

Noise Reduction in Nano-Raman Spectroscopy Using Principal Component Analysis

Jane Elisa Guimarães, Rafael Nadas, Wenjin Zhang, Takahiko Endo, Kenji Watanabe, Takashi Taniguchi, Riichiro Saito, Yasumitsu Miyata, and Ado Jorio*

Tip-enhanced Raman spectroscopy (TERS) combined with principal component analysis (PCA) offers a robust approach for enhancing signal quality and uncovering spectroscopic features otherwise concealed by noise. This study demonstrates that integrating TERS with PCA in large-scale datasets effectively reduces noise and enhances the extraction of weak Raman signals that are often obscured by random spectral fluctuations. The methodology was applied to hyperspectral datasets acquired from MoSe₂ monolayers exhibiting nanoscale surface features. Through this approach, previously hidden nano-Raman peaks were successfully isolated, enabling reliable chemical identification at the nanoscale. The combined use of TERS and PCA significantly improves sensitivity and resolution in the spectroscopic analysis of 2D materials, advancing their characterization with respect to interfacial and environmental effects.

application of techniques such as principal component analysis (PCA).^[9,10] PCA is particularly advantageous in this context, as it efficiently extracts the relevant spectral components and substantially minimizes noise contributions.^[11–13] Consequently, this approach becomes a powerful tool for highlighting the most significant spectroscopic features and interpreting the results with greater accuracy and confidence.

When applied to large datasets, the PCA method seeks to identify a set of principal components (PCs) that capture the most significant variance in the data.^[10,13,14] This process involves computing the eigenvalues and eigenvectors of the covariance matrix of the variables. Each eigenvector

defines the direction of a PC, while its corresponding eigenvalue quantifies the amount of explained variance. The components are then ranked in decreasing order of variance, allowing the selection of the most significant ones. By projecting the original data onto the new set of axes that capture the largest variation, defined by the number of PCs chosen, the dataset can be effectively represented in a lower-dimensional space that retains the essential structure and variability of the original data.

The integration of Raman spectroscopy with PCA has proven efficiency in extracting detailed chemical information from Raman mappings across a wide range of materials.^[13,15,16] This work presents a reproducible method for reducing noise in large nano-Raman hyperspectral datasets. By applying PCA


1. Introduction

When tip-enhanced Raman spectroscopy (TERS) mapping is performed, the acquired hyperspectral data, comprising hundreds or even thousands of individual spectra, enables robust statistical analysis. TERS combines Raman spectroscopy with scanning probe microscopy (SPM), using a metallic tip to enhance local electromagnetic fields and achieve nanoscale spatial resolution.^[1–4] This technique enables chemical characterization beyond the diffraction limit.^[5–8]

In this work, 4096 spectra were collected in a single TERS measurement, corresponding to a 64 × 64 pixel map from a monolayer MoSe₂ flake. This extensive dataset allows the

J. E. Guimarães, A. Jorio
Departamento de Física
Universidade Federal de Minas Gerais
Belo Horizonte 31270-901, MG, Brazil
E-mail: ado.jorio@fisica.ufmg.br

R. Nadas
Institut für Physik
Humboldt-Universität zu Berlin
Newtonstraße 15, 12489 Berlin, Germany

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/pssb.202500291>.

© 2025 The Author(s). physica status solidi (b) basic solid state physics published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/pssb.202500291

W. Zhang, T. Endo, R. Saito, Y. Miyata
Department of Physics
Tokyo Metropolitan University
Tokyo 192-0397, Japan

W. Zhang, T. Endo, T. Taniguchi, Y. Miyata
Research Center for Materials Nanoarchitectonics
National Institute for Materials Science
Tsukuba 305-0044, Japan

K. Watanabe
Research Center for Electronic and Optical Materials
National Institute for Materials Science
Tsukuba 305-0044, Japan

R. Saito
Department of Physics
Tohoku University
Sendai 980-8578, Japan

and carefully selecting the appropriate number of PCs for data reconstruction, it becomes possible to enhance the signal-to-noise ratio and reveal spectral features that were previously obscured. This approach not only facilitates the detection of weak or hidden peaks but also improves the reliability and interpretability of complex spectroscopic data.

