

ROBERT KOCH INSTITUT



Originally published as:

**Lasch, P.**

**Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging**

**(2012) Chemometrics and Intelligent Laboratory Systems, 117, pp. 100-114.**

**DOI: 10.1016/j.chemolab.2012.03.011**

This is an author manuscript.

The definitive version is available at: <http://www.sciencedirect.com/>

# Spectral Pre-processing for Biomedical Vibrational Spectroscopy and Microspectroscopic Imaging

Peter Lasch <sup>§</sup>

Centre for Biological Security (ZBS 6) "*Proteomics and Spectroscopy of Highly Pathogenic Microorganisms*" at the Robert Koch-Institut (RKI), Nordufer 20, D-13353 Berlin, Germany

<sup>§</sup> corresponding author. e-mail: LaschP@rki.de, phone: +49-30-18754 2259, fax: +49-30-18754 2606

**Keywords:** spectral pre-processing, chemometrics, biomedical vibrational spectroscopy, Raman spectroscopy, infrared spectroscopy, hyperspectral imaging

**Abbreviations:** CCD, charged-coupled device; COW, correlation optimized warping; CRM, confocal Raman microspectroscopy; DTF, detector transfer function; DTGS, Deuterated Triglycine Sulfate; DTW, dynamic time warping; EMSC, Extended Multiplicative Signal Correction; FPA, Focal Plane Array; FSD, Fourier self-deconvolution; FT, Fourier transform; FT-IR, Fourier transform infrared spectroscopy; MCT, Mercury Cadmium Telluride, MIR, mid-infrared; MNF, Minimum Noise Fraction; MSC, Multiplicative Scatter Correction; OTF, optical transfer functions; PSF, point spread function; SERDS, shifted excitation Raman difference spectroscopy; SG, Savitzky-Golay; SNR, signal-to-noise ratio; SR, spatial resolution; SNV, Standard Normal Variate; UHCA, unsupervised hierarchical cluster analysis

## 1. Introduction

With the development of modern analytical technologies such as infrared (IR) and Raman spectroscopy the capabilities of both generating and collecting data has been tremendously increased. Time-resolved vibrational spectroscopy, microspectroscopy, and vibrational hyperspectral imaging for example are now routinely employed in many areas of industry, technology development and scientific research. The advancements in IR and Raman instrumentation has led to an explosive growth in stored or transient data and has generated an urgent need for new and automated methods of spectral data analysis.

Spectral data pre-processing is an important first step in the workflow of IR and Raman spectra analysis which involves specific processing procedures performed on the raw data. Pre-processing has been shown to be of crucial importance for subsequent data mining tasks. In fact, it is now widely recognized that quantitative and classification models developed on the basis of pre-processed data generally perform better than models that solely use raw data [1-3].

With this review it is intended to explore the concepts and techniques of pre-processing methods and to discuss the applicability of distinct pre-processing techniques in the field of biomedical IR and Raman spectroscopy.

The main goals of data pre-processing can be summarized as follows:

- (i) Improvement of the robustness and accuracy of subsequent quantitative or classification analyses
- (ii) Improved interpretability: raw data are transformed into a format that will be better understandable by both humans and machines
- (iii) Detection and removal of outliers and trends
- (iv) Reduction of the dimensionality of the data mining task. Removal of irrelevant and redundant information by feature selection.

For systematic reasons it is useful to subdivide data pre-processing procedures into at least eight different categories. These groups can be used to perform the following operations:

1. *Exclusion* (cleaning) - this class of data pre-processing methods is used to detect and eliminate spectral outliers. Tests for spectral quality are important examples and aim at the detection of outlier spectra prior to further data mining tasks. Removal of outlier spectra can be achieved by labeling them as NaNs (Not a Number), for example as “bad” pixel spectra in hyperspectral imaging applications. Another way of outlier removal in hyperspectral imaging applications is the interpolation of bad pixel spectra by using spectral data from neighboring locations.

2. *Normalization* - normalization is used to scale the spectra within a similar range. In vibrational spectroscopy this is useful to compensate for the differences in sample quantity or a different optical pathlength. In data mining tasks involving spectral distance measurements normalization has been shown to improve the accuracy and efficiency of the models [1-3].

3. *Filtering* - in frequency analyses filtering is known as a group of methods that removes unwanted frequencies from the signals to be analyzed. Examples for high-pass filters widely applied in vibrational spectroscopy are derivative filters. These filters can be regarded as pre-processing methods that minimize broad baseline effects and amplify high-frequency signals (noise) in IR and Raman spectra. Low pass filters like smoothing filters have the opposite effect: low frequency components in the spectra are retained while the high-frequency components (noise) are attenuated. Band pass filters are combinations

of high and low pass filters. A good example of a band pass filter is the Fourier self-deconvolution filter which can be used for resolution enhancement and noise reduction at the same time.

4. *Detrending* - detrending is defined as a statistical or mathematical operation that removes underlying trends from series. Such trends may strongly superimpose, or obscure the changes of interest. In spectral time series, for example, detrending is often applied to remove long-term spectral changes. A simple and straightforward method of detrending spectral time series that are non-stationary in mean is the subtraction of the mean values of the spectra. More complex detrending routines aim at decomposing the spectral changes into a slow, or low frequency component (the trend) and the spectral changes of interest (which is often represented by the higher frequency components).

5. *Transformations*: Within the present review the term “transformation” is used to describe a group of pre-processing methods that are based on a well-defined physical model. Examples of such pre-processing routines are the ATR correction method, the Kramers-Kronig transformation, or the conversion routines between IR absorbance and transmission spectra.

6. *Feature selection* - feature selection techniques are known as pre-processing methods useful for both, quantitative and classification analyses. Contrary to other dimensionality reduction methods (PCA) the original representation of the variables is not altered [4]. Feature selection thus involves a process in which a subset of the original variables is selected for a specific purpose (e.g. to train a neural network). In this way, the original context of the variables is preserved which allows in vibrational spectroscopy interpretation of the selected spectral features by a spectroscopy expert [4]. Feature selection methods are, however, not a subject of this review.

7. *Folding/Unfolding*: Folding/unfolding is an operation in which an original  $n$ -dimensional spectral data matrix is reshaped such that the number of dimensions is modified while the data values and the number of elements remain constant. An IR or a Raman spectrum can be considered a 1-way data vector with spectral variables, usually an absorbance value (IR), or a Raman intensity value, being dependent on one argument  $\lambda$  (wavelength), or  $\nu$  (wavenumber). Examples of 2-way data are spectral time series in which the variable of interest is measured as a function of two arguments  $\nu$  and  $t$  (time). Hyperspectral imaging data can be considered 3-way data. IR absorbance values, or Raman intensities, are measured as a function of three variables, the spectral coordinate  $\nu$  and two spatial coordinates  $x$  and  $y$ . Imaging data can be easily transformed into a 2-way data format. This operation is termed unfolding; it involves rearrangement of the 3-way data matrix into a 2-way matrix where the original  $x$ - and  $y$ -modes are encoded by a new mixed spatial mode.

8. *Other methods*: This heterogeneous group comprises various pre-processing methods that have been specifically developed for vibrational spectroscopic applications: methods for the removal of spectral contributions from atmospheric water vapor, buffer subtraction routines for the analysis of IR spectra from proteins in aqueous solutions, techniques for correcting cosmic ray artifacts and methods for removing the fluorescence background in Raman spectra are important examples of this group of pre-processing routines.

In a broader sense, the FT operation which is required to transform an interferogram into an interpretable spectra format can be also considered a spectral pre-processing method. Parameters of the Fourier operation, the phase correction method, the apodization function and the zero filling factor evidently have an impact on the effective spectral resolution and the resulting signal-to-noise ratio (SNR). As an

exhaustive discussion of the expressions and relationships of operations in Fourier space is, however, beyond the scope of this review, the reader is referred for further details to the existing literature [5, 6].

The main purpose of this review is to make spectroscopy experts aware of the necessity and benefits of applying pre-processing techniques to their data. It is clear from the listing above that the discussion of the methods cannot be both, complete and comprehensive. The present article is thus focused on selected pre-processing methods popular in biomedical mid-infrared (MIR) spectroscopy (section 2) and on pre-processing techniques specifically developed for Raman spectroscopic applications (section 3). In addition to these sections, the review gives also an overview about selected routines for pre-processing 2-way, or multi-way spectral data (section 4). The section is complemented by a short introduction into combined pre-processing approaches (see section 5). With the present review it is intended to address the main issues and problems of spectral pre-processing from the point of view of a practitioner. Practical considerations were discussed rather than focusing on the exact mathematics of the given pre-processing method. It is however hoped that the interested reader will find useful information and some guidance to more detailed discussions of this important aspect of spectral data analysis.

## 2. Data pre-processing in biomedical IR spectroscopy

### 2.1. Quality tests

Among tests for spectral repeatability and reproducibility, quality tests are important first steps in any workflow of data analysis in vibrational spectroscopic studies. Quality tests can be considered as outliers tests; they were introduced to the field of biomedical vibrational spectroscopy in the late 1980's within the context of a project for FT-IR spectroscopic identification and classification of intact microorganisms [7]. In this project the quality assessment of the raw experimental spectra was carried out by defining certain criteria, or thresholds, of specific absorbance values, the SNR, intensities of infrared water vapor bands, optical fringes, and more. Since then, quality tests have been adapted and expanded for IR and Raman microspectroscopic imaging and usually comprise the following independent quality checks [8]:

1. Test for absorption bands of atmospheric water vapor (MIR only): based on second derivative spectra in the 1750-1900  $\text{cm}^{-1}$  region intensities thresholds of water vapor lines are defined.
2. A *so-called* test for sample thickness: in transmission type IR measurements integrated absorbance values can be considered as rough estimates for sample thickness. The test relies on upper and lower thresholds for intensities of selected vibrational bands (e.g. the amide I band), or of large spectral regions, respectively.
3. The test of the spectral signal-to-noise ratio (SNR): The SNR test allows to obtain the SNR for individual spectra and to eliminate those spectra that do not fulfill a certain threshold. When analyzing biomedical samples the signal is usually obtained in the amide I region (1620-1690  $\text{cm}^{-1}$ ). The standard deviation of the absorbance values in the signal-free spectral region between 1800-1900  $\text{cm}^{-1}$  is a popular means to assess the spectral noise level.

Other tests like the "test for a specific band" or the "bad pixel test" [9] allow to automatically remove spectra showing contributions from unwanted compounds (example: tissue embedding medium exhibiting strong absorption features from carbonyl esters), or to eliminate spectra acquired by dead pixels elements of a focal plane array (FPA) detector.

## 2.2. Water vapor correction

Variations of the atmospheric water vapor content during the time of sample and background collection are known to cause intense and sharp water vapor absorption features in the mid-IR (MIR) spectral region between 1350-1950  $\text{cm}^{-1}$  [10]. Water vapor rotational bands are of low intrinsic half-width and their precise wavenumber position and intensities are dependent on parameters like partial pressures, or the temperature. Spectrum (a) of figure 1 exemplary illustrates typical water vapor absorption features in the MIR spectral region between 1350-1950  $\text{cm}^{-1}$ . This spectrum, and the corresponding second derivative spectrum (see trace (b) of figure 1) demonstrate that atmospheric water vapor lines may significantly obscure important spectral details of MIR sample spectra.

