Analytica Chimica Acta 1000 (2018) 123–131

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Exploring hyperspectral imaging data sets with topological data analysis

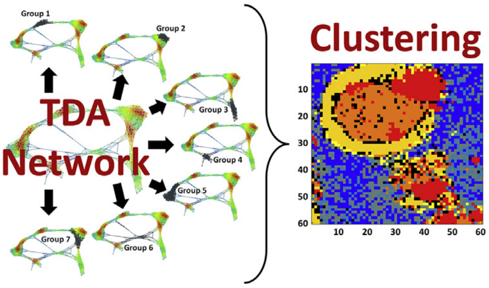
Ludovic Duponchel

LASIR CNRS UMR 8516, Université Lille 1, Sciences et Technologies, 59655 Villeneuve d'Ascq Cedex, France

HIGHLIGHTS

- First use of topological data analysis in spectroscopic imaging.
- Detection of minor compounds in a multiphase chemical system.
- TDA: a new paradigm for cluster analysis in the framework of imaging.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 2 October 2017
Received in revised form
16 November 2017
Accepted 17 November 2017
Available online 22 November 2017

Keywords:
Hyperspectral imaging
Spectroscopic imaging
Topological data analysis
Clustering
Raman spectroscopy

ABSTRACT

Analytical chemistry is rapidly changing. Indeed we acquire always more data in order to go ever further in the exploration of complex samples. Hyperspectral imaging has not escaped this trend. It quickly became a tool of choice for molecular characterisation of complex samples in many scientific domains. The main reason is that it simultaneously provides spectral and spatial information. As a result, chemometrics has provided many exploration tools (PCA, clustering, MCR-ALS …) well-suited for such data structure at early stage. However we are today facing a new challenge considering the always increasing number of pixels in the data cubes we have to manage. The idea is therefore to introduce a new paradigm of Topological Data Analysis in order explore hyperspectral imaging data sets highlighting its nice properties and specific features. With this paper, we shall also point out the fact that conventional chemometric methods are often based on variance analysis or simply impose a data model which implicitly defines the geometry of the data set. Thus we will show that it is not always appropriate in the framework of hyperspectral imaging data sets exploration.

1. Introduction

Hyperspectral imaging is today a choice tool for characterising complex samples in different scientific domains. It is obvious that instrumental developments have first contributed to this potential but multivariate data analysis tools have really revealed its potential. Many chemometric algorithms have been developed in order to explore hyperspectral data cubes without *a priori* such as Principal Component Analysis (PCA), clustering techniques, Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) but we must admit that we are always more hampered by the increasing number of pixels (i.e. spectra) in our data sets. Although these tools allow us to extract valuable information about major compounds, it is still difficult to extract information about minor ones, all the more so due to the bad signal to noise ratio we often

E-mail address: ludovic.duponchel@univ-lille1.fr.

https://doi.org/10.1016/j.aca.2017.11.029
0003-2670/© 2017 Elsevier B.V. All rights reserved.

observe in molecular spectroscopy. It is in this sense that we need to explore new paradigms such as topological data analysis.

Mathematicians usually use topology in order to study shape of abstract objects. Nevertheless they have discovered a short time ago that it could be used for the exploration of real-world data sets and topological data analysis (TDA) was born [1–4]. Since then, many papers have demonstrated the great potential of the concept. The first domain taking advantage of the method is certainly biology at large. We thus find many papers in genomics [5] with, for example, the development of a new visualisation tool for expression Quantitative Trait Locus (eQTL) [6], analysis of immune response [7,8], viral evolution [9], antibiotic resistance [10], bacteria [11], protein interaction networks [12,13], intestinal development in preterm and term infants [14], DNA repairing [15], evolutionary processes [16,17] and cellular development [18]. TDA is also used for the study of different diseases such as pulmonary embolism [19], diabetes [20,21], autism [22], asthma [23–25], infections [26,27], cardiac septal defects and cardiomyopathy [28], ovarian cancer [29], precision oncology [30], knee osteoarthritis [31], oral squamous cell Carcinomas [32] and chronic fatigue syndrome [33]. It is also used in neuroscience for the study of the activity in the visual cortex [34], the analysis of traumatic brain injury [35,36] and the identification of neuroimaging biomarkers for patients with serious mental illness [37]. The first paper in analytical chemistry is dedicated to the analysis of more than 20.000 samples in order to reveal pedogenetic principles of European topsoil system [38]. This article demonstrates the good scalability of the approach able to explore and summarize quite big data sets. Another very interesting paper is focused on finding the best nanoporous materials for gas storage [39]. It was not until 2016 that a first TDA paper is published in physical-chemistry and, more specifically, in molecular spectroscopy [40]. This article is dedicated to Raman analysis of single bacteria. It is then demonstrated that TDA is able to classify spectra of different bacteria strains considering different experimental conditions wherever classical clustering methods fail. We come to realize that TDA seems to have nice properties for the analysis of spectroscopic data sets. With this in mind, proposing an article dedicated to the development of TDA for the analysis of hyperspectral data cubes is a natural extension. The first part of the paper will introduce the TDA concept in the framework of hyperspectral imaging and provide details about the data set used in this study. A second part will give details about TDA network construction and results concerning clustering with TDA compared with Kmeans algorithm.

