engineering proceedings

MDPI

Proceeding Paper

Preprocessing and Analyzing Raman Spectra Using Python [†]

Eleftherios Pavlou and Nikolaos Kourkoumelis *

Laboratory of Medical Physics, University of Ioannina, 451 10 Ioannina, Greece; nkourkou@uoi.gr
* Correspondence: nkourkou@uoi.gr
[†] Presented at the 4th International Electronic Conference on Applied Sciences, 27 October–10 November 2023; Available online: https://asec2023.sciforum.net/.

Abstract: The inherent complexity of the Raman spectra of biomedical samples reflects the intricate molecular composition and intermolecular interactions of these diverse systems. Unraveling the complexities of biological Raman spectra is essential for bioscience and bioengineering research because it provides insight into cellular processes, disease states, and drug interactions. For the effective analysis of such complex data, robust and cutting-edge software is required that provides sophisticated algorithms for data preprocessing, thereby enhancing signal-to-noise ratio and revealing hidden spectral information. In addition, novel applications of this type may include machine learning algorithms for automated clustering analysis, enabling the identification of biomolecules and their conformational changes in diverse biological specimens. We present a Python 3 package built around popular scientific Python libraries that aims to provide Raman spectroscopists with user-friendly programming tools for the analysis of complex biomedical Raman data.

Keywords: Raman; spectroscopy; Python; preprocessing; analysis; software

Citation: Pavlou, E.; Kourkoumelis, N. Preprocessing and Analyzing Raman Spectra Using Python. *Eng. Proc.* **2023**, *52*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: date

1. Introduction

Raman spectroscopy is a type of vibrational spectroscopy that relies on the inelastic scattering of light (Raman scattering) upon its interaction with the vibrational modes of a Raman-active molecule. It is a technique that is widely used because it allows for the non-destructive study and molecular characterization of both organic and inorganic materials, in either solid or liquid state or in solutions, with minimal to no preparation prior to measurement [1]. The ability to perform rapid measurements and the low interference from water molecules make it an excellent technique for studying wet tissues and also suitable for in vivo measurements [2]. Raman spectra contain characteristic vibrational information that can be used for the identification and quantification of compounds present in a sample, as well as for the determination of its chemical composition. This information provides, in essence, the molecular fingerprint of the substance. Due to its versatility, ease of use, and ability to provide both qualitative and quantitative results, Raman spectroscopy has evolved to a valuable analytical tool in biomedical applications.

Raman scattering is a weak phenomenon, with only a fraction of the photons incident to a substance undergoing Raman scattering. This results in the signal being weak and hard to distinguish from background noise, which can occur due to the instruments and detectors used, the environment, and sample impurities among other factors [3]. Additionally, Raman spectra are influenced by fluorescent molecules that are present in a sample. This influence has the form of background signal, which can be stronger than the Raman signal and overlaps with it, obscuring and deforming the Raman peaks [4]. Both factors can lead to spectra with low signal-to-noise ratio (SNR).

Addressing the issue of fluorescence background in biological samples may require several strategies. These include selecting an appropriate excitation wavelength, using the anti-Stokes segment of a Raman spectrum, and utilizing techniques such as photobleaching, in which the sample is subjected to prolonged irradiation. However, it is worth noting that the latter approach is not commonly employed with biological samples mostly due to the fact that photobleaching may harm the samples due to the high radiation intensity required and the extended duration of irradiation [5]. As a standard practice, computational techniques are employed to denoise and eliminate the fluorescence background from Raman spectra. This procedure, commonly referred to as "preprocessing", is undertaken with the objective of enhancing SNR before engaging in further analysis [6].

Here we present a Python 3 package built around some of the most popular scientific Python libraries that aims to provide Raman spectroscopists with user-friendly programming tools for the preprocessing and analysis of complex biomedical Raman data. To briefly demonstrate the package's usage, we will use Raman spectra of bone, collected from the tibias of healthy and osteoporotic rabbits.