Effective minimization of unwanted spectral contributions from water vapor can be achieved by two means [11]. First and foremost, the instrument and sample area should be purged by dry air [12,13]. Secondly, water vapor lines can be computationally removed by subtracting a weighted spectrum of pure water vapor from a sample spectrum (see figure 1). Such numerical routines require to measure a high-quality water vapor spectrum and to calculate a weighting, or water vapor correction factor. The water vapor correction routine which is illustrated next makes use of a specific feature of water vapor bands, their intrinsically low half-width.

1. Firstly, a high-quality absorbance spectrum of pure water vapor is measured. This spectrum should be obtained by using the same instrument and measurement parameters as those for sample measurement. The spectrum is converted to a second derivative spectrum.
2. A second derivative spectrum is obtained from the sample spectrum.
3. At the precise wavenumber positions of water vapor lines, preferably in the spectral region between 1750-1900  $\text{cm}^{-1}$ , second derivative intensities are extracted for both, the sample and water vapor spectrum.
4. The correction factor is then calculated by the ratio between the derivative intensity values of the water vapor and the sample spectrum. In case that more than one water vapor line was selected, the correction factor is determined by averaging the ratios.
5. Finally, water vapor correction is carried out by subtracting the product of the water vapor absorbance spectrum and the correction factor from the respective sample spectrum.

The routine has been successfully applied to compensate for spectral contributions of atmospheric water vapor in spectral time series [13, 14] and in FT-IR imaging measurements of tissues [15, 16]. The results of water vapor correction is exemplary illustrated by figure 1 which shows an original absorbance (a) and the corresponding second derivative IR spectrum from the human colon mucosa (b). Both spectra exhibit substantial contributions from atmospheric water vapor. Figure 1 illustrates also a spectrum of pure water vapor (c), the corrected sample spectrum (d) and the second derivative spectrum thereof (e).

## 2.3. Normalization

The use of IR spectra in classification analysis (library search) typically requires some form of normalization that allows an effective comparison across heterogeneous sets of samples [17]. Normalization has been thus identified as one of the most important pre-processing method which is commonly applied to minimize the effects of varying optical pathlengths on the data, or to compensate for intensity variations of the source (example: IR synchrotron) to mention one of the possible instrumental

causes. The result of normalization is a spectrum which is scaled and offset corrected at the same time. Normalization methods can be subdivided into two main groups. The first group of rather simple normalization methods requires only the information from the spectrum to be normalized. The second group comprises normalization methods such as Multiplicative Scatter Correction (MSC) [18], and Extended Multiplicative Signal Correction (EMSC) [19] which either require the presence of *so-called* collective (2-way, or multi-way) spectral data matrices, or of reference spectra. As the latter group of normalization methods will be presented in large detail in a separate contribution of this special issue the description is focused on the first group of normalization methods.

*Min-Max normalization:* Min-Max normalization is by far the most simple normalization method. In Min-Max normalization, spectra are first offset-corrected by setting the minimum intensity of the whole spectrum, or of a defined spectral region, to zero. Spectra are then scaled with the maximum intensity value equaling one.

*1-norm:* The second normalization method is sometimes referred to as 1-norm normalization. The first step of this method is mean centering; that is the average spectral intensity is subtracted from the spectrum. Mean centered spectra are then scaled such that the sum of the absolute values of all intensities equals one.

*Vector normalization:* Another popular normalization method is vector normalization, also called 2-norm. Mean-centered spectra are divided by the square root of the sum of the mean-centered intensities squared. In this way a spectrum is obtained in which the sum of all intensity values squared equals one.

*Standard normal variate (SNV):* SNV normalization is often used in near-infrared spectroscopy. Like in the case of vector-normalization the procedure starts with mean centering. SNV normalization is then achieved by dividing mean-centered spectra by the standard deviation over the spectral intensities [20] giving the resulting spectra a unit standard deviation of one.

## 2.4. Baseline correction

In transmission, or transfection type IR spectroscopy, spectral baselines can be distorted as a result of scattering (i), absorption by the supporting substrate (ii), changing conditions during data collection (iii), or the variableness due to instrumental factors (iv). Such baseline distortions are critical, particularly in quantitative analyses when e.g. absorbance values are systematically evaluated. Subtracting the estimation of a background from the un-processed spectrum leads to a more interpretable signal, allowing to determine spectral parameters (band positions, intensity values) more accurately [21]. A large variety of different methods for background estimation and correction has been suggested. Although baseline correction methods may rely on distinct principles and algorithms, they have the common objective of minimizing unwanted spectral offsets, broad baseline distortions, positive or negative slopes, and other baseline effects in vibrational spectra.

An illustration of popular baseline correction routines is given by figure 2. In this figure black-colored curves (a) denote a raw FT-IR absorbance microspectrum obtained from a cytoplasmic region of a human skin fibroblast. In this spectrum the baseline distortions are a result of moderate Mie-type scattering [22,23]. Baseline curves (b) obtained by four different methods are shown in the red color. Baseline correction (traces (c)) is achieved by subtracting (b) from (a).

*Offset correction:* This is one of the most simplest baseline correction methods. In offset correction, a straight horizontal baseline is subtracted from the spectrum (cf. figure 2A). The offset value is chosen such that at least one point of the corrected spectrum equals zero. Spectra are not scaled in this mode.

*Piecewise baseline correction:* A baseline is obtained by a number of user-defined points which are connected by straight lines. Correction is achieved by subtracting the baseline from the sample spectrum (see figure 2B).

*Polynomial baseline correction:* Instead of connecting user-defined baseline points by a straight line, a  $n^{\text{th}}$ -order polynomial function is used to fit the spectral data points (see figure 2C). Low-order polynomials should be preferred to avoid baseline correction artifacts. Polynomial baseline correction is extensively used in analysis of Raman spectra to flatten the baseline from contributions of fluorescence (cf. also section 3.2.).

*Savitzky-Golay (SG) baseline correction:* This baseline correction technique makes use of extensive smoothing of spectra. Spectral baseline curves can be obtained by using a zeroth-order SG smoothing/derivation filter with an extremely high number of smoothing points [9, 24] (see section 2.5 for details of the SG smoothing/derivative filter). The method allows to exclude spectral regions with strong signals and is recommended for use only with spectra containing a few sharp bands (see figure 2D).

Spectral baseline distortions can be considered also as low-frequency components of vibrational spectra. High pass filters have been successfully applied in many studies to eliminate spurious baseline components from vibrational spectra. Two high pass filter functions, derivatives filters and Fourier transform methods (see next section) have been systematically compared with a variety of other baseline correction methods in an article by Schulze et al.. For details of this comparison the reader is referred to this excellent feature article [25].

## **2.5. Spectral filtering (smoothing/derivatives, Fourier-self deconvolution)**

Spectral filtering techniques are widely employed in biomedical IR spectroscopy. Popular filters are noise filters for de-noising and smoothing, SG smoothing/derivative filters for smoothing/resolution enhancement, and various types of frequency filters such as Fourier self-deconvolution (FSD) employed in the interferogram, or Fourier domain.

*Noise filters:* Noise filters can be considered as specific low pass filters which are popular means for smoothing spectra. Such filters can be used to reduce random noise but have the drawback that depending on the type and the amount of smoothing the SNR is increased at the expenses of distorting the signal. Among others, popular smoothing filters are the zeroth-order SG smoothing/derivative filter [24], the binomial filter [26] and the moving average filter, also known as sliding average filter. These filters can be applied to 1-way data and do not rely on Fourier transformation as the spectral data points are replaced by some kind of local average of the surrounding data points. The application of noise filters to 2-way, or  $n$ -way data will be described separately (see section 4.1.).

*Derivative filters:* Derivative filters are popular means to enhance the resolution of infrared spectra. The filters are thus routinely employed to resolve and identify overlapping band components in complex spectral profiles. Another advantage of derivative spectroscopy is that contributions of baseline offsets or slopes are minimized. In this way the complexity of the spectra is reduced which facilitates spectral curve fitting by reducing the number of fit parameters [27]. Unfortunately, derivative spectroscopy requires a

high SNR. This is sometimes hard to achieve, for example in cases where IR microspectra are acquired at a diffraction-limited spatial resolution.

The *so-called* Savitzky-Golay (SG) smoothing/derivative method [24] represents one of the most popular filters. The technique has been suggested in 1962 by Marcel J.E. Golay and Abraham Savitzky and applies in its initial version to equally spaced and continuous data. The advantage of the method is that computation of the derivatives and smoothing is carried out in one single step. The fundamental concept of the SG smoothing and derivative method is the idea of a convolute and of a convolution function [24]. Convolution functions can be regarded as a vector of convolution integers and a normalization factor. The convolute of a given spectral data point  $x_0$  is obtained by computation of the dot product between the vector of convolution integers and a spectral segment of equal length with the midpoint being  $x_0$ . The result of this operation is then divided by a normalization factor. In SG filtering the length of the convolution vector is commonly referred to as the number of smoothing points. The operation has to be repeated for each single data point of a spectrum.

Savitzky and Golay have shown in their work that convolution vectors for smoothing and  $n^{\text{th}}$  order derivatives can be derived from the coefficients of least square-fit formulas [24]. Furthermore, they provided numerical tables of these vectors and demonstrated how convolution vectors can be used to obtain smoothed  $n$ -order derivatives in a single convolution operation. Although the original paper contained a number of typographical errors that were subsequently corrected by Steinier et al. [28] it became a classic and is now one of the most widely cited papers in the journal *Analytical Chemistry* [29].

*Fourier self-deconvolution:* Fourier self-deconvolution (FSD) is an alternative method for resolution enhancement. The FSD method was initially presented by Stone [30] and has been developed further as a technique to computationally resolve overlapped IR bands from spectra of condensed phase samples [31-34]. FSD has been employed in countless studies to reduce the degree of overlap between two adjacent bands, particularly in the field of secondary structure analysis of proteins [35-37]. An example of FSD application is given by figure 3. Trace (a) of figure 3A shows an infrared spectrum of the model protein RNase A between 1500-1800  $\text{cm}^{-1}$ . The protein was measured in a solution of heavy water ( $\text{D}_2\text{O}$ ). Complete H/D exchange was obtained after incubation of the protein in  $\text{D}_2\text{O}$  at 80° [14]. Furthermore, spectral contributions of the buffer were eliminated by the weighted subtraction method (cf. section 2.6). As mentioned earlier, the use of a Savitzky-Golay second derivative filter allows analysis of overlapped bands, such as amide I band components in the spectral region of 1620-1690  $\text{cm}^{-1}$ . For example, the second derivative spectrum of RNase A exhibits four important amide I band components which all of them attributed to distinct secondary structure elements (cf. trace (b) of figure 3A). Amide I band components identified in the second derivative spectrum can be in principle detected also by FSD filtering as the alternative resolution enhancement technique (see figure 3B).

