Analytica Chimica Acta 704 (2011) 47–56

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

How to pre-process Raman spectra for reliable and stable models?

Thomas Bocklitz[a], Angela Walter[a], Katharina Hartmann[a], Petra Rösch[a], Jürgen Popp[a,b,*]

[a] Institute of Physical Chemistry and Abbe-Center of Photonics, Helmholtzweg 4, Friedrich-Schiller University, D-07743 Jena, Germany
[b] Institute of Photonic Technology, Albert-Einstein-Straße 9, D-07745 Jena, Germany

ARTICLE INFO

Article history:
Received 13 September 2010
Received in revised form 31 March 2011
Accepted 21 June 2011
Available online 31 July 2011

Keywords:
Raman spectroscopy
Quantitative analysis
Calibration
Pre-processing
Classification
Genetic algorithm

ABSTRACT

Raman spectroscopy in combination with chemometrics is gaining more and more importance for answering biological questions. This results from the fact that Raman spectroscopy is non-invasive, marker-free and water is not corrupting Raman spectra significantly. However, Raman spectra contain despite Raman fingerprint information other contributions like fluorescence background, Gaussian noise, cosmic spikes and other effects dependent on experimental parameters, which have to be removed prior to the analysis, in order to ensure that the analysis is based on the Raman measurements and not on other effects.

Here we present a comprehensive study of the influence of pre-processing procedures on statistical models. We will show that a large amount of possible and physically meaningful pre-processing procedures leads to bad results. Furthermore a method based on genetic algorithms (GAs) is introduced, which chooses the spectral pre-processing according to the carried out analysis task without trying all possible pre-processing approaches (grid-search). This was demonstrated for the two most common tasks, namely for a multivariate calibration model and for two classification models. However, the presented approach can be applied in general, if there is a computational measure, which can be optimized. The suggested GA procedure results in models, which have a higher precision and are more stable against corrupting effects.

1. Introduction

During the last decade a large number of light-matter interaction phenomena like fluorescence microscopy [1], second harmonic generation microscopy [2], absorption spectroscopy [3] or Raman based methods (e.g. micro-Raman [4,5], resonance Raman [6], SERS [7,8], TERS [9,10], CARS [11], and SRS) were utilized in order to get a deeper insight into biological samples. In particular Raman spectroscopy as a non-invasive and marker-free technique, which is not corrupted by the presence of water, is an ideal tool for dealing with biological problems like the investigation of biological cells [12–14].

One challenge in Raman microspectroscopy is the analysis of the spectra. Since Raman spectra of biological cells exhibit all very similar Raman spectra special methods to properly analyze the data, e.g. statistical/chemometrical methods are necessary. The proper application of these powerful classification and calibration techniques is restricted however to the case where all side effects which might influence the Raman spectra are separated from those. Such effects which corrupt Raman spectra are the appearance of a fluorescence background, CCD background noise, Gaussian noise and cosmic noise. All of these effects contribute to a certain wavenumber region of the experimentally recorded Raman spectra, and therefore have to be removed. Beside these corrections it is often necessary to correct for varying sampling geometries and reject highly redundant variables.

Here we establish a methodology for choosing the best pre-processing, in order to exploit the unique potential of Raman spectroscopy together with chemometrics for studying biological samples. The workflow is as follows: first all suitable pre-processing steps must be arranged and a order in which they should be applied has to be chosen. This order must be derived from principles of the effects, which should be removed from the experimentally recorded Raman spectra. Afterwards a genetic algorithm (GA) is applied, which is able to combine different pre-processing procedures, but with the restrictions given by the selected order. The GA is utilized in order to optimize the internal validation of statistical methods. This was carried out for the commonly applied methods, a multivariate calibration procedure and a classification technique. Besides the optimization of the pre-processing with the genetic algorithm the outcomes of the statistical method for all possible pre-processing combinations were calculated. This operation, which is called grid-search, was utilized to judge the time saving

* Corresponding author at: Institute of Physical Chemistry, Helmholtzweg 4, Friedrich-Schiller University, D-07743 Jena, Germany. Tel.: +49 3641948320.
E-mail address: juergen.popp@uni-jena.de (J. Popp).

0003-2670/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2011.06.043

Table 1
Reference substances together with their physical properties are given.

Name	Density, ρ (g cm^{-3})	Molar mass, M (g mol^{-1})
Ethanol	0.79	46.07
2-Propanol	0.78	60.10
DMSO	1.1	78.13
1-Octanol	0.83	130.23

and the quality the genetic algorithm was achieving compared to the calculation of all possible solutions.

In this study three datasets are generated and investigated. First the Raman spectra for the multivariate calibration experiment (dataset 1) consist of four substances with varying concentrations. This experiment is constructed to simulate the problem, which is occurring if online measurements of some substances are carried out [7]. The other two datasets were used for classification experiments. The artificial classification dataset (dataset 2) was constructed out of two components which were mixed in three fixed concentration ratios. The resulting three-class problem was designed to match problems found within the biology of fungi [15]. The last dataset (dataset 3) consists of bacteria spectra of two species and is termed real-world classification dataset. The problem is inspired by the assumption, that both *Streptomyces* species change their composition, if they grow in different media [16,17].

2. Experimental

In the following the preparation of all three datasets is described and the measurement conditions are given.

2.1. Reference substances

The calibration samples (dataset 1) are assembled by mixing the four components, which were supplied by Sigma Aldrich and are given in Table 1. The four solvents are mixed in steps of 200 μL and always 5 parts are combined in all possible combinations. Therefore the total volume of one sample is approximately 1 mL. Because the Raman intensity is proportional to the number of scatterers [18], the volume percentage has to be converted into a relative number of scatterers N_i.

$$N_i = \frac{\rho_i \cdot V_i}{M_i} \quad (1)$$

where ρ_i is the density, V_i is the volume and M_i corresponds to the molar mass of the i-th component. The number N_i is then proportional to the influence of every reference spectrum on the composition spectrum. This conversion also corrects for the effect that the mixture can exhibit a smaller volume than the sum of the volumes of the different contributions. The four solvents, ethanol, 2-propanol, DSMO and 1-octanol, were chosen, because they only mix and do not react with each other. Additionally their boiling point is above 78 °C, so the error introduced by vaporization of the solvents is minimized.

Table 2
Overview of the three datasets.

Dataset number	Dataset name	Number of spectra	Property
1	Calibration dataset	1180	4 substances
2	Artificial classification dataset	150	3 groups
3	Real-world classification dataset	192	4 groups

In order to study classification tasks two datasets were investigated: an artificial one and a real-world dataset. The artificial classification samples (dataset 2) were generated from glucan and chitin for three fixed ratios. The mixture ratios were determined to match biological classification tasks, e.g. different regions in fungi exhibit different glucan–chitin ratios. The solid substances glucan (β-glucan from *Saccharomyces cerevisiae*, Sigma Aldrich) and chitin (from crab shell, Sigma Aldrich) are mixed in 9:10, 1:1 and 10:9 ratios and prepared as pellets for the measurements.

