Oleg Ryabchykov[1] / Shuxia Guo[1] / Thomas Bocklitz[1]

Analyzing Raman spectroscopic data

[1] Institute of Physical Chemistry and Abbe Center of Photonics (IPC), Friedrich-Schiller-University, Jena, Germany, E-mail: Oleg.Ryabchykov@uni-jena.de, Shuxia.Guo@uni-jena.de, Thomas.Bocklitz@uni-jena.de

Abstract:
This chapter is a short introduction into the data analysis pipeline, which is typically utilized to analyze Raman spectra. We empathized in the chapter that this data analysis pipeline must be tailored to the specific application of interest. Nevertheless, the tailored data analysis pipeline consists always of the same general procedures applied sequentially. The utilized procedures correct for artefacts, standardize the measured spectral data and translate the spectroscopic signals into higher level information. These computational procedures can be arranged into separate groups namely data pre-treatment, pre-processing and modeling. Thereby the pre-treatment aims to correct for non-sample-dependent artefacts, like cosmic spikes and contributions of the measurement device. The block of procedures, which needs to be applied next, is called pre-processing. This group consists of smoothing, baseline correction, normalization and dimension reduction. Thereafter, the analysis model is constructed and the performance of the models is evaluated. Every data analysis pipeline should be composed of procedures of these three groups and we describe every group in this chapter. After the description of data pre-treatment, pre-processing and modeling, we summarized trends in the analysis of Raman spectra namely model transfer approaches and data fusion. At the end of the chapter we tried to condense the whole chapter into guidelines for the analysis of Raman spectra.

Keywords: Raman spectroscopic data analysis, spectral preprocessing, spectral standardization, machine learning for spectral data, data analysis workflow

DOI: 10.1515/psr-2017-0043

1 General analysis pipeline

Since the early 70s the potential of Raman spectroscopy for the characterization of biological samples like DNA, proteins and lipids was recognized. Nevertheless, it took since the 2000s until the potential could be utilized. One reason was the development of components needed for the application of Raman spectroscopy for biological samples (see chapter 2). The other reason is that powerful statistical and computational methods are needed in order to translate the Raman spectral signals into meaningful bio-medical information. To do so powerful computers are needed, which can deal with large data sets. Additionally, tailored data analysis pipelines for the analysis of Raman spectra must be developed [1], which allow the application of Raman spectroscopy for real-world tasks. Possible applications include crystals and minerals (chapter 5), pharmacy (chapter 6), fine particles (chapter 7), medicine (chapter 8) archeological investigations (chapter 9) and forensics (chapter 10).

The data analysis pipeline must be tailored to the specific application of interest and is composed of computational procedures to correct for artefacts, standardize the measured spectral data and to translate the spectroscopic signals into higher level information. The procedures can be arranged in separate groups namely data pre-treatment, pre-processing and modeling. The pre-treatment aims to correct for non-sample-dependent artefacts, like spikes and contributions of the measurement device. This will be described in Section 2. The next block of procedures to be applied is called pre-processing and include smoothing, baseline correction, normalization and dimension reduction. These methods are described in Section 3. Thereafter, the analysis model is constructed and it is evaluated, which will be described in Section 4. Every data analysis pipeline is composed of procedures of these three groups and the whole data pipeline to analyze Raman spectra is sketched in Figure 1. In Section 5 we will shortly describe new trends in the analysis of Raman spectra and in Section 6 a summary will be given, which focuses on dos and don'ts.

Thomas Bocklitz is the corresponding author.
Oleg Ryabchykov and Shuxia Guo share the main authorship.

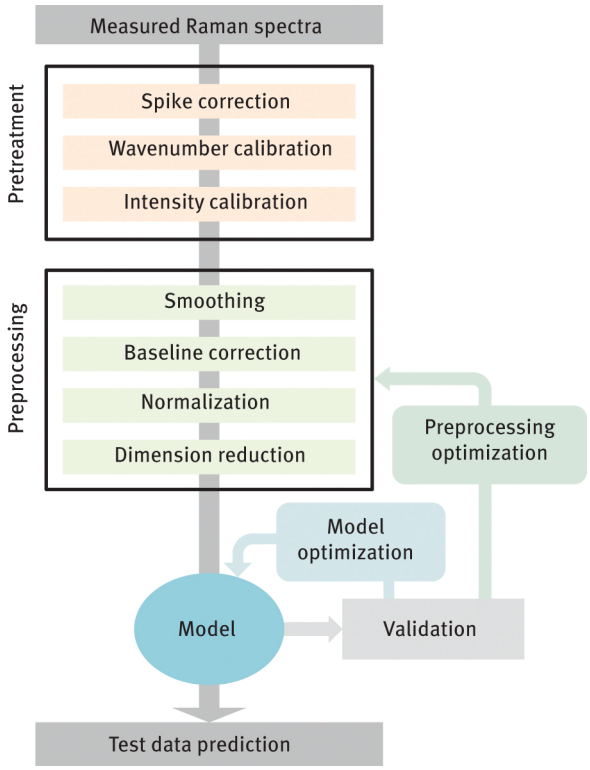

Figure 1: Data analysis pipeline for Raman spectra. A data pipeline used to analyze Raman spectroscopic data is shown. It is composed of pre-treatment, pre-processing and analysis procedures. The pretreatment steps remove corrupting effects which are not related to the sample and the preprocessing steps standardize the data by removing sample related contributions from the data. At the end of the pipeline, statistical models or machine learning approaches are constructed. These models are evaluated and there may be a parameter optimization based on the model outcome. All these steps aim a robust prediction of the constructed model.

2 Data pretreatment

Like described earlier, the analysis of Raman spectra starts with a pre-treatment of the measured Raman spectra, which is necessary because the measured Raman spectra contain disturbing contributions and artefacts. These contributions corrupt the spectral information of the sample and prevent a reliable analysis. Thus, correction procedures need to be carried out. The most disturbing contributions within Raman spectra, which are not sample dependent, are cosmic spikes and contributions caused by experimental parameters and/or the measurement device itself. In order to deal with these contributions a spike correction, a wavenumber and an intensity calibration are needed. These methods are described in the following subsections. The description starts with the spike correction, which should be carried out at the beginning of the data pipeline for Raman spectra. Thereafter a wavenumber calibration needs to be done and an optional intensity calibration might be performed.

2.1 Spike correction

In contrast to other corrupting effects in Raman spectroscopy, a presence of cosmic ray spikes in the spectroscopic data does not depend on the sample, laser, or spectrometer. Spikes appear at random positions in the data as values with large intensities (Figure 4). They occur when high-energetic cosmic particles hit the detector. In the CCD these particles generate electrons, which are read out along with the charges induced by the energy of Raman scattered photons. Therefore, cosmic particles introduce narrow features of high intensity, called spikes. Their positions are random and do not correspond to the wavenumbers directly. Typically, a spike appears within just one pixel or a few pixels.

The cosmic ray noise may affect the subsequent analysis, especially the normalization step. Therefore, many commercial devices perform multiple measurements in order to calculate an average spectrum. This approach applies when a small data set is acquired because it decreases all types of noise, including spikes. Unfortunately,

it leads to the measurement time increase that is unsuitable for the acquisition of large spectral maps. Another method which works for small data sets is a manual removal of spikes found by visual inspection of the spectra. A drawback of manual spikes removal is that massive data sets cannot be processed by an operator within a reasonable time. To overcome this issue and automate the spike removal in massive data sets, specialized computational approaches were developed.

The simplest spike correction procedure is a median smoothing of the spectra. Unfortunately, besides spikes, this method filters and changes sharp spectral bands. Another option is a wavelet or Gabor transform with a suppressing of the coefficients corresponding to the spikes. Because spectral bands and spikes share the same frequencies, this approach can also corrupt the spectral information [2].