Mathematically, FSD can be regarded as a specific band pass filter operation involving a deconvolution function as the high pass filter and a smoothing function as the low pass filter. For FSD, the measured spectrum is firstly transformed into an interferogram (Fourier domain spectrum) by an inverse Fourier transformation (FT). It is known that in the interferogram domain convolution reduces to a multiplication and deconvolution to a division. For deconvolution, the interferogram can be thus divided by the deconvolution function, (or multiplied by the inverse thereof). For Lorentzian band shapes, the inverse deconvolution function is an exponential function, which is multiplied by a smoothing, or damping

function. For recovering the FSD spectrum the product is subsequently processed by a forward FT operation. This sequence of steps is schematically illustrated by figure 3C-F: the interferogram given by the solid line of figure 3D is the inverse Fourier transform of the buffer corrected spectrum of RNase A of figure 3C (trace (a)). This signal is subsequently multiplied by a specific band pass filter function (figure 3D, dotted line) in the interferogram, or Fourier domain (see figure 3E). The forward Fourier transform of the product (figure 3E) gives then the FSD filter spectrum (figure 3F).

When applying FSD to real data one should be aware of the fact that the actual shape of the FSD filter function defines the factor by which the deconvolved bands are narrowed. Furthermore, FSD filter functions determine the shape of deconvolved bands and the SNR degradation in the FSD spectra [38]. Inadequate FSD filter parameters may result in under- or over-deconvolution, with the latter one characterized by noise amplification and the appearance of large negative side-lobes (see [31,38,39] for details).

## 2.6. Other pre-processing methods

In the previous section it was attempted to provide a comprehensive overview on spectral pre-processing techniques that found broad application in biomedical infrared spectroscopy. Naturally this overview cannot be both, detailed and comprehensive. The description was thus restricted to the most common pre-processing routines with the exception of MSC/EMSC [18,19] and the procedure to correct for spectral contribution of resonant Mie scattering [23,40-43]. Both procedures will be dealt with in separate articles of this special issue.

*Spectral subtraction:* Contamination of the samples of interest often result in additional bands and/or spectral distortions which can pose a serious problem for subsequent multivariate analyses. Spectral subtraction is applied in cases where infrared sample spectra have to be cleansed from spectral contributions of unwanted compounds ("contaminants"). Example for such contaminations are spectral contributions from the supporting substrate (e.g. from diamond windows or polymer films), of ingredients in powder mixtures, from the solvent or buffer, or from atmospheric water vapor and carbon dioxide. The basic principle of spectral subtraction has been already outlined on the example of the computational removal of water vapor bands (cf. section 2.2): first, a high-quality spectrum (high SNR) of the pure contaminant is obtained. Then, based on selected spectral parameters of the contaminant in the sample spectrum, a weighting factor is determined. Spectral correction for the contaminant is then carried out by subtracting the spectrum of the pure contaminant multiplied by the weighting factor from the sample spectrum. The procedure is simple, fast and efficient, but has important requirements: The first precondition is that measurements should be carried out under conditions where the detector transfer function (DTF), a measure of the detector output versus the IR intensity (signal input), is linear. This criterion involves several sub-criteria such as the maximum value for the product of optical pathlength, extinction coefficient and the concentration of absorbing species, or the type of the detector (MCT or DTGS), etc.. The second requirement for applying spectral subtraction is the independence of signals. To give an example, computational removal of spectral features from unwanted compounds will be successful only in cases where no molecular interaction between sample constituents and the contaminant takes place. Thirdly, spectral subtraction requires the absence of optical effects other than absorption (scattering). Spectral subtraction will be valid only in cases where these conditions are fully satisfied.

A practical application of spectral subtraction is given by figure 4. This example illustrates the standard pre-processing workflow in secondary structure analysis of proteins measured in aqueous solutions. Trace (a) of figure 4 shows the original un-processed infrared absorbance spectrum of the model protein RNase A which was dissolved in a 100 mM cacodylate buffer of heavy water ( $D_2O$ ). Infrared spectral measurements have been carried out in MIR transmission mode using a custom designed IR cell ( $CaF_2$ ) of an optical pathlength of approximately 45  $\mu m$ . The  $D_2O$  buffer spectrum (b) was obtained under identical experimental conditions (temperature, pH, IR measurement parameters) using an IR cell of a slightly reduced optical pathlength. The first step of pre-processing consisted in subtracting the buffer spectrum (b) from the spectrum of the protein solution (a). In this difference spectrum (c) the strong absorption features of  $D_2O$  such as the OD stretching band at  $\sim 2490\text{ cm}^{-1}$  and the DOD deformation band  $\sim 1210\text{ cm}^{-1}$  are largely compensated. On the other hand, the resulting difference spectrum (c) still exhibits broad OH stretching features ( $\sim 3380\text{ cm}^{-1}$ ) and a HOD deformation band ( $\sim 1450\text{ cm}^{-1}$ ) hidden under the amide II' contour [44]. These bands are due to the presence of residual protons in the  $D_2O$  buffer and the formation of HOD. A spectrum of this rare molecular species can be obtained by a second subtraction involving two  $D_2O$  buffer spectra varying slightly by H content (see trace (d) of figure 4; a varying H content of  $D_2O$  buffers can be attained by exposing  $D_2O$  to the open atmosphere for a short time). Correction for spectral contributions of HOD in the amide II' region (N-D bending vibration coupled to C-N stretching) at  $\sim 1448\text{ cm}^{-1}$  can be achieved by a third subtraction operation between the difference spectra (c) and (d). In this way, a protein spectrum corrected for spectral contributions of  $D_2O$  and HOD is obtained (cf. trace (e) of figure 4) [12]. Prior to structural interpretation the double difference spectra are usually further processed using resolution enhancement techniques such as Savitzky-Golay second derivatives (see trace (f)) or Fourier self-deconvolution.

### 3. Raman spectral pre-processing

Generally, most of the pre-processing methods discussed in the previous section can be successfully applied also to pre-process Raman spectra. Although there are a few exceptions to this rule (water vapor correction, resonant Mie scattering correction) the majority of pre-processing methods outlined in section 2 have shown their usefulness also in studies dealing with Raman spectroscopy as the analytical method. Many studies show no principal differences between the applicability of quality tests, normalization methods, baseline correction or spectral filtering techniques to Raman and IR spectra. However attention should be paid to lower SNR frequently found in Raman spectra of biological compounds. The latter fact is of particular importance when applying frequency filtering techniques (smoothing, derivative filtering, FSD).

The following section is thus focused on the description of pre-processing procedures that can be specifically applied for processing non-resonant Raman spectra. Such routines include methodologies to remove cosmic ray artifacts and strong fluorescence backgrounds, and to handle the wavelength calibration problem in Raman spectroscopy when using dispersive instruments.

#### 3.1. Removal of cosmic ray artifacts

Raw data from sensitive integrating detectors such as charged-coupled devices (CCD) commonly used in dispersive Raman spectrometers may contain artifacts originating from high-energy cosmic particles

hitting CCD detector elements. Cosmic ray artifacts manifest themselves as non-reproducible, sharp and intense features superimposed on the Raman signals. As these events can corrupt important parts of the Raman spectra and mislead subsequent multivariate analyses they are required to be either replaced by a local estimate or flagged as invalid.

A number of methods have been suggested for finding and eliminating cosmic ray artifacts [45-49]. The most simple approach of cosmic ray detection makes use of the temporal randomness of the artifacts and is thus based on the comparison of two consecutive Raman sample measurements. For example, cosmic ray artifacts can be routinely identified based on local correlations established between two consecutively measured sample spectra. Such a procedure is advantageous even in the case where no cosmic rays are found as it allows to increase the SNR via signal averaging. On the other hand two-, or multi-frame comparison methods have the shortcoming that they are based on the premise that actual spectral features stay unchanged over time [47].

Purely computational procedures sought to take advantage of the considerably lower half-widths of cosmic ray "bands". To give an example Phillips et al. suggested a method in which cosmic ray artifacts are identified by their deviations from the trends of the surrounding data, relative to a robust estimate of the standard deviation [48]. The authors suggested furthermore *so-called* missing point filters which could be successfully employed to replace cosmic spike features by interpolated data. The method combines spike correction and smoothing and can be used to effectively remove cosmic ray artifacts in 1-way data. The latter concept has been further developed by Hill and Rogalla [50], which proposed to perform spike correction and smoothing separately and extended the algorithm for multiple spike corrections [50].

An example for cosmic spike correction in 1-way data is given by figure 5. The upper red spectrum (red color) was obtained from a confocal Raman microspectroscopic (CRM) imaging data set obtained from the gray matter of a hamster brain section. In this example the sharp and intense cosmic ray artifact at approximately  $1795\text{ cm}^{-1}$  has been identified by comparing two consecutively measured (neighboring) Raman pixel spectra. The lower trace of figure 5) (blue color) shows the spectrum in which the cosmic ray feature has been replaced by interpolated data.

In the last decade, the advent of Raman imaging has led to the development of dedicated cosmic ray rejection methods which specifically apply to large hyperspectral imaging data sets containing hundreds or thousands of individual spectra. Because of the importance of this topic a separate section (4.3.) is devoted to the description of cosmic ray rejection in 2-way or n-way Raman data matrices.

### 3.2. Removal of the fluorescence background

Although Raman spectroscopy has been proven to be a powerful tool for biomedical and microbiological applications it has been severely limited in its applicability by fluorescence [51]. Fluorescence is characterized as a broad band emission that occurs in the same wavelength interval as the Raman signal. In some cases the fluorescence background can be  $10^6$ - $10^8$  times more intense [52] than Raman scattering so that the Raman signal may be entirely obscured.

A variety of different methods have been proposed to overcome the fluorescence problem. These methods can be roughly subdivided into three main categories: methods that aim at reducing the fluorescence signal and/or at enhancing the Raman signal: photobleaching, fluorescence quenching [53], removal of the fluorophores e.g. by sample washing, or filtration [54], the use of UV or near-infrared lasers which do not

stimulate fluorescence [52, 55], and resonantly enhanced Raman scattering [56] are examples of popular methods for improving the Raman-fluorescence intensity-ratio. The second group of techniques makes use of different physical properties of the Raman scattered light and the fluorescence: time resolved, or time-gated Raman spectroscopy can be used to separate between the almost instantaneously Raman scattering light and the fluorescence signal which is a comparatively long lasting effect [57, 58]. The fluorescence signal can be also separated by using anti-Stokes Raman spectroscopy [59] or by employing different polarization properties of fluorescence and Raman scattered light [60]. The next option for fluorescence rejection is known as shifted excitation Raman difference spectroscopy (SERDS) which makes use of the fact, that the fluorescence background does not change whereas the Raman scattering is frequency-shifted when the laser excitation wavelength is periodically modulated [51, 61, 62].

The third group of methods comprises purely mathematical or chemometric techniques to mitigate the fluorescence components of Raman spectra. Such methods include (fluorescence) baseline subtraction procedures using polynomial fittings [63-65], the use of first or second derivative filters [52, 66], the shifted-spectra technique [52], PCA analysis [67], wavelet transformations [68, 69], and the application of FT frequency filters [25, 52,]. Though each of these methods has been shown to be useful in certain situations, they are not without limitations.

The main advantage of software-based methods for fluorescence background correction is that these techniques do not require additional optical or electronic components or other complex instrument hardware modifications. Depending on the intensity and the spectral range of the fluorescence signal computational methods for fluorescence rejection have been proven to be efficient, inexpensive and relatively easy-to-perform.