2. Materials and methods

2.1. Image data set

The paper of Andrew and al [41]. is the origin of the hyperspectral data set we use in this study. An oil in-water emulsion system is then explored with Raman spectroscopy. The complexity of this multicomponent/multiphase system explains the fact that we have selected it. Moreover it has been already explored in different chemometric papers which is very interesting for comparison purpose [42,43]. The hyperspectral data set contains 3600 preprocessed spectra (i.e. a data cube of 60 pixels by 60 pixels) with a spectral range from 950 cm^{-1} to $1800^{-1}$and thus 253 spectral variables. Spectral preprocessing has been used prior data analysis in order to suppress fluorescence effects.

2.2. Topological data analysis of a hyperspectral data cube

The main aim of TDA is first to generate a topological network which represents the intrinsic shape of the explored data set. Fig. 1 presents its use in the framework of the analysis of a hyperspectral data cube. Because, like most of the chemometric methods, TDA is not suited for a direct analysis of a 3D hyperspectral data cube, an unfolding procedure is used to generate a more convenient 2D matrix with the size (60 x 60, 253) considering the selected Raman data set. The TDA is then decomposed in 8 different steps: (1) given the unfolded data set, we observe each row (i.e. spectrum) through a lens. In fact, all functions that produce a number from a spectrum can be a lens. It may originate from different domains such as statistics (min, max, mean, variance, density …), geometry (centrality, curvature …) and chemometrics (PCA scores, Support Vector Machine (SVM) distance from hyperplane …) without being exhaustive. At the end of this step, we have a lens value per spectrum of the data set and, by extension, a lens value scale. (2) Then we divide the lens scale into overlapping subsets. We will observe the impact of the number of subsets and the percentage of overlap in the 'Results and discussion' section of the paper. (3) Next we then use scale subsets to partition the data set. Because of overlaps, it is possible to retrieve simultaneously a spectrum in different pixel subsets. (4) In this step we consider each pixels subset separately. Indeed we apply a cluster analysis on each. In general single linkage algorithm [44] is used but other techniques can be implemented. At this point starts the construction of the topological network. Each cluster in each analysis is then represented by a node. (5) We connect nodes with edges when corresponding clusters have at least one spectrum (i.e. pixel) in common. (6) Nodes are colored depending on the number of spectra they contain. (7) The network is split in different subparts or groups of pixels considering variations of nodes density and/or particular features of the network shape. We generate here different classes of pixels. (8) In the last step, a clustering map is generated considering the coordinates of each pixel in the sample plane and its class membership.

In this article, Topological Data Analysis was performed with the Ayasdi software platform (ayasdi.com, Ayasdi Inc., Menlo Park CA). An in-house Python script has been developed in connection with the Ayasdi Python SDK in order to generate the clustering map.

2.3. Kmeans clustering

Many clustering methods have been used for the exploration of hyperspectral data cubes. However this would not make any sense to compare TDA results with every possible approach. We have therefore decided to select K-Means [45,46] (KM) clustering because it is one of the most popular unsupervised classification methods. Very briefly, the goal of KM is to separate a set of n unlabelled data points (defined in a d-dimensional space) into k clusters. Each cluster is represented by its barycentre called centroid. As a first step, the algorithm selects at random k initial points as centroids in the d-dimensional space. We have then an iterative process defined in two steps: (i) given a chosen distance, each point of the data set is assigned to the nearest centroid; (ii) Considering this new partition, centroids are updated for each cluster. These two steps are repeated until convergence, that is to say when no more changes are observed in the assignment of all data points. In this study, the point-to-centroid distance has been calculated with the Euclidean one. Because partitions from KM algorithm are known to be very sensitive to the initialisation step, partitioning has been replicated 100 times. The idea here was to repeat clustering using each time new initial cluster centroid positions. The partition having the lowest within-cluster sum of point-to-centroid distances for all the data points is considered as the optimal one. All KM calculations in this paper have been developed with MATLAB environment version R2016a (The MathWorks Inc., Natick, MA, USA) and the Statistics and Machine Learning Toolbox version 10.2 (The MathWorks Inc., Natick, MA,