A three-dimensional principal component analysis approach for exploratory analysis of hyperspectral data: identification of ovarian cancer samples based on Raman microspectroscopy imaging of blood plasma

Camilo L. M. Morais[1,*], Pierre L. Martin-Hirsch[2] and Francis L. Martin[1,*]

[1]*School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston PR1 2HE, United Kingdom*

[2]*Department of Obstetrics and Gynaecology, Lancashire Teaching Hospitals NHS Foundation Trust, Preston PR2 9HT, United Kingdom*

*cdlmedeiros-de-morai@uclan.ac.uk or flmartin@uclan.ac.uk

Abstract

Hyperspectral imaging is a powerful tool to obtain both chemical and spatial information of biological systems. However, few algorithms are capable of working with full three-dimensional images, in which reshaping or averaging procedures are often performed to reduce the data complexity. Herein, we propose a new algorithm of three-dimensional principal component analysis (3D-PCA) for exploratory analysis of complete 3D spectrochemical images obtained through Raman microspectroscopy. Blood plasma samples of ten patients (5 healthy controls, 5 diagnosed with ovarian cancer) were analysed by acquiring hyperspectral imaging in the fingerprint region (~780–1858 cm^{-1}). Results show that 3D-PCA can clearly differentiate both groups based on its scores plot, where higher loadings coefficients were observed in amino acids, lipids and DNA regions. 3D-PCA is a new methodology for exploratory analysis of hyperspectral imaging, providing fast information for class differentiation.

1. Introduction

In spectrochemical imaging, a spectrum is generated for each pixel in the original image, where both spatial and chemical information are considered. The data are represented by three-dimensional (3D) arrays for each sample measured, where the spatial coordinates are present in the *x*- and *y*-coordinates and the wavenumbers in the *z*-coordinate. Thus, each wavenumber response (a 2D image) is stacked up one above the other in a manner similar to paper sheets in a book in order to form a 3D object,[1] informally called a "data cube".

There are many types of instrumental techniques that generate 3D spectrochemical imaging (*i.e.*, multispectral or hyperspectral imaging), *e.g.*, near-infrared (NIR), infrared (IR), Raman and mass spectrometry (MS).[2-5] Several matrices have been analysed by using spectrochemical imaging, *e.g.*, food,[6,7] soil,[8] atmospheric particulate matter,[9] and tissues.[10] Many chemometric techniques can be used for analysing this type of data, such as principal component analysis (PCA), partial least squares (PLS), multivariate curve resolution (MCR), among others;[11,12] however, in many cases, reshaping, averaging procedures, and data compression are performed in order to reduce dimensionally.[11,13] Recently, some adaptations of first-order algorithms used for classical spectroscopy data, such as linear discriminant analysis (LDA) and PCA, were produced for 2D data obtained via excitation-emission matrix (EEM) fluorescence spectroscopy.[13,14] These algorithms, named 2D-LDA and 2D-PCA, are found to have excellent performance using 2D data without using previous dimensional reduction techniques,[13,14] hence its usage could be extended for chemical imaging.

One of the imaging techniques that has found increasingly applications is Raman microspectroscopy.[15,16] Raman imaging has been used in a wide range of applications,

including investigation of drug delivery systems,[17] pharmaceutical analysis,[18] food quality control,[19] and analysis of biological materials.[16,20] For instance, in cancer detection, Raman imaging has been applied to diagnose breast,[4] skin,[21] cervical,[22] lung,[22] and brain cancers.[23] A major advantage is that the use of Raman imaging provides both chemical and structural information of the sample being analysed with minimum water interference.

Ovarian cancer affects some 7,300 women in the UK alone per year and results in around 4,100 deaths per year.[24] For standard ovarian cancer diagnosis, women with symptoms undergo a pelvic examination followed by measurement of serum cancer antigen (CA-125). If symptoms persist in the absence of raised CA-125 levels, an abdominal and transvaginal ultrasound is performed.[24,25] However, ovarian cancer often presents late symptoms in which the cancer has already metastasized within the abdomen, resulting in late-stage and poor prognoses.[24,25] Besides these limitations, the diagnosis tends to be extremely invasive, expensive and time-consuming. Therefore, alternative methodologies to detect ovarian cancer that can reduce these drawbacks are of major importance, especially towards early-stage diagnosis. Herein, we propose a new algorithm of 3D principal component analysis (3D-PCA) for hyperspectral image analysis, exemplified in the exploratory analysis of plasma samples of healthy controls and ovarian cancer patients analysed by Raman microspectroscopy imaging.