Frequency-domain techniques such as FT filtering or the wavelet transform technique [70-75] make use of the fact that the fluorescence background is often composed of lower frequency components than the Raman signals. This allows in principle to separate the fluorescence signal from the Raman scatter (and noise). The main drawback of FT or wavelet filtering is under- or over-filtering, for example in cases when the frequency components of fluorescence and Raman scatter are not well separated [64]. Secondly, the results of FT filtering or wavelet transformation strongly depend on a number of complex parameters which introduces some subjectivity to the analysis [64]. Because of the complexity of frequency-domain filtering such methods have not been implemented in commercial software packages. The methods have thus found only limited application in practical setups.

Differentiation techniques as the next group of methods can be also effectively used to remove the fluorescence background. However as some implementations require extended computer modeling (i.e. fitting) of each spectral line [52], derivative methods often rely on complex mathematical fitting algorithms [64] and are thus also prone to subjectivity.

Among mathematical methods for fluorescence background removal, polynomial baseline fitting is the most commonly used [64,65]. This is due to the speed, simplicity and convenience of this technique. Another principal advantage of polynomial fitting over frequency or derivative filtering is that traditional Raman line shapes are preserved which facilitates interpretation of the Raman spectra in terms of chemical composition and structure.

Fluorescence background correction with polynomials relies on the assumption that a fluorescence background can be modeled by a polynomial. Using this approach the polynomial coefficients are

traditionally estimated in a least-square manner by using a set of baseline points that are either user-defined or set automatically [9]. It is important to define the polynomial order sufficiently low to reduce the complexity of the fit model and to avoid over-fitting. Meaningful values for polynomial order vary between 4 and 6 [65, 76].

When evaluating the least-square polynomial baseline correction method Lieber et al. have pointed out that the definition of baseline points should ideally consider spectral regions containing only fluorescence background and ignore regions containing Raman signals of interest [64]. This criterion, however cannot be always satisfied when baseline points are defined automatically. Manual definition of baseline points requires on the other hand user intervention and could be troublesome and time consuming in cases where many spectra have to be processed [21]. In that respect, sophisticated methods for automated polynomial baseline correction were suggested [21,64,65]. These techniques do not rely on user intervention and have been shown highly effective in removing the fluorescence background. The modified polyfit function for fluorescence subtraction [64], for example uses an iterative polynomial fitting approach to minimize a classical least square error in which peaks are eliminated. Mazet et al. suggested an alternative method that minimizes non-quadratic cost functions specifically designed for optical spectroscopy [21]. It could be demonstrated that asymmetric truncated quadratic cost functions are insensitive to large Raman peaks. Furthermore, the method of Mazet et al. allows to consider also the noise level which makes the technique particularly useful for fluorescence correction of Raman spectra [21, 76].

### 3.3. Wavelength calibration in dispersive Raman spectroscopy

Similarly to FT-IR spectroscopy, the accuracy of the wavelength position can be precisely maintained in FT-Raman spectroscopy by internal laser calibration. This section thus deals with wavelength calibration procedures that are required when using non-FT, i.e. dispersive Raman instrumentation.

Most of the modern dispersive Raman instruments utilize gratings in combination with sensitive CCD detectors for collecting the weak Raman signals. One of the main problems of such multichannel Raman spectrometers is the problem of wavelength stability [77,78]. Wavelength inaccuracies are often times inevitable and may occur as a result of instrumental factors such as source/grating changes, misalignment of the collection optics, thermal changes, and other factors. [79]. These inaccuracies can cause notable band shifts in the resulting Raman spectra making a detailed spectral analysis difficult, if not impossible. Wavelength accuracy is of particular importance in Raman difference spectroscopy, that is when a reference spectrum is subtracted from an experimental sample spectrum. It is well known that wavelength shifts between sample and reference spectra may result in derivative-like features which can render the Raman difference spectrum unrecognizable [80]. Other critical applications requiring increased wavelength stability include search-match applications (library searches) where Raman spectra of unknown compounds are systematically compared with spectral data bases by pattern recognition methodologies.

Many instrument manufacturers address the wavelength stability problem by offering menu-driven software protocols for wavelength calibration in which either absolute wavelength standards (atomic emission lines) or *so-called* Raman shift standards are utilized. While in the first approach one requires precise measurements of the laser line position [81], the use of a Raman shift standard will produce

Raman bands with a known shift relative to the laser line. Raman shift standards can be therefore used without knowing the precise laser line position.

Raman shift standards are available from the American Society for Testing and Materials (ASTM) as ASTM E 1840 shift standards [82]. The choice of the individual standard will depend on the number of Raman bands observed in the wavelength range of interest. For accurate wavelength calibration it is furthermore important that the material of the shift standard exhibits Raman bands across the full spectral range. In this way, wavelength uncertainties can be estimated at multiple locations within a spectrum.

Wavelength calibration by Raman shift standards is usually achieved by applying polynomial fit functions that aim at establishing a relation between the column numbers of the CCD detector and the individual Raman lines [83]. In this context, Carter et al. systematically compared first, second and third order polynomials for wavelength calibration of dispersive instruments. The first conclusion that was drawn was that calibration models require independent, or external validation. External validation reveals overfitting and ideally involves testing of an established calibration model by a second Raman shift standard. The authors found furthermore, that the simplest calibration model, the linear model, gave reasonably good and robust calibration results. For example, linear models performed comparatively well in situations where erroneous band positions were purposely introduced into the calibration data [83]. Finally, linear models were found particularly valuable in case of extrapolation, i.e. when significant parts of the spectrum to be calibrated lie beyond the initial calibration limits. In such cases the use of second or third order polynomials should be avoided [83].

### **3.4. Other pre-processing methods specific to Raman spectroscopy**

Additional methods employed to pre-process spectra from dispersive Raman instruments involve procedures to correct for (i) dark current of the CCD detector, (ii) the optical response of the spectrometer and (iii) the detector response [76]. Correction of the dark current can be easily achieved by subtracting a CCD signal measured without laser light, sample and slide from the Raman sample spectrum. To compensate for the optical system response it is first mandatory to obtain the signal under laser illumination but with the sample and slide being absent. Similarly to the previous case the corrected Raman spectrum can be then obtained by subtracting this signal from the raw data. The detector response function is required to correct for the frequency response of the CCD. This function can be obtained by illuminating the CCD by a calibrated polychromatic source (see ref [76] for further details).

Gobinet et al. suggested furthermore a pre-processing routine for peak width homogenization [76]. The basic element of this routine is a filter which is designed to transform a peak into a version that best fits a reference peak. For this purpose a convolution-deconvolution scheme was proposed that aims at minimizing the difference between selected reference and target peaks. The authors noted, however, that the suggested algorithm for peak width homogenization works only locally on the recorded spectra [76]. Although other techniques like correlation optimized warping (COW) or dynamic time warping (DTW) [84, 85] are considered to be potentially useful to homogenize peaks shapes over the entire data vector they could not be successfully applied in the study of Gobinet et al. [76].

## **4. Specialized spectral pre-processing routines for 2-way and n-way data**

Classical data pre-processing techniques employed in single point spectroscopy (1-way data) can be also applied to process 2-way, or n-way data matrices. Normalization, baseline correction, spectral filtering, to mention a few of the 1-way pre-processing methods, are routinely utilized to pre-process spectra from time series experiments or hyperspectral imaging data matrices. Aside from these methods there are a number of specialized pre-processing routines available that take advantage of existing correlations between spectral features in collective 2-way, or n-way data sets. To give an example, PCA-based noise reduction of 2-way spectral time series is carried out under the premise that spectra are recorded at almost identical measurement conditions with only one physical parameter (temperature, pressure, etc.) being varied over time. It is furthermore assumed that spectra of 2-way data matrices have a comparable spectral quality (SNR) with spectral outliers being removed.

The focus of the following section will be therefore on pre-processing methods that aim at the separation of physical and chemical/structural information in 2-way or n-way sample spectra.

#### 4.1. De-noising of 2-way data

*PCA-based noise reduction:* The PCA noise filtering technique is based on a principal component transform of a 2-way spectral data matrix. Unlike Savitzky-Golay smoothing, binomial, or moving average filtering, PCA based noise reduction cannot be applied to 1-way data and is thus employed for noise rejection in spectra from time series or imaging measurements. The PCA method itself is defined as a orthogonal linear transformation that can be used to decompose 2-way data into orthogonal vectors, or principal components (PCs). The number of PCs is equal to the number of spectral data points. The principal components describe the variance between the spectra and are ordered by the amount of variance they explain. PCA thus reorders data in decreasing order of variance, i.e. the first PC describes the majority of the spread of the data, the second PC explains the (independent) second-largest variance in the data, and so on. In consequence, low-order PCs represent most of the signal, whereas high-order PCs are supposed to contain mostly unexplained variance and noise. As each spectrum of a 2-way data matrix can be reconstructed by a linear combination of PCs, the basic principle of PCA-based noise reduction is to omit the noise content contained in the high-order PCs. This is usually achieved by neglecting or smoothing high-order PCs when reconstructing the 2-way data matrices. The details of the PCA-based noise-filtering process, including the determination of the number of PCs that are used for reconstructing the spectra has been reported by others [86-88].

*Minimum noise fraction (MNF) transform:* An alternative approach for noise rejection of 2-way data matrices has been introduced by Green et al. [89]. When analyzing remotely sensed multispectral imaging data, these authors found that the trend of decreasing signals/increasing noise with increasing PC index is not always obeyed [89]. To overcome this problem Green et al suggested a two-step cascaded method termed minimum noise fraction (MNF) transform. The basic idea of MNF is to introduce a noise-based ordering to the data. Instead of maximizing the variance (as in PCA), MNF attempts to order the data according to the SNR. The first step of MNF transformation thus involves normalization and de-correlation of the noise by using an estimated noise covariance matrix. This step is followed by a standard PCA of the noise-normalized data. The application of the MNF transformation requires knowledge of or an estimate of the true signal and the noise dispersion matrices [90]. The result of a MNF transformation is a partitioning of the data into a part represented by large eigenvalues and coherent eigenvectors of the

signals, and a second "noise" part of near-unity eigenvalues and noise-dominated eigenvectors [91]. For reconstructing the spectra noise can be then segregated from the data similarly to the PCA approach by using only the first (coherent) portions of the data.

Examples for the application of the MNF based noise reduction in hyperspectral IR imaging have been published by Bhargava and co-workers [92,93]. In both studies a modified version of the MNF approach [94] was utilized.

*Other methods for noise cancellation:* Wavelet transformation [70-75] has been reported as an effective alternative for noise cancellation in 2-way, or n-way vibrational spectroscopic data which clearly improved the accuracy of subsequent classification analysis [95,96]. Compared with other noise reduction techniques the wavelet-based approach was found to produce visually more appealing results in the neighborhood of peak patterns [96]. In this way the interpretation of the vibrational spectra is facilitated.

#### **4.2. Pixel binning in vibrational spectral imaging**

In statistics binning describes an operation in which the number of intervals or classes in a frequency distribution of continuous variables (histogram) is reduced. During the binning process the original variable values are replaced by new values representative for the binned intervals. Such variables are often obtained by interpolation, summation or by averaging of the original data values.

The term binning is also used in digital imaging, where pixels from a given detector array are grouped into larger "bins", or super-pixel units. Pixel binning is routinely carried out in a number of technical applications. For example, commercial digital camera and video systems offer options for combining the charge from adjacent pixels of a CCD on-chip during readout. In this way the signal intensity is increased and the SNR is improved. Furthermore, binning of digital images is applied to reduce the number of signals which are required to be processed. This allows higher frame rates, albeit at the expense of reduced spatial resolution [97].