2. Methodology

The sample studied in this work was prepared using the dry-stamping technique, where MoSe₂ grains, grown on a SiO₂/Si substrate via salt-assisted chemical vapor deposition, were retrieved with an hexagonal boron nitride flake. The resulting heterostructure was then transferred onto a glass slide using a polymer stamp.^[17,18]

Hyperspectral data were obtained by TERS in the porto laboratory prototype system, which operates in an atomic force microscope in non-contact mode using a tuning fork and Plasmon tunable tip pyramids (PTTP) probes.^[19–22] The excitation source was a He–Ne radially polarized laser, and the spectrometer used was an Andor shamrock 303i with a 600 l/mm grating.

PCA was conducted using an automated routine implemented in the portoflow analysis software. The objective was to enhance data quality by reconstructing the dataset using only the five PCs carrying the largest variational contribution, thus preserving the most relevant information within the data. The following section details the method used to select the specific number of PCs. Subsequently, the original dataset was reconstructed using the inverse transformation as a linear combination of only the five selected PCs.

Before the PCA-based noise reduction process, spikes resulting from cosmic rays or experimental artifacts were carefully removed from each individual spectrum presenting such anomalies. This step was crucial, as it prevents these anomalies from affecting the decomposition into the PCs. After applying PCA, both the full spectra and relevant spectral regions were selected using the software, and background subtraction was applied to each region using a built-in routine that sets minimum values to zero, thereby minimizing baseline contributions. This processing resulted in spectra with well-defined peaks, enabling the generation of corresponding intensity maps.

3. Results

The region analyzed was a $1 \times 1 \mu\text{m}^2$ area of monolayer MoSe₂,^[23] scanned as a 64×64 pixel map, resulting in a hyperspectral dataset comprising 4096 spectra. A representative spectrum is shown in Figure 1a. After applying the PCA-based noise reduction procedure that will be described here, the same spectrum appears as shown in Figure 1b. The signal-to-noise ratio, considering the most intense A_{1g} mode of MoSe₂, was initially 48. After applying the PCA-based noise reduction procedure, it increased to 1040, representing a signal-to-noise improvement by a factor of ≈ 21.7 .

To evaluate whether the five components used to reconstruct the data are sufficient for accurate Raman hyperspectral imaging, we can compare a Raman map built from the MoSe₂ data before and after the PCA-based noise reduction. The resulting difference between these maps, as shown in Figure 1c, consists only of noise, which confirms that the reconstructed spectra indeed keep the spectral features that truly define the map image, and only noise remains after subtraction. Therefore, the essential features of the data were effectively captured by the five selected PCs.

Although the well-known peaks of MoSe₂ were readily observed even in the as-measured data prior to PCA, the aim was to visualize features that could be associated with the nanometric structures, namely the nanoprotuberances, since the presence of noise can obscure weak peaks that might otherwise be detected, particularly those not related to the intense, well-known MoSe₂ features below 650 cm^{-1} . To address this, PCA was applied to decompose the dataset into components representing the main sources of spectral variation and potential artifacts, thereby enhancing the detection of weak TERS peaks that would otherwise remain hidden. Figure 2a shows the proportion of variance considering the first 10 PCs, from PC1 to PC10. The plot indicates that the first five components (PC1 to PC5) together account for more than 90% of the total variance, which is the rationale behind selecting these five. If the spectra is reconstructed with the inverse PCA transformation, using only these five components, most of the representative spectral features will be kept, while the random spectral variation related to noise will be removed. One way to illustrate this is to observe the spectral variance of PC2 to PC5 in PC1. Figures 2b–f present the projections

