Hybrid approach combining chemometrics and likelihood ratio framework for reporting the evidential value of spectra☆

Agnieszka Martyna[a,b], Grzegorz Zadora[b,c,*], Tereza Neocleous[d], Aleksandra Michalska[b], Nema Dean[d]

[a] *Jagiellonian University in Krakow, Faculty of Chemistry, Department of Analytical Chemistry, 3 Ingardena, Krakow 30-060, Poland, Tel.: +48-12-663-20-14, Fax: +48-12-634-05-15*

[b] *Institute of Forensic Research in Krakow, 9 Westerplatte, Krakow 31-033, Poland, Tel.: +48-12-618-57-00, Fax: +48-12-422-38-50*

[c] *University of Silesia in Katowice, Institute of Chemistry, Chemometric Research Group, 9 Szkolna, Katowice 40-006, Poland, Tel.: +48-32-359-15-45, Fax: +48-32-259-99-78*

[d] *University of Glasgow, School of Mathematics and Statistics, 15 University Gardens, Glasgow G12 8QW, United Kingdom, Tel.: +44-141-330-6820, Fax: +44-141-330-4814*

Abstract

Many chemometric tools are invaluable and have proven effective in data mining and substantial dimensionality reduction of highly multivariate data. This becomes vital for interpreting various physicochemical data due to rapid development of advanced analytical techniques, delivering much information in a single measurement run. This concerns especially spectra, which are frequently used as the subject of comparative analysis in e.g. forensic sciences. In the presented study the microtraces collected from the scenarios of hit-and-run accidents were analysed. Plastic containers and automotive plastics (e.g. bumpers, headlamp lenses) were subjected to Fourier transform infrared spectrometry and car paints were analysed using Raman spectroscopy. In the forensic context analytical results must be interpreted and reported according to the standards of the interpretation schemes acknowledged in forensic sciences using the likelihood ratio approach. However, for proper construction of LR models for highly multivari-

*Corresponding author

Email addresses: rzepecka@chemia.uj.edu.pl (Agnieszka Martyna), gzadora@ies.krakow.pl (Grzegorz Zadora), tereza.neocleous@glasgow.ac.uk (Tereza Neocleous), amichalska@ies.krakow.pl (Aleksandra Michalska), nema.dean@glasgow.ac.uk (Nema Dean)

ate data, such as spectra, chemometric tools must be employed for substantial data compression. Conversion from classical feature representation to distance representation was proposed for revealing hidden data peculiarities and linear discriminant analysis was further applied for minimising the within-samples variability while maximising the between-samples variability. Both techniques enabled substantial reduction of data dimensionality. Univariate and multivariate likelihood ratio models were proposed for such data. It was shown that the combination of chemometric tools and the likelihood ratio approach is capable of solving the comparison problem of highly multivariate and correlated data after proper extraction of the most relevant features and variance information hidden in the data structure.
Keywords: dimension reduction, likelihood ratio, infrared and Raman spectroscopy, polymer, car paint, forensic science

1. Introduction

Recent developments in the field of instrumental analytical chemistry enable recording of many physicochemical features which extensively characterise the analysed samples in one single measurement. Spectroscopic methods are examples of such. Such techniques, generating the signal reflecting the nature of interaction between sample and light (e.g. intensity of absorbed, scattered or reflected light), are frequently applied for investigating the chemical features (e.g. functional groups) of the samples, often with complex chemical composition of the matrices.

Many scientific fields consider spectroscopic data as the basis of comparative analysis. It is also the case in the forensic sciences, where spectroscopy is employed for characterising for example plastics used for car body elements production (e.g. bumpers, headlamp lenses) and automotive paints collected from the scenarios of hit-and-run car accidents. Fourier transform infrared spectrometry (FTIR) can be applied for characterising organic compounds of polymers while Raman spectroscopy (RS) may be utilised for pigments identification in

car paints. For making inferences about the connections between the scene of a car accident and the suspected car, the spectra of the material collected from the car accident scenario (so-called recovered samples, whose source is unknown) are compared with the spectra of the known-source control material collected e.g. from the suspected car. Even though forensic scenarios are the illustrative examples used here in discussing the methodology for solving the so-called comparison problem, the workflow may be utilized in any field of chemistry, where the issue of comparing physicochemical features is raised. Moreover, any scientist with a specialist knowledge in some field can be asked by the court representatives to express an opinion on the casework. The analytical results must then be interpreted and reported according to the standards of the interpretation schemes acknowledged in forensic sciences [1].

Visual overlying of the spectra is unfortunately still the most frequent method for commenting on their similarity. However, despite focusing on the discrepancies in the spectras' general shapes and location of absorption bands, peaks etc., such a naked-eye comparison can only be credible for visually distinguishable spectra. In the case of very similar spectra, the resulting conclusion must be supported by more reliable tools. Moreover, when the comparison problem is addressed in the forensic sciences, the evidential value of the observed similarities and differences in the spectra must be reported. This can be expressed by the likelihood ratio (LR) approach being a well documented method for assessing the evidential value of the physicochemical data [2, 3, 4].

LR expresses the data in the context of two contrasting hypotheses. In the comparison problem they may state that:

- H_1: compared recovered and control materials come from the same source (e.g. suspected car),
- H_2: compared recovered and control materials do not come from the same source.

Due to its dichotomic nature, LR can be regarded as a reliable and objective test for making inference about the common provenance of the compared

samples based on their physicochemical data by investigating the data from two contrasting perspectives given by the LR expression:

$$LR = \frac{\Pr(E|H_1)}{\Pr(E|H_2)}. \quad (1)$$

Values of LR above 1 support H_1, while values of LR less than 1 support H_2. A value of LR equal to 1 does not provide support for either proposition. The strength of support towards each of the hypotheses is determined by the LR value itself. The larger (lower) the value of LR, the stronger the support for H_1 (H_2).

When the LR is computed for original features such as for instance elemental content of the samples (so called feature-based approach) it accounts for:

- the similarity of the features,
- the rarity of the observed features,
- the sources of uncertainty including the within- and between-objects (samples) variability in the relevant population (e.g. plastics or car paints),
- correlation between the measured features.

Including the rarity information is what makes the LR approach more suitable for assessing the evidential value than any other tests for comparing two datasets such as the t-test. The LR assigns greater support for the relevant hypothesis when the similarity is observed between rare features than when it is detected for quite common characteristics. It is worth noting that for the models in which features are replaced by e.g. distances between samples (so called score-based approach [5, 6, 7]) the rarity refers rather to the frequency of observing a particular distance than a feature.

LR models are widely developed and easily constructed for data sets described by a limited number of variables such as in the case of glass fragments characterised by their elemental composition [3] concerning only oxygen, sodium, magnesium, aluminium, silicon, potassium, calcium and iron. Similarly to most of the methods strongly embedded in statistics, LR also reveals

some limitations when dealing with highly multidimensional data, such as spectra. The main problem relates to the inability to reliably estimate the relevant parameters for LR calculations (means, variances, covariances) for data sets consisting of less samples than the number of variables they are described by. This issue is known as the *curse of dimensionality*.

Representing the spectra in the form of the so-called peaks table comprising of the areas below the limited number of spectra peaks is the easiest way for reducing their dimensionality [8]. However, this method seems to be quite time-consuming and too subjective, causing some troubles especially when establishing the boundaries of the peaks. Moreover, for some spectroscopic methods it becomes difficult to exactly identify the chemical compound responsible for the specific peak appearance.

A more convenient solution may be investigation of the dependencies between variables allowing for grouping them in clusters of highly correlated variables. This idea is the basis of graphical models, which in the forensic field are extensively used for glass elemental composition data [3, 9, 10]. In the presented approach the multidimensional problems were split into multiple problems of lower dimensions, for which LR models are more credible. However, the applied methodology is only successful for the physicochemical data sets described by only a few variables [3, 11]. Even though application of the graphical models is reported also for highly multivariate data [12, 13], the LR models based on their outcomes may not be reliable enough for legal processing.

The so-called *naïve* LR approach [14, 15] was also applied to various kinds of physicochemical data described by many variables. It assumes that the final LR value can be easily computed by multiplying all univariate LR values based on each of the variables. The justified criticism of this approach stems from the fact that it ignores the correlations between the variables. The issues with this misuse are more severe, the higher the correlation is.

In contrast to pure statistical tools, chemometrics enable investigators to extract the most relevant features from complex structured data in the original multidimensional space and represent them compactly in the form of a limited

number of variables. Therefore, the application of chemometric tools in the forensic science has been gaining importance in recent years [16, 17, 15, 14, 18, 19, 20]. In spite of their usefulness in the field of data mining, outcomes of chemometric tools cannot be directly translated and interpreted for forensic purposes, as they do not always account for all essential aspects listed above (e.g. physicochemical features rarity information, within- and between-object sources of variability and correlation structure). These, in turn, are addressed in likelihood ratio approach. However, to the authors' best knowledge, hardly any publications can be found in the literature that discuss the issue of an application of LR models accompanied by chemometric tools for highly multidimensional data such as spectra.