2. Methods

2.1 Samples

Ten plasma samples of five healthy controls and five patients diagnosed with ovarian cancer were analysed by a Renishaw InVia Basis Raman spectrometer coupled to a confocal microscope (Renishaw plc, UK). All experiments were performed in accordance with Royal

Preston Hospital Guidelines, and approved by the ethics committee at Royal Preston Hospital UK (16/EE/0010). Informed consents were obtained from all human participants of this study. For analysis, 50 μL of plasma were deposited on aluminium covered glass slides and left to air-dry overnight. Samples were analysed with an acquisition area of 50 μm × 50 μm using 50× magnification and a laser power of 100% at 785 nm with 0.1 ms exposure time. Hyperspectral images were acquired via StreamHR™ imaging technique (high confocality mode) with a grid area of 57 × 57 pixels, resulting in 3,249 spectra in the range of ~780–1858 cm^{-1} generated for each image (1 cm^{-1} data spacing, 1,016 wavenumbers per spectrum). Thus, each sample's image was composed by a data array with dimension 57 × 57 × 1016.

2.2 Software

The Raman images were converted into suitable .txt files using Renishaw WiRE software; and processed using MATLAB R2014b (MathWorks, Inc., USA) with lab-made routines. All the samples' images were pre-processed by cosmic rays (spikes) removal and Savtizky-Golay smoothing (window of 9 points, 2nd order polynomial fitting). All data were mean-centred before further data analysis. A personal computer (16 GB of RAM memory, Intel® CoreTM i7 processor 2.81 GHz) was used for data processing.

2.3 3D-PCA

PCA is an exploratory analysis technique characterized by the decomposition of a given spectral data matrix **X** into a few number of principal component (PCs) responsible for the majority of the original data variance. Each PC is orthogonal to each other, being composed of scores (projections of the samples on the PC direction) and loadings (angle cosines of the variables projected on the PC direction).[26-28] The PCA decomposition of a spectral matrix **X** into scores (**T**), loadings (**P**) and residuals (**E**) takes the form:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \quad (1)$$

The scores **T** represent the variability on sample direction; the loadings **P** the variability on variables (e.g., wavenumbers) direction; and the residuals **E** the unexplained data after decomposition. **T** is used for assessing similarities/dissimilarities among the samples in an exploratory analysis context, whereas **P** contains the weights for each variable in the decomposition.

In 3D-PCA, a regular PCA decomposition (eqn.1) using nonlinear iterative partial least squares (NIPALS) algorithm is applied to each point (*i*,*j*) on the surface of the hyperspectral image data set. However, before PCA, each point in the image is transformed into a temporary 2D structure $\mathbf{X}_{ij}^{*}$ having *s* rows (samples) and *k* columns (variables) in order to keep the scores and loadings with their original meanings:

$$\mathbf{X}_{ij}^{*} = \mathbf{T}_{ij}\mathbf{P}_{ij}^{\mathrm{T}} + \mathbf{E}_{ij} \quad (2)$$

The number of PCs is selected based on the singular values obtained by singular value decomposition (SVD)[26] of the hyperspectral imaging, in a similar manner as described by Morais and Lima for florescence data.[13] After the number of PCs is selected, the scores $\mathbf{T}_{ij}$ and loadings $\mathbf{P}_{ij}$ are combined for all points (*i*,*j*) and separated for each PC. Hence, new three-dimensional arrays $\underline{\mathbf{T}}_{c}$ ($s \times n \times m$) and $\underline{\mathbf{P}}_{c}$ ($k \times n \times m$) are created for each PC, *c*. Figure 1 illustrate the 3D-PCA graphically.

[Insert Figure 1 here]

3. Results and Discussion

Ten plasma samples (5 health controls and 5 of patients diagnosed with ovarian cancer) were analysed by Raman microspectroscopy imaging. Their hyperspectral

images were generated with dimension of 57 × 57 × 1016, accounting 3,300,984 data points for each sample. The hyperspectral images for healthy controls and ovarian cancer samples are depicted in Figure 2 and Figure 3, respectively. Notably, each image presents distinct visual features, characterized by physical differences, such as dents and surface anomalies, of the samples analysed. However, chemically they should be grouped into at least two clusters (healthy vs. cancer).

[Insert Figure 2 here]

[Insert Figure 3 here]

The images were acquired in the spectral range of ~780–1858 cm^{-1}, which includes the fingerprint region; therefore, encompassing Raman signals of the major biochemical molecules present in the samples.[29] 3D-PCA was applied to the pre-processed images using only 2 PCs (34.23% cumulative explained variance) (Table 1). The 3D-PCA took approximately 1 min to run the entire data set, which accounted to more than 33 million of data points (10 images × 3,300,984 data points/image), using a standard personal computer. The 3D-PCA scores on PC1 and PC2 are shown in Figure 4.

