Improving discrimination of Raman spectra by optimising preprocessing strategies on the basis of the ability to refine the relationship between variance components

Improving discrimination of Raman spectra by optimising preprocessing strategies on the basis of the ability to refine the relationship between variance components

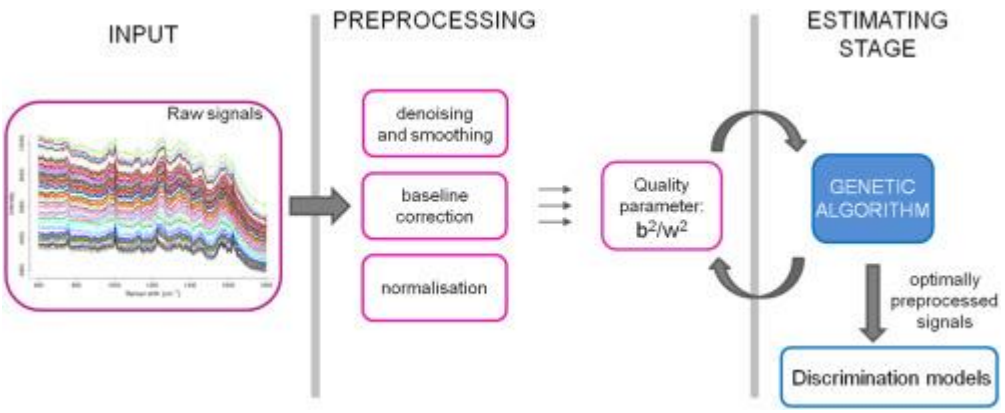
Author links open overlay panel[AgnieszkaMartyna](#)[a][AlicjaMenżyk](#)[a][AlessandroDamin](#)[b][AleksandraMichalska](#)[c][GianmarioMartra](#)[b][EugenioAlladio](#)[bde][GrzegorzZadora](#)[ac]
Show more

Abstract

Discrimination of the samples into predefined groups is the issue at hand in many fields, such as medicine, environmental and forensic studies, etc. Its success strongly depends on the effectiveness of groups separation, which is optimal when the group means are much more distant than the data within the groups, i.e. the variation of the group means is greater than the variation of the data averaged over all groups. The task is particularly demanding for signals (e.g. spectra) as a lot of effort is required to prepare them in a way to uncover interesting features and turn them into more meaningful information that better fits for the purpose of data analysis. The solution can be adequately handled by using preprocessing strategies which should highlight the features relevant for further analysis (e.g. discrimination) by removing unwanted variation, deteriorating effects, such as noise or baseline drift, and standardising the signals. The aim of the research was to develop an automated procedure for optimising the choice of the preprocessing strategy to make it most suitable for discrimination purposes. The authors propose a novel concept to assess the goodness of the preprocessing strategy using the ratio of the between-groups to within-groups variance on the first latent variable derived from regularised MANOVA that is capable of exposing the groups differences for highly multidimensional data. The quest for the best preprocessing strategy was carried out using the grid search and much more efficient genetic algorithm. The adequacy of this novel concept, that remarkably supports the discrimination analysis, was verified through the assessment of the capability of solving two forensic comparison problems - discrimination

between differently-aged bloodstains and various car paints described by Raman spectra - using likelihood ratio framework, as a recommended tool for discriminating samples in the forensics.

Graphical abstract



Keywords

Signals preprocessing
Regularised MANOVA
Discrimination
Raman spectra
Likelihood ratio

1. Introduction

Discrimination of the samples into predefined categories (groups, classes) is one of the leading issues in chemometric analysis in the field of food analysis, environmental studies, medical applications, forensics, etc. The aim is to develop the rules for assigning new samples for which the group membership is unknown, based on a few latent variables (e.g. linear combinations of original variables) summarising multivariate data structure. The latent variables are found to expose the groups separation, which is optimal when the group means are much more distant than the data within the groups, i.e. the variation of the group means is greater than the variation of the data in the groups, averaged over all groups. There are numerous

methods routinely used for discrimination purposes such as linear discriminant analysis, partial least squares discriminant analysis, logistic regression, to name a few [1].

Effective data grouping attracts considerable interest also in the forensics if the task is to assess whether the two fragments of evidence materials collected during the criminal investigations, such as car paints, glass fragments, polymer materials etc., may be two pieces of the same object, called the *source*. Comparing the features of the recovered sample, coming from an unknown source, and control sample, from the known source, helps to establish the links between the suspect, victim and the crime place. Concluding on common, or uncommon, source of samples is actually similar to the concept of discrimination since the task is to judge if the recovered sample features resemble the features of a particular source so much that it can be considered as originating from this source. Conclusions are drawn in the light of features describing other available potential sources of the recovered material, e.g. collected in a database storing the characteristics of a variety of samples of this material. Reliable assessment of the samples similarity is successful only when the sources are uniquely defined, i.e. means of the features, characterising the sources, are sufficiently distinct (i.e. between-source variation is maximised, b^2) and the variation of the data within each source (w^2) is minimised. The task, however, differs from the classical discrimination in that it is only decided if the recovered sample may share the same origin with the indicated source and it does not assign the membership to any other remaining sources. Even though one may argue that this is rather a classification issue, it is not, as the other sources are also clearly defined. Moreover, the match between the compared materials is always judged on the basis of both the similarity and uniqueness of their features (section 2.5) in regard to similarity and uniqueness of features in other available sources.

Evidence materials are typically analysed by spectroscopic or chromatographic methods and thus characterised by signals such as spectra or chromatograms. Despite the ease of visualisation, such data requires a lot of effort to uncover interesting features and turn them into more meaningful information that better fits for the purpose of data analysis. This applies above all to appropriately tailored preparation of the signals, called preprocessing [[2], [3], [4]], and then adequate data dimensionality reduction, since working with lower-dimensional data is advisable to

reveal interesting features. The aim of preprocessing is to highlight the features relevant for further analysis, e.g. discrimination, by removing unwanted variation, deteriorating effects, such as noise or baseline drift, and standardising the signals. It consists of denoising, smoothing, baseline correction and normalisation/scaling/standardisation. Adequate choice of the preprocessing strategy is a key to improve statistical models performance. However, there is no optimal preprocessing strategy as it is heavily dependent on the data and the purpose of the analysis.

Engel et al. [2] aptly summarised the paths for optimisation of the preprocessing strategy. As mentioned, attempts for choosing the optimal preprocessing strategy are often limited to visual inspection of the signals graphical representation. The preprocessing strategy is then deemed satisfactory if the picture looks more legible (e.g. certain features unique for the groups are more noticeable) and unwanted artifacts are effectively eliminated. This tactics is subjective, user-dependent and does not guarantee that the most appealing results will also prove well for statistical models. The optimal strategy may also be the one producing the data for which best performing statistical models (regression, discrimination, classification, etc.) are constructed. This approach, however, is time-consuming and computationally demanding as it requires training, validating and testing of the statistical models. Therefore an objective criterium based on *quality parameters* may be proposed as an alternative. Quality parameters can be considered markers that quantify the preprocessing strategy effectiveness, i.e. evaluate the suitability of the data for the purpose of further analysis based on the experts experience. The optimal preprocessing solution is found when quality parameters take their extremes (maximum or minimum).

A recent review of the literature on the area of preprocessing optimisation revealed that many researchers have undertaken this issue using either the grid search process, where a defined quality parameter is computed for each preprocessing strategy, or using less time-consuming heuristic alternative such as genetic algorithms [[5], [6], [7]] (section 2.3), which do not try out every strategy to find the most promising strategy for the purpose of their analyses [[8], [9], [10]]. In both concepts the optimal strategy is found as the one yielding the best quality parameter. There are numerous attempts to design the quality parameters to measure the effectiveness of

the preprocessing. Their main downside, however, is that they might not entirely be suitable for discrimination purposes.

We offer a novel concept that remarkably supports the discrimination analysis of the signals owing to appropriately conducted optimisation of the preprocessing strategy. Our idea is to define the quality parameter as a ratio of the between-source and within-source variation ($b2/w2$) for the preprocessed data to select the preprocessing strategy that best exposes the differences between sources (i.e. groups) and minimises the casual variations within sources. $b2/w2$ will be estimated from regularised MANOVA (rMANOVA [11]) which defines a limited number of latent variables that maximise the ratio of between-source variance and the within-source variance. In this sense, rMANOVA reduces data dimensionality in a way that is beneficial for the data analysis goal, i.e. discrimination. Regularisation of the method makes it feasible for handling singularity problems of variance-covariance matrices for highly multidimensional data. The grid search process as well as the genetic algorithm are used to find the optimal strategy. The adequacy of the results found in both approaches is judged by evaluating the performance of the statistical likelihood ratio models (LR, section 2.5) [[12], [13], [14]] for concluding if the samples may share common origins. Fig. 1 briefly summarises this concept.

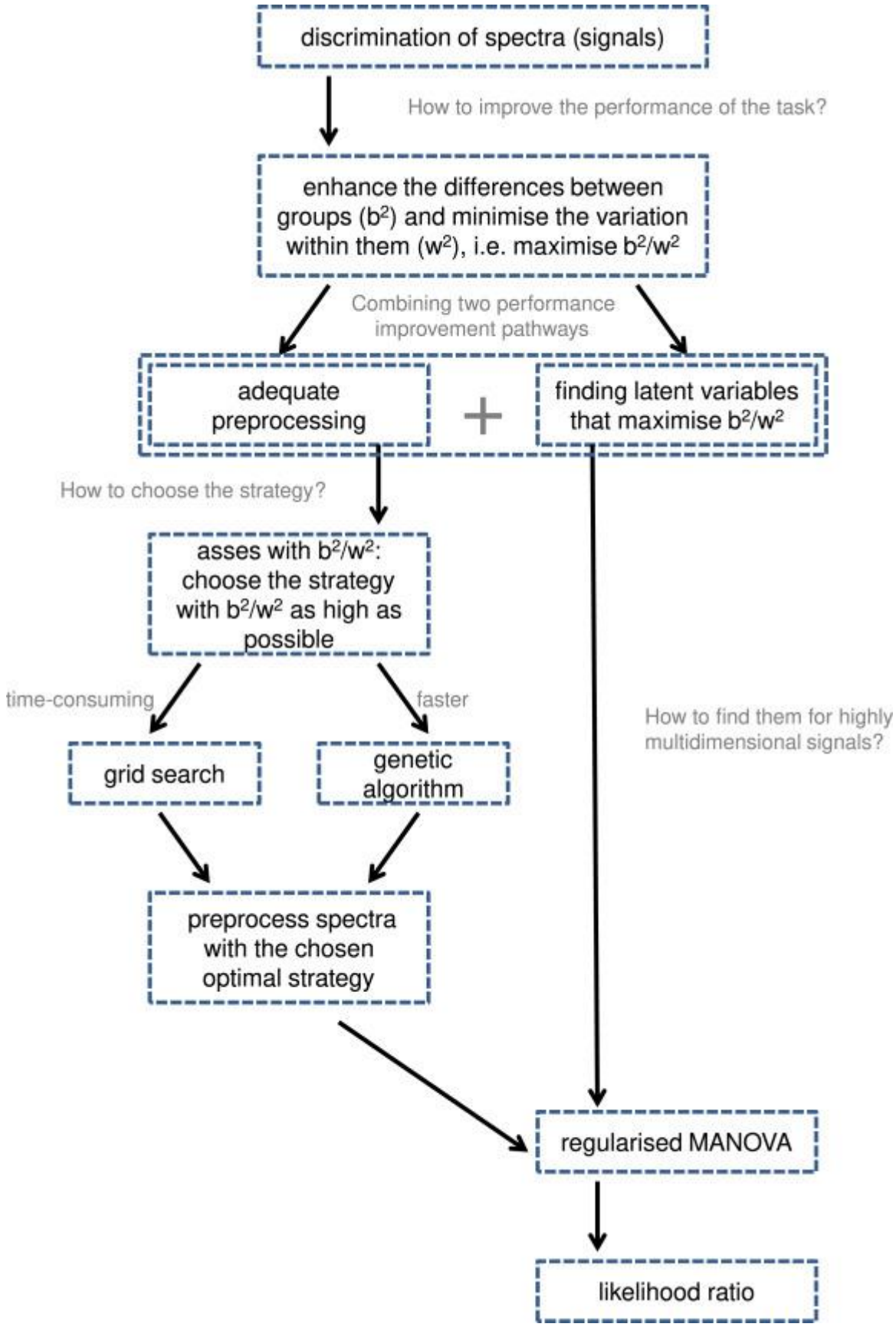

discrimination of spectra (signals)
How to improve the performance of the task?
enhance the differences between groups (b^2) and minimise the variation within them (w^2), i.e. maximise b^2/w^2
Combining two performance improvement pathways
adequate preprocessing
+
finding latent variables that maximise b^2/w^2
How to choose the strategy?
asses with b^2/w^2: choose the strategy with b^2/w^2 as high as possible
time-consuming
faster
grid search
genetic algorithm
How to find them for highly multidimensional signals?
preprocess spectra with the chosen optimal strategy
regularised MANOVA
likelihood ratio

Fig. 1. The concept of the studies.