In [21] wavelets [22, 23, 24] were proposed for representing the spectra in a shorter form of reduced number of wavelet transform coefficients, for which univariate and multivariate LR models were constructed. The method focuses on the local spectra features as for example bands associated with particular bond vibrations, which are especially important from the chemical perspective. Limiting the relevant spectra features to those extracted from the wavelet transform may preserve sufficient information by keeping the spectra original shape deprived of irrelevant details. Even though the wavelet transform was effective in solving the comparison problem of spectra within FTIR database of polypropylene samples and Raman spectra database of car paints [21], the authors aim to study other chemometric methods also regarded as useful in dealing with high dimensionality data.

In this work, various aspects of hidden data structure were addressed using a sequence of chemometric techniques, the outcome of which was adopted as the input for LR models. The studies presented herein were aimed at verifying the suitability of joining distance representation and linear discriminant analysis for generating lower dimensional data without ignoring relevant data features. The objective was to construct LR models for the comparison problem of FTIR spectra for polypropylene and Raman spectra of solid and metallic car paints. The presented LR model can be viewed as a hybrid between feature and score

based models, as each object becomes described by a set of distances (scores) from the reference samples, acting now as new features. For clarity, the computed scores do not refer to the pairwise distances between compared samples, but constitute vectors of characteristics describing the proximity of each sample to the preselected prototypes (section 2.2.2). Moreover it is worth noting that linear discriminant analysis, a method developed and commonly used for classification, is applied here to the comparison problem.

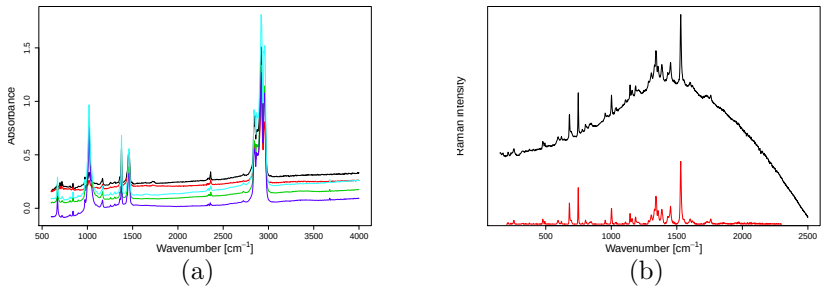
2. Materials and methods

2.1. Samples and equipment

Eleven polypropylene samples (also containing traces of other compounds such as polyethylene) used for the production of car body elements (e.g. bumpers, headlamp lenses) were the subject of Fourier transform infrared spectrometry analysis. Additionally FTIR spectra were recorded for 13 polypropylene containers which were packages for products commonly used in our daily lives (e.g. cosmetics). Due to the high probability of their occurrence on the scene of investigation (e.g. car accident), they supplemented the database. The true distribution of the non-vehicle and vehicle samples is unknown, however, more or less equal proportions for both groups were assumed. To the authors' best knowledge there is no substantial difference between the frequency of collecting the automotive plastics and containers found on the scene of car accident. Both materials are more or less equally likely to occur, which is the assumption followed in the research.

The FTIR signal was collected in the range 600-4000 cm^{-1} and for the purpose of calculations was limited to only 700-3000 cm^{-1}, capturing all the relevant chemical information. The equipment used was an FTS 40Pro Fourier transform infrared spectrometer (Bio-Rad/Digilab, Marlborough, MA), coupled with a UMA 500 microscope. For each of the 24 samples three FTIR spectra in the transmission mode were recorded from four distinct parts of the samples (with a total of $n = 12$ measurements per sample).

Thirty solid and 30 metallic blue car paints originating from bodies of damaged cars were subjected to Raman spectroscopy using a Renishaw inVia spectrometer equipped with a Leica microscope and near infrared semiconductor laser (785 nm) as an excitation source. The laser beam was focused on samples by 50x (N.A=0.75) objective lens, which gives a theoretical spot size of approximately 2 μm. The laser power applied while recording Raman spectra used 1% or 0.5% of its maximum power (300±30 mW). The spectrometer collected light in the back-scattering mode, which was dispersed on a 1200 grooves/mm grating and was focused on a Peltier-cooled charged coupled device (CCD). Spectral data was processed with Renishaw Wire 3.2 software. Spectra of all samples were recorded in situ in the region of 200-2500 cm^{-1} with an acquisition time of 10 s and collection of five accumulations. For the purpose of the calculations within this research the spectrum range was finally limited to 380-2300 cm^{-1}, which still covered all the relevant pigments Raman bands. Each paint sample was measured in three to seven spots.



(a) (b)

Figure 1: Examples of (a) FTIR polypropylene spectra, (b) Raman solid car paint spectrum before (black) and after (red) the baseline removal using the continuous wavelet transform.

2.2. Chemometric and statistical tools

The comparison problem of the recorded spectra was addressed for samples within each of the three databases:

1. $24 \times 12 = 288$ FTIR spectra recorded for 24 polypropylene samples (containing some additional compounds, e.g. polyethylene) from car body parts (headlamp lenses, bumpers) and plastic containers,
2. 100 Raman spectra recorded for 30 blue solid automotive paints,
3. 97 Raman spectra recorded for 30 blue metallic automotive paints.

2.2.1. Signal preprocessing

FTIR polypropylene spectra

For the sake of making all of the spectra comparable (Figure 1a) despite the varying thicknesses of samples, each of the 288 FTIR spectra recorded for 24 polypropylene samples in the range 700-3000 cm^{-1} was normalised using the standard normal variate (SNV) method [25]. SNV was proposed as a normalisation technique due to its independence of the whole database in contrast to other often applied techniques such as multiplicative scatter correction or probabilistic quotient normalisation [25].

Raman solid and metallic car paints spectra

Each single recorded Raman spectrum constituted the set of data points measured for a series of wavenumbers. Due to indivertible instrumental settings, the set of wavenumbers by which Raman signal intensities were measured differed between spectra. The sequence demonstrated changeable starting and ending points and varying step between subsequent elements. Such an inconvenience precludes the application of various statistical or chemometric methods aiming at interpreting the spectra similarity as they are only suitable for making inferences when samples are described by the same sets of variables (here: wavenumbers). This needed to be sorted out by reconstructing the spectra so that they reflect the Raman signal intensities measured by equally spaced wavenumbers, consistent between spectra. 1024 data points sampled every ca. 2.053 cm^{-1} were supposed to be generated in the reconstructed spectra spanning the range of 200-2300 cm^{-1}. Spline functions [26] with cubic polynomials were used for interpolating (fitting) the spectrum and finding the Raman intensities

for the new set of 1024 equally spaced wavenumbers.

Most of the Raman spectra were distorted by a fluorescence effect lifting the baseline, which made them difficult to interpret. To deal with the problem, the continuous wavelet transform (CWT) was applied for each of the analysed spectra separately (Figure 1b). Thanks to the CWT properties, the true signal can be successfully separated from the low frequency background (mainly attributed to the fluorescence effect) [27]. The baseline drift of each individual spectrum was removed using the Mexican hat wavelet [22] with the same set of CWT parameters for each spectrum (Fig. 1b). The reason for creating the number of 1024 points per spectrum in the reconstruction step was the CWT, which operates on 2^N long signals.

For the sake of calculations the spectra were limited to the range 380-2300 cm^{-1} in order to exclude the initial part of spectrum for which CWT hardly managed to separate the baseline from the true signal. They were then normalised using the standard normal variate (SNV) method.

2.2.2. Reduction of data dimensionality

Each of the three databases consists of m objects (for FTIR spectra it is $m_{IR} = 24$, for solid and metallic Raman spectra databases $m_R = 30$), each measured n_i times (n_i is constant, $n = 12$, for FTIR spectra database and differs between 3 and 7 in Raman spectra databases) and described by ν variables (understood as signal intensities measured for a set of ν wavenumbers). After preprocessing steps (see section 2.2.1) the FTIR polypropylene spectra database is a matrix of size 288 spectra $\times$ 1193 variables (signal intensities measured by 1193 wavenumbers), the solid car paints Raman spectra database is a matrix of size 100$\times$936 and the metallic car paints Raman spectra database is a matrix of size 97$\times$936.

The most natural and powerful way for reducing the dimensionality of physicochemical data is by principal component analysis (PCA). Despite of its great capability for effective dimension reduction, the main disadvantage of the PCA is that it fails in distinguishing between the within- and between samples variabil-

ity as illustrated in Figure 2. Its aim is to search for the directions that maximise the overall variance of the data without considering any division into various objects described by multiple spectra (dashed red line in Figure 2). As a consequence the generated space is not optimal for maximising the between-objects variability while minimising the within-objects variability. Since LR models perform effectively only when the between-object variance is much greater than the within-object variance, preserving the relation $\mathbf{C} \gg \mathbf{U}$ is crucial from the perspective of LR model efficacy. Here $\mathbf{C}$ denotes the between-sample variance-covariance matrix (Equation 6) and $\mathbf{U}$ the within-sample variance-covariance matrix (Equation 5). Otherwise, the distributions of the variables describing objects overlap, making it difficult to distinguish between them.