Besides application as filtering methods, wavelet or Gabor transform can be used to obtain a quantitative marker, related to the sharpness of features in the spectra. Then, spikes can be detected based on that marker and eliminated. Similarly, the marker can be obtained from nearest neighbors comparisons within the spectrum, or by applying a discrete Laplace operator [3]. If time series or scans are analyzed, the changes from spectrum to spectrum are typically small and the spikes detection methods can utilize the extra dimensions. Hence, 2-dimensional wavelets, comparison with the nearest neighbors within the spectral matrix or 2-dimensional Laplacian operator can be applied to enhance the reliability of the cosmic ray noise markers.

After extracting the quantitative marker, the spikes need to be located by setting a threshold. To estimate the threshold, the marker values can be compared to their standard deviation. For more robust comparison, the lowest and largest values may be excluded from the calculation of the standard deviation. Typically, a spike is considered to be present, if the response is higher than a preset threshold like three times the standard deviation. However, for some noise characteristics and sharpness of the Raman bands, this threshold may not be optimal. To optimize cosmic ray noise removal, the threshold should be selected depending on the distribution of the marker values [4].

2.2 Spectrometer calibration

The next step in the pre-treatment of Raman spectra is the spectrometer calibration. Ideally, the Raman spectroscopic signals should have high reproducibility and consistency. That means the measurement should be independent of the device and its instrumental configurations. In reality, however, a recorded Raman spectrum is affected by measurement conditions and does not solely reflect the sample. Variations in instruments, temperature, physical and chemical states of samples (e. g. solid or liquid) can lead to substantial spectral changes like wavenumber shifts and intensity variations [5, 6]. Furthermore, the instrumental response function changes over time. Raman spectra measured with a time delay can be different, even if they are measured on identical samples and under the same conditions. A calibration procedure is often required to reduce the introduced spectral changes, which includes wavenumber calibration and intensity calibration, as described in the following [5–7].

2.2.1 Wavenumber calibration

A CCD detector is typically used in a Raman spectrometer to collect Raman scattered photons at different frequencies (i. e. wavenumbers) with a different pixel (Figure 7a). Each spectrometer has a defined relation between wavenumber and pixel position [8]. However, the relation is sensitive to environmental and instrumental changes, like variations of temperature, the replacement of an instrumental component, or drifts of the instrument over-time. As a consequence, a CCD pixel can correspond to a different wavenumber. Hence the Raman spectrum is not recorded with the correct wavenumber axis, which manifests itself as wavenumber shifts compared to the theoretical values. This is shown by the two spectra in Figure 7a. The wavenumber shifts make it problematic to compare Raman spectra measured under different conditions or analyze them together.

A wavenumber calibration is conducted to find the correct wavenumber axis and thus remove the wavenumber shifts. To do so, a standard material with well-defined Raman bands is measured before measuring real samples. As shown in Figure 7b, the positions of these known Raman bands are located on the measured Raman spectrum. Thereafter the wavenumber differences between the observed and theoretical Raman bands are calculated. Based on these differences a parametric function is fitted and interpolated to all recorded wavenumber positions forming the wavenumber axis. Provided the Raman spectra of the standard material and the sample spectra share the same wavenumber axis, the wavenumber axis of the Raman spectra can be corrected by the obtained correction function, which removes the undesired wavenumber shifts.

The precision of wavenumber calibration is dependent on several factors [7]. Among those a careful selection of the standard material is highly important. The known Raman bands of the standard material need to be

densely distributed and cover the whole spectral range of interest. For biological applications the bands of the standard material should cover the fingerprint region and the CH-stretching region. Additionally, the number of Raman bands needs to be sufficient to stably interpolate the correction function to the whole spectral range. Materials that can be used for wavenumber calibration and their tabulated Raman bands are available in [9, 10]. One widely applied example in biological studies is 4-acetamedophenol (Figure 7b).

Other influential factors for a precise wavenumber calibration are the quality of the peak searching and the subsequent fitting of the correction function. For the first aspect, the wavenumber positions of tabled Raman bands are located by interrogating a neighborhood of a given Raman band. The result can be the peak point of the neighborhood or the peak point of a Gaussian or Lorenz curve fitted to this neighborhood. For the second aspect, the correction function is typically fitted as a polynomial with a degree of three to five. Polynomials with higher degrees are not recommended.

2.2.2 Intensity calibration

The intensity of a recorded Raman spectrum is the product of the true Raman scattered intensity (including baseline intensities) and the intensity response function of an instrument (Figure 7c) [7, 11]. In principle, quantitative and qualitative studies can be performed without considering this fact as long as the intensity response function remains unchanged over the measurements. However, the intensity response function does change with instrumental and environmental factors like excitation wavelength change, detector replacement, changes of the sampling geometry, temperature, and so forth. The intensity response function of the same instrument can also change over time. Consequently, the intensities of recorded Raman spectra vary from instrument to instrument, condition to condition, and time to time. Such intensity variations can be ignored in most qualitative studies. For quantitative analyses and the comparison with a spectral library, the influence of the intensity variations becomes a critical issue and has to be corrected.

Therefore, intensity calibration is required, which corrects the recorded Raman spectrum with the intensity response function. Similar to wavenumber calibration, the intensity calibration also needs a standard material with known emission at different frequencies. The intensity response function is considered unchanged during the measurements of the standard material and actual samples, given they are measured under the same condition. As illustrated in Figure 7c and d, the procedure includes three steps. (1) The emission of a standard material is measured under the same condition as for the samples of interest. (2) The intensity response function is calculated as the ratio between the measured and theoretical emission of this standard material. (3) The Raman intensities of actual samples are corrected by dividing with the calculated intensity response function [5].

The efficiency of intensity calibration largely relies on the standard material. An ideal standard material should be homogeneous and give reproducible emission over broad wavenumber range. Existent standard materials (SRM) can be either a black-body radiator or a luminescence standard; both are available at NIST (National Institute of Standards & Technology, Gaithersburg, MD, USA) [12]. Black-body radiators are less often employed due to stability issues and difficulties to duplicate the sampling condition. The luminescence standards are more widely utilized in Raman spectroscopy. More details are beyond the scope of this chapter and interested readers are referred to [7, 11].

Above all, spectrometer calibration is proven to improve the results of the statistical analysis for datasets measured under different conditions, thanks to the improved spectral consistency over different measurements [13]. However, calibration cannot completely remove all undesired conditional relevant spectral variations [6, 14], which originates from multiple reasons including inaccessibility of a perfect standard material and unavoidable changes for measuring the standard material and the samples. The remaining spectral variations can still be disturbing for subsequent analysis and have to be handled. This leads to the topic of model transfer, which will be described in Section 5.1.

3 Data preprocessing

After the pre-treatment is carried out and artefacts of the measurement are corrected for, the pre-processing needs to be performed [15]. In this part sample related artifacts and sample dependent spectral contributions are corrected, leading to a standardization of the spectra. The most important correction procedure is a baseline correction, because the fluorescence background might be orders of magnitude stronger than the Raman signal. Before (or after) the baseline correction a smoothing can be carried out to correct noise contributions, but this is rather an optional step. Nevertheless, a few baseline correction procedures need a smoothed spectrum to

construct a reliable baseline estimate. After these corrections are carried out, a normalization is performed to statistical standardize the spectra and a dimension reduction is done. The last step can also be done directly within the statistical model, e. g. some methods doing this implicitly. Nevertheless, in principle, it is advisable to work with a lower dimensional representation of the spectra. In the following, we describe all parts of the pre-processing starting with smoothing procedures.

3.1 Smoothing

There are various types of noise that corrupt Raman spectra. They can be categorized into groups according to the source of the noise or according to the appearance in the data. However, as the cosmic ray noise stands out from the other types and it was already described in Section 2.1, we will not discuss it here.

In contrast to cosmic spike noise, the random noise in a Raman spectrum can be additive or intensity dependent. The additive noise has a Gaussian distribution and does not depend on the signal intensity. It corresponds mostly to the detector's dark current and readout noise. On another side, the intensity dependent noise increases with increasing signal intensity and follows the Poisson distribution. To suppress this intensity dependent noise, it is important to plan the experiment in a way to keep the intensity of a fluorescence background low.