The need to link signal preprocessing strategies with reducing their dimensionality in a way that maximises differences between groups and minimises differences within them has already been raised by the authors, e.g. in data analysis for the forensic aims [15,16]. In these studies the preprocessing strategies dealt mostly with fluorescence background in Raman spectra of car paints but no attention was paid to choose these which maximise b2 and minimise w2. This task was accomplished in a separate step. These aspects also apply to other research fields and thus the proposed framework may be found useful not only in the forensics but also medical, environmental and food analysis applications, where the grouping of signals is the issue at hand.

2. Materials and methods

2.1. Samples

This study attempted to facilitate the solution of two distinct forensic problems - one of them involving the discrimination between differently-aged blood traces, and the other connected with differentiating car paint samples. Both data sets consisted of Raman spectra, which were often obscured by the strong fluorescence interference. Raman spectroscopy is a powerful technique providing an insight into the molecular structure and functional groups, which in contrast to infrared spectroscopy, is not limited by the presence of water in biological samples. For this reason Raman spectra are frequently registered for samples with the aim of their differentiation not only in the forensics but also medical, environmental and biological applications.

2.1.1. Blood traces

Estimation of bloodstains age is one of the most challenging (and hence still unsolved) forensic task. Once the bloodstain is created, a cascade of physicochemical processes takes place, which include hemoglobin as the dominant component of dried red blood cells [17,18], leading to changes of bloodstains' properties. These changes can be tracked using e.g. Raman spectroscopy and subsequently used for distinguishing between differently-aged bloodstains [19].

Bloodstains used in this study were created by depositing 20 μl aliquots of capillary blood without preservatives originating from a single donor (to reduce the inter-

personal blood composition variations) on aluminum sample pans, that do not give Raman signal. Bloodstains were left to dry for 2 h before first spectrum collection, in stable laboratory conditions (temperature: 23.6±2.0 °C, relative humidity: 30±4%) and stored for the next three weeks. Samples were analysed every 2 h (from two up to 8 h elapsed since bloodstain formation, when the degradation process is remarkably fast) and then almost daily for the period of three weeks. In each of 18 time points, assumed to constitute 18 different evidence time-related sources, the bloodstains were measured six times [19]. The task is to judge if the features of the recovered bloodstain are close enough to the features of a bloodstain of a known age (time-related source) to conclude that their age is the same.

The spectra were recorded in the range 300–1800 cm^{-1} using a Renishaw inVia Raman Microscope spectrometer with near infrared semiconductor laser (785 nm) as an excitation source and Peltier-cooled charge-coupled device (CCD). The laser beam was focused on the samples surface through 5x NIR optimized objective (N.A. = 0.1), the final power density at the sample being so 0.16 $mW/\mu m^2$ (about 10% of the total emission and considering a spot with diameter of 9.57 μm). The Raman spectra were recorded using rotating mode to prevent sample damage due to excessive point laser irradiation [19].

2.1.2. Car paints

The aim of comparing features of car paints is to establish a link between e.g. car and the victim in hit-and-run cases. The task is to judge if the paint features are close enough to the features of a particular source that it can be considered as originating from this source. 30 blue solid car paints, assumed to constitute 30 different evidence sources were subjected to Raman analysis. Each sample was measured in situ three times in three different locations [16]. Raman spectra were recorded in the range 200–2500 cm^{-1} using Renishaw inVia Raman Microscope spectrometer with near infrared semiconductor laser (785 nm) as an excitation source and Peltier-cooled charge-coupled device as a detector. The laser beam was focused on the samples surface through 50x N Plan objective (N.A. = 0.75), the final power density at the sample being so 0.52 or 0.26 $mW/\mu m^2$ (about 1% or 0.5% of the total emission and considering a spot with diameter of 1.28 μm).

2.2. Preprocessing methods

This sections provides a brief summary of the applied preprocessing methods. We did not intend to review the methods, but only introduce them and provide adequate bibliography positions for the readers who might not be familiar with them. Throughout this section the signals subjected to any of the preprocessing steps will be vectors $s=(s1,s2,\ldots,sJ)$.

2.2.1. Denoising and smoothing

Noise is an inherent component of any measured signal. Denoising and smoothing of the signals are widely applied to handle various noise types. Smoothing is used for removing high frequency components while denoising eliminates only the signal components with a limited amplitude. The aim of both is to make the signals more legible and visually pleasing.

Savitzky-Golay filter. The method is well adapted both for smoothing and differentiation of the signals [20]. For a subset of signal points, called window, least squares procedure is applied for fitting a low degree polynomial to smooth the signal. A fitted polynomial value is kept for a central point of the window. The window is then shifted one point and the fitting is repeated until the window moves to the end of the signal.

Discrete wavelet transform, DWT. Wavelet transform (WT), like Fourier transform (FT), assumes that noise, baseline and true signal components are well separated in the frequency domain. This is because usually baseline varies at the lowest rates, whilst the frequency of signal noise is the highest. Unlike FT, which represents the signal as a linear combination of sinusoids and cosinusoids only, WT engages a great variety of wavelet functions (e.g. Daubechies [21], Coiflet, Symmlet) localised in time and frequency. WT is therefore more efficient as it requires much less wavelets to reproduce the signal than FT.

Wavelet transform projects the signal onto the basis of functions - wavelets. They are derived from one function called *mother wavelet* Ψ by its dilation or contraction in the frequency domain (controlled by scaling parameter a) and shifting in the time domain

(determined by localisation parameter *b*) to cover the whole frequency and time information:(1)$\Psi(x)=a-1/2\Psi(x-ba),a,b\in R,a\neq 0.$
Restricting scaling parameter *a* to $2j$ and localisation parameter *b* to $2jk$, with *j* being the resolution or decomposition level, is the core concept of the discrete wavelet transform, DWT.
It is convenient to demonstrate DWT in the form of Mallat pyramid algorithm [22] as a series of low and high pass filters applied to the analysed signal. High pass filter, H, defined by mother wavelet, extracts the highest frequencies in the signal, usually associated with noise fraction. Low pass filter, L, fixed by scaling function, passes lower frequencies containing baseline and true signal. The output of H is a set of details coefficients (Wj) mostly representing the high frequency noise. L generates approximation coefficients (Vj) portraying the smoothed signal, deprived of noise. At each level *j* the details part is kept and the approximations are decomposed using the same pair of filters into the approximation and details part of twice lower resolution.
DWT found a variety of applications in analytical chemistry [23] and until today it is widely applied for smoothing (removing high frequency coefficients) and denoising (removing only the coefficients with a limited amplitude) since the details coefficients attributed to highest frequencies may be easily suppressed [24,25]. For denoising the truncation of details coefficients is usually applied using hard or soft thresholding policies. Hard thresholding sets all the coefficients absolute values below a threshold value *t* to 0 and keeps the remaining:(2)$Whardj=\{0,\text{if } |Wj|<tWj,\text{if } |Wj|\geq t.$
In soft thresholding the coefficients absolute values below the threshold are set to 0 and the remaining are suppressed by this value:(3)$Wsoftj=\{0,\text{if } |Wj|<tsgn(Wj)(|Wj|-t),\text{if } |Wj|\geq t.$
t may be computed using a variety of possibilities, briefly summarised e.g. in Ref. [24]. Universal threshold is one of the most commonly applied:(4)$t=s2\log N,$where *s* is the measure of the *N* wavelet coefficients dispersion expressed as their standard deviation or more robustly with median absolute deviation ($1.4826\cdot MAD(W)$).
Once denoised, the signal is reconstructed using inverse DWT.

2.2.2. Baseline correction

Raman spectra are often corrupted by broad and intense bands of fluorescence which is a competing process to relatively weak Raman scattering effect. If fluorescence is more intense than the true Raman signal and obscures the Raman peaks, some experimental techniques applied during signal collection (photobleaching process, fluorescence quenching, removal of fluorophores, changing the laser source or using time gated Raman spectroscopy and resonantly enhanced Raman scattering technique) should be applied [26]. Baseline effects arising, among others, due to fluorescence, that do not cover the true Raman signal totally, may be appropriately handled either during the signal collection or using computational methods after signal collection, concisely described in this section.

Polynomial methods. The traditional polynomial methods for baseline correction fit the polynomial curve to the user defined baseline points using least squares method. As laborious, highly subjective and immensely time-consuming procedure, especially facing the ease of measuring vast amount of data that need fast and effective preprocessing, it was upgraded by the automated methods such as modified polyfit (ModPoly) [27] and improved modified polyfit (IModPoly) [28]. In ModPoly procedure the polynomial (w) of a fixed but adjustable degree is initially fit to the original signal in a least squares manner. This obviously involves both the baseline and signal peaks and requires a modification to eliminate the true signal (peaks) from the fit. For this purpose peaks are gradually eliminated in an iterative process, where in each turn polynomial fitting is applied to a new signal generated as the minimum between the polynomial fitted in the previous round and the original signal. The procedure is repeated until convergence, when further iterations (t) do not improve the fitting, i.e. $|(wt-wt-1)/wt-1|<0.01$, or maximum number of iterations is reached.

For noisy signals the results of ModPoly may appear inadequate as noise regions may imitate the signal. Moreover, the method is prone to variations for signals with a few major peaks, which take the control over the entire polynomial fitting. To address these limitations IModPoly algorithm removes the major peaks in the first iteration and iteratively composes the baseline with a slight modification in regard to ModPoly. In each iteration it fits a polynomial to the signal being the minimum of the signal to which the polynomial was fitted in the previous round and this polynomial plus the standard deviation of the least squares model residuals as a measure of noise level (DEV). When the procedure converges the baseline is interpolated in the major peaks

regions and subtracted from the original signal. The convergence is reached when in two subsequent iterations (t) $|(DEVt-DEVt-1)/DEVt|<0.01$ or maximum number of iterations is reached.

Asymmetric penalised least squares methods. The foundations for these methods are borrowed from Whittaker smoothing algorithm [29,30]. It is a procedure that smooths the signal by controlling the balance between two conflicting goals, the fidelity of the smoothed curve to the signal and its roughness [30]. The fidelity is a lack of fit measured as the sum of squared differences between the smoothed curve (z) and the signal (s):(5)$F=\sum_{i=1}^{J}(s_i-z_i)^2$.