Furthermore a dataset of samples characteristic features (e.g. in the form of spectra) may not always be the most straightforward and informative. Dealing with so many variables (features) as the number of wavenumbers for which intensities of spectroscopic signals were measured is both inconvenient and prevents easy observation of the genuine spectra similarities and differences. As suggested in [28, 29, 30] the so-called distance representation has many advantages over classical feature representation. It not only enables for substantial data compression, but also easily copes with the non-linear problems [28]. Nevertheless, the methodology causes some information loss especially regarding the influence of the original features on the results but it is always a matter of finding a compromise between losing some information and reducing data dimensionality so that they are adjusted to further processing steps. Applying data dimensionality reduction techniques may deteriorate the results when applied prior to the classification [31], however, in the case of spectra initial data dimensionality reduction is absolutely inevitable as without this step hardly any method for their evaluation can be used, the likelihood ratio approach at all. Therefore using distance representation instead of feature space must be considered as a *quid pro quo* methodology, which is assumed to preserve enough information for solving the stated problem.

In the distance representation approach, each sample is described by the

distance measure (Manhattan, Euclidean, etc.) from the predefined reference samples (e.g. mean spectrum, reference material spectrum). Then the matrix originally containing P wavenumbers by which intensities were measured for M spectra (matrix size $M \times P$) is reduced to a matrix of $M \times R$, where R stands for the number of reference samples spectra the distances are measured to. In distance representation each sample (spectrum) is described by features that are its distances to the reference objects. In the presented research the suitability of five different distance metrics including Manhattan, Euclidean, squared Euclidean, correlation-based and Chebyshev distance was examined. The correlation based distance metric between two spectra A and B is calculated as $1 - r_{AB}$, where r_{AB} is the Pearson's correlation coefficient between the two spectra. For this reason the values it takes fall between 0 and 2, whereas the remaining considered distance metrics produce non-negative values.

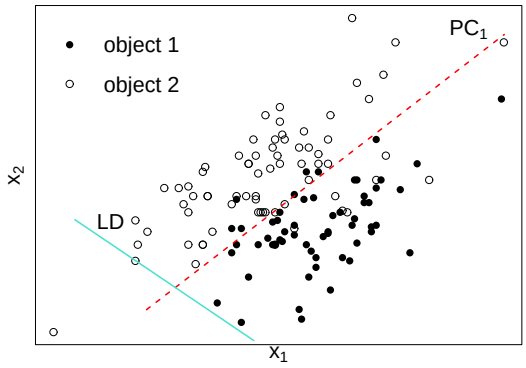
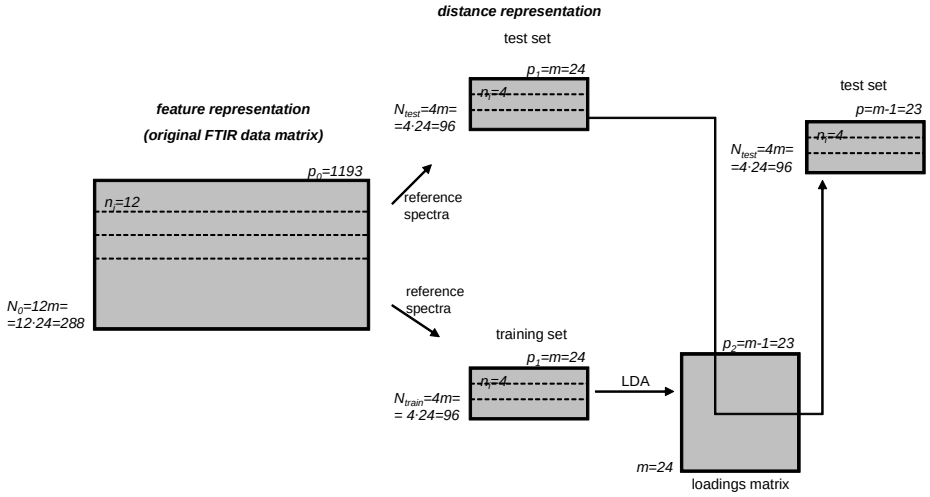
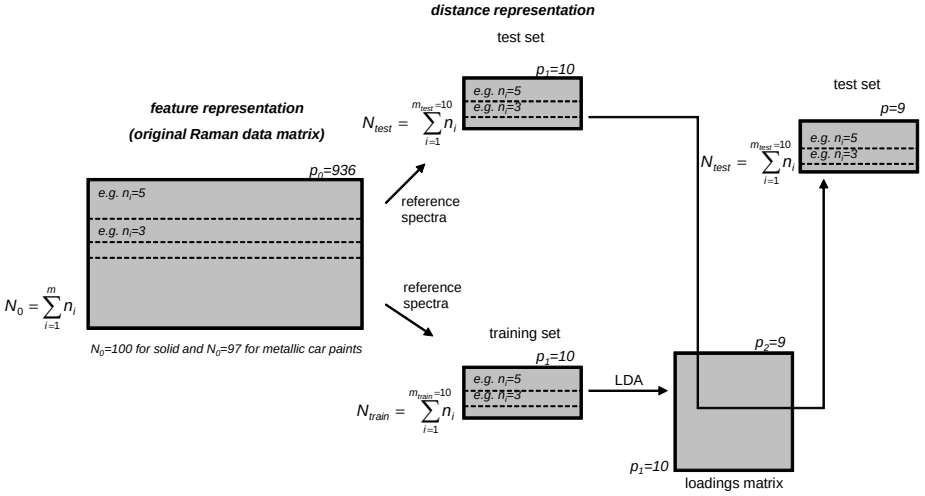


Figure 2: The illustration of differences between the variance aspects PCA and LDA address.

However, for meeting the condition of much greater between- than within-objects variability, $\mathbf{C} \gg \mathbf{U}$, linear discriminant analysis (LDA) [32] was applied to data in distance representation prior to LR models construction. Even though the method is known for classification purposes, it can be successfully applied for maximising the ratio of between- to within-*object* variabilities (instead of between- to within-*class* variabilities) when objects are regarded as separate classes. Such an approach enables finding the best direction in which the vari-



(a)



(b)

Figure 3: The scheme illustrating the validation procedure for (a) FTIR polypropylene spectra database, (b) Raman solid or metallic car paints databases. Notation: m-number of objects, N_0 - number of original spectra in the considered FTIR and Raman spectra databases (subscripts *test* and *train* refer to the test and training sets respectively), n_i-number of multiple spectra recorded for each of m samples (constant for FTIR spectra database, $n_i = 12$ and varying between 3 and 7 for Raman spectra databases), p-number of variables considered at each step indicated by a subscript.

ance related to the differences between objects is captured and the random within-object variability is diminished (solid blue line in Figure 2). LDA can be used for compressing data as new directions are found in the sequence of decreasing ratio of between- to within-classes (here objects) variance. Therefore accounting for first few LDs will enable for capturing the greatest part of variance associated with data grouping. LDA has also one more advantage stemming from the fact that eigenvectors (LDA directions) are scaled to generate unit within-object variances. This is seen as an additional merit, since LR models also assume the same and constant within-object variance for all objects.

As a consequence, LDA enabled for further substantial reduction of data dimensionality in distance representation. The projections of the data on linear discriminant directions were further considered as variables for LR models construction.

According to the mathematical principles of LDA, the projections representing the spectra in the LDA space are both orthogonal and uncorrelated. This property enables adopting the easiest solution to create multivariate LR models (section 2.2.3). The idea is to convert the p-dimensional problem into p univariate models. The solution involves the construction of univariate LR models based on single linear discriminants (LDs) and their multiplication for getting the final LR value (Figures 3 and 4, Equations 2-4). Such an approach is known as *naïve*, in which it is assumed that the contribution of each of the variables in the support for the relevant hypothesis (H_1 or H_2) is independent of the other variables. Its only limitation relates to the assumption of the lack of correlation between the variables and escalates with increasing correlation [6, 18]. However, for the examined data, the assumption still holds thanks to orthogonalisation and centering of the LDA space. The *naïve* LR models constructed from LDs are only reasonable when they account for the LDs with the greatest ability to distinguish between all samples. For selecting such LDs the algorithm proposed by Raftery and Dean in [33] was applied. It is based on the use of the Bayes factor and Bayesian Information Criterion (BIC) for selecting the variables with

the greatest ability to put each single sample into a separate group [33, 34]. The algorithm selects LDs by iteratively proposing keeping/removing an LD one at a time. Through a stepwise procedure, each LD is considered for retention or removal from consideration given the set of LDs already retained. BIC scores are used to compare the model where the LD under consideration is assumed to show clustering in the data (where clustering indicates grouping replicate measurements of samples into groups which are regarded as separate samples) along with the other currently retained LDs and the competing model where the already retained LDs provide clustering but the LD under consideration has a non-clustering model. This comparison basically indicates whether the new LD provides an improvement in separating the samples over and above the LDs already selected. Since the same set of LDs is being used in both models, the BIC scores can be compared and the model that represents the best BIC score will indicate the relevant decision to be made regarding retention or removal of the LD of interest. Iterating this procedure until convergence results in a set of selected LDs.

