Title: Vibrational spectra, principal components analysis and the horseshoe effect

Author: P.D. Lewis G.E. Menzies

PII: S0924-2031(15)30018-7
DOI: http://dx.doi.org/doi:10.1016/j.vibspec.2015.10.002
Reference: VIBSPE 2460

To appear in: *VIBSPE*

Received date: 20-7-2015
Revised date: 19-9-2015
Accepted date: 8-10-2015

Please cite this article as: P.D.Lewis, G.E.Menzies, Vibrational spectra, principal components analysis and the horseshoe effect, Vibrational Spectroscopy http://dx.doi.org/10.1016/j.vibspec.2015.10.002

Vibrational spectra, principal components analysis and the horseshoe effect

Lewis PD*, Menzies GE

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

* Corresponding author

Address: Institute of Life Science, Swansea University Medical School, Swansea University, Swansea, UK, SA2 8PP

Tel: +44 1792 295222

Email: p.d.lewis@swansea.ac.uk

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.

Keywords: vibrational spectroscopy; principal components analysis; horseshoe effect.

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 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 it’s 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*. [15] after that protein had been subjected to 2^{o}C increments between 25 and 70^{o}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 4cm^{-1} for a temperature range between 15 and 85^{o}C and a temperature increment of 5^{o}C. Thus, our dataset is not intended to replicate that published by Wang *et al*. [15] 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*. [15], varying the full-width at half-height and standard deviation.

All data analysis was performed using the R Statistical Programming Environment [22]. Principal components analysis was performed using the 'prcomp' function with default parameters.

3. Results and Discussion

3.1. Simulated RNase A spectral dataset

The 15 simulated absorbance spectra with 72 wavenumber data points between 1600 and 1700 cm^{-1} and of a 5^{o}C increment between the range of 15 and 85^{o}C are shown in Figure 1. At 15^{o}C the spectrum shows a strong peak at 1641 cm^{-1} and weaker peak within the shoulder at around 1653 cm^{-1}. As the temperature increases, the absorbance maximum decreases and shifts from 1641 cm^{-1} whereas absorbance at the band peaking around 1653 cm^{-1} increase in intensity. Very small absorbance fluctuations between spectra exist at the two weak bands around 1615 and 1689 cm^{-1}. There are a number of important observations to be made about these spectra regardless of response to temperature increment. Firstly, the spectra have a multimodal distribution although the spectra for the highest temperatures appear almost unimodal. Secondly, for each spectrum, the absorbance values for contiguous wavenumbers at each end are very similar to each other. Thirdly, there is very little variation between spectra for absorbance at each end.

3.2. PCA and the non-linearity relationship between spectra

PCA was applied to the 15 spectra dataset using absorbance values for all 72 wavenumbers between 1600 and 1700 cm^{-1}. A plot of scores for each spectrum on principal components one (PC1) and two (PC2) is shown in Figure 2A. The order of spectra by scores on PC1 appears to mirror the temperature gradient and we can confidently interpret PC1 as explaining the variance due to temperature. The spectra do however display a clear horseshoe shaped distribution as the PC2 axis curves on that for PC1. The spectra of lowest and highest temperatures, i.e. those at the each end of the spectral gradient, begin to twist back characteristic of the horseshoe effect. The loadings of each wavenumber on PC2 are shown in Figure 2B. The negative loadings of wavenumbers 1630 and 1669 cm^{-1} on PC2 coincide with the negative scores of the spectra at the ends of the temperature gradient. Taking all this information into account one would conclude from PCA that similarities exist between the lower and higher temperature spectra influenced by a similar pattern of

absorbance around 1630 and 1669 cm^{-1}. In reality, a quick look at Figure 1 shows that spectra vary in opposite directions by absorbance level at these two wavenumbers. There is certainly no interpretable reason why spectra corresponding to temperatures at 55 and 60°C would be at the opposite end of a gradient, as suggested by PC2, to those spectra at 15, 20, 25, 30, 75, 80 and 85°C.

