Mining large Raman spectroscopic data beyond the shot noise limit

Shimada, Rintaro; Ozawa, Takeaki

2020

Shimada, R., & Ozawa, T. (2020). Mining large Raman spectroscopic data beyond the shot noise limit. Proc. Of the 7th Asian Spectroscopy Conference (ASC 2020).
doi:10.32655/ASC_8-10_Dec2020.77

<https://hdl.handle.net/10356/144412>

https://doi.org/10.32655/ASC_8-10_Dec2020.77

Mining Large Raman Spectroscopic Data Beyond the Shot Noise Limit

Rintaro Shimada and Takeaki Ozawa

*Department of Chemistry, School of Science, the University of Tokyo,
7-3-1 Hogo, Bunkyo-ku, Tokyo, Japan
E-mail: rintaro@chem.s.u-tokyo.ac.jp*

Abstract

Various multivariate chemometric techniques have been proven to be robust in analyzing complex Raman hyperspectral datasets obtained from chemically diverse biological samples for classification, quantification, and exploratory studies. Among various techniques, singular value decomposition (SVD) and its mathematical akin, principal component analysis (PCA), are particularly useful because they provide objective summarization of a given dataset by approximation of observed spectra into linear combinations of a finite number of significant basis spectra along with effective signal-noise separation. The number of significant components retained by SVD essentially indicates the wealth of information retracted from the spectroscopic dataset; therefore, increasing the number is deemed important for improving analytical performance, especially when minor species are under investigation. Generally, the use of a larger dataset is considered to yield more significant components. However, the dataset size relationship to the efficacy of SVD summarization has not been extensively studied. In this presentation, we will report results from the systematic study on SVD analysis of Raman hyperspectral datasets of various size, from which the presence of fundamental limitation which prevents recovery of minor signals is unraveled. Furthermore, a possible workaround using variance stabilization transform is demonstrated to overcome the limitation.

SVD and PCA both decomposes the input data matrix based on its covariance statistics. In SVD, repeatedly or concurrently appearing physically relevant signals have high covariance in both spectral and spatial dimensions and are thus expected to stand out from random noise components that have ideally no covariance with the signal. However, systematic SVD analysis on Raman hyperspectral datasets, obtained from a biological specimen of various size up to 10^5 spectra, revealed the deteriorated signal-noise separation efficiency of SVD for a larger dataset (above 10^4 spectra), which results in the saturation of the increase in the number of significant components against input data size. We suspected shot noise to be responsible for this limitation through the intrinsic correlation between signal amplitude and its variance. To overcome this limitation, we applied generalized Anscombe transform [1,2] (GAT) to the Raman spectroscopic dataset to stabilize variance, i.e., decorrelate shot noise from its amplitude. GAT-SVD processed datasets exhibited a monotonic increase in the number of components throughout the entire data size range we have processed. A computational simulation of a spectroscopic data (up to 10^6 spectra) with artificial Poisson (shot) and Gaussian noises confirmed the plateaued data-size dependence of the signal-noise separability by SVD and the improvement by GAT-SVD. The proposed method opens up the possibility for vibrational analysis of minor species that had not been detectable due to the shot noise interference.

Keywords: Denoising, chemometrics, hyperspectral data analysis

References:

- [1] F.J. Anscombe, *Biometrika*, **1948**, 35, 246
- [2] J.L. Starck, et al. "*Image Processing and Data Analysis, The Multiscale Approach*," U.K.:Cambridge Univ. Press, **1998**.