The roughness of the curve is quantified by computing the squared sum of differences between neighbouring points:(6)$R=\sum_{i=1}^{J-1}(z_i-z_{i+1})^2=\sum_{i=1}^{J-1}(\Delta z_i)^2$.

Most often, however, squared second differences are applied. In its most general form for m-th differences Equation (6) becomes $R=\sum_{i=1}^{J-m}(\Delta^m z_i)^2$.

z is found with penalised least squares to minimise the expression(7)$Q=F+\lambda R$,where λ is the penalty arbitrary assigned by the user. λ is a tuning parameter to control the contribution of the roughness term to Q and makes z smoother as λ grows at the expense of fidelity.

To adapt this method for baseline estimation, z has to be found to fit the baseline regions only, excluding the signal peaks. For this purpose appropriate asymmetric weights w_i are introduced that weigh the positive deviations from the baseline estimate (mostly peaks) much less than the negative deviations. The fidelity is then modified to(8)$F=\sum_{i=1}^{J}w_i(s_i-z_i)^2=(s-z)^T W(s-z)$in matrix notation, where W is a diagonal $J\times J$ matrix with w on the diagonal. Once the solution is found for the system of equations(9)$(W+\lambda D^T D)z=Ws$(where D is the difference matrix, $Dz=\Delta z$) using initial weights, the weights can be updated and the procedure continues until convergence, when the weights cease to change and the baseline estimate is no longer significantly improved. The final baseline is computed from(10)$z=(W+\lambda D^T D)^{-1}W s$and then subtracted from the signal.

There are many asymmetric least squares methods differing in the way the weights are assigned. The most trivial assigns small p or large $1-p$ weights for peak regions (when $s_i>z_i$) and baseline segments (when $s_i\leq z_i$), respectively:(11)$\{w_i=p,\text{if } s_i>z_i w_i=1-p,\text{if } s_i\leq z_i$.

The method converges when weights do not change in two subsequent iterations or maximum number of iterations is reached.

In [31] the authors propose an automatic weights assignment in an adaptive iteratively reweighted penalised least squares (airPLS) algorithm. Here the weights depend on the previous baseline approximation and are iteratively recomputed to eliminate peaks from baseline estimation. In *t*-th iteration the weights are given as:(12){wit=0,if si≥zit−1wit=exp(t|si−zit−1|/|dt|),if si<zit−1,where dt contains the negative s−zt−1 values. The idea is to assign 0 weight for peaks regions to totally eliminate them from the baseline estimation. The method converges when |dt|<0.001·|s| or maximum number of iterations is reached.

Informative peak regions may also be identified using continuous wavelet transform (CWT) as suggested in Ref. [32]. CWT using the Haar wavelet proved to be successful in establishing an exact position and width of the peaks. The terminal points of the peaks are connected by a straight line and the PLS algorithm is applied for estimating the baseline in the remaining segments.

The concept of stiffness of the estimated baseline in the peak regions is followed in Ref. [33] in the method referred to as doubly weighted spline. The method assumes that the roughness term should more contribute to the baseline estimation in peak regions than in baseline segments. Thus maximum stiffness γmax is assigned to peak regions and takes minimum γmin for baseline regions. Instead of Equation (7), the cost function to be minimised is then expressed as:(13)Q=∑i=1Jwisi−zi2+γmax∑i=1J−m1−ηwiΔmzi2,where η=(γmax−γmin)/γmax and the weights are expressed according to Equation (12). The method converges when |dt|<0.001·|s| or maximum number of iterations is reached.

The asymmetric penalised least squares algorithm published in Ref. [34] for baseline estimation should receive special attention due to its ability to reduce the variations between replicate signals after the baseline correction. The core concept of this methodology is the clever introduction of an additional penalty to penalise remarkable differences between the corrected replicate signals, which should be obviously as similar as possible after correction of the baseline.

Statistics-sensitive non-linear iterative peak-clipping, SNIP. Originally proposed for correcting baseline effects in PIXE spectra of geological samples [35], SNIP proved

to be an efficient method for handling baseline variations for other signals as well. The algorithm is initialised with a low statistics digital filter to account for possible large differences in signal magnitude and transforms each signal intensity according to the equation $y_i=\log(\log(s_i+1)+1)$. The baseline is estimated then in an iterative process from 1 to *W* iterations, where *W* is the size of the *clipping window*. In *w*-th iteration each intensity point, y_i, becomes a central $2w+1$-length interval point, which is replaced by a minimum of the mean of intensities at the both interval ends and the point intensity itself: $g_i=\min(y_i,1/2\cdot(y_{i+w}+y_{i-w}))$. The final baseline is estimated using inverse transform $z_i=\exp(\exp(g_i-1)-1)$ and subtracted from the signal.

Quantile regression based methods. Polynomial or spline quantile regression (QR) methods fit the baseline when small quantiles are assumed (e.g. 0.01) [36]. The methods may also be upgraded in weighted quantile regression models with weights automatically and iteratively assigned according to Equation (12).

Robust baseline estimation, RBE. RBE [37], that is closely related to LOWESS procedure (locally weighted scatter plot smoother [38]), assumes that (i) the signal points in peak regions are outliers in regard to the ordinary points that belong to baseline segments and (ii) that with an undefined functional shape, the baseline can only be estimated locally, in adequately small fragments using e.g. linear models. To meet the above assumptions robust local regression methods are most suitable that will robustly approximate the baseline only in small signal fractions ignoring the points in peak regions. These signal fractions are specified by kernel functions. The residuals of the local regression models are then used for establishing small weights (Equation (14)) for signal points with large residuals (peaks) and unit weights for baseline region points.(14)$$\{w_i=\max 1-s_i-z_i/\sigma b2,02,\text{if } s_i-z_i/\sigma\geq 0 w_i=1,\text{if } s_i-z_i/\sigma<0,$$where *b* is the robustness parameter that controls the influence outliers and ordinary points have on baseline estimation, σ is the scale parameter estimated as median absolute deviation, $\sigma=1.4826\cdot\text{median}(|s_i-z_i|)$. The baseline is then iteratively recomputed using weighted least squares regression models with kernels until convergence.

2.2.3. Normalisation

The compulsion for normalisation arises from registering signals under unstable conditions, such as fluctuating laser power in Raman spectroscopy. Thus in most cases normalisation relies on multiplying the signal by a scaling value to make the

corresponding intensities comparable across spectra which should not theoretically pose any differences. Normalisation techniques are either model-based or dedicated to individual signals.

Probabilistic quotient normalisation, PQN. PQN was originally proposed to correct for the dilution of urine samples measured by NMR [39]. It assumes that the differences in the intensity of the majority of signal peaks result from the dilution of the samples rather than alterations of the single constituents concentrations. The normalisation factor for each signal, is then the most probable quotient of this signal and the reference, usually selected to be the median quotient as a robust summarising value. Median, mean signal or a *golden standard* is usually adopted as a reference. For normalisation each *i*-th signal intensity is divided by the defined quotient, *q*, as a normalisation factor, $s_{i,norm}=s_i/q$.

Vector normalisation. The normalisation factor is computed as a square root of the sum of squared signal intensities, $q=\sum_{i=1}^{J} s_i^2$. Then, each of *J* signal intensities is divided by *q*, $s_{i,norm}=s_i/q$.

Standard normal variate, SNV. Each signal intensity is reduced by mean signal intensity and then divided by its standard deviation, $s_{i,norm}=\frac{s_i-mean(s)}{sd(s)}$. It effectively eliminates the constant offset and multiplicative differences between spectra.

Multiplicative signal correction methods, MSC. The family of model-based MSC methods aims at getting the largest possible similarity between the spectrum and the reference by accounting for various physical and chemical sources of variation in vibrational spectra, using ordinary or weighted least squares procedure [[40], [41], [42]]. MSC methods serve as a perfect tool for normalisation of signals by correcting the additive, multiplicative, wavenumber-dependent variations between spectra and the reference as well as physical effects related to temperature, samples thickness, etc. [[42], [43], [44], [45]].

The concept of basic MSC is founded in the Lambert-Beer law and models the spectrum with respect to a reference (usually mean spectrum) according to the equation(15)$s_{\tilde{\nu}}=a+bm_{\tilde{\nu}}+E_{\tilde{\nu}}$,where *a* is the constant offset between the spectrum $s(\tilde{\nu})$ and reference $m(\tilde{\nu})$, *b* represents the multiplicative effect between $s(\tilde{\nu})$ and $m(\tilde{\nu})$, arising mostly from variations in laser intensity in Raman

spectroscopy, and $E(\tilde{\nu})$ are model residuals reflecting the unmodeled differences between spectra. After the model parameters are estimated in ordinary or weighted least squares procedure, the corrected spectrum is given as(16)$$corr\tilde{\nu}=s\tilde{\nu}-a/b.$$
As shown above, MSC only eliminates constant baseline and scaling effects between spectra. However, typically in Raman spectra the baseline effects cannot be portrayed with a straight line but are much more complex. Thus extended MSC (EMSC) is intended to include the wavenumber-dependent variations of fluctuating baseline using the polynomials with increasing degree [40,41]. EMSC approximates the spectrum as(17)$$s\tilde{\nu}=a+bm\tilde{\nu}+d1\tilde{\nu}+d2\tilde{\nu}2+\ldots+dn\tilde{\nu}n+E\tilde{\nu},$$where $d1\tilde{\nu}$, $d2\tilde{\nu}2$ and $dn\tilde{\nu}n$ are linear, quadratic and higher polynomial degree baseline effects. The corrected spectrum is then found from(18)$$corr\tilde{\nu}=s\tilde{\nu}-a-d1\tilde{\nu}-d2\tilde{\nu}2-\ldots-dn\tilde{\nu}n/b.$$
Basic version of EMSC applies only linear and quadratic terms.

EMSC may be further improved to account for the variations between replicate spectra of the same sample [41,42,45]. Inter-replicate variations are summarised using only a small number of PCA components and subsequently removed through incorporation of the orthogonal subspace model in EMSC model in the following procedure:

(1)
build an EMSC model for each set of replicate spectra, correct the replicate spectra with these local EMSC models and mean-center them within the replicate sets;

(2)
concatenate all replicate sets in one data matrix and summarise the between-replicate variance using a few orthogonal PCA components.

In EMSC with replicates correction each spectrum is represented as(19)$$s\tilde{\nu}=a+bm\tilde{\nu}+d1\tilde{\nu}+d2\tilde{\nu}2+\ldots+dn\tilde{\nu}n+\Sigma k=1Kgkpk\tilde{\nu}+E\tilde{\nu},$$where pk is the *k*-th from *K* most significant loading vectors and gk are the corresponding fitted parameters.

2.3. Genetic algorithm

Genetic algorithm (GA) [[5], [6], [7]] is embedded in the Darwin's evolution theory, where the nature determines the survivability of individuals based on their adaptation to life. In this sense it can be considered an optimisation process, in which the best solution is found, that in nature setting is an equivalent of an individual with best accommodation to living in a specified environment. Only a limited number of individuals with better fitness to the environment are more likely to survive and procreate to transmit their profitable genetic material to the next generations.
When moving the concept of the algorithm from nature settings to applications in the field of optimisation, the following relations hold:

- adaptation to the environment acts as a response function;
- genetic material that is responsible for good or bad fitness to the environment becomes a particular solution from a set of them under optimisation;
- genes building the chromosomes are the variables in each solution;
- nitrogen bases, as the basic element of the genetic material, are known as bits to encode the variable value.