The real world classification dataset (dataset 3) consists of two bacteria, *Streptomyces galilaeus* HKI 22 and *Streptomyces chartreusis* DSM 41255, which were cultured on agar plates containing minimal medium (0.5 g asparagine, 11 g glucose monohydrate, 0.2 g $MgSO_4 \cdot 7H_2O$, 0.01 g $FeSO_4 \cdot 7H_2O$, 0.66 g $K_2HPO_4 \cdot 3\ H_2O$ in 1 L distilled water) and complex medium (10 g starch, 1 g casamino acids (aMReSCO), 0.5 g K_2HPO_4 in 1 L distilled water) for 1–10 days and stored in an incubator at 28 °C. The samples were prepared every day for ten days. For sample preparation 1 mL distilled water was added to the plate, the biomass was removed with an inoculation loop, collected and centrifuged. The biomass was washed three times with distilled water and then smeared on a fused silica objective slide for the Raman measurements.

2.2. Raman spectroscopy

Raman spectra of the real-world classification dataset (dataset 3) and of the calibration samples (dataset 1) were excited with visible laser light at 532 nm from a frequency-doubled Nd:YAG laser (Coherent Compass) and recorded in 180° backscattering geometry. The laser light was coupled into a 2 mL cuvette. A LabRam HR800 spectrometer (JobinYvon) equipped with a 300 lines mm^{-1} grating was used to analyze the scattered light. The spectra were recorded with a Peltier-cooled charge coupled device (CCD) camera. The spectral resolution is about 7 cm^{-1}. For the calibration dataset (dataset 1) each concentration series was measured in a time series of 100 spectra with 0.5 s integration time. In total 11,800 spectra were acquired and every 10 spectra were averaged, i.e. the dataset consists after this procedure of 1180 Raman spectra. The time series option of the spectrometer is used for producing a given number of spectra of the same sample. These spectra are used to check, if a chemical reaction is going on or not. No indication of a chemical reaction was found. For the real-world classification dataset (dataset 3) 192 spectra of two bacteria grown on two media are measured, where the integration time was 60 s. This dataset is featuring 4 groups (two bacteria grown on two media).

The Raman spectra of the artificial classification dataset (dataset 2) were recorded with a Fourier transform Raman setup (MultiRam RFS27; Bruker) combined with a Raman-Modul (RAM II-Raman compartement). The excitation wavelength is 1064 nm provided by a diode-pumped Nd:YAG-laser (KLASTECH Laser Serie DENICAF 1064-xxx (R513-1000/R), Karpushko Laser Technologies GmbH, Dortmund (Germany)) with a spot size of about 0.1 mm. The Raman scattered light was collected by a nitrogen cooled Germanium-detector (D418-T/R27) with a spectral resolution of 4 cm^{-1}. The laser power was set to 1 W and the number of scans to 32. For each glucan–chitin ratio 10 pellets were prepared. Five FT-Raman spectra per pellet were recorded and the pellets were rotated after every measurement. This procedure results in 150 spectra, which are grouped in 3 groups.

Table 2 summarizes the names of the dataset, the number of spectra and its properties, while in Fig. 1 the mean Raman spectra of all three datasets are plotted.

2.3. Computation

The calculations were performed on a commercially available PC system (Intel(R) Core(TM) 2Duo CPU, E67502.66GHz, 1.97 GB

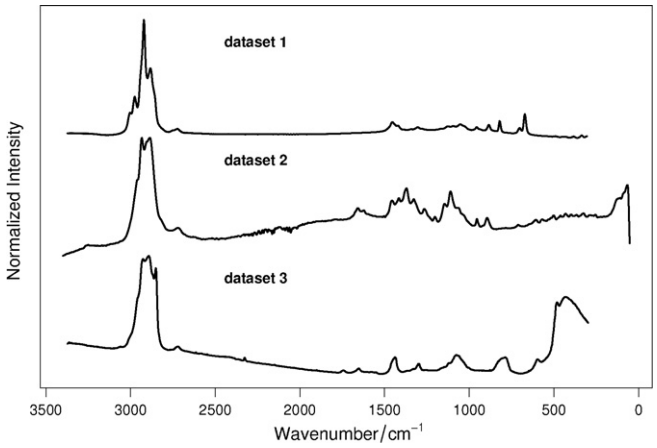

Fig. 1. Mean spectra of all three datasets.

RAM). The computations were done in Gnu R [19] a statistical programming language. The utilized packages were *KernSmooth* [20], *pls* [21], *MASS* [22], *Rgenoud* [23] and *Peaks* [24].

2.4. Pre-processing

The experimentally recorded Raman spectra are compositions of various contributions (see Fig. 2). Only contributions originating from Raman modes of the molecules in the laser focus are named in the following as Raman spectrum. This 'true' Raman spectrum is corrupted by cosmic spikes originating from high energy particles hitting the charged-coupled device (CCD), Gaussian distributed noise originating from uncorrelated processes and a high-intense background. The latter is caused by fluorescence of the sample and the baseline of the CCD itself. For all of these effects a correction procedure has to be applied in order to quantify the analysis of the Raman spectra. These procedures are grouped in four categories (background, filtering, scaling and dimension reduction), which reflect the eliminated effects, and are shortly reviewed below. No removal procedure for cosmic spikes is introduced, because this was done by the measurement software.

2.4.1. Background

The background correction procedures are minimizing the effect of a varying background caused by fluorescence of the sample or thermal fluctuations on the CCD. Theses procedures can be divided into two groups. Estimating procedures like the SNIP algorithm [25] or a polynomial background fit introduced by Lieber et al. [26] are estimating the unknown background. Another possibility is calculating the derivative, which can be done by a Savitzky–Golay algorithm [27] or by a kernel estimate [20]. All of these methods are highly dependent on the chosen parameters. Table 3 summarizes these background procedures together with the used parameters.

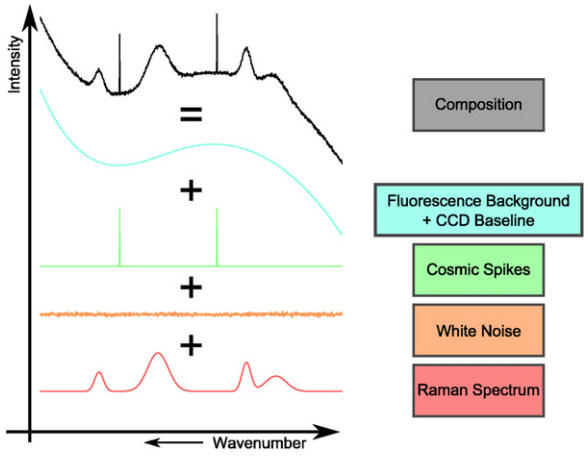

Fig. 2. Spectrum composition. The measured Raman spectra are suffering from different side effects, like fluorescence background, cosmic spikes and white noise. All contributions have to be rejected prior the analysis.

Table 3
Background block. All methods concerning rejecting background contributions within the spectra (fluorescence background, CCD baseline and so forth) are put together.

