Temperature-robust multivariate calibration

Wülfert, F.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
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.

Linear techniques to correct for temperature induced spectral variation in multivariate calibration.

**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\textsuperscript{30}.

\textit{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\textsuperscript{31, 32, 33} regression models, which are inverse calibration models of the type:

\[ Y = XB + F \]

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.

\textbf{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
Linear techniques to correct for temperature induced spectral variation in multivariate calibration.

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\textsuperscript{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\textsuperscript{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).
Linear techniques to correct for temperature induced spectral variation in multivariate calibration.

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.](image)

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.
Linear techniques to correct for temperature induced spectral variation in multivariate calibration.

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.

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

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

Table 3-2: Results for data set B.

<table>
<thead>
<tr>
<th>Method</th>
<th>No. of IV's</th>
<th>RMSEP Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>6</td>
<td>0.00324</td>
</tr>
<tr>
<td>Reference auto scaled</td>
<td>6</td>
<td>0.00356</td>
</tr>
<tr>
<td>Method 1.</td>
<td>6</td>
<td>0.00363</td>
</tr>
<tr>
<td>Method 2.</td>
<td>6</td>
<td>0.00410</td>
</tr>
<tr>
<td>Method 3.</td>
<td>1 ; 5</td>
<td>0.01303</td>
</tr>
<tr>
<td>Method 4.</td>
<td>6</td>
<td>0.00414</td>
</tr>
</tbody>
</table>

* 1st value (1) for preprocessing model, 2nd value (5) for calibration model.
Reference method, Global model: In a previous article\textsuperscript{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.

![Parity Plot](image)

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
Linear techniques to correct for temperature induced spectral variation in multivariate calibration.

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
Linear techniques to correct for temperature induced spectral variation in multivariate calibration.

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
Linear techniques to correct for temperature induced spectral variation in multivariate calibration.

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: 1.9x10^{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.](image)

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
Linear techniques to correct for temperature induced spectral variation in multivariate calibration.

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.

56
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.
Linear techniques to correct for temperature induced spectral variation in multivariate calibration.

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\textsuperscript{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.
Linear techniques to correct for temperature induced spectral variation in multivariate calibration.

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 $\Delta$. In general a signal is expressed on the Dirac basis with the orthonormal basis vectors $\delta(x)$ and coefficients $a_i$ and $b_i$. The basis vectors $\delta(x)$ are the Kronecker delta functions which are one for $x=i$ and zero for $x\neq i$:

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

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:

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

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

\[ \int_{-\infty}^{\infty} f_n f_m dx = 1 \quad ; \quad \int_{-\infty}^{\infty} \delta_n f_m dx = 1 \quad \text{for} \quad n = m \]

The following integral:

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

\[ = c_1 \int_{-\infty}^{\infty} f_1(x) f_1(x) dx + c_2 \int_{-\infty}^{\infty} f_2(x) f_1(x) dx + 1 \quad 0 \]

\[ + d_3 \int_{-\infty}^{\infty} f_3(x) f_1(x) dx + d_4 \int_{-\infty}^{\infty} f_4(x) f_1(x) dx \quad 0 \quad 0 \]

\[ = c_1 \]

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

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

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

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

This can only be equal to \( c_1 \) for all \( c_i \in \mathbb{R} \) when:
Linear techniques to correct for temperature induced spectral variation in multivariate calibration.

\[ \int_{-\infty}^{\infty} f_1(x - \Delta) f_1(x) dx = 1; \int_{-\infty}^{\infty} f_2(x - \Delta) f_1(x) dx = 0 \]  \[ \int_{-\infty}^{\infty} f_3(x - \Delta) f_1(x) dx = 0; \int_{-\infty}^{\infty} f_4(x - \Delta) f_1(x) dx = 0 \]

Combining equations 3 and 6 results in the following equality:

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

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


Linear techniques to correct for temperature induced spectral variation in multivariate calibration.


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


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


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


Linear techniques to correct for temperature induced spectral variation in multivariate calibration.