GA is initialised with a formation of the original population by random selection of a specified number of individuals described by their genetic material (one of the solutions for optimisation). The individuals that are best fit to the environment mate and their genes are shuffled in the crossover process. In this way good genetic material is propagated, while the bad one disappears and the fitness is improving through the generations. In optimisation framework this means that the profitable solutions are selected based on the response function and their variables are mixed and spread to set up better solutions and eliminate the worst.

While reproduction leads just to a combination of the genetic material of the parents, mutations remarkably change the genetic material content by introducing minor

changes at the nitrogen bases level. This is equivalent to changing the variables values. The process of reproduction and mutation is repeated to create new generations that always have better average adaptation to the environment than their ancestors. This corresponds to mixing the variables and eventually changing their values slightly to receive better solutions than previously. To increase the GA effectiveness, a number of best individuals is kept and preserved to the next generation according to the elitism rule to prevent from losing their most profitable genetic material if they die. This immortality rule is a consequence of the fact that in some cases new best solutions are not necessarily better than the best in the previous set, even though the average response of the new set is improved. For this reason a specified number of best solutions from each set is kept and propagated to appear finally as the most optimal solutions that were ever found.

2.4. Regularised MANOVA

Regularised MANOVA [11] is a modification of classical MANOVA (an extension of ANOVA for multivariate sets). Similar to LDA, MANOVA works with the matrices of between-groups variability (B) and within-groups variability (W), so it accounts for the covariance structure of the data. However, both LDA and MANOVA fail for highly multidimensional data when the number of variables extends the number of samples. This is due to the inability to compute the inverse of the variance-covariance matrices that do not have full rank or their instability when the number of variables is comparable to the number of samples. Regularisation of the method, achieved by introducing suitable parameters, is the effective solution for handling singularity issues of variance-covariance matrices. Its objective is to find the eigenvectors of the matrix $((1-\delta)W+\delta T)-1B$. These are the directions along which the between-groups variance is the highest and the within-groups variance the lowest. T is the target matrix which is either $T=1ptr(W)$ when the variances of p variables for each group are equal or $T=diag(W)$ when the variances for each group are unique. δ is dependent on the chosen target and expresses the variance of the W matrix components according to the Ledoit-Wolf theorem [11].

2.5. Likelihood ratio

In the forensics verifying which of the two contrasting hypotheses stating that samples have the same source (also time-related source as in bloodstains age determination) ($H1$), or two different sources ($H2$), is more likely, is actually a discrimination task that apart from the similarity of the samples takes into account the uniqueness of their measured features with respect to other available sources. Thus establishing the potential (un)common samples origins is something more than only likeness of the data but should also include their frequency. When the data of both samples are typical they may match just by chance. The conclusions are therefore more persuasive if the similarity is observed between rare data than when it is detected between typical features. The risk of coincidental match between typical features escalates with increasing data frequency. Thus the evidential value of the match between samples increases with uniqueness of the features. Even though many chemometric tools designed for discrimination purposes attempt to expose the most unique features for each source in a few latent variables, they tend to ignore the features typicality when assigning samples membership. The likelihood ratio (LR) framework [[12], [13], [14]] quantifies the strength of samples similarity which escalates with their increasing typicality and indicates how strongly they are alike to establish whether the samples share common origins. Basically, the LR is computed as the probability of recording the physicochemical data for the samples (E), given the propositions ($H1$ and $H2$):$$(20) LR=Pr(E|H1)Pr(E|H2).$$

$H1$ is supported by the LR values larger than 1 and the support is strengthening with increasing LR. Conversely, the $H2$ is more likely when LR is below 1 and the support for this hypothesis reinforces with the LR values approaching 0. Both hypotheses are equally likely when $LR=1$.

Current solutions attempt to construct (train) LR models on databases with J variables for I measurements from M sources, each measured N times ($I=MN$) and use them to compare two samples, each described by a mean vector of J variables. When $I<J$, the LR models fail due to singularity of the variance-covariance matrices and adequate data dimensionality reduction is requisite. The obvious concept is to apply hybrid LR models [15,16,46] where conventional LR models are constructed for a limited number of latent variables derived from chemometric tools (e.g. rMANOVA) with least variability within each source and maximal variability between sources to enhance the LR models performance. In hybrid LR models the likeness of the samples

is studied by the LR framework for appropriately compressed data by chemometric tools that are believed to best describe the individual sources and preserve their most unique features.

According to Equation (20), the LR numerator evaluates the support towards the H1. It accounts for the similarity of the samples means, $\bar{y}1$ and $\bar{y}2$ with k1 and k2 replicate measurements, as well as the similarity of their weighted average, $\bar{y}*=k1\bar{y}1+k2\bar{y}2k1+k2$, to means of each of M sources $\bar{x}m$ of training data. The denominator of the LR formula corresponds with H2. Then both contributions from the samples $\bar{y}1$ and $\bar{y}2$ are assumed independent [[12], [13], [14]].

When the between-source distribution is assumed normal, then LR expression is given as in Ref. [[12], [13], [14]]. When the data cannot be assumed normally distributed, the kernel density estimation (KDE) procedure estimates the underlying distributions by averaging over all sources means instead of the general mean as adopted in Gaussian distribution. The smoothing parameter is set as $h=(4M(2p+1))1p+4$, where *p* is the number of considered variables. Then LR is given as a product of the following multivariate normal distributions (MVN) [[12], [13], [14]]:[21]$LR=MVN(\bar{y}1-\bar{y}2|0,Wk1+Wk2)\cdot 1M\sum m=1MMVN(\bar{y}*|\bar{x}m,Wk1+k2+h2B)1M\sum m=1MMVN(\bar{y}1|\bar{x}m,Wk1+h2B)\cdot 1M\sum m=1MMVN(\bar{y}2|\bar{x}m,Wk2+h2B)$

For univariate data matrices or vectors (e.g. W, $\bar{x}$) become scalars (w2, $\bar{x}$).

LR models quality diagnostics primarily include the levels of false positive (LR>1 when H2 is true) and false negative responses (LR<1 when H1 is true). Even though these rates only indicate which of the hypotheses is supported, but disregard the magnitude of this support, this paper is limited only to this form of reporting LR models performance.

3. Experimental

The original signals were subjected to preprocessing starting with denoising/smoothing, then baseline correction followed by normalisation. Denoised/smoothed signals, a, were additionally transformed with log-centered transform to compensate heteroscedastic noise [47] that grows with signal intensity:(22)$s=\log 10a-1/J\sum i=1J\log ai=\log 10a(J\prod i=1Jai)$.

This was the only reasonable sequence since many baseline correction methods are successful only for at least partially denoised/smoothed signals with homoscedastic noise and normalisation must be preceded by the removal of baseline. The MSC methods were an exception as they provide both baseline correction and normalisation if the mean centered signals are then subjected to statistical models. Therefore, these methods were the last link in some preprocessing strategies, preceded only by denoising/smoothing. The space of available parameters for each preprocessing method was limited in visual inspection by looking at the data after preprocessing and controlling if the unwanted artifacts were eliminated. The groups of parameters for which the graphical visualisation was pleasing were then selected for optimisation. They are listed in Table 1, Table 2, Table 3, which also provide useful details such as the R packages for implementing the methods and source literature positions introducing them. 16 denoising/smoothing strategies were tested based on discrete wavelet transform and Savitzky-Golay filter. 64 baseline correction strategies involved asymmetric penalised least squares (5 methods), robust baseline estimation, statistics-sensitive not-linear iterative peak-clipping, multiplicative signal correction (3 methods), polynomials (2 methods) and quantile regression (3 methods). Due to unsatisfying visual results, IModPoly was skipped for preprocessing bloodstains Raman spectra and SNIP was ignored for car paints Raman spectra. 16 normalisation strategies were based on standard normal variate, probabilistic quotient normalisation, vector normalisation and multiplicative signal correction (3 methods).

Table 1. Details of denoising and smoothing strategies. BS stands for the database of Raman spectra of bloodstains and CP for car paints.

group of methods	abbrev.	parameters	parameter values	R package	literature
Savitzky-Golay	SG	*p*-polynomial degree	p=3,4,5,6	*signal*	[20]
		w-window size	w=17 for BS and w=7 for CP		
discrete wavelet transform	DWT	*W*-wavelet type	*W* = Daubechies Least Asymmetric 4,8, Coiflets 1,5	*wavethresh*	[21], [22], [23], [24], [25]
		d-decomposition level for denoising	d=10		

group of methods	abbrev.	parameters	parameter values	R package	literature
		t-threshold estimation	*t* = universal, SURE		
		c-thresholding policy	*c* = hard, soft		
		sd-dispersion estimate	sd = MAD		

Table 2. Details of baseline correction strategies. BS stands for the database of Raman spectra of bloodstains and CP for car paints.

group of methods	abbrev.	parameters	parameter values	R package	literature
asymmetric penalised least squares					
	pAsWPLS	*m*-order of differences	m=2	–	
		λ-penalty	λ=6·105,8·105,106		
		w-weights	w=0.0005,0.005,0.001		
	CWTAsWPLS	*m*-order of differences	m=2	*baselineWavelet*	[32]
		λ-penalty	λ=7·107,8·107,9·107,108 for BS		
			λ=3·107,5·107,7·107,108 for CP		
	airPLS	*m*-order of differences	m=2	*airPLS*	[31]
		λ-penalty	λ=6·104,7·104,8·104,9·104		
	2WAsPLS	*m*-order of differences	m=2	–	[33]
		γmax-penalty	γmax=6·104,9·104		
		r=γmin/γmax– penalties ratio	r=0.7,0.9		
	multiWAsPLS	*m*-order of differences	m=2	–	[34]
		λ-penalty term	λ=10,100 for BS		
			λ=1000,10000 for CP		
		μ-penalty term	μ=107,108 for BS		
			μ=108 for CP		
robust baseline estimation	RBE	*b*-robustness parameter	b=2,2.5 for BS	*baseline*	[37]

group of methods	abbrev.	parameters	parameter values	R package	literature
			b=2.5,3 for CP		
		h-proportion of signal points	h=0.3,0.4		
		for local regression			
statistics-sensitive	SNIP	*w*-clipping window	w=25,30 only for BS	*MALDIquant*	[35]
non-linear iterative peak-clipping					
multiplicative signal correction methods					
-multiplicative signal correction	MSP	–	–	*pls*	[40]
-extended multiplicative signal correction	EMSC	*p*-polynomial degree	p=3,4,5,6	*EMSC*	[40]
-extended multiplicative signal correction	repEMSC	*p*-polynomial degree	p=3,4,5,6	*EMSC*	[40]
with replicates correction		pc-proportion of the explained	pc=0.9,0.95		
		replicates variance			
polynomial methods					
-modified polynomial	ModPoly	*p*-polynomial degree	p=3,4,5,6	*baseline*	[27]
-improved modified polynomial	IModPoly	*p*-polynomial degree	p=3,4,5,6 only for CP	–	[28]
quantile regression methods					
-polynomial quantile regression	polyQR	*p*-polynomial degree	p=5,6 for BS	*quantreg*	[36]
			p=6,7 for CP		
		q-quantile	q=0.05,0.01,0.001 for BS		
			q=0.05,0.01,0.1 for CP		
-spline quantile regression	splineQR	*q*-quantile	q=0.1,0.05,0.01,0.001 for BS	*cobs*	[36]

group of methods	abbrev.	parameters	parameter values	R package	literature
			q=0.1,0.01 for CP		
		λ-penalty	λ=0 for BS		
			λ=1,−1 for CP		
-reweighted quantile regression	reweightedQR	*p*-polynomial degree	p=5,6 for BS	*quantreg*	[36]
			p=6,7 for CP		
		q-quantile	q=0.05,0.01,0.1		

Table 3. Details of normalisation strategies.