The aggregation of pixels is usually achieved by combining quadratic pixel patterns. In 2×2 binning, for example, an array of 4 neighboring pixels is formed. Other binning patterns include 3×3 and 4×4 arrays, but alternative irregular patterns of pixel clusters are also possible.

Another common application of binning is spectroscopy. In modern Raman spectrometers, for example, dispersive elements, usually gratings, and dedicated non-quadratic CCDs (e.g. with 1024×128 pixel elements) are employed. In such instruments the wavelength-separated light is dispersed along the longer dimension of the CCD array. The larger horizontal dimension of the CCD array defines the spectral resolution (1024 wavelength channels in this example). On-chip vertical binning (1×128 binning: full vertical binning) of the Raman signals improves the SNR without any deterioration in spectral resolution.

Although on-chip pixel binning generally results in a better SNR compared with off-chip pixel binning, the latter pre-processing routine enjoys increasing popularity in infrared and Raman hyperspectral imaging. Reasons for this are three-fold: first and foremost off-chip pixel binning is purely software-based and therefore relatively easy to perform. While on-chip analog pixel binning always requires a specific hardware design of the CCD, or the FPA, off-chip digital binning offers much more flexibility in terms of size and shape of the binning patterns, and of the underlying binning algorithm. Secondly, software binning is like on-chip binning, a popular means for improving the SNR. A third important reason for performing binning is that the resulting data matrix contains a smaller number of pixel spectra. The latter

fact is of particular importance in many practical applications; binning has been thus identified as an important pre-processing method in vibrational spectroscopic imaging that helps to reduce time and efforts of computation, particularly in cases where multivariate cluster imaging methods are employed for image segmentation [98,99].

Software-based aggregation of neighboring pixel signals can be easily achieved by averaging or summation of pixel spectra. An alternative method for pixel binning includes 2-dimensional  $(x,y)$  interpolation of the chemical maps. Two-dimensional interpolations of the Raman, or IR chemical maps allow in principle to produce 3D hyperspectral data cubes of any  $(x,y)$  size and  $(x,y)$  aspect ratio. [9].

### 4.3. Cosmic ray removal in Raman hyperspectral imaging

Cosmic ray rejection in hyperspectral imaging can be regarded as a specific adaptation of the multi-frame comparison method for outlier detection. The basic idea of such comparison methods is that the local similarity (correlations, spectral distances) between neighboring pixel spectra can be used as a measure for the presence of cosmic spike artifacts. For example, Behrend and co-workers suggested a despiking method which works in the following way [100]:

1. Firstly, the Raman hyperspectral data matrix is preliminary de-spiked by smoothing with a median filter.
2. Correlations between an original un-smoothed central pixel spectrum and the smoothed  $3 \times 3$  neighborhood spectra are established.
3. The most highly correlated spectrum of the local  $3 \times 3$  neighborhood is then used to locate cosmic ray artifacts in the central pixel spectrum.
4. Finally, spectral regions showing cosmic ray artifacts are replaced by using the results of a polynomial interpolation.

This spike removal method can be used to successively correct all pixel spectra of a Raman hyperspectral imaging measurement. As pointed out by the authors the algorithm does not require sequential spectra acquisition and shows its strength at sharp boundaries between regions of high chemical contrast [100].

An alternative method for cosmic spike removal in hyperspectral imaging data also uses initial smoothing [9].

1. Raman spectra of the hypercube  $A$  are initially smoothed by employing a zeroth order 7-point Savitzky-Golay smoothing/derivative filter yielding the smoothed data matrix  $\bar{A}$ .
2. The difference matrix  $D$  is calculated as the difference between the original Raman intensities  $A$  and smoothed hyperspectral data  $\bar{A}$ .
3. Then, the mean standard deviation of  $D_i$  in each image plane  $std(D_i)$  is calculated and utilized for normalization:  $D_{norm,i} = D_i/std(D_i)$  with  $i$  being the wavenumber index.
4. To identify cosmic spikes, each of the planes  $D_{norm,i}$  is systematically screened for outliers. For this purpose, a threshold  $S$  is defined and spatial  $(x,y)$  indices of elements of  $D_{norm,i}$  with  $D_{norm,i} \geq S$  are determined. In the CytoSpec [9] implementation, the threshold  $S$  is obtained by the equation  $S = 10/sens$  with  $sens$  being a variable called sensitivity.
5. The final step is an operation in the spectral domain. Cosmic spike artifacts are excised and replaced by linearly interpolated Raman intensity values.

#### 4.4. Deconvolution in vibrational hyperspectral imaging

This group of computationally intensive pre-processing methods aims at increasing the quality of the imaging data: deconvolution is useful to remove or reverse blurring and to increase spatial resolution and image contrast. From the literature of conventional/confocal microscopy and image processing it is well-known that the resolution can be computationally increased using a variety of deconvolution algorithms. The first class of algorithms is represented by de-blurring methods. De-blurring methods apply to 2-way data ( $(x,y)$  chemical images); each chemical image of a hyperspectral data matrix is treated separately. In contrast, image restoration techniques involve deconvolution operations on 3-way  $(x,y, \nu)$  data matrices. Similar to Fourier self-deconvolution of 1-way data (see section 2.5.), deconvolution in vibrational hyperspectral imaging relies on the application of 2D, or 3D band pass functions that can be used to filter unwanted spatial and/or spectral frequencies. Successful application of deconvolution methods requires certain degree of SNR and an adequate sampling rate according to the Nyquist-Shannon sampling theorem [101]. Furthermore, the design and properties of the band pass filter is of crucial importance to the performance of the deconvolution algorithm. In cases where the response function of the optical system to a point source, the point spread function (PSF) is exactly known one can utilize the product of the PSF and a noise filter as the band pass filter function for deconvolution.

Nasse et al. recently presented a de-blurring method for FT-IR hyperspectral imaging matrices in which deconvolution was carried out by the help of the instrument-specific PSF [102]. The suggested deconvolution technique relied on the knowledge of the wavelength-dependent PSF: this function was determined experimentally by MIR transmittance measurements of a 2  $\mu\text{m}$  pinhole [102]. To avoid the introduction of additional noise, the pinhole data were fitted using the diffraction pattern of a Schwarzschild objective as the fit model [102,103]. Deconvolution first involved a series of 2D Fourier transforms (FT) of both, experimental image data and the PSF fits. Then, the FTs of the chemical images were divided by the corresponding FTs of the PSF (the FT of the PSF is called optical transfer function, or OTF). To these ratios a frequency-dependent Hanning filter was applied for noise suppression. The 2D inverse Fourier transforms of the filtered data were then computed. The final step of the deconvolution procedure consisted in rescaling (normalization) on the basis of the integrated intensities of the original chemical maps. Nasse et al. could show that deconvolution resulted in an increased image contrast and improved spatial resolution. The deconvolution algorithm was tested on several different test samples (tissue, polystyrene beads). These tests confirmed that the spatially deconvolved spectra preserved the original spectral features [102].

In cases where the PSF is only poorly determined, or entirely unknown deconvolution can be performed also using an estimate of the point spread function. Methods based on such approaches are called blind deconvolution techniques. Blind deconvolution can be performed in an iterative, or non-iterative way.

The *so-called* 3D-Fourier self-deconvolution (3D-FSD) technique suggested by Lasch et al. [99] represents an application of non-iterative blind deconvolution image restoration to FT-IR hyperspectral imaging data. This computational approach is an extension of the Fourier self-deconvolution technique developed by Kauppinen et al. [31-33] to 3-way data and assumes Lorentzian band profiles in the spectral and the spatial domains (see ref. [99] for details).

An example of the application of the blind deconvolution technique to a MIR hyperspectral data matrix (cryosection of the human colon mucosa) is given by figure 7. Figure 7A shows for comparison purposes

the Nomarski contrast image of a cross-sectioned individual crypt from the mucosa. In the crypts mucin producing goblet cells (G) are clustered around a central lumen (L) of these tubular glands. IR sample spectra were measured in transmission mode at the National Synchrotron Light Source in Brookhaven utilizing a confocal IR microscopy setup [99]. In the measurement the total sample area was  $140 \times 140 \mu\text{m}^2$  and  $36 \times 36$  point spectra were collected using a microscope stage step size of  $4 \mu\text{m}$  in  $x$ - and  $y$ -direction, respectively. A rectangular aperture of a size of  $8 \times 8 \mu\text{m}^2$  was used which gave a spatial oversampling factor of about 2. Figure 7B illustrates the chemical image produced on the basis of integrated absorbance values of the amide I band ( $1620\text{-}1680 \text{ cm}^{-1}$ ). In figure 7B the main morphologic structures such as the crypts in total, or the central lumen of the crypts can be visually identified whereas individual goblet cells are unresolved. As seen by the morphologic patterns of the individual goblet cells in the corresponding deconvolved map (see figure 7C) blind deconvolution was helpful in increasing contrast and resolution in MIR hyperspectral imaging data.

The main disadvantage of the 3D-FSD approach is that a wavelength dependence of the deconvolution function is not considered. Furthermore, the suggested algorithm requires to define 6 independent deconvolution parameters which makes the results of 3D-FSD subjective and the results of the procedure strongly dependent on the experience of the investigator.

#### **4.5. Correction of chromatic aberration artifacts in confocal Raman microspectroscopy (CRM)**

Due to the advent of a new generation of confocal Raman microspectrometers equipped with ultrasensitive CCD detectors the field of Raman microscopy could be extended to applications requiring a lateral and axial spatial resolution better than one micron. The high level of spatial resolution achievable with modern CRM technology allowed, for example, the Raman spectroscopic characterization of microorganisms on a single cell basis [104] and enabled to establish a rapid identification technique for clinically and technically relevant pathogens [105]. Other applications of CRM of single microbial cells include the characterization of the phenotypic heterogeneity in genetically homogeneous microbial cultures [106].

The technological advancements that allowed to improve the spatial resolution led on the other hand to a type of artifacts that were new to the field of Raman microspectroscopy. Examples of such artifacts are chromatic aberrations.

Recently, Lasch et al. suggested a simple pre-processing algorithm to correct for the axial component of the chromatic aberration in confocal Raman microspectroscopy. This study involved the acquisition of 4-dimensional ( $x, y, z, \nu$ ) Raman intensities from single bacterial endospores [107]. The correction method is based on measuring a vertical series of confocal Raman images by a high numerical aperture Raman microscope. Raman data were corrected by rearranging measured Raman intensities according to the known characteristics of the wavelength-dependent focal shift function of the optical system [107]. The results of the study suggested that correction of chromatic aberration distortions is mandatory for a comprehensive understanding of the information contained in the spectra. As uncritical interpretation of uncorrected spectral data would lead to wrong conclusions, the improved spatial resolution in CRM nowadays requires to address optical artifacts in the same way as it has been done for a long time in multicolor confocal fluorescence microscopy [107].