Although adjusted measurement conditions can minimize the noise, a completely noise-free spectrum cannot be measured. The noise in measured spectra can affect the baseline correction, normalization and detection of peak positions. To reduce the noise and make the interpretation of spectra easier, spectral smoothing can be applied.

Prominent smoothing procedures are Savitzky-Golay, mean, Gaussian, and median filtering. Each method has its own specifics. The Savitzky-Golay smoothing [16], which is based on the least square fitting, is the most effective in preserving the peaks from corruption. On the other side, mean and Gaussian filters allow an efficient de-noising, and the median filter allows removing outliers from the spectrum. Although any filtering may remove parts of useful spectral information along with the noise, the corruption of the spectra can be avoided completely, if the size of the data set is large enough. In large data sets, a smooth appearance of spectra can be obtained by averaging over a large number of spectra. The influence of noise on the further analysis can also be reduced by a dimension reduction. To preserve the data from corruption, filtering should be avoided if there is no specific reason for filtering and large data sets are analyzed.

3.2 Background correction

There are two different types of baseline correction procedures in Raman spectroscopy: (1) subtracting the signal with the shutter closed from the spectrum and (2) subtracting the mathematically estimated baseline. For further discussion and differentiation between these both methods, the second method is referred to as a baseline correction. The baseline correction is of high importance for standardization of the spectral data when the samples feature a fluorescent background.

Estimating of the fluorescent background mathematically is based on the fact that fluorescence signal is broader than Raman spectral bands. Based on this property, a variety of algorithms for baseline correction were developed. Among the most typical ones (Figure 2a) we can highlight the modified polynomial fit [17], the asymmetric least squares baseline estimation [18], and the statistics-sensitive non-linear iterative peak-clipping (SNIP) algorithm [19]. The last one, in contrast to the others, does not lead to oscillations of the baseline at the edges of the spectral interval. For a better performance of the SNIP baseline algorithm, the estimation should be carried out based on a smoothed spectrum. However, this estimated baseline can be subtracted from the non-smoothed spectrum, preserving the features that could be filtered out by smoothing.

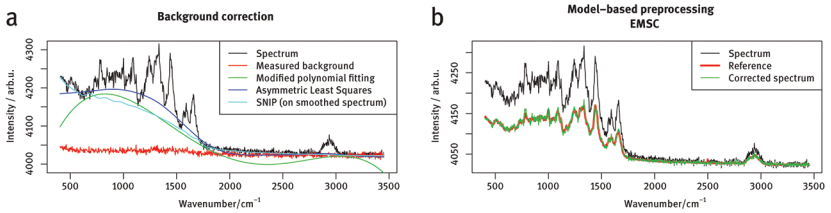

Figure 2: (a) Examples of background estimation. An example spectrum is shown in black color and the measured background signal is shown in red color. The other lines depict estimation of the fluorescence background by means of various baseline correction methods. (b) Model-based preprocessing. The spectrum after the data pretreatment is shown in green. For standardizing the spectra, an extended multiplicative scatter correction (EMSC) can be applied. The reference, which is typically an average spectrum over the data set, is depicted in red color, and the corrected spectrum is shown in green color.

If the variations of the Raman signals within the data set are expected to be small, a model-based preprocessing approach can be used. For example, an extended multiplicative scatter correction (EMSC) [20] is a powerful preprocessing tool that standardizes spectra according to chosen reference spectrum (Figure 2b). An additional advantage of this method is that further normalization is not required and the replicate variations within the data can be taken into account.

Both approaches, namely model-based preprocessing and baseline correction methods, should eliminate the background contribution of the spectra. Therefore, it is highly important to optimize their parameters based on the complexity of the background. Commonly the corrected spectra are investigated visually to estimate the goodness of the correction. A more robust approach is the introduction of a quantitative marker for the quality of baseline correction [21]. This marker should be based on expert knowledge about the spectroscopic data, such as regions where no background is expected and where the Raman spectroscopic signals should be located. If these regions influence the parameter differently, the correction approach, which features the extremum (maximum or minimum value) of the marker, would correspond to the optimal preprocessing.

3.3 Normalization

After the baseline correction, the Raman spectra become more standardized and in some cases can be analyzed directly. Unfortunately, the intensity variations of Raman spectra between investigated samples and even within spectral maps can be dramatic due to the changes in focusing and other experimental factors. An elimination of this effect is possible by applying a normalization step. Out of a huge range of normalization techniques, a few methods are commonly applied: vector normalization, normalization to the integrated spectral intensity, standard normal variate (SNV) and min-max scaling [22].

The vector normalization is performed by dividing the spectrum by the square root of the sum of the squared spectral intensities. It is conceptually similar to the root mean square normalization. This normalization can be also performed separately for different spectral regions, which could be needed in specific cases. Thus, in the analysis of biological samples, such as bacteria or fungi, better performance may be achieved by normalizing the fingerprint region and the CH-stretching region separately. As well as vector normalization, the normalization to the integrated spectral intensity, or area normalization, can be performed separately for different wavenumber regions. Besides that, the entire spectrum can be normalized to the intensity of a specific band, which is stable within the dataset. Furthermore, the l_1-normalization [23] is similar to the area normalization but operates with absolute values. So, in the case of l_1-norm, normalization to the integrated absolute spectral intensity is performed.

The next typical normalization approach is SNV scaling. It is performed by subtracting the mean intensity from the spectra and then dividing the result by the standard deviation of the spectrum. This method removes the constant background from the data. Thus, SNV is suitable to be applied without preliminary baseline correction in cases of a simple constant background.

Another scaling which eliminates the constant background is min-max normalization. It is performed by subtracting the minimum value of the spectrum and then dividing it by the maximum value of the resulting spectrum. This scaling approach is easy to use, but it is more sensitive to noise than other normalization methods.

Alternatively to normalization and scaling approaches, a model-based preprocessing, such as EMSC can be used, which was already mentioned in the Section 3.2 (Figure 2b). This approach does not require additional normalization because it combines both baseline correction and normalization.

3.4 Dimension reduction

Raman spectral datasets are mostly composed of a large number of variables, which poses challenges for a statistical analysis in terms of generalization performance as well as computational effort. A dimension reduction is needed to seek for a lower-dimensional representation of the original dataset without significantly losing key information [24]. The most straightforward way is to choose only the peak positions of Raman bands (probably together with their neighboring data points). This can be combined with a peak fitting procedure. However,

this approach is not applicable if the Raman bands are unknown a priori, which is often the case. Hereby we briefly introduce more advanced approaches that have been widely applied.

Existent dimension reduction approaches can be categorized based on two properties. On the one hand, a lower-dimensional representation can be found with or without the presence of response variables such as group information or concentration of certain chemical components. The respective approaches are termed supervised and unsupervised dimension reduction approaches. On the other hand, the lower-dimensional representation can be based on the same variable space as the original data or based on a transformed space of the original data. Accordingly, the approaches are termed feature selection and feature extraction, respectively. A dimension reduction technique features both aspects, for instance, a supervised feature selection method.

