions at the same temperature.

Comparing with the prediction errors of the local models it can be stated that the temperature correction model removes the temperature effect almost completely for prediction of the alcohols, as the results are

comparable, but not for water where the prediction still is slightly worse than for local models.

The difference in the results for the alcohols and water can be explained from the sources of variation present in the spectra. Water has higher absorption coefficients and larger temperature effects than the alcohols and therefore the variance present in the spectra is dominated by water and its temperature effect. Since the PLS calibration model maximizes the covariance between spectra and mole fractions, water is easier to predict, as can be observed from the local models. Furthermore, the temperature influence on water is so large, that it is to some extent indicative for water presence in mixtures. This can explain the good performance of global models for water. While the high temperature effects and absorbances for water aid the prediction of water itself, they represent a large interference to models for prediction of both alcohols. Therefore mainly the alcohol predictions will benefit from a removal of the temperature effects by the CPDS method.

Conclusions

It has been shown that spectroscopic measurements under influence of temperature can very well be used for analytical purposes by applying a temperature correction model. This is achieved by transforming the discrete PDS calibration transfer method into the continuous CPDS model by finding a relation between the transformation matrices of the discrete PDS solutions. Choosing a non-linear model as a function of temperature for these relations provides the means to find a model that can correct for non-linear and non-discrete effects. The found solution is superior to implicit or explicit inclusion of temperature by the calibration model itself, as evidenced by the lower prediction errors found.

Additionally the CPDS approach makes it possible to combine in one single model the conventional use of PDS calibration transfer with the correction for continuous temperature effects. The calibration set can be measured with high precision instrumentation under well controlled laboratory conditions while the standardization samples (except for those at calibration temperature) are measured on a more robust industrial process system.

A limiting factor for the usage of a temperature correction model is the requirement to measure and know the temperature for every spectrum, not only for building the correction model but also for samples to be predicted. This is not necessary for global models because of the treatment of temperature as an unknown interferent. Furthermore, it is necessary for a CPDS model that the standardization measurements are made for exactly the same standardization samples. Global models do not require this, enabling the use of e.g. historical process data.

From the encouraging results of this study it can be concluded that the combination of spectroscopic analysis with the right chemometric tools can lead to more a prominent role for analytical chemistry in technical application fields such as industrial process monitoring and control.

References

[1] Blaser, W. W.; Bredeweg, R. A.; Harner, R.S.; LaPack, M.A.; Leugers, A.; Martin, D. P.; Pell, R. J.; Workman, J., Jr.; Wright, L. G. *Anal. Chem.* **1995**, *67*, 47R-70R.

[2] Cooper, J.B., Wise, K.L., Welch, W.T., Sumner, M.B., Wilt, B.K., Bledsoe, R.R. *Appl. Spectrosc.* **1997**, *51*, 1613-1620.

[3] Dailey, Wm.V. *Proc. Contr. & Qual.* **1992**, *3*, 99-106.

[4] DeThomas, F. A.; Hall, J. W.; Monfre, S.L. *Talanta* **1994**, *41*, 425-431.

[5] Frank, I. E.; Feikema, J.; Constantine, N.; Kowalski, B. R. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 20-24.

[6] Hall, J. W.; McNeil, B.; Rollins, M. J.; Draper, I.; Thompson, B. G.; Macaloney, G. *Appl. Spectrosc.* **1996**, *50*, 102-108.

[7] Siesler, H.W. *Landbauforschung Völkenrode* **1989**, *107*, 112-118.

[8] Geladi, P. *J. Chemom.* **1988**, *2*, 231-246.

[9] Geladi, P.; Kowalski, B.R. *Anal. Chim. Acta* **1986**, *185*, 1-17.

[10] Höskuldsson, A., *J. Chemom.* **1988**, *2*, 211-228.

[11] DeBraekeleer, K.; Cuesta Sánchez, F.; Hailey, P.A.; Sharp, D.C.A.; Pettman, A.J.; Massart, D.L. *J. Pharmac. Biomed. Analysis* **1998**, *17*, 141-152.

[12] Yalvac, E.D.; Seasholtz, M.B.; Beach, M.A.; Crouch, S.R. *Appl. Spectrosc.* **1997**, *51*, 1565-1572.

[13] Wülfert, F.; Kok, W.Th.; Smilde, A.K. *Anal. Chem.* **1998**, *70*, 1761-1767.

[14] Wülfert, F.; Kok, W.Th.; Noord, O.E.de; Smilde, A.K. *Chemom. Intell. Lab. Syst.* **2000**,. 51, 189-200

[15] Cho, T.; Kida, I.; Ninomiya, J.; Ikawa, S. *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 103-107.

[16] Czarnecki, M.A.; Czarnecka, M.; Ozaki, Y.; Iwahashi, M. *Spectrochim. Acta* **1994**, *50*, 1521-1528.

[17] Czarnecki, M.A.; Liu, Y.; Ozaki, Y.; Suzuki, M.; Iwahashi, M. *Appl. Spectrosc.* **1993**, *47*, 2162-2168.

[18] Finch, J. N.; Lippincott,k E. R. *J. Chem. Phys* **1956**, *24*, 908-909.

[19] Finch, J. N.; Lippincott,k E. R. *Phys. Chem.* **1957**, *61*, 894-902.

[20] Hazen, K. H.; Arnold, M. A.; Small, G. W. *Appl. Spectrosc.* **1994**, *48*, 477-483.

[21] Iwata, T.;Koshoubu, J.; Jin, C.; Okubo, Y. *Appl. Spectrosc.* **1997**, *51*, 1269-1275.

[22] Kamiya, N.; Sekigawa, T.; Ikawa, S. *J. Chem. Soc. Faraday Trans.* **1993**, *89*, 489-493.

[23] Libnau, F.O.; Kvalheim, O.M.; Christy, A. A.; Toft, J. *Vib. Spectrosc.* **1994**, *7*, 243-254.

[24] Lin, J.; Brown, C.W. *Appl. Spectrosc.* **1993**, *10*, 1720-1727.

[25] Liu, Y. ; Czarnecki, M.A.; Ozaki, Y. *Appl. Spectrosc.* **1994**, *48*, 1095-1101.

[26] Liu, Y.; Czarnecki, M.A.; Ozaki, Y.; Suzuki, M.; Iwahashi, M. *Appl. Spectrosc.* **1993**, *47*, 2169-2171.

[27] Noda, I. ; Liu, Y.; Ozaki, Y.; Czarnecki, M.A. *J. Phys. Chem.* **1995**, *99*, 3068-3073.

[28] Okuyama, M.; Ikawa, S. *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 3065-3069.

[29] Ozaki, Y.;Liu ,Y.; Noda, I. *Appl. Spectrosc.* **1997**, *51*, 526-535.

[30] Pegau, W. S.; Zaneveld, J. R. V. *Limnol. Oceanogr.* **1993**, *38*, 188-192.

[31] Wang, F.C.; Feve, M.; Lam, T. M.; Pascault, J. P. *J. Polym. Sci.: Phys.* **1994**, *32*, 1305-1313.

[32] Wang, Y.; Veltkamp, D.J.; Kowalski, B.R. *Anal. Chem.* **1991**, *63*, 2750-2756.

[33] Bouveresse, E.; Massart, D.L. *Chemom. Intell. Lab. Syst.* **1996**, *32*, 201-213.

[34] Lin, J. *Appl. Spectrosc.* **1998**, *52*, 1591-1596.

[35] Noord, O.E.de *Chemom. Intell. Lab. Syst.* **1994**, *25*, 85-97.

[36] Sales, F.; Callao, M.P.; Rius, F.X. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 63-73.

[37] Swierenga, H.; Haanstra, W.G.;Weijer, A.P.de; Buydens, L.M.C. *Appl. Spectrosc.* **1998**, *52*, 7-16.

[38] Wang, Y.; Kowalski, B.R. *Anal. Chem.* **1993**, *65*, 1301-1303.

[39] Wang, Y.; Lysaght, M.J.; Kowalski, B.R. *Anal. Chem.* **1992**, *64*, 562-564.

5. DEVELOPMENT OF ROBUST CALIBRATION MODELS IN NIR SPECTROSCOPIC APPLICATIONS

Abstract

When spectral variation caused by factors different from the parameter to be predicted (e.g. external variations in temperature) is present in calibration data, a common approach is to include this variation in the calibration model. For this purpose, the calibration sample spectra measured under standard conditions and the spectra of a smaller set measured under changed conditions are combined into one dataset and a global calibration model is calculated. However, if highly nonlinear effects are present in the data, it may be impossible to capture this external variation in the model. Recently, a new technique based on selection of robust variables was proposed for constructing robust calibration models. In this technique, a calibration model is developed which uses a subset of spectral values that are insensitive to external variations.

This new technique is compared to global calibration models for constructing robust models in spectroscopic applications. Both techniques are applied to two different NIR spectroscopic applications. The first application is the determination of the ethanol, water, and iso-propanol concentrations in a ternary mixture of these components and the second application is the determination of the density of heavy oil products. In both applications the calibration set spectra have been measured at standard sample temperature, and a subset has been measured at sample temperatures deviating from the standard temperature. It has been found that models based on robust variable selection are similar or sometimes better than global calibration models with respect to their predictive ability at different sample temperatures.

Based on: Swierenga, H.; Wülfert, F.; de Noord, O.E.; de Weijer, A.P.; Smilde, A.K.; Buydens, L.M.C.; *Anal. Chim. Acta* **2000**, 411, 121-135.