## 5. The combination of pre-processing methods

In a practical application, the data analysis workflow usually involves more than only one pre-processing step. In fact, in most of the biomedical IR- or Raman spectroscopic studies, pre-processing consists of a specific combination of two or more sequentially executed pre-processing steps [7,16]. The main advantage of such an approach is that many different requirements of successive analyses can be addressed simultaneously. To give an example, a typical sequence of pre-processings in MIR spectroscopy starts with an outlier, or spectral quality test which is followed by spectral filtering (SG smoothing/derivative filter), normalization (2-norm) and some kind of data reduction and/or feature selection [108]. This sequence can be complemented by procedures specific to MIR spectroscopy like water vapor correction, transmission-absorbance conversion, or ATR correction. In this way, pre-processing ensures at the same time outlier removal, dimensionality reduction, interpretability and improvements of robustness and/or accuracy of subsequent data mining applications.

It is one of the main data analysis tasks to adapt and optimize the pre-processing workflow to the specific needs of subsequent quantitative or classification analysis procedures. Even though there are a few studies available in which the effectiveness of different ways of pre-processing was systematically investigated [1,109-113,], the definition and optimization of the pre-processing workflow is still more an art rather than a science. It is the observation of the author, that the design and compilation of an efficient pre-processing workflow, and the optimization of the parameters thereof, is often based on experience and intuition of the investigator rather than on objective criteria.

The latter statement will be illustrated by an example. In a study on classification and identification of bacteria using MIR spectroscopy [114] (in which the authors used mainly hierarchical clustering for classification) the pre-processing workflow consisted in the following procedures: bacterial spectra were first checked by some quality tests (SNR, water vapor, etc.) and then subjected to first or second derivative SG smoothing/derivative filtering. After that, a pre-selection of five spectral windows was carried out considering the specific information content and discrimination power [114]. In addition the spectral information of individual spectral windows was rated by weighting factors which were intended to account for the specific contributions of cellular compounds such as fatty acids of the membrane or polysaccharides of the cell wall [114]. The choice between first or second derivatives, of spectral windows boundaries and of the weighting factors can be considered a subjective process in which the experience and the spectroscopic expertise of the investigator plays an decisive role.

The pre-processing workflow developed for an application of MIR microspectroscopic imaging on colorectal adenocarcinoma cyrosections [16] will be illustrated as a second example (see figure 8). The main goal of the study was to build up a database of spatially resolved point spectra which was used to teach and validate supervised neural network classification models. In this way the spectral information contained in the database of point spectra was evaluated by a neural network and served at a later stage for segmentation of 3-way MIR data from tissue sections of an unknown pathohistological status.

The first step of preprocessing consisted in the extraction of point spectra from the 3-way FT-IR imaging data sets (see figure 8, step). Selection and class assignment of point spectra was carried out on the basis of a precise pathohistological assessment of the carcinoma tissue sections under study. The totality of the extracted spectra was in the following subjected to quality tests in which the integrated intensity of a broad spectral region, the intensity of specific water vapor lines, and the SNR were systematically

investigated (figure 8, step 2). Point spectra, that have successfully passed the quality tests were further processed by means of a 7-point 1<sup>st</sup> derivative SG smoothing/derivative filter (cf. figure 8, step 3). After vector-normalization in the spectral range of 950-1480 cm<sup>-1</sup> the amount of data was reduced by extracting the spectroscopically relevant information between 950-1800 cm<sup>-1</sup> and 2800-3100 cm<sup>-1</sup> (see figure 8, steps 4 and 5; the regions of 1800-2800 cm<sup>-1</sup> and 3100-4000 cm<sup>-1</sup> contain spectral information of only secondary importance). The final pre-processing step consisted in feature selection. For spectral feature selection a defined number of discriminative spectral features (60, or 85) was selected on the basis of the covariance between features from labeled spectral subsets (see refs. [16,114] for details).

## 6. Conclusions

In biomedical IR and Raman spectroscopy data pre-processing has been identified as an essential component in an effective data analysis workflow. Its purposes include outlier rejection, reducing dimensionality, removal of irrelevant or redundant information and improvements of interpretability, robustness and accuracy of subsequent quantitative or classification analysis tasks. With this review article it was intended to explore the concepts and techniques of a large variety of individual pre-processing methods and to discuss the applicability of different pre-processing techniques in the context of practical applications of biomedical vibrational spectroscopy. It is hoped that this article not only serves as a starting point for beginners in the field, but also as a source of reference for more experienced spectroscopists.

## 7. Acknowledgements

The author is thankful to D. Naumann and H. Fabian for useful discussion and comments.

## 8. Legends to the figures

### Figure 1

Example of water vapor correction of FT-IR sample spectra.

- (a) - Raw FT-IR absorbance spectrum from the human colon mucosa showing spectral contaminations due to atmospheric water vapor in the spectral region of  $1350\text{-}1950\text{ cm}^{-1}$ .
- (b) - Second derivative spectrum of (a). Spectral features of the tissue are strongly superimposed by features of water vapor.
- (c) - FT-IR absorbance spectrum of pure water vapor.
- (d) - Corrected sample spectrum. Spectral contributions of water vapor have been minimized by weighted subtraction with  $(d) = (a) - cf \times (c)$  with  $cf$  being the correction factor.
- (e) - Second derivative spectrum of (d), spectra are shifted along the y-axis for clarity

### Figure 2

Baseline correction of FT-IR spectra collected from the cytoplasm region of a human skin fibroblast.

- A - offset correction, (a) - raw absorbance spectrum; (b) - baseline curve; (c) - corrected spectrum
- B - piecewise baseline correction (7 user-defined baseline points)
- C - polynomial baseline correction (order of the polynomial: 6; number of user-defined baseline points: 10)
- D - Savitzky-Golay (SG) baseline correction (zeroth-order SG smoothing/derivative filter with 199 smoothing points, the spectral region of  $1462\text{-}1780\text{ cm}^{-1}$  has been linearly interpolated).

### Figure 3

Resolution enhancement by Fourier self-deconvolution (FSD) and second derivatives

- A - Infrared spectrum of RNase A in heavy water, corrected for spectral contributions of the  $\text{D}_2\text{O}$  buffer and HOD (a) and the corresponding second derivative spectrum (b).
- B - The FSD-spectrum of trace (a)
- C-F - Illustration of the FSD technique (see text for details)

### Figure 4

Pre-processing of FT-IR protein spectra as an essential prerequisite for secondary structure analysis.

- (a) The infrared spectrum of RNase A (25 mg/mL) dissolved in a cacodylat  $\text{D}_2\text{O}$  buffer (100 mM Na-cacodylat at pH 7.0). Optical pathlength of the IR cell:  $45\text{ }\mu\text{m}$ .
- (b) FT-IR spectrum of the  $\text{D}_2\text{O}$  buffer measured at the same conditions in an IR cell of a slightly smaller optical pathlength.
- (c) Difference spectrum between (a) and (b) magnified by a factor of  $2 (\times 2)$ .
- (d) HOD spectrum obtained by subtracting two cacodylat  $\text{D}_2\text{O}$  buffer spectra of varying H-content ( $\times 2$ ).
- (e) Double difference MIR spectrum of RNase A corrected for spectral contributions of the  $\text{D}_2\text{O}$  buffer and for residual HOD:  $(e) = (c) - wf \times (d)$ ,  $wf$  - weighting factor, ( $\times 2$ ).
- (f) Second derivative spectrum of RNase A in the spectral region of  $1500\text{-}1750\text{ cm}^{-1}$  (water vapor corrected, SG second derivative filter with 9 smoothing points).

Spectra are shifted along the y-axis.

#### *Figure 5*

Example of cosmic ray correction of Raman spectra.

Red trace - Raman spectrum obtained from the cerebellum of a Syrian hamster showing a strong cosmic ray feature around  $1795\text{ cm}^{-1}$ . Blue: the corrected Raman spectrum (spectra are shifted along the y-axis for clarity)

#### *Figure 6*

Removal of the fluorescence background in confocal Raman microspectroscopy

(a) - raw Raman spectrum; (b) - polynomial baseline; (c) - fluorescence corrected spectrum. Order of the polynomial: 7; number of user-defined baseline points: 10.

#### *Figure 7*

Example of resolution enhancement by 3D-Fourier self-deconvolution (3D-FSD).

Panel A: Nomarski contrast image of a thin section from the colon mucosa. The upper-central area displays a cross sectioned crypt with a central lumen (L) and individual goblet cells (G).

Panel B: chemical image assembled on the basis of baseline corrected absorbance spectra using the integrated absorbance of the amide I band ( $1620\text{-}1680\text{ cm}^{-1}$ ).

Panel C: chemical image assembled after applying 3D-FSD. Imaging parameter: baseline correction, area of the amide I band ( $1620\text{-}1680\text{ cm}^{-1}$ ).

Panel D: Comparison of intensity profiles obtained from original data (black curve) and 3D-FSD data (red curve).

(reprinted from Lasch et al. [99] with permission)

#### *Figure 8*

Combination of pre-processing steps: the pre-processing workflow in a FT-IR microspectroscopic imaging study on colorectal adenocarcinoma diagnosis [16]. The raw spectral data, 3-way imaging data, were processed using the following sequence of pre-processing steps: (i) extraction of point spectra; (ii) quality tests; (iii) SG smoothing/derivative filtering; (iv) normalization; (v) spectral window selection; (vi) feature selection (spectra may be shifted along the y-axis, see also text for details)