The studies examined the performance of the following *naïve* LR models:

- univariate LR models:

$$LR(LD_{BIC_1}) \tag{2}$$

- bivariate LR models:

$$LR(LD_{BIC_1}) \cdot LR(LD_{BIC_2}) \tag{3}$$

- trivariate LR models:

$$LR(LD_{BIC_1}) \cdot LR(LD_{BIC_2}) \cdot LR(LD_{BIC_3}) \tag{4}$$

Validation of the procedure

Two different approaches were applied for creating the training and test sets for validation purposes (Figure 3):

- for the FTIR polypropylene spectra database there were $s = 100$ training and test sets randomly generated using the *leave-p-out* method. In each iteration 4 randomly selected spectra from each of the m_{IR} samples were averaged and constituted a set of m_{IR} reference spectra. Another 4 randomly selected spectra recorded for each of the m_{IR} samples formed the training set for LDA space construction. The remaining 4 from each sample formed the test set for LR models (Figure 3a). Both training and test sets consisted of $m = 24$ exactly the same samples, but different measurements. Finally, the test and training sets in distance representation were matrices of size $4{\cdot}24{=}96{\times}24$ since the distances from 96 spectra were computed to 24 averaged reference spectra. LDA performed on the training set reduced the number of variables to the minimum of the number of groups (here samples, whose number is equal to 24) minus one and the number of variables, which is 24. This produces 24-1=23 new orthogonal features. The final test set matrix used in the LR calculations was developed by applying the LDA model parameters on the distance representation of the test set spectra and was described by 96 rows referring to the spectra of the 24 samples and 23 columns referring to linear discriminant directions acting as new variables.

- for the Raman car paints spectra databases, due to limited and differing number of spectra recorded for each car paint sample (so-called unbalanced data), the validation sets were constructed in a samples-driven way. Ten samples spectra were randomly selected and averaged to create 10 reference spectra, the training set consisted of 10 randomly selected car paint samples (with all their spectra) and the 10 samples left formed the test set (Figure 3b). Such test and training sets were created $s = 100$ times. Since the number of spectra recorded for each sample (n_i) differs, the dimensionalities of test and training sets are not straightforward. Nevertheless, the test and training sets in distance representation were matrices of size $10{\cdot}n_i{\times}10$ since the distances from $10{\cdot}n_i$ spectra were computed to 10 av-

eraged reference spectra. LDA performed on the training set reduced the number of variables to the minimum of the number of groups (here samples, whose number is equal to 10) minus one and the number of variables, which is 10. This produces 10-1=9 new orthogonal features. The final test set matrix was developed by applying the LDA model parameters on the distance representation of the test set spectra and was described by $10 \cdot n_i$ rows referring to the spectra of the 10 samples and 9 columns referring to the linear discriminant directions acting as new variables.

The relevant dimensionality of the FTIR and Raman databases in the distance representation and after performing LDA is given schematically in Figure 3. From now on these matrices of lower dimensions will serve as databases for LR models construction.

2.2.3. Likelihood ratio

In each of $s = 100$ iterations (for validating LDA models) one final test set was constructed separately for FTIR and Raman databases. This test set is supposed to be the foundation for LR models construction. For validating the LR models new training and test sets were derived from the final test sets of FTIR and Raman data matrices after performing LDA on the spectra distance representation. When different samples were under comparison, they constituted a single LR test set. The corresponding LR training set for estimating the LR models parameters (means, variances, covariances) consisted of the remaining samples. The procedure was repeated until all possible comparisons in the database were completed (including comparison of data for the two samples coming from the same object). New defined LR test and training sets will be denoted with relevant subscripts.

Each i-th ($i = 1, \ldots, m_{train}$) object from the training set is described by n_i vectors ($j = 1, \ldots, n_i$) containing data of p variables being projections of distance representations of spectra on the linear discriminants, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$. Their number is equal to $p = m = 24$ for FTIR spectra database and $p = 10$ for Raman spectra databases.

The mean vector of p variables for each object is given as $\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$. The overall mean μ is estimated from $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{m_{train}} n_i\bar{\mathbf{x}}_i$, where $N = \sum_{i=1}^{m_{train}} n_i$. The distribution of $\mathbf{x}_{ij}$ given $\bar{\mathbf{x}}_i$ is assumed normal and characterised by the within-object variance $\mathbf{U}$ that is for all objects equal: $(\mathbf{x}_{ij}|\bar{\mathbf{x}}_i, \mathbf{U}) \sim N(\bar{\mathbf{x}}_i, \mathbf{U})$. The within-object variance-covariance estimate ($\hat{\mathbf{U}}$) is expressed by:

$$\hat{\mathbf{U}} = \frac{1}{N}\sum_{i=1}^{m}\frac{n_i}{n_i - 1}\mathbf{S}_{wi}, \tag{5}$$

where:

$$\mathbf{S}_{wi} = \sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T.$$

The objects means $\bar{\mathbf{x}}_i$ are distributed around μ estimated as $\bar{\mathbf{x}}$ with dispersion given by the between-object variability ($\mathbf{C}$). The distribution can be assumed normal $(\bar{\mathbf{x}}_i|\mu, \mathbf{C}) \sim N(\mu, \mathbf{C})$ or is estimated by kernel density estimation (KDE) procedure using Gaussian kernels with bandwidth parameter calculated as $h = \left(\frac{4}{m_{train}(2p+1)}\right)^{\frac{1}{p+4}}$ [35].

The between-object variance-covariance estimate ($\hat{\mathbf{C}}$) is expressed as:

$$\hat{\mathbf{C}} = \frac{N}{N^2 - \sum_{i=1}^{m_{train}} n_i^2}\left(\mathbf{S}_b - (m_{train} - 1)\hat{\mathbf{U}}\right), \tag{6}$$

where:

$$\mathbf{S}_b = \sum_{i=1}^{m_{train}} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T.$$

LR models constructed using the parameters ($\mathbf{U}$, $\mathbf{C}$, μ, $\bar{\mathbf{x}}_i$, $\bar{\mathbf{x}}$, h) estimated from the training set are subsequently applied for comparing each two objects from the test set, known as recovered and control samples.

For control object $\mathbf{y}_1$ there are k_1 observations of the p variables:

$$\mathbf{y}_{1j} = (y_{1j1}, \ldots, y_{1jp})^T$$

with mean vector $\bar{\mathbf{y}}_1 = \frac{1}{k_1}\sum_{j=1}^{k_1}\mathbf{y}_{1j}$, coming from normal distribution $(\bar{\mathbf{y}}_1|\mu, \mathbf{U}, \mathbf{C}) \sim N\left(\mu, \frac{\mathbf{U}}{k_1} + \mathbf{C}\right)$.

For recovered object $\mathbf{y}_2$ there are k_2 observations of the p variables:

$$\mathbf{y}_{2j} = (y_{2j1}, \ldots, y_{2jp})^T$$

with mean vector $\bar{\mathbf{y}}_2 = \frac{1}{k_2}\sum_{j=1}^{k_2} \mathbf{y}_{2j}$, coming from normal distribution $(\bar{\mathbf{y}}_2|\mu, \mathbf{U}, \mathbf{C}) \sim N\left(\mu, \frac{\mathbf{U}}{k_2} + \mathbf{C}\right)$.

The weighted mean of $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ is given as $\bar{\mathbf{y}}^* = \frac{k_1\bar{\mathbf{y}}_1 + k_2\bar{\mathbf{y}}_2}{k_1+k_2}$ and comes from normal distribution $(\bar{\mathbf{y}}^*|\mu, \mathbf{U}, \mathbf{C}) \sim N\left(\mu, \frac{\mathbf{U}}{k_1+k_2} + \mathbf{C}\right)$.

In the case of univariate data all matrices or vectors in LR expressions become scalars, e.g. $\mathbf{U}$ becomes u^2 and $\bar{\mathbf{x}}$ is reduced to $\bar{x}$.

In the LR numerator it is assumed that the compared samples originate from the same object (H_1). Thus, it can be shown that the numerator can be expressed by explicitly taking into account the evaluation of the difference of physicochemical data between compared objects, as well as their rarity, the latter being expressed by the distance of the weighted mean $\bar{\mathbf{y}}^*$ from the overall mean $\bar{\mathbf{x}}$.