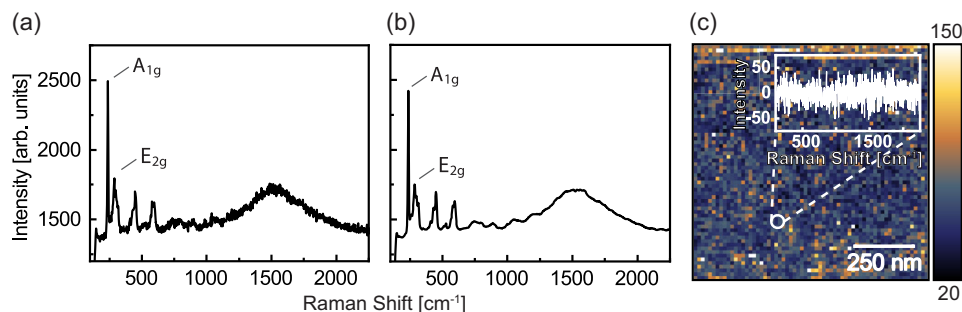


Figure 1. a) Representative as-measured spectrum from MoSe₂, indicated in (c) by the open white circle. b) The same spectrum after PCA-based noise reduction. c) Integrated Raman intensity map after PCA-based noise reduction, subtracted from the map built using the as-measured data. The inset shows a typical subtracted spectrum, where one can only see residual noise.

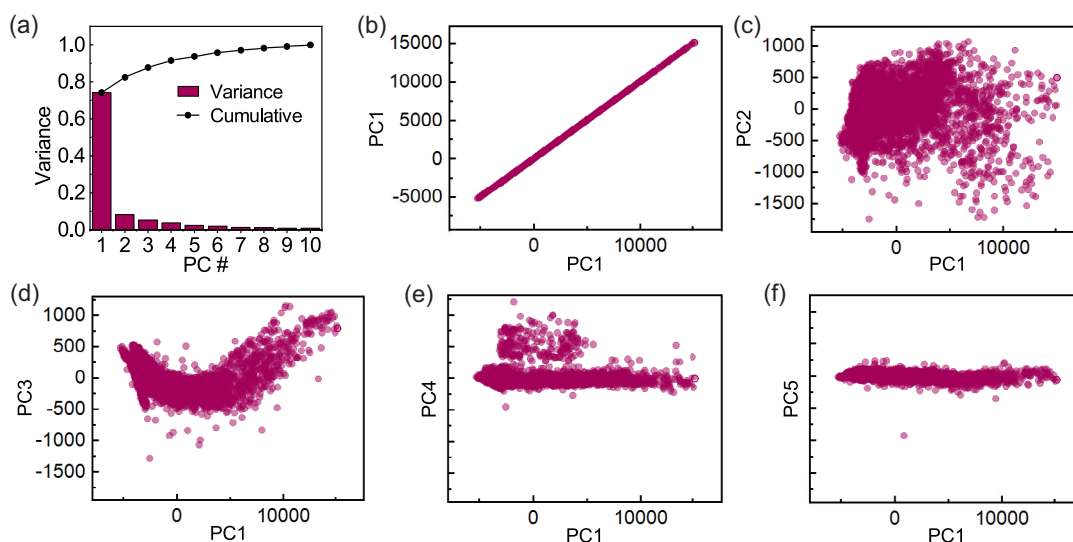


Figure 2. a) Normalized variance considering the first ten PCs obtained from the PCA decomposition of the hyperspectral TERS data, and respective cumulative variance. b–f) Projections of the dataset onto the first five PCs, with each plot representing a specific PC (PC1–PC5) against PC1. Scale bars are the same for (c–f).

of the 4096 data points onto each of the first five PCs, with each plot showing a specific PC plotted against PC1. Figure 2b displays PC1 plotted against itself. Similarly, Figure 2c presents PC2 versus PC1, and so forth for the subsequent components. This analysis demonstrates that the variance associated with each component progressively diminishes up to PC5, which is the last component included in the dataset reconstruction.