group of methods	abbrev.	parameters	parameter values	R package	literature
standard normal variate	SNV	–	–	–	
probabilistic quotient normalisation	PQN	–	–	–	[39]
vector normalisation	VN	–	–	–	
multiplicative signal correction methods					
-multiplicative signal correction	MSC	–	–	*pls*	[40]
-extended multiplicative signal correction	EMSC	*p*-polynomial degree	p=3,4,5,6	*EMSC*	[40]
-extended multiplicative signal correction	repEMSC	*p*-polynomial degree	p=3,4,5,6	*EMSC*	[40]
with replicates correction		pc-proportion of the explained			
		replicates variance	pc=0.9,0.95		

Within each of the preprocessing methods all parameter values combinations listed in Table 1, Table 2, Table 3 were tested giving in total 13264 possible preprocessing strategies. DWT was the only exception as SURE thresholding may be applied in R only with soft policy. All 13264 preprocessing strategies were subjected to optimisation in the grid search process and using the genetic algorithm. It should be emphasised that the entire preprocessing strategies consisting of denoising/smoothing, baseline correction and normalisation were the subject of optimisation, rather than individual preprocessing steps. This is a consequence of the fact that the suitability of the preprocessing steps strongly depends of their coupling and the effect is not a simple resultant sum of the contributing components. The quality parameter in the

grid search and response function in genetic algorithm was the ratio of the between-source to within-source variance ($b2/w2$) on the first rMANOVA latent variable, LV1. The chromosome in the genetic algorithm consisted of three genes corresponding with denoising/smoothing, baseline correction and normalisation methods. The initial generation consisted of 50 randomly selected preprocessing strategies, the chance of mutations was 0.1, elitism level was set at 5% and the algorithm converged if 5 subsequent solutions were identical. The target matrix in rMANOVA expressed equal variances for each source to remain in line with the statistical LR models assumption. The relevance of the proposed methodology was verified through the development of LR models (in order to meet forensic interpretation requirements) for concluding if the samples may share common origins (as in car paints example) or have the same age (as in bloodstains example). LR models were trained and tested according to Equation 21 for a single variable being the first latent variable from rMANOVA, LV1. Their performance was reported with the false positive and false negative rates (section 2.5). The LR values for estimating the false positive rates were computed for test samples from two different sources (car paints) or of different age (bloodstains). Any value above 1 was a false positive indication. The LR values for computing false negative rates were yielded for test samples from the same source (car paints) or with the same age (bloodstains). Any value below 1 was a false negative indication.
The calculations were carried out in R software [48] using home-written scripts and available R packages listed in Table 1, Table 2, Table 3.

4. Results and discussion

All 13264 preprocessing strategies are summarised and ordered by increasing $b2/w2$ on the rMANOVA first latent variable LV1 as demonstrated in Fig. 2. The colours in the plots correspond to results observed when various preprocessing methods within denoising, baseline correction and normalisation steps were applied, regardless of their parameters. For instance in Fig. 2b black points show $b2/w2$ for all the preprocessing strategies including pAsWPLS method. From the graphs we can easily note that the range of $b2/w2$ obtained for different preprocessing strategies reaches two or three orders of magnitude for the databases of Raman spectra for bloodstains (BS) and car paints (CP) respectively. Moreover, for the BS ca. 13%

of the preprocessing strategies result in lower variance between sources than within them, which is completely useless for developing well performing discrimination models. These findings emphasise the fact that preprocessing has an influential effect on variance components and considerable insight into this area is essential and may become a noteworthy clue in improving discrimination models. The diagrams referring to denoising and baseline correction methods (Fig. 2a and b) practically do not present any trend which may point out that any of the applied methods is clearly better than the others. Due to the poverty of their informativeness, they are presented only for Raman spectra of bloodstains. Some tendency is observable only for normalisation methods (Fig. 2c and d), from which EMSC with replicates correction (repEMSC) appears to be indisputably the worst. When using other methods $b2/w2$ rises drastically, which is visible as a steep slope starting in the middle of Fig. 2c and d. These findings should not come as a surprise as only normalised signals are fully able to reveal the proper within- and between-groups variance structure. Poor performance of repEMSC method may, however, seem surprising at first glance. The method is known to be successful in increasing $b2/w2$ thanks to reduction of the variations between replicate signals (i.e. marked as belonging to particular groups we try to discriminate the samples between) after the correction by modeling and removal of the differences between them. The reason for lower $b2/w2$ observed in the studies should be seen, however, as a consequence of applying proper validation schemes for forensic investigations that force to treat any two samples as two different sources *a priori* to follow the principle of the presumption of innocence. According to this validation scheme, each source is always composed of two smaller sets that are individually preprocessed. If the preprocessing strategies are applied individually for each signal, this division has no meaning. It matters, however, for supervised preprocessing strategies that use the information about all signals in a group to correct the baseline or normalise them. repEMSC may serve as an example. If we use repEMSC for each set separately, the replicates are made maximally close within each set and naturally more diversed between sets. Then the variation within sources (each composed of two sets) rises, making $b2/w2$ automatically lower in regard to other methods that do not intend to reduce $w2$ unduly. Nevertheless, for BS there are a few preprocessing strategies involving repEMSC that yield very high $b2/w2$. This in turn is the result of a random

selection of the signals for the validation sets that is beneficial for achieving high b2/w2. By coincidence, the preprocessing strategies involving repEMSC may make the LR models overfitted, with poor performance (high false positive and false negative rates), as will be shown later.

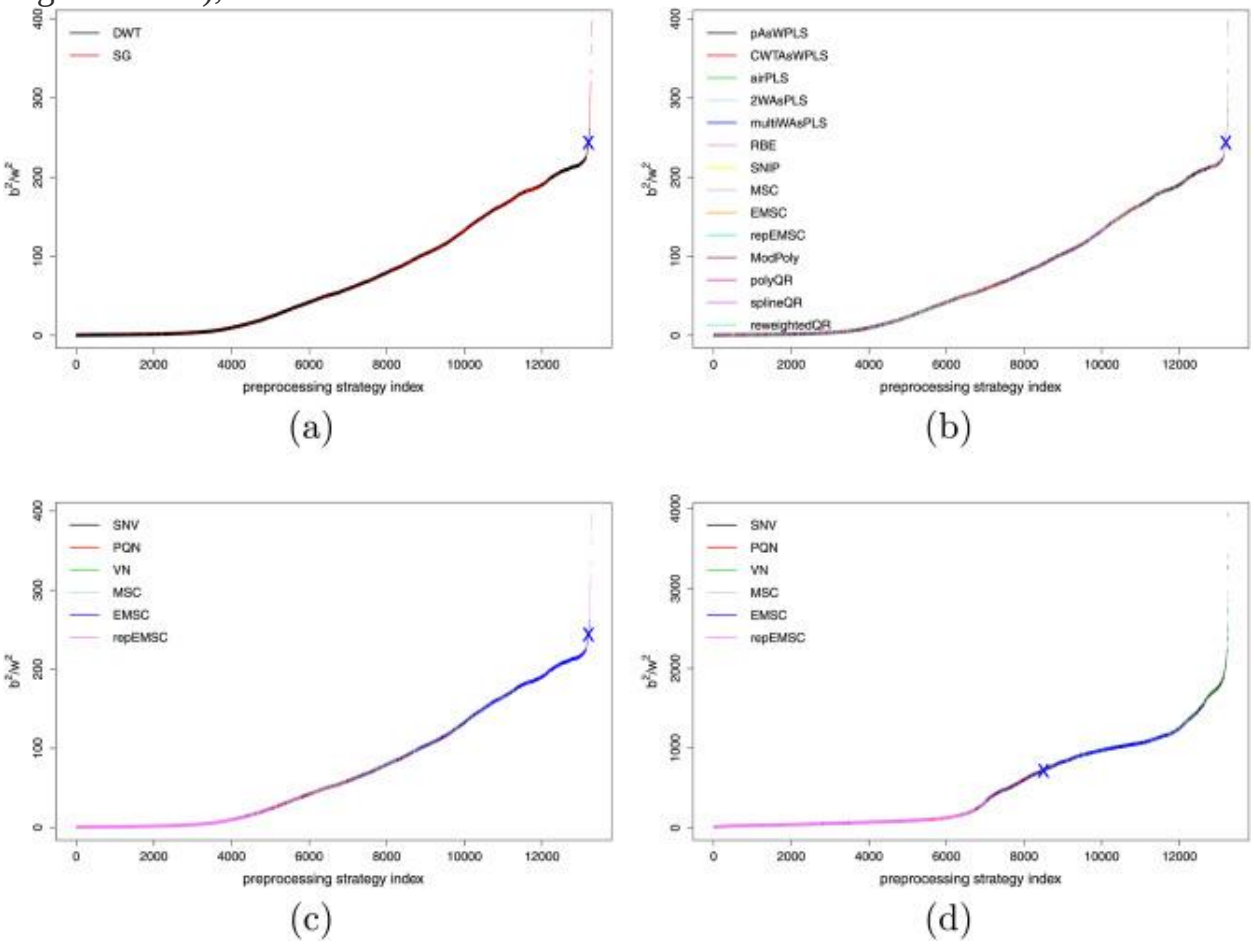


Fig. 2. b2/w2 values computed for all 13264 preprocessing strategies. Colours refer to strategies using various (a) denoising techniques for Raman spectra of bloodstains, (b) baseline correction techniques for Raman spectra of bloodstains, normalisation techniques for (c) Raman spectra of bloodstains, (d) Raman spectra of car paints.

The blue X signs in the pictures in Fig. 2 show the preprocessing strategies found best using the genetic algorithm. The solution found using the GA is the 68th solution in descending order per 13264 in total for Raman spectra of blood traces and 4766/13264 for Raman spectra of car paints. The optimal solutions found in genetic algorithm were obtained in several dozen times shorter time than using the grid search. The algorithm converged in 6th and 14th generation for both databases respectively, after having found the same optimal solution in five subsequent

generations. Table 4 records the best, the worst preprocessing strategies observed in the grid search process as well as the winning solutions found using the genetic algorithm.