## 9. References

- [1] P. Heraud, B.R. Wood, J. Beardall, D. McNaughton, Effects of pre-processing of Raman spectra on in vivo classification of nutrient status of microalgal cells, *J. Chemom.* 20 (2006) 193-197.
- [2] K.Z. Liu, K.S. Tsang, C.K. Li, R.A. Shaw, H.H. Mantsch, Infrared spectroscopic identification of beta-thalassemia, *Clin Chem.* 49 (2003) 1125-32.
- [3] A. Kohler, A.N. Kristian, H. Martens, Chemometrics in Biospectroscopy, in: Handbook of Vibrational Spectroscopy, PB - John Wiley & Sons, Ltd 2006.
- [4] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, *Bioinformatics.* 23 (2007) 2507-2517.
- [5] P.R. Griffiths, J.A. de Haseth, Fourier Transform Infrared Spectroscopy, (Second Edition), John Wiley and Sons, New York, 2007.
- [6] W. Herres, J. Gronholz, Understanding FT-IR data processing, Reprint Vols. 1(84), 3(85) Dr. Alfred Huethig Publishers (obtainable from Bruker Optics, Rheinstetten).
- [7] D. Naumann, FT-IR spectroscopy of microorganisms at the Robert Koch Institute: experiences gained during a successful project, in: A. Mahadevan-Jansen, W. Petrich; R.R. Alfano; A. Katz, (Eds), Biomedical Optical Spectroscopy, Proceedings of SPIE, 2008, 68530F.
- [8] P. Lasch P, W. Petrich, Data Acquisition and Analysis in Biomedical Vibrational Spectroscopy, in: D. Moss (Ed), Biomedical Applications of Synchrotron Infrared Microspectroscopy: A Practical Approach, *RSC Analytical Spectroscopy Series* v. 11, 2011, pp. 192-225.
- [9] <http://www.cytospec.com> [March 1, 2012].
- [10] F.W. Dalby, H.H. Nielsen, Infrared Spectrum of Water Vapor. Part I—The 6.26  $\mu\text{m}$  Region, *J. Chem. Phys.* 25 (1956), 934-940.
- [11] H. Fabian, C.P. Schultz, D. Naumann, O. Landt, U. Hahn, W. Saenger, Secondary structure and temperature-induced unfolding and refolding of ribonuclease T1 in aqueous solution. A Fourier transform infrared spectroscopic study, *J. Mol. Biol.* 232 (1993) 967-81.
- [12] H. Fabian, C.P. Schultz, Fourier transform infrared spectroscopy in peptide and protein analysis, in: R.A. Meyers (Ed.), Encyclopedia of Analytical Chemistry, John Wiley & Sons Ltd, Chichester, 2000, pp. 5779-5803.
- [13] P. Lasch, C.P. Schultz, D. Naumann, The Thermotropic and Ionotropic Phase Behaviour of Mixed Vesicles Made up from Lipopolysaccharides, 1,2-Dimyristoylphosphatidylethanolamine-d54, and OmpF from Escherichia coli. FT-IR-spectroscopic Investigations, *Biophys. J.* 75 (1998) 840-852.
- [14] P. Lasch, T. Petras, O. Ullrich, J. Backmann, D. Naumann, T. Grune, Hydrogen Peroxide Induced Damage of RNase A is Followed by Degradation by Proteasome and FT-IR Spectroscopy, *J. Biol. Chem.* 276 (2001) 9492-9502.
- [15] P. Lasch, L. Chiriboga, H. Yee, M. Diem, Infrared Spectroscopy of Human Cells and Tissue: Detection of Disease, *Technol. Cancer Res. Treat.* 1 (2002) 1-5.
- [16] P. Lasch, M. Diem, W. Hänsch, D. Naumann, Artificial Neural Networks as Supervised Techniques for FT-IR Microspectroscopic Imaging, *J. Chemometrics* 20 (2006) 209-220.
- [17] T.W. Randolph, Scale-based normalization of spectral data, *Cancer Biomark.* 2 (2006) 135-44.
- [18] P. Geladi, D. MacDougall, H. Martens, Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat, *Appl. Spectrosc.* 39 (1985) 377-562.
- [19] H. Martens, J.P. Nielsen, S.B. Engelsen, Light scattering and light absorbance separated by extended multiplicative signal correction. application to near-infrared transmission analysis of powder mixtures, *Anal. Chem.* 75 (2003) 394-404.
- [20] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra, *Appl. Spectrosc.* 43 (1989) 772-777.
- [21] V. Mazet, C. Carteret, D. Brie, J. Idier, B. Humbert. Background removal from spectra by designing and minimising a non-quadratic cost function, *Chemom. Intell. Lab. Syst.* 76 (2005) 121-133.
- [22] G. Mie. Contribution to the optical properties of turbid media, in particular of colloidal suspensions of metals, *Ann. Phys (Leipzig).* 25 (1908) 377-452.
- [23] B. Mohlenhoff, M. Romeo, M. Diem, B.R. Wood, Mie-type scattering and non-B Beer-Lambert absorption behavior of human cells in infrared microspectroscopy, *Biophys. J.* 88 (2005) 3635-40.
- [24] A. Savitzky and M. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures, *Anal. Chem.* 36 (1964) 1627.
- [25] G. Schulze, A. Jirasek, M.M. Yu, A. Lim, R.F. Turner, M.W. Blades, Investigation of selected baseline removal techniques as candidates for automated implementation, *Appl. Spectrosc.* 59 (2005) 545-74.
- [26] P. Marchand, L. Marmet, Binomial smoothing filter: A way to avoid some pitfalls of least square polynomial smoothing, *Rev. Sci. Instrum.* 54 (1983) 1034-41.

- [27] F. Holler, D.H. Burns, J.B. Callis, Direct Use of Second Derivatives in Curve-Fitting Procedures, *Appl. Spectrosc.* 43 (1989) 877-882.
- [28] J. Steinier, Y. Termonia, J. Deltour, Smoothing and differentiation of data by simplified least square procedure, *Anal Chem.* 44 (1972) 1906-1909.
- [29] [http://en.wikipedia.org/wiki/Savitzky-Golay\\_smoothing\\_filter](http://en.wikipedia.org/wiki/Savitzky-Golay_smoothing_filter) [November 1, 2011].
- [30] H. Stone, Mathematical resolution of overlapping bands, *J. Opt. Soc. Am.* 52 (1962) 998-1003.
- [31] J.K. Kauppinen, D.J. Moffat, H.H. Mantsch, D.G. Cameron, Fourier Self-Deconvolution: a method for resolving intrinsically overlapped bands, *Appl. Spectrosc.* 35 (1981) 271-276.
- [32] J.K. Kauppinen, D.J. Moffat, H.H. Mantsch, D.G. Cameron, Self-deconvolution and first order derivatives using Fourier transforms, *Anal Chem.* 53 (1981) 1454-1457.
- [33] J.K. Kauppinen, D.J. Moffat, H.H. Mantsch, D.G. Cameron, Noise in Fourier self-deconvolution, *Appl Optics.* 20 (1981) 1866.
- [34] D.J. Moffat, J.K. Kauppinen, H.H. Mantsch, Computers and infrared spectroscopy: evolution and revolution, *Can J Chem.* 69 (1991) 1781-1785.
- [35] W.J. Yang, P.R. Griffiths, M. Byler, H. Susi, Protein conformation by infrared spectroscopy: resolution enhancement by Fourier self-deconvolution, *Appl. Spectrosc.* 39 (1985) 281-28.
- [36] W.K. Surewicz, H.H. Mantsch, D. Chapman, Determination of protein secondary structure by Fourier transform infrared spectroscopy: a critical assessment, *Biochemistry.* 32 (1993) 389-394.
- [37] A. Barth, C. Zscherp, What vibrations tell us about proteins, *Q. Rev. Biophys.* 35 (2002) 369-430.
- [38] V.A. Lórenz-Fonfría, E. Padrós, The role and selection of the filter function in Fourier self-deconvolution revisited, *Appl. Spectrosc.* 63 (2009) 791-9.
- [39] L. Smeller, K. Goossens, K. Heremans, How to Minimize Certain Artifacts in Fourier Self-Deconvolution, *Appl. Spectrosc.* 49 (1995) 1538-1542.
- [40] P. Bassan, H.J. Byrne, J. Lee, F. Bonnier, C. Clarke, P. Dumas, E. Gazi, P. Gardner, Reflection contributions to the dispersion artefact in FTIR spectra of single biological cells, *Analyst* 134 (2009) 1171-1175.
- [41] P. Bassan, H.J. Byrne, F. Bonnier, J. Lee, P. Dumas, P. Gardner, Resonant Mie scattering in infrared spectroscopy of biological materials - Understanding the 'dispersion artefact', *Analyst* 134 (2009) 1586-1593.
- [42] P. Bassan, A. Kohler, H. Martens, J. Lee, H.J. Byrne, P. Dumas, E. Gazi, P. Gardner, Resonant Mie Scattering (RMieS) correction of infrared spectra from highly scattering biological samples, *Analyst* 135 (2010) 268-277.
- [43] P. Bassan, A. Kohler, H. Martens, J. Lee, E. Jackson, N. Lockyer, P. Dumas, P. Gardner, RMieS-EMSC correction for infrared spectra of biological cells: Extension using full Mie theory and GPU computing, *J. Biophotonics* 3 (2010) 609-620.
- [44] J.J. Max, C. Chapados, Isotope effects in liquid water by infrared spectroscopy, *J. Chem. Phys.* 116 (2002) 4626-4642.
- [45] Z.D. Zhang, D. Ben-Amotz, Removal of Cosmic Spikes from Hyper-spectral Images Using a Hybrid Upper-Bound Spectrum Method, *Appl. Spectrosc.* 56 (2002) 91-98.
- [46] C.L. do Lago, V.F. Juliano, C. Kascheres, Applying moving median digital filter to mass spectrometry and potentiometric titration, *Analytica Chimica Acta*, 310 (1995) 281-288.
- [47] J. Zhao, Image curvature correction and cosmic removal for high-throughput dispersive Raman spectroscopy, *Appl. Spectrosc.* 57 (2003) 1368-75.
- [48] G.R. Phillips, J.M. Harris, Polynomial filters for data sets with outlying or missing observations: application to charge-coupled-device-detected Raman spectra contaminated by cosmic rays, *Anal Chem.* 62 (1990) 2351-2357.
- [49] S Li, L Dai, An Improved Algorithm to Remove Cosmic Spikes in Raman Spectra for Online Monitoring, *Appl. Spectrosc.* 65 (2011) 1300-1306.
- [50] W. Hill, D. Rogalla, Spike-correction of weak signals from charge-coupled devices and its application to Raman spectroscopy, *Anal. Chem.* 64 (1992) 2575-2579.
- [51] A.C. De Luca, M. Mazilu, A. Riches, C.S. Herrington, K. Dholakia, Online fluorescence suppression in modulated Raman spectroscopy, *Anal. Chem.* 82 (2010) 738-745.
- [52] P.A. Mosier-Boss, S.H. Lieberman, R. Newbery, Fluorescence Rejection in Raman Spectroscopy by Shifted-Spectra, Edge Detection, and FFT Filtering Techniques, *Appl. Spectrosc.* 49 (1995) 630-638.
- [53] K. Kamogawa, T. Fujii, T. Kitagawa, Improved Fluorescence Rejection in Measurements of Raman Spectra of Fluorescent Compounds, *Appl. Spectrosc.* 42 (1988) 248-254.
- [54] D. Rohleder, W. Kiefer, W. Petrich, Quantitative analysis of serum and serum ultrafiltrate by means of Raman spectroscopy, *Analyst.* 129 (2004) 906-11.
- [55] K.P.J. Williams, D.L. Gerrard, The use of 752.5 and 799.3 nm laser excitation in Raman spectroscopy, *Opt. Laser Technol.* 17 (1985) 245-248.
- [56] S.P.A. Fodor, T.G. Spiro, Ultraviolet resonance Raman spectroscopy of DNA with 200-266-nm laser excitation, *J. Am. Chem. Soc.* 108 (1986) 3198-3205.
- [57] R.P. Van Duyne, D.L. Jeanmaire, D. F. Shriver, Mode-Locked Laser Raman Spectroscopy-A New Technique for the Rejection of Interfering Background Luminescence Signals, *Anal Chem.* 46 (1974) 213-222.