2. Materials and Methods

2.1. Python Package Overview

Written for Python 3.7 and later versions, the code allows the user to preprocess Raman spectra and deconvolute complex Raman bands, as well as apply Principal Components Analysis (PCA) and Partial Least Squares Regression (PLSR) to Raman data. The package is primarily developed to be run in Jupyter Notebooks and depends on pandas 1.0+ [7], matplotlib 3.0+ [8], NumPy 1.19+ [9], seaborn 0.11+ [10], SciPy 1.5.0+ [11], scikit-learn 0.23+ [12], and Rohan Isaac's (rohanisaac) spc module [13].

The core component of the code is the Pandas dataframe. Dataframes are versatile data structures that provide methods for reading from and writing to various file types, as well as a significant number of methods that allow for advanced data manipulation and visual representation. Although they may lack in terms of processing speed and memory efficiency when compared to NumPy arrays, pandas dataframes in combination with Jupyter Notebooks offer great data inspection capabilities and interactivity, which is of utmost importance when performing exploratory data analysis, as usually required in Raman spectroscopy. They also allow for easily performing batch-processing actions on spectra, which is essential when handling large amounts of data, such the ones usually obtained from Raman experiments.

The package contains methods for file operations on Raman data and for preprocessing spectra. The preprocessing functionalities contain methods for despiking and smoothing spectra, interpolating, differentiating, background subtraction using the SNIP algorithm [14], and various normalization options. All preprocessing methods are available for use on either NumPy arrays or pandas dataframes. Additionally, classes for clustering analysis and modeling have been implemented. More specifically, the package includes a Principal Components Analysis (PCA) class built around the decomposition.PCA scikit-learn class and a Partial Least Squares (PLS) class built around the cross_decomposition.PLSRegression scikit-learn class. Both classes can also be used for dimensionality reduction and provide methods that facilitate the creation of publication-ready visualizations. The PLS class can also be used both for regression (PLSR) and two-class discriminant analysis (PLS-DA). Finally, a peak deconvolution module is included that allows for fitting complex Raman bands with Gaussian or Lorentzian functions, allowing for the extraction of additional information from Raman spectra.

2.2. Samples

Samples were obtained from the left tibias of 5 healthy and both the right and left tibias of 2 osteoporotic, 8 months old, female New Zealand rabbits. Inflammation-mediated osteoporosis was induced to the osteoporotic rabbits by following the method described by Kourkoumelis et al. [15]. Six slices were obtained from the diaphyses (mainly consisting of cortical bone) of each tibia, symmetrically towards the proximal and distal epiphyses. Raman spectra from three different points of the transverse surface of each slice, separated approximately by 120°, were obtained using a BWTEK i-Raman Plus spectrometer, operating at 785 nm, with a power output of 200 mW at the probe and signal collection time of 6 s. In total, 36 healthy and 36 osteoporotic Raman spectra were collected.

3. Discussion

The required preprocessing steps and the subsequent application of PCA will be briefly described. Both healthy and osteoporotic rabbit spectra were combined in a single dataframe with the Raman shift as the dataframe's index and the sample names as the dataframe's columns. The spectra were subsequently cropped to the 380–1800 cm^{-1} region (fingerprint region), treated to remove spikes that may have occurred mainly due to cosmic rays and detector artifacts, and smoothed using a Savitzky-Golay filter. The fluorescence background of each spectrum was then calculated using the SNIP algorithm and each calculated background was subtracted from the respective spectrum. Normalization of each spectrum to the maximum intensity of the respective phosphate (v_1 PO$_4^{-3}$) Raman peak between 955 cm^{-1} and 965 cm^{-1} [16], concludes the preprocessing procedure, leading to spectra with good SNR and most of the fluorescence background removed. Preprocessing is a crucial step in Raman analysis and the quality of the subsequent results strongly depends on it.

