Hindawi
Journal of Spectroscopy
Volume 2018, Article ID 5037572, 14 pages
https://doi.org/10.1155/2018/5037572

Research Article

Dynamic Localized SNV, Peak SNV, and Partial Peak SNV: Novel Standardization Methods for Preprocessing of Spectroscopic Data Used in Predictive Modeling

Emily Grisanti,[1,2] **Maria Totska**,[2] **Stefan Huber**,[2] **Christina Krick Calderon**,[2] **Monika Hohmann**,[2] **Dominic Lingenfelser**,[2] **and Matthias Otto**[1]

[1]*Institute of Analytical Chemistry, TU Bergakademie Freiberg, Leipziger Str. 29, 09599 Freiberg, Germany*
[2]*Robert Bosch GmbH, Renningen, 70465 Stuttgart, Germany*

Correspondence should be addressed to Emily Grisanti; emily.grisanti@de.bosch.com

Received 15 March 2018; Accepted 10 July 2018; Published 28 October 2018

Academic Editor: K. S. V. Krishna Rao

An essential part of multivariate analysis in spectroscopic context is preprocessing. The aim of preprocessing is to remove scattering phenomena or disturbances in the spectra due to measurement geometry in order to improve subsequent predictive models. Especially in vibrational spectroscopy, the Standard Normal Variate (SNV) transformation has become very popular and is widely used in many practical applications, but standardization is not always ideal when performed across the full spectrum. Herein, three different new standardization techniques are presented that apply SNV to defined regions rather than to the full spectrum: Dynamic Localized SNV (DLSNV), Peak SNV (PSNV) and Partial Peak SNV (PPSNV). DLSNV is an extension of the Localized SNV (LSNV), which allows a dynamic starting point of the localized windows on which the SNV is executed individually. Peak and Partial Peak SNV are based on picking regions from the spectra with a high correlation to the target value and perform SNV on these essential regions to ensure optimal scatter correction. All proposed methods are able to significantly improve the model performance in cross validation and robustness tests compared to SNV. The prediction errors could be reduced by up to 16% and 29% compared with LSNV for two regression models.

1. Introduction

Chemometric approaches are becoming increasingly popular as they enable more comprehensive extraction of relevant information out of complex data provided by modern instrumental analytics. At the same time, advances in data analysis make it possible to reduce the size of the instrument hardware by compensating for the missing measurement quality of miniaturized instruments. In combination with multivariate calibration, the development of models based on low-cost analytics, such as vibrational spectroscopy, allows the development of models that predict parameters usually determined with cost-intensive measuring instruments or complex methods. Monitoring the alcoholic fermentation [1] and determining the viscosity of engine oil [2, 3] or proteins in milk [4] by spectroscopic means become thus feasible. It has also been possible to determine specific viscosity modifiers and pour point depressant additive compounds in engine oils [5] by FTIR, which is due to the fact that the concentration of a component follows, according to the Lambert–Beer Law, a linear dependency on the light absorbance of the medium [6, 7].

Preprocessing methods play a decisive role for the performance of these models, as spectra can be influenced by various disturbing factors that interfere with the significance of the measurement [8–11]. The main influence comes from the measuring geometry, which includes the sample thickness, the distance from the detector to sample, the contact pressure, and the angle from the light source to sample [12, 13]. The elimination of scattering effects by particles of different size and distribution also plays a major role in preprocessing.

Different spectroscopic measurement techniques suffer from different major disturbing factors. In near-infrared spectroscopy, it is usually a constant or linear baseline

offset due to scattering light, Raman spectra often show polynomial fluorescence background, and for mid-infrared spectra, the sample thickness and thus the spectroscopic response plays a crucial role [14, 15]. The information about the sample is present in the shape of the spectrum and independent of the offset (additive effect) and the scaling of the absolute signal intensity (multiplicative effect). The task of preprocessing is to remove these interfering factors from the informative part of the spectrum, and there are different approaches for this.

A method for eliminating constant offset terms is to calculate the first derivative [9]. This procedure can be extended to higher-order derivatives also eliminating offset terms with linear or quadratic baseline curves. The disadvantage of calculating the deviation of a spectrum is that noise effects are amplified.

Multiplicative signal correction (MSC) is another tool which can deal with the two major effects. A reference spectrum, in most cases represented by the mean spectrum of the calibration data set, is defined, and the spectra are corrected for the baseline and the multiplicative amplification effects [16, 17]. The approach is associated with the Kubelka–Munk theory, which takes optical phenomena caused by light scattering into account [18, 19]. For each spectrum, the two correction parameters are estimated via a least squares regression calculation.

Standard normal variate (SNV) removes a constant offset term by subtracting the mean value of the full spectrum and brings all spectra to the same scale by subsequent division by the standard deviation of the full spectrum [20]. Due to its simplicity, SNV is a popular preprocessing method [21]. SNV and MSC usually yield similar results and are often regarded as exchangeable [22]. Since no extra regression step is needed for the SNV transformation to estimate the correction parameters, in the following, the focus lies on SNV as the models should be kept as simple as possible.

Some efforts have been made to optimize standardization techniques. A piecewise MSC (PMSC) method has been proposed by Isaksson and Kowalski [23], which significantly improved the predictive power of several regression models based on near-infrared transmittance spectra. A Localized SNV (LSNV) approach has been introduced by Bi et al. performing the SNV not on the full spectrum but on subsequent sequences [24]. This strategy also yielded very promising results in several regression cases based on benchmark NIR data sets. In the following, a dynamic version of the LSNV algorithm, called DLSNV, is presented. By allowing for a dynamic starting point of the first and subsequent SNV windows, it is more flexible to align the SNV to important vibrational bands in the spectra. PSNV and PPSNV are based on the idea that the standardization can be optimized when performed on distinct wavenumber windows across highly specific regions of the spectrum.

2. Experimental

As a sample set, data originated from an investigation about aging and interaction phenomena in Automatic Transmission Fluids (ATF) were used. Many ATF samples have been stored for different periods at several temperatures to produce artificially aged samples.

The aim of the presented study was to transfer information coming from a highly specific, costly, and complex measurement method (High-Performance Liquid Chromatography coupled with Quadrupole Time-of-Flight-Mass Spectrometry (HPLC-QToF-MS)) to data measured with a low-cost, flexible tabletop instrument (Fourier-Transform Infrared (FTIR) spectrometer). This was achieved by analyzing each sample coming from the storage experiment and determining the additive response signals in these samples by HPLC-QToF-MS. By using these additive responses as reference values, a calibration model was created in order to be able to predict the concentration of the additive compounds in the samples by evaluating the FTIR spectra. The new standardization techniques proposed here are being tested for the regression models.

2.1. Additive Compounds. Two additive compounds from two different ATF oils were analyzed:

Within ATF A: an unsaturated ethoxylated amine known as friction modifier

Within ATF B: a bis-*tert*-butyl-hydroxytoluene (BHT) derivate known as phenolic antioxidant

2.2. Samples and Experiments. For the investigation of degradation phenomena in ATFs, a comprehensive storage experiment had been set up. The effects of different materials on ATFs and the impact of temperature on oil aging should be analyzed. Therefore, the ATFs were stored under various conditions in an oven. Three parameters had been varied: the storage temperature, the storage time, and added materials. The storage times had been adjusted to the temperatures so that a comparable load, according to Arrhenius Law, could be expected. The parameters are listed in Table 1.

For all time/temperature combinations, three interaction experiments have been conducted:

(i) storage with pure oil

(ii) storage with oil plus copper alloy chips

(iii) storage with oil plus chips from copper alloy, iron, and PA66

The samples were prepared by storing 100 ml fresh oil in a glass jar with a screw cap. The lid had been manipulated with a central hole that allowed air exchange.

2.3. Sample Measurements

2.3.1. FTIR. The FTIR spectra were collected in transmission with a Bruker Alpha instrument in combination with the QuickSnapTM transmission sample compartment in the wavenumber region ranging from 4000 to 600 cm^{-1} with a spectral resolution of 4 cm^{-1}.

The samples were measured without any special sample preparation with two different setups: (1) a droplet of ATF between two potassium bromide (KBr) discs separated by

Table 1

Temperature (°C)	Storage time (h)			
120	500	1000	2000	3000
140	105	210	415	625
160	25	50	105	165

a teflon spacer with the thickness of about 50 μm, and (2) fixed KBr cuvette of 100 μm thickness filled with ATF.

After each sample measurement, the KBr discs and the cuvette were rinsed several times with petroleum ether in order to prevent cross contamination. The cuvette was dried with N_2 gas after rinsing, and the KBr discs were dried under ambient air. For the measurement type (1), 4 spectra per sample were recorded, and for type (2), one spectrum per sample was recorded.