[Insert Table 1 here]

[Insert Figure 4 here]

The scores on PC1 and PC2 across the x-axis (Figure 4A and 4B, respectively) show a separation tendency between healthy controls and ovarian cancer patients. However, a across the y-axis, the scores on both PC1 and PC2 are very noisy (Figure 4C and 4D, respectively); although, a separation pattern is observed on the scores on PC2 (Figure 4D). Combining the average scores on PC1 and PC2, the PC1 vs. PC2 scores plot (Figure 4E) shows a clear formation of two clusters separated along both PC1 and PC2. Healthy control patients are

located in the bottom-right side of the graph, while ovarian cancer patients on the upper-left side. Only one ovarian cancer sample is within the healthy control cluster. Figure 5 shows the boxplots for comparing the 3D-PCA scores individually along the axis and averaged. In all cases, statistical difference between healthy controls and ovarian cancer patients were observed at a 95% confidence level ($p < 0.05$): $p \approx 10^{-25}$ for scores on PC1 across *x*-axis (Figure 5A); $p \approx 10^{-27}$ for scores on PC2 across *x*-axis (Figure 5B); $p \approx 10^{-46}$ for scores on PC1 across *y*-axis (Figure 5C); $p \approx 10^{-100}$ for scores on PC2 across y-axis (Figure 5D); $p \approx 0.004$ for average scores on PC1 (Figure 5E); and $p \approx 0.002$ for average scores on PC2 (Figure 5F).

[Insert Figure 5 here]

The loadings profiles show larger coefficients around the Raman shift at 1400 cm^{-1} for PC1 (Figure 6A), a region containing N-H in-plane deformation and (C=O)-O- stretching in amino acids; and at ~1800 cm^{-1} and ~825 cm^{-1} representing C=O stretching in lipids and O-P-O stretching vibration in DNA, respectively.[30] Vibrations around 820 cm^{-1} and 1400 cm^{-1} have been reported as protein biomarkers for cervical tumours.[30,31]

[Insert Figure 6 here]

The fast data processing and clear scores segregation between healthy controls and ovarian cancer patients depicts the power of 3D-PCA as an exploratory analysis method for assessing between-samples differences in hyperspectral images. Even being an unsupervised method, statistical differences were found at a 95% confidence level between the 3D-PCA scores of the two different classes, indicating its potential usage towards classification applications. However, to build proper classification models in this case, a large cohort should be analysed by means of supervised

classification techniques, which can be easily adapted to 3D-PCA by employing discriminant analysis techniques[32] or support vector machines[33] to the 3D-PCA scores.

4. Conclusion

This paper reports a new 3D-PCA algorithm applied for exploratory analysis of plasma samples of healthy controls and ovarian cancer patients. Ten samples (5 healthy controls and 5 ovarian cancer) were analysed by Raman miscrospectrocopy imaging in the region of ~780–1858 cm^{-1}, generating data tensors with size of 57 × 57 × 1016 data points. 3D-PCA was applied to the whole dataset, generating scores showing clear differences between the two classes on both PC1 and PC2; and the loadings profiles on these components indicate that the main biomarker contributing for class differentiation are amino acids, lipids and DNA. 3D-PCA provided fast exploratory analysis for hyperspectral data, having potential for future applications in other types of spectrochemical imaging techniques.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

CLMM would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Brazil (grant 88881.128982/2016-01) for financial support.

References

1 D. Porro-Muñoz, R. P. W. Duin, I. Talavera and M. Orozco-Alzate, *Signal Processing*, 2011, **91**, 2520–2529.

2 B. Aguayo, S. J. Blackband, J. Schoeniger, M. A. Mattingly and M. Hintermann, *Nature*, 1986, **332**, 190.

3 A. R. Buchberger, K. DeLaney, J. Johnson and L. Li, *Anal. Chem.*, 2018, **90**, 240–265.

4 H. Abramczyk and B. Brozek-Pluska, *Chem. Rev.*, 2013, **113**, 5766–5781.

5 S. Türker-Kaya and C. W. Huck, *Molecules*, 2017, **22**, 168.

6 J. M. Amigo, I. Martí and A. Gowen, Chapter 9 - Hyperspectral Imaging and Chemometrics: A Perfect Combination for the Analysis of Food Structure, Composition and Quality. In: F. Marini (Ed.), *Chemometrics in Food Chemistry*, Elsevier, 2013, 343–370.

7 J. A. Fernández Pierna, P. Vermeulen, O. Amand, A. Tossens, P. Dardenne and V. Baeten, *Chemom. Intell. Lab. Syst.*, 2012, **117**, 233–239.