Procedure	No.B	Parameter
RAW	00	–
SNIP ($w = 9$)	01	Order = 2
	02	Order = 4
	03	Order = 6
	04	Order = 8
Polynomial fit	05	Degree = 0
	06	Degree = 1
	07	Degree = 2
	08	Degree = 3
	09	Degree = 4
	10	Degree = 5
Kernel derivative	11	drv = 1, bw = $\Delta\tilde{\nu}$
	12	drv = 2, bw = $\Delta\tilde{\nu}$
	13	drv = 3, bw = $\Delta\tilde{\nu}$
	14	drv = 4, bw = $\Delta\tilde{\nu}$
Savitzky–Golay derivative	15	Points = 7, degree = 5, drv = 1
	16	Points = 7, degree = 5, drv = 2
	17	Points = 7, degree = 5, drv = 3
	18	Points = 7, degree = 5, drv = 4

2.4.2. Filtering

Filtering methods are used to remove uncorrelated noise. This can be done by the same procedures used for derivatives like the Savitzky–Golay method [27] or the kernel smoothing [20]. In Table 4 the filtering methods and their parameters are listed.

2.4.3. Scaling

Scaling is applied in order to ensure that the outcome of the analysis is independent of different Raman scattering collection geometries. This is especially important for a quantitative comparison of micro-Raman spectra, because in such spectra the variation of the Raman intensity is higher since the focal volume of the analyte of interest is strongly varying. Here a collection of different scaling types is used. First the normalization on the l_p-Norm

$$\left\|\vec{S}\right\|_p = \sqrt[p]{\sum_{i=1}^{N} S_i^p} \tag{2}$$

Table 4
Filtering block. All techniques dealing with filtering white noise are composed.

Procedure	No.F	Parameter
RAW	00	–
Kernel smoothing	01	bw = $\Delta\tilde{\nu}$
	02	bw = $2 \cdot \Delta\tilde{\nu}$
	03	bw = $3 \cdot \Delta\tilde{\nu}$
	04	bw = $4 \cdot \Delta\tilde{\nu}$
	05	bw = $5 \cdot \Delta\tilde{\nu}$
	06	bw = $6 \cdot \Delta\tilde{\nu}$
	07	bw = $7 \cdot \Delta\tilde{\nu}$
Savitzky–Golay smoothing	08	Points = 5, degree = 2, drv = 0
	09	Points = 5, degree = 3, drv = 0
	10	Points = 7, degree = 4, drv = 0
	11	Points = 7, degree = 5, drv = 0

Table 5
Scaling block. All procedures, which try to minimize the influence of laser power fluctuation or changes in geometric properties are arranged.

Procedure	No.S	Parameter
RAW	00	–
l_p-Norm	01	$p=1$
	02	$p=2$
	03	$p=3$
	04	$p=4$
	05	$p=5$
	06	$p=10$
	07	$p=\infty$
M.-M.-Norm.	08	–
M.-V.-Norm.	09	–
Peak-Norm.	10	$\tilde{\nu} = 3000 \pm 100\,\mathrm{cm}^{-1}$

is utilized, where $\vec{S}$ stands for a Raman spectrum, i is the index of the used wavenumber point and p is a positive integer. It should be noted that for $p=\infty$ the l_p-Norm is equivalent with the maximum norm. Another widely used scaling procedure is the min–max-normalization (M.-M.-Norm.), where the minimum is scaled to 0 and the maximum to 1. The centering to zero mean and scaling to unit variance (M.-V.-Norm.) is mostly applied for derivative spectra, while in case an internal standard is present, the probably best scaling technique is the normalization on a peak (Peak-Norm.), which is connected with that standard. Here, normalization on the C–H-stretching vibration in the wavenumber region 2900–3100 cm^{-1} is used. Since the investigated datasets are of biological nature the intensity of the C–H-stretching vibration is roughly proportional to the 'biomass' and can therefore be utilized as internal standard. All of theses scaling possibilities are summarized together in Table 5.

2.4.4. Dimension reduction

Due to the fact that most statistical methods are based on optimization criteria, it is advisable to reduce the dimension of the problem. This dimension reduction results in decreasing computational costs and increasing probability of finding the best model representing the data. For this purpose often a principal component analysis (PCA) [28]

$$\mathrm{Y} = \mathrm{L} \cdot \mathrm{S}^T + \mathrm{E} \tag{3}$$

is utilized, where the last $N-n$ scores are rejected. In Eq. (3) **Y** is the matrix of spectra, **S** is the score matrix, **L** is the matrix of loadings and **E** is the error matrix. Another often used approach is the truncation of the wavenumber region to the essential region. As truncation intervals the fingerprint region 600–1800 cm^{-1}, the C–H-stretching region 2700–3200 cm^{-1} and the combination of both are utilized. In Table 6 all of these methods are listed together. For reasons, which will be explained later, the numbering is not ascending.

2.5. Statistical techniques

In the following the basic idea of genetic algorithms (GAs) for optimization problems is outlined. Thereafter, two statistical models are introduced. The first procedure is called support vector machine (SVM) and is utilized as a class modeling technique, which learns the significant difference between groups. This model can be used as a diagnostic technique. The other technique is partial-least-squares (PLS) modeling, which is used to calculate a multivariate calibration model. At the end common methods for judging the quality of such statistical methods are introduced and the used method is explained.

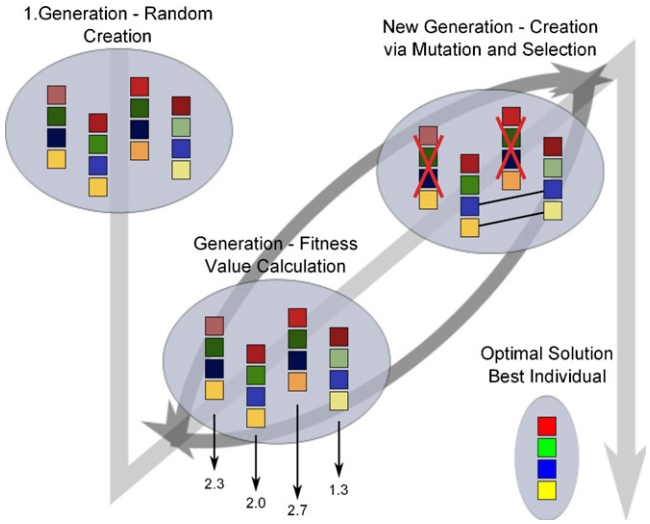

Fig. 3. *Genetic algorithm.* A genetic algorithm is a heuristic optimization procedure. The first generation is chosen randomly, then the optimization function is determining the fitness of the individuals and some are selected out, while new individuals are created by some mutations. This is done a predefined number of iterations and at the end the best individual is the optimization vector.

2.5.1. Genetic algorithms

Genetic algorithms are heuristic optimization methods, which can be applied to complex optimization problems, where basic optimization methods break down. Their mode of operation is described in detail by Lucasius and Kateman [29]. First it is necessary to define a function, which should be optimized, and a subset of the parameter space of that function. This subset is called the parameter population and the algorithm is allowed to chose a number of individuals out of this population. The optimization function is evaluated and the part of the chosen individuals, which have the lowest values, are terminated. Afterwards different mutation operators are applied and the next generation is generated. This procedure is done for a fixed number of generations and at the end the best parameter vector of the last generation represents the optimization vector (see Fig. 3).