The programmed PCA class was then employed as a technique for the discrimination of the two bone classes (healthy and osteoporotic). The result of PCA is displayed in a summarizing plot that contains a scree plot, the loadings plots for the first three principal components (PCs) and a 3 by 3 plot containing PC scores plots for the non-diagonal elements and kernel density estimate (KDE) plots for the diagonal elements (Figure 1). The PC scores plots also include the 95% confidence ellipses of each sample class. The scree plot indicates that the first three PCs explain most of the observed variance of the data (77.91%). Adding more than three PCs does not represent a significant contribution to the total variance. The PC1-PC2 and PC2-PC3 scores plots show clear discrimination of the healthy and osteoporotic samples along the PC2 axis. This is also especially obvious in the KDE plots, where the PC2 KDE distributions for the healthy and osteoporotic samples are clearly discriminated, while the KDE plots for PC1 and PC3 overlap heavily.

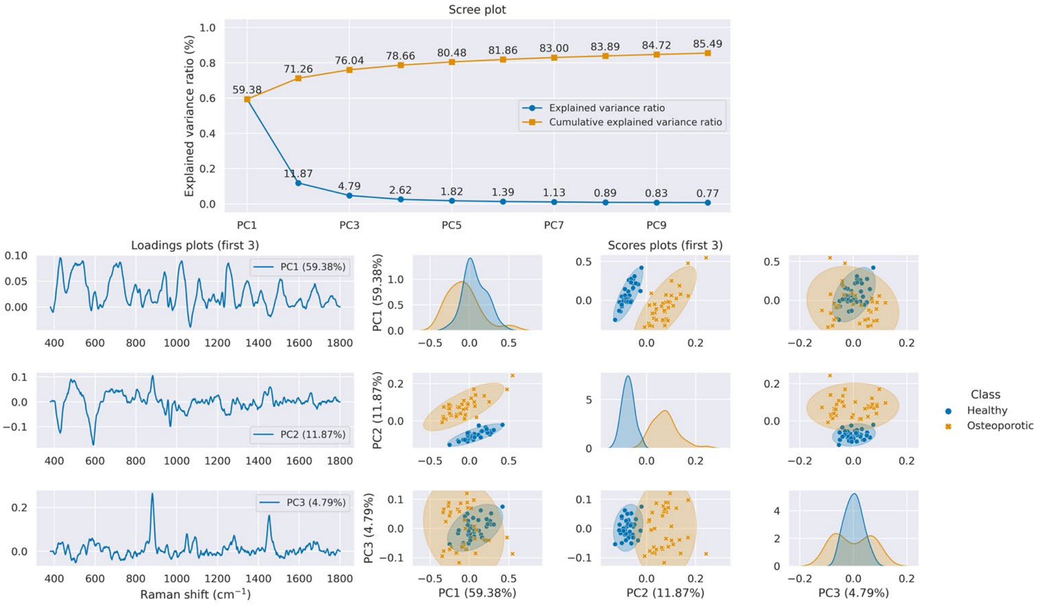

Figure 1. PCA summary plot containing a scree plot and the first three PC loadings scores plots. The shaded ellipses represent the 95% confidence ellipses of the classes, colored with their respective colors. The diagonal elements of the scores plots are the kernel density estimate (KDE) plots of the respective PC.

4. Conclusions

In this paper we presented a Python 3 package that utilizes popular scientific Python libraries with the goal of providing Raman scientists a user-friendly but mostly versatile and expandable programming tool for preprocessing and analyzing Raman data. Using Raman spectra of healthy and osteoporotic rabbit bones, we briefly described the basic functionality of the package and showed how it can be used to apply principal components analysis under a concise scheme of relevant scores and loadings plots.

Author Contributions: Conceptualization, E.P. and N.K.; methodology, E.P and N.K.; software, E.P.; validation, E.P. and N.K.; formal analysis, E.P.; investigation, E.P.; resources, N.K.; data curation, E.P.; writing—original draft preparation, E.P.; writing—review and editing, N.K.; visualization, E.P.; supervision, N.K.; project administration, N.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of the University of Ioannina.