This horseshoe pattern results from a composition of several such non-linear relationships between spectra. The non-linear relationship between wavenumbers at the extreme ends of the spectral gradient i.e. for 15 and 85°C is obvious when just aligning those spectra (Figure 3A). The relative degree of non-linearity between any two spectra can be visualized by plotting the absorbance values for each wavenumber as a scatterplot. Figure 3B shows scatterplots of absorbance between the spectrum at 15°C and each of the remaining spectra in order of increasing temperature. As expected, a scatterplot showing the joint distribution of wavenumber absorbance for the two spectra of 15C and 20°C (i.e. very similar spectral shape) have a linear relationship. As the temperature increases, we see the scatterplots between 15 and 25°C through the temperature gradient to 85°C show an increasingly stronger curvilinear relationship. Thus, the non-linearity increases as the spectral distribution modes are shifted along the wavenumber axis.

3.3. Relationship between horseshoe effect and spectrum length

To examine the contribution of wavenumbers at the end of spectral regions to the horseshoe effect, PCA was carried out on spectra with decreasing lengths. Starting with a full length of 72 wavenumbers, all spectra were iteratively shortened by a wavenumber at each end and the remaining contiguous wavenumbers entered into PCA. This shortening was repeated until 30 wavenumber data points had been removed from each end leaving a shortest spectrum of 11 data points spanning the middle of the original spectra. Figure 4 shows scatterplots of spectral scores on PC1 and PC2 for the range of increasingly shortened spectra. The first plot shows the characteristic horseshoe as observed in Figure 2A when the spectral range comprised of all wavenumbers. As the range decreases, it appears that the horseshoe effect remains until at least

25 wavenumbers have been removed from either end (a total of 50) of the spectral data points.

To test whether the PCA scores for the variable length spectra on PC1 and PC2 conformed to a horseshoe or arch shape, both quadratic and linear curves were fitted to the data. A one-way ANOVA was used to determine if a quadratic model fit was a significant improvement on a linear model fit in each case. If the *P-value* was greater than 0.05 then the score distribution was assumed to not follow a quadratic curve and thus the horseshoe effect was rejected. The line plot in Figure 5 shows the *P-values* generated for each length decrease. A quadratic is a significantly better fit until the spectral length is decreased by 26 wavenumbers from each end. This range corresponds to wavenumbers between 1636 to 1664 cm^{-1} meaning that the horseshoe effect was minimized by analysing a shorter range of 20 wavenumbers. Interestingly, when the proportion of variance explained by PC2 was assessed for the varying length spectra the values remained constant at around 0.017 until the end length decrease of 26 wavenumbers after which the variance increases. On this evidence we can further conclude that inclusion of long wavenumber ranges for PCA does not increase the information explained by PCA.

4. Conclusion

Vibrational spectra have unimodal or multimodal distributions of absorbance across wavenumbers. As such, there can often be a shape dependent non-linear relationship between different spectra from different sources. In this paper, we've highlighted how application of PCA, a widely used data reduction technique, to regions of contiguous wavenumbers in vibrational spectra can lead to artifacts which could lead to misinterpretation of the relationships between the samples from which they are derived. The length of the spectral wavenumber region analyzed by PCA is clearly critical. Long, approximately unimodal spectra entered into PCA with little variation at the ends will lead to the horseshoe effect. Shorter wavenumber ranges are more likely to approximate a linear relationship between spectra and reduce distortion of the inter-spectral similarities. It is a

simple process to iteratively shorten the spectra to the point where a quadratic model is not a significantly better fit than a linear model.

In practice, many vibrational spectroscopy datasets of multiple spectra display a more complex pattern of absorbance distribution across wavenumbers. Such spectra are multimodal and this will add further complexity to the pattern of distortion observed when plotting component scores and loadings. Such complexity renders these artifacts difficult to detect. There are modifications of PCA that attempt to limit the horseshoe effect and other multivariate analysis techniques that are applicable to unimodal data. Detrended principal components analysis attempts to correct the horseshoe pattern by projecting the objects onto a single axis following a regression of component 1 onto component 2 [6]. In his study comparing different multivariate analysis methods applied to artificial species response data, Minchin concludes that non-multidimensional scaling (NMDS), using the Bray-Curtis dissimilarity coefficient, despite showing some curvilinear distortion, is the most robust and effective of the methods compared to PCA, principal co-ordinates analysis, detrended correspondence analysis and Gaussian ordination [19]. Other methods should be evaluated on vibrational spectra.

