Analytica Chimica Acta 931 (2016) 34–46

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

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

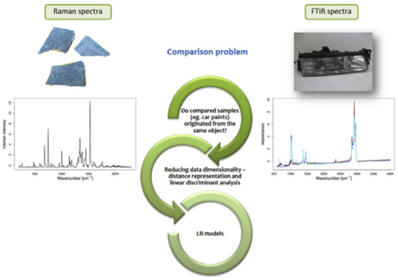
Agnieszka Martyna [a, b], Grzegorz Zadora [b, c, *], Tereza Neocleous [d], Aleksandra Michalska [b], Nema Dean [d]

[a] Jagiellonian University in Krakow, Faculty of Chemistry, Department of Analytical Chemistry, 3 Ingardena, Krakow 30-060, Poland
[b] Institute of Forensic Research in Krakow, 9 Westerplatte, Krakow 31-033, Poland
[c] University of Silesia in Katowice, Institute of Chemistry, Chemometric Research Group, 9 Szkolna, Katowice 40-006, Poland
[d] University of Glasgow, School of Mathematics and Statistics, 15 University Gardens, Glasgow G12 8QW, United Kingdom

HIGHLIGHTS

- The comparison problem of infrared spectra of polymers and Raman spectra of car paints was investigated for forensic purposes.
- Likelihood ratio with combination of chemometric tools for data compression was applied for reporting the evidential value.
- The differences between spectra expressed in the distance representation were captured using the linear discriminat analysis.
- False positive, false negative rates and empirical cross entropy plots were used for assessing the models performance.

GRAPHICAL ABSTRACT


Raman spectra
FTIR spectra
Comparison problem
Do compared samples (eg. car paints) originated from the same object?
Reducing data dimensionality – distance representation and linear discriminant analysis
LR models
Raman intensity
Wavenumber [cm⁻¹]
Absorbance
Wavenumber [cm⁻¹]

ARTICLE INFO

Article history:
Received 17 December 2015
Received in revised form
25 March 2016
Accepted 17 May 2016
Available online 24 May 2016

Keywords:
Dimension reduction
Likelihood ratio
Infrared and Raman spectroscopy
Polymer
Car paint
Forensic science

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 multivariate 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-sample variability while maximising the between-sample variability. Both techniques enabled substantial reduction of data dimensionality. Univariate and multivariate likelihood ratio models

[*] Corresponding author. Institute of Forensic Research in Krakow, 9 Westerplatte, Krakow 31-033, Poland.

E-mail addresses: rzepecka@chemia.uj.edu.pl (A. Martyna), gzadora@ies.krakow.pl (G. Zadora), tereza.neocleous@glasgow.ac.uk (T. Neocleous), amichalska@ies.krakow.pl (A. Michalska), nema.dean@glasgow.ac.uk (N. Dean).

http://dx.doi.org/10.1016/j.aca.2016.05.016
0003-2670/© 2016 Elsevier B.V. All rights reserved.

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.

© 2016 Elsevier B.V. All rights reserved.

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 utilized for pigment 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 [2–4].

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:

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

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-object (sample) 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 [5–7]) 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

Download English Version:

https://daneshyari.com/en/article/1162718

Download Persian Version:

https://daneshyari.com/article/1162718

Daneshyari.com