Informed Consent Statement:

Data Availability Statement: Raman data are available upon reasonable request.

Acknowledgments: This research is co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the projects: "Strengthening Human Resources Research Potential via Doctorate Research" (MIS-5000432) and "Reinforcement of Postdoctoral Researchers-2nd Cycle" (MIS-5033021), both implemented by the State Scholarships Foundation (IKY).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Das, R.S.; Agrawal, Y.K. Raman Spectroscopy: Recent Advancements, Techniques and Applications. *Vib. Spectrosc.* **2011**, *57*, 163–176. https://doi.org/10.1016/j.vibspec.2011.08.003.
2. Cordero, E.; Latka, I.; Matthäus, C.; Schie, I.W.; Popp, J. In-Vivo Raman Spectroscopy: From Basics to Applications. *JBO* **2018**, *23*, 071210. https://doi.org/10.1117/1.JBO.23.7.071210.
3. Smulko, J.; Wróbel, M.S.; Barman, I. Noise in Biological Raman Spectroscopy. In Proceedings of the 2015 International Conference on Noise and Fluctuations (ICNF), Xi'an, China, 2–6 June 2015; pp. 1–6.
4. Kostamovaara, J.; Tenhunen, J.; Kögler, M.; Nissinen, I.; Nissinen, J.; Keränen, P. Fluorescence Suppression in Raman Spectroscopy Using a Time-Gated CMOS SPAD. *Opt. Express* **2013**, *21*, 31632. https://doi.org/10.1364/OE.21.031632.
5. Zięba-Palus, J.; Michalska, A. Photobleaching as a Useful Technique in Reducing of Fluorescence in Raman Spectra of Blue Automobile Paint Samples. *Vib. Spectrosc.* **2014**, *74*, 6–12. https://doi.org/10.1016/j.vibspec.2014.06.007.
6. Bocklitz, T.; Walter, A.; Hartmann, K.; Rösch, P.; Popp, J. How to Pre-Process Raman Spectra for Reliable and Stable Models? *Anal. Chim. Acta* **2011**, *704*, 47–56. https://doi.org/10.1016/j.aca.2011.06.043.
7. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3July 2010; pp. 56–61.
8. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.
9. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array Programming with NumPy. *Nature* **2020**, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.
10. Waskom, M.; Botvinnik, O.; Gelbart, M.; Ostblom, J.; Hobson, P.; Lukauskas, S.; Gemperline, D.C.; Augspurger, T.; Halchenko, Y.; Warmenhoven, J.; et al. *Mwaskom/Seaborn: V0.11.0 (Sepetmber 2020);* Zenodo: 2020.
11. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.
12. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
13. Rohan, I. Spc. Available online: https://github.com/rohanisaac/spc/ (accessed on 12 September 2023).
14. Morháč, M.; Kliman, J.; Matoušek, V.; Veselský, M.; Turzo, I. Background Elimination Methods for Multidimensional Coincidence γ-Ray Spectra. *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers Detect. Assoc. Equip.* **1997**, *401*, 113–132. https://doi.org/10.1016/S0168-9002(97)01023-1.

15. Kourkoumelis, N.; Lani, A.; Tzaphlidou, M. Infrared Spectroscopic Assessment of the Inflammation-Mediated Osteoporosis (IMO) Model Applied to Rabbit Bone. *J. Biol. Phys.* **2012**, *38*, 623–635. https://doi.org/10.1007/s10867-012-9276-6.
16. Khalid, M.; Bora, T.; Alghaithi, A.; Thukral, S.; Dutta, J. Raman Spectroscopy Detects Changes in Bone Mineral Quality and Collagen Cross-Linkage in Staphylococcus Infected Human Bone. *Sci. Rep.* **2018**, *8*, 9417. https://doi.org/10.1038/s41598-018-27752-z.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.