Principal component analysis (PCA) is the most commonly applied unsupervised feature extraction method and the PCA model can be written as $\mathbf{X} = \mathbf{T}\mathbf{V}^T + \mathbf{e}$. The original dataset $\mathbf{X} \in \mathbb{R}^{\boldsymbol{m,q}}$ is mapped onto $r(r = \min(m, q))$ uncorrelated vectors $\boldsymbol{V}_j$, namely the principal components (PC) or loadings. Each PC represents a different source of variances in $\mathbf{X}$, with the largest variance in the first PC, the second largest in the second PC, and so forth. The calculation is achieved by a singular value decomposition on $\mathbf{X}$ or an Eigen value decomposition on the covariance matrix ($\mathbf{X}^T\mathbf{X}$). The loadings $\boldsymbol{V}_j$ $(n \leq j \leq r)$ are usually removed, because they mainly correspond to noise and are irrelevant to further analysis. In this way, the dataset $\mathbf{X}$ is represented by a lower dimensional score matrix ($\mathbf{T}^{m,n}$) and the corresponding error is denoted as $\mathbf{e}$. Besides PCA, other unsupervised feature extraction approaches include independent component analysis and non-negative matrix factorization [25]. In particular, multivariate curve resolution alternating least squares (MCR-ALS) has shown its power in spectral analysis due to its capability of decomposing spectroscopic mixtures into multiple pure components and their concentrations. The concentration maxtrix can be used as scores for subsequent qualitative and quantitative analysis. By applying different constraints like non-negativity, unimodality and local rank, MCR-ALS can provide physically and chemically meaningful decomposition [26, 27]. Another commonly applied dimension reduction method is partial least squares (PLS) modeling. It is a supervised feature extraction method and bears some relation to PCA. Hereby the matrices of predictors ($\mathbf{X} \in \mathbb{R}^{\boldsymbol{m,q}}$) and responses ($\mathbf{Y} \in \mathbb{R}^{\boldsymbol{m,p}}$) are decomposed as $\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{e}_1, \mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{e}_2$. The decomposition is performed so that the covariance between $\mathbf{T}$ and $\mathbf{U}$ is maximized. The dataset $\mathbf{X}$ is transferred into a lower-dimensional score matrix $\mathbf{T}^{m,n}$ $(n < \min(m, q))$ by using the first n latent variables. All these described methods share the similarity that they decompose the observed dataset as a linear combination of N vectors and are called factor methods [28]. Other feature extraction methods like wavelet transform are also used in some investigations [29].

Unlike feature extraction, feature selection works by choosing variables that perform the best according to a predefined metric [24]. It has been well proven that a variable that is completely useless by itself can be significantly useful in combination with other variables. Hence a subset of variables is often selected simultaneously in practice. Approaches of subset selection include three categories: wrapper, filtering, and embedded methods. With wrapper methods, the optimal feature subset is selected to obtain the best prediction on data independent of the training data. Wrapper methods are computationally expensive due to the requirement of model training. Filter methods select a feature subset according to a certain metric that is independent of subsequently applied statistical models, for example, mutual information, Pearson's correlation coefficient, Fisher's discriminant ratio, or results from statistical tests. Filter methods are advantageous in terms of computational cost but they are less powerful to build a good predictive model. Nonetheless, filter methods can be used as a pre-selection prior wrapper methods. Embedded methods conduct feature selection as a part of model construction. This can be achieved by enforcing most coefficients of the model to be zero, like in the cases of LASSO [30] and sparse PLS [31]. In addition, feature selection can also be performed based on the variable weights/importance given by statistical models like support vector machine and random forest (RF). In all cases of feature selection approaches, a search procedure for feature subsets has to be utilized, be it genetic algorithms, simulated annealing or greedy search (i. e. forward/backward feature selection). For wrapper and embedded approaches, a way of assessing the prediction performance has to be known as well, which overlaps with the issue of model selection and is outlined in the following section.

Besides typical factor methods and feature selection, it is worth to note that nonlinear dimension reduction approaches have also been reported, for example, Isomap [32], locally linear embedding (LLE) [33, 34], feature learning with auto-encoder and other neural network framework [35, 36].

A common question in terms of dimension reduction is to find the best dimension, which typically refers to the optimal number of components in feature extraction or the best feature subset in feature selection. This task leads to the topic of model selection/optimization. In practice, it is achieved by searching for a trade-off between error and variance [37]. That is to say, to find a compromise between the error on the training data as well as a good generalization performance on test data. A routine procedure is to split the available dataset into training and validation data and optimize the model by minimizing the prediction error of the validation data. Figure 3 illustrates an example of dimension reduction conducted with a factor method, where the original dataset ($\mathbb{R}^{k\times m}$) is reduced to $\mathbb{R}^{k\times d}$. The training and validation data is represented by blue and green blocks, respectively.

Dimension reduction and statistical modeling are conducted with the training data only. The validation data is predicted afterward. The performance of the prediction is benchmarked by a pre-determined metric. To get a more stable optimization, a cross-validation is often applied [38], where the dataset is re-split into training and validation datasets for several times. The metric of the prediction is averaged over the splits to determine the optimal dimension (d).

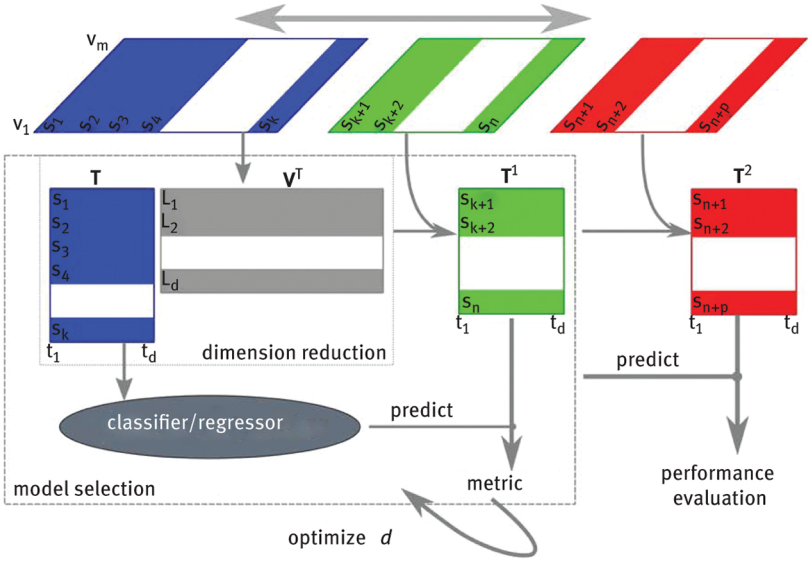

Figure 3: Workflow of dimension reduction and statistical modeling. The dimension reduction can be done with factor method like PCA. The three data blocks shown in blue, green, and red represent training, validation, and testing dataset, respectively. The training dataset is used to build the model. Parameters of the model (such as the number of principal components of the PCA, d) are optimized based on the prediction on the validation dataset (green). Afterwards, the model is evaluated according to the prediction on the testing dataset (red).

A crucial issue to find the best dimension is that the optimal result of dimension reduction is dependent on the model applied subsequently. Therefore, the optimization has to be conducted for dimension reduction and the subsequent model altogether, like in Figure 3. Another critical issue is that after the optimization, an additional validation is necessary to evaluate the performance of the optimized model, namely, the external validation. The normal way is to predict data that is not used as training or validation data (red block in Figure 3). In this case, the unknown data must be excluded during the model construction and optimization, especially if supervised dimension reduction methods are employed [39]. More details on this topic can be found in Section 4.3 of this chapter.

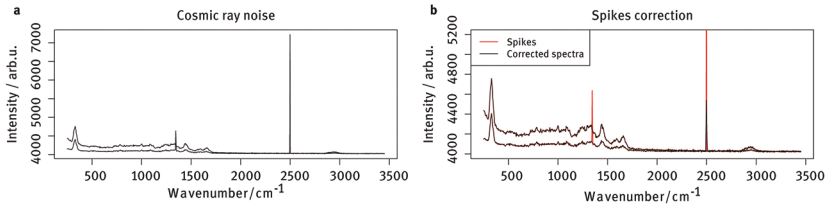

Figure 4: Cosmic ray noise. On the left side unprocessed Raman spectra containing spikes are shown. On the right side these spectra are shown in red and the corrected spectra are plotted over them in black. Therefore, only the spikes are visible in red color.

4 Models

After the Raman spectra are pretreated and preprocessed, statistical models are applied in order to extract high-level information, such as concentrations of substances, distribution of substances, disease markers or sample types. The so-called statistical methods aim to translate the standardized Raman spectra into high-level information of interest, which can be further used by chemists, biologists and physicians. As most of these methods have a statistical origin we call them statistical methods even though a few are developed within the framework of machine learning [40].