Due to the sample layer thickness, the hydrocarbon bands are saturated, and therefore, the spectra had to be cut in the wavenumber regions between 3000 and 2815 cm^{-1} (C-H stretching mode) and between 1491 and 1424 cm^{-1} (C-H bending and rocking mode). Additionally, the CO_2 bands were eliminated by cutting out the region from 2387 to 2285 cm^{-1} as well. The spectra of ATF A are shown in Figure 1 in transmission without any preprocessing as measured, in Figure 1(b) after truncation and SNV transformation, and in Figure 1(c), SNV transformed after calculating the absorbance spectra by using $A = -\log(T)$. In Figure 2, the same diagrams are shown for ATF B. In both cases, two series of curves can be discriminated from the raw spectra by the eye. The blue series comes from measurement type (1), and the red set comes from the cuvette measurements (2). To combine the two data sets from the measurement setups (1) and (2) are challenging tasks for a predictive model as the main variance is due to the thickness variation. The data set demonstrates the importance of suitable and sophisticated preprocessing methods in order to eliminate the difference in the spectra induced by the varying sample thickness. The standardization techniques presented here are able to meet this need.

2.3.2. Liquid Chromatography Coupled with Mass Spectrometry. The measurements for the determination of the additive compound signals were performed with an Agilent liquid chromatograph 1260 coupled with a high-resolution QToF 6540 mass spectrometer with methanol/water/ammonium acetate and isopropanol as an eluent. Ionization was carried out by means of electrospray (ESI). The final compound peak area data set was created using the Agilent MassHunter Qualitative Analysis B. 06.00 analysis software.

The response signals of the additive compounds are standardized by subtracting mean and dividing by standard deviation in order to bring all signal values on the same scale. The standardized signals are depicted in Figure 3.

3. Methods

3.1. Implementation. The proposed novel standardization methods and respective optimization processes were implemented via Python scripts.

3.2. Regression Algorithm—Ridge. For the prediction, the ridge regression estimator implemented in the Python scikit-learn framework for machine learning applications was used [25]. It is a linear model which solves a regression task via the least squares loss function $J(w)$ with L_2 regularization [26]. Regularization is an approach to minimize the issue of overfitting, which is particularly important for high-dimensional data such as FTIR spectra, by controlling the quadratic sum of the model coefficient w. This is done by adding the penalizing term L_2 weighted by the hyper parameter λ.

$$\lambda \|w\|^2 = \lambda \sum_{j=1}^{m} w_j^2. \tag{1}$$

Thus, the loss function is defined as

$$J(w) = \sum_{i=1}^{n} \left(y_i - y_{i,\text{pred}}\right)^2 + \lambda \|w\|^2, \tag{2}$$

where y_i stands for the reference value of the ith sample and $y_{i,\text{pred}}$ for the prediction of this sample. Since the performance of the preprocessing methods has to be assessed independently from the actually used predictive regression model, the same regression model with identical hyper-parameter λ was applied to the various preprocessed data sets. For the regression of the friction modifier compound of ATF A, $\lambda = 5$, and for the antioxidant of ATF B, $\lambda = 3$ was used. These parameters turned out to be the best choices regarding cross validation and robustness for the SNV transformed data set in a previously conducted internal study.

3.3. Model Performance Evaluation. To assess the performance of our models, two different approaches were chosen, namely, the predictive power under cross validation and noise addition.

3.3.1. Cross Validation. For cross validation, the mean from the different measurements of one sample was calculated. The sample set was randomly divided 50 times into a calibration and validation set by taking 70% of the data as training samples and 30% as test samples in each validation iteration with different combinations. Each separation run was provided with a unique random seed to ensure that the data set was split into the same training and test sets for each model, enabling better comparability of results between the different models.

3.3.2. Robustness against Noise. In order to assess the model performance under noisy input spectra, the model was calibrated by the full original data set. Random Gaussian-distributed white noise was added to each data point. These perturbed samples were predicted by the model and the prediction error was monitored. This was done for different noise levels. The random numbers added to each data point were generated by a standard normal distributed (mean: $\mu = 0$ and standard deviation: $\sigma = 1$) random

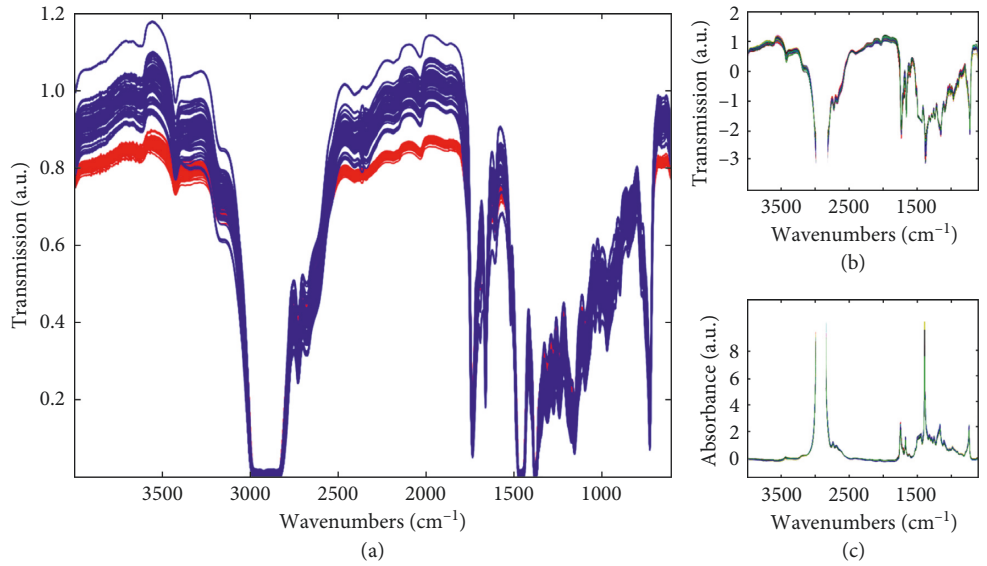

Figure 1: FTIR spectra of ATF. (a) Raw full transmission spectra without any preprocessing. The two data sets with different measurement setups can be discriminated by eye. The blue spectra originate from measurement type (1) with two KBr discs separated by a Teflon spacer, and the red set of curves originates from the cuvette measurement (2). (b) SNV-transformed transmission spectra after truncation of the saturated C-H vibrational regions and CO_2 areas and (c) SNV-transformed absorbance spectra after truncation.

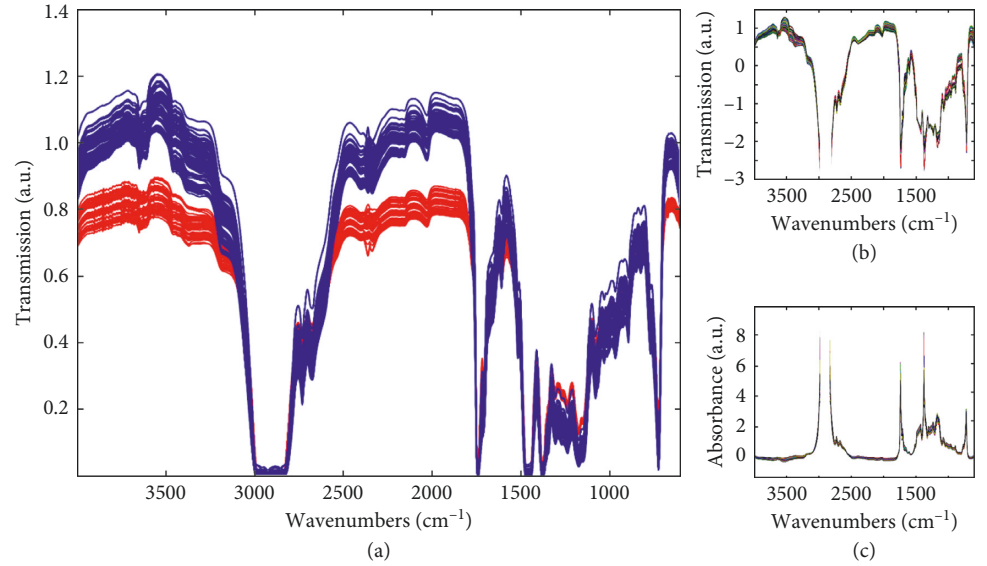

Figure 2: FTIR spectra of ATF B. (a) Raw full transmission spectra without any preprocessing. (b) SNV-transformed transmission spectra after truncation of the saturated C-H vibrational regions and CO_2 areas and (c) SNV-transformed absorbance spectra after truncation.

number generator. The noise levels were defined by the factors (0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, and 0.45), which were multiplied with the output of the random number generator. For each noise level, 50 simulated noisy data sets were generated and predicted by the pretrained model in order to be able to make well-founded statements about the model performance under noise perturbation.