2.5.2. Support-vector-machines

Support vector machines (SVMs) [30] are powerful statistical techniques which can be used as regression tools as well as classification tools. In the 'C-classification' mode a linear SVM tries to find the hyperplane, which is separating a dataset in two given classes and the corresponding margin is maximal. The hyperplane is con-

Table 6
Dimension reduction block. Techniques which truncate the size of the variable space are put together.

Procedure	No.D	Parameter
RAW	06	–
PCA dimension reduction	00	$d=2$
	01	$d=3$
	02	$d=4$
	03	$d=5$
	04	$d=7$
	05	$d=10$
	10	$d=50$
	11	$d=100$
	12	$d=150$
	13	$d=250$
Spectral window	07	600–1800 cm^{-1}
	08	2700–3200 cm^{-1}
	09	600–1800, 2700–3200 cm^{-1}

structed with a subset of the dataset and the spectra in this subset are called support vectors $\vec{x}^{(i)}$. The output function is given by:

$$F(\vec{x}) = \sum_{i=1}^{N} y_i \alpha_i (\vec{x}^{(i)} \cdot \vec{x}) + b, \quad (4)$$

which is converted by a threshold in a classification function. In Eq. (4) $\vec{x}$ is a spectrum and $\vec{x}^{(i)}$ are the support vectors, while y_i is the label of i-th support vector and α_i is the corresponding coefficient. Since their development in 1995, SVMs gained a lot of importance and significant improvements were made. Especially the kernel trick was incorporated, in order to allow non-linear hyperplanes. Here we focus on linear SVMs, but the generalization to non-linear models is straight forward [31].

2.5.3. Partial-least-squares modeling

Partial-least-squares modeling [21] can be applied as classification model or as multivariate calibration model. Here it is used as a calibration tool. The PLS model calculates a linear relationship between two matrices:

$$\mathrm{Y} = \mathrm{Q} \cdot \mathrm{X} + \mathrm{P}, \quad (5)$$

where **Y** and **X** are the concentration matrix and spectra matrix, respectively. The matrices **Q** and **P** are the regression coefficients. If **Y** has only one column, Eq. (5) can be interpreted as calculating the linear spectral response (rhs), if the concentration (lhs) is changed or vice-versa.

The quality of a PLS model is often measured by the mean-squared-error-of-prediction (MSEP), which reflects the averaged error rate. The PLS model is used to predict the concentrations from the corresponding Raman spectra and afterwards the differences between the predicted concentrations and the true concentrations are calculated. From this matrix it is possible to calculate different indicators for the predictive power of the model. Here we use the mean-squared-error-of-prediction (MSEP), which is the mean over all squared concentration differences.

2.5.4. Evaluation procedures

For the estimation of the prediction ability of statistical models like, SVMs or PLS models, a few estimation algorithms exist, which all feature advantages and disadvantages. The most widely used procedures are a holdout estimation and a cross validation approach [32]. In a holdout the dataset is split in a test set and a training set, where the latter is used for building the model and the performance is checked for the test set. The result of the algorithm is unbiased, but strongly dependent on the splitting. This dependency is called in-completeness of the algorithm and to get rid of that an average of a number of instances has to be done. The second algorithm presented here is the cross validation approach, which divides the data set in a number of parts called 'folds'. All folds except of one are utilized as training set with which the other fold is predicted. After leaving out every fold once, the resulting values (accuracy or MSEP) are averaged over the number of folds. This estimation procedure is biased (too optimistic), but not so dependent on the splitting. Only the leave-one-out cross validation is not depending on the splitting and therefore complete. Here we use a ten times averaged 10 fold cross validation, which was found to be a compromise between the in-completeness and the bias of the algorithm [32]. It should be mentioned that the evaluation presented in this contribution was also done for a ten times averaged holdout estimate (data not shown), but it was found that more averages have to be done in order to suppress the dependency on the splitting. In the presented work we were not using independent test sets, because the aim of this study is the investigation of the influence of the pre-processing procedure. Therefore the biological and chemical variation due to the preparation is neglected and the outcome of the models is only influenced by the pre-processing. Nevertheless, the presented methodology is capable when independent test sets are in use, but here we interpreted the cross validation estimate.

3. Results

In order to investigate the influence of the pre-processing procedure on the outcome of various statistical methods, a 'grid-search' was performed in a first step. This is the calculation of the outcome of the internal validation of a method for all possible pre-processing combinations. The grid-search is a time consuming operation and was done to judge the quality of the pre-processing optimization. Because the outcome of all possible combinations is known, it is possible to check, if the genetic algorithm converges to a pre-processing combination, which gives good results or not. Another criteria, which can be used for ranking the optimization results, is the computational time difference between the grid-search and the evaluation of the genetic algorithm. A judgment of the optimization quality is only possible if both indicators, time and optimization criteria (accuracy or MSEP), are utilized together.

In order to investigate the influence of the pre-processing procedure on multivariate calibration procedures and classification methods three datasets are investigated (see Table 2). For all three datasets, the multivariate calibration case and both classification cases, a grid-search algorithm was used to determine the accuracies or the MSEP of all possible pre-processing combinations. The grid-search was done over all possible choices of No.B, No.F, No.S, and No.D, which stand for the corresponding pre-processing procedures listed in Table 3–6, respectively. Because the tackled problems were different the region where No.D was allowed to vary was different. The calibration experiment consists of four substances, why for the PLS model also four components were chosen. This requires a dataset with minimal four variables, and therefore, No.D was restricted to No.D > 1. For the artificial and real-world classification experiment No.D < 13 was fixed, because both datasets consist of less than 250 spectra (150 spectra, 192 spectra).

3.1. Multivariate calibration model

First the influence of different pre-processing procedures on a multivariate calibration experiments is investigated (dataset 1). For this purpose a set of 1180 Raman spectra with 4 components is tested (see Table 1). The aim is to predict the concentrations of all constituents based on the corresponding Raman spectra. As analyzing technique a PLS model is applied and the quality of the model is monitored by the mean-squared-error-of-prediction (MSEP). The square root of the MSEP is a medial error of the concentration values. These concentration values are ranging from 0 to 1, while the smallest concentration was 0.085.

In Fig. 4a the ten times averaged MSEP of a 10 fold cross-validation is plotted against the ordered pre-processing index. It is obvious that a high number of pre-processing combinations exist, which are semi-optimal, e.g. the MSEP is slightly decreased compared with unprocessed spectra (indicated by a red cross). This fact results from the measurement in the cuvette, which allows no variation in the setup and light path. The scaling procedures have a higher impact for the pre-processing of spectra taken in a setup in which the focal volume of the analyte of interest is strongly varying. Hence the scaling procedure is of major importance for Raman microspectroscopy while for macro Raman setups using a cuvette it is of less importance. The calculation of all MSEPs for all pre-processing combinations took approximately 6 days, while the genetic algorithm needed 7 min for calculation and received a MSEP of 0.007. This is an improvement of 0.004 compared with the initial