Introduction

Multivariate calibration models are often associated with vibrational spectroscopic techniques in order to predict physical or chemical sample properties from the spectra. To construct a multivariate calibration model, the spectra and corresponding properties of many samples need to be measured in order to capture the variation in the sample properties to be predicted. Once the model has been developed, it is supposed to be valid for a long period of time. This implies that after this period the model's prediction error is not significantly different from the prediction error obtained during calibration. However, there may be various reasons why the model makes erroneous predictions: replacement of the instrument or part of it, ambient changes such as temperature, and changes in physical sample conditions.[1]

If the calibration model loses its validity, a new calibration model needs to be constructed. Therefore, a set of calibration samples, representative of the original calibration samples should be remeasured under the changed conditions. If the original calibration samples are not stable, this calls for collecting or preparing new samples, measuring of the reference values, and measuring the corresponding spectra, which may involve a large amount of work. Recently, more efficient methods, known as multivariate calibration standardization methods, became available to establish a new calibration model.[1] Multivariate calibration standardization methods can be divided into two categories: 1) Improvement of robustness of the calibration model; and 2) Adaptation of the calibration model.[2] The first category aims to improve the selectivity of the calibration model by data preprocessing (e.g. variable selection), the incorporation of measurement conditions into the calibration model (global calibration models) and/or the application of robust multivariate calibration techniques such as IVS-PLS.[3] The second category includes techniques that transform the measured spectra, the model’s regression parameters or the predictions by the calibration model (e.g. bias/slope correction, direct standardization and piecewise direct

standardization). One of the disadvantages of this category is that the same sample subset needs to be measured in both the old and the new situation, which is not possible when unstable samples are involved. Another disadvantage of techniques of category two is that they are only applicable to discrete situations such as instrumental changes. Frequently, however, external conditions (e.g. sample temperature) which influence the model's predictions change continuously and, consequently, techniques of category two cannot be applied. In the applications studied in this chapter, the sample temperature is a continuously changing condition, and we therefore focused on two techniques of the first category: a) Global calibration models;[4] and b) Robust variable selection models.[5]

Although often not recognized, global calibration models are frequently used. The construction of a global calibration model involves measurement of calibration samples under normal conditions, measurement of these samples or a sample subset under changed conditions and the combination of the data to one dataset. Besides spectral variation caused by the variation in the reference parameter, this dataset includes external spectral variation introduced by the new situation. Subsequently, a new calibration model is calculated on the basis of the joint dataset. Thus, global calibration models try to model the external spectral variation and implicitly include the external variation into the calibration model.

Recently, a new technique based on variable selection was presented in order to enhance the robustness of a calibration model.[5] Instead of using the whole spectral range for modeling, this technique uses a subset of spectral values which is not sensitive to the changing conditions and rejects those spectral regions that are sensitive to these changing conditions. There are various reasons why the predictive ability and the robustness of a calibration model are enhanced by variable selection: 1) some spectral regions related to the parameter of interest may contain large variation caused by external influences such as temperature variations or interferents; 2) there may be spectral regions whose intensities

(absorbances) are not linearly related to the parameter to be predicted; and 3) there may be spectral regions which exhibit an indirect correlation with the parameter of interest (apparent causalities). This makes variable selection especially suitable for situations in which the spectral variation caused by external changes are localized in the spectra. Thus, instead of modeling the external variation, robust variable selection excludes external spectral variation before modeling.

In this chapter global calibration models are compared to the new technique for enhancement of model robustness, namely calibration models based on robust variable selection. In order to select the robust variables, simulated annealing was used. Both techniques were applied to two different NIR spectroscopic applications. The first application is the determination of the ethanol, water, and iso-propanol concentration in a ternary mixture of these components and the second application is the determination of the density of heavy oil products. In both applications the model's predictions should be insensitive to sample temperature variations within a predefined temperature range. In this chapter only partial least squares (PLS) regression models are considered, but the above-mentioned techniques can be applied to other multivariate calibration techniques as well.

Theory

Global calibration models

Global models try to include implicitly the variation due to external effects in the model, in much the same way as unknown chemical interferents can be included in an inverse calibration model. As long as the interfering variation is present in the calibration set, an inverse calibration model can, in the ideal case of additivity and linearity, easily correct for the variation due to the unknown interferents. It is assumed in global calibration models that the new sources of spectral variation can be modeled by including a limited number of additional PLS factors.[4] Due to the increase of the calibration model's dimensionality, it becomes necessary to measure a large number of samples under changed conditions in order to make a good estimation of the additional parameters.[6] When highly nonlinear effects are present in the spectra, a lot of additional PLS factors will be necessary to model the spectral differences while, sometimes, it is not even possible to model these spectral differences. Therefore, other strategies need to be used to make modeling of nonlinear data possible.[7]

Robust variable selection models

Whereas global models try to capture the external variation into the model, robust variable selection attempts to exclude the external variation before modeling. Basically, it selects those spectral regions that are important for the parameter to be predicted and those that can correct for the spectral differences caused by external conditions, at the same time rejecting those regions that are sensitive to the spectral differences caused by the external variations. It is assumed that a calibration on the robust wavelengths will be free of influences by external factors and may be more parsimonious, as it only needs to model the spectral variation caused by the parameter of interest. However, it is difficult to compare variable selection with other calibration models with respect to parsimony because it is difficult to assess

the degrees of freedom lost in the selection of the robust variables (a lot of models are calculated during optimization).[8]

While global calibration models are straightforward and the model calculations can be performed in a short time by commercially available software packages, the robust variable selection by simulated annealing requires more sophisticated software and faster computers. Furthermore, some additional parameters need to be optimized for the simulated annealing algorithm (number of PLS factors, the number of variables selected, representation of problem, length of Markov chain, initial temperature, or control parameter). Therefore, special expertise about the simulated annealing techniques is necessary.

Since the number of selected variables will be seriously reduced and a lot of models are calculated during optimization, there is a possibility of overfitting; the selected variable subset should not include irrelevant noise-containing variables and overfitting should be prevented. Recently, Jouan-Rimbaud *et al.* developed a method to evaluate the performance of variable selection by genetic algorithms (GAs) with respect to overfitting.[9] For this purpose, they added random variables to the original spectral data matrix and performed a GA run using this extended dataset. The amount of selected random variables is a measure of the selection of randomly correlated variables from the original spectral data. Leardi *et al.* proposed a stopcriterion for a variable subset search by a genetic algorithm in order to prevent overfitting.[10] This stopcriterion is based on a random permutation test of the original Y variables.

All problems associated with robust variable selection result from the use of simulated annealing and not from the principle of using a spectral subset for PLS modeling instead of using the whole spectral range. If knowledge were available about the relation between external variables and the spectral intensities, these problems would disappear. Usually, however, no physical model is available for estimating the influence of external variations on the

spectral variables. As a result, variable selection techniques need to be used to make the spectral subset selection.

Simulated annealing

Since no prior knowledge is available, the selection of the robust variables from the whole spectral range is a large optimization problem which can be solved by optimization techniques such as simulated annealing or genetic algorithms. In this chapter, simulated annealing is used for variable selection. Simulated annealing is a probabilistic global optimization technique based on the physical annealing process of solids. In contrast to deterministic optimization techniques (e.g. simplex optimization), probabilistic optimization techniques allow acceptance of an inferior solution during optimization. Consequently, probabilistic optimization techniques have the ability to escape from a local optimum and find the global optimal solution. More detailed description about simulated annealing can be found in ref. [5] and [11].

A simulated annealing solution is represented as a numerical string containing *k* values (integers) representing the variables to be selected from the whole spectral range of *N* variables. These *k* variables are selected from the calibration set spectra and in combination with the reference values of the corresponding samples, a PLS model with a predefined number of factors is calculated. Subsequently, the same *k* variables are selected from the standardization set spectra (spectra measured under changed circumstances) and the reference parameters are predicted using the calibration set and the standardization set variable subset spectra. On the basis of these prediction results an error value is calculated. This error value comprises the predictive ability of the model at the standard temperature and the predictive ability of the model when it is used at different temperatures. The goal of the simulated annealing is to minimize the error value; which implies that the prediction error of the model is minimized at all temperatures. In order to find the proper *k* value, various simulated annealing runs are performed using different values for *k*.

Comparison of predictive accuracy of models

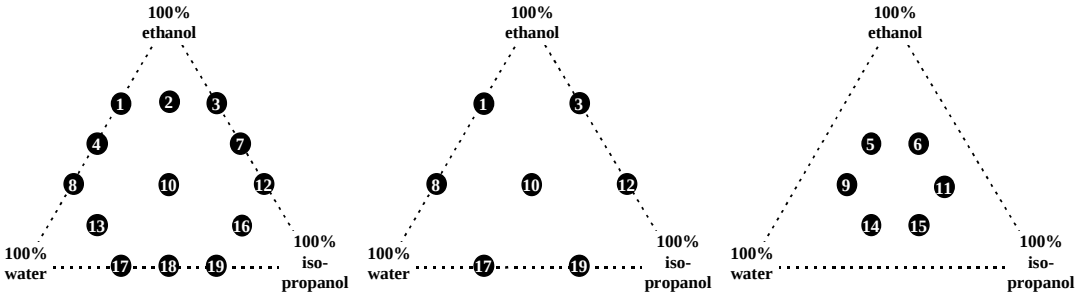
Usually, two models are compared with respect to their predictive ability on a representative independent data set (i.e. dataset not used for model calculation). Frequently, the predictive ability of a model is expressed in the mean squared error of prediction (MSEP). During the development of a calibration model, a minimal MSEP value is aimed at. Recently, Swierenga *et al*. proposed a strategy which uses the prediction error and, simultaneously, the sensitivity to external variations for selecting a multivariate calibration model.[12] Van der Voet proposed a randomization *t*-test to compare the predictive accuracy of two models using the distribution of prediction errors.[13] In this chapter this randomization *t*-test is applied in order to compare the predictive ability of global models and robust variable selection models.

Experimental

Dataset A: Ternary mixture of ethanol, water, and iso-propanol

The mixtures (19 samples) were prepared from p. a. quality alcohols and subboiled water according to a mixture design (Figure 5-1).[4] Short-wave NIR measurements (580 to 1091 nm, 1 nm resolution, 20 s integration time) were performed on a Hewlett Packard HP 8453 spectrophotometer with a thermostatically controlled cell holder and cell stirring module. Closed quartz cells with 1 cm path length were used with an external Pt-100 sensor immersed in the sample linked to a circulator bath for temperature control and measurement. Instrumental baseline drift and offset of the spectra was corrected with straight line fits using the wavelength range 749-849 nm. The data analysis was performed on the region 850-1049 nm.

