The Baseline Harmonization Algorithm: A Dataset-Level Strategy for Robust Raman Spectroscopy Classification in Liquid Biopsy Analysis

Elisa Grassi[a], **Carlo Liberale* [a,b]

[a] Biological and Environmental Science and Engineering (BESE) Division, King Abdullah University of Science and Technology (KAUST)

[b] Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST)

* *carlo.liberale@kaust.edu.sa*

Baseline variability remains one of the dominant sources of distortion in Raman spectroscopy, arising from sample autofluorescence, wavelength-dependent scattering, and instrument-dependent spectral response [1]. Such variability can overshadow weak but relevant vibrational signatures and severely impair reproducibility in classification applications, particularly in medical applications such as liquid-biopsy analysis for disease detection. Currently used baseline correction algorithms—including SNIP [2] and EMSC [3]—operate on each spectrum independently and often suffer from incomplete background removal or Raman peak distortion, especially when baseline shapes vary markedly across samples. Moreover, their reliance on user-defined parameters introduces additional variability and subjectivity into the preprocessing workflow. Here we introduce the Baseline Harmonization Algorithm (BHA), a pre-processing method that equalizes the baseline across measurements, by suitably minimizing variations among spectra so that the remaining differences reflect only Raman-specific information. A key novelty of BHA is the modeling of inter-sample baseline variability using a representation that combines polynomials and a spline component. This formulation allows BHA to capture both broad baseline curvature and local undulations without producing artifacts such as oscillatory rippling or Raman peak distortion. The performance of BHA was compared with SNIP and EMSC using synthetic Raman datasets designed to reproduce realistic measurement conditions and baseline variability. Classification performance was evaluated using with PCA–SVM models with both internal and external validations. Across all simulations, BHA provided higher classification accuracies, particularly in external validations. The method was then applied to Raman spectra of blood plasma from cancer patients and matched controls. Following BHA preprocessing and PCA–SVM classification, sensitivities and specificities approached 0.95. Models trained on BHA-corrected spectra consistently outperformed those where data were pre-processed using SNIP or EMSC. Importantly, BHA showed markedly improved performance in external validation on independent datasets measured under different conditions. This configuration represents a particularly stringent test for classification workflows, thus successful performance in this scenario provides strong evidence of an algorithm's robustness and generalization capability.

References

[1] F. Bonnier, S. M. Ali, P. Knief, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T. Lee, F. M. Lyng, and H. J. Byrne, "Analysis of human skin tissue by Raman microspectroscopy: dealing with the background," Vibrational Spectroscopy, vol. 61, pp. 124–132, 2012.

[2] D. D. Burgess and R. J. Tervo, "Background estimation for gamma-ray spectrometry," Nuclear Instruments and Methods in Physics Research, vol. 214, no. 2-3, pp. 431–434, 1983.

[3] L. Kerr and B. Hennelly, "A multivariate statistical investigation of background subtraction algorithms for Raman spectra of cytology samples recorded on glass slides," Chemometrics and Intelligent Laboratory Systems, vol. 158, pp. 61–68, 2016.