Therefore, when kernel density estimation is applied for modelling the between-object distribution, the numerator could be expressed by [4, 2, 36]:

$$\begin{aligned} f(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2|\mathbf{U}, \mathbf{C}, \bar{\mathbf{x}}, H_1) = f(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2|\mathbf{U}, H_1) \times \frac{1}{m_{train}} \sum_{i=1}^{m_{train}} K(\bar{\mathbf{y}}^*|\bar{\mathbf{x}}_i, \mathbf{U}, \mathbf{C}, h, H_1) = \\ = (2\pi)^{-p/2} \left|\frac{\mathbf{U}}{k_1} + \frac{\mathbf{U}}{k_2}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^T \left(\frac{\mathbf{U}}{k_1} + \frac{\mathbf{U}}{k_2}\right)^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)\right\} \times \\ \times (2\pi)^{-p/2} \left|\frac{\mathbf{U}}{k_1+k_2} + h^2\mathbf{C}\right|^{-\frac{1}{2}} \frac{1}{m_{train}} \sum_{i=1}^{m_{train}} \exp\left\{-\frac{1}{2}(\bar{\mathbf{y}}^* - \bar{\mathbf{x}}_i)^T \left(\frac{\mathbf{U}}{k_1+k_2} + h^2\mathbf{C}\right)^{-1} (\bar{\mathbf{y}}^* - \bar{\mathbf{x}}_i)\right\}. \end{aligned} \tag{7}$$

In the denominator of the likelihood ratio, $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ are taken to be independent as the data are assumed to originate from different objects (H_2). When KDE is applied for modelling the between-object distribution, it takes the form [4, 2, 36]:

$$\begin{aligned} f(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2|\mathbf{U}, \mathbf{C}, \bar{\mathbf{x}}, H_2) = \frac{1}{m_{train}} \sum_{i=1}^{m_{train}} f(\bar{\mathbf{y}}_1|\bar{\mathbf{x}}_i, \mathbf{U}, \mathbf{C}, h, H_2) \times \frac{1}{m_{train}} \sum_{i=1}^{m_{train}} f(\bar{\mathbf{y}}_2|\bar{\mathbf{x}}_i, \mathbf{U}, \mathbf{C}, h, H_2) = \\ = (2\pi)^{-p/2} \left|\frac{\mathbf{U}}{k_1} + h^2\mathbf{C}\right|^{-\frac{1}{2}} \frac{1}{m_{train}} \sum_{i=1}^{m_{train}} \exp\left\{-\frac{1}{2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{x}}_i)^T \left(\frac{\mathbf{U}}{k_1} + h^2\mathbf{C}\right)^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{x}}_i)\right\} \times \\ \times (2\pi)^{-p/2} \left|\frac{\mathbf{U}}{k_2} + h^2\mathbf{C}\right|^{-\frac{1}{2}} \frac{1}{m_{train}} \sum_{i=1}^{m_{train}} \exp\left\{-\frac{1}{2}(\bar{\mathbf{y}}_2 - \bar{\mathbf{x}}_i)^T \left(\frac{\mathbf{U}}{k_2} + h^2\mathbf{C}\right)^{-1} (\bar{\mathbf{y}}_2 - \bar{\mathbf{x}}_i)\right\}. \end{aligned} \tag{8}$$

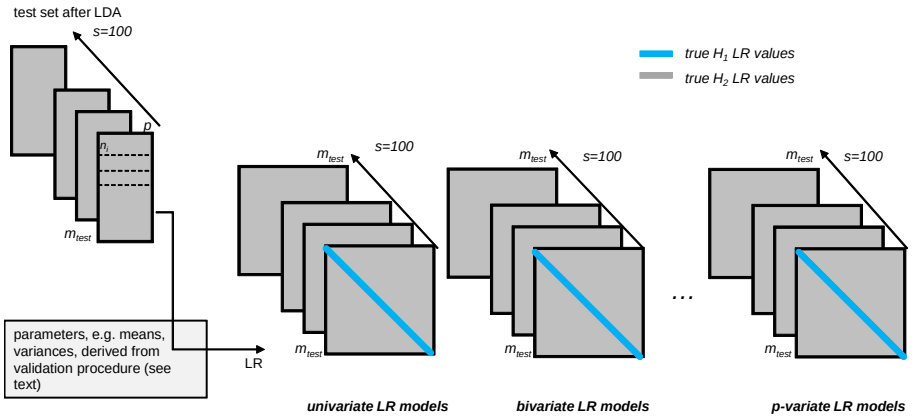


Figure 4: The structure of the symmetric matrices of LR results (size $m_{test} \times m_{test}$, where m_{test} refers to the number of objects in the test set of FTIR or Raman spectra databases after LDA, Figure 3) obtained for each of the $s = 100$ test sets within particular *naïve* LR models. The diagonals of the matrices archive the true H_1 LR values (expected to be greater than 1) for comparing the objects sharing the same origins (estimation of false negative answers), whereas gray areas below or above the diagonals store the true H_2 LR values (expected to be lower than 1) for comparing the objects with different origins (estimation of false positive answers). n_i refers to the number of multiple spectra recorded for each of m_{test} samples (constant for FTIR spectra database, $n_i = 12$, and varying between 3 and 7 for Raman spectra databases), p denotes the number of variables considered after performing LDA.

2.2.4. Assessing the LR models performance - false positive and false negative rates and empirical cross entropy approach

When constructing new LR models, special care should be taken for assessing their performance and correctness of the yielded results. Providing the levels of correct LR models responses belongs to the most coarse methods for commenting on the LR models effectiveness. In the proposed research the levels of false positive and false negative answers were under examination [4]. For their estimation the following comparisons were simulated:

(1) LR is computed for two samples sharing common origin, i.e. one sample from the test set was split into two samples (one acting as recovered and the second one as control sample). The samples with even number of spectra were divided into two equally numerous parts. For samples with odd number of spectra, the control sample consisted of one more observation than the recovered one. The correct model response should support H_1 (hence $LR > 1$). Each value of $LR < 1$ is then a false negative answer. The number of comparisons performed within this experiment is then $N_1 = m_{IR} = 24$ for each of $s = 100$ test sets of FTIR polypropylene spectra and $N_1 = m_{R,test} = 10$ for Raman spectra databases.

(2) LR is computed for each two objects from the test set. Since they do not share common origin, H_2 should be supported (hence $LR < 1$) and each value supporting the H_1 ($LR > 1$) is known as the false positive answer. Total number of the comparisons performed in this experiment for each of $s = 100$ test sets is given by $N_2 = \frac{m_{IR} \cdot (m_{IR}-1)}{2} = \frac{24 \cdot (24-1)}{2} = 276$ for each test set of FTIR polypropylene spectra and $N_2 = \frac{m_{R,test} \cdot (m_{R,test}-1)}{2} = \frac{10 \cdot (10-1)}{2} = 45$ for Raman spectra databases.

When considering only the levels of false positive and false negative answers the information concerning the strength of the support for particular hypothesis is hidden. For instance, false positive answers of $LR = 5$ and $LR = 500$ are both regarded in favour of H_1, even though the strength of the support towards this hypothesis varies significantly for both values. Therefore, information on

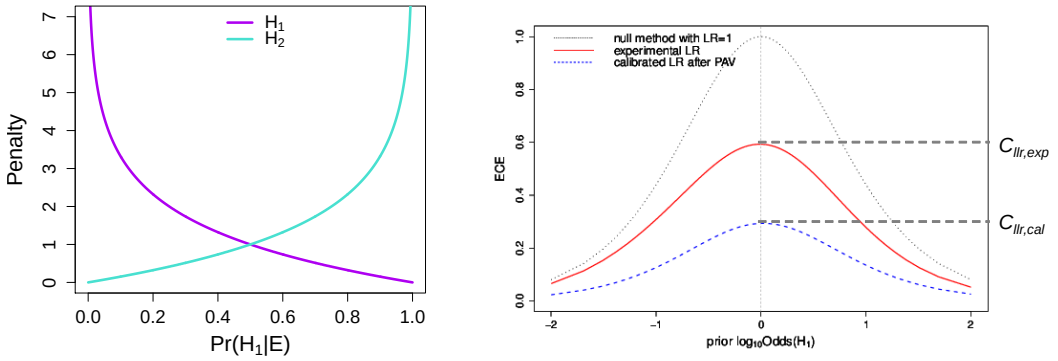
false positive and false negative rates is usually insufficient for portraying an extensive and credible assessment of the performance of the LR models.

In hypothesis testing it is not only desirable to support the correct hypothesis but also that this support should be as strong as possible (i.e. $LR >> 1$ when H_1 is correct and $LR << 1$ when H_2 is correct). If an incorrect hypothesis is supported by LR value (i.e. $LR < 1$ when H_1 is true and $LR > 1$ when H_2 is true) then LR value should concentrate around 1, supporting the incorrect hypothesis very weakly. Therefore LR models, capable of pointing out the strength of support towards one of the hypotheses, should be extensively exploited by taking advantage of this information, instead of only the indication of the supported hypothesis. This information is crucial from the perspective of Bayesian theory (Equation 9), in which LR values modify the prior assumptions ($\Pr(H_1)$ and $\Pr(H_2)$) about the evidence stated before its analysis. Such a modification generates final results in the form of ratio of conditional probabilities $\Pr(H_1|E)$ and $\Pr(H_2|E)$, namely posterior probabilities.

$$\frac{\Pr(H_1)}{\Pr(H_2)} \cdot \frac{\Pr(E|H_1)}{\Pr(E|H_2)} = \frac{\Pr(H_1)}{\Pr(H_2)} \cdot LR = \frac{\Pr(H_1|E)}{\Pr(H_2|E)}. \tag{9}$$

Bayes theorem became the foundation of the empirical cross entropy (ECE) approach for assessing the LR models performance enclosing the values of LR, prior and posterior probabilities [37, 38, 10, 9, 4]. It is a method based upon the system of rewarding and penalising the obtained LR values. The LR models responses supporting the incorrect hypothesis are penalised according to logarithmic strictly proper scoring rules (Fig. 5a). The higher the support for the incorrect hypothesis, the greater penalty is assigned to the model's response, i.e. $-\log_2 \Pr(H_1|E)$, when H_1 is true and $-\log_2 \Pr(H_2|E)$, when H_2 is true.