8 D. Eylenbosch, B. Bodson, V. Baeten and J. A. Fernández Pierna, *J. Chemom.*, 2017, **32**, e2982.

9 J. Ofner, K. A. Kamilli, E. Eitenberger, G. Friedbacher, B. Lendl, A. Held and H. Lohninger, *Anal. Chem.*, 2015, **87**, 9413–9420.

10 V. Olmos, M. Marro, P. Loza-Alvarez, D. Raldúa, E. Prats, F. Padrós, B. Piña, R. Tauler and A. de Juan, *J. Biophotonics.*, 2017, **11,** e201700089.

11 J. M. Amigo, H. Babamoradi and S. Elcoroaristizabal, *Anal. Chim. Acta.*, 2015, **896**, 34–51.

12 N. Mobaraki and J. M. Amigo, *Chemom. Intell. Lab. Syst.*, 2018, **172**, 174–187.

13 C. L. M. Morais and K. M. G. Lima, *Chemom. Intell. Lab. Syst.*, 2017, **170**, 1–12.

14 A. C. da Silva, S. F. C. Soares, M. Insausti, R. K. H. Galvão, B. S. F. Band and M. C. U. de Araújo, *Anal. Chim. Acta.*, 2016, **938**, 53–62.

15 Y. Zhang, H. H. and W. Cai, *Curr. Pharm. Biotechnol.*, 2010, **11**, 654–661.

16 H. J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N. J. Fullwood, B. Gardner, P. L. Martin-Hirsch, M. J. Walsh, M. R. McAinsh, N. Stone and F. L. Martin, *Nat. Protoc.*, 2016, **11**, 664.

17 G. P. S. Smith, C. M. McGoverin, S. J. Fraser and K. C. Gordon, *Adv. Drug Deliv. Rev.*, 2015, **89**, 21–41.

18 Z. Tian, N. Bing, L. Xie, L. Wang and H. Yuan, *2011 Third Int. Conf. Meas. Technol. Mechatronics Autom.*, 2011, 943–947.

19 T. Yaseen, D. -W. Sun and J. -H. Cheng, *Trends Food Sci. Technol.*, 2017, **62**, 177–189.

20 S. Lohumi, M. S. Kim, J. Qin and B. -K. Cho, *TrAC Trends Anal. Chem.*, 2017, **93**, 183–198.

21 H. Lui, J. Zhao, D. McLean and H. Zeng, *Cancer Res.*, 2012, **72**, 2491–2500.

22 M. Diem, A. Mazur, K. Lenau, J. Schubert, B. Bird, M. Miljković, C. Krafft and J. Popp, *J. Biophotonics.*, 2013, **6**, 855–886.

23 M. Kirsch, G. Schackert, R. Salzer and C. Krafft, *Anal. Bioanal. Chem.*, 2010, **398**, 1707–1713.

24 M. Paraskevaidi, C. L. M. Morais, K. M. G. Lima, K. M. Ashton, H. F. Stringfellow, P. L. Martin-Hirsch and F. L. Martin, *Analyst*, 2018, **143**, 3156–3163.

25 G. C. Jayson, E. C. Kohn, H. C. Kitchener and J. A. Ledermann, *Lancet*, 2014, **384**, 1376–1388.

26 R. Bro and A. K. Smilde, *Anal. Methods*, 2014, **6**, 2812–2831.

27 P. Geladi and B. R. Kowalski, *Anal. Chim. Acta.*, 1986, **185**, 1–17.

28 M. C. D. Santos, C. L. M. Morais, Y. M. Nascimento, J. M. G. Araujo and K. M. G. Lima, *TrAC Trends Anal. Chem.*, 2017, **97**, 244–256.

29 J. G. Kelly, J. Trevisan, A. D. Scott, P. L. Carmichael, H. M. Pollock, P. L. Martin-Hirsch and F. L. Martin, *J. Proteome Res.*, 2011, **10**, 1437–1448.

30 Z. Movasaghi, S. Rehman and I. U. Rehman, *Appl. Spectrosc. Rev.*, 2007, **42**, 493–541.

31 U. Utzinger, D. L. Heintzelman, A. Mahadevan-Jansen, A. Malpica, M. Follen and R. Richards-Kortum, *Appl. Spectrosc.*, 2001, **55**, 955–959.

32 L. F. S. Siqueira, R. F. Araújo Júnior, A. A. de Araújo, C. L. M. Morais and K. M. G. Lima, *Chemom. Intell. Lab. Syst.*, 2017, **162**, 123–129.