- [58] M.D. Morris, P. Matousek, M. Towrie, A.W. Parker, A.E. Goodship, E.R. Draper, Kerr-gated time-resolved Raman spectroscopy of equine cortical bone tissue, *J Biomed Opt.* 10 (2005) 14014.
- [59] H.J. Bowley, D.L. Gerrard, The use of diode array detectors in conjunction with continuous wave gas lasers for anti-Stokes Raman spectroscopy, *Opt. Laser Technol.* 18 (1986) 33-35.
- [60] S.M. Angel, M.K. DeArmond, K.W. Hanck, D.W. Wertz, Computer-controlled instrument for the recovery of a resonance Raman spectrum in the presence of strong luminescence, *Anal Chem.* 56 (1984) 3000-3001.
- [61] A.P. Shreve, N.J. Cherepy, R.A. Mathies, Effective Rejection of Fluorescence Interference in Raman Spectroscopy Using a Shifted Excitation Difference Technique, *Appl. Spectrosc.* 46 (1992) 707-711.
- [62] J. Zhao, M.M. Carrabba, F.S. Allen, Automated Fluorescence Rejection Using Shifted Excitation Raman Difference Spectroscopy, *Appl. Spectrosc.* 56 (2002) 834-845.
- [63] A. Mahadevan-Jansen, R.R. Richards-Kortum, Raman spectroscopy for the detection of cancers and precancers, *J. Biomed. Opt.* 1 (1996) 31-70.
- [64] C.A. Lieber, A. Mahadevan-Jansen, Automated method for subtraction of fluorescence from biological Raman spectra, *Appl Spectrosc.* 57 (2003) 1363-7.
- [65] J. Zhao, H. Lui, D.I. McLean, H. Zeng, Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. *Appl. Spectrosc.* 61 (2007) 1225-1232.
- [66] D. Zhang, D. Ben-Amotz, Enhanced Chemical Classification of Raman Images in the Presence of Strong Fluorescence Interference, *Appl. Spectrosc.* 54 (2000) 1379-1383.
- [67] T. Hasegawa, J. Nishijo, J. Umemura, Separation of Raman spectra from fluorescence emission background by principal component analysis, *Chem Phys Letters.* 317 (2000) 642-646.
- [68] Y. Hua, T. Jiang, A. Shen, W. Lib, X. Wang, J. Hub, A background elimination method based on wavelet transform for Raman spectra. *Chemom. Intell. Lab. Syst.* 85 (2007) 94-101.
- [69] Z.M. Zhang, S. Chen, Y.Z. Liang, Z.X. Liu, Q.M. Zhang, L.X. Ding, F. Ye, H. Zhou, An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy, *J Raman Spectrosc.* 41 (2009) 659-669.
- [70] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Trans. Pattern Anal. Machine Intell.* 11 (1989). 674-693.
- [71] A. Graps, An introduction to wavelets, *IEEE Comput. Sci. Eng.* 2 (1995) 50-61.
- [72] Y. Mallet, D. Coomans, O. De Vel, Recent developments in discriminant analysis on high dimensional spectral data, *Chemometr Intell Lab Syst.* 35 (1996) 157-173.
- [73] B.K. Alsberg, A.M. Woodward, D.B. Kell, An introduction to wavelet transforms for chemometricians: a time-frequency approach, *Chemometr Intell Lab Syst.* 37 (1997) 215-239.
- [74] B. Walczak, D.L. Massart, Noise suppression and signal compression using the wavelet packet transform, *Chemometr. Intell. Lab. Syst.* 36 (1997) 81-94.
- [75] B. Walczak, B. Bogaert, D.L. Massart, Application of wavelet packet transform in pattern recognition of near-IR data, *Anal Chem.* 68 (1996) 1742-1747.
- [76] C. Gobinet, V. Vrabie, M. Manfait, O. Piot, Preprocessing methods of Raman spectra for source extraction on biomedical samples: application on paraffin-embedded skin biopsies, *IEEE Trans Biomed Eng.* 56 (2009) 1371-82.
- [77] S.T. Wollman, P.W. Bohn, Evaluation of Polynomial Fitting Functions for Use with CCD Arrays in Raman Spectroscopy, *Appl Spectrosc.* 47 (1993) 125-126.
- [78] J.V. Simpson, O. Oshokoya, N. Wagner, J. Liu, R.D. JiJi, Pre-processing of ultraviolet resonance Raman spectra, *Analyst.* 136 (2011) 1239-47.
- [79] G.R. Hopkinson, T.M. Goodman, S.R. Prince SR, A guide to the use and calibration of detector array equipment, *SPIE - The International Society for Optical Engineering*, Bellingham, Washington 2004.
- [80] Automated wavelength & intensity calibration routines significantly improves accuracy of recorded spectra, Princeton Instruments Spectroscopy Group. Technical Note. 2011.
- [81] C.H. Tseng, J.F. Ford, C.K. Mann, T.J. Vickers, Wavelength Calibration of a Multichannel Spectrometer, *Appl. Spectrosc.* 47 (1993) 1808-1813.
- [82] <http://www.astm.org> [November 1, 2011].
- [83] D.A. Carter, W.R. Thompson, R. Wade, E.T. Chad, J.E. Pemberton, Frequency/Wavelength Calibration of Multipurpose Multichannel Raman Spectrometers. Part II: Calibration Fit Considerations and Calibration Standards, *Appl. Spectrosc.* 49 (1995) 1561-1576.
- [84] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, *J. Chromatography A* 805 (1998) 17-35.
- [85] G. Tomasi, F. van den Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, *J. Chemometrics.* 18 (2004) 231-241.
- [86] I. Haq, B.Z. Chowdhry, J.B. Chaires, Singular value decomposition of 3-D DNA melting curves reveals complexity in the melting process, *Eur. Biophys. J.* 26 (1997) 419-426.
- [87] H.J. van Manen, M. Kraan, Y.D. Roos, C. Otto, Intracellular Chemical Imaging of Heme-Containing Enzymes Involved in Innate Immunity Using Resonance Raman Microscopy, *J. Phys. Chem. B.* 48 (2004) 18762-18771.

- [88] D.D. Turner, R.O. Knuteson, H.E. Revercomb, C. Lo, R.G. Dedecker, Noise reduction of atmospheric emitted radiance interferometer (AERI) observations using principal component analysis, *J. Atmos. Oceanic Technol.* 23 (2006) 1223-1238.
- [89] A.A. Green, M. Berman, P. Switzer, M.D. Craig, A Transformation for Ordering Multispectral Data in Terms of Image Quality with Implications for Noise Removal, *IEEE Transactions on Geoscience and Remote Sensing*, 26 (1988) 65-74.
- [90] A.A. Nielsen, Analysis of Regularly and Irregularly Sampled Spatial, Multivariate, and Multi-temporal Data, Dissertation, IMM/Technical University of Denmark, Lyngby 1994.
- [91] F.A. Kruse, J.W. Boardman, A.B. Lefkoff, Extraction of compositional information for trafficability mapping from hyperspectral data, *Proc. SPIE*. 4049 (2000) 262.
- [92] Bhargava R, Wang SQ, Koenig JL. Route to Higher Fidelity FT-IR Imaging, *Appl. Spectrosc.* 54 (2000) 486-495.
- [93] R.K. Reddy, R. Bhargava, Accurate histopathology from low signal-to-noise ratio spectroscopic imaging data. *Analyst.* 135 (2010) 2818-25.
- [94] J.W. Boardman, F.A. Kruse, Automated spectral analysis: A geologic example using AVIRIS data, north Grapevine Mountains, Nevada: in Proceedings of the Tenth Thematic Conference on Geologic Remote Sensing, Environmental Research Institute of Michigan, Ann Arbor, MI, 1994, Vol 1: pp. 407-418.
- [95] T.T. Cai, D. Zhang, D. Ben-Amotz, Enhanced Chemical Classification of Raman Images Using Multiresolution Wavelet Transformation, *Appl. Spectrosc.* 55 (2001) 1124-1130.
- [96] Y.P. Wang, Y. Wang, P. Spencer, Fuzzy clustering of Raman spectral imaging data with a wavelet-based noise-reduction approach, *Appl. Spectrosc.* 60 (2006) 826-832.
- [97] <http://www.andor.com> [March 1, 2012].
- [98] P. Lasch, W. Haensch, D. Naumann, M. Diem, Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis, *Biochim Biophys Acta.* 1688 (2004) 176-86.
- [99] P. Lasch, D. Naumann, Spatial resolution in infrared microspectroscopic imaging of tissues, *Biochim Biophys Acta.* 1758 (2006) 814-29.
- [100] C.J. Behrend, C.P. Tarnowski, M.D. Morris, Identification of Outliers in Hyperspectral Raman Image Data by Nearest Neighbor Comparison, *Appl. Spectrosc.* 56 (2002) 1458-1461.
- [101] H. Nyquist, Certain topics in telegraph transmission theory, *AIEE Trans.* 47 (1928) 617-644.
- [102] M.J. Nasse, M.J. Walsh, E.C. Mattson, R. Reininger, A. Kajdacsy-Balla, V. Macias, R. Bhargava, C.J. Hirschmugl, High-resolution Fourier-transform infrared chemical imaging with multiple synchrotron beams, *Nat. Methods.* 8 (2011) 413-416.
- [103] G.L. Carr, Resolution limits for infrared microspectroscopy explored with synchrotron radiation, *Rev. Sci. Instrum.* 72 (2001) 1613-1619.
- [104] K.C. Schuster, I. Reese, E. Urlaub, J.R. Gapes, B. Lendl, Multidimensional information on the chemical composition of single bacterial cells by confocal Raman microspectroscopy, *Anal Chem.* 72 (2000) 5529-5534.
- [105] P. Rösch, M. Harz, M. Schmitt, K.D. Peschke, O. Ronneberger, H. Burkhardt, H.W. Motzkus, M. Lankers, S. Hofer, H. Thiele, J. Popp, Chemotaxonomic identification of single bacteria by micro-Raman spectroscopy: application to clean-room-relevant biological contaminations, *Appl. Environ. Microbiol.* 71 (2005) 1626-37.
- [106] A. Hermelink, A. Brauer, P. Lasch, D. Naumann. Imaging phenotypic heterogeneity within microbial populations at the single cell level by confocal Raman microspectroscopy, *Analyst.* 134, (2009) 1149-53.
- [107] P. Lasch, A. Hermelink, D. Naumann, Correction of axial chromatic aberrations in confocal Raman microspectroscopic measurements of a single microbial spore, *Analyst.* 134 (2009) 1162-70.
- [108] P. Lasch, J. Schmitt, M. Beekes, T. Udelhoven, M. Eiden, H. Fabian, W. Petrich, D. Naumann, Antemortem Identification of Bovine Spongiform Encephalopathy from Serum Using Infrared Spectroscopy, *Anal. Chem* 75 (2003) 6673-6678
- [109] L. Xu, Y.P. Zhou, L.J. Tang, H.L. Wu, J.H. Jiang, G.L. Shen, R.Q. Yu, Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration, *Anal Chim Acta.* 616 (2008) 138-43.
- [110] S.R. Delwiche, J.B. 3rd. Reeves, A graphical method to evaluate spectral preprocessing in multivariate regression calibrations: example with Savitzky-Golay filters and partial least squares regression, *Appl Spectrosc.* 64 (2010) 73-82.
- [111] S.R. Delwiche, R.A. Graybosch, Examination of spectral pretreatments for partial least-squares calibrations for chemical and physical properties of wheat. *Appl Spectrosc.* 57 (2003) 1517-27.
- [112] A. Candolfi, R. De Maesschalck, D. Jouan-Rimbaud, P.A. Hailey, D.L. Massart, The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra. *J Pharm Biomed Anal.* 21 (1999) 115-32.
- [113] N. K. Afseth, V. H. Segtnan, J. P. Wold, Raman Spectra of Biological Samples: A Study of Preprocessing Methods, *Appl Spectrosc.* 60 (2006) 1358-1367.
- [114] D. Helm, H. Labischinski, G. Schallehn, D. Naumann, Classification and identification of bacteria by Fourier-transform infrared spectroscopy, *J Gen Microbiol.* 137 (1991) 69-79.
- [115] T. Udelhoven, M. Novozhilov, J. Schmitt, The NeuroDeveloper(R): a tool for modular neural classification of spectroscopic data, *Chemom. Intell. Lab. Syst.* 66 (2003) 219-226.

