The noise robustness workflow is a very helpful tool to investigate whether a good calibration error is a real advantage or if the model ran into overfitting. Using the same regression algorithm twice with different regularization parameters λ, the lower regularized model will generate a lower initial calibration error than the more stringent regularized model. But if the models are tested for robustness, the latter tends to have a lower error slope when the noise level increases.

3.4. Evaluation Metrics. The built-in functions R^2 score and mean squared error (MSE) of the scikit-learn framework were used as performance metrics.

3.4.1. Mean Squared Error (MSE). The mean squared error (MSE) of a prediction is calculated by the squared differences between the predicted value $y_{i,\mathrm{pred}}$ and the reference value y_i

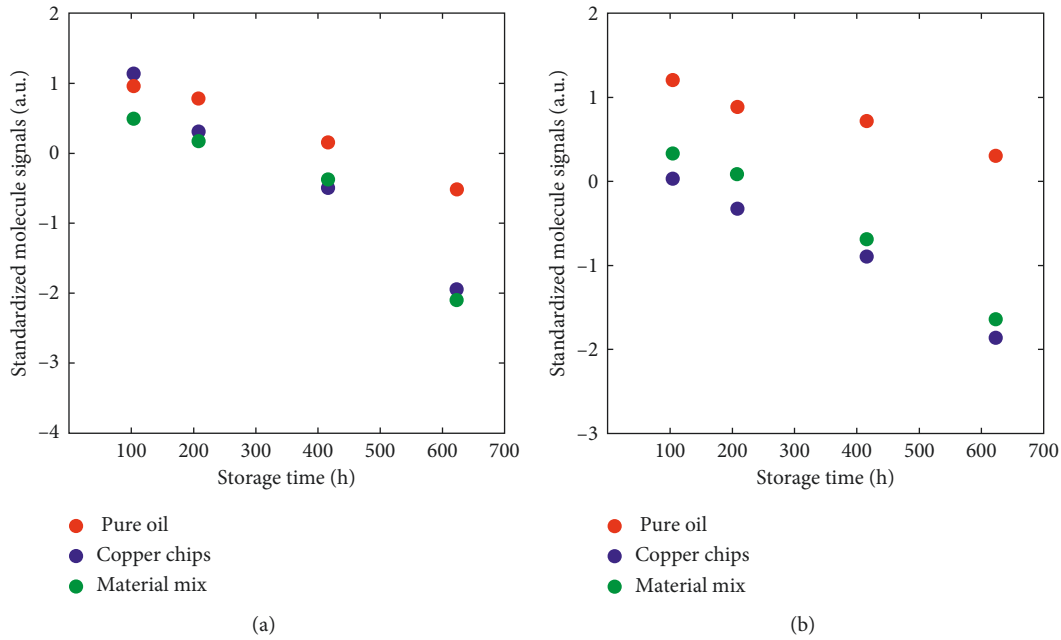

FIGURE 3: Standardized additive responses used as target value for the FTIR regression model for (a) the friction modifier compound and (b) the antioxidant plotted against time for storage temperature 140°C for all three storage experiments measured by HPLC-QToF.

of the ith sample. For a given data set with n samples, the MSE is the average value over all samples. It follows the following formula [27]:

$$\mathrm{MSE}\left(y, y_{\mathrm{pred}}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_{i,\mathrm{pred}}\right)^2. \tag{3}$$

The best possible MSE value is 0, and small values are desirable as the deviation from the correct prediction is low. From MSE, the root-mean-squared error (RMSE) was calculated by taking the square root. The RMSE value has the same dimension as the original reference target values.

3.4.2. R^2 Coefficient of Determination. R^2 describes the portion of the variance in the target values (dependent variables) that can be predicted from the spectra (independent variables) by the model [28]. The best possible score for R^2 is 1.0. R^2 gets 0.0 for a constant model which predicts a constant value disregarding of the input features. For linear regression modeling with intercept, R^2 is equal to the square of Pearson correlation coefficient between predicted and reference target values [29]. For a data set comprising n samples, the R^2 score is given as

$$R^2\left(y, y_{\mathrm{pred}}\right) = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_{i,\mathrm{pred}}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_{\mathrm{pred}}\right)^2}, \tag{4}$$

where $y_{i,\mathrm{pred}}$ is the model prediction of the ith sample which has a reference value y_i, and $\overline{y}_{\mathrm{pred}}$ is the mean value of all predictions.

$$\overline{y}_{\mathrm{pred}} = \frac{1}{n}\sum_{i=1}^{n} y_{i,\mathrm{pred}}. \tag{5}$$

3.5. Standard Normal Variate. Each spectrum $x = (x_1, x_2, \ldots, x_k)$ with k measured data points is transformed to the standardized form $z = (z_1, z_2, \ldots, z_k)$ by bringing the spectra to zero mean and unit variance. For this purpose, the mean spectrum $\overline{x}$ is subtracted from each data point x_i and divided by the standard deviation.

$$z_i = \frac{x_i - \overline{x}}{\sqrt{\sum_j^k (x_i - \overline{x})^2/k}}, \tag{6}$$

with

$$\overline{x} = \frac{1}{k}\sum_{j}^{k} x_j. \tag{7}$$

3.5.1. Dynamic Localized SNV (DLSNV). The DLSNV workflow is based on the SNV-transformed spectra data set (Figure 4(a)). To calculate the DLSNV data, the spectra are divided into multiple regions. On each of these regions, standardization is performed. To adjust the windows to important areas in the spectrum, a starting point can be defined. In Figure 4(b), the DLSNV spectra are shown, with a starting point of 100 and a window size of 300 pixels.

DLSNV algorithm

(i) Perform SNV on a window of the spectrum ranging from first data point to the s^{th} one
(ii) Subdivide spectra from s^{th} data point into windows of all the same size ws

To optimize the two parameters, *window size* ws and *starting point* s, a three-step approach is performed. In each step, the predictive power of the model is assessed via the

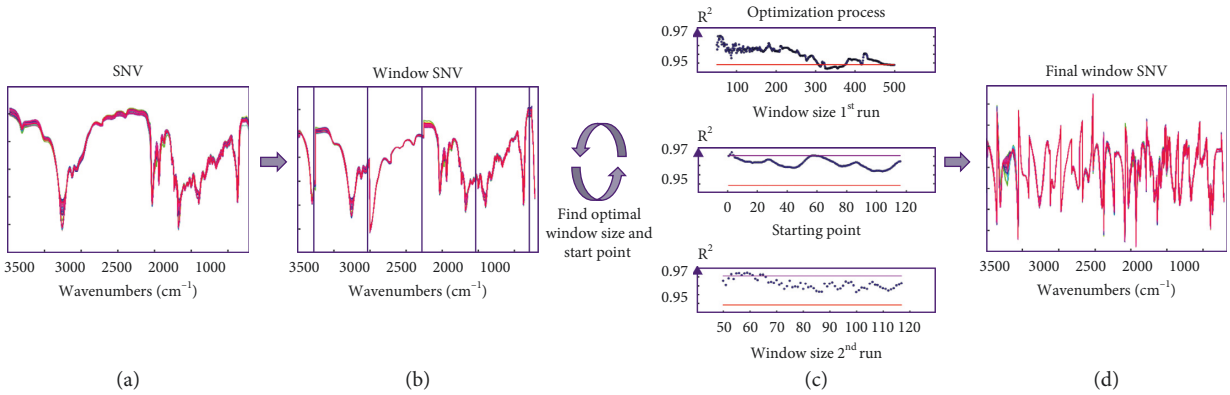

FIGURE 4: Demonstration of the workflow and optimization process for Dynamic Localized SNV. (a) Single SNV, (b) Dynamic Localized SNV with starting point 100 and window 300 for visualization, (c) three-stage optimization process for window, starting point and final window optimization, and (d) optimized DLSNV.

coefficient of determination R^2. The prediction performance of the chosen *window size* in combination with the regression model is benchmarked by fitting the same model to the single SNV data, indicated by a red line in Figure 4(c). The optimization steps can be summarized as follows:

(1) Perform LSNV with window sizes from 50 to 500 pixels, and determine R^2 for all window sizes. Find the optimal window size ws_{opt1}.