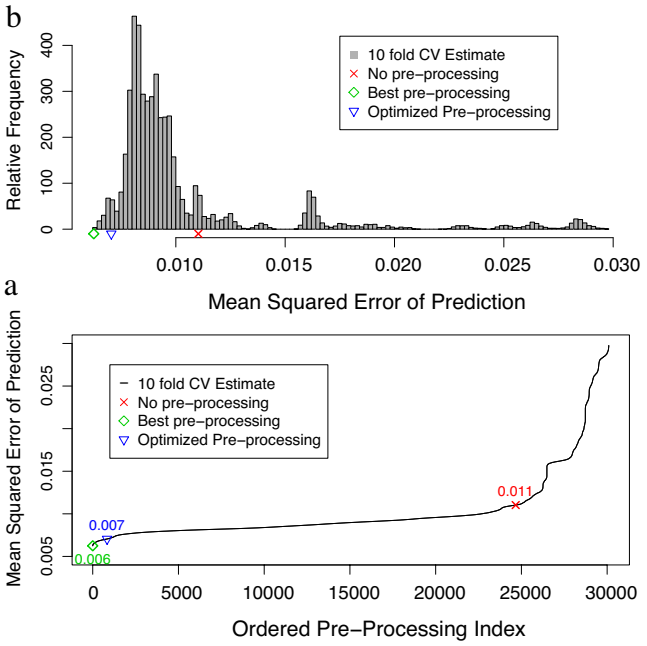

Fig. 4. (a) MSEP of the calibration experiment (dataset 1) is plotted versus the ordered pre-processing combination index. The ten times averaged 10 fold cross validated MSEP is plotted together with a holdout estimate of the MSEP. The optimal pre-processing is marked with a square, while no pre-processing is indicated by a cross. (b) Histogram of the MSEP of the calibration experiment (dataset 1) of all pre-processing combinations is visualized. It is obvious that there is a variety of quasi-optimal solutions, which is due to the experimental setup, which is not allowing variations in the geometric and physical parameters, due to the measurements in a cuvette.

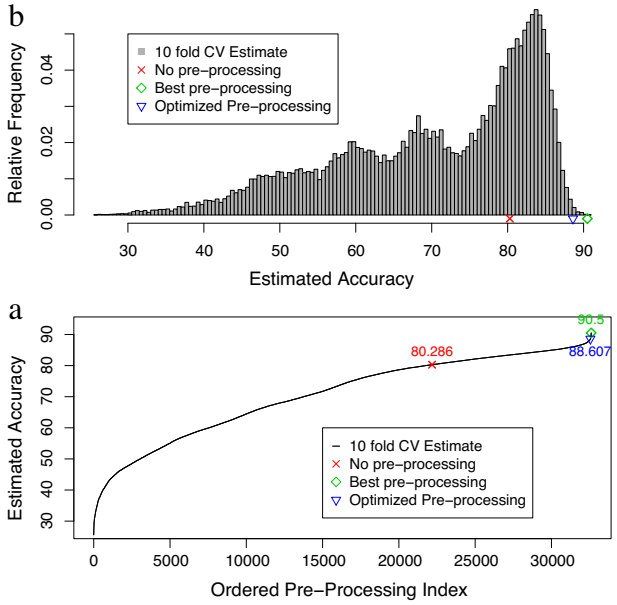

Fig. 5. (a) The accuracy of the artificial classification problem (dataset 2) is plotted versus the ordered pre-processing combination index. The ten times averaged 10 fold cross validated accuracy is plotted together with a holdout estimate of the accuracy. The optimal pre-processing is marked with a square, while no pre-processing is indicated by a cross. (b) Histogram of the mean accuracy of all pre-processing combinations is given (dataset 2). It is obvious that there are quite a lot of bad solutions, which indicate the necessity of choosing a good pre-processing by an external criteria.

MSEP of 0.011. The best solution has a MSEP of 0.006, which is in a comparable range with the optimization result.

Fig. 4b visualizes the calculated MSEPs as a histogram. The cumulation around 0.09 indicates only minor improvements, but the genetic algorithm achieved a solution, which is reasonably good (indicated by a blue triangle). The real optimal solution is only a bit better and is marked by a green square. No pre-processing results in a model, which leads to a MSEP of 0.011 and is marked by a red cross in Fig. 4b.

3.2. Classification models

In the following the impact of different pre-processing procedures on the outcome of the classification models is investigated. Therefore two classification tasks are tested. The first one, an artificial task, consists of FT-Raman spectra of three different mixtures of glucan and chitin (dataset 2). The second task consists of Raman spectra from two bacteria species, which were grown on two different media (dataset 3). As classification model a linear support vector machine was used and the model was evaluated using a ten fold cross validation. In order to produce significant results this procedure was carried out ten times and the result was averaged. All following accuracies are calculated in that way.

3.2.1. Artificial classification task

The first classification problem is inspired by the biology of fungi (dataset 2). In micro-biology it is discussed if during hyphal tip growth of fungal hyphae the glucan–chitin ratio of the cell wall changes [15]. Here, it is tested whether Raman spectroscopy together with statistical analysis is capable to distinguish between different glucan–chitin compositions. For a first try, glucan and chitin as pure substances are mixed in different ratios (1:1, 10:9, 9:10) to verify the separation potential and to investigate the best pre-processing combination for this task.

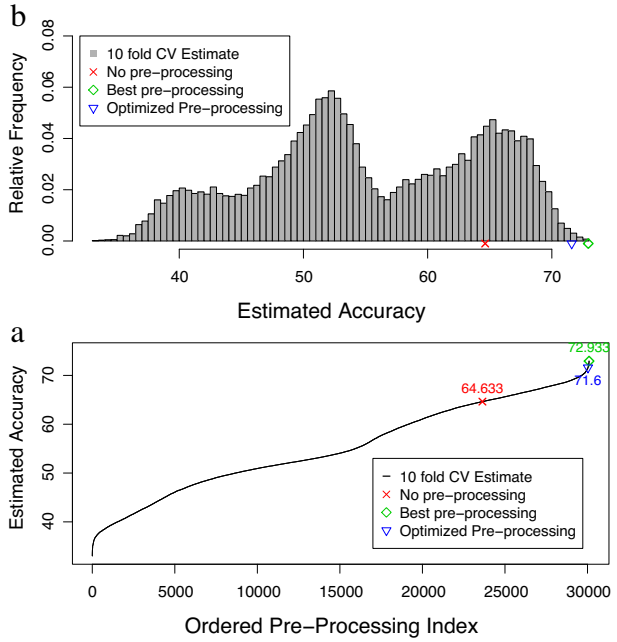

Fig. 6. (a) The accuracy of the real-world classification task (dataset 3) is plotted versus the ordered pre-processing combination index. The ten times averaged 10 fold cross validated accuracy is plotted together with a holdout estimate of the accuracy. The optimal pre-processing is marked with a square, while no pre-processing is indicated by a cross. (b) Histogram of the mean accuracy of all pre-processing combinations of the real-world classification problem (dataset 3) is visualized. It is obvious that there are a high number of bad pre-processing combinations, which indicate that selecting a pre-processing can have a negative influence on the outcome.