33 C. Cortes and V. Vapnik, *Mach. Learn.*, 1995, **20**, 273–297.

Captions for Figures

Figure 1: Illustration of data processing using 3D-PCA. *d* represents the z-axis coordinate dimension with size of *k* (number of wavenumbers) × *s* (number of images); *n* the number of pixels in the x-axis coordinate; *m* the number of pixels in the y-axis coordinate; and *c* the number of principal components (PCs).

Figure 2: Raman hyperspectral images of healthy control samples.

Figure 3: Raman hyperspectral images of ovarian cancer samples.

Figure 4: 3D-PCA scores plot. (A) Scores on PC1 and (B) PC2 across *x*-axis; (C) scores on PC1 and (D) PC2 across *y*-axis; (E) average scores on PC1 *versus* PC2. HC: healthy controls (in blue); OC: ovarian cancer (in red).

Figure 5: Boxplots for 3D-PCA scores. (A) Scores on PC1 across *x*-axis ($p = 1.903\times10^{-25}$); B) scores on PC2 across *x*-axis ($p = 4.884\times10^{-27}$); C) scores on PC1 across *y*-axis (6.118×10^{-46}); (D) scores on PC2 across *y*-axis (6.239×10^{-100}); (E) average scores on PC1 ($p = 0.004$); (F) average scores on PC2 ($p = 0.002$). HC: healthy controls; OC: ovarian cancer.

Figure 6: 3D-PCA loadings. (A) Loadings on PC1; (B) loadings on PC2.

Tables

Table 1. Explained variance for 3D-PCA.

PC	Explained variance (%)	Cumulative explained variance (%)
1	20.78	20.78
2	13.45	34.23
3	11.34	45.57
4	10.32	55.88
5	9.61	65.50
6	9.12	74.62
7	8.73	83.34
8	8.46	91.80
9	8.20	100