The spectra of these nineteen ternary mixtures of ethanol, water and iso-propanol were measured at 50°C. The dataset was split into a calibration set (Figure 5-1A) containing the samples 1, 2, 3, 4, 7, 8, 10, 12, 13, 16, 17, 18, 19 and a test set (Figure 5-1C) containing the samples 5, 6, 9, 11, 14, 15. The calibration set will be denoted as X_{cal}^{50} and the test set as X_{test}^{50}. A subset of the calibration set (Figure 5-1B) containing the samples 1, 3, 8, 10, 12, 17, 19 was measured at 30, 40, 60, and 70°C and will be denoted as X_{stand}^{30}, X_{stand}^{40}, X_{stand}^{60}, X_{stand}^{70}, respectively. The test set samples were measured at the same temperatures and will be denoted as X_{test}^{30}, X_{test}^{40}, X_{test}^{60}, X_{test}^{70}, respectively. These datasets were used to calculate and validate the global calibration model and the calculation of a model containing robust wavelengths.



A) **B)** **C)**

Figure 5-1 Construction of datasets used for multivariate calibration
A) calibration set measured at one temperature (50°C)
B) standardization set (subset of calibration set) measured at 30, 40, 60, and 70°C
C) test set measured at 30, 40, 50, 60, and 70°C

Local calibration model

Local models were built to evaluate the influence of temperature on the model's predictions if temperature effects are not taken into account at all. A PLS1 calibration model was calculated based on the spectra X_{cal}^{50}. The number of PLS factors is four, which was determined by leave-one-out cross-validation. The datasets X_{test}^{30}, X_{test}^{40}, X_{test}^{50}, X_{test}^{60}, and X_{test}^{70} were used as independent test sets.

Global calibration model

Datasets X_{stand}^{30}, X_{stand}^{40}, X_{cal}^{50}, X_{stand}^{60}, and X_{stand}^{70} were used to calculate global calibration (PLS1) models, and datasets X_{test}^{30}, X_{test}^{40}, X_{test}^{50}, X_{test}^{60}, and X_{test}^{70} were used as independent test sets. The number of PLS factors for the model was determined by leave-one-sample-out cross-validation and the optimal model complexity is seven factors for all three components in the ternary mixtures.

Robust variable selection

Out of the whole set of possible variables (200 variables), a subset of *k* variables was proposed as a possible solution by the simulated annealing algorithm. This subset of variables was selected from dataset X_{cal}^{50}, and a PLS model was calculated using four PLS factors (number of factors for local models). Subsequently, this model was used to make predictions of the contents using the spectra from sets X_{stand}^{30}, X_{stand}^{40}, X_{stand}^{60}, and X_{stand}^{70}. On the basis of the prediction results, an error value was calculated representing the predictive ability of the model at various temperatures (30, 40, 50, 60, and 70°C). During a simulated annealing run this error value was minimized. At the end of the simulated annealing search the calculated model was tested using the independent datasets X_{test}^{30}, X_{test}^{40}, X_{test}^{50}, X_{test}^{60}, and X_{test}^{70}. Ten random initialized simulated annealing runs were performed at a certain *k* value.

Dataset B: Density of heavy oil products

NIR spectra (6206 - 3971 cm^{-1}, 1.9 cm^{-1} data point and 3.8 cm^{-1} spectral resolution) of the heavy oil products were measured on a Bomem MB 160 FTNIR spectrometer in a temperature controlled flow cell. The density measurements were performed following the ASTM D4052 method. Baseline offset correction of the spectra was applied by subtracting the average absorbance in the range 4810-4800 cm^{-1}. The last 400 variables (4740-3971 cm^{-1}) were used for the data analysis.

The spectra of 42 heavy oil samples were measured at 100°C. This calibration set of 42 samples will be denoted as X_{cal}^{100}. Subsequently, 15 samples were selected from the calibration set using the Kennard Stone algorithm [14, 15] and this subset was measured at 95 and 105°C. These standardization sets will be denoted as X_{stand}^{95} and X_{stand}^{105}, respectively.

Furthermore, a test set containing 35 samples was measured at 95, 100, and 105°C and the test set spectra will be denoted as X_{test}^{95}, X_{test}^{100}, and X_{test}^{105}, respectively.

Local model

Dataset X_{cal}^{100} was used to calculate the local model for the prediction of the density. The model complexity was determined by leave-one-out cross-validation and was set to five factors. Datasets X_{test}^{95}, X_{test}^{100}, and X_{test}^{105} were used as independent test sets.

Global calibration model

Datasets X_{stand}^{95}, X_{cal}^{100}, and X_{stand}^{105} were used to calculate a global calibration model and datasets X_{test}^{95}, X_{test}^{100}, and X_{test}^{105} as independent test sets. The number of PLS factors for the model was determined by leave-one-sample-out cross-validation and the optimal model complexity is six factors.

Robust variable selection

Out of the whole set of possible variables (400 variables), a subset of *k* variables was proposed as a possible solution by the simulated annealing algorithm. This subset of variables was selected from dataset X_{cal}^{100}, and a PLS model was calculated for these spectral variables and the corresponding density using five PLS factors. Subsequently, this model was used to make predictions about the density using the spectra of sets X_{stand}^{95}, and X_{stand}^{105}. On the basis of the prediction results, an error value is calculated which represents the predictive ability of the model at various temperatures (95, 100, and 105°C). During a simulated annealing run this error value was minimized. At the end of the simulated annealing search the calculated model was tested using the independent datasets X_{test}^{95}, X_{test}^{100}, and X_{test}^{105}. Ten random initialized simulated annealing runs were performed at a certain *k* value.

Model validation

The test sets were used to validate the constructed calibration models (dataset A: X_{test}^{30} , X_{test}^{40} , X_{test}^{50} , X_{test}^{60} , X_{test}^{70} and dataset B: X_{test}^{95} , X_{test}^{100} , and X_{test}^{105}). These datasets were used to predict the component concentrations in the ternary mixtures (dataset A) and the density in the oil samples (dataset B). The difference between the predicted and the reference values is expressed in the prediction error:

$$RMSEP = \left(\sum_{n=1}^{N} \frac{\left(\hat{y}_n - y_n \right)^2}{N} \right)^{\frac{1}{2}}$$

where *N* is the number of samples in the test set; $\hat{y}_n$ and y_n are the predictions and the reference values of the samples of the test set, respectively.

Software and algorithms

For local and global models Matlab™ [16] and the PLS Toolbox [17] for Matlab™ were used. For the robust variable selection a simulated annealing toolbox has been written in ANSI C. Additionally, some PLS routines from the PLS Toolbox for Matlab™ were integrated, using the MATCOM compiler (version 2). The programs were compiled for the DOS/windows operating system using DJGPP, version 2.01. The configuration of the simulated annealing for the different datasets are shown in Table 5-1. A detailed description of the configuration can be found in a previous paper.[5]

Table 5-1 Parameters for the different simulated annealing runs used in this chapter

Simulated annealing parameter [5]	Application Component concentrations in ternary mixture	Density of heavy oil products
Number of spectral variables to select from (*N*)	200	400
Disturbance generation	N(0,5)	N(0,5)
Initial control parameter (c_1)	0.05	0.01
Cooling schedule	geometric with $\alpha = 0.90$	geometric with $\alpha = 0.85$
Length of Markov Chain	1,000	1,000
Exit Markov Chain	minimum number of accepted transitions (250) or maximum number of transitions tested (1,000)	
Exit Simulated Annealing	minimum control parameter $c = 1*10^{-6}$ or minimum acceptance ratio $\chi = 0$	
Acceptance criterion	Metropolis	Metropolis

Randomization t-*test*

The randomization *t*-test was performed in order to compare the test set prediction results of global models and models based on robust variable selection. To this end, for e.g. the ternary mixtures, the component concentrations were predicted from spectra measured at various temperatures by the two models to be compared (global and robust variable selection model). Subsequently, two vectors were constructed from these predictions: one containing the global model predictions for one component at various temperatures and one containing the variable selection model predictions for the same component at various temperatures (vector containing number of temperatures times *N* elements, where *N* is the number of samples in test set). These vectors, along with the known reference values, were used for the randomization *t*-test.

Results and discussion

Dataset A: Ternary mixture of ethanol, water, and iso-propanol

Temperature influence on vibrational spectra

A NIR spectrum consists of overtones and combination bands (resulting from the interaction between two or more different vibrations of neighboring bonds). These absorption bands provide information about features such as: chemical nature (e.g. bond types and functional groups) and molecular conformation (e.g. *gauche* and *trans* conformations). It provides information about the individual molecular bonds and information about the interaction between different types of molecules (intermolecular bonds). Since the molecular vibrations are influenced by these intermolecular interactions, absorption bands in mixtures change in relation to pure analytes. Usually, the intermolecular interaction such as hydrogen bonding is very weak and can be broken by increasing the temperature. Consequently, the vibrational spectrum will change due to these temperature changes. In Figure 5-2, the temperature effect on the pure water spectrum is shown; an increase in the temperature results in an intensity increase, peak shift towards lower wavelengths, and band narrowing. As mentioned in ref. 4, an increase in temperature results in a decrease of the amount of hydroxyl groups involved in a hydrogen bonding and, consequently, the absorption band of “free” hydroxyl increases. Also, the second overtone absorption band of the hydroxyl group in ethanol and iso-propanol (~970 nm) increases as the sample temperature increases. On the other hand, in both the ethanol and iso-propanol spectrum the third overtone C-H stretch vibration (~910 nm) of the CH_3 group and the C-H stretch vibration (third overtone at ~920-930 nm) of the CH_2 group in ethanol change slightly due to temperature changes. Some increase in the C-H combination band of the CH_3 group (~1020 nm) in ethanol and iso-propanol is observed in the spectra when the temperature is increased.

Determination of ethanol content

The test set prediction results of various PLS models for determination of the ethanol content are shown in Table 5-2.

Table 5-2 Prediction results for determination of ethanol content

Model type	# vars	#PLS factors	RMSEP					
			30°C	40°C	50°C	60°C	70°C	mean
local	200	4	0.018	0.011	0.017	0.010	0.011	0.014
Local (50°C)	200	4	0.063	0.028	0.017	0.043	0.079	0.051
Global	200	7	0.014	0.012	0.037	0.016	0.014	0.021
var. sel. [a]	30	4	0.007	0.011	0.023	0.015	0.009	0.014

[a] From ten SA runs the best model (smallest overall prediction error in standardization sets) is selected.