The calculation of all possible pre-processing combinations (grid-search) took approximately 2 days and the result of the

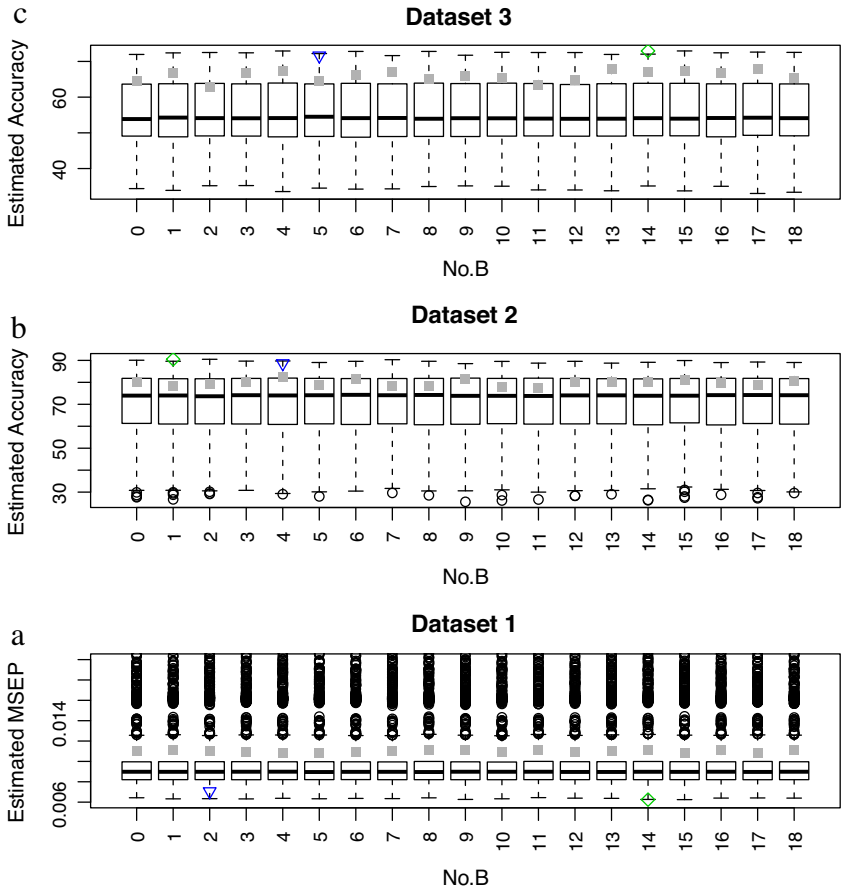

Fig. 7. A boxplot of the outcome of all three tasks is visualized. It is obvious that, a determination of the pre-processing steps subsequently, cannot be done. Because the gray squares, which correspond to the baseline correction itself, do not correlate with the best group. Also the optimized solution is not in the 'best' category, but for longer optimization times that would be the case.

grid-search algorithm is plotted in Fig. 5a. The optimization procedure needed 2 min 1 s for improving the accuracy from 80.29% for no pre-processing (red cross) to 88.61% (blue triangle). The best solution had an accuracy of 90.5% and is indicated by a green square (see Fig. 5a). From Fig. 5a it can be seen that only a small improvement is visible as compared to using the raw spectra, but a lot of combinations yield a decrease of the classification rate. This example demonstrates that false pre-processing can be negative for the outcome of a classification model.

Fig. 5b displays a histogram of the calculated accuracies. The cumulation around 84% indicates, that a lot of semi-optimal pre-processing combinations exist. The tail (below 80%) leads to pre-processing combinations, which are lowering the accuracy rate and should therefore be avoided. The optimal solution (green square) as well as the optimized result (blue triangle) are located right of the cumulation of semi-optimal solutions (see Fig. 5b).

Real-world classification task

The motivation of the second classification task (dataset 3) is the fact that the composition of lipids in *Streptomyces* varies for different growing media. *Streptomyces* produce lipids in different compositions depending on age and nutrient availability [16,17]. Since lipids have a high impact on Raman spectra the question arises if those spectra are applicable after spectral treatment for classification and identification purposes on species level.

In order to study the influence of the pre-processing approach on such a classification problem (involving two bacteria species grown on two media) a grid-search algorithm was carried out. The calculation of the outcome of all pre-processing combinations took approximately 2 days and the results are plotted in Fig. 6a. The ten times averaged estimated accuracy is ranging from 33.02%, which is almost guessing (25%), to 72.93% in the best case. The best solution is indicated by a green square in Fig. 6a. The best estimated accuracy was 72.93%, while no pre-processing yielded 64.63% classification rate (red cross in Fig. 6a).

Fig. 6b displays a histogram over all accuracies determined by the grid search. From Fig. 6b it is obvious that a lot of pre-processing combinations result in a model which features bad classification properties, e.g. the accuracy is decreased as compared to no pre-processing. The latter is indicated by a red cross in Fig. 6b. The optimized solution by GA was calculated in 4 min 36 s and resulted in an accuracy of 71.6% (blue triangle), which is almost optimal (see Fig. 6b).

3.3. Discussion

In the previous paragraph the dependency of two statistical methods on the used preprocessing was demonstrated. In the following our investigations on the independence of the pre-processing steps and the properties of the 'good' and the 'bad' pre-processing combinations are summarized.

In order to check, if the pre-processing steps are independent of each other a boxplot showing some statistical characteristics is utilized [33]. First of all the median of the MSEP or accuracy is displayed in Fig. 7 (thick line) as function of the index of the background algorithm. Additionally the lower and upper quartile, which is represented by a box, is shown. The whisker is defined by 1.5 times the inter-quartile range and consists of almost all data points,

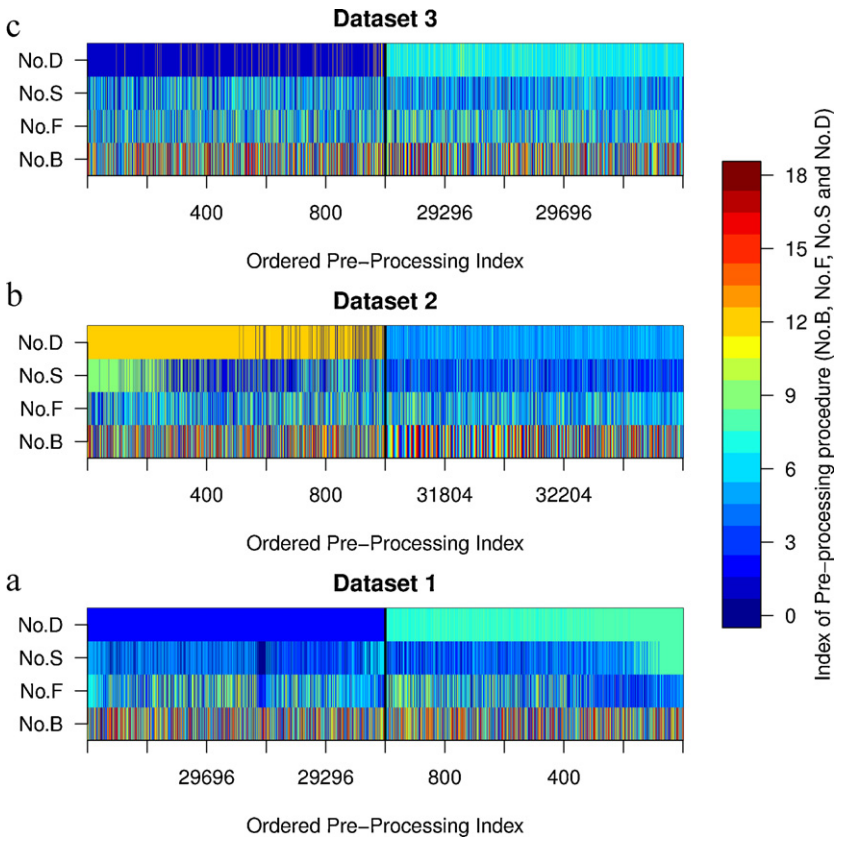

Fig. 8. The 1000 best (right side of every panel) and worst combinations (left side) for every task is visualized in false colors. It is obvious that the first three steps (background correction, filtering and scaling) do not exhibit similarities. Only in the last step (dimension reduction step) similarities can be seen, which reflects the optimal dimension of the problem (right side) and the cutting out of essential informations (left side).