The statistical methods applied for the analysis of Raman spectra are standard techniques and we group them according to their application scenario. We will first introduce clustering algorithms and unmixing procedures utilized for image generation. It should be noted here that the methods for dimension reduction can be utilized for the image generation as well. Thereafter we describe supervised methods including classification models for diagnostics and regression procedures for analytics [1].

4.1 Clustering and unmixing for imaging

There are two major clustering algorithm types: hard clustering and fuzzy clustering. In the former a spectrum belongs to one certain cluster and no any other clusters. The latter methods are related to unmixing. A spectrum belongs to multiple clusters to a certain extent, which is called cluster membership. Both types of clustering methods are widely used in Raman spectroscopy, especially for imaging purposes to produce an overview. The most often applied clustering algorithms are *k*-means clustering and hierarchical clustering [41]. The *k*-means clustering starts with a random cluster distribution of *k* clusters. Then the distance of all spectra to the cluster mean spectra are evaluated and the spectra are resorted corresponding to the minimal distance. The procedure should be performed for multiple times because the algorithm is greedy. The most common version of hierarchical clustering is the agglomerative clustering, where Raman spectra are merged to clusters until only one cluster exists [40]. As described above there are also fuzzy clustering versions like fuzzy c-means clustering [42], which is the extension of the hard *k*-means clustering. If the task is to determine mixture compositions, unmixing methods are the ideal tools. To extract pure components from the data without a training dataset the so-called end-member extraction methods were developed. They extract the most "extreme" spectra in a specific sense from the data. Methods that are commonly applied for end-member extraction are N-FINDR [41] and Vertex Component Analysis (VCA) [41]. Besides these techniques the multivariate curve resolution-alternating least squares (MCR-ALS) method gains more and more attention due to the incorporation of different constraints and additional knowledge about the data [27].

4.2 Classification for diagnostics and regression models for analytics

If no training data with reference values are available, clustering or unmixing are the only techniques which can be applied. If training data with reference values are available, supervised machine learning algorithms, like regression or classification models, can be applied to extract high-level information [43]. Linear classification and regression methods are frequently applied due to their simplicity and robustness. Even though linear models often perform well, in many cases more powerful techniques are needed. Among those the most often utilized ones are kernel support vector machines (SVMs) and artificial neural networks [43]. Especially deep artificial neural network are powerful emerging techniques for classification and regression [44]. Another powerful classification and regression algorithm is the random forest (RF) model, which is an ensemble based method [45]. RF constructs a pre-defined number of random decision trees and every tree is able to predict. The output of the whole RF is generated by a voting procedure at the end. While certain models like SVMs or ANNs can be used intrinsically as classification and regression model, pure regression models can be converted to classification models using pseudo-concentrations. Among the most often applied regression models are principal component regression and PLS regression [46].

4.3 Evaluation procedures

A statistical model with perfect predictive performance is only possible if the training data is a complete representation of the population under investigation. However, this is usually not the case in real-world tasks and the data at hand is always a limited sampling of the population. The property of the population has to be estimated from the (limited) sampled data. In terms of chemometrics, the estimation refers to constructing a statistical model for a given classification or regression task. Due to the incomplete sampling, errors almost always occur when predicting new data with a trained model. In extreme cases, the model may fit the training data perfectly but cannot be generalized to unknown data. This is termed overfitting. To avoid overfitting, a procedure is required to evaluate the generalization performance of the model before it can be used in practice. To do so, the model is used to predict data that has not been used during model construction [37, 47]. The performance of the prediction can be benchmarked by different metrics, as is described in the following.

For the evaluation of regression models the prediction performance is measured as differences between the true and predicted values, for instance, the root-mean-squared-error of prediction (RMSEP; RMSEP =

$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$) or the mean-squared-error of prediction (MSEP; MSEP = RMSEP^2). To quantify the performance of classification/clustering models, the metrics can be used are accuracy, sensitivity, specificity and Cohen's kappa. These values are derived from a confusion matrix, which is the combination of the predicted and true group assignments (see Table 1). A model can be evaluated with a combination of multiple metrics according to certain fusion regimes like the sum of ranking difference (SRD) [48]. This combination can provide more stable and reliable model evaluations than using a single metric alone.

Table 1: Confusion table. The table compares the prediction against the correct group assignment. From this table a number of classification characteristics, like accuracy, sensitivity and specificity, can be calculated.

		Predicted	
		P	˜P
True	P	**a**	b
	˜P	c	**d**

Accuracy = $100\% \times \frac{a+b}{a+b+c+d}$, percentage of correctly classified samples.
Sensitivity = $100\% \times \frac{a}{a+b}$, percentage of true positive.
Specificity = $100\% \times \frac{d}{c+d}$, percentage of true negative.
$\kappa = \frac{\text{Accuracy} - p_e}{1-p_e}$, agreement between truth and prediction.

As shown in Figure 3, the model evaluation requires a prediction of new data that is independent of the training data. This requires additional samples to be measured, which can be expensive, time-consuming, or even impossible. A more practical solution is resampling, where the datasets for model training and prediction are independently sampled from the accessible data. The most widely applied resampling regime is cross-validation [49]. Thereby the accessible dataset is split into training and testing data, used for model training and evaluation, respectively. The splitting procedure is repeated for several times to get a stable validation. The results of the prediction over all splits are averaged to benchmark the generalization performance of the statistical model. According to the data splitting scheme, cross-validation can be conducted in different ways, including leave-*p*-out cross-validation, *k*-fold cross-validation, and Monte Carlo cross-validation [50]. A special case of cross-validation is holdout validation, where the data split is done only once without repetition. Another important resampling method is bootstrapping, which is a resampling procedure with replacement. No matter how data split is performed, the proportional composition of classes (or concentrations) in every split should be consistent to the composition of the population, because the constructed model can be influenced by the relative compositions of different classes in the training data. One way to achieve this is the Latin-partition method [51]. Specifically, bootstrapping with Latin partition was reported, which constructs multiple Latin partitions with a bootstrap. This allows getting the relation between the prediction and the composition of the training data as well as the optimization of the statistical model [52, 53].

There are two issues extremely crucial to model evaluation [38, 39, 54]. The first issue is the independence requirement between training and testing data. In practice, this is ensured by resampling the data on the highest level of sampling hierarchy, which might be the biological replicate, cultivation, or patient. With *k*-fold cross-validation, for example, the folds should be arranged according to the highest level of sampling hierarchy. Otherwise, the information of testing data is implicitly used during model construction and the prediction on the testing data does not reveal the true generalization performance. As a result, the statistical model is overestimated. In addition, the evaluation should be conducted involving both the dimension reduction and the statistical model. This is especially important if a supervised dimension reduction is applied. In special cases, if parameters of a pre-processing procedure are optimized according to the output of the statistical models, the evaluation loop should include the pre-processing steps as well.

Noteworthy, it requires special attention when both the model selection and model evaluation are conducted with cross-validation. In this case, a two-layer cross-validation is needed [39]. As shown in Figure 3, the dataset is split into training (blue and green blocks) and testing data (red block) within the outer-layer cross-validation. The testing data is taken aside and the training data is fed into the inner-layer cross-validation for model optimization as described in Section 3.4. Thereafter the testing data is predicted with the optimized model, of which the results are used to evaluate the performance of the model. The inner-layer and outer-layer cross-validation are termed internal and external validation, respectively.

5 New trends

As almost all applications of Raman spectroscopy are only possible, if an adequate data analysis pipeline is utilized, the research area developing new analysis methods and tools is active. A full summary of trends is beyond the scope of this chapter, but two topics, which emerged recently, should be summarized here. These both topics are model transfer and data fusion. Model transfer is dealing with the use of models, which were constructed based on training data different from the test data. Such a model transfer issue arises, if a chemometric model should be utilized for diagnostic purposes and the device on which the test data are measured in clinics is different compared to the device utilized for measuring the training data. This issue is important because it is linked to real-world applications of Raman spectroscopy. The next active research area is related to data fusion, where Raman spectra are computationally combined with other data types in order to extract more information as it would be possible from the Raman spectra alone. In that manner, complementary information to the information extracted from Raman spectra can be analyzed together with the Raman based information.