The ECE as a proper measure of performance is then proposed by taking into account the mean penalties computed from N_1 and N_2 experiments performed in the aim to estimate the rates of false negative and false positive answers (under H_1 and H_2 hypotheses), which are weighted by the relevant prior probabilities $\Pr(H_1)$ and $\Pr(H_2)$:



(a) (b)

Figure 5: (a) Logarithmic strictly proper scoring rules, (b) empirical cross entropy (ECE) plot (description in the text).

$$\mathrm{ECE} = \frac{\Pr(H_1)}{N_1}\sum_{i=1}^{N_1}\log_2\left[1+\frac{\Pr(H_2)}{\mathrm{LR}_i\Pr(H_1)}\right] + \frac{\Pr(H_2)}{N_2}\sum_{j=1}^{N_2}\log_2\left[1+\frac{\mathrm{LR}_j\Pr(H_1)}{\Pr(H_2)}\right]. \quad (10)$$

The ECE is computed for the set of possible prior probability quotients $\Pr(H_1)/\Pr(H_2)$ since assigning any particular values to the *a priori* probabilities $\Pr(H_1)$ and $\Pr(H_2)$ is not the task of the forensic expert. Therefore the ECE outcome is illustrated in the form of a diagram presenting the ECE values for a set of $\Pr(H_1)/\Pr(H_2)$ commonly referred to as *prior odds* in favour of H_1: $\log_{10} Odds(H_1)$. Each diagram consists of three curves, which relative location represents the LR model performance (Fig. 5b):

(a) the solid (red) curve (named *experimental*) - represents the ECE values calculated using the LR model, which performance is under investigation (Equation 10),

(b) the dashed (blue) curve (named *calibrated*) - represents the ECE values obtained for the experimental LR values transformed with Pool Adjacent Violators algorithm (PAV) [39, 40, 4]. The discriminating power, expressed by the levels of false positive and false negative answers, of the

calibrated set of LRs remains unchanged, even though it boasts the best performance (i.e. LR values strongly support the correct hypotheses and give weak support for the incorrect). Therefore, the observed differences between the *calibrated* curve and the ECE curve for the experimental LR set indicate the possibility of improving the model performance,

(c) the dotted (black) curve (named *null*) - represents the performance of a model, which does not support any of the hypotheses ($LR = 1$). This model always produces identical curves, acting as reference curves for estimating the model performance.

The relative location of the ECE curve for the experimental set of LR values (solid, red line) to the remaining two (dashed, blue and dotted, black lines) illustrates the performance of the LR model as a way for commenting on the evidential value. If there is still too much uncertainty within the model about the correct hypothesis, the ECE curve for experimental LR values will grow, and more information will be needed in order to identify the true hypothesis. The best models for the interpretation of the evidence under analysis are those for which the ECE curve lies as low as possible and as close as possible to the *calibrated* curve. If the curve appears to have greater values than the ones in the *null* method, the evidence evaluation introduces more misleading information than when assuming evidence neutrality ($LR = 1$). For the purposes of this research, the information about the reduction of information loss due to the analysis of evidence always refers to the point of $log_{10}Odds(H_1) = 0$ (Figure 5), although it could also be compared for any value of the prior odds.

2.3. Software

All the calculations including the spectra pretreatment, chemometric methods application and LR calculations were performed using R software [41], including the *MASS* [42] and *mclust* [43] packages implemented in the environment.

3. Results and discussion

3.1. Descriptive statistics

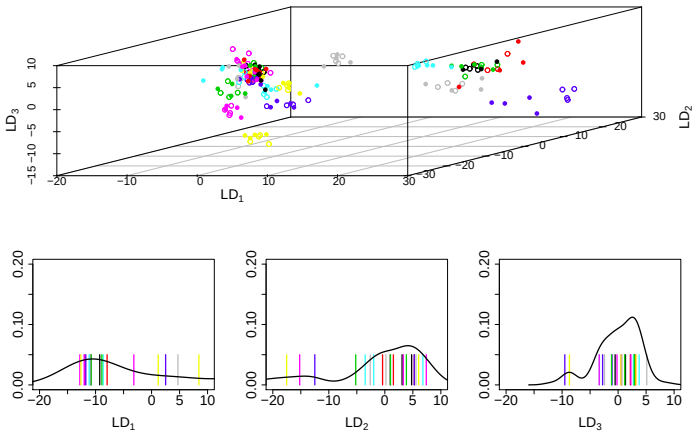
The assumption of normal within-sample distribution and constant within-sample variance was made. However, it cannot be verified formally as there are only a few replicate measurements within each sample, which is not enough for any statistical test for normality.

Figure 6 demonstrates the projections of spectra features in the distance representation on the first three linear discriminant directions. The plots in Figure 6 refer to only single randomly selected test sets. The diagrams visualise the effect of applying LDA for maximising the between-objects variability and minimising the within-object variability. This is portrayed by well separated groups of observations coming from the same samples of the training set (see same colour empty dots in Figure 6). The projections of test data (see same colour full dots in Figure 6) in most cases also generate clusters of points representing the spectra recorded for the same sample. However, their within-object distributions seem to be more dispersed than for training data. This finding is most likely due to the fact that the training sets are not as numerous as they should be for being representative of the whole population and in many cases may be insufficient for extensive description of the data genuine structure.

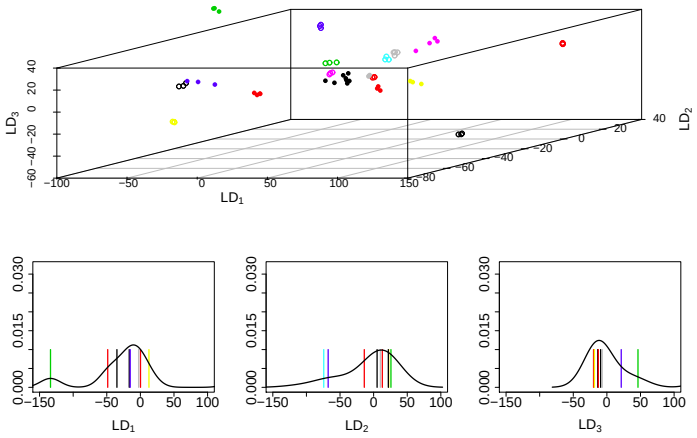
The plots in Figure 6 outline also the underlying distribution of data for first three linear discriminant directions, which cannot be assumed to be normal. For this reason, kernel density estimation was applied for modelling the between-object distribution in a non-parametric way (see section 2.2.3).

3.2. LR models performance

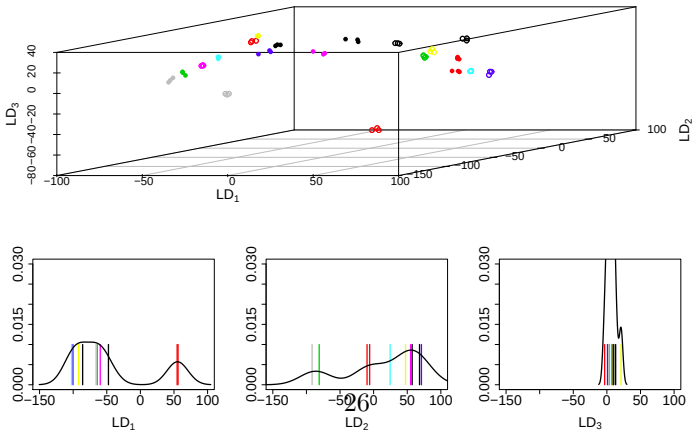
In previous sections it was shown that usually only first few LDs seem to be credible variables for LR models construction, i.e. delivering $\mathbf{C} \gg \mathbf{U}$. They can be selected using the BIC criterium and then used for constructing univariate, bivariate, trivariate etc. *naïve* LR models. The presented results of LR calculations account only for univariate, bivariate and trivariate LR models. For



(a)



(b)



(c)

Figure 6: (a) FTIR polypropylene spectra, (b) Raman solid car paints spectra, (c) Raman metallic car paints distance representation of spectra, in the space of first three linear discriminants. The plots refer to randomly selected pairs of training and test sets.