the rest is treated as outlier. From Fig. 7 it is obvious that the boxes and whisker are similar, when compared with boxes and whiskers of the other underground procedures. This can be explained by the dependence of every pre-processing steps on the other three steps. Fig. 7 also depicts the optimized and the optimal solution, together with pre-processing, which only apply the background step (gray box). From this values choosing a background procedure without checking all following combinations is not possible and the pre-processing cannot be determined subsequently. Nevertheless, it can be also seen that choosing some procedure for the baseline and optimizing the last 3 steps would lead also to semi-optimal results. Also the optimized and the optimal solutions are not in the same baseline category and lead to almost the same outcome (MSEP or accuracy).

The next question, which should be addressed, is if there are similarities of the appropriate and not appropriate combinations. Therefore a visualization of the 1000 best and worst combinations is given in Fig. 8. In this false color plot the corresponding index for the steps (No.B, No.F, No.S, No.D) is coded by a unique color. At the right side the first 1000 appropriate combinations are visualized, while the left side corresponds to the non appropriate pre-processing combinations (last 1000 combinations). The good and bad ones do not feature any similarities, except of the dimension reduction step, which is obvious from the uniform color. The almost uniform coloring in the last step indicates a way to reduce the dimension, which is problem specific and is not depending on the previous steps. This is clear as every problem features a unique number of hidden independent variables in the Raman spectra and the pre-processing should extract them. This number should be used as truncation threshold for a dimension reduction based on a PCA. It is also possible to use a spectral window, which exhibits no information for the analysis task, or the threshold is too low and so the corresponding combinations (regardless of the first 3 steps) lead to a decrease of the outcome of the statistical model. This is the reason for the unique coloring on the side of the "not appropriate" combinations (left side).

Table 7
Summarized results of all three experiments.

Experiment	Type	No.B	No.F	No.S	No.D	Estimate	SD
Calibration	No pre-processing	0	0	0	6	0.0110	0.0024
	Best pre-processing	9	7	5	2	0.0062	0.0018
	Optimized pre-processing	3	5	6	8	0.0070	0.0020
Artificial classification	No pre-processing	0	0	0	6	80.29%	14.79%
	Best pre-processing	2	6	3	5	90.50%	11.78%
	Optimized pre-processing	5	5	5	5	88.61%	12.67%
Real-world classification	No pre-processing	0	0	0	6	64.63%	14.22%
	Best pre-processing	15	5	1	6	72.93%	13.77%
	Optimized pre-processing	6	5	6	6	71.6%	13.82%

Overall it was shown, that the composing of the pre-processing steps is not straight forward and every step depends on the other three steps. Therefore all steps have to be optimized in one procedure rather than subsequently. Furthermore it was shown, that for a background procedure almost all methods can be applied, as long as the following steps are chosen in an appropriate way.

4. Conclusions

Here we report about the influence of the spectral pre-processing of Raman spectra on the outcome of statistical methods. This was done by using two artificial datasets (datasets 1 and 2), where every chemical and physical parameter could be controlled and one real-world dataset (dataset 3). The work-flow started with collecting a set of pre-processing methods which were combined in a physical meaningful way. First a background correction was applied, followed by a filtering and a scaling procedure, while at the end a dimension reduction step was carried out. All possible combinations of these four procedures were investigated within the work presented here, which result in over 30,000 combinations.

In summary the pre-processing procedures have a strong influence on the outcome of the applied statistical methods (see Table 7). It was possible to enhance the model by using "appropriate" combinations and so the accuracy was increased or the MSEP was decreased. But more remarkable is the decline of the model's outcome, which was introduced by "not appropriate" pre-processing combinations. In the two classification experiments almost half of all combinations were yielding worse results as compared to no pre-processing. This fact is curious due to the design of the pre-processing procedures, which is based on physical principles. Therefore all possibilities should work quiet well, but a few combinations did not.

Because it is not known what a "good" pre-processing for a special task is and trying all combinations is too time consuming (grid-search), a methodology based on genetic algorithms is proposed, which calculates optimized pre-processing combinations for special tasks. This is done without calculating all possibilities and so a time improvement of a factor 1/1500 is gained. The solutions of the genetic algorithm are comparable to the optimal solution, e.g. the accuracy and MSEP of optimization result and optimal solution are differing only minor (see Table 7).

At the end it should be mentioned that it is also possible to optimize the pre-processing combinations for other criteria than accuracy or MSEP. For example in a classification task it can be more desirable to derive a model with a high sensitivity or selectivity rather than a high accuracy. This is the case for classification models, which are used as diagnosis models. In a calibration experiment the MSEP of one substance can be of more interest than an other, especially for drug tests where all other substances are only modeling the surrounding matrix. These cases can be also tackled by the proposed methodology, but not the accuracy must be optimized rather the sensitivity or selectivity of the model. In the multivariate calibration case the MSEP of all substances has to be switched against the MSEP of the desired substance. However, not only the quality measure can be changed, but also the estimation algorithm can be altered. It is possible to use a holdout estimator, a jackknife estimator or calculate the quality of a model with an independent test set. With the latter it is possible to design pre-processing combinations, which correct for variations in the sample not corresponding to the analysis task (e.g. not linked with substances of a multivariate calibration experiment or not connected to the groups of interest).

However, the field of application of the proposed methodology is much wider than this. It can be applied to every physical/spectroscopic measurement, where a pre-treatment is necessary and a statistical analysis should be done. Especially the methods presented in Section 1 [1–14] can be tackled by the here introduced methodology.

Acknowledgements

The funding of the research project Exprimage (FKZ13N9364) within the framework Biophotonik from the Federal Ministry of Education and Research, Germany (BMBF) is gratefully acknowledged. The comments of the unknown referees are highly acknowledged.