5.1 Model transfer

A well-known challenge in chemometrics is the substantially inferior prediction quality of a pre-trained chemometric model if it is applied to newly measured data [55]. This issue gets more important if the new data is significantly different to training data. In Raman spectroscopy, such differences manifest itself as wavenumber shifts and intensity variations (Figure 5a). One of the major reasons for such spectral deviations is the instrumental change over-time or after replacement of a component. The wavenumber/intensity calibration (see Section 2) helps to reduce such instrument induced spectral variations but cannot completely remove them. The remaining spectral variations can still mask the spectral differences of interest and thus corrupt the prediction, which is very common in biological studies. Besides the instrument variations, other experimental changes can also disturb the reproducibility, for example, cultivation conditions cannot be exactly identical for all replicates and differences over measurement of different replicates are resulting. These spectral variations cannot be tackled with calibration at all. That is why an existent model cannot successfully predict the newly measured data. A simple but labor-extensive solution is to train another model for this new data. However, this is not possible if new training samples are inaccessible, which might be the case in disease diagnosis. Therefore, a method is needed to enable the prediction of the new data based on the existent model. This is achieved with model transfer approaches, as described in the following [14, 56, 57].

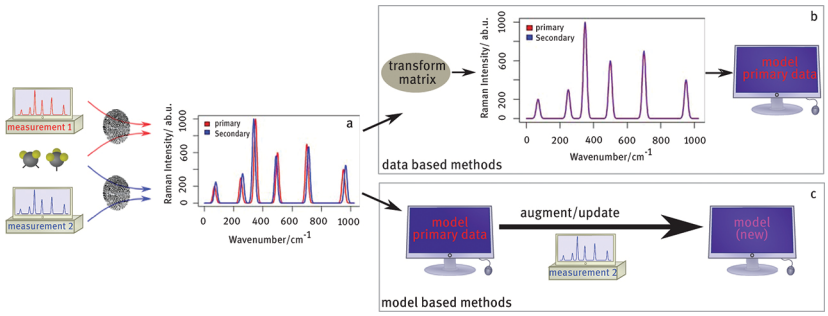

Figure 5: Overview of model transfer. (a) Training (primary) and testing (secondary) datasets can be significantly different if they are measured from different replicates or on different devices. Hence the chemometric model constructed with the primary dataset can fail to predict the secondary dataset. This can be tackled with model transfer approaches according to two mechanisms: data based methods (b) and model based methods (c).

In the terminology of model transfer, the (old) training and the new data are termed primary and secondary data, respectively. There are two types of model transfer approaches: data based (Figure 5b) and model based methods (Figure 5c) [58]. In the former case, the primary and secondary data are transformed to make them more similar. In the latter case, an existent model is updated to improve the prediction on the secondary data. Model transfer can be applied in a supervised or an unsupervised manner. Unsupervised model transfer is conducted without the knowledge of response variables (class information or concentration) of the secondary data. For supervised model transfer, a few secondary samples with known responses are needed; but the required sample size is much smaller compared to the construction of a new model. Unsupervised methods do not need the response information of the secondary data, making them superior to supervised model transfer in the cases where the response information is not accessible. A typical example of this case is bio-medical diagnostics, where the disease level of a new patient should be predicted and is unknown.

Data based model transfer aims to remove the spectral variations between secondary and primary data. Typically applied methods are Procrustes analysis and piecewise direct standardization (PDS) [59], where a transformation matrix is calculated to map the secondary to the primary data. Other approaches include warping methods, which adjust the peak misalignment between primary and secondary data [60]. These methods are typically conducted based on spectra of standard samples measured under primary and secondary conditions. Then the resulting correction function is applied to the spectra of the secondary samples. The bottleneck of these methods is that the standard and secondary samples have to be measured under identical conditions. In addition, the correction function only corrects the instrumental deviations and cannot tackle disturbing effects like the variation of cultivation conditions. Hence the secondary and primary data of secondary samples cannot be perfectly matched. Another option is to calculate the transformation matrix based on the secondary samples itself. However, this works only if the samples for primary and secondary measurements share the same chemical components. In classification and regression tasks, the samples belong to multiple classes or feature different concentrations of their components. It is almost impossible to ensure that both the primary and the secondary samples feature the same classes or identical concentrations. In this case, the transformation matrix does not only model the undesired changes but also the spectral differences of interest. It is advisable to calculate transformation matrix separately for each class or concentration, i. e. in a supervised manner.

Model based model transfer seeks for a compromise between the primary and secondary data. The first scheme is to build a global statistical model involving experimental variables responsible for undesired spectral changes [58]. This procedure requires to know and to include all potential influential experimental variables, which makes it less feasible for model transfer. An alternative regime is to build a model on the primary data with features robust to the experimental changes. In ref [61] the authors proposed a sample-wise spectral multivariate calibration approach by penalizing and desensitizing features that strongly differ between the primary and secondary data. This method is less powerful if the secondary conditions differ strongly from the primary conditions. The third scheme is local modeling, where the model is built only with the primary samples that are the nearest neighbors to the secondary data [62]. It is crucial in local modeling to determine the number of nearest neighbors and the similarity metric to select the nearest neighbors. Another model transfer approach is model augmentation. Thereby, the training dataset is enlarged with several additional secondary samples, the so-called transfer samples [63].

So far model transfer has been mostly investigated for near-infrared spectroscopy and regression problems. Model transfer of Raman spectroscopy and classification tasks is rather new and only a few studies exist. Recently, a model transfer approach was developed for Raman spectroscopy using Tikhonov regularization based partial least squares regression (TR-PLSR) [14]. However, the method does not work if the response variables of the secondary data are unavailable. To deal with this issue, unsupervised model transfer approaches for Raman spectroscopy were also developed recently [57].

5.2 Data fusion

In cases when the Raman data does not yield sufficient information, it can be complemented by additional measured data. For example, if correlated imaging is performed, several types of spectroscopic or spectrometric data are measured and can be combined. Another example is that beside the Raman spectra of the sample other additional information can be used, like the patient's laboratory values, gender, age, physical parameters, and known medical conditions. These values can be used along with Raman spectra to improve the performance of a chemometric model. Often the different data types feature different dynamic range, dimensionality, and the number of observations per sample. Therefore, the question arises how different data types can be combined [64]. This process of combining different types of data is called data fusion.

The data fusion can be performed on different levels [65] of the analysis pipeline (Figure 6). The combination can be done on a low-level directly after the preprocessing, possibly even before the dimension reduction. This data fusion scheme is called centralized data fusion. It is performed by merging the data from different sources into a single data matrix with subsequent simultaneous analysis. However, dealing with different data types within a low-level data fusion approach requires accounting for different scaling, dynamic range and dimensionality of the data in order to balance the contributions of different data types against each other. To account for these differences, it may be important to rescale the data before combining them [66].

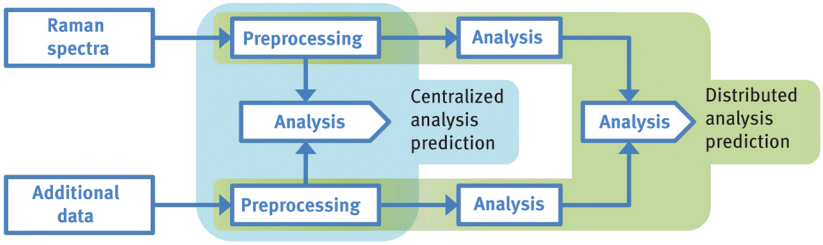

Figure 6: Schematic representation of data combination with a fusion center at different levels of the data analysis pipeline. The low-level (centralized) data fusion is highlighted in blue color, and the high-level (distributed) data fusion is highlighted in green color.