Figure 1

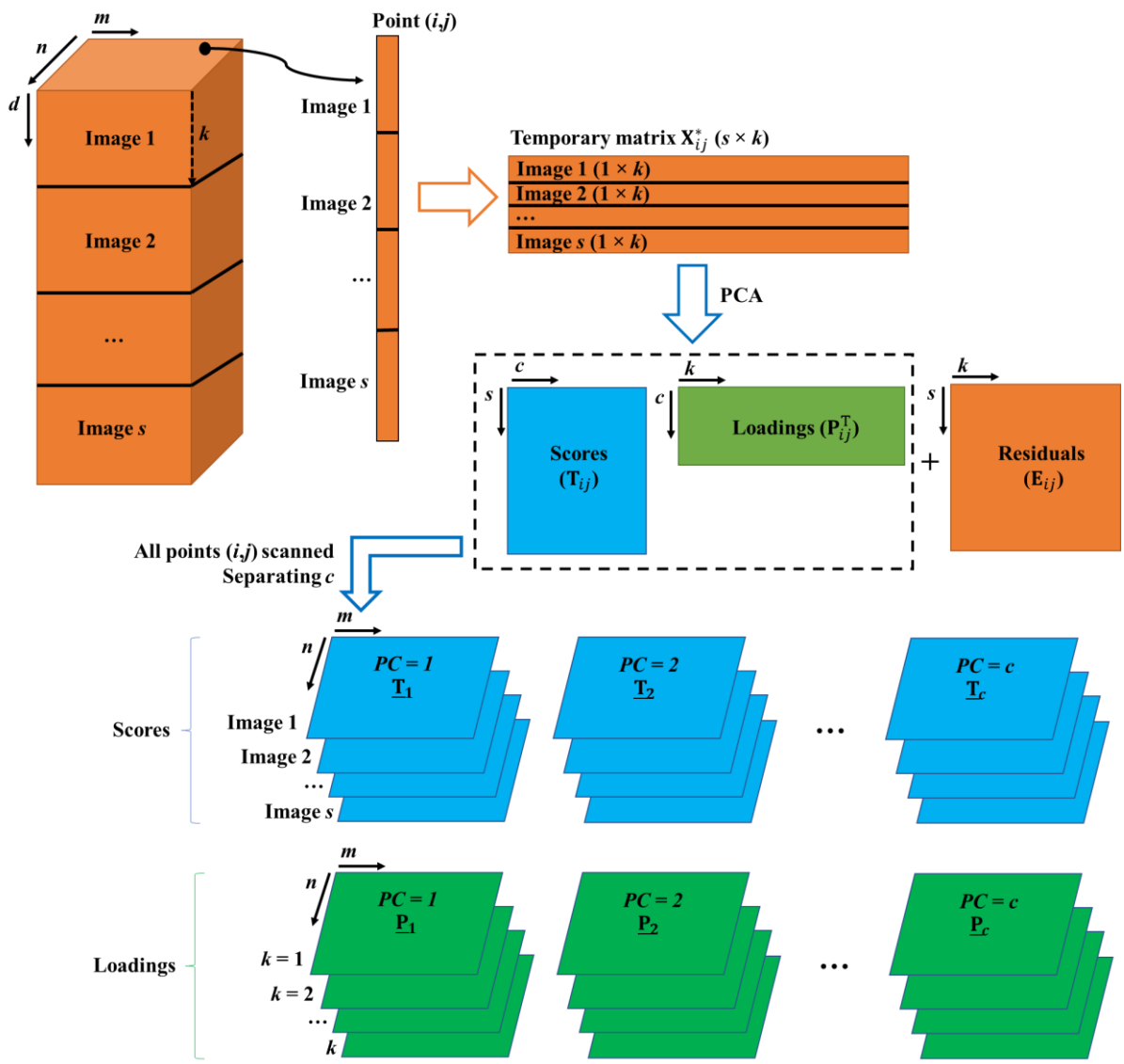

m
n
d
Image 1
k
Image 2
...
Image s
Point (i,j)
Image 1
Image 2
...
Image s
Temporary matrix X^*_{ij} (s × k)
Image 1 (1 × k)
Image 2 (1 × k)
...
Image s (1 × k)
PCA
c
s
Scores (T_{ij})
k
c
Loadings (P^T_{ij})
+
k
s
Residuals (E_{ij})
All points (i,j) scanned
Separating c
m
n
Scores
Image 1
Image 2
...
Image s
PC = 1
T_1
PC = 2
T_2
...
PC = c
T_c
m
n
Loadings
k = 1
k = 2
...
k
PC = 1
P_1
PC = 2
P_2
...
PC = c
P_c