(2) Perform LSNV with optimal window size of step 1 $ws_{\text{opt 1}}$, vary the starting point from 0 to $2{\cdot}ws_{\text{opt 1}}$, and select the optimal starting point s_{opt}.

(3) Perform LSNV with optimal starting point s_{opt} with window sizes from 50 to $2{\cdot}ws_{\text{opt 1}}$ in order to find the best combination of window size $ws_{\text{opt 2}}$ and starting point s_{opt}.

In Figure 4(d), the final DLSNV spectra after optimization are shown. Note that jumps can occur between the individual standardization windows since the mean value of this current window is subtracted for each window. However, this does not affect the regression model.

In Figure 5(a), the ATF A samples are shown with SNV performed on the entire spectral region, and in Figure 5(b), the same spectra are depicted after DLSNV optimization. Figure 5(c) shows a zoom-in view of the highlighted region of Figure 5(a), and in Figure 5(d), the same region is depicted after DLSNV optimization. The baseline is removed for the exact spectra sequence, and thus, peaks are aligned in a way that the different aging levels of the samples can already be recognized by eye. The shown snipped spectrum is the phenolic antioxidant region. Thus, the decrease of this band can be associated with the aging level. Magenta indicates (relatively) fresh samples, whereas red indicates a strong degradation level.

3.5.2. Peak SNV. The idea behind the Peak SNV method is to standardize the important areas of the spectrum independently of each other. The optimization workflow for PSNV is shown in Figure 6, starting from the single SNV transformed data set. Data points with a high correlation with the target values (points of interest, POI) are selected (Figure 6(a)), and the SNV transformation is performed on windows around the centroids. Once the POIs are identified, the PSNV transformation is conducted as follows:

PSNV algorithm

(i) Subdivide spectra into sequences ranging from half the distance from the previous POI to half the distance to the next one (Figure 6(b)). SNV is performed across these windows.

To find the POI, an initial regression model is fitted to the data. In order to identify important regions of the spectra, the model coefficients are assessed. The normalized absolute values of the coefficient vector are fed into a peak-picking algorithm. Since it may occur that POIs are in close proximity, an agglomeration of the POIs is conducted in order to prevent from very narrow standardization windows. Peak centroids are calculated via the mean value of the combined POIs. The task for the optimization process is to find the best window for POI agglomeration, agg_{opt}, which is done by analyzing the calibration R^2 for each agglomeration window and picking the window size with maximal correlation between the predicted and reference target (Figure 6(c)). The steps are summarized as follows:

(1) Fit the data set to the target values (only calibration)

(2) Pick peaks from the normalized model coefficient vector ($|w|/\max(|w|)$), threshold for peaks = 0.1

(3) Combine peaks which are within a certain window agg, and calculate the centroid of the agglomerated POIs

(4) Perform PSNV across the centroid of the POIs

(5) Evaluate performance via R^2 for agg between 10 and 50 data points, and choose agg_{opt} according to maximal R^2

After optimization, each window has an individual window size and range over the peak centroid of important signals in the spectrum. On these windows, SNV transformation provides an optimal baseline and scatter effect removal. The optimized spectrum is shown in Figure 6(d).

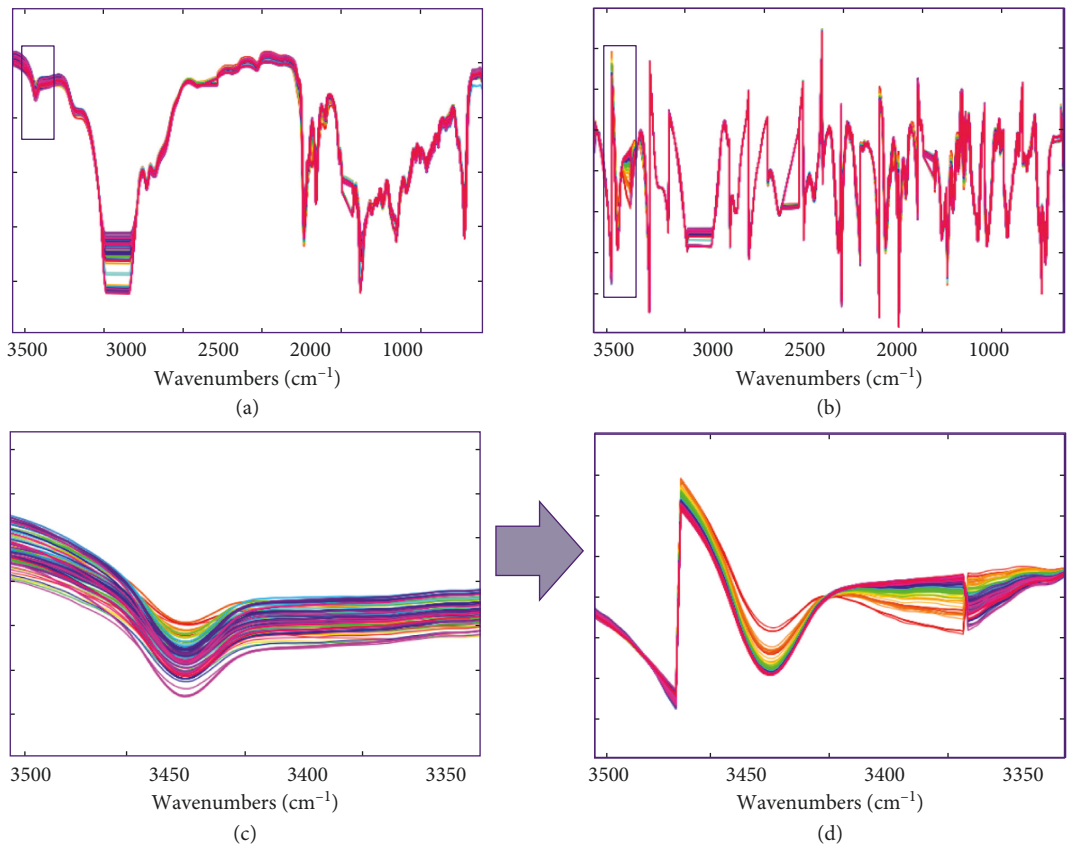

Figure 5: Demonstration of the improvement of peak alignment for DLSNV. (a) SNV-transformed transmission spectra with marked zoom level of (c). (b) Optimized DLSV spectra with marked zoom area of (d). Magenta indicates (relatively) fresh samples, whereas red indicates a strong degradation level.

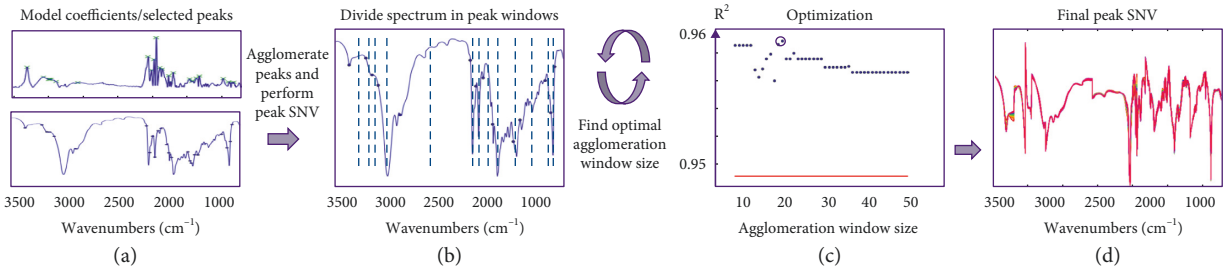

Figure 6: Demonstration of the workflow and optimization process for Dynamic Localized SNV. (a) Picked peaks of normalized absolute coefficient vector and indication of the POIs in one spectrum, (b) spectrum separation according to agglomerated peaks, (c) optimization process in order to find optimal agglomeration window size, and (d) optimized PSNV.