Table 4. The best, the worst preprocessing strategies observed in the grid search process as well as the best solutions found using genetic algorithm (GA).

	best	worst	GA
Raman spectra of bloodstains			
b2/w2	395	nearly 0	244
denoising	SG polynomial degree p=6	SG polynomial degree p=4	DWT Coiflets 1 decomposition level for denoising d=10 threshold estimation *t* = universal thresholding policy *c* = hard dispersion estimate sd = MAD
baseline correction	ModPoly polynomial degree p=4	pAsWPLS order of differences m=2 penalty term λ=106 weights w=0.005	reweightedQR polynomial degree p=6 quantile q=0.05
normalisation	repEMSC polynomial degree p=3 prop. of the explained replicates variance pc=0.9	repEMSC polynomial degree p=3 prop. of the explained replicates variance pc=0.95	EMSC polynomial degree p=5
Raman spectra of car paints			
b2/w2	3922	4	2633
denoising	SG polynomial degree p=6	SG polynomial degree p=5	DWT Daubechies Least Asymmetric 8 decomposition level for denoising d=10 threshold estimation *t* = SURE thresholding policy *c* = soft dispersion estimate sd = MAD
baseline correction	multiWAsPLS	repEMSC	pAsWPLS
	order of differences m=2	polynomial degree p=3	order of differences m=2
	penalty term λ=104	prop. of the explained replicates variance	penalty λ=6·105
	penalty term μ=108	pc=0.9	weights p=0.005
Normalisation	SNV	–	EMSC
			polynomial degree p=4

Table 4 clearly shows that Savitzky-Golay filter with the polynomial of 6th degree delivers the most satisfying b2/w2 for both databases. SG filters with lower polynomial degrees were found as the least preferable. According to Fig. 2a the usefulness of SG or DWT is not that clear and must always be judged in view of the baseline correction and normalisation methods applied afterwards.
For Raman spectra of car paints asymmetric penalised least squares methods that introduce an additional penalty to penalise remarkable differences between the corrected replicate signals, which should be obviously as similar as possible after correction of the baseline (multiWAsPLS), deliver the most promising results. This is not surprising on the one hand, as the method helps in reducing the variations between replicate signals after the baseline correction and thus, it reduces diversity of the samples within the groups, making b2/w2 automatically higher. But on the other hand, the multiWAsPLS method is similar to repEMSC in that it also takes care of removing the differences between the replicates. As explained above, the method should rather produce overfitted LR models, but it does not. Thus we suspect that it presumably is not as successful as repEMSC in reducing b2/w2 and acts more like a method applied to single signals than to a group of them. However, this method does not guarantee the best results for Raman spectra of bloodstains, for which modified polynomial method (ModPoly) scores the highest. repEMSC is for both databases producing the worst results, as presumed. Surprisingly, it is also a normalisation method of the best preprocessing strategy for BS. This is rather a coincidence producing overfitted LR models wrongly stating that samples of the same age pose different age in even 60% of cases. The solutions found using genetic algorithm include EMSC method for both databases as a normalisation strategy.
Fig. 3, Fig. 4 portray the capability of the preprocessing strategies in exposing the differences between groups and hiding the diversity within the groups of spectra. It is clear that the worst preprocessing strategies fail to correct baseline properly by cutting off some important parts as evidently visible in Fig. 3c. The picture definitely improves when preprocessing strategies selected using the genetic algorithm were applied (Fig. 3, Fig. 4d,e). Despite less efficient denoising strategy and thus lower legibility of the images in Fig. 3, Fig. 4e, using the strategy from GA instead of the best preprocessing strategy translates in a much shorter period of time into a well

preprocessed spectra where group differences are only referring to Raman bands and do not arise from baseline artifacts.

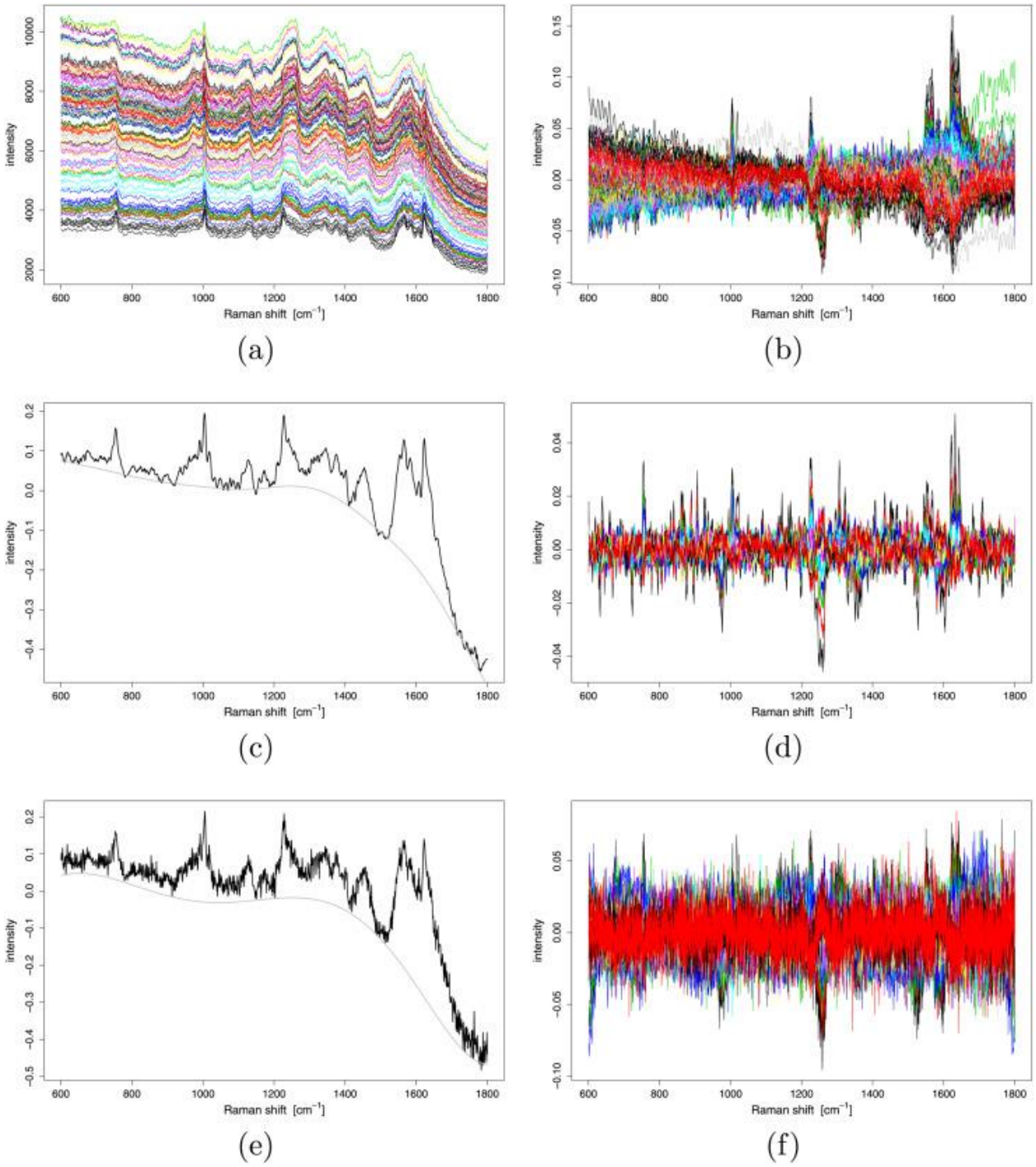


Fig. 3. (a) Denoised (using SG from the best preprocessing strategy), (b) log transformed and mean centered, (c) baseline corrected (example, with pAsWPLS as in the worst preprocessing strategy), (d) normalised with repEMSC (worst preprocessing strategy) and mean centered, (e) baseline corrected (example, with reweightedQR as in a preprocessing

strategy found using genetic algorithm), (f) normalised with EMSC (preprocessing strategy found using genetic algorithm) and mean centered Raman spectra of bloodstains.

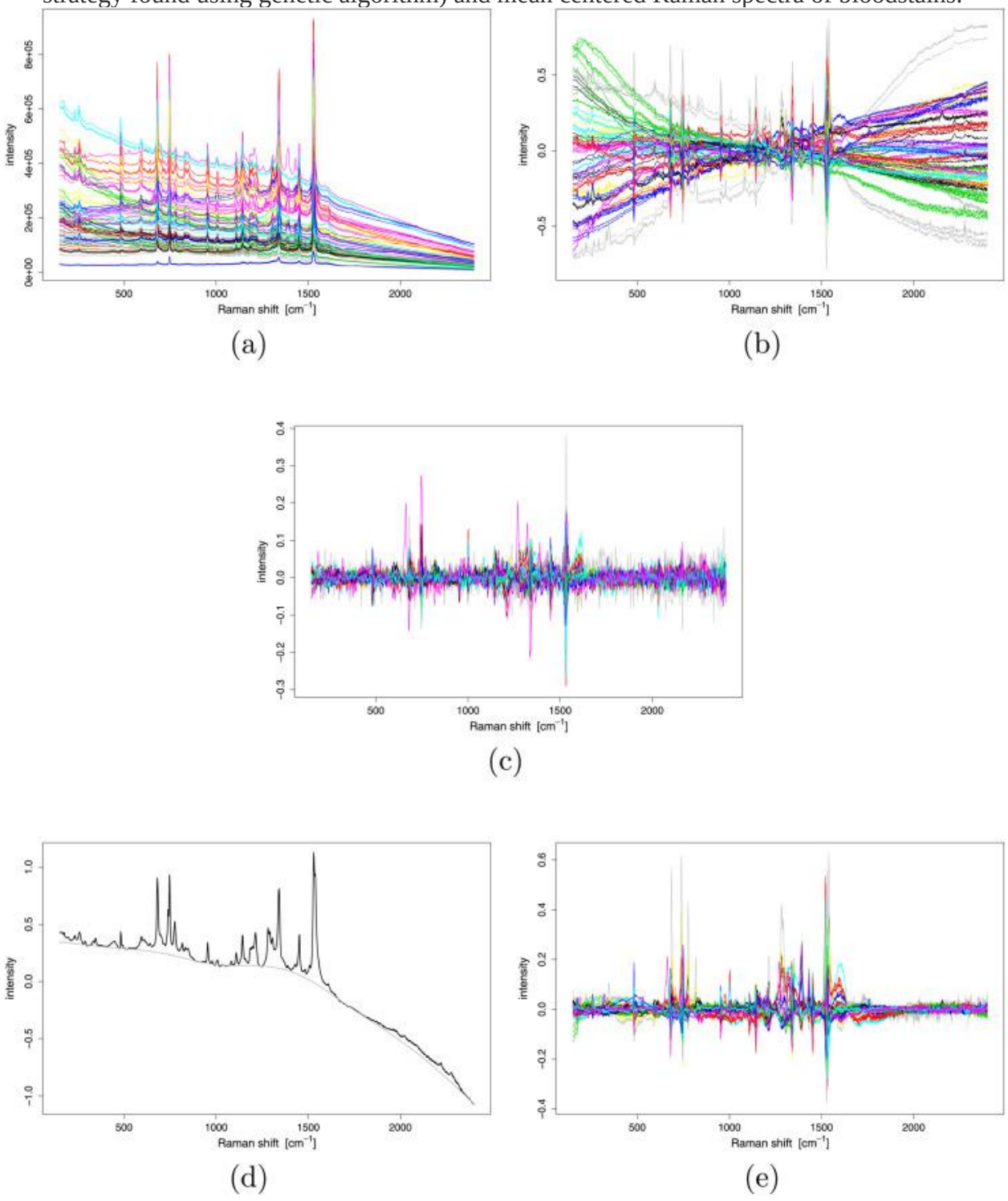


Fig. 4. (a) Denoised (using SG from the best preprocessing strategy), (b) log transformed and mean centered, (c) normalised with repMSC (worst preprocessing strategy) and mean centered, (d) baseline corrected (example, with pAsWPLS as in a preprocessing strategy

found using genetic algorithm) and (e) normalised with EMSC (preprocessing strategy found using genetic algorithm) and mean centered Raman spectra of car paints.

Fig. 5, Fig. 6 illustrate the capability of rMANOVA to maximise b2/w2. The loadings of the first latent variable (LV1; Fig. 5, Fig. 6a) follow the shape of the original Raman spectrum in the sense that the extreme loadings correspond with most crucial Raman peaks. This proves that rMANOVA successfully describes the differences between samples arising from the changes in their chemical structure. The effect is less pronounced for the next latent variable for Raman spectra of bloodstains as shown in Fig. 5b since subsequent latent variables take care a lot less about b2/w2 (note the differences in the scale). Diagrams in Fig. 5c and d are the confirmation of these observations as the mean centered spectra reconstructed using only LV1 much better illustrate the differences between groups of spectra in the Raman peaks position than for subsequent latent variables. LV2 is, however, quite significant and explains much of b2/w2 for Raman spectra of car paints as Fig. 6b portrays. However, as Fig. 5d displays rather chaotic reconstruction of the signals using LV2, it was decided to use only LV1 in both databases as the variables for LR models. Finally, Fig. 5, Fig. 6e,f plainly show that the abilities of rMANOVA to maximise b2/w2 are strongly dependent on the preprocessing strategy that prepares the data before rMANOVA is applied. The projections of single spectra within each of the groups in the LV1-LV2 space that were prepared using preprocessing strategies chosen in the genetic algorithm are very close and form separate groups (indicated by the same colours and shapes of the points), while these prepared using the worst preprocessing strategies demonstrate much greater variability.