We conclude that PCA should be used with extreme caution in the analysis and interpretation of vibrational spectra when using contiguous wavenumbers. Further work is required to determine best practice of applying multivariate statistical methods to vibrational spectra for optimal interpretation of results.

Acknowledgements

This research is supported by Lung Research Wales through funding by the National Institute for Social Care and Health Research (NISCHR).

References

[1] L. Wang, B. Mizaikoff, Anal. Bioanal. Chem. 391 (2008) 1641–54.

[2] D. Reich, A.L. Price, N. Patterson, Nat. Genet. 40 (2008) 491–2.

[3] I. Jolliffe. in: B.Everritt, D.Howell (Ed.), Encyclopedia of Statistics in Behavioral Science, John Wiley & Sons Ltd, London, UK, 2005, pp 1580-1584.

[4] D. Groth, S. Hartmann, S. Klie, J. Selbig. in: B. Reisfeld, A.N. Mayeno (Ed.), Methods in Molecular Biology: Computational Toxicology, Vol. 930, New York, USA: Humana Press, 2013, pp. 527-547.

[5] J. Love, Process Automation Handbook: A Guide to Theory and Practice, Springer, London, UK, 2007, pp 827-836 .

[6] J.A. Ludwig, J.F. Reynolds, Statistical Ecology: A Primer on Methods and Computing, John Wiley & Sons Ltd, London, UK, 1988, pp 223-242.

[7] P.D. Lewis, K.E. Lewis, R. Ghosal, S. Bayliss, A.J. Lloyd, J. Wills, et al., BMC Cancer. 10 (2010) 640.

[8] G.R. Lloyd, L.E. Orr, J. Christie-Brown, K. McCarthy, S. Rose, M. Thomas, et al., Analyst. 138 (2013) 3900–8.

[9] T. Nakamura, J.G. Kelly, J. Trevisan, L.J. Cooper, A.J. Bentley, P.L. Carmichael, et al., Mol. Vis. 16 (2010) 359–68.

[10] M.L. Paret, S.K. Sharma, L.M. Green, A.M. Alvarez, Appl. Spectrosc. 64 (2010) 433–41.

[11] G.L. Monferrere, S.M. Azcarate, M.Á. Cantarelli, I.G. Funes, J.M. Camiña, J. Food Sci. 77 (2012) C1018–22. s

[12] A. Barth, Biochim. Biophys. Acta - Bioenerg. 1767 (2007) 1073–1101.

[13] M.N. Kinalwa, E.W. Blanch, A.J. Doig, Accurate determination of protein secondary structure content from raman and raman optical activity spectra, in: Anal. Chem., 2010, pp. 6347–6349.

[14] S.P. Lewis, A.T. Lewis, P.D. Lewis, Vib. Spectrosc. 69 (2013) 21–29.

[15] B. Yuan, K. Murayama, Y. Wu, R. Tsenkova, X. Dou, S. Era, et al., Appl. Spectrosc. 57 (2003) 1223–1229.

[16] L.-X. Wang, F. Meersman, Y. Wu, J. Mol. Struct. 883-884 (2008) 79–84.

[17] H. Abdi, L.J. Williams, Wiley Interdiscip. Rev. Comput. Stat. 2 (2010) 433–459.

[18] N.R. Clark, A. Ma'ayan, Sci. Signal. 4 (2011) tr3.

[19] P.R. Minchin, Vegetatio. 69 (1987) 89–107.

[20] M. Palmer, Botany Department, Oklahoma State University, "Ordination Methods for Ecologists", http://ordination.okstate.edu/PCA.htm [accessed Aug 02, 2013].

[21] P. Legendre, E.D. Gallagher, Oecologia. 129 (2001) 271–280.

[22] R Core Team, (2012) http://www.R-project.org. [accessed Aug 02, 2013].

Figure Legends

Figure 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}.