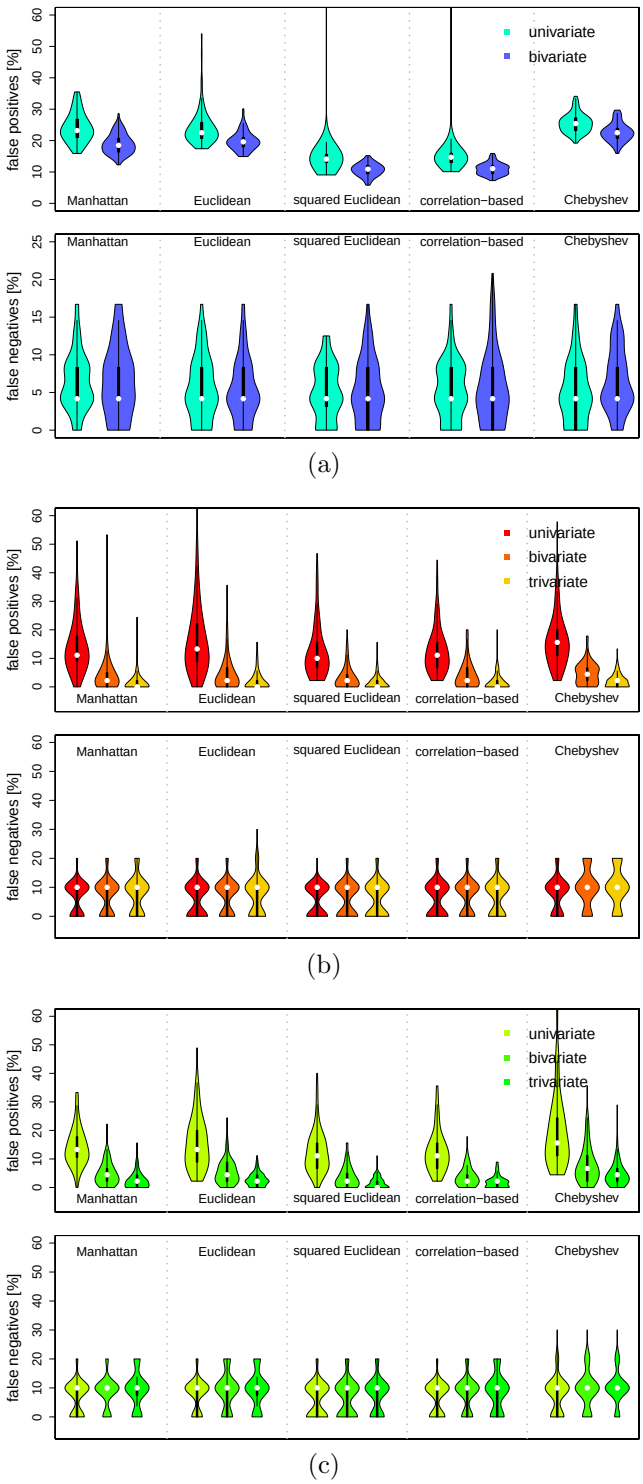


Figure 7: Illustration of the levels of false positive and false negative answers for univariate (Equation 2), bivariate (Equation 3) and trivariate (Equation 4) *naïve* LR models in regard to the distance metric for (a) FTIR polypropylene spectra, (b) Raman solid car paints spectra, (c) Raman metallic car paints spectra. Each violin plot accounts for all generated s=100 test sets. For the FTIR spectra database there are only uni- and bivariate LR models presented due to very limited number of results for trivariate ones.

the FTIR spectra database there are only univariate and bivariate LR models presented due to insufficient number of results for trivariate LR models. It appears as a consequence of the BIC criterion that usually chose only two LDs that proved useful.

The experiments for estimation of the levels of false positive and false negative answers were designed for the test sets according to the scheme described in section 2.2.4 (Figure 4). For each comparison between two samples there were $s = 100$ LR values (being consequence of $s = 100$ random selections of test and training sets) within each LR model expressing the evidential value of their similarity. Figure 7 illustrates the overall performance of the LR models under investigation in regard to the various distance metrics. Each violin plot, being a hybrid of a boxplot and a kernel density plot curved along the boxplot sides, corresponds with the false positive or false negative outcomes yielded from $s = 100$ test sets (Figure 4).

The diagrams in Figure 7 clearly demonstrate the dependency between the false positive or false negative rates observed for various LR models. Regardless of the distance metric it was noted that accounting for more LDs in *naïve* LR models leads to decreasing of the false positive rates. This is not observed for false negative rates, which rather stabilise around 5% for FTIR spectra of polypropylene and 10% for Raman spectra of car paints. As demonstrated in Figure 7a the rates of incorrect models responses seem to be the lowest (below 20% for univariate LR models and around 10% for bivariate ones) when squared Euclidean and correlation based distance metrics were used. The outcomes for Raman spectra databases are more stable and insensitive to the chosen distance metrics and oscillate between ca. 20% for univariate LR models to less than 5% for trivariate ones. The general impression indicates that on the basis of the LR approach Raman spectra for car paints are more effectively differentiated than FTIR spectra for polymers.

The results of ECE plots accounting for the strength of the support towards both hypotheses also confirm above findings. Table 1 demonstrates the ECE values for particular LR models ($C_{llr,exp}$, Figure 5b) and the calibrated ECE

Table 1: Empirical cross entropy results concerning $C_{llr,exp}$ (corresponding with experimental solid red curve in Figure 5b at $log_{10}Odds(H_1) = 0$) or $C_{llr,cal}$ (corresponding with calibrated dashed blue curve in Figure 5b at $log_{10}Odds(H_1) = 0$) yielded for the proposed LR models. For FTIR spectra database there are only univariate and bivariate LR models presented due to insufficient number of results for trivariate LR models.

FTIR polypropylene spectra						
	$C_{llr,exp}$ [%]	$C_{llr,cal}$ [%]	$C_{llr,exp}$ [%]	$C_{llr,cal}$ [%]	$C_{llr,exp}$ [%]	$C_{llr,cal}$ [%]
	univariate		bivariate		trivariate	
Manhattan	41.8	33.7	38.5	26.9	-	-
Euclidean	41.7	30.3	46.0	27.2	-	-
squared Euclidean	27.7	17.7	22.3	14.5	-	-
correlation-based	27.8	19.9	22.7	15.3	-	-
Chebyshev	49.2	36.6	47.4	27.9	-	-
Raman spectra of solid car paints						
	$C_{llr,exp}$ [%]	$C_{llr,cal}$ [%]	$C_{llr,exp}$ [%]	$C_{llr,cal}$ [%]	$C_{llr,exp}$ [%]	$C_{llr,cal}$ [%]
	univariate		bivariate		trivariate	
Manhattan	5.8	0.0	0.7	0.0	0.1	0.0
Euclidean	5.9	0.0	0.7	0.0	0.09	0.0
squared Euclidean	4.7	0.0	0.4	0.0	0.05	0.0
correlation-based	4.9	0.0	0.5	0.0	0.05	0.0
Chebyshev	8.4	0.0	1.4	0.0	0.3	0.0
Raman spectra of metallic car paints						
	$C_{llr,exp}$ [%]	$C_{llr,cal}$ [%]	$C_{llr,exp}$ [%]	$C_{llr,cal}$ [%]	$C_{llr,exp}$ [%]	$C_{llr,cal}$ [%]
	univariate		bivariate		trivariate	
Manhattan	6.7	0.0	1.1	0.0	0.02	0.0
Euclidean	5.9	0.0	0.8	0.0	0.1	0.0
squared Euclidean	5.6	0.0	0.6	0.0	0.1	0.0
correlation-based	5.6	0.0	0.7	0.0	0.1	0.0
Chebyshev	7.6	0.0	1.4	0.0	0.3	0.0

values ($C_{llr,cal}$, Figure 5b), both referring to $log_{10}Odds(H_1) = 0$ (Figure 5b). Each value corresponds to the ECE results from the median set of LR values (regarded as a kind of a representative LR set) for each comparison obtained from all $s = 100$ sets of LR values yielded for particular LR model. Table 1 clearly illustrates that the LR models for FTIR spectra are most effective when squared Euclidean or correlation-based distance metrics are used with the lowest $C_{llr,exp}$ and $C_{llr,cal}$ values. For Raman spectra databases it is much more difficult to clearly indicate any distance metric yielding the most promising results, since all of them prove that the performance of the proposed LR models is very good. Furthermore it is observed that the $C_{llr,exp}$ values decrease with the dimensionality of the LR models, indicating better performance and calibration of the LR models.

Significant differences between the experimental ($C_{llr,exp}$) and calibrated ($C_{llr,cal}$) ECE curves for FTIR spectra database (Table 1) point out great opportunities for improving the models performance. An improvement may involve expanding the database to allow better capturing of all the relevant features characteristic for the whole population of the analysed samples.

An additional merit of the proposed methodology is that, despite the very small database, the results obtained are satisfactory. Although the small size of the database necessitates caution when drawing conclusions and when making generalisations for legal processing, it is still more advantageous to apply the LR approach, even using a small database, than to perform subjective visual comparison of spectra, where the database features are completely ignored.