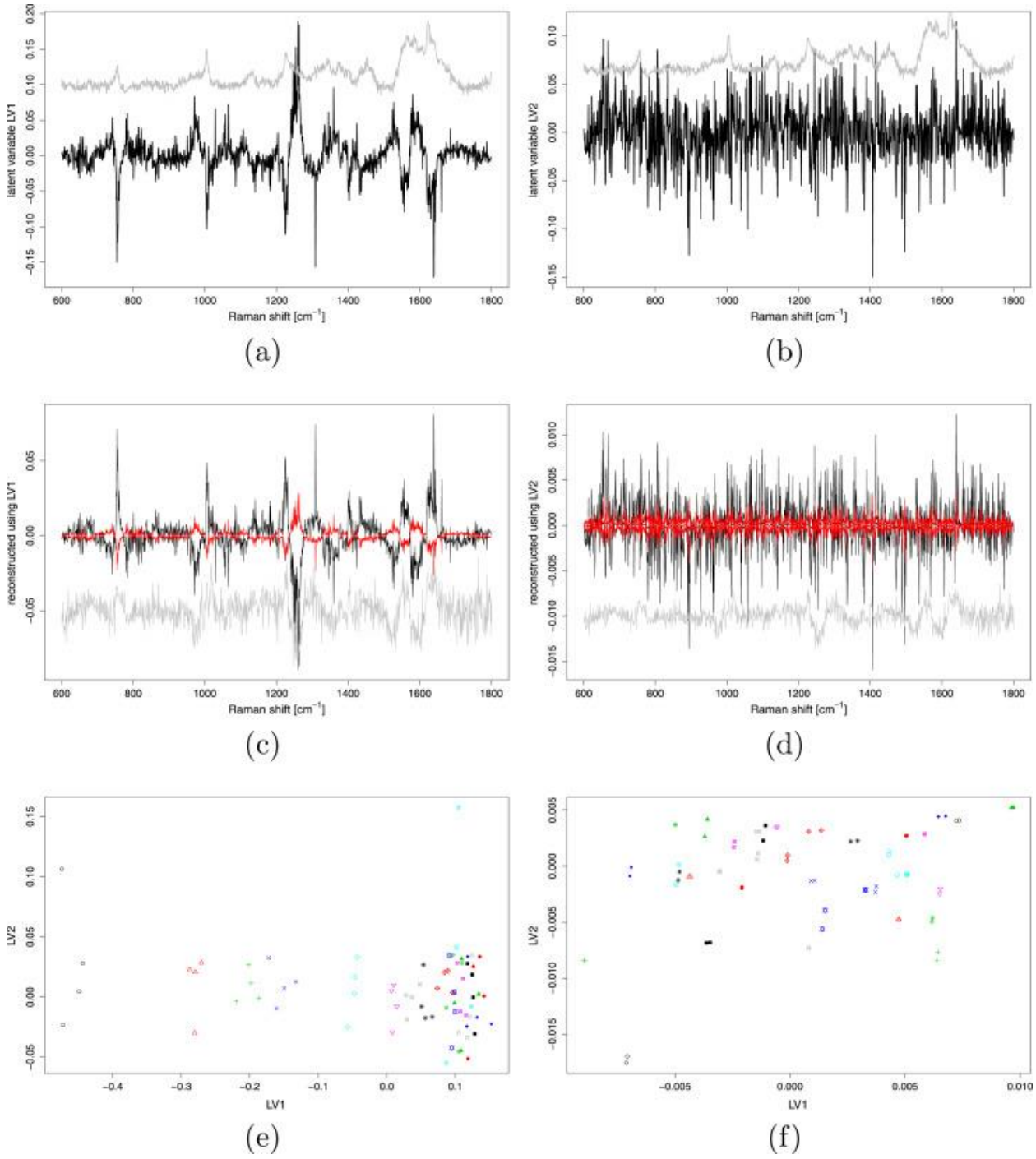


Fig. 5. rMANOVA loadings of the (a) first latent variable (LV1), (b) second latent variable (LV2), mean centered 2 groups of Raman spectra of bloodstains prepared using the preprocessing strategy found in genetic algorithm and reconstructed using (c) LV1, (d) LV2, (e) their projections in the LV1-LV2 space and (f) projections in the LV1-LV2 space of the spectra prepared using the worst preprocessing strategy (data for each time-related source are indicated by the same colours or signs). An example of the original spectrum (mean centered in (c) and (d)) is plotted in gray. (For interpretation of the

references to colour in this figure legend, the reader is referred to the Web version of this article.)

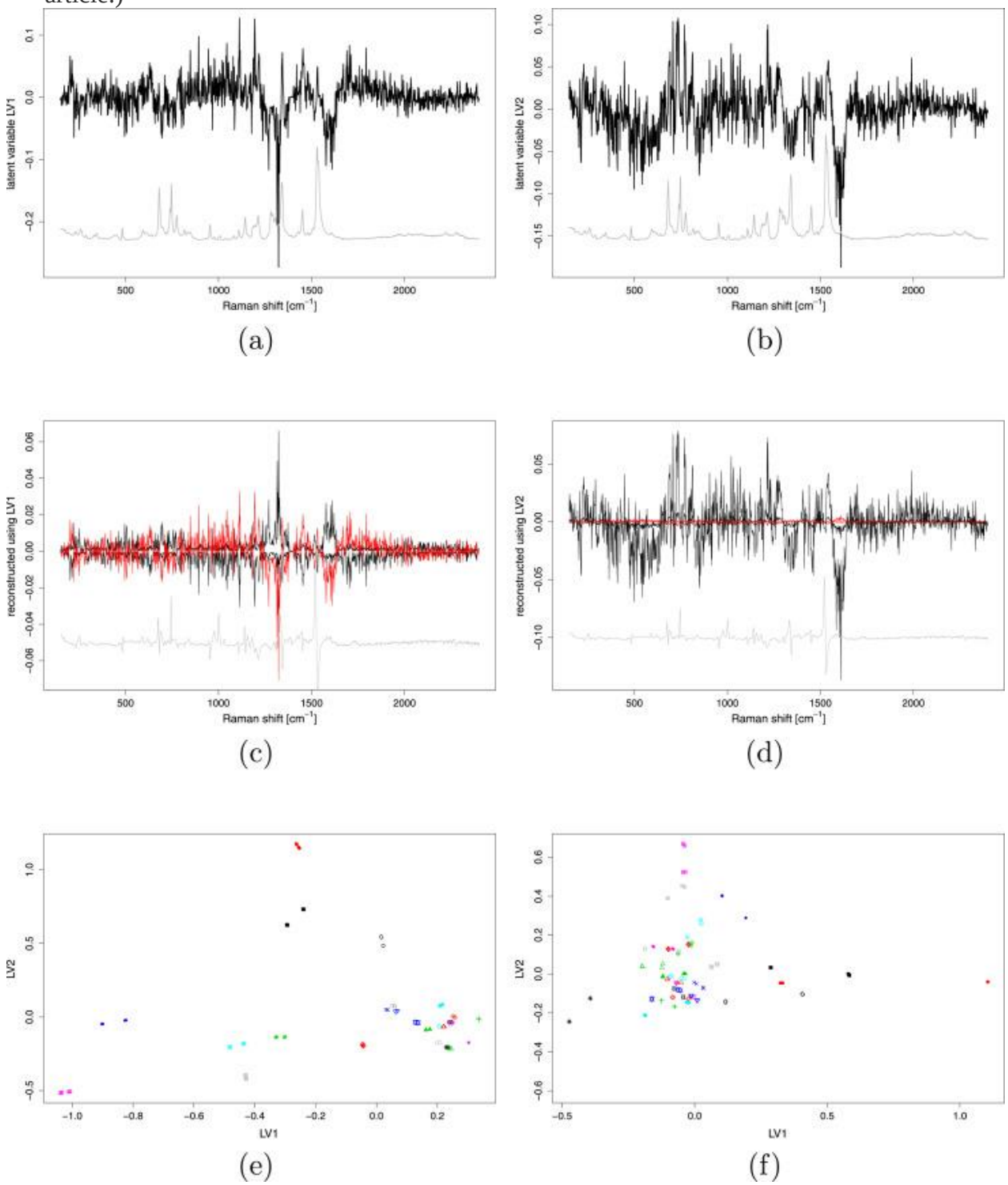


Fig. 6. rMANOVA loadings of the (a) first latent variable (LV1), (b) second latent variable (LV2), mean centered 2 groups of Raman spectra of car paints prepared using the preprocessing strategy found in genetic algorithm and reconstructed using (c) LV1,

(d) LV2, (e) their projections in the LV1-LV2 space and (f) projections in the LV1-LV2 space of the spectra prepared using the worst preprocessing strategy (data for each source are indicated by the same colours or signs). An example of the original spectrum (mean centered in (c) and (d)) is plotted in gray. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

The suitability of the proposed methodology for discrimination tasks was tested using the LR approach in three cases, i.e. when the data were prepared using the best preprocessing strategy (denoted as LR_{best}), the worst one (LR_{worst}) and the one selected using the genetic algorithm (LR_{GA}). The levels of false positive and false negative responses (Fig. 7) were the highest for the LR_{worst}, as expected. The lowest false rates were observed for LR_{GA} models. False positive answers oscillated around 24% and false negative answers for Raman spectra of bloodstains were 3%. 13% of false positive and no false negative answers for Raman spectra of car paints were observed. The results for the LR_{GA} models were thus not inferior to the best ones, especially that LR_{best} for Raman spectra of bloodstains were overfitted due to preprocessing with repEMSC method.

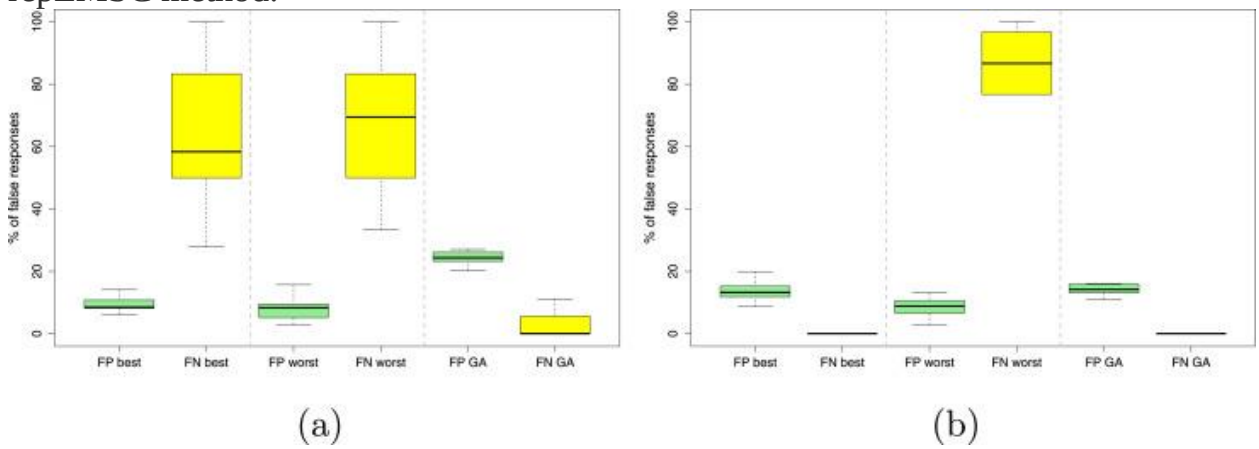


Fig. 7. The levels of false positive (FP) and false negative (FN) responses of LR models constructed for (a) Raman spectra of bloodstains and (b) Raman spectra of car paints prepared using the best, the worst preprocessing strategies and the one selected in genetic algorithm (GA).