3.5.3. Partial Peak SNV. The idea behind Partial Peak SNV is similar to PSNV: picking the regions of the spectrum which show a high correlation with the target values, agglomerating POIs in close proximity, and standardizing these important spectral features (Figure 7(a)). But unlike for PSNV, not only the whole spectrum is finally taken into account but also a small window around the POI. It may occur that the same data point appears several times in different standardizations (see overlapping regions in Figures 7(b) and 7(d)). Due to this workflow, the PPSNV spectrum may have more data points (due to overlapping) or less (because not the entire spectrum is taken into account) than those of the original spectrum. A PPSNV spectrum is calculated as follows:

PPSNV algorithm

(i) Perform SNV across the POIs with a left and right margin of pw

The optimization focuses on the adjustment of the window size pw around the POIs in which the SNV is applied for maximal predictive power in calibration (Figure 7(c)). The optimization process is divided into the following steps:

(1) Fit the data set to the target values (only calibration)

(2) Pick peaks from the normalized model coefficient vector $(|w|/\max(|w|))$, threshold for peaks = 0.1

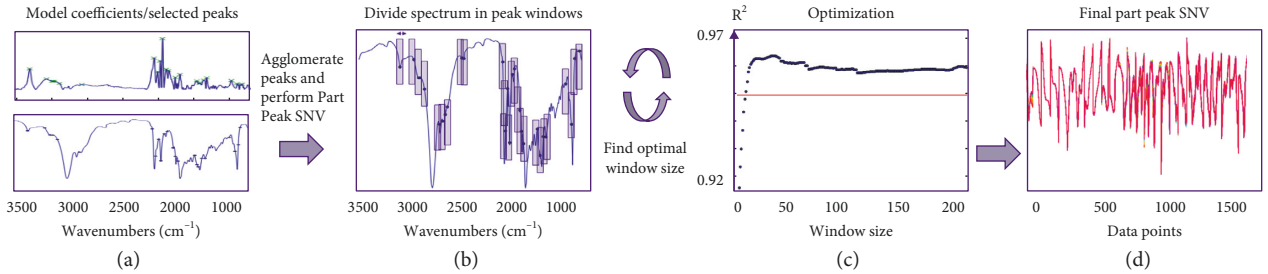

Figure 7: Demonstration of the workflow and optimization process for Dynamic Localized SNV. (a) Picked peaks of normalized absolute coefficient vector and indication of the POIs in one spectrum, (b) spectrum separation according to agglomerated peaks, (c) optimization process in order to find optimal window size around the POIs, and (d) optimized PPSNV.

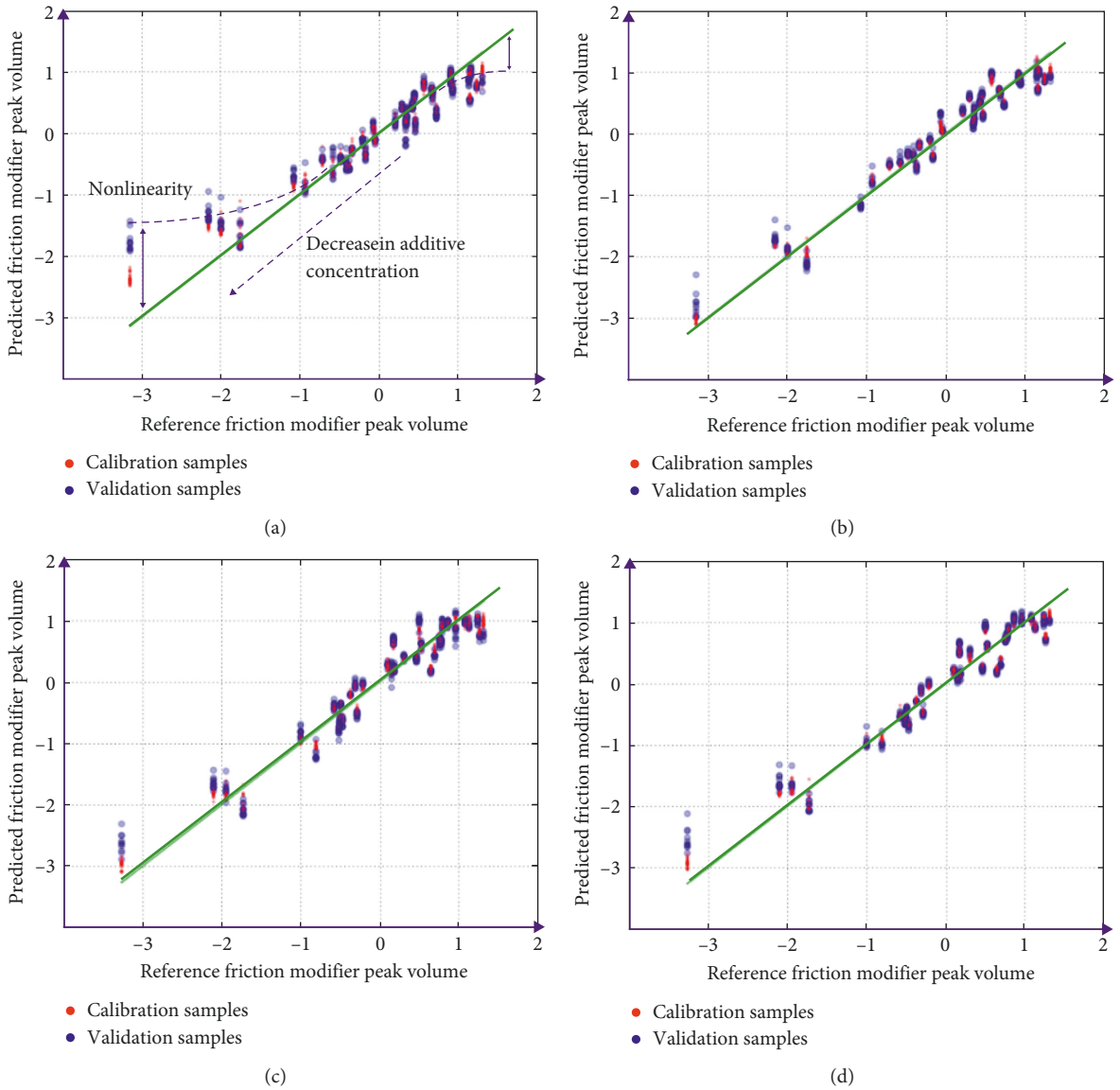

Figure 8: Cross validation results for (a) SNV, (b) Dynamic Localized SNV, (c) Peak SNV, and (d) Partial Peak SNV preprocessed spectra in the regression case of the friction modifier additive. The reference values were measured by HPLC-QToF. Red dots indicate prediction of calibration samples, and blue dots represent prediction of the hold-out validation samples. The dashed arrow in (a) indicates the increasing degradation of the samples with increasing time as the friction modifier additive concentration gets lower.

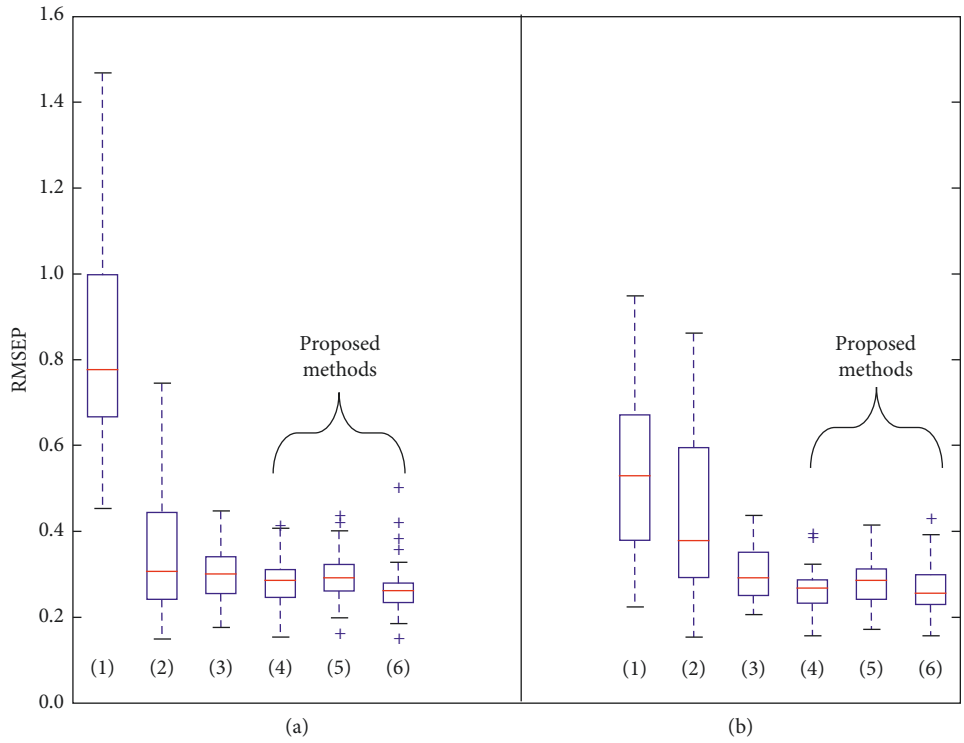

Figure 9: Box-and-whisker plot representation of the root-mean-squared error of prediction of the cross validation strategy (50 folds, random train test split of 70/30% of the data) for the friction modifier compound of ATF A. In (a), the RMSEP values for the transmission spectra are shown, and in (b), the RMSEP values for the absorbance spectra are shown. Boxplot (1) is without standardization, (2) is with a single SNV transformation on the full spectrum, (3) is with optimized LSNV, (4) is with optimized Dynamic DLSNV, (5) is with optimized PSNV, and (6) is with optimized PPSNV.