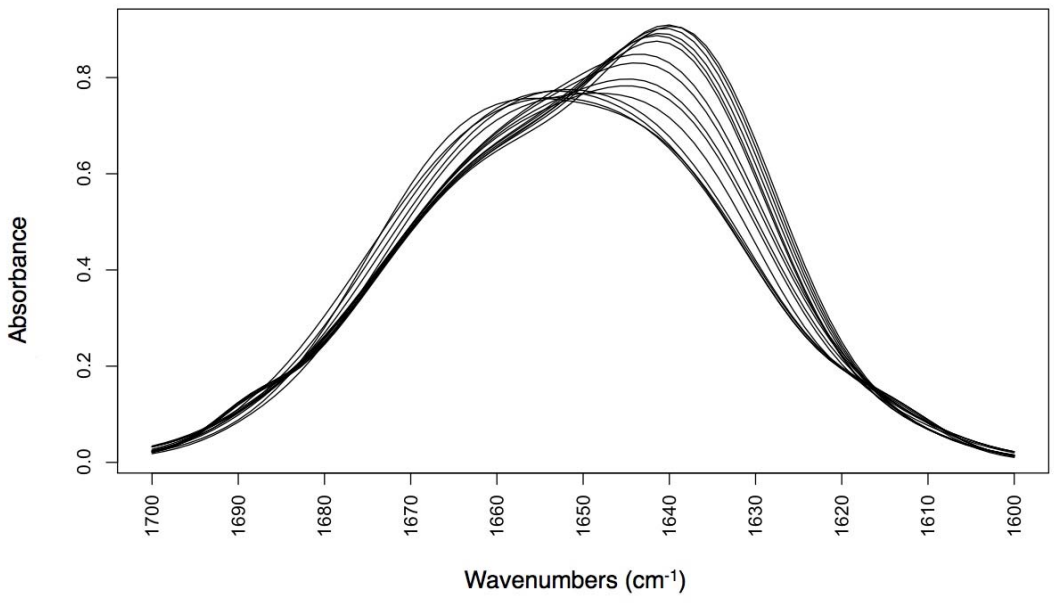

Figure 2. PCA of simulated spectra. A. Scatterplot of scores for each spectrum (labeled as 15 to 85) on principal components one (PC1) and two (PC2). The horseshoe shape can be observed as the PC2 axis curves on that for PC1. B. Line plot of the loadings of each wavenumber on PC2.

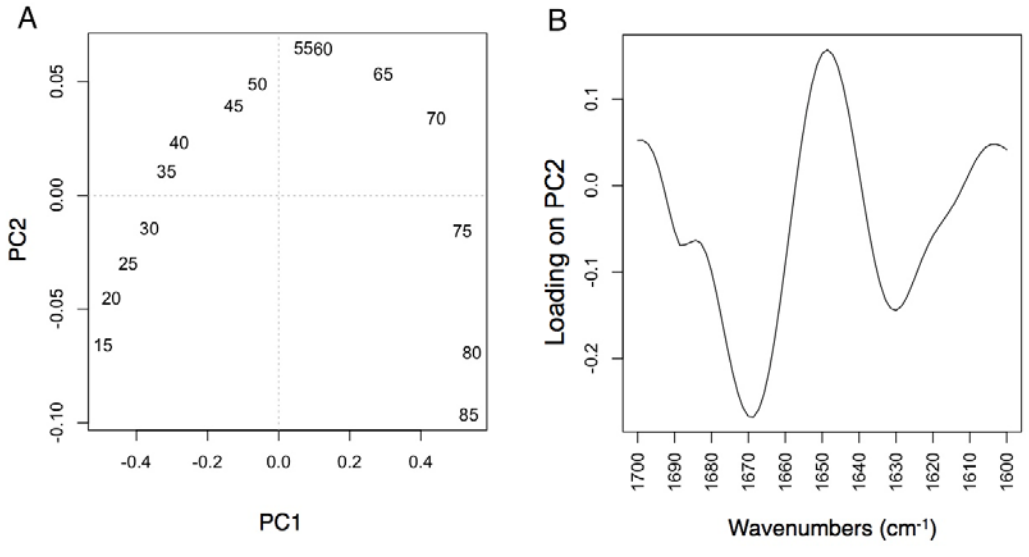

Figure 3. The non-linear relationship between spectra. A. The non-linear relationship between wavenumbers at the extreme ends of the spectral gradient for 15 and 85°C. B. Scatterplots of absorbance between the spectrum at 15°C (x-axis) and each of the remaining spectra (y-axis for each plot), in order of increasing temperature, highlighting the relative degree of non-linearity.