5. Conclusions

In this paper we have outlined a novel concept that remarkably supports the discrimination analysis of the Raman signals owing to appropriately conducted preprocessing steps. The idea is based on using the genetic algorithm to find the

optimal preprocessing strategy yielding the highest ratio of the between-source and within-source variation ($b2/w2$) for the first latent variable computed from rMANOVA, as a quality parameter. Assessing the preprocessing strategy with this quality parameter computed on the rMANOVA first latent variable, as the most discriminating variable, ensures that the selected preprocessing strategy exposes best the differences between sources (i.e. groups) and minimises the casual variations within sources. Thus this research investigates the applicability of the rMANOVA as a mean for development of the criterium for fast and automatic selection of the most appropriate signal preprocessing tool when the discrimination of the highly multidimensional data is the problem at hand. Using the GA instead of the grid search substantially saves the time without prejudice to the final statistical models performance compared to the results produced for the data prepared using the best preprocessing strategies found in the grid search process.
Our findings emphasise the fact that preprocessing has an influential effect on variance components and considerable insight into this area is essential and may become a noteworthy clue in improving discrimination models. The preprocessing strategies best suited for our forensic applications should definitely skip the methods that overfit the statistical models, such as repEMSC. We have succeeded in showing that EMSC models deliver most pleasing results, however, they should work as a normalisation technique rather than both baseline correction and normalisation tool. They seem to be more successful when preceded by appropriate baseline correction methods. The selection of optimal preprocessing strategy is thus a matter of establishing the sequence of the methods for denosing/smoothing, baseline correction and normalisation and fixing of their most appropriate parameters.

Finally, it is also worth noting that the presented framework may be found useful not only in the forensics but also medical, environmental and food analysis applications, where the grouping of samples is the issue at hand. And even though our findings may not always be transferable to any datasets, we have developed a framework for enhancing the discrimination models performance for signals affected by fluorescence or any other distortions (such as for instance Mie scattering).

CRediT authorship contribution statement

Agnieszka Martyna: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Visualization. **Alicja Menżyk:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - review & editing. **Alessandro Damin:** Methodology, Validation, Investigation, Resources, Writing - review & editing. **Aleksandra Michalska:** Validation, Investigation, Resources, Writing - review & editing. **Gianmario Martra:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision. **Eugenio Alladio:** Writing - review & editing. **Grzegorz Zadora:** Conceptualization, Formal analysis, Writing - review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Authors wish to thank Prof. Marco Vincenti (Dipartimento di Chimica, Universita degli Studi di Torino; Centro Regionale Antidoping e di Tossicologia “A. Bertinaria”) for his professional guidance, valuable support and constructive recommendations on this project. Special thanks should be given to Universita degli Studi di Torino, Centre for Nanostructured and Interfaces and Surfaces for offering the resources in running the research.

References

[1]
R. Tauler, B. Walczak, S. Brown**Comprehensive Chemometrics**
Elsevier (2009)
[2]
J. Engel, J. Gerretzen, E. Szymanska, J. Jansen, G. Downey, L. Blanchet, L. Buydens**Breaking with trends in pre-processing**
Trends Anal. Chem., 50 (2013), pp. 96-106
[3]
P. Lasch**Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging**
Chemometr. Intell. Lab. Syst., 117 (2012), pp. 100-114
[4]

R. Gautam, S. Vanga, F. Ariese, S. Umapathy**Review of multidimensional data processing approaches for Raman and infrared spectroscopy**
EPJ Techniques and Instrumentation, 2 (2015), pp. 8-45

[5]

R. Leardi**Genetic algorithms in chemistry**
J. Chromatogr. A, 1158 (2007), pp. 226-233

[6]

R. Wehrens, L. Buydens**Evolutionary optimisation: a tutorial**
Trends Anal. Chem., 17 (1998), pp. 193-203

[7]

D. Hibbert**Genetic algorithms in chemistry**
Chemometr. Intell. Lab. Syst., 19 (1993), pp. 277-293

[8]

T. Bocklitz, A. Walter, K. Hartmann, P. Rösch, J. Popp**How to pre-process Raman spectra for reliable and stable models?**
Anal. Chim. Acta, 704 (2011), pp. 47-56

[9]

N. Afseth, V. Segtnan, J. Wold**Raman spectra of biological samples: a study of preprocessing methods**
Appl. Spectrosc., 60 (2006), pp. 1358-1367

[10]

G. Shuxia, T. Bocklitz, J. Popp**Optimization of Raman-spectrum baseline correction in biological application**
Analyst, 141 (2016), pp. 2396-2404

[11]

J. Engel, L. Blanchet, B. Bloemen, L. Heuvel, U. Engelke, R. Wevers, L. Buydens**Regularized MANOVA (rMANOVA) in untargeted metabolomics**
Anal. Chim. Acta, 899 (2015), pp. 1-12

[12]

C. Aitken, F. Taroni**Statistics and the Evaluation of Evidence for Forensic Scientists**
Wiley (2004)

[13]

G. Zadora, A. Martyna, D. Ramos, C. Aitken**Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data**
Wiley, Chichester, UK (2014)

[14]

C. Aitken, D. Lucy**Evaluation of trace evidence in the form of multivariate data**
Applied Statistics, 53 (2004), pp. 109-122

[15]

A. Martyna, G. Zadora, T. Neocleous, A. Michalska, N. Dean**Hybrid approach combining chemometrics and likelihood ratio framework for reporting the evidential value of spectra**
Anal. Chim. Acta, 931 (2016), pp. 34-46

A. Martyna, A. Michalska, G. Zadora**Interpretation of FTIR spectra of polymers and Raman spectra of car paints by means of likelihood ratio approach supported by wavelet transform for reducing data dimensionality**
Anal. Bioanal. Chem., 407 (2015), pp. 3357-3376

R. Bremmer, K. de Bruin, M. van Gemert, T. van Leeuwen, M. Aalders**Forensic quest for age determination of bloodstains**
Forensic Sci. Int., 216 (2012), pp. 1-11

G. Zadora, A. Menżyk**In the pursuit of the holy grail of forensic science – spectroscopic studies on the estimation of time since deposition of bloodstains**
Trends Anal. Chem., 105 (2018), pp. 137-165

Menżyk A., Damin A., Martra G., Martyna A., Alladio E., Vincenti M., Zadora G., Toward a novel framework for bloodstains dating by Raman spectroscopy: how to avoid sample photodamage and subsampling errors, Talanta 209 (n.d.).

A. Savitzky, M. Golay**Smoothing and differentiation of data by simplified least squares procedures**
Anal. Chem., 36 (1964), pp. 1627-1639

I. Daubechies**Ten Lectures on Wavelets**

CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia, USA (1992)

S. Mallat**A theory for multiresolution signal decomposition: the wavelet representation**
IEEE Trans. Pattern Anal. Mach. Intell., 7 (1989), pp. 674-693

B. Walczak, D. Massart**Wavelets-something for analytical chemistry?**
Trends Anal. Chem., 16 (1997), pp. 451-463

B. Alsberg, A. Woodward, M. Winson, J. Rowland, D. Kell**Wavelet denosing of infrared spectra**
Analyst, 122 (1997), pp. 645-652
View Record in ScopusGoogle Scholar

V. Barclay, R. Bonner**Application of wavelet transforms to experimental spectra: smoothing, denoising, and data set compression**
Anal. Chem., 69 (1997), pp. 78-90

D. Wei, S. Chen, Q. Liu**Review of fluorescence suppression techniques in Raman spectroscopy**
Appl. Spectrosc. Rev., 20 (2015), pp. 387-406

C. Lieber, A. Mahadevan-Jansen**Automated method for subtraction of fluorescence from biological Raman spectra**
Appl. Spectrosc., 57 (2003), pp. 1363-1367

J. Zhao, H. Lui, D. McLean, H. Zeng**Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy**
Appl. Spectrosc., 61 (2007), pp. 1225-1232

E. Whittaker**On a new method of graduation**
Proc. Edinb. Math. Soc., 41 (1922), pp. 63-75

P. Eilers**A perfect smoother**
Anal. Chem., 75 (2003), pp. 3631-3636

Z. Zhang, S. Chen, Y. Liang**Baseline correction using adaptive iteratively reweighted penalised least squares**
Analyst, 135 (2010), pp. 1138-1146

Z. Zhang, S. Chen, Y. Liang, Z.-X. Liu, Q.-M. Zhang, L.-X. Ding, F. Ye, H. Zhou**An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy**
J. Raman Spectrosc., 41 (2010), pp. 659-669

P. Cadusch, M. Hlaing, S. Wade, S. McArthur, P. Stoddart**Improved methods for fluorescence background subtraction from Raman spectra**
J. Raman Spectrosc., 44 (2013), pp. 1587-1595

J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, J. Tan**Asymmetric least squares for multiple spectra baseline correction**
Anal. Chim. Acta, 683 (2010), pp. 63-68

C. Ryan, E. Clayton, W. Griffin, S. Sie, D. Cousens**SNIP, a statistic-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications**
Nucl. Instrum. Methods Phys. Res., B34 (1988), pp. 396-402

L. Komsta**Comparison of several methods of chromatographic baseline removal with a new approach based on quantile regression**
Anal. Bioanal. Chem., 406 (2014), pp. 1985-1998

A. Ruckstuhl, M. Jacobson, R. Field, J. Dodd**Baseline subtraction using robust local regression estimation**
J. Quant. Spectrosc. Radiat. Transf., 68 (2001), pp. 179-193

W. Cleveland**Robust locally weighted regression and smoothing scatterplots**
J. Am. Stat. Assoc., 74 (1979), pp. 829-836

F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn**Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics**
Anal. Chem., 78 (2006), pp. 4281-4290

H. Martens, E. Stark**Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy**
J. Pharmaceut. Biomed. Anal., 9 (1991), pp. 625-635

N. Afseth, A. Kohler**Extended multiplicative signal correction in vibrational spectroscopy, a tutorial**
Chemometr. Intell. Lab. Syst., 117 (2012), pp. 92-99

A. Kohler, N. Afseth, H. Martens

J. Chalmers, E. Li Chan, P.R. Griffiths (Eds.), Chemometrics in Biospectroscopy, Wiley (2010), pp. 89-106

A. Kohler, C. Kirschner, A. Oust, H. Martens**Extended multiplicative signal correction as a tool for separation and characterisation of physical and chemical infromation in Fourier Transform Infrared Microscopy images in cryo-sections of beef loin**
Appl. Spectrosc., 59 (2005), pp. 707-716

H. Martens, S. Bruun, I. Adt, G. Sockalingum, A. Kohler**Pre-processing in biochemometrics: correction for path-length and temperature effects of water in FTIR bio-spectroscopy by EMSC**
J. Chemometr., 20 (2006), pp. 401-417

K. Liland, A. Kohler, N. Afseth**Model-based pre-processing in Raman spectroscopy of biological samples**
J. Raman Spectrosc., 47 (2016), pp. 643-650

A. Bolck, H. Ni, M. Lopatka**Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison**
Law Probab. Risk, 14 (2015), pp. 243-266

O. Kvalheim, F. Brakstad, Y.-Z. Liang**Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise**
Anal. Chem., 66 (1994), pp. 43-51

R Core Team**R: A Language and Environment for Statistical Computing**
R Foundation for Statistical Computing, Vienna, Austria (2018)
URL