In the first row of Table 5-2, the test set prediction errors of the individual local models at each temperature are shown. These values are taken from ref. 3. Similar test set prediction results are observed in the models. Subsequently, the local model based on spectra measured at 50°C is used to make predictions from spectra measured at temperatures deviating from 50°C (second row Table 5-2). The prediction error increases if the predictions are performed with samples measured at temperatures deviating from 50°C. Thus the sample temperature influences the NIR spectra and, consequently, the model's predictions.

In order to make the model insensitive to temperature variations, a global model and robust variable selection models are constructed. For both the global model and the variable selection model the prediction errors of the test set samples measured at different temperatures are shown (third and fourth row Table 5-2). As ten simulated annealing runs were performed, ten robust variable selection models were obtained. From these models, the model that possesses the smallest overall prediction error in the standardization sets is selected. The test set prediction errors (RMSEP values) obtained using the global model at different temperatures are compared to the predictions obtained using the variable selection model at those temperatures. The overall prediction error obtained using the variable

selection model (overall RMSEP is 0.014) turns out to be significantly smaller than the overall prediction error obtained using the global model (overall RMSEP is 0.021) according to the randomization *t*-test (1999 trials and $\alpha = 0.05$).[13]

Besides having a smaller test set prediction error, the robust variable selection model is based on a smaller number of variables (30 instead of 200) and uses four PLS factors instead of seven as for the global models. It is difficult to say whether the SA model is really more parsimonious, because the variable selection part of the SA model takes away degrees of freedom.[8] However, on the basis of the prediction error, the robust variable selection model may be preferred.

Determination of water content

Table 5-3 shows the prediction results for the models used to predict the water content of the NIR spectra.

Table 5-3 Prediction results for determination of water content

Model type	# vars	# PLS factors	RMSEP					
			30°C	40°C	50°C	60°C	70°C	mean
Local	200	4	0.009	0.007	0.011	0.004	0.004	0.008
Local (50°C)	200	4	0.053	0.023	0.011	0.014	0.028	0.030
Global	200	7	0.015	0.007	0.009	0.008	0.005	0.009
Var. sel. [a]	30	4	0.009	0.004	0.011	0.008	0.009	0.009

[a] From ten SA runs the best model (smallest overall prediction error in standardization sets) is selected.

In ref. 4 separate models are calculated using calibration samples measured at various temperatures (calibration models at 30, 40, 50, 60, or 70°C). The test set prediction errors of these models are shown in Table 5-3 (first row). Similar prediction results are obtained for the models. Subsequently, the local 50°C model is used to make predictions of the water content using test set spectra measured at temperatures other than 50°C. The prediction error in the test set measured at temperatures other than 50°C increases compared to the prediction error obtained at 50°C.

In order to make the calibration model insensitive to temperature changes, global and robust variable selection models were constructed. The test set prediction results are shown in Table 5-3. If the test set predictions of the best variable selection model (overall RMSEP is 0.009) and the global model (overall RMSEP is 0.009) at various temperatures are compared, no significant difference is observed between the models according to the randomization *t*-test (1999 trials and $\alpha = 0.05$). The models are therefore comparable with respect to their predictive power.

Determination of iso-propanol content

Table 5-4 shows the prediction results of the iso-propanol content for the local, global and robust variable selection models.

Table 5-4 Prediction results for determination of iso-propanol content

Model type	# vars	# PLS factors	RMSEP					
			30°C	40°C	50°C	60°C	70°C	mean
Local	200	4	0.012	0.009	0.022	0.008	0.015	0.014
Local (50°C)	200	4	0.055	0.028	0.022	0.048	0.088	0.054
Global	200	7	0.011	0.016	0.042	0.017	0.015	0.023
Var. sel. [a]	10	4	0.009	0.020	0.035	0.014	0.010	0.020

[a] From ten SA runs the best model (smallest overall prediction error in standardization sets) is selected.

In the first row of Table 5-4, the test set prediction results of the local models (test samples and calibration samples measured at the same temperature for each model) are shown. Similar test set prediction errors of the various models are obtained. Only the measurements at 50°C show a systematically higher prediction error, even when predicted from a model constructed at the same temperature. This is most probably due to minor instrumental difficulties with the temperature control during the measurement at 50°C. The spectra do not show a visible deterioration and it was wrongly assumed that it would most probably not affect the quality of models.

The calibration model based on calibration spectra measured at 50°C is used to predict the iso-propanol content of samples measured at temperatures other than 50°C (second row); the prediction error in the test set measured at temperatures other than 50°C is larger than the prediction error at 50°C. Therefore, the model's predictions are sensitive to sample temperature variations.

Subsequently, robust variable selection and global calibration models were calculated in order to develop robust models. If the test set predictions of the best variable selection model (smallest overall RMSEP value) and the global model at various temperatures are compared, the overall prediction error of the robust variable selection model (RMSEP is 0.020) is significantly smaller than the overall prediction error of the global model (RMSEP is 0.023) according to the randomization *t*-test.

Interpretation of variable selection results

Generally, several spectral regions can be distinguished in vibrational spectra of mixtures of chemical compounds with external variation included. These spectral regions can be classified into the following categories:

Regions which only show variation due to variation in the reference parameter (e.g. spectral variation caused by variations in water, ethanol and iso-propanol content in ternary mixtures).

Regions which only show variation caused by an external factor and no variations caused by changes in the parameter of interest (in this study spectral variation caused by sample temperature variations).

Regions which both contain variation due to the parameter of interest and variation caused by external factors.

Regions which do not contain variations of spectral region category one or two (e.g. spectral baseline).

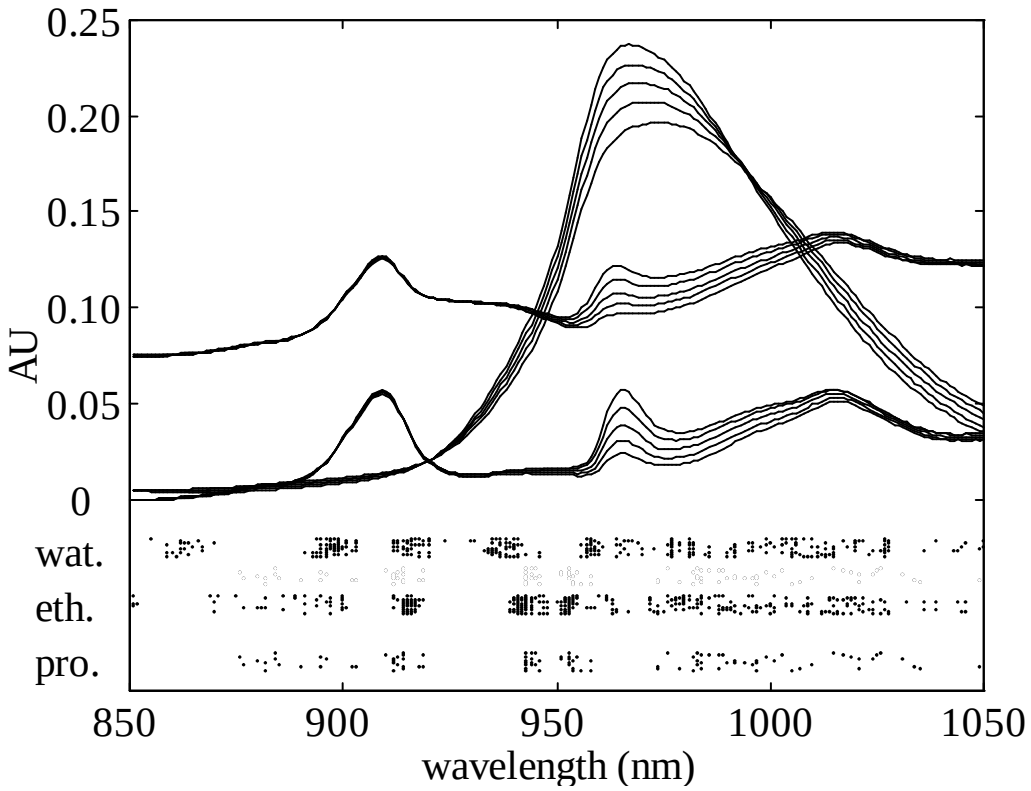


Figure 5-2 Selected variables for the determination of ethanol, water, and iso-propanol content in ternary mixtures.
Lower dots (pro.) = selected variables (10) for determination of iso-propanol (ten SA runs)
Center dots (eth.) = selected variables (30) for determination of ethanol (ten SA runs)
Upper dots (wat.) = selected variables (30) for determination of water (ten SA runs).
Additionally, the NIR spectra of pure ethanol (—), water (–·–) and iso-propanol (––) measured at 30, 40, 50, 60, and 70°C are plotted. A baseline of 0.075 AU was added to the ethanol spectra for visualization purposes

The variable selection/rejection results for the ethanol, water, and iso-propanol models are shown in Figure 5-2. In this figure a selected variable (wavelength) is represented as a dot. For every component in the mixture, ten simulated annealing runs are performed (10 rows each containing *k* dots). As can be seen in Figure 5-2, almost the entire range of the water spectrum belongs to category three, i.e. spectral regions containing information about the water concentration and showing variations caused by sample temperature. Since water is present in all samples (Figure 5-1), almost all spectral regions contain variations caused by sample temperature variations. The variables found by the SA are a combination of the selection of the informative variables for the parameter of interest, the rejection of variables influenced by external factors and the selection of spectral regions which can compensate for selected informative variables possibly affected by external variations. Therefore, interpretation of the variable selection and rejection results is very difficult. However, some selected and rejected spectral regions for the determination of the ethanol, water and iso-propanol concentration can be assigned.

The spectral region between 1029 and 1050 nm is hardly ever selected in any model. This region is very “noisy” compared to the other regions in the NIR spectra. Selection of this region may lead to an increased prediction error. Since variable selection is based on minimization of the prediction errors at different temperatures, this region is sparsely sampled. In both alcohol models, the region around 970 nm, which corresponds to the second overtone absorption band of the hydroxyl group, is rejected (ethanol: 964 to 972 nm; and iso-propanol: 958 to 975 nm). As can be seen in Figure 5-3, the intensity of this absorption band is proportional to the temperature. Therefore, this temperature-sensitive region is rejected from the entire spectral range. A very densely sampled region for both the ethanol and iso-propanol model is observed at ~915 nm, which corresponds to the third overtone C-H stretch vibration of the CH_3 group. This region (911 to 918 nm) shows almost no variation caused by changes in sample temperature and possesses ethanol/iso-propanol concentration information.