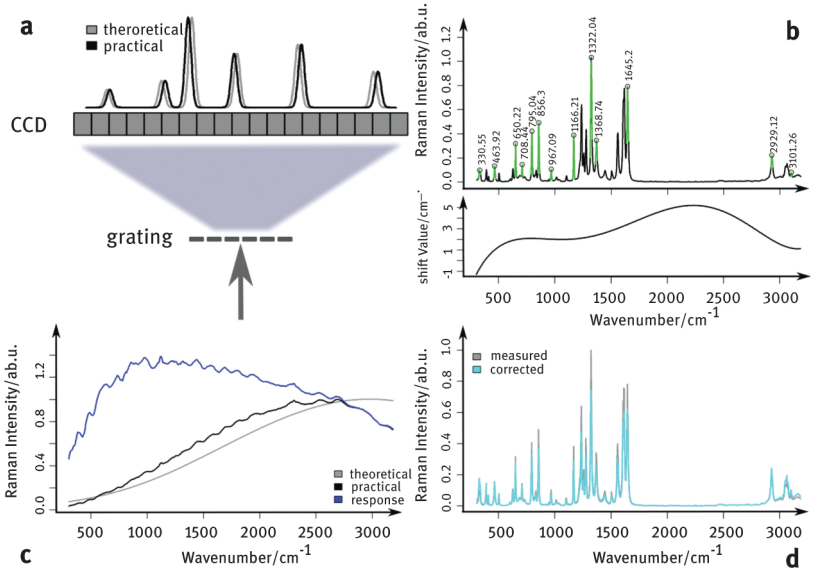

Figure 7: Workflow of wavenumber (a–b) and intensity calibration (c–d). (a) The relation between wavenumber and pixel positions can change, leading to wavenumber misalignment between measured and theoretical Raman spectra. (b) The wavenumber misalignment is corrected based on Raman spectra of a known standard material. (c) The intensity response function of the device is calculated as the ratio between measured and theoretical emission of a known standard material. (d) Intensity axis of measured Raman spectra is corrected by the calculated intensity response function.

Another possible data fusion scheme is a high-level data fusion, also called a distributed data fusion. In this scheme, each data type is analyzed separately and the scores are combined at the final step of the analysis [67]. The advantage of the high-level data fusion is that it is computationally less costly and allows dealing with the different data types in an easier manner.

Besides the low-level and high-level data fusion, a decentralized data fusion approach [68] or hierarchical data fusion can be used. For example, in correlated imaging, the hierarchical data fusion allows using one imaging technique for finding areas of the interest. The other imaging technique is then utilized to study these regions [69, 70]. In that manner, deeper insights into the investigated areas can be gathered.

6 Summary: dos and don'ts in analyzing Raman spectra

In this section we would like to summarize the sections above. We would like to give the summary in terms of a Dos and Don'ts list. In that manner, we tried to condense the content of the whole chapter into guidelines and rules. In order to allow a further reading, the recommended practices discussed in the sections about data pretreatment (Section 2), data preprocessing (Section 3) and models (Section 4) are marked in the table with a respective chapter number in the parentheses.

Do's	Dont's
Avoid fluorescence if possible (3)	
Check instrumental drift every day before measurement (2)	

Measure standard material for calibration every time before measuring real samples (2)	**Measure a spectrum of standard material for calibration at different conditions (days) compared to real samples (2)**
Wavenumber calibration and/or intensity calibration (2)	**Direct application of a model to data of another device (2)**
Model transfer between different devices/replicates (4)	**Model transfer between datasets measured from different classes (4)**
Apply baseline correction methods prior modeling (3)	
Smoothing before SNIP baseline correction (3)	
Normalization (after baseline correction) (3)	**Normalization before baseline correction (3)**
Involve procedures to be optimized like dimension reduction (pre-processing, if necessary) inside the CV loop (4)	**Perform dimension reduction outside of CV loop, especially for supervised dimension reduction approaches (4)**
Evaluate the model with independent dataset (e. g. external CV) (4)	**Evaluate the model with data already used during model construction/optimization (4)**
Use data from the same sample (replicate) exclusively as training or testing data (4)	**Split the data into training and testing data regardless of the replicate information (4)**
Leave-one-sample (replicate)-out CV (4)	**Leave-one-spectrum-out CV (4)**

References

[1] Bocklitz TW, Guo S, Ryabchykov O, Vogler N, Popp J. Raman based molecular imaging and analytics: a magic bullet for biomedical applications!? Anal Chem. 2016;88:133–51.
[2] Ehrentreich F, Sümmchen L. Spike removal and denoising of Raman spectra by wavelet transform methods. Anal Chem. 2001;73:4364–73.
[3] Schulze HG, Turner RF. A two-dimensionally coincident second difference cosmic ray spike removal method for the fully automated processing of Raman spectra. Appl Spectrosc. 2014;68:185–91.
[4] Ryabchykov O, Bocklitz T, Ramoji A, Neugebauer U, Foerster M, Kroegel C, et al. Automatization of spike correction in Raman spectra of biological samples. Chemometrics Intell Lab Syst. 2016;155:1–6.
[5] Dörfer T, Bocklitz T, Tarcea N, Schmitt M, Popp J. Checking and improving calibration of Raman spectra using chemometric approaches. Z Phys Chem Int J Res Phy Chem Chem Phy. 2011;225:753–64.
[6] Bocklitz T, Dörfer T, Heinke R, Schmitt M, Popp J. Spectrometer calibration protocol for Raman spectra recorded with different excitation wavelengths. Spectrochimica Acta A: Mol Biomol Spectrosc. 2015;149:544–9.
[7] McCreery RL. Raman spectroscopy for chemical analysis Vol. 157. New York: John Wiley & Sons, 2000
[8] Berg RW, Nørbygaard T. Wavenumber calibration of CCD detector Raman spectrometers controlled by a sinus arm drive. Appl Spectrosc Rev. 2006;41:165–83.
[9] Carrabba MM. Wavenumber standards for Raman spectrometry. In: Griffiths P, Chalmers JM, editor(s). Handbook of vibrational spectroscopy. Chichester: Wiley online library, 2006
[10] E1840-96, A., Standard guide for Raman shift standards for spectrometer calibration. ASTM International, West Conshohocken, PA, 2014. 03.06.
[11] Fryling M, Frank CJ, McCreery RL. Intensity calibration and sensitivity comparisons for CCD/Raman spectrometers. Appl Spectrosc. 1993;47:1965–74.
[12] Davis W, Forney G, Bukowski R. National institute of standards and technology, Gaithersburg MD, USA.
[13] Rodriguez JD, Westenberger BJ, Buhse LF, Kauffman JF. Standardization of Raman spectra for transfer of spectral libraries across different instruments. Analyst. 2011;136:4232–40.
[14] Guo S, Heinke R, Stöckel S, Rösch P, Bocklitz T, Popp J. Towards an improvement of model transferability for Raman spectroscopy in biological applications. Vib Spectrosc. 2017;91:111–8.
[15] Bocklitz T, Walter A, Hartmann K, Rosch P, Popp J. How to pre-process Raman spectra for reliable and stable models? Anal Chim Acta. 2011;704:47–56.
[16] Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. Anal Chem. 1964;36:1627–39.
[17] Lieber CA, Mahadevan-Jansen A. Automated method for subtraction of fluorescence from biological Raman spectra. Appl Spectrosc. 2003;57:1363–7.
[18] Eilers PH, Boelens HF. Baseline correction with asymmetric least squares smoothing. Leiden Univ Med Centre Rep. 2005;1:1.
[19] Ryan CG, Clayton E, Griffin WL, Sie SH, Cousens DR. SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. Nucl Instrum Methods Phys Res B: Beam Interact Mater Atoms. 1988;34:396–402.
[20] Martens H, Stark E. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. J Pharm Biomed Anal. 1991;9:625–35.
[21] Guo S, Bocklitz T, Popp J. Optimization of Raman-spectrum baseline correction in biological application. Analyst. 2016;141:2396–404.
[22] Gautam R, Vanga S, Ariese F, Umapathy S. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. EPJ Tech Instrum. 2015;2:8.
[23] Black MJ, Anandan P. The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. Comput Vis Image Understand. 1996;63:75–104.