Once the dataset analyzed in this work reaches a sufficient volume, the application of the PCA method significantly

enhances the detection of spectral peaks associated with contaminants, as shown in **Figure 3**. Before the analysis, these peaks were obscured by noise, even after background subtraction, making it difficult to distinguish them from the random variations originating from noise. After applying PCA, several peaks become prominent and, when their intensity is mapped, a distinct pattern is revealed, corresponding to the spatial distribution of the contaminants. This improved spectral clarity and enabled a more precise correlation between the chemical features and the observed morphological structures.

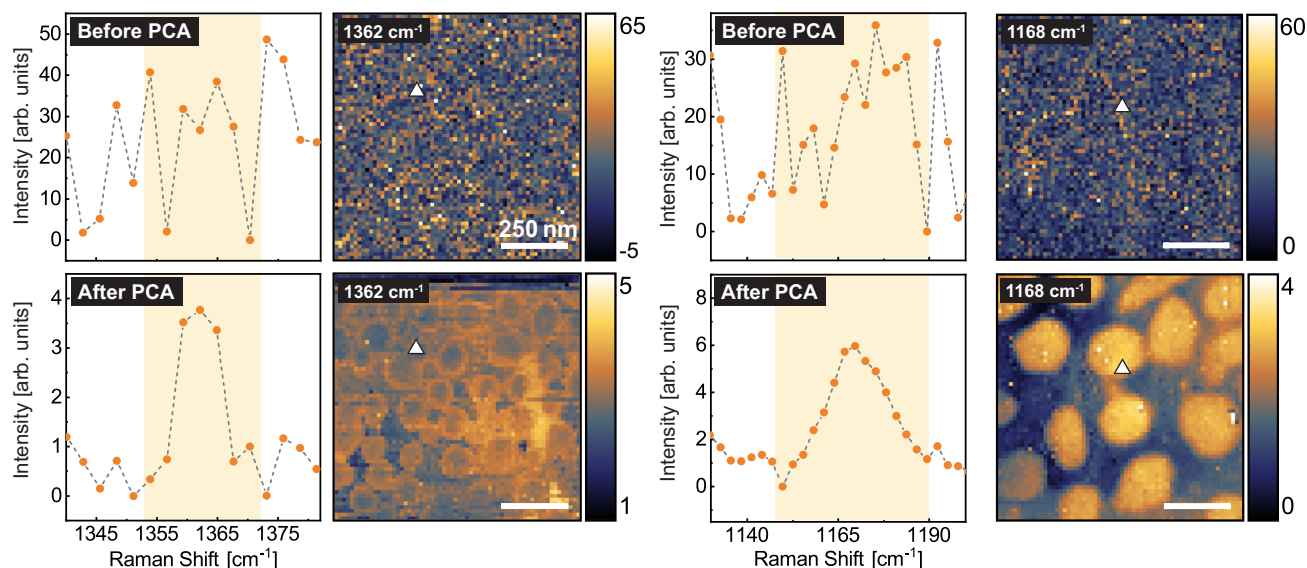


Figure 3. Comparison between as-measured (top row, Before PCA) and PCA-based noise reduction processed (bottom row, After PCA) TERS spectra for two distinct peaks. The intensity maps were obtained by selecting the spectral region highlighted in yellow in the spectra. Triangles represent the pixels from which the spectra were acquired. The scale bars correspond to 250 nm in all images, and the color-coded bars represent the intensities in arbitrary units. The maps were acquired from two different regions of the sample.

Figure 3 presents a direct comparison between the spectra before (first row) and after (second row) PCA-based noise reduction, clearly demonstrating that noise initially obscures relevant features, which only become discernible after data processing. The corresponding intensity maps obtained for each region before and after PCA-noise reduction reveal relevant features associated with chemical signatures likely originating from contamination, further validating our methodology and findings.