Figure 2

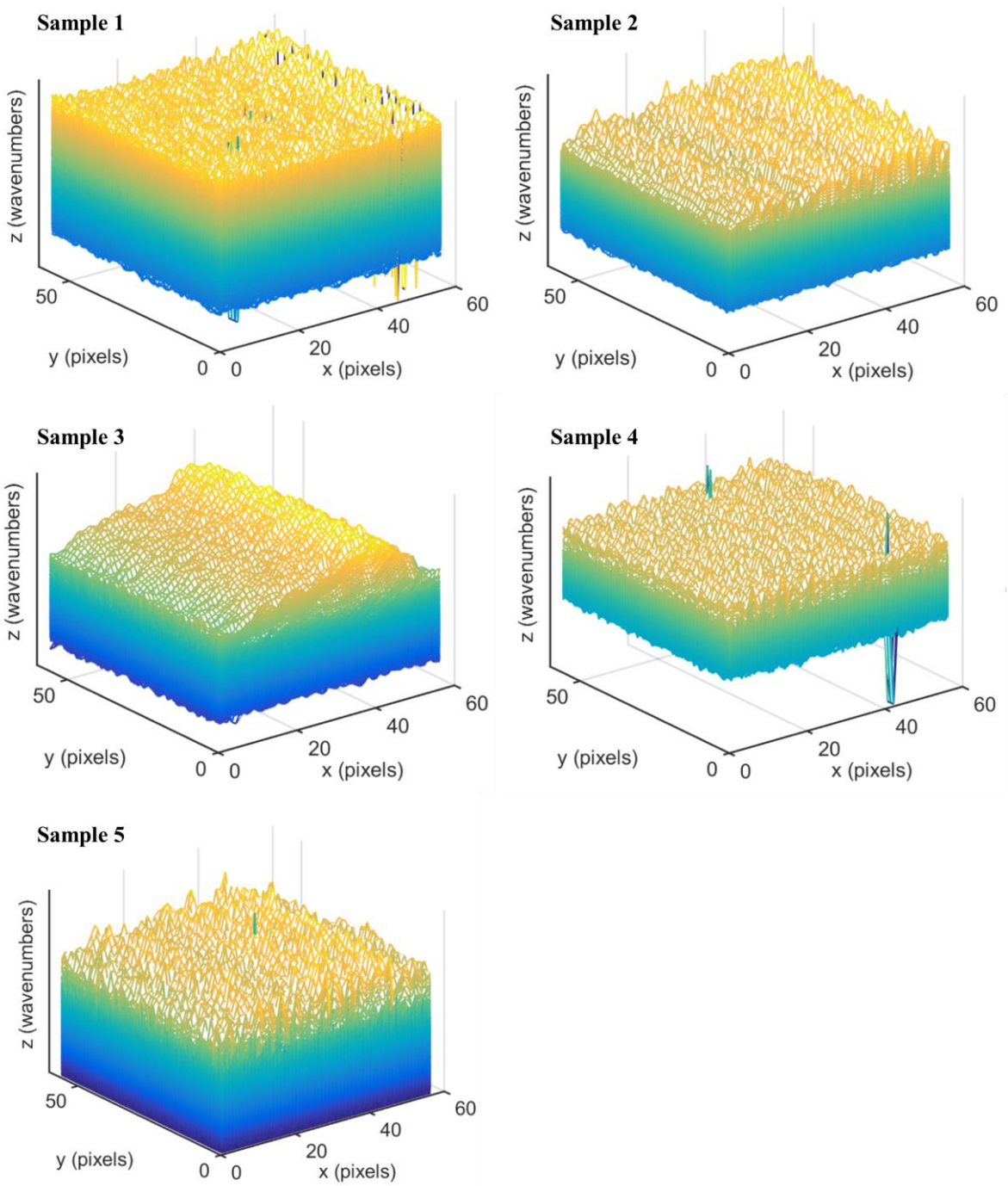

Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
z (wavenumbers)
y (pixels)
x (pixels)
0
50
0
20
40
60

Figure 3

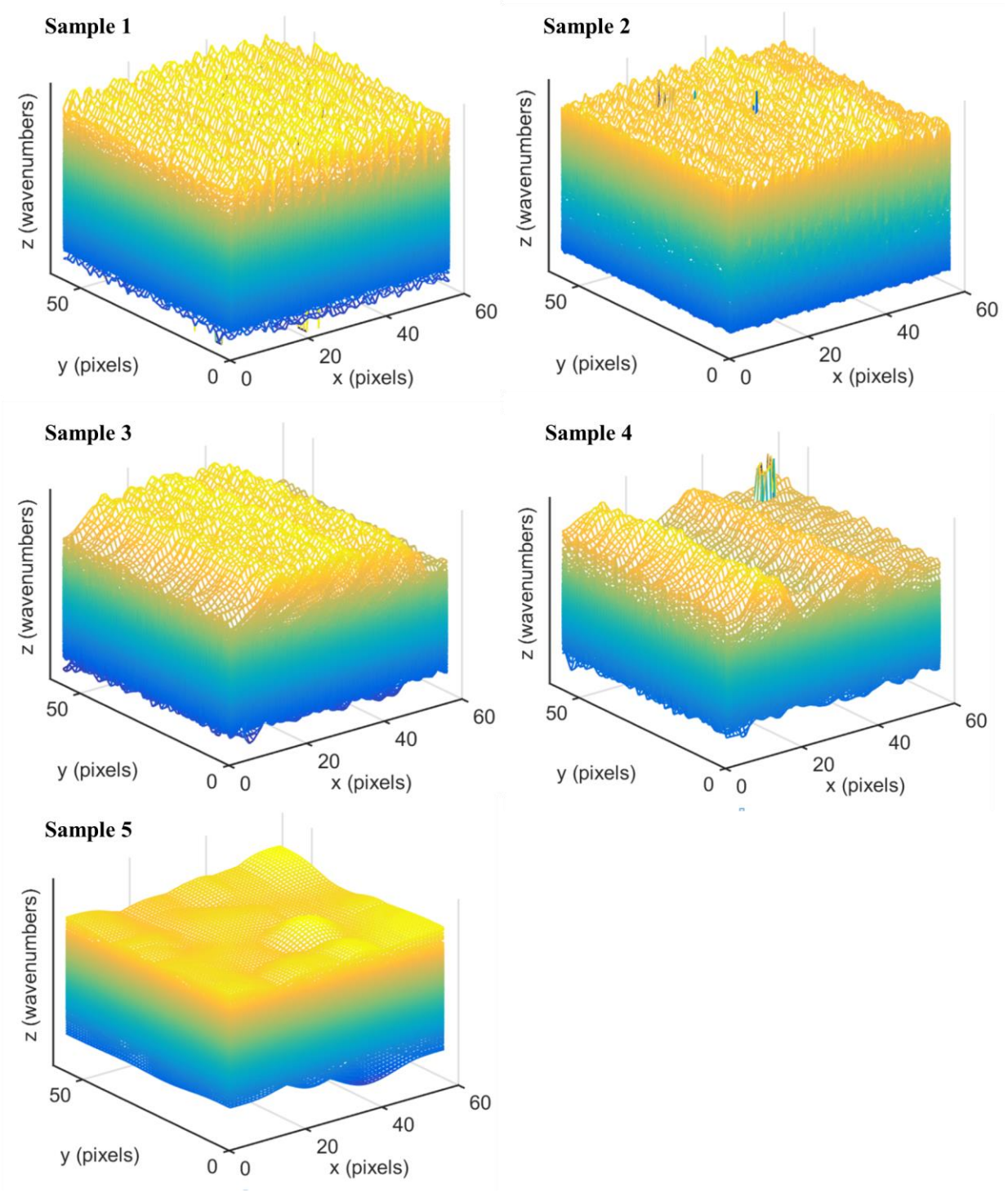

Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
z (wavenumbers)
y (pixels)
x (pixels)
0
20
40
50
60