[24] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
[25] Malinowski ER. Factor analysis in chemistry, 2 ed. New York: Wiley, 1991
[26] Zhang X, Tauler R. Application of multivariate curve resolution alternating least squares (MCR-ALS) to remote sensing hyperspectral imaging. Anal Chim Acta. 2013;762:25–38.
[27] Piqueras S, Krafft C, Beleites C, Egodage K, Von Eggeling F, Guntinas-Lichius O, et al. Combining multiset resolution and segmentation for hyperspectral image analysis of biological tissues. Anal Chim Acta. 2015;881:24–36.
[28] Brereton RG, Jansen J, Lopes J, Marini F, Pomerantsev A, Rodionova O, et al. Chemometrics in analytical chemistry – Part I: history, experimental design and data analysis tools. Anal Bioanal Chem. 2017;409:5891–9.
[29] Bruce LM, Koger CH, Li J. Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. IEEE Trans Geosci Remote Sens. 2002;40:2331–8.
[30] Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B Methodol. 1996;58:267–88.
[31] Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc Series B Stat Methodol. 2010;72:3–25.
[32] Zhang Z, Chow TW, Zhao M. M-Isomap: orthogonal constrained marginal isomap for nonlinear dimensionality reduction. IEEE Trans Syst Man Cybern. 2013;43:180–91.
[33] Silva VD, Tenenbaum JB. Global versus local methods in nonlinear dimensionality reduction. In: Advances in neural information processing systems. Cambridge, MA, USA: MIT Press, 2003.
[34] Shan R, Cai W, Shao X. Variable selection based on locally linear embedding mapping for near-infrared spectral analysis. Chemometrics Intell Lab Syst. 2014;131:31–6.
[35] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313:504–7.
[36] Wang W, Huang Y, Wang Y, Wang L. Generalized autoencoder: a neural network framework for dimensionality reduction. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2014.
[37] Kalivas JH, Palmer J. Characterizing multivariate calibration tradeoffs (bias, variance, selectivity, and sensitivity) to select model tuning parameters. J Chemom. 2014;28:347–57.
[38] Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Stat Surv. 2010;4:40–79.
[39] Guo S, Bocklitz T, Neugebauer U, Popp J. Common mistakes in cross-validating classification models. Anal Methods. 2017;9:4410–7.
[40] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning; data mining, inference and prediction. New York: Springer, 2008.
[41] Hedegaard M, Matthäus C, Hassing S, Krafft C, Diem M, Popp J. Spectral unmixing and clustering algorithms for assessment of single cells by Raman microscopic imaging. Theor Chem Acc. 2011;130:1249–60.
[42] Bezdek JC, Ehrlich R, Full W. FCM: the fuzzy c-means clustering algorithm. Comput Geosci. 1984;10:191–203.
[43] Bocklitz T, Putsche M, Stüber C, Käs J, Niendorf A, Rösch P, et al. A comprehensive study of classification methods for medical diagnosis. J Raman Spectrosc. 2009;40:1759–65.
[44] Acquarelli J, Van Laarhoven T, Gerretzen J, Tran TN, Buydens LM, Marchiori E. Convolutional neural networks for vibrational spectroscopic data analysis. Anal Chim Acta. 2017;954:22–31.
[45] Breiman L. Random forests. Mach Learn. 2001;45:5–32.
[46] Mevik B-H, Wehrens R, Liland KH. pls: partial least squares and principal component regression. R Package Version. 2011;2(3).
[47] Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput. 2017;27:1413–32.
[48] Héberger K. Sum of ranking differences compares methods or models fairly. TrAC Trends Anal Chem. 2010;29:101–9.
[49] Refaeilzadeh P, Tang L, Liu H. Cross-validation, in Encyclopedia of database systems. New York: Springer, 2009:532–8.
[50] Xu QS, Liang YZ, Du YP. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. J Chemom. 2004;18:112–20.
[51] Wan C, Harrington PDB. Screening GC-MS data for carbamate pesticides with temperature-constrained-cascade correlation neural networks. Anal Chim Acta. 2000;408:1–12.
[52] De Boves Harrington P. Statistical validation of classification and calibration models using bootstrapped Latin partitions. TrAC Trends Anal Chem. 2006;25:1112–24.
[53] Qi N, Zhang Z, Xiang Y, Yang Y, Liang X, Harrington PD. Terahertz time-domain spectroscopy combined with support vector machines and partial least squares-discriminant analysis applied for the diagnosis of cervical carcinoma. Anal Methods. 2015;7:2333–8.
[54] Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform. 2014;6:10.
[55] Copas JB. Regression, prediction and shrinkage. J R Stat Soc Series B Methodol. 1983;45:311–54.
[56] Shahbazikhah P, Kalivas JH. A consensus modeling approach to update a spectroscopic calibration. Chemometrics Intell Lab Syst. 2013;120:142–53.
[57] Guo S, Heinke R, Stöckel S, Rösch P, Popp J, Bocklitz T. Model transfer for Raman-spectroscopy-based bacterial classification. J Raman Spectrosc. 2018;49:627–37.
[58] Kalivas JH, Siano GG, Andries E, Goicoechea HC. Calibration maintenance and transfer using Tikhonov regularization approaches. Appl Spectrosc. 2009;63:800–9.
[59] Liang C, Yuan H-F, Zhao Z, Song C-F, Wang J-J. A new multivariate calibration model transfer method of near-infrared spectral analysis. Chemometrics Intell Lab Syst. 2016;153:51–7.
[60] Bloemberg TG, Gerretzen J, Lunshof A, Wehrens R, Buydens LM. Warping methods for spectroscopic and chromatographic signal alignment: a tutorial. Anal Chim Acta. 2013;781:14–32.
[61] Kalivas JH, Brownfield B, Karki BJ. Sample-wise spectral multivariate calibration desensitized to new artifacts relative to the calibration data using a residual penalty. J Chemom. 2017;31:e2873
[62] Bevilacqua M, Marini F. Local classification: locally weighted-partial least squares-discriminant analysis (LW-PLS-DA). Anal Chim Acta. 2014;838:20–30.

[63] Kalivas JH. Overview of two-norm (L2) and one-norm (L1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance. J Chemom. 2012;26:218–30.
[64] Castanedo F. A review of data fusion techniques. Sci World J. 2013;2013:19.
[65] Márquez C, López MI, Ruisánchez I, Callao MP. FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud. Talanta. 2016;161:80–6.
[66] Teglia CM, Azcarate SM, Alcaráz MR, Goicoechea HC, Culzoni MJ. Exploiting the synergistic effect of concurrent data signals: low-level fusion of liquid chromatographic with dual detection data. Talanta. 2018;186:481–8.
[67] Borràs E, Ferré J, Boqué R, Mestres M, Aceña L, Busto O. Data fusion methodologies for food and beverage authentication and quality assessment – a review. Anal Chim Acta. 2015;891:1–14.
[68] Durrant-Whyte H, Stevens M, Nettleton E. Data fusion in decentralised sensing networks. In: 4th International Conference on Information Fusion, 2001.
[69] Bocklitz T, Bräutigam K, Urbanek A, Hoffmann F, Von Eggeling F, Ernst G, et al. Novel workflow for combining Raman spectroscopy and MALDI-MSI for tissue based studies. Anal Bioanal Chem. 2015;407:7865–73.
[70] Bocklitz T, Crecelius AC, Matthaus C, Tarcea N, Von Eggeling F, Schmitt M, et al. Deeper understanding of biological tissue: quantitative correlation of MALDI-TOF and Raman imaging. Anal Chem. 2013;85:10829–34.