4. Conclusions

PCA proved to be an effective tool for removing noise from TERS hyperspectra. Together, these two techniques enable the detection of weak nano-Raman peaks that would otherwise remain undetectable due to both spectral noise and the limitations of conventional optical microscopy in visualizing features at the nanoscale. The comparison of spectra before and after PCA application demonstrates an improvement in data quality. Furthermore, reconstruction using only five PCs was sufficient to preserve all relevant information, indicating that the evaluation of cumulative variance is a reliable criterion to determine the appropriate number of components. This is further supported by the observation that subtracting the original data from the reconstructed yields only residual noise. This methodological approach significantly enhances the capacity for nanoscale chemical and morphological characterization and can be extended to other complex systems, for which weak signals are typically masked by noise. Therefore, PCA stands out as a powerful and versatile tool for large-scale data analysis.

Although PCA is a powerful tool for enhancing these weak signals, it is important to note that subtle spectral features with low variance, represented by discarded PCs, may be lost along with the noise. While these features may not represent significant variance overall, they can be crucial for capturing specific details in the sample.^[12] Furthermore, it is important to emphasize that PCA has limitations in capturing complex non-linear variances, and its effectiveness is highly dependent on the data volume to obtain reliable results. Moreover, PCs are mathematical constructs rather than actual spectra, which limits the direct physical interpretation of their significance.^[11,13] In conclusion, all data processing should always be performed with due care.

Acknowledgements

The authors thank financial support by FAPEMIG (APQ - 04852-23, APQ - 01860-22, RED - 00081-23, APQ-01402-23, RED-00079-23), the Japan Science and Technology Agency (JST), the JST FOREST Program (grant no. JPMJFR213X), the CREST (grant no. JPMJCR24A5), Kakenhi Grants-in-Aid (grant nos. JP21H05232, JP21H05233, JP21H05234, JP22H00283, JP22H04957, and JP23H02052) from the Japan Society for the Promotion of Science (JSPS), World Premier International Research Center Initiative (WPI), MEXT, Japan, and software resources and technical assistance provided by FabNS. R.S. acknowledges a JSPS KAKENHI Grant (grant no. JP22H00283), Japan, and the Yushan Fellow Program by the Ministry of Education (MOE), Taiwan.

The Article Processing Charge for the publication of this research was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) (ROR identifier: 00x0ma614).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

MoSe₂, nano-Raman spectroscopy, principal component analysis, tip-enhanced Raman spectroscopy

Received: May 30, 2025

Revised: July 13, 2025

Published online:

- [1] R. M. Stockle, Y. D. Suh, Y. D. Suh, V. Deckert, R. Zenobi, *Chem. Phys. Lett.* **2000**, 318, 1.
- [2] N. Hayazawa, Y. Inouye, Z. Sekkat, S. Kawata, *Opt. Commun.* **2000**, 183, 1.
- [3] M. S. Anderson, *Appl. Phys. Lett.* **2000**, 76, 21.
- [4] N. Kumar, S. Mignuzzi, W. Su, D. Roy, *EPJ Tech. Instrum.* **2015**, 2, 1.
- [5] A. C. Gadelha, D. A. Ohlberg, C. Rabelo, E. G. Neto, T. L. Vasconcelos, J. L. Campos, J. S. Lemos, V. Ornelas, D. Miranda, R. Nadas, F. C. Santana, *Nature* **2021**, 590, 7846.
- [6] F. B. Sousa, R. Nadas, R. Martins, A. P. Barboza, J. S. Soares, B. R. Neves, I. Silvestre, A. Jorio, L. M. Malard, *Nanoscale* **2024**, 16, 27.
- [7] A. Jorio, R. Nadas, A. G. Pereira, C. Rabelo, A. C. Gadelha, T. L. Vasconcelos, W. Zhang, Y. Miyata, R. Saito, M. D. Costa, L. G. Cançado, *2D Mater.* **2024**, 11, 3.
- [8] C. Höppener, J. Aizpurua, H. Chen, S. Gräfe, A. Jorio, S. Kupfer, Z. Zhang, V. Deckert, *Nat. Rev. Methods Primers* **2024**, 4, 1.
- [9] I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, NY **2002**.
- [10] H. Abdi, L. J. Williams, *Wiley Interdiscip. Rev.: Comput. Statist.* **2010**, 2, 4.
- [11] G. Rusciano, G. Zito, R. Isticato, T. Sirec, E. Ricca, E. Bailo, A. Sasso, *ACS Nano* **2014**, 8, 12300.
- [12] S. Jiang, X. Zhang, Y. Zhang, C. Hu, R. Zhang, Y. Zhang, Y. Liao, Z. J. Smith, Z. Dong, J. G. Hou, *Light: Sci. Appl.* **2017**, 6, e17098.
- [13] J. L. E. Campos, H. Miranda, C. Rabelo, E. Sandoz-Rosado, S. D. Pandey, J. Riikonen, A. G. Cano-Marquez, A. G. Cano-Márquez, A. Jorio, *J. Raman Spectrosc.* **2018**, 49, 1.
- [14] M. Greenacre, P. J. Groenen, T. Hastie, A. I. d'Enza, A. Markos, E. Tuzhilina, *Nature Rev. Methods Primers* **2022**, 2, 1.
- [15] H. Shinzawa, K. Awa, W. Kanematsu, Y. Ozaki, *J. Raman Spectrosc.* **2009**, 40, 12.
- [16] Y. Luo, X. Zhang, Z. Zhang, R. Naidu, C. Fang, *Anal. Chem.* **2022**, 94, 7.

- [17] S. Masubuchi, M. Sakano, Y. Tanaka, Y. Wakafuji, T. Yamamoto, S. Okazaki, K. Watanabe, T. Taniguchi, J. Li, H. Ejima, T. Sasagawa, *Sci. Rep.* **2022**, 12, 1.
- [18] H. Naito, Y. Makino, W. Zhang, T. Ogawa, T. Endo, T. Sannomiya, M. Kaneda, K. Hashimoto, H. E. Lim, Y. Nakanishi, K. Watanabe, T. Taniguchi, K. Matsuda, Y. Miyata, *Nanoscale Adv.* **2023**, 5, 18.
- [19] T. L. Vasconcelos, B. S. Archanjo, B. Fragneaud, B. S. Oliveira, J. Riikonen, C. Li, D. S. Ribeiro, C. Rabelo, W. N. Rodrigues, A. Jorio, C. A. Achete, *ACS Nano* **2015**, 9, 6.
- [20] T. L. Vasconcelos, B. S. Archanjo, B. S. Archanjo, B. S. Oliveira, R. Valaski, R. C. Cordeiro, H. G. Medeiros, C. Rabelo, A. R. Ribeiro, A. R. Ribeiro, P. Ercius, C. A. Achete, A. Jorio, L. G. Cançado, *Adv. Opt. Mater.* **2018**, 6, 20.
- [21] C. Rabelo, H. Miranda, T. L. Vasconcelos, L. G. Cançado, A. Jorio, in *2019 4th Int. Symp. on Instrumentation Systems, Circuits and Transducers (INSCIT), São Paulo, Brazil 2019*, pp. 1–6.
- [22] H. Miranda, C. Rabelo, T. L. Vasconcelos, L. G. Cançado, A. Jorio, *Phys. Status Solidi-rapid Res. Lett.* **2020**, 14, 9.
- [23] J. Guimarães, R. Nadas, R. Alves, W. Zhang, T. Endo, K. Watanabe, T. Taniguchi, R. Saito, Y. Miyata, B. Neves, A. Jorio, **2025**, ArXiv pre-print arXiv:2505.19224.