(3) Perform PPSNV across the peaks with the window size pw

(4) Evaluate the performance via R^2 for pw between 1 and 200 data points, and choose pw_{opt} according to maximal R^2

4. Results and Discussion

4.1. Cross Validation. In Figure 8(a), the cross validation recovery function for predictions of the SNV preprocessed spectra of the regression on the friction modifier compound is shown. A 50-fold cross validation strategy with a calibration/validation splitting of 70%/30% was used. Red dots represent the prediction of calibration, and blue dots represent validation samples. It is obvious that the linear model struggles to predict the high and low compound intensity regions correctly. The nonlinearity is visualized by an arrow and a dashed line to guide the eye.

In Figure 8 also, the cross validation recovery function for predictions after Dynamic Localized SNV (Figure 8(b)), Peak SNV (Figure 8(c)), and Partial Peak SNV (Figure 8(d)) optimization are shown. The saturation effect in the low intensity area of the compound response is almost completely removed in the latter three cases. It is also notable that the scattering around the green bisecting line is significantly reduced. Thus, the confidence interval for the predictions is improved.

The RMSEP values during the cross validation of the regression of the friction modifier component are summarized in Figure 9 in a box-and-whisker plot representation. The red line indicates the median, within the boxes, the interquartile range (IQR) (contains 50% of the data) is depicted, and the margins of the whiskers represent $Q_1 - 1.5 \cdot \text{IQR}$ and $Q_3 + 1.5 \cdot \text{IQR}$ for the lower and upper bound, respectively (Q_1 means the smallest 25% of the data set are smaller than this value and Q_3 means the smallest 75% are smaller than this value). Subplot Figure 9(a) refers to the transmission spectra and Figure 9(b) refers to the absorbance spectra. The labels are associated with (1) without standardization, (2) single SNV transformation on the full spectral range, (3) Localized SNV, (4) Dynamic Localized SNV, (5) Peak SNV, and (6) Partial Peak SNV.

It is noticeable that the RMSEP is very poor in case of the crude transmission spectra and that SNV has a very useful impact on them, whereas the improvement after SNV is low for absorbance spectra.

For all sophisticated optimized standardization approaches DLSNV, PSNV, and PPSNV, the median and the scattering around the median of RMSEP decreases drastically with respect to the SNV-transformed full spectra but also LSNV seems to be a reasonable choice. DLSNV on absorbance spectra is characterized by the lowest median and the smallest scattering confirmed by Table 2, summarizing the mean values and standard deviation of RMSEP.

Table 2: Summary of the model performances described by the mean value and the standard deviation of the RMSEP values during cross validation. The relative improvements and respective p-values compared with LSNV are also listed.

Method	Friction modifier ATF A		Antioxidant ATF B	
	Transmission	Absorbance	Transmission	Absorbance
Raw	0.83 ± 0.22	0.53 ± 0.16	0.90 ± 0.14	0.70 ± 0.11
Single SNV	0.34 ± 0.14	0.42 ± 0.17	0.61 ± 0.11	0.70 ± 0.13
LSNV	0.30 ± 0.07	0.31 ± 0.07	0.24 ± 0.04	0.24 ± 0.04
DLSNV	0.28 ± 0.06	$\mathbf{0.26 \pm 0.05}$	0.21 ± 0.04	0.21 ± 0.04
Rel. improvement	9% ($p < 0.05$)	**16**% ($p < 0.001$)	13% ($p < 0.001$)	13% ($p < 0.001$)
PSNV	0.28 ± 0.08	0.28 ± 0.06	0.33 ± 0.06	0.31 ± 0.07
Rel. improvement	8% ($p > 0.05$)	9% ($p < 0.05$)	−41% ($p < 0.001$)	−33% ($p < 0.001$)
PPSNV	0.26 ± 0.06	0.26 ± 0.06	$\mathbf{0.17 \pm 0.04}$	0.24 ± 0.05
Rel. improvement	13% ($p < 0.01$)	15% ($p < 0.001$)	**29**% ($p < 0.001$)	−3% ($p > 0.05$)

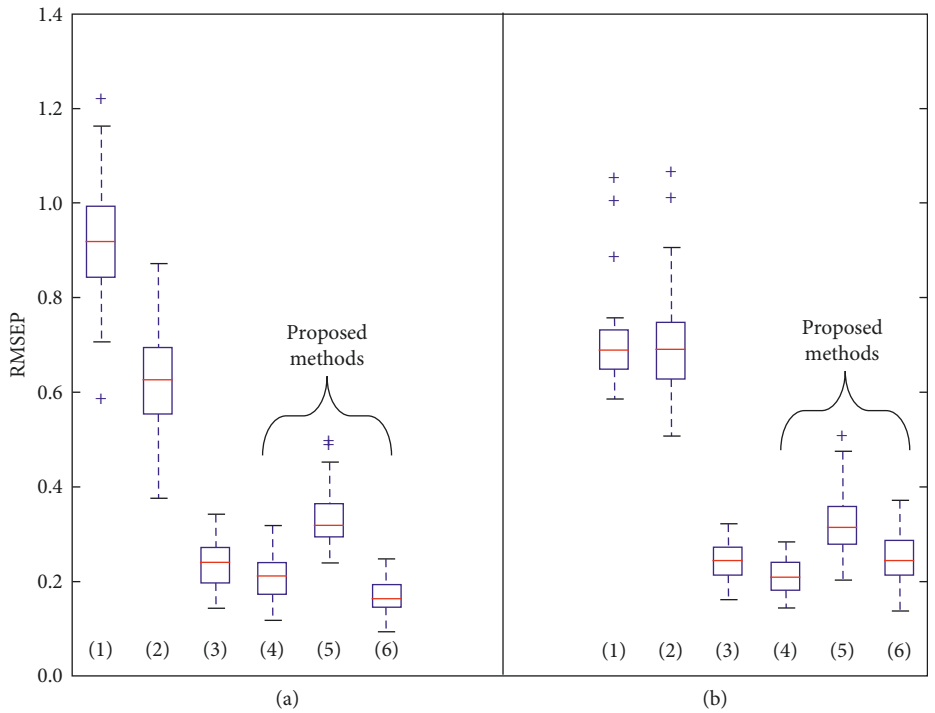

Figure 10: Box-and-whisker plot representation of the root-mean-squared error of prediction of the cross validation strategy for the phenolic antioxidant compound of ATF B. In (a), the RMSEP values for the transmission spectra are shown, and in (b), the RMSEP values for the absorbance spectra are shown. Boxplot (1) is without standardization, (2) is with a single SNV transformation on the full spectrum, (3) is with optimized LSNV, (4) is with optimized Dynamic DLSNV, (5) is with optimized PSNV, and (6) is with optimized PPSNV.

The summarized RMSEPs of the regression to the antioxidant additive of ATF B are shown in Figure 10 in a box-and-whisker plot representation where Figure 10(a) refers to the transmission spectra and Figure 10(b) to the absorbance spectra. In this case, DLSNV, PSNV, and PPSNV reduce both the median and the scattering around the median enormously when compared with SNV on full spectra. The best performance is achieved by PPSNV conducted on the transmission spectra confirmed by Table 2. On transmission spectra, DLSNV and PPSNV perform better than LSNV, but PSNV only has a positive effect when compared to SNV. In relation to LSNV, using PSNV, the predictive power is reduced. The fact that PPSNV is the best choice for this regression use case suggests that it is beneficial to only use spectral regions with high correlation with the target value and drop regions without or low correlation.