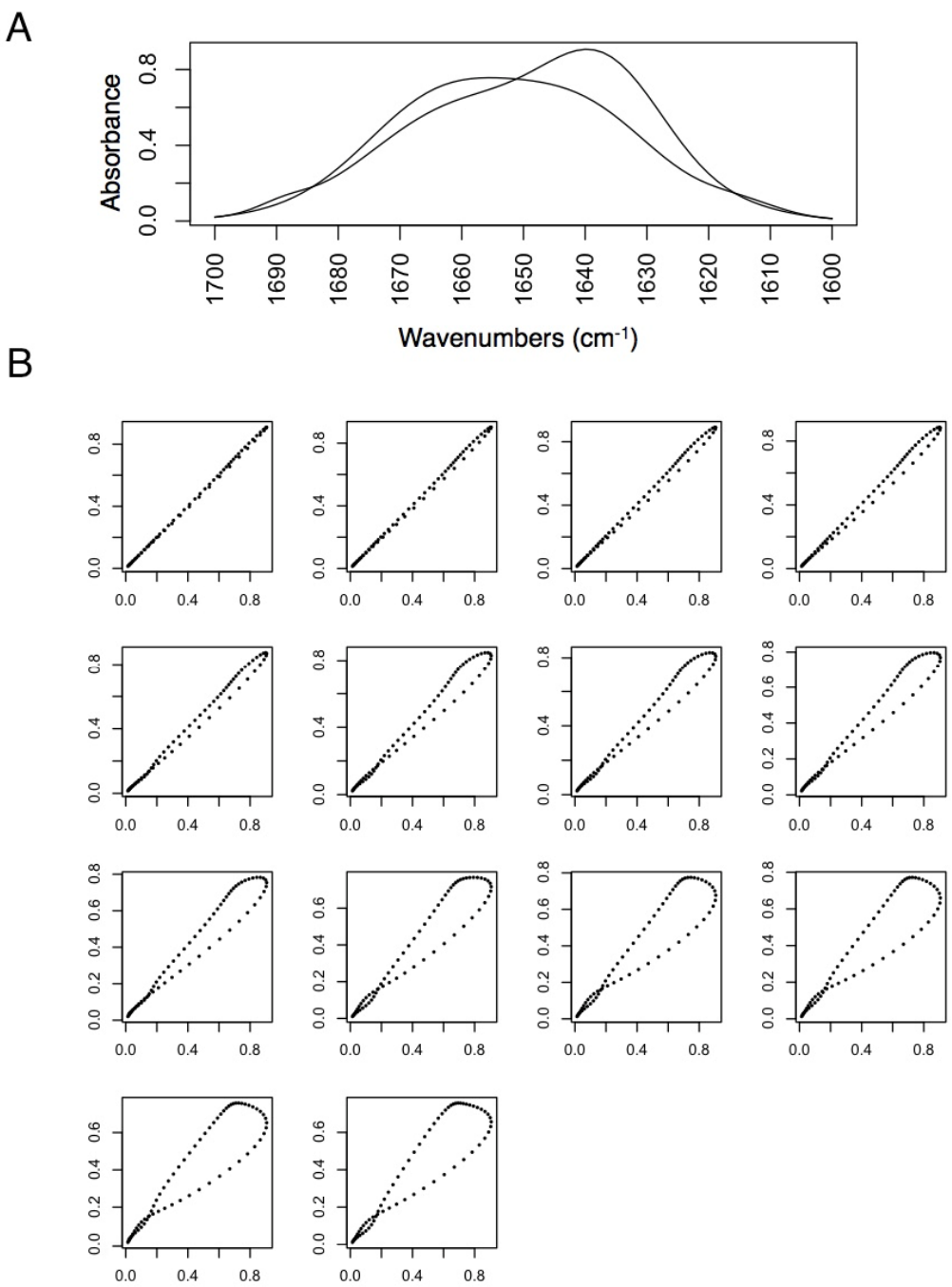

Figure 4. Series of plots of PCA scores for spectra on component 1 (x-axis) and component 2 (y-axis) as the length of wavenumbers input into PCA is decreased. The first plot shows scores for full length spectra. Each plot then shows scores as the spectra decrease in length at each end. Numbers indicate the number of data points removed at each end.