Furthermore, this region is located at a peak wing. As peak wings are less sensitive to temperature variations, this region is preferred. Additionally, the spectral regions 939 to 947 nm and 950 to 954 nm are densely sampled for the ethanol model. The former region, which corresponds to the C-H stretch vibration of CH_2 in ethanol, may be selected in order to distinguish between ethanol and iso-propanol. The latter region (950 to 954 nm) may be selected to compensate for the temperature sensitivity of the water hydroxyl band in this region. Similar regions are selected for the iso-propanol model in order to distinguish between alcohols and compensate for temperature influences.

For the water model, some very densely sampled spectral regions can be distinguished: 891 to 902 nm, 911 to 919 nm, 934 to 940 nm, 858 to 866 and 956 to 959 nm. The first two spectral regions (891 to 902 nm and 911 to 919 nm) are probably selected to compensate for the alcohol hydroxyl contribution to the water hydroxyl absorption band as can be seen in Figure 5-2 and the loading plot shown in Figure 5-3.

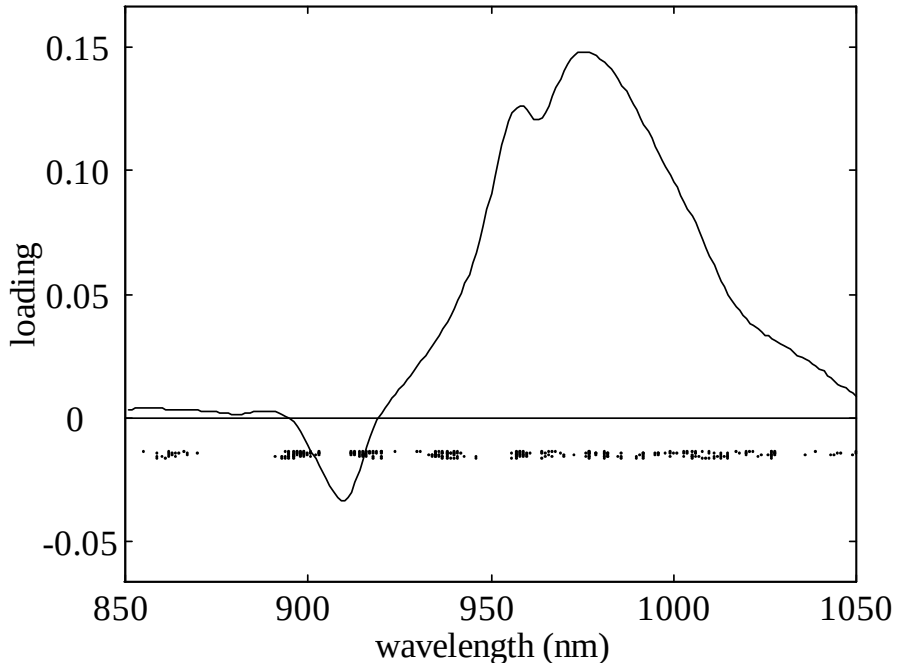


Figure 5-3 Loading plot of first PLS factor in model for water content determination. First factor captures 97% of variance in X and 94% of variance in Y. Furthermore, the selected variables for the determination of water content are shown (same SA runs as shown in Figure 5-2)

The variables are selected in a spectral region which shows negative loading values in the first PLS factor. This region corresponds to the third overtone C-H stretch vibration of the CH_3 group in ethanol and iso-propanol (~915 nm). These spectral regions, especially the second one, are also selected in the ethanol and iso-propanol model and hardly any spectral variation caused by variation in the water content is present in this region. Furthermore, in this region the intensities are not very sensitive to variations in the sample temperature. Therefore, the spectral variables (891 to 902 nm and 911 to 919 nm) located at important ethanol and iso-propanol absorption bands are probably used to compensate for the alcohol hydroxyl absorption band at the water hydroxyl band. Especially, the wings of the

absorption bands are selected because the wings are less sensitive to intensity variation caused by temperature variations.

Other selected regions for determination of the water content are located at the hydroxyl absorption band. These selected regions have a large variation due to water content and sample temperature variations (e.g. 956 to 959 nm). In order to compensate for these temperature variations, some additional regions are selected (e.g. 934 to 940 nm). An explanation for this compensation selection can be found in a paper by Wülfert *et al*.[18] They developed a UVE-PLS model for predicting the water content and a UVE-PLS model for predicting the temperature of the same ternary mixtures as used in the current study (more details about this method can be found in the next section). It was found that the region of 956 to 959 nm was used for both the water content model and temperature model while the region of 934 to 940 nm was used for the temperature model.

In conclusion, both global calibration and robust variable selection models can be used to calculate calibration models that are less sensitive to external variations and more selective for the parameter of interest. For the water model, the techniques are comparable (the prediction results of both techniques are not significantly different). Long-term validation in practice should indicate which technique works better. For the alcohol models, robust variable selection yields significantly better results than the global calibration model. To what extent robust variable selection yields better results, is probably determined by the relative amount of contribution of the above-mentioned categories. For robust variable selection, spectral regions of category one, two, and four are preferred. Especially in the water models, almost the entire spectral range belongs to category three (spectral intensities show variation caused by variation in water content and temperature variations) or category four (no variation in intensities) and the variable selection does not yield better results than the global models. On the other hand, in the ethanol models the prediction results of the robust variable selection model are better than those of the global model. In the

spectra of ethanol, spectral regions of all categories are present and, consequently, the majority of the selected variables belongs to category one. However, there are variables selected from the other categories as well.

Comparison with other robust variable selection technique

Recently, Wülfert *et al*.[18] applied uninformative-variable-elimination by PLS (UVE-PLS) to select robust variables. UVE-PLS, originally developed by Centner *et al.*,[19] eliminates variables from PLS models by judging a criterion based on the regression vector. In UVE-PLS, the variables are eliminated on the basis of the quotient of the regression coefficient and the uncertainty in the calculated regression coefficients (confidence limits are estimated by leave-one-out jackknifing). Variables that give smaller quotients than a certain threshold value are considered to be uninformative. The threshold value is estimated by adding artificial random spectral variables to the original spectral data and calculating the above-mentioned quotients for these random variables. The maximum absolute quotient is taken as the threshold value. In the variable selection method of Wülfert *et al*.,[18] a UVE-PLS model is constructed for predicting the parameter of interest (concentration) and another UVE-PLS model is constructed for predicting the parameter causing the external spectral variation (temperature). The variables that are selected in the model of the parameter of interest and rejected in the external variation model are supposed to be robust. Category 1 variables are selected in the concentration UVE-PLS model and rejected in the temperature UVE-PLS model. Consequently, category 1 variables are selected by the robust UVE-PLS model. As category 2 variables are rejected in the concentration UVE-PLS model and selected in the temperature UVE-PLS model, they are rejected in the robust UVE-PLS model. Category 3 variables can be selected or rejected in the concentration UVE-PLS model and/or the temperature UVE-PLS model dependent on the ratio between spectral variations caused by temperature and concentration in the variables. As a result, only those variables that are both selected in the concentration model and rejected in the temperature

model are selected in the robust UVE-PLS model. Category 4 variables are rejected in the concentration model and rejected in the temperature model. Consequently, these variables are rejected by the robust UVE-PLS model.

In this study, we have calculated models based on the variables and number of PLS factors found in ref. 18. For each component in the ternary mixtures, these variables were selected from the datasets X_{stand}^{30}, X_{stand}^{40}, X_{cal}^{50}, X_{stand}^{60}, and X_{stand}^{70}. Subsequently, from the joint dataset a four factor PLS1 model (determined by cross-validation on selected variables) was calculated for each component and the datasets X_{test}^{30}, X_{test}^{40}, X_{test}^{50}, X_{test}^{60}, and X_{test}^{70} were used as independent test sets. The prediction results are shown in Table 5-5.

Table 5-5 Prediction results for determination of component concentration in ternary mixtures using models based on variable subset selection

Model type	# vars	# PLS factors	RMSEP 30°C	40°C	50°C	60°C	70°C	Mean
***Ethanol*:**								
SA [a]	30	4	0.007	0.011	0.023	0.015	0.009	0.014
UVE-PLS [b]	44	4	0.013	0.010	0.028	0.024	0.035	0.024
***Water*:**								
SA [a]	30	4	0.009	0.004	0.011	0.008	0.009	0.009
UVE-PLS [b]	32	4	0.022	0.011	0.009	0.007	0.010	0.013
***iso-propanol*:**								
SA [a]	10	4	0.009	0.020	0.035	0.014	0.010	0.020
UVE-PLS [b]	45	4	0.020	0.016	0.032	0.026	0.039	0.028

[a] From ten SA runs the best model (smallest overall prediction error in standardization sets) is selected.

[b] PLS1 model based on spectral subset from the datasets: X_{stand}^{30}, X_{stand}^{40}, X_{cal}^{50}, X_{stand}^{60}, and X_{stand}^{70}

Using the randomization *t*-test (1999 trials and $\alpha = 0.05$), the independent test set prediction results of these UVE-PLS models are compared to the calibration models based on a spectral subsets found by simulated annealing. For predicting the water content, the SA variable selection model performs significantly better than the UVE-PLS model with respect to the

prediction error at different temperatures. For predicting the ethanol content, the model based on variables selected by simulated annealing is significantly better than the UVE-PLS based selection. Finally, the SA variable selection model for prediction of iso-propanol content has a significantly smaller overall prediction error than the UVE-PLS models at various temperatures.