4.1.1. Noise Robustness. In Figure 11, the performance of the regression model for the prediction of noisy spectra is shown for the friction modifier. In subplot Figure 11(a), the curves for all preprocessings, are depicted and in Figure 11(b), a zoomed view is shown. Without any preprocessing, the initial calibration error for both transmission and absorbance spectra is very poor and rises very fast with the increasing noise level factor. Although the sophisticated preprocessing methods LSNV, DLSNV, PNSV, and PPSNV show a lower initial calibration error, the slope of the error is lower than for SNV. In Figure 11(b), the trend of the transmission spectra having low error steepness is visible. One may say that the three proposed standardization techniques show a very similar noise robustness behavior and are significantly better than none or

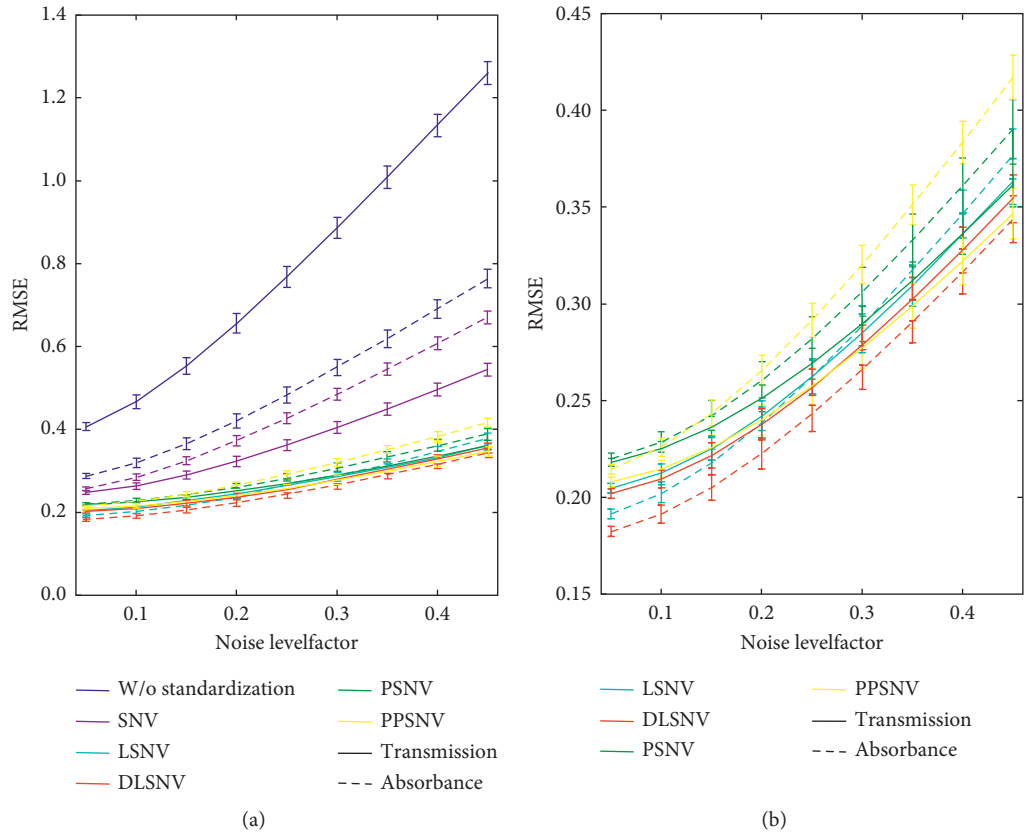

Figure 11: RMSE as a function of the noise level factor for the regression of the friction modifier compound of ATF A calibrated by the unperturbed full data set. The error bars represent the standard deviation of the prediction error calculated from the statistics of 50 repetitions of noise addition. In subplot (a), all curves are shown, and in (b), the sophisticated standardizations are shown.

SNV preprocessing, indicating that the model did not overfit the data as mentioned in Section 3.3.2.

In Figure 12, the performance of the regression model of the antioxidant for the prediction of noisy spectra is shown. The absorbance spectra with or without preprocessing show a similar noise trend as the SNV-transformed spectra. In Figure 12(b), the localized versions are shown in a zoomed view. The PPSNV preprocessing on the transmission spectra is characterized by the flattest noise dependency. These resuts demonstrate the superiority of the PPSNV method in this use case. As mentioned above, PSNV is not advantageous in this application and shows the lowest noise immunity, but it is preferable to the SNV across the whole spectrum.

4.2. Summary. The optimized parameters for the preprocessing methods are summarized in Table 3. The LSNV optimization process selects the same window size as DLSNV. Thus, the second window size run has no influence on the final result in these two cases but the starting point produces an improvement.

As already mentioned in Table 2, the cross validation performances of the tested methods are summarized as mean values and standard deviation for all cross validation runs. For the friction modifier, the performances of DLSNV, PSNV, and PPSNV are very similar. Table 2 also lists the relative improvements against the benchmark preprocessing, LSNV, accompanied by corresponding p values from a two-sided t-test, which tests the significance of the mean values being different (the deviation for relative improvements when the same mean value is given due to the fact that the improvements were calculated from exact values rather than rounded values).

The best mean RMSEP value for the regression model for the friction modifier of 0.26 is produced by DLSNV based on absorbance spectra. The antioxidant compound is modeled best by PPSNV preprocessing of the transmission spectra and yields a very low prediction error of 0.17.

To summarize, one may say that all proposed methods performed very well reducing both mean and standard deviation of the cross validation error compared with SNV. PSNV is not reasonable for the antioxidant additive as the performance is poor compared with the benchmark preprocessing method LSNV.

Which preprocessing method is the best depends on the actual regression use case, but in general, it is shown that PPSNV outperforms PSNV. This suggests that it is beneficial to drop spectral regions showing low or no dependency on the target value and to only consider highly correlated peaks.

For the antioxidant compound, PPSNV yielded an enormous improvement. This could be explained as the

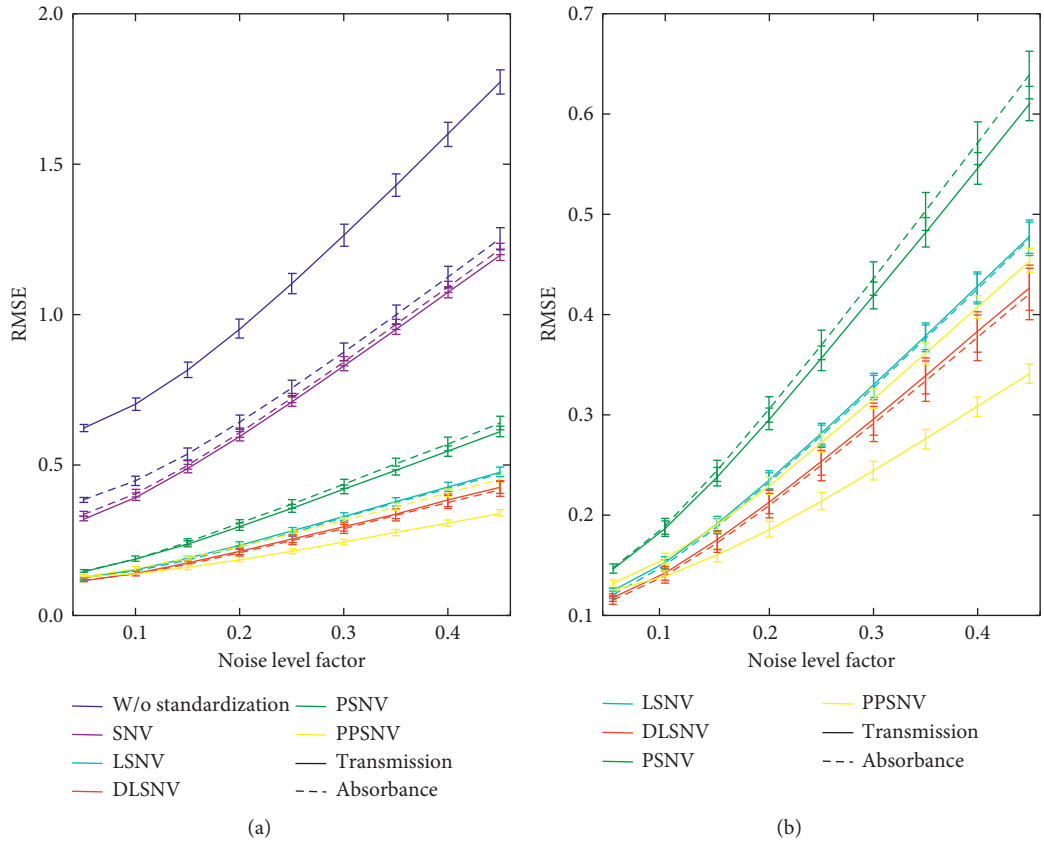

Figure 12: RMSE as a function of the noise level factor for the regression of the antioxidant compound of ATF B calibrated by the original full data set. The error bars represent the standard deviation of the prediction error calculated from the statistics of 50 repetitions of noise addition. In subplot (a) all curves are shown, and in (b), the sophisticated standardizations are shown.

phenolic aging inhibitor is a compound with very narrow vibrational band in the ATF B and thus does not have a great impact when the SNV is carried out across the entire spectrum. This may lead to a suboptimal alignment of this band. In case of novel standardizations, the SNV is optimized to the high correlative bands, and scatter effects can be compensated for these exact regions.