References

[1] E.M. Sevick-Muraca, J.C. Rasmussen, Molecular imaging with optics: primer and case for near-infrared fluorescence techniques in personalized medicine, J. Biomed. Opt. 13 (4) (2008) 041303-01–041303-16, doi:10.1117/1.2953185.
[2] F. Tiaho, G. Recher, D. Rouède, Estimation of helical angles of myosin and collagen by second harmonic generation imaging microscopy, Opt. Express 15 (19) (2007) 12286–12295, http://www.opticsexpress.org/abstract.cfm?URI=oe-15-19-12286.
[3] Y. Sulub, G.W. Small, Spectral simulation protocol for extending the lifetime of near-infrared multivariate calibrations, Anal. Chem. 81 (3) (2009) 1208–1216, http://dx.doi.org/10.1021/ac80174n.
[4] T. Bocklitz, M. Putsche, C. Stüber, J. Käs, A. Niendorf, P. Rösch, J. Popp, A comprehensive study of classification methods for medical diagnosis, J. Raman Spectrosc. 40 (2009) 1759–1765, http://dx.doi.org/10.1021/ac80174n.
[5] M. Harz, M. Kiehntopf, S. Stöckel, P. Rösch, E. Straube, T. Deufel, J. Popp, Direct analysis of clinical relevant single bacterial cells from cerebrospinal fluid during bacterial meningitis by means of micro-Raman spectroscopy, J. Biophoton. 1 (2009) 1–11.
[6] S. Tschierlei, B. Dietzek, M. Karnahl, S. Rau, F.M. MacDonnell, M. Schmitt, J. Popp, Resonance Raman studies of photochemical molecular devices for multielectron storage, J. Raman Spectrosc. 39 (5) (2008) 557–559, http://dx.doi.org/10.1002/jrs.1954.
[7] A. März, K.R. Ackermann, D. Malsch, T. Bocklitz, T. Henkel, J. Popp, Towards a quantitative SERS approach – online monitoring of analytes in a microfluidic system with isotope-edited internal standards, J. Biophoton. 2 (2009) 232–242, doi:10.1002/jbio.200910069.
[8] K. Hering, D. Cialla, K. Ackermann, T. Dörfer, R. Möller, H. Schneidewind, R. Mattheis, W. Fritzsche, P. Rösch, J. Popp, SERS: a versatile tool in chemical and biochemical diagnostics, Anal. Bioanal. Chem. 390 (2008) 113–124.
[9] R. Böhme, M. Richter, D. Cialla, P. Rösch, V.D. und, J. Popp, Towards a specific characterisation of components on a cell surface – combined TERS-investigations of lipids and human cells, J. Raman Spectrosc. 40 (2009) 1452–1457, doi:10.1002/jrs.2433.
[10] E. Bailo, V. Deckert, Tip-enhanced Raman scattering, Chem. Soc. Rev. 37 (2008) 921–930, doi:10.1039/b705967c.
[11] T. Meyer, D. Akimov, N. Tarcea, S. Chatzipapadopoulos, G. Muschiolik, J. Kobow, M. Schmitt, J. Popp, Three-dimensional molecular mapping of a multiple emulsion by means of CARS microscopy, J. Phys. Chem. B 113 (2008) 1420–1426, doi:10.1021/jp709643h.
[12] 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) 1626–1637.
[13] A. Walter, S. Erdmann, T. Bocklitz, E.-M. Jung, N. Vogler, D. Akimov, D. Dietzek, P. Rösch, E. Kothe, J. Popp, Analysis of the cytochrome distribution via linear and nonlinear Raman spectroscopy, Analyst 135 (2010) 908–917, doi:10.1039/B921101B.
[14] U. Neugebauer, J.H. Clement, T. Bocklitz, C. Krafft, J. Popp, Identification and differentiation of single cells from peripheral blood by Raman spectroscopic imaging, J. Biophoton. 3 (2010) 579–587, doi:10.1002/jbio.2010000201.
[15] G. Steinberg, Hyphal growth: a tale of motors, lipids, and the Spitzenkorper, Eukaryotic Cell 6 (3) (2007) 351–360, arXiv:http://ec.asm.org/cgi/reprint/6/3/351.pdf, doi:10.1128/EC.00381-06, http://ec.asm.org.
[16] Y. Li, G. Florova, K.A. Reynolds, Alteration of the fatty acid profile of *Streptomyces coelicolor* by replacement of the initiation enzyme 3-ketoacyl acyl carrier protein synthase III (FabH), J. Bacteriol. 187 (11) (2005) 3795–3799, arXiv:http://jb.asm.org/cgi/reprint/187/11/3795.pdf, doi:10.1128/JB.187.11.3795-3799.2005, http://jb.asm.org/cgi/content/abstract/187/11/3795.
[17] C. Schauner, A. Dary, A. Lebrihi, P. Leblond, B. Decaris, P. Germain, Modulation of lipid metabolism and spiramycin biosynthesis in *Streptomyces ambofaciens* unstable mutants, Appl. Environ. Microbiol. 65 (6) (1999) 2730–2737, arXiv:http://aem.asm.org/cgi/reprint/65/6/2730.pdf, http://aem.asm.org/cgi/content/abstract/65/6/2730.
[18] C. Krafft, B. Dietzek, J. Popp, Raman and CARS microspectroscopy of cells and tissues, Analyst 134 (6) (2009) 1046–1057, doi:10.1039/b822354h.
[19] R Development Core Team, in: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2007, http://www.R-project.org.

[20] S. original by Matt Wand. R port by Brian Ripley, KernSmooth: Functions for Kernel Smoothing for Wand & Jones (1995), R package version 2.22-22, 2008.
[21] R. Wolthuis, G.C.H. Tjiang, G.J. Puppels, T.C.B. Schut, Estimating the influence of experimental parameters on the prediction error of pls calibration models based on Raman spectra, J. Raman Spectrosc. 37 (2006) 447–466.
[22] W.N. Venables, B.D. Ripley, in: Modern Applied Statistics with S, 4th edition, Springer, New York, 2002, http://www.stats.ox.ac.uk/pub/MASS4.
[23] W.R. Mebane Jr., J.S. Sekhon, Rgenoud: R version of GENetic Optimization Using Derivatives, R package version 5.4-7, 2007. http://sekhon.berkeley.edu/rgenoud/.
[24] M. Morhac, Peaks: Peaks, R package version 0.2 (2008).
[25] C. Ryan, E. Clayton, W. Griffin, S. Sie, D. Cousens, Snip, a statistics-sensitive background treatment for the quantitative analysis of pixe spectra in geosience aplications, Nucl. Instrum. Methods Phys. Res. B 34 (1988) 396–402.
[26] C.A. Lieber, A. Mahadevan-Jansen, Automated method for substraction of fluorescence from biological Raman spectra, Appl. Spectrosc. 57 (2003) 1363–1367.
[27] A. Savitzky, M. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[28] K. Pearson, On lines and planes of closest fit to systems of points in space, Philos. Mag. 2 (1901) 559–572.
[29] C.B. Lucasius, G. Kateman, Understanding and using genetic algorithms. Part 1. Concepts, properties and context, Chemometr. Intell. Lab. Syst. 19 (1) (1993) 1–33, doi:10.1016/0169-7439(93)80079-W, http://www.sciencedirect.com/science/article/B6TFP-44GGKJ0-95/2/2741ca4b638885ebe4b95e259f22796f.
[30] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[31] T. Dörfer, W. Schumacher, N. Tarcea, M. Schmitt, J. Popp, Quantitative mineral analysis using Raman spectroscopy and chemometric techniques, J. Raman Spectrosc. 41 (2010) 684–689.
[32] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1995, pp. 1137–1143.
[33] M. Frigge, D.C. Hoaglin, B. Iglewicz, Some implementations of the boxplot, Am. Stat. 43 (1) (1989) 50–54, http://www.jstor.org/stable/2685173.