Figure 4

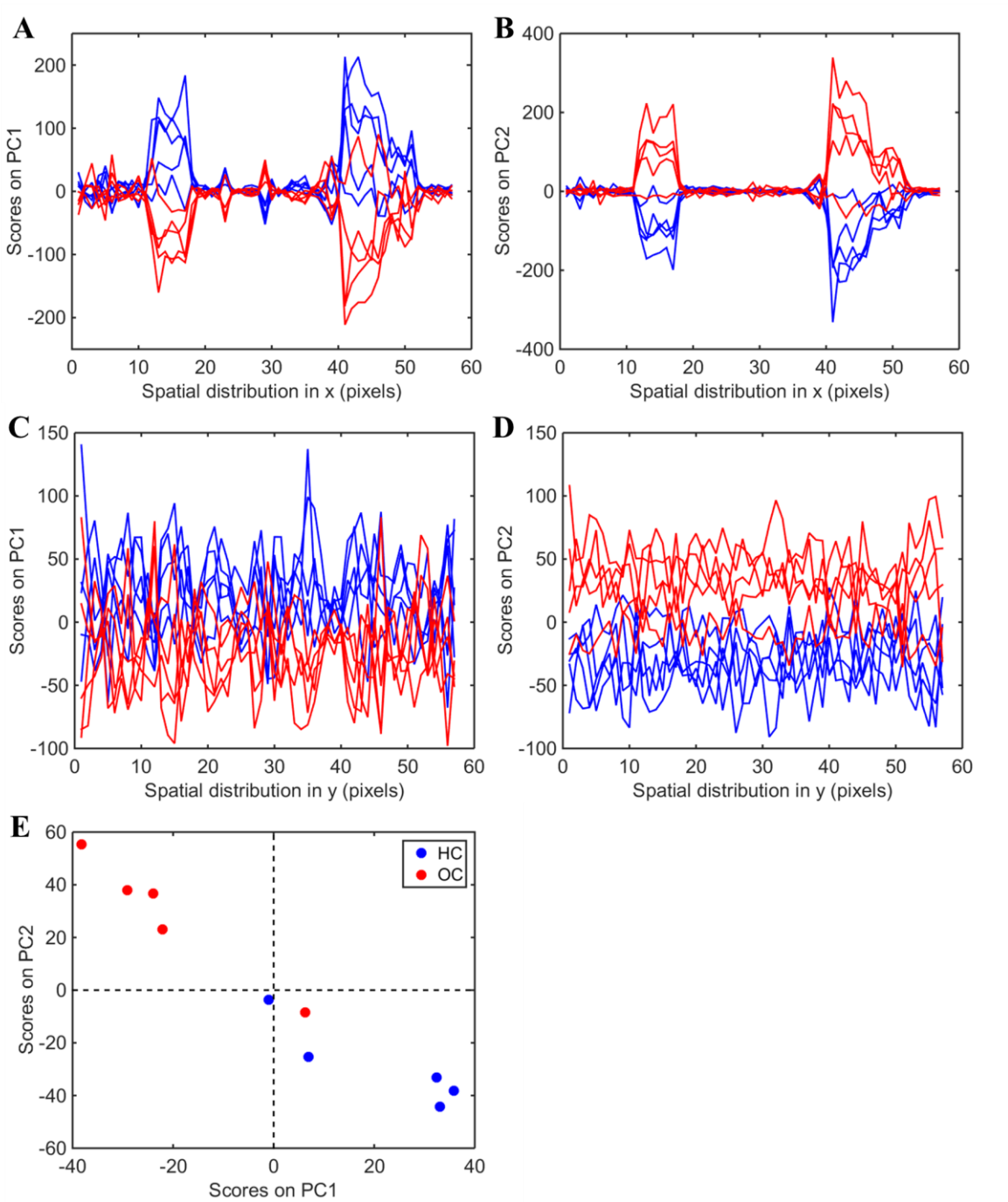

A
Scores on PC1
Spatial distribution in x (pixels)
B
Scores on PC2
Spatial distribution in x (pixels)
C
Scores on PC1
Spatial distribution in y (pixels)
D
Scores on PC2
Spatial distribution in y (pixels)
E
HC
OC
Scores on PC2
Scores on PC1

Figure 5