The fact that PSNV is unsuccessful for the antioxidant may be because the POIs are not centered to the middle of the SNV window and may have large left and right margins if they are far away from other POIs. As a result, they may not be optimally standardized. This is shown in Figure 6(b), where the single POI at about 2700 cm^{-1} has a large single SNV window.

The study provides an overview of model performances when using transmission or absorbance spectra suggesting that both cases can lead to valid regression models. However, for quantitative models built on transmission spectra, the SNV is vital, whereas in the absorbance case, the predictive power does not depend on SNV transformation. In absorbance spectra, the influence of the baseline constant is reduced because high transmission values are converted into low absorbance values.

To conclude, DLSNV, PSNV, and PPSNV were able to improve both transmission and absorbance predictive models. The scattering around the mean values are also drastically reduced because the model does not have to learn how to compensate for the baseline shift in each cross validation step leading to more reproducible results. Each vibrational band is optimally aligned so that the additive depletion trend is encoded in the absolute signal intensity, and the model does not have to weigh a data point as background correction.

Table 3: Summary of the preprocessing parameters. For DLSNV window size and starting point, for PSN agglomeration window, and for PPSNV, window width around POI is shown.

Method	ATF A		ATF B	
	Transmission	Absorbance	Transmission	Absorbance
LSNV	$w_{opt} = 52$	$w_{opt} = 52$	$w_{opt} = 51$	$w_{opt\,2} = 51$
DLSNV	$w_{opt\,2} = 52$	$w_{opt\,2} = 52$	$w_{opt\,2} = 51$	$w_{opt\,2} = 51$
	$s_{opt} = 5$	$s_{opt} = 7$	$s_{opt} = 48$	$s_{opt} = 48$
PSNV	$agg_{opt} = 5$	$agg_{opt} = 14$	$agg_{opt} = 6$	$agg_{opt} = 10$
PPSNV	$pw_{opt} = 17$	$pw_{opt} = 25$	$pw_{opt} = 15$	$pw_{opt} = 38$

5. Conclusion

The results presented in this study demonstrate the outperformance of the proposed novel standardization strategies Dynamic Localized SNV, Peak SNV, and Partial Peak SNV to improve both the mean and scatter of RMSEP

values in cross validation and the robustness against noise drastically with respect to SNV transformation executed on the entire spectrum. Against the benchmark LSNV, an enhancement of the predictive power of a ridge regression model by up to 16% and 29% could be achieved for the friction modifier and the antioxidant compound, respectively. The demonstrated optimization workflows for performing SNV on specific regions of the spectrum have been introduced here for the first time. Therefore, the standardization methods used in this paper are capable of eliminating nonlinearities by flexible rescaling in defined areas. To our knowledge, such standardization techniques have not been presented elsewhere.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. Zeaiter, J. M. Roger, and V. Bellon-Maurel, "Dynamic orthogonal projection. a new method to maintain the on-line robustness of multivariate calibrations. application to nir-based monitoring of wine fermentations," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, no. 2, pp. 227–235, 2006.

[2] R. M. Balabin and R. Z. Safieva, "Motor oil classification by base stock and viscosity based on near infrared (nir) spectroscopy data," *Fuel*, vol. 87, no. 12, pp. 2745–2752, 2008.

[3] A. Borin and R. J. Poppi, "Multivariate quality control of lubricating oils using fourier transform infrared spectroscopy," *Journal of the Brazilian Chemical Society*, vol. 15, no. 4, pp. 570–576, 2004.

[4] J. Huang, D. Brennan, L. Sattler, J. Alderman, B. Lane, and C. O'Mathuna, "A comparison of calibration methods based on calibration data size and robustness," *Chemometrics and Intelligent Laboratory Systems*, vol. 62, no. 1, pp. 25–35, 2002.

[5] M. A. Al-Ghouti and L. Al-Atoum, "Virgin and recycled engine oil differentiation: a spectroscopic study," *Journal of Environmental Management*, vol. 90, no. 1, pp. 187–195, 2009.

[6] R. Kellner, J. M. Mermet, M. Otto, M. Valcárcel, and H. M. Widmer, *Analytical Chemistry*, John Wiley & Sons Australia Limited, Milton, Australia, 2004.

[7] B. G. Osborne, T. Fearn, P. T. Hindle, and B. G. Osborne, *Practical nir Spectroscopy with Applications in Food and Beverage Analysis*, Longman Scientific & Technical, Wiley, Harlow, Essex, UK, 1993.

[8] J. D. Donald and D. D. Kevin, *Interpreting Diffuse Reflectance and Transmittance*, NIR, Chichester, UK, 2007.

[9] A. Rinnan, F. van den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009.

[10] E. Arendse, O. Amos Fawole, L. Samukelo Magwaza, N. Helene, and U. Linus Opara, "Comparing the analytical performance of near and mid infrared spectrometers for evaluating pomegranate juice quality," *LWT*, vol. 91, pp. 180–190, 2018.

[11] J. C. Machado, M. A. Faria, I. M. P. L. V. O. Ferreira, R. N. M. J. Pascoa, and J. A. Lopes, "Varietal discrimination of hop pellets by near and mid infrared spectroscopy," *Talanta*, vol. 180, pp. 69–75, 2018.

[12] H. W. Siesler, "Vibrational spectroscopy," in *Reference Module in Materials Science and Materials Engineering*, Elsevier, New York, NY, USA, 2016.

[13] A. Paula Craig, B. G. Botelho, L. S. Oliveira, and A. S. Franca, "Mid infrared spectroscopy and chemometrics as tools for the classification of roasted coffees by cup quality," *Food Chemistry*, vol. 245, pp. 1052–1061, 2018.

[14] S. R. Khandasammy, M. A. Fikiet, E. Mistek et al., "Bloodstains, paintings, and drugs: raman spectroscopy applications in forensic science," *Forensic Chemistry*, vol. 8, pp. 111–133, 2018.

[15] J. Engel, G. Jan, E. Szymanska et al., "Breaking with trends in pre-processing?," *TrAC Trends in Analytical Chemistry*, vol. 50, pp. 96–106, 2013.

[16] H. Martens, S. Jensen, and P. Geladi, "Multivariate linearity transformation for near-infrared reflectance spectrometry," in *Proceedings of Nordic symposium on Applied Statistics*, pp. 205–234, Stavanger, Norway, June 1983.

[17] P. Geladi, D. MacDougall, and H. Martens, "Linearization and scatter-correction for near-infrared reflectance spectra of meat," *Applied Spectroscopy*, vol. 39, no. 3, pp. 491–500, 1985.

[18] P. Kubelka and F. Munk, "Ein beitrag zur optik der farbanstriche," *Zeitschrift fur Technische Physik*, vol. 12, pp. 593–601, 1931.

[19] A. Claus Andersson, "Direct orthogonalization," *Chemometrics and Intelligent Laboratory Systems*, vol. 47, no. 1, pp. 51–63, 1999.

[20] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Applied Spectroscopy*, vol. 43, no. 5, pp. 772–777, 1989.

[21] M. Zeaiter, J.-M. Roger, and V. Bellon-Maurel, "Robustness of models developed by multivariate calibration. part ii: the influence of pre-processing methods," *TrAC Trends in Analytical Chemistry*, vol. 24, no. 5, pp. 437–445, 2005.

[22] T. Fearn, C. Riccioli, A. Garrido-Varo, and J. Emilio Guerrero-Ginel, "On the geometry of SNV and msc," *Chemometrics and Intelligent Laboratory Systems*, vol. 96, no. 1, pp. 22–26, 2009.

[23] T. Isaksson and B. Kowalski, "Piece-wise multiplicative scatter correction applied to near-infrared diffuse transmittance data from meat products," *Applied Spectroscopy*, vol. 47, no. 6, pp. 702–709, 1993.

[24] Y. Bi, K. Yuan, W. Xiao et al., "A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation," *Analytica Chimica Acta*, vol. 909, pp. 30–40, 2016.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] S. Raschka, *Python Machine Learning*, Packt Publishing, Birmingham, UK, 2015.

[27] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, Springer Texts in Statistics Springer, New York, NY, USA, 2003.

[28] O. Heinisch, "Steel, R. G. D., and J. H. Torrie: principles and procedures of statistics. (with special reference to the biological sciences) mcgraw-hill book company, New York,

Toronto, London 1960, 481 s., 15 abb.; 81 s 6 d," *Biometrische Zeitschrift*, vol. 4, no. 3, pp. 207-208, 1962.

[29] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, no. 1, pp. 240–242, 1895.