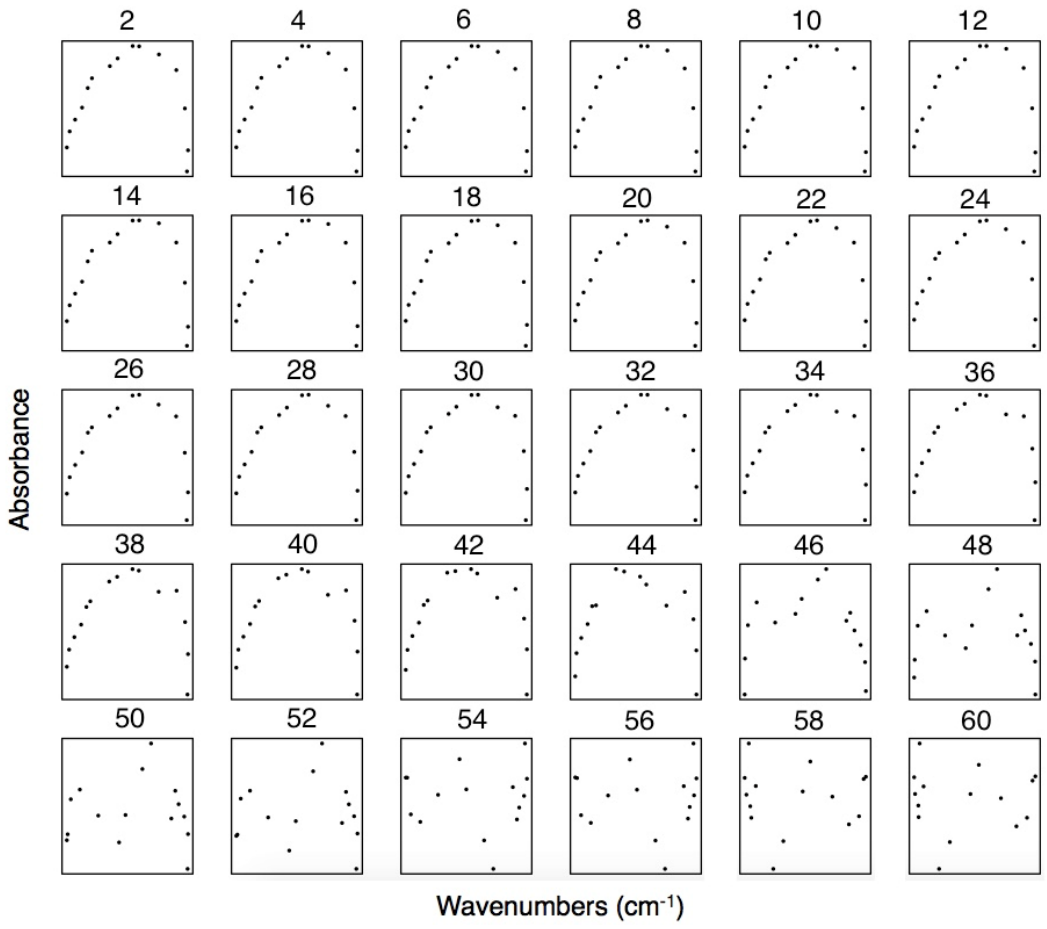

Figure 5. Plot to show the significance of the fit of a quadratic line through the data plotted in Figure 4. The y-axis represents the *P-value* of a one-way ANOVA applied to a quadratic model fit for every spectrum of decreasing size to determine if the quadratic model was a significant improvement on a linear model fit. The x-axis represents the decrease in spectral length from each end. The horizontal dashed line signifies a *P-value* of 0.05 and the vertical dashed line the length of decrease where a linear fit is better than a quadratic. C. The dashed lines enclose the region of spectra where a quadratic fit was not a significant improvement on a linear fit.