4. Conclusions

The presented studies were focused mainly on verifying the applicability of LR approach for commenting on the evidential value of highly multidimensional data for polypropylene and car paints in the form of FTIR and Raman spectra. From the results already seen, chemometric tools seem to provide noteworthy solutions for effective data dimensionality reduction, which is indispensable prior

to LR calculations. This combination of techniques addresses various aspects of multidimensional data analysis and provides the compact solution to the stated comparison problem of spectra. It is especially important as chemometric tools outcomes cannot be directly interpreted for forensic purposes and their application has to be followed by the LR approach. Therefore, combining the chemometric tools results with the LR approach has been gaining importance in recent years [15, 14] due to the more complex data structures delivered by advanced analytical equipment.

Comprehending the results of LDA for distance representation of spectra for solving their comparison problem using LR approach enables for capturing the relevant between-object variability, which is usually lost when only for example PCA is applied. Moreover, converting the space from feature to distance representation and using LDA leads to the most effective spectra compression by extracting only the information corresponding with characteristic objects features and neglecting the irrelevant noise or negligible variability.

From the results shown, it is evident that practically only the first few LDs will play an important role and constitute credible variables for solving the comparison problem with the application of *naïve* LR approach. The findings confirm that it is always a matter of finding a compromise between the LR model complexity (the number of variables it accounts for) and the magnitude of the misleading results it yields.

With the diversity of analysed microtraces growing rapidly, researchers are often faced with limited databases due to the time and resources required for data collection. The results presented here show that the proposed methodology works even for such small databases, giving reliable conclusions in the comparison problem. Caution is needed, however, when generalizing these conclusions, as small databases may underrepresent the relevant population.

As was demonstrated in this paper, the likelihood ratio approach is capable of solving the comparison problem of highly multivariate and correlated data after proper extraction of the most relevant variance information hidden in the data structure. On the basis of the promising findings presented, work on the

remaining issues is being continued and will be presented in future publications.

5. Acknowledgements

The experiments were undertaken within the National Science Centre in Poland (Preludium 6 no. 2013/11/N/ST4/01547) and the Institute of Forensic Research projects VI K/2013-15 and IV K/2015-17.

References

[1] ENFSI . ENFSI Guideline for Evaluative Reporting in Forensic Science. Strengthening the Evaluation of Forensic Results across Europe. European Network of Forensic Science Institutes; 2015.

[2] Aitken C, Taroni F. Statistics and the Evaluation of Evidence for Forensic Scientists. Chichester: Wiley; 2004.

[3] Aitken C, Zadora G, Lucy D. A two-level model for evidence evaluation. Journal of Forensic Sciences 2007;52(3):412–9.

[4] Zadora G, Martyna A, Ramos D, Aitken C. Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data. Wiley; 2014.

[5] Meuwly D. Forensic individualisation from biometric data. Science and Justice 2006;46:205213.

[6] Hepler A, Saunders C, Davis L, Buscaglia J. Score-based likelihood ratios for handwriting evidence. Forensic Science International 2012;219:129140.

[7] Gonzalez-Rodriguez G, Drygajlo A, Rasmos-Castro D, Garcia-Gomar M, Ortega-Garcia J. Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Computer peech and Language 2006;20:331355.

[8] Michalska A, Martyna A, Zieba-Palus J, Zadora G. Application of a likelihood ratio approach in solving a comparison problem of Raman spectra recorded for blue automotive paints. Journal of Raman Spectroscopy 2015;46:772–83.

[9] Ramos D, Zadora G. Information-theoretical feature selection using data obtained by scanning electron microscopy coupled with an energy dispersive x-ray spectrometer for the classification of glass traces. Analytica Chimica Acta 2011;705:207–17.

[10] Zadora G, Ramos D. Evaluation of glass samples for forensic purposes – an application of likelihood ratios and an information-theoretical approach. Chemometrics and Intelligent Laboratory Systems 2010;102:63–83.

[11] Zieba-Palus J, Zadora G, Milczarek J. Differentiation and evaluation of evidence value of styrene acrylic urethane topcoat car paints analysed by pyrolysis-gas chromatography. Journal of Chromatography A 2008;1179:47–58.

[12] Dahlhaus R. Graphical interaction models for multivariate time series. Metrika 2000;51:157–72.

[13] Allen R, Mills D. Signal analysis. Time, frequency, scale and structure. Wiley; 2004.

[14] Wlasiuk P, Martyna A, Zadora G. A likelihood ratio model for the determination of the geographical origin of olive oil. Analytica Chimica Acta 2015;853:187–99.

[15] Martyna A, Zadora G, Stanimirova I, Ramos D. Wine authenticity verification as a forensic problem: An application of likelihood ratio test to label verification. Food Chemistry 2014;154:287–95.

[16] Wiklund S, Johansson E, Sjostrom M, Mellerowicz E, Edlund U, Shockcor J, et al. Visualization of gc/tof-ms-based metabolomics data for identi-

fication of biochemically interesting compounds using opls class models. Analytical Chemistry 2008;80:115–22.

[17] Ahlinder J, Nordgaard A, Lindstrm S. Chemometrics comes to court: evidence evaluation of chembio threat agent attacks. Journal of Chemometrics 2015;29:267276.

[18] Bolck A, Ni H, Lopatka M. Evaluating score-and feature-based likelihood ratio models for multivariate continuous data: Applied to forensic mdma comparison. Law, Probability and Risk 2015;14:243–66.

[19] Muehlethaler C, Massonnet G, Esseiva P. The application of chemometrics on infrared and raman spectra as a tool for the forensic analysis of paints. Forensic Science International 2011;209:173–82.

[20] Muehlethaler C, Massonnet G, Esseiva P. Discrimination and classification of ftir spectra of red, blue and green spray paints using a multivariate statistical approach. Forensic Science International 2014;244:170–8.

[21] Martyna A, Michalska A, Zadora G. 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. Analytical and Bioanalytical Chemistry 2015;407:3357–76.

[22] Daubechies I. Ten Lectures on Wavelets. Philadelphia: CBMS-NSF Regional Conference Series in Applied Mathematics; 1992.

[23] Walczak B. Wavelets in Chemistry. (eds). Elsevier; 2000.

[24] Alsberg B, Woodward A, Kell D. An introduction to wavelet transforms for chemometricians: A time-frequency approach. Chemometrics and Intelligent Laboratory Systems 1997;37:215–39.

[25] Tauler R, Walczak B, Brown S. Comprehensive Chemometrics. Elsevier; 2009.

[26] Hazewinkel eM, Subbotin Y. Encyclopedia of Mathematics. Springer; 2001.

[27] Wee E, Grayden D, Zhu Y, Petkovic-Duran K, Smith D. A continuous wavelet transform algorithm for peak detection. Electrophoresis 2008;29:4215–25.

[28] Zerzucha P, Walczak B. Concept of (dis)similarity in data analysis. Trends in Analytical Chemistry 2012;38:116–28.

[29] Porro D, Duin R, Talavera I, Hernandez N. Alternative representations of sepctral data for classification. Proc ASCI 2009, 15th Annual Conf of the Advanced School for Computing and Imaging (Zeewolde, June 3-5, 2009) 2009;.

[30] Porro-Munoz D, Talavera I, Duin R, Hernandez N, Orozco-Alzate M. Dissimilarity representation on functional spectral data for classification. Journal of Chemometrics 2011;25:476486.

[31] Bouveyron J. Probabilistic model-based discriminant analysis and clustering methods in chemometrics. Journal of Chemometrics 2013;27:433–46.

[32] Varmuza K, P. F. Introduction to Multivariate Statistical Analysis in Chemometrics. Wiley; 2009.

[33] Raftery A, Dean N. Variable selection for model-based clustering. Journal of the American Statistical Association 2006;101:168–78.

[34] Murphy T, Raftery A, Dean N. Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. Annals of Applied Statistics 2012;6:396–421.

[35] Silverman B. Density estimation for statistics and data analysis. London, UK: Chapman and Hall; 1986.

[36] Aitken C, Lucy D. Evaluation of trace evidence in the form of multivariate data. Applied Statistics 2004;53:109–22.

[37] Brümmer N, du Preez J. Application independent evaluation of speaker detection. Computer Speech and Language 2006;20:230–75.

[38] Ramos D, Gonzalez-Rodriguez J, Zadora G, Aitken C. Information-theoretical assessment of the performance of likelihood ratio computation methods. Journal of Forensic Sciences 2013;58:1503–18.

[39] Ayer M, Brunk H, Ewing G, Reid W, Silverman E. An empirical distribution function for sampling with incomplete information. The Annals of Mathematical Statistics 1955;26:641–7.

[40] Best M, Chakravarti N. Active set algorithms for isotonic regression; a unifying framework. Math Program 1990;47:425–39.

[41] R Core Team . R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria; 2012. URL: http://www.R-project.org/; ISBN 3-900051-07-0.

[42] Ripley B, Venables B, Bates D, Hornik K, Gebhardt A, Firth D. Package MASS; 2015. URL: https://cran.r-project.org/web/packages/MASS/MASS.pdf.

[43] Fraley C, Raftery A, Scrucca L, Murphy T, Fop M. Package mclust; 2015. URL: https://cran.r-project.org/web/packages/mclust/mclust.pdf.