Vibrational Spectroscopy 81 (2015) 62–67

Contents lists available at ScienceDirect

Vibrational Spectroscopy

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

Vibrational spectra, principal components analysis and the horseshoe effect

P.D. Lewis*, G.E. Menzies

Institute of Life Science, Swansea University Medical School, Swansea University, UK

ARTICLE INFO

Article history:
Received 20 July 2015
Received in revised form 19 September 2015
Accepted 8 October 2015
Available online 22 October 2015

Keywords:
Vibrational spectroscopy
Principal components analysis
Horseshoe effect

ABSTRACT

Vibrational spectroscopy studies often generate datasets containing multiple spectra that are categorized into distinct groups according to similarity. Principal components analysis (PCA) is one of the most frequently used multivariate analysis methods for data reduction of vibrational spectra and visualization of potential groupings between subjects. Vibrational spectra usually display unimodal or multimodal distribution patterns of absorbance or transmittance across wavenumbers. PCA, requires that a linear relationship exists between data distributions of the objects under analysis otherwise the method is prone to a serious artifact known as the 'horseshoe effect'. This artifact, well known in other fields of science, manifests as a serious distortion of the pattern of how objects group according to the most important principal components leading to misinterpretation of the relationships between the samples from which they are derived. In this paper, using a simulated mid-infrared spectral dataset, we investigate for the first time the potential for the PCA horseshoe effect on vibrational spectra and the why this artifact occurs. We show that when comparing large regions of contiguous wavenumbers between multiple spectra there can be a non-linear relationship between distributions of different spectra. Such non-linearity causes the horseshoe effect and we demonstrate that the degree of distortion of how spectra map on the first two components is related to the region size. We further show that reducing the size of spectra analyzed by PCA can minimize the horseshoe effect. We conclude that PCA should be used with caution in the analysis and interpretation of vibrational spectra and the application of more robust methods should be explored.

1. Introduction

Vibrational spectroscopy studies often generate datasets containing multiple spectra that are categorized into distinct groups according to similarity. This is especially true of studies using mid/near-infrared or raman spectroscopy for characterization of molecular composition or structure and biomedical diagnosis using fluids or solid tissue [1]. Simple statistical analysis of complex biological sample spectra is often not appropriate due to the sampling of multiple spectra, differentiation of spectra by group and strong overlapping spectral features. Therefore, multivariate analysis methods are usually deployed to large datasets to assist in visualization of relationships of spectral features either within or between groups of spectra.

Multivariate analysis methods are a group of statistical procedures used to simultaneously analyze three or more variables. One of the oldest and most commonly used of these methods is principal components analysis (PCA). Although a major data reduction tool in chemometrics, this technique has been widely used in many scientific fields as diverse as molecular biology [2], the behavioural sciences [3] computational toxicology [4], industrial chemistry [5] and ecology [6]. We can cite but a few of many studies in the literature where PCA has been used on vibrational spectroscopic data to develop discriminatory models for diverse objectives such as disease diagnosis [7,8], cell type characterization [9], bacterial strain differentiation [10] as well as seed varieties [11]. Another important application of vibrational spectroscopy is in determination of single biomolecule structure, particularly protein secondary structure [12–14]. PCA has been used to associate protein absorbance change and shift in frequencies due to altering environmental conditions such as temperature for both near [15] and mid infrared spectroscopy [16].

The main objective of PCA is to extract important information from a table of high dimensional data with inter-correlated

* Corresponding author at: Institute of Life Science, Swansea University Medical School, Swansea University, Swansea SA2 8PP, UK.
E-mail address: p.d.lewis@swansea.ac.uk (P.D. Lewis).

http://dx.doi.org/10.1016/j.vibspec.2015.10.002
0924-2031/

variables into new and fewer uncorrelated variables. These reduced variables reveal trends in the data that are otherwise difficult to visualize. For vibrational spectroscopy data, this table of absorbance would be comprised of rows of individual spectra where each column represented the wavenumbers in the spectra. The table is often mean centred prior to PCA. A covariance matrix is calculated from the data table from which eigenvalues (variance explained) and eigenvectors are found and eigenvectors ranked according to the size of eigenvalue. This new matrix of data describes a multidimensional coordinate system where the axes are rotated so as to align with the greatest variation of the data. The first axis, or principal component, captures the most variation as this eigenvector has the largest eigenvalue. The second component captures the second highest variance, independent from the first, and so on. The eigenvector is a series of weights (loadings) for each wavenumber on a component. Linear combinations of these weights with the original data can be summed to give scores for each spectrum on a component. Thus, the relationships between spectra according to how they co-vary by wavenumber absorbance may be visualized in this coordinate system by plotting the scores of each spectrum on the most important components. In this way, potential groupings of spectra by similarity are easily visualized using scatterplots. For a full introduction into the concepts and detailed steps of PCA, readers are referred to some popular introductions [17,18].

PCA is however prone to a serious artifact that can lead to false interpretation of how objects under consideration could group. In the field of ecology it is well known that when species abundance data along a sampling gradient (species response curve) are analyzed by PCA then the resulting scatterplot of species scores on the first two components will often show a distortion [19]. This distortion occurs when the second axis is curved and twisted relative to the first as an arch or "horseshoe" pattern of objects and is not a true secondary gradient. The cause of the horseshoe effect relates to the fact that species response curves are unimodal in distribution (like a Gaussian curve) especially over a long gradient where there are few species sampled at the ends of the gradient. Even though species may be truly different in abundance along a gradient, they may all share low or zero abundance at the tail ends. This shared zero abundance is meaningless in terms of how species vary but PCA assumes species similarity at these points which manifests in similar scores on component two and its arch over component one [20,21]. The horseshoe effect is also not confined to unimodal species response curves in ecology but also with more complex multimodal models [18]. PCA is only really useful when objects are linearly related to each other as monotonic distributions or when the gradient assessed is short.

An absorbance (or equally transmittance) band in a vibrational spectrum displays a unimodal distribution and is analogous to a unimodal species response curve in ecology. A whole absorbance spectrum is analogous to a multimodal species response curve. This suggests that PCA applied to regions of contiguous wavenumbers across a series of spectra with variable band peak positions (a common practice), could be susceptible to the horseshoe effect and misinterpretation of results.

In this paper, using a simulated mid-infrared spectral dataset, we investigate for the first time the potential for the PCA horseshoe effect on vibrational spectra and the why this artifact occurs. We show how use of regions of contiguous spectral wavenumbers causes the horseshoe effect and the degree of distortion is related to the region size.

2. Methods

A dataset was created for 15 spectra in the mid-infrared spectral range. The simulated spectra were loosely based on the series of 25 temperature dependent spectra for Bovine pancreatic ribonuclease A (RNase A) reported by Wang et al. [16] after that protein had been subjected to 2 °C increments between 25 and 70 °C. In that study, the RNase A spectra between 1600 and 1700 cm^{-1} showed absorbance at 1641 cm^{-1} that weakens as temperature increases with the formation of a band at 1653 cm^{-1}. Two weak bands were also observed at 1615 and 1689 cm^{-1} that varied with temperature.

In our artificial dataset we generated 15 simulated spectra between 1600 and 1700 cm^{-1} at a resolution of 4 cm^{-1} for a temperature range between 15 and 85 °C and a temperature increment of 5 °C. Thus, our dataset is not intended to replicate that published by Wang et al. [16] but to be one that simply shows a similar pattern of variable multimodal spectra over a 72 data point wavenumber range of 100 cm^{-1}. Each spectrum was created by curve fitting at each of the four wavenumbers described by Wang et al. [16], varying the full-width at half-height and standard deviation.

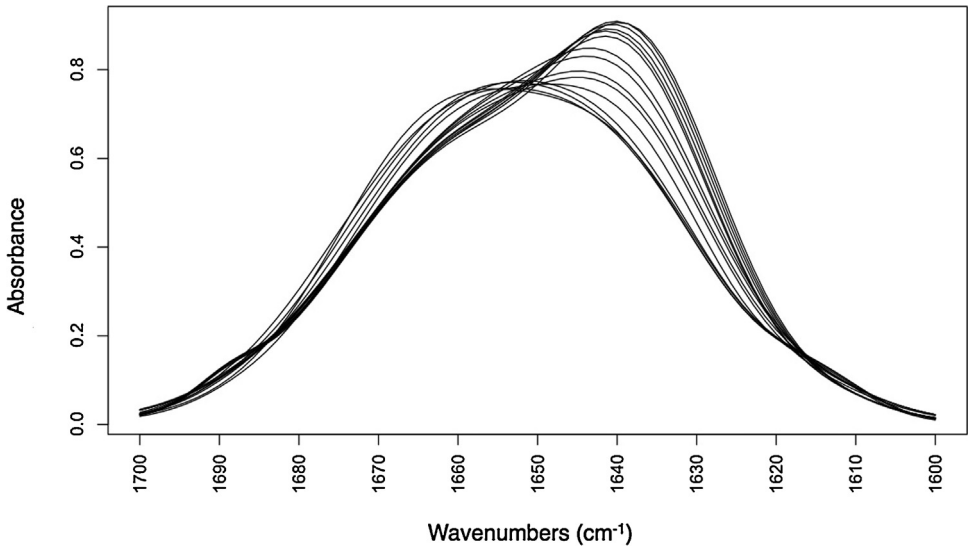

Fig. 1. The 15 absorbance spectra simulating absorbance of RNase A between 1600 and 1700 cm^{-1} with a 2 °C increment between the range of 25 and 70 °C. As the temperature increases the absorbance maximum weakens and shifts from 1641 cm^{-1} to around 1653 cm^{-1}.