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

Agnieszka Martyna a,*, Alicja Menzyk a, Alessandro Damin b, Aleksandra Michalska c, Gianmario Martra b, Eugenio Alladio d,e, Grzegorz Zadora a,c

a University of Silesia in Katowice, Faculty of Science and Technology, Institute of Chemistry, Forensic Chemistry Unit, 9 Szkolna, Katowice, 40-006, Poland
b Dipartimento di Chimica, Universita degli Studi di Torino, 7 Via Pietro Giuria, Torino, 10125, Italy
c Institute of Forensic Research in Krakow, 9 Westerplatte, Krakow, 31-033, Poland
d Centro Regionale Antidoping e di Toxicologia "A. Bertinari", 10/1 Regione Gonzole, Orbassano, 10043, Torino, Italy
e Reparto CC Investigazioni Scientifiche di Roma, Sezione di Biologia, 119 Viale Tor di Quinto, Rome, 00191, Italy

ARTICLE INFO

Keywords:
Signals preprocessing
Regularised MANOVA
Discrimination
Raman spectra
Likelihood ratio

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.

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 group separations, 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. 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 [2–4], 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 criterion 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 [5–7] (section 2.3), which do not try out every strategy to find the most promising strategy for the purpose of their analyses [8–10]. 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 (\( b^2/w^2 \)) 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. \( b^2/w^2 \) 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) [12–14] for concluding if the samples may share common origins. Fig. 1 briefly summarises this concept.

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 \( b^2 \) and minimise \( w^2 \). 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 then 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 then 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 then 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.
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 ev-

dence 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 50x N Plan 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 \(\mu\)m). 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 be-
tween 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 sub-
ject to Raman analysis. Each sample was measured in situ three times in
different locations [16]. Raman spectra were recorded in the range 200–2500 cm\(^{-1}\) using Renishaw inVia Raman Microscope spec-

trometer with near infrared semiconductor laser (785 nm) as an excita-
tion 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 \(\mu\)m).

2.2. Preprocessing methods

This section 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 \(s = (s_1, s_2, \ldots, s_J)\).

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 am-
plitude. 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 combina-
tion of sinusoids and cosinuoids 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 \(\Psi\) by
its dilatation 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:

\[
\Psi(x) = a^{-1/2} \Psi \left( \frac{x - b}{a} \right), \quad a, b \in \mathbb{N}, a \neq 0.
\]

Restricting scaling parameter \(a\) to \(2^j\) and localisation parameter \(b\) to 2\(k\), 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 frac-
tion. Low pass filter, \(L\), fixed by scaling function, passes lower fre-
quencies containing baseline and true signal. The output of \(H\) is a set of
details coefficients \(W(j)\) mostly representing the high frequency noise. \(L\)
generates approximation coefficients \(V(j)\) 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:

\[
W(j)_{\text{hard}} = \begin{cases} 0, & \text{if } |W(j)| < t \\ W(j), & \text{if } |W(j)| \geq t. \end{cases}
\]

In soft thresholding the coefficients absolute values below the
threshold are set to 0 and the remaining are suppressed by this value:

\[
W(j)_{\text{soft}} = \begin{cases} 0, & \text{if } |W(j)| < t \\ \text{sgn}(W(j))|W(j)| - t, & \text{if } |W(j)| \geq t. \end{cases}
\]

\(t\) may be computed using a variety of possibilities, briefly summarised
e.g. in Ref. [24]. Universal threshold is one of the most commonly
applied:

\[
t = \sqrt{2 \log N},
\]

where \(s\) is the measure of the \(N\) wavelet coefficients dispersion expressed
as their standard deviation or more robustly with median absolute de-
viation (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 scat-
tering technique) should be applied [26]. Baseline effects arising, among
others, due to fluorescence, that do not cover the true Raman signal
totaly, 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. \( |w^t - w^{t-1}| / |w^{t-1}| < 0.01 \), 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 polynomial plus the standard deviation of the least squares model residual in the peak regions. The terminal points of the peaks are connected by a straight line and the PLS algorithm is applied for estimating the baseline variations for signals with a few major peaks, which take the control over the entire polynomial fitting. The method converges when weights do not change in two subsequent iterations 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 \((\hat{z})\) and the signal \((s)\):

\[
F = \sum_{i=1}^{J} (s_i - \hat{z}_i)^2.
\]

The roughness of the curve is quantified by computing the squared sum of differences between neighbouring points:

\[
R = \sum_{i=1}^{J-1} (\Delta z_i)^2 = \sum_{i=1}^{J} (\Delta z_i)^2.
\]

Most often, however, squared second differences are applied. In its most general form for \(m\)-th differences Equation (6) becomes

\[
R = \sum_{i=1}^{J} (\Delta^m z_i)^2.
\]

\(z\) is found with penalised least squares to minimise the expression

\[
Q = F + \lambda R,
\]

where \(\lambda\) is the penalty arbitrary assigned by the user. \(\lambda\) is a tuning parameter to control the proportion of the roughness term to \(Q\) and makes \(z\) smoother as \(\lambda\) 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

\[
F = \sum_{i=1}^{J} w_i (s_i - \hat{z}_i)^2 = (s - z)^t W (s - z)
\]

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

\[
(W + \lambda D^2)z = W_{\text{final}} \quad (9)
\]

(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

\[
z = (W + \lambda D^2)^{-1} W_{\text{final}} \quad (10)
\]

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 \(s_i > \hat{z}_i\)) and baseline segments (when \(s_i \leq \hat{z}_i\)), respectively:

\[
\left\{ \begin{array}{ll}
  w_i = p, & \text{if } s_i > \hat{z}_i \\
  w_i = 1 - p, & \text{if } s_i \leq \hat{z}_i
\end{array} \right. \quad (11)
\]

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:

\[
\left\{ \begin{array}{ll}
  w_i^t = 0, & \text{if } s_i > \hat{z}_i^t \\
  w_i^t = \exp(\gamma(s_i - \hat{z}_i^t - 1)/|d'|), & \text{if } s_i < \hat{z}_i^t
\end{array} \right. \quad (12)
\]

where \(d'\) contains the negative \(s - \hat{z} \cdot 1\) values. The idea is to assign 0 weight for peaks regions to totally eliminate them from the baseline estimation. The method converges when \(|d'| < 0.001 \cdot |s|\) 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 \(\gamma_{\text{max}}\) is assigned to peak regions and takes minimum \(\gamma_{\text{min}}\) for baseline regions. Instead of Equation (7), the cost function to be minimised is then expressed as:

\[
Q = \sum_{i=1}^{J} w_i (s_i - \hat{z}_i)^2 + \gamma_{\text{min}} \sum_{i=1}^{J} (1 - \eta w_i)(\Delta^m s_i)^2,
\]

where \(\eta = (\gamma_{\text{max}} - \gamma_{\text{min}})/\gamma_{\text{max}}\) and the weights are expressed according to Equation (12). The method converges when \(|d'| < 0.001 \cdot |s|\) 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
The baseline is then iteratively recomputed using weighted least squares to correct for the dilution of urine samples measured by NMR. It is usually selected to be the median quotient as a robust summarising value.

\[ s(\hat{v}) = a + bm(\hat{v}) + E(\hat{v}), \]

where \( a \) is the constant offset between the spectrum \( s(\hat{v}) \) and reference \( m(\hat{v}) \), \( b \) represents the multiplicative effect between \( s(\hat{v}) \) and \( m(\hat{v}) \), arising mostly from variations in laser intensity in Raman spectroscopy, and \( E(\hat{v}) \) 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

\[ s_{\text{corr}}(\hat{v}) = \left( s(\hat{v}) - a \right) / b. \]

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

\[ s(\hat{v}) = a + bm(\hat{v}) + d_1 \hat{v} + d_2 \hat{v}^2 + \ldots + d_n \hat{v}^n + E(\hat{v}), \]

where \( d_1, d_2, \ldots, d_n \) are linear, quadratic and higher polynomial degree baseline effects. The corrected spectrum is then found from

\[ s_{\text{corr}}(\hat{v}) = \left( s(\hat{v}) - a - d_1 \hat{v} - d_2 \hat{v}^2 - \ldots - d_n \hat{v}^n \right) / b. \]

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

\[ s(\hat{v}) = a + bm(\hat{v}) + d_1 \hat{v} + d_2 \hat{v}^2 + \ldots + d_n \hat{v}^n + \sum_{k=1}^{K} p_k(\hat{v}) + E(\hat{v}), \]

where \( p_k \) is the \( k \)-th from \( K \) most significant loading vectors and \( g_k \) are the corresponding fitted parameters.

### 2.3. Genetic algorithm

Genetic algorithm (GA) [5–7] 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 potential (un)common samples origins is something more than only

\[
LR = \frac{Pr(E|H_1)}{Pr(E|H_2)}
\]  

(20)

\(H_1\) is supported by the LR values larger than 1 and the support is strengthening with increasing LR. Conversely, the \(H_2\) 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 \(H_1\). It accounts for the similarity of the samples means, \(y_1\) and \(y_2\), with \(k_1\) and \(k_2\) replicate measurements, as well as the similarity of their weighted average, \(y^* = \frac{k_1 y_1 + k_2 y_2}{k_1 + k_2}\), to means of each of \(M\) sources \(y_m\) of training data. The denominator of the LR formula corresponds with \(H_2\). Then both contributions from the samples \(y_1\) and \(y_2\) are assumed independent [12–14].

When the between-source distribution is assumed normal, then LR expression is given as in Ref. [12–14]. 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 \(h = \left(\frac{2}{p \sigma^2} \right)^{1/5}\), where \(p\) is the number of considered variables. Then LR is given as a product of the following multivariate normal distributions (MVN) [12–14]:

\[
LR = \frac{MVN \left(y_1 - \bar{y}, \frac{1}{2} \sum_{m=1}^{M} MVN \left(y_m - \frac{1}{k_m} \bar{y}, \frac{1}{k_m} \sigma^2 \right) + h^2 B \right)}{MVN \left(y_2 - \bar{y}, \frac{1}{2} \sum_{m=1}^{M} MVN \left(y_m - \frac{1}{k_m} \bar{y}, \frac{1}{k_m} \sigma^2 \right) + h^2 B \right)}
\]  

(21)

For univariate data matrices or vectors (e.g. \(W, x\)) become scalars (\(w^2, x\)).

LR models quality diagnostics primarily include the levels of false positive (LR > 1 when \(H_2\) is true) and false negative responses (LR < 1 when \(H_1\) 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:

\[
 s = \log_{10} a - 1 / \left( \sum_{i=1}^{J} \log a_i - \log_{10} \frac{a}{\sqrt{\prod_{i=1}^{J} a_i}} \right).
\]  

(22)

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

Fig. 1. The concept of the studies.
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 Tables 1–3, 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, ImModPoly was skipped for preprocessing bloodstains Raman spectra and SNIP was ignored for car paints Raman spectra. 16 normalisation strategies were tested giving in total 13264

To further investigate the parameters, Table 1 lists details of denoising and smoothing strategies. BS stands for the database of Raman spectra of bloodstains and CP for car paints. The group of methods abbreviations, parameters, their values, R package and literature are shown.

<table>
<thead>
<tr>
<th>group of methods</th>
<th>abbrev.</th>
<th>parameters</th>
<th>parameter values</th>
<th>R package</th>
<th>literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Savitzky-Golay</td>
<td>SG</td>
<td>p-polynomial degree</td>
<td>( p = 3.4.5.6 )</td>
<td>signal</td>
<td>[20]</td>
</tr>
<tr>
<td>discrete wavelet transform</td>
<td>DWT</td>
<td>W-wavelet type</td>
<td>( W = \text{Daubechies Least} )</td>
<td>wavethresh</td>
<td>[21–25]</td>
</tr>
</tbody>
</table>

Table 2

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

<table>
<thead>
<tr>
<th>group of methods</th>
<th>abbrev.</th>
<th>parameters</th>
<th>parameter values</th>
<th>R package</th>
<th>literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>asymmetric penalised least squares</td>
<td>pAsWPLS</td>
<td>m-order of differences</td>
<td>( m = 2 )</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>multiplicative signal correction</td>
<td>CWTAisWPLS</td>
<td>m-order of differences</td>
<td>( m = 2 )</td>
<td>baselineWavelet</td>
<td>[32]</td>
</tr>
<tr>
<td>non-linear iterative peak-clipping</td>
<td>airPLS</td>
<td>m-order of differences</td>
<td>( m = 2 )</td>
<td>airPLS</td>
<td>[31]</td>
</tr>
<tr>
<td>selective clipping</td>
<td>2WAisPLS</td>
<td>m-order of differences</td>
<td>( m = 2 )</td>
<td>–</td>
<td>[33]</td>
</tr>
<tr>
<td>statistics-sensitive</td>
<td>multiWAisPLS</td>
<td>m-order of differences</td>
<td>( m = 2 )</td>
<td>–</td>
<td>[34]</td>
</tr>
<tr>
<td>non-linear iterative peak-clipping</td>
<td>RBE</td>
<td>b-robustness parameter</td>
<td>( b = 2.2.5 ) if BS</td>
<td>baseline</td>
<td>[37]</td>
</tr>
<tr>
<td>multiplicative signal correction</td>
<td>SNIP</td>
<td>w-clipping window</td>
<td>( w = 25.30 ) only for BS</td>
<td>MALDIquant</td>
<td>[35]</td>
</tr>
<tr>
<td>multiplicative signal correction</td>
<td>MSP</td>
<td>p-polynomial degree</td>
<td>( p = 3.4.5.6 )</td>
<td>pls</td>
<td>[40]</td>
</tr>
<tr>
<td>multiplicative signal correction</td>
<td>EMSC</td>
<td>p-polynomial degree</td>
<td>( p = 3.4.5.6 )</td>
<td>EMSC</td>
<td>[40]</td>
</tr>
<tr>
<td>multiplicative signal correction</td>
<td>repEMSC</td>
<td>pc-proportion of the explained replicates variance</td>
<td>( pc = 0.9.95 )</td>
<td>EMSC</td>
<td>[40]</td>
</tr>
<tr>
<td>polynomial methods</td>
<td>ModPoly</td>
<td>p-polynomial degree</td>
<td>( p = 3.4.5.6 ) only for CP</td>
<td>baseline</td>
<td>[27]</td>
</tr>
<tr>
<td>polynomial methods</td>
<td>IModPoly</td>
<td>p-polynomial degree</td>
<td>( p = 3.4.5.6 ) only for CP</td>
<td>–</td>
<td>[28]</td>
</tr>
<tr>
<td>polynomial methods</td>
<td>polyQR</td>
<td>p-polynomial degree</td>
<td>( p = 5.6 ) for BS ( p = 6.7 ) for CP</td>
<td>quantreg</td>
<td>[36]</td>
</tr>
<tr>
<td>polynomial methods</td>
<td>splineQR</td>
<td>q-quantile</td>
<td>( q = 0.05.0.01 ) for BS ( q = 0.05.0.01 ) for CP</td>
<td>cobs</td>
<td>[36]</td>
</tr>
<tr>
<td>polynomial methods</td>
<td>reweightedQR</td>
<td>p-polynomial degree</td>
<td>( p = 5.6 ) for BS ( p = 6.7 ) for CP</td>
<td>quantreg</td>
<td>[36]</td>
</tr>
</tbody>
</table>

Within each of the preprocessing methods all parameter values combinations listed in Tables 1–3 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 ($b^2/w^2$) 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 Tables 1–3.

### Table 3

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

Fig. 2. $b^2/w^2$ 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.
A. Martyna et al. Chemometrics and Intelligent Laboratory Systems 202 (2020) 104029

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

<table>
<thead>
<tr>
<th>Raman spectra of bloodstains</th>
<th>best</th>
<th>worst</th>
<th>GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>b²/w²</td>
<td>395</td>
<td>nearly 0</td>
<td>244</td>
</tr>
<tr>
<td>denoising</td>
<td>SG polynomial degree p = 6</td>
<td>SG polynomial degree p = 4</td>
<td>DWT</td>
</tr>
<tr>
<td>baseline correction</td>
<td>ModPoly polynomial degree p = 4</td>
<td>pAsWPLS</td>
<td>Coiflets 1 decomposition level for denoising d = 10</td>
</tr>
<tr>
<td>normalisation</td>
<td>repEMSC</td>
<td>repEMSC</td>
<td>polynomial degree p = 5</td>
</tr>
<tr>
<td>prop. of the explained replicates variance</td>
<td>pc = 0.9</td>
<td>pc = 0.95</td>
<td>penalties</td>
</tr>
<tr>
<td>Raman spectra of car paints</td>
<td>3922</td>
<td>4</td>
<td>2633</td>
</tr>
<tr>
<td>b²/w²</td>
<td></td>
<td></td>
<td>DWT</td>
</tr>
<tr>
<td>denoising</td>
<td>SG polynomial degree p = 6</td>
<td>SG polynomial degree p = 5</td>
<td>Daubechies Least Asymmetric 8 decomposition level for denoising d = 10</td>
</tr>
<tr>
<td>baseline correction</td>
<td>multiWAsPLS</td>
<td>repEMSC</td>
<td>polynomial degree p = 3</td>
</tr>
<tr>
<td>order of differences m = 2</td>
<td>penalty term λ = 10⁷</td>
<td>prop. of the explained replicates variance</td>
<td>weights p = 0.005</td>
</tr>
<tr>
<td>normalisation</td>
<td>SNV</td>
<td></td>
<td>EMSC</td>
</tr>
</tbody>
</table>

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 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 diversified 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.

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 dozens 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 clearly shows that Savitsky-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 (multi-WASPLS), 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 $\frac{b^2}{w^2}$ 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 $\frac{b^2}{w^2}$ 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 blood-stains, 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.

Figs. 3 and 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. 3e and f and 4d,e). Despite less efficient denoising strategy and thus lower legibility of the images in Figs. 3f and 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.

Figs. 5 and 6 illustrate the capability of rMANOVA to maximise $b^2/w^2$. The loadings of the first latent variable (LV1; Figs. 5a and 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 $b^2/w^2$ (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 $b^2/\omega^2$ 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. 5e and f and 6e,f plainly show that the abilities of rMANOVA to maximise $b^2/\omega^2$ 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 or signs), while these prepared
using the worst preprocessing strategies demonstrate much greater variability.

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 $LR_{GA}$), 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.

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 ($b^2 / w^2$) for the first latent variable computed from rMANOVA, as a quality parameter. Assessing the preprocessing strategies 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 criterion 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).

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.

CRediT authorship contribution statement

Agnieszka Martyna: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Visualization. Alicja Menzyk: 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.

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