A disadvantage of the UVE-PLS based method is that PLS (or UVE-PLS) must be capable of modeling external variations. Frequently, external factors cause complex effects on the spectra, which may be difficult to model by PLS. Furthermore, robust variable selection based on UVE-PLS selects those variables which are kept in the parameter of interest model but rejected in the external variation model (category 1 regions) as well. As a result, problems may arise from the fact that some spectra only contain regions belonging to spectral region category 3 (both spectral variation caused by variations in parameter of interest and external variations). In such case it may be possible that no robust variable is maintained in the final robust model. On the other hand, robust variable selection as described in ref. 5 and this chapter can select regions of category 3 and compensate the external effects in these regions by selecting other regions of category 3. Another disadvantage of robust variable selection based on UVE-PLS is due to the fact that the temperature is modeled. As a consequence, the temperature of the calibration and standardization samples need to be known with a high degree of accuracy.

Dataset B: Density of heavy oil products

In Figure 5-4, the mean of the calibration set spectra of heavy oil products measured at 100°C is shown. The major components in crude oil are hydrocarbons including aromatics, paraffins and naphtenes. The bands at ~4350, ~4260 and ~4065 cm^{-1} are CH_2 and CH_3 combination bands and the spectral region between 4550 and 4650 cm^{-1} is assigned to the vibration of the aromatic C-H bonds.

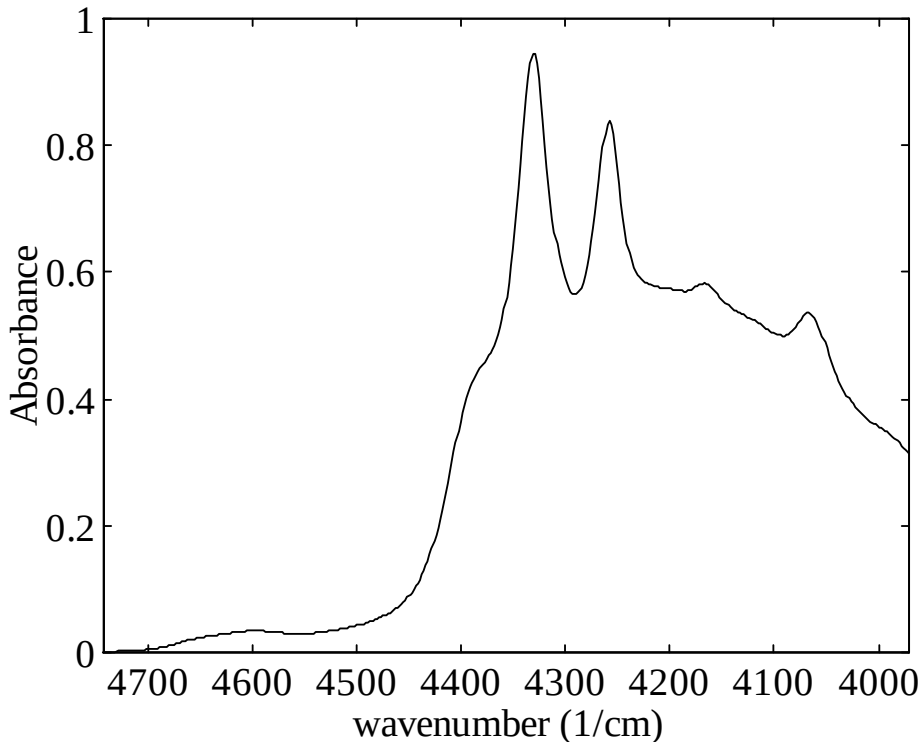


Figure 5-4 Mean spectrum of heavy oil calibration samples measured at 100°C

The prediction results of the local model, the global model, and the robust variable selection model for the density determination of heavy oil products from their NIR spectrum are shown in Table 5-6.

Table 5-6 Prediction results for density determination of heavy oil products

Model type	# vars	# PLS factors	RMSEP			
			95°C	100°C	105°C	mean
local (100°C)	400	5	0.0101	0.0027	0.0067	0.0072
Global	400	6	0.0035	0.0033	0.0032	0.0033
var. sel. [a]	25	5	0.0030	0.0026	0.0022	0.0026

[a] From ten SA runs the best model (smallest overall prediction error in standardization sets) is selected.

From the calibration spectra measured at 100°C and corresponding densities, a local calibration model is calculated. This model is used to predict the density of the test set samples from the corresponding spectra measured at 95°C, 100°C, and 105°C. The prediction error in the test set samples measured at 95°C and 105°C increases, compared to the prediction error of these samples measured at 100°C.

In order to make the models predictions insensitive to temperature variations, a global calibration model and models based on robust variables are constructed. The prediction results of these models are shown in Table 5-6. Using the randomization *t*-test (1999 trials and $\alpha = 0.05$), the best robust variable selection model is compared to the global calibration model with respect to their prediction errors at various temperatures. The robust variable selection model gives significantly better overall prediction results (RMSEP is 0.0026) than the prediction results of the global model (RMSEP is 0.0033).

Interpretation of variable selection results

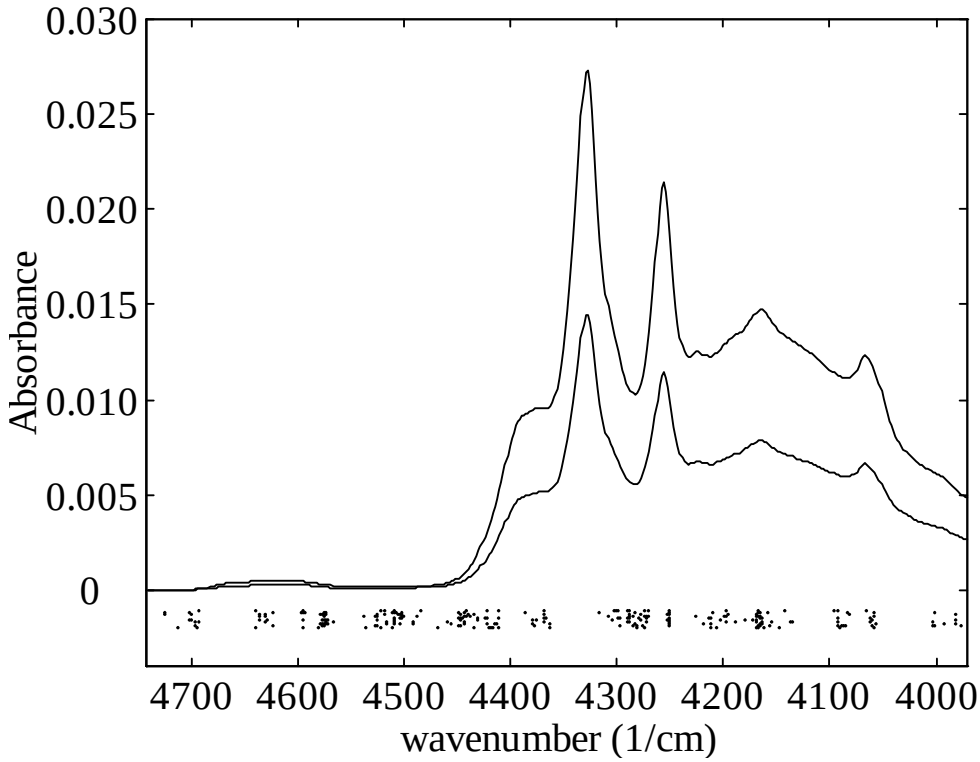


Figure 5-5 Selected variables for the density determination of heavy oil products. Plotted spectra are "difference spectra" between the mean test set spectra measured at 95°C and 105°C (—) and the mean test set spectra measured at 95°C and 100°C (—). Dots represent the selected variables of ten SA runs at k = 25

In Figure 5-5 the variable selection results of the simulated annealing algorithm are presented (ten random initialized simulated annealing runs). The variables selected by the simulated annealing algorithm are represented as dots (10 rows containing 25 dots). Furthermore, the difference spectra between the mean test set spectra measured at 95°C and the other temperatures are plotted. It can be seen from this figure that the spectral differences caused by temperature variations are mainly

intensity variations. As the heavy oil products are complex mixtures containing many types of hydrocarbons, it is very difficult to interpret the variable selection results. However, some selected and rejected regions can be assigned. A large region which is rejected by robust variable selection is 4362 to 4318 cm^{-1}. If the temperature increases, the absorbance in this region decreases and the absorbance peak shifts slightly to higher wavenumbers. It is known that a peak shift can result in erroneous model predictions, and therefore this region is rejected from the whole spectral range. In the other regions, the main difference in absorbance between the different sample temperatures shows a multiplicative effect (Figure 5-4 and Figure 5-5). Consequently, important peaks for density determination are selected (e.g. 4252 cm^{-1}). Other regions are selected to correct for this selection (e.g. 4375 to 4365 cm^{-1}, 4308 to 4272 cm^{-1}, or 4171 to 4156 cm^{-1}). Another densely sampled region is the one between 4452 and 4411 cm^{-1}. Probably, this region on the wing of an absorbance peak is selected because the wings of a peak are less sensitive to changes in peak intensities than those at the top of an absorbance peak. This is also observed in the above-mentioned alcohol models.

Comparison with other robust variable selection technique

In order to compare robust variable selection by simulated annealing with variable selection based on UVE-PLS, the models presented in ref. 18 were used. A six factor PLS1 model is calculated using the variables selected by UVE-PLS and the datasets X_{stand}^{95}, X_{cal}^{100}, and X_{stand}^{105}. Subsequently, this model is used for predicting the density of the independent datasets X_{test}^{95}, X_{test}^{100}, and X_{test}^{105}. The prediction results are shown in Table 5-7.

Table 5-7 Prediction results for density determination using models based on variable subset selection

Model type	# vars	# PLS factors	RMSEP			
			95°C	100°C	105°C	mean
SA [a]	25	5	0.0030	0.0026	0.0022	0.0026
UVE-PLS [b]	157	6	0.0066	0.0042	0.0044	0.0052

[a] From SA runs the best model (smallest overall prediction error in standardization sets) is selected.

[b] PLS1 model based on spectral subset from the datasets: X_{stand}^{95}, X_{cal}^{100}, and X_{stand}^{105}.

Using the randomization *t*-test, the UVE-PLS based model is compared with the model based on a variable subset found by simulated annealing. The SA variable selection model possesses a significantly smaller overall prediction error (RMSEP is 0.0026) than the UVE-PLS model (RMSEP is 0.0052) at the different temperatures.

Conclusions

In this chapter, robust variable selection models are compared to global calibration models for different applications in order to decrease the influence of temperature variations on the model's predictions. It is shown that models based on robust variable selection are sometimes better than or similar to global calibration models with respect to prediction errors at different sample temperatures. However, a disadvantage related to the simulated annealing approach used for variable selection is that special expertise and software are needed.

It is shown that robust variable selection models are less complex, because they are based on a smaller number of variables and use fewer PLS factors than global calibration models. However, it is difficult to say whether the robust variable selection models are more parsimonious, because many degrees of freedom are lost during variable selection. Therefore, long-term validation in practice is necessary to indicate which method works best.

References

[1] O.E. de Noord, Chemom. Intell. Lab. Syst., **25** (1994) 85.

[2] H. Swierenga, W.G. Haanstra, A.P. de Weijer and L.M.C. Buydens, Appl. Spectrosc., **52** (1998) 7.

[3] F.Lindgren, P. Geladi, S. Ränner and S. Wold, J. of Chemom., **8** (1994) 349.

[4] F. Wülfert, W. Th. Kok and A.K. Smilde, Anal. Chem., **70** (1998) 1761.

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

[6] D. Ozdemir, M. Mosley and R. Williams, Appl. Spectrosc., **52** (1998) 599.

[7] T. Næs and T. Isaksson, NIR News, **5** (1994) 4.

[8] O.E. de Noord, Chemom. Intell. Lab. Syst., **23** (1994) 65.

[9] D. Jouan-Rimbaud, D.L. Massart and O.E. de Noord, Chemom. Intell. Lab. Syst., **35** (1996) 213.

[10] R. Leardi and A. Lupiáñez González, *Chemom. Intell. Lab. Syst*., **41** (1998) 195.

[11] E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines*. (Wiley, Chichester, 1989) pp. 13-31.

[12] H. Swierenga, A.P. de Weijer, R.J. van Wijk and L.M.C. Buydens, Chemom. Intell. Lab. Syst., **49** (1999) 1.

[13] H. van der Voet, Chemom. Intell. Lab. Syst., **25** (1994) 313.

[14] E. Bouveresse, C. Hartmann, D.L. Massart, I.R. Last and K.A. Prebble, Anal. Chem., **68** (1996) 982.

[15] R.W. Kennard and L.A. Stone, Technometrics, **11** (1969) 137.

[16] Matlab, version 4.2, The MathWorks Inc., Matick, USA.

[17] PLS Toolbox for Use with Matlab, version 1.5, Eigenvector Technologies,West Richland, USA.

[18] F. Wülfert, W. Th. Kok, O.E. de Noord and A.K. Smilde, . Chemom. Intell. Lab. Syst. **51** (2000), 189.

[19] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste and C. Sterna, Anal. Chem., **68** (1996) 3851.

6. SUMMARY AND GENERAL CONCLUSIONS

The influence of temperature on spectra and consequently on multivariate calibration models has been studied. This is intended to serve as an example for the more general problem of influence of external factors in the practical or industrial application of spectroscopic measurements for multivariate process analysis, monitoring and control.

Different methods have been applied for two data sets exhibiting temperature influences that reduce the predictive quality of calibration models:

Table 6-1: Models and strategies used

Category	Model Type	Chapter	Temp. known for	
			Calibration	Prediction
Implicit inclusion	Global	2	✗	✗
Explicit inclusion in calibration model	Local + interpolation	2	✓	✓
	Incl. in X	3	✓	✓
	Incl. in Y	3	✓	✗
Linear Data pre-processing	2-step PLS	3	✓	✗
	Basis projection	3	✗	✗
Non-linear Data pre-processing	Var. selection PLS-UVE	3	✓	✗
	CPDS	4	✓	✓
	Var. selection SA	5	✗	✗

✓: Knowledge of temperature is required in order to be able to use the model.

✗: Temperature is not required but in case of calibration it should be possible to assume, by e.g. the size of the dataset, that the temperature variation is well spread in order to be excluded as a confounding factor.

Table 6-1 summarizes the characteristics of the applied methods and models in terms of strategies used and necessity to know the interfering temperature. The strategies to handle the temperature influences can be grouped into 4 different categories:

1. Global models: These models can include the temperature implicitly in the calibration model through experimental design. They are simple to set up and performed remarkably well. However, global models become more complex in terms of latent variables and are therefore more prone to instabilities. Another drawback lies in the assumption that the temperature is fully covered and balanced in the calibration design. Already a slight deviation from this assumption can provoke large prediction errors, since the regression model will confound the temperature effect with the analyte to be predicted. This issue did not prove to be a major issue in the present work because only well designed calibration sets were used, but this cannot be guaranteed in the daily practice of e.g. industrial applications.

2. The explicit inclusion of temperature in the calibration model itself forms the second category and has been tried in three different ways: The correction on values predicted by local models (models at one temperature), the inclusion of the temperature in the predicting X- and the inclusion in the predicted Y-block. The correction of local model does not work due to the non-linearity of the temperature effect. A response shift is fundamentally non-linear and additionally the effect of the temperature on the spectra does depend on the analyte-concentrations too. Therefore, it is impossible to correct for temperature influences by a simple correction factor after prediction. The inclusion into the X-block does intend to use the temperature as a corrective input and therefore to aid in the prediction of the analyte. But, - by the experimental design - , there is no correlation between the analyte concentration and temperature what leads to only a small influence of the temperature in the model. Additionally, the correlation between the temperature and its

effects on the spectra is, as explained earlier, very complex. The PLS-regression models are therefore not able to use the temperature information when using it as predictor-variable. The third approach in this category, using the temperature as a y-variable and predicting it simultaneously with the analyte, is more logical since, - like the analyte - , the temperature also generates the spectrum. The inclusion in y should therefore enable the PLS model to use the most temperature-affected spectral regions for temperature prediction and predict with the less affected regions the analyte. In most cases the performance is therefore comparable to the global models. Only in the case where the temperature interference overlaps strongly with the spectral region of the analyte (i.e. water) this strategy demonstrates an inferior performance. This is probably due to the model using up large part of the variance for the temperature prediction, leaving too little to predict the analyte as well.

3. The third category uses linear data preprocessing to correct for temperature interference. One approach is to lead further the last mentioned idea of correction through temperature prediction (inclusion in y) and build a separate temperature regression model prior to building the analyte regression model on the remaining variance. Even more extremely than in the case of water prediction, the temperature model now uses up most of the variance. Consequently, the prediction of the analyte can be only very poor. Another linear preprocessing strategy is to express the data on a new basis on which the analyte signal and temperature interference could be separated. The growth of wavelet transform applications led to a closer inspection of the possibility to apply it to the “temperature problem”. However, this inspection led to the conclusion that, since the wavelet transform is a linear operation which transforms from one orthogonal to another orthogonal basis, the non-linearity cannot be separated from the analyte signal in a better way than in the original domain.

4. The fourth and final category studied in the presented work is the correction for temperature with non-linear data preprocessing methods. Three methods, namely a variable elimination method based on predictive stability (PLS-UVE), a simulated annealing based variable selection method and a continuous spectra-correction method (CPDS) were chosen as exponents of this category. The non-linear strategies worked well, compensating for the largest part the temperature effects on the calibration accuracy and leading to less complex calibration models. However it should be noted that, while the calibration models become less complex in terms of the number of latent variables used, the complexity of the method does not decrease as it incorporates either a non-linear correction or variable selection. The methods differ on several points and have their own strengths and weaknesses. The PLS-UVE method makes only use of fast linear algorithms (although a variable selection or elimination is non-linear by itself) but needs the temperature to be known during calibration. The simulated annealing based variable selection does not need the temperature to be measured for calibration and prediction. On the other hand, the algorithm is a probabilistic approach, leading to different solutions when repeated. The CPDS method is very straightforward and leads to a very interpretable correction result but will only work when the effect to be corrected for is continuous.

The strategies presented in this work show, that multivariate calibration models can be made robust enough to handle the external non-linear interferences that can be expected when applying the models in an industrial environment. In the end it will depend on the problem at hand which technique will be the most suitable. A global model may be used for its ease of implementation. In the case of a strongly non-linear interference in combination with a need for higher prediction accuracy one of the non-linear strategies may be more appropriate. In any case, the large differences in predictive quality imply that the choice of the calibration

algorithm becomes secondary to the successful correction for interferences and good calibration design.

Samenvatting en Algemene Conclusies

De invloed van de temperatuur op spectra en daarmee op multivariate kalibratie modellen is onderzocht. Dit kan als voorbeeld gezien worden voor het algemenere probleem van invloed van externe factoren in de praktijkgerichte of industriële toepassing van spectroscopische meetmethoden in de multivariate procesanalyse en monitoring.

Verschillende methoden zijn toegepast op 2 datasets met temperatuurinvloeden die de voorspellingskwaliteit van de kalibratiemodellen aantasten.

Tabel 6-2: Gebruikte modellen en strategieën

Categorie	Model Type	Hoofdstuk	Temp. bekend voor: Kalibratie	Temp. bekend voor: Voor-spelling
Impliciete opname	Globaal	2	✘	✘
Expliciete opname in kalibratie model	Lokaal + interpolatie	2	✓	✓
	Incl. in X	3	✓	✓
	Incl. in Y	3	✓	✘
Lineaire Data voorbewerking	2-staps PLS	3	✓	✘
	Basis projectie	3	✘	✘
Niet-lineaire Data voorbewerking	Var. selection PLS-UVE	3	✓	✘
	CPDS	4	✓	✓
	Var. selection SA	5	✘	✘

✓: Temperatuur moet bekend zijn om model te kunnen gebruiken.

✘: Temperatuur hoeft niet bekend te zijn maar in geval van kalibratie moet ervan uit gegaan kunnen worden (b.v. door de grootte van de dataset) dat de temperatuursvariatie goed is opgespannen en dat deze niet verwisselt kan worden met de analiet variatie

Tabel 6-2 vat de gebruikte methoden en modellen samen in termen van gebruikte strategieën en noodzakelijkheid van kennis van de interfererende temperatuur. De strategieën kunnen in 4 verschillende categorieën worden ingedeeld:

1. Globale modellen: Deze modellen kunnen de temperatuur incorporeren door middel van het experimental design. Zij zijn makkelijk te implementeren en hebben redelijk goed gepresteerd. Deze prestatie gaat ten koste van een grotere complexiteit van het kalibratie model (in termen van aantallen latente variabelen) en zal daardoor eerder instabiel kunnen worden. De eis dat de temperatuur door middel van een goede proefopzet (design) is opgespannen leidt tot het gevaar dat ook een kleine afwijking ervan tot grote fouten kan leiden. In het gepresenteerde onderzoek was dit geen probleem omdat een volledig design is gebruikt, in de praktijk kan dit echter lang niet altijd gegarandeerd worden.

2. De expliciete opname van de temperatuur in het kalibratie model vormt de tweede categorie en is op drie manieren uitgevoerd: Een temperatuurscorrectie achteraf op de voorspelling van kalibratiemodellen bij maar één temperatuur, de opname van de temperatuur als voorspellende variabele (in het X-block) of de opname in y als te voorspellen variabele. De niet-lineariteit van de temperatuursinvloeden op de spectra maken een goede correctie achteraf onmogelijk. De door de temperatuur geïnduceerde signaalverschuiving is een niet-lineair verschijnsel dat bovendien afhankelijk is van de analietconcentratie. Daardoor is het onmogelijk om achteraf met een simpele correctiefactor voor een correcte voorspelling te zorgen. De opname van de temperatuur in het X-block moet zorgen voor een correctieve variabele in het kalibratie model. Door de proefopzet is er echter geen correlatie tussen de analietconcentratie en de temperatuur, waardoor het gewicht van de temperatuur in het model laag zal zijn. Bovendien is het verband tussen de temperatuur en de

spectrale invloeden ervan uiterst complex zoals al eerder aangegeven. Daardoor is een PLS algoritme uiteindelijk niet in staat om de temperatuur als correctieve X-variabele te benutten. De derde manier, het opnemen van de temperatuur in het te voorspellen Y block is logischer. De temperatuur ligt immers – net als de analiet – ten grondslag aan het spectrum. Een PLS model kan daardoor het spectrale gedeelte dat het meest wordt beïnvloed door de temperatuur ook voor de temperatuursvoorspelling gebruiken, en wordt daardoor in staat gesteld om de analietconcentratie juist met de andere gebieden te voorspellen. Alleen in gevallen waarin de temperatuur een sterke invloed op het analiet spectrum heeft zal deze grote overlap ertoe zorgen dat er te weinig variantie voor een deugdelijke analiet-voorspelling is, zoals dat met de water modellen ook inderdaad gebeurde.

3. De derde categorie gebruikt lineaire datavoorbewerking om voor temperatuurseffecten te corrigeren. Één mogelijkheid hiervoor is het verder doorvoeren van laatstgenoemde correctie door temperatuursvoorspelling (temperatuur in y block). Nu wordt echter vooraf een apart model gebruikt om de temperatuur te voorspellen en daarna met alleen de resterende variantie voor de analiet gekalibreerd. Maar hier geldt in nog veel sterkere mate dan bij de bovengenoemde moeilijkheden bij de water voorspelling dat er bij lange na niet genoeg variantie overblijft voor een goede analiet voorspelling. Een andere mogelijkheid tot lineaire datavoorbewerking is het uitdrukken van de data op een andere basis om het analiet signaal beter van de temperatuur interferentie te scheiden. De groeiende toepassingsgebieden van wavelettransform leidden tot een verdere evaluatie van de mogelijkheden om een wavelettransform op de temperatuur-beïnvloede spectra uit te voeren. Maar omdat een wavelet transform alleen een lineaire operatie is die van de ene naar de andere ortoghonale basis projecteert, is het onmogelijk om op de waveletbasis

een betere scheiding tussen signaal en temperatuurinterferentie te verkrijgen.

4. De vierde en laatste categorie die binnen dit onderzoek aan de orde is gekomen is de temperatuurscorrectie door middel van niet-lineaire datavoorbewerking. Drie methoden, een op stabiliteit gebaseerde variabelen eliminatie methode (UVE-PLS), een op simulated annealing gebaseerde variabelen selectie en een continue spectrale correctie methode zijn als exponenten van deze categorie gekozen. Deze niet-lineaire preprocessing methoden zijn in staat om voor een groot gedeelte de temperatuur invloeden te compenseren en tot simpelere kalibratie modellen te leiden. Er moet echter opgemerkt worden dat het data analyse model in zijn totaliteit natuurlijk niet simpeler wordt, omdat de preprocessing stappen op zich ook modelcomplexiteit en daarmee een verlies aan vrijheidsgraden betekenen. De voorgestelde methoden verschillen op een aantal punten en hebben elk zowel voor- als nadelen. De PLS-UVE methode maakt alleen gebruik van snelle lineaire algoritmen (ook al is een variabele selectie methode zelf per definitie niet lineair) maar heeft de meting van de temperatuur voor de kalibratie nodig. De op simulated annealing gebaseerde variabelenselectie heeft de temperatuur noch gedurende de kalibratie fase noch tijdens de voorspellingsfase nodig. Aan de andere kant is dit algoritme een probabilistische methode en geeft deze bij herhaling verschillende oplossingen. De CPDS methode daarentegen is rechttoe rechtaan en geeft goed interpreteerbare correctieresultaten maar zal alleen functioneren als het te corrigeren effect continu is.

De in dit werk gepresenteerde strategieën geven weer dat multivariate modellen robuust gemaakt kunnen worden voor externe interferenties die te verwachten zijn wanneer de modellen in een industriële omgeving toegepast worden. Uiteindelijk zal het van het voorhanden zijnde probleem afhankelijk zijn welke techniek het meest geschikt is. Een globaal model kan gekozen worden vanwege zijn toepassingsgemak. Als de interferentie sterk

niet-lineair is in combinatie met een noodzaak voor hoge voorspellingsprecisie kan ook een van de twee niet-lineaire strategieën geschikter zijn. Duidelijk wordt hoe dan ook geïllustreerd dat ondanks dat er iedere keer hetzelfde kalibratie-algoritme gebruikt wordt, de uiteindelijke voorspellingskwaliteit vooral van een succesvolle correctie voor de temperatuurs interferenties afhangt.

ACKNOWLEDGMENTS

This is the place and finally the time to thank a lot of people. Safest would be to just give a general thanks to everybody avoiding the danger of forgetting someone:

THANX EVERYBODY!

But this wouldn't do justice to the pleasant and interesting time I had at the Laboratory for Analytical Chemistry, so I will try to be more specific.

First, I want to thank the 'gang in room 4.15', Sabina, Frans and Renger. The coexistence we had was extremely varied and versatile: the jokes, matches of 'office-football', conversations and discussions over a wide range of subjects (even serious ones) and the open companionship surely are big part of the positive memories I have of that time. Another special group of people are Raivo, Ad, Jaap (K) and Hans (Boelens, there were a lot of Hans'). During my last year as a Masters and later as a Ph.D. student I could always count on their help, advice and friendship.

As I already started with one Hans, I will complete the list: Hans Kragten introduced me to the beauty of statistics (you are not with us anymore, that's a great loss but you certainly will never be forgotten). Hans Poppe was always a source of inspiring discussions and Hans van der Moolen a source of humor (especially with our "Who is the biggest rat?"-competition).

A very special acknowledgement should go to Her: he introduced me during my last year as a Masters student to chemometrics and the beauty of models. The knowledge and insight I learned under his guidance are still helping me to the present days.

The necessary measurements for this work would never have been successful without the great help and assistance of the real lab experts: Gjalt, Rob, Wim (O), Jaap (E), Sytske and Dini. But my recollection of them

is certainly not limited to the analytical field only. I remember discussions about the major challenges in life, society and tropical frogs with Rob and Gjalt. Sometimes we would reach philosophical heights, - or at least it seemed so after a few beers in Kraak's Booze Corner.

Which leeds me to Johan: a warm fatherly figure to the whole department (as much as an acoustic phenomenon). Every Friday afternoon we would sit in front of his office for the traditional 'borrel', maybe a pizza and Grand Marnier afterwards. Johan: you were always available for a talk on whatever subject, you made everyone feel welcome, you played a crucial role in creating an atmosphere of friendship, which made the lab a special place.

Furthermore, I want to thank all the people not yet mentioned that I had the pleasure to meet at the lab and outside: Alejandro, Anna, Arian, Edward (always in a good mood, always helping), Edwin (a true philosopher), Ellen (trying to keep some organization in this lab sure wasn't easy), Gerard, Gerrit, Huub (thanks for giving a sound mathematical basis concerning the (im)possibility of finding a general linear transform able to filter out temperature effects), Johan (W), Jonas (introducing pizza at lunch), Judith (there is never too much pink), Koen (for his work on the wavelet packet transform of the simulated data set), Rasmus (more pizza at lunch accompanied by a beer and a cigarette), Remco (St), Remco (Sw), Ricardo, Roelant (a true brother to me), Sandra and Sylvia. I also I want to thank you, Erik, for our cooperation and most inspiring discussions. It was a pleasure to work with you and you are the proof that one can work at different groups, on similar problems, but still inspire each other instead of simply compete.

The saying "last but not least" surely holds true for the last two: my (co)-promoters Wim and Age. A discussion with Wim is always challenging, his advice is to the point, his humor is witty and sharp, which makes working with him an interesting and very enjoyable experience. Needless to say that he too was one of the regulars at the Friday-'borrel'.

So, my final acknowledgement goes to Age: your door was always open and your interest genuine. Your constant demand for visualizations and schematics opened for me a new way of looking at (data)-analytical problems (and greatly improved my drawing skills). Thanks also for the confidence you had in the finishing of this work. More than once I thought that if you still believe in it, then who am I to doubt it. Either you are very patient or very good in psychology.

One last thing now that I (hopefully) acknowledged everybody who was directly involved and helped me with my work and study (and more). I want to thank my parents to whom I dedicate this book. You have been a source of enormous and unconditional support during all my life and surely I haven't been the easiest one (to put it mildly). Your love, inspiration, encouragement and confidence in good times and especially also in bad times makes you the kind of parents one can only wish to have. I'm lucky to have you and hope that I will continue having you for many, many years. Millions of thanks to you, folks.

Pai, Mãe: um enorme abraço para vocês. ***Obrigado****.*

APPENDIX

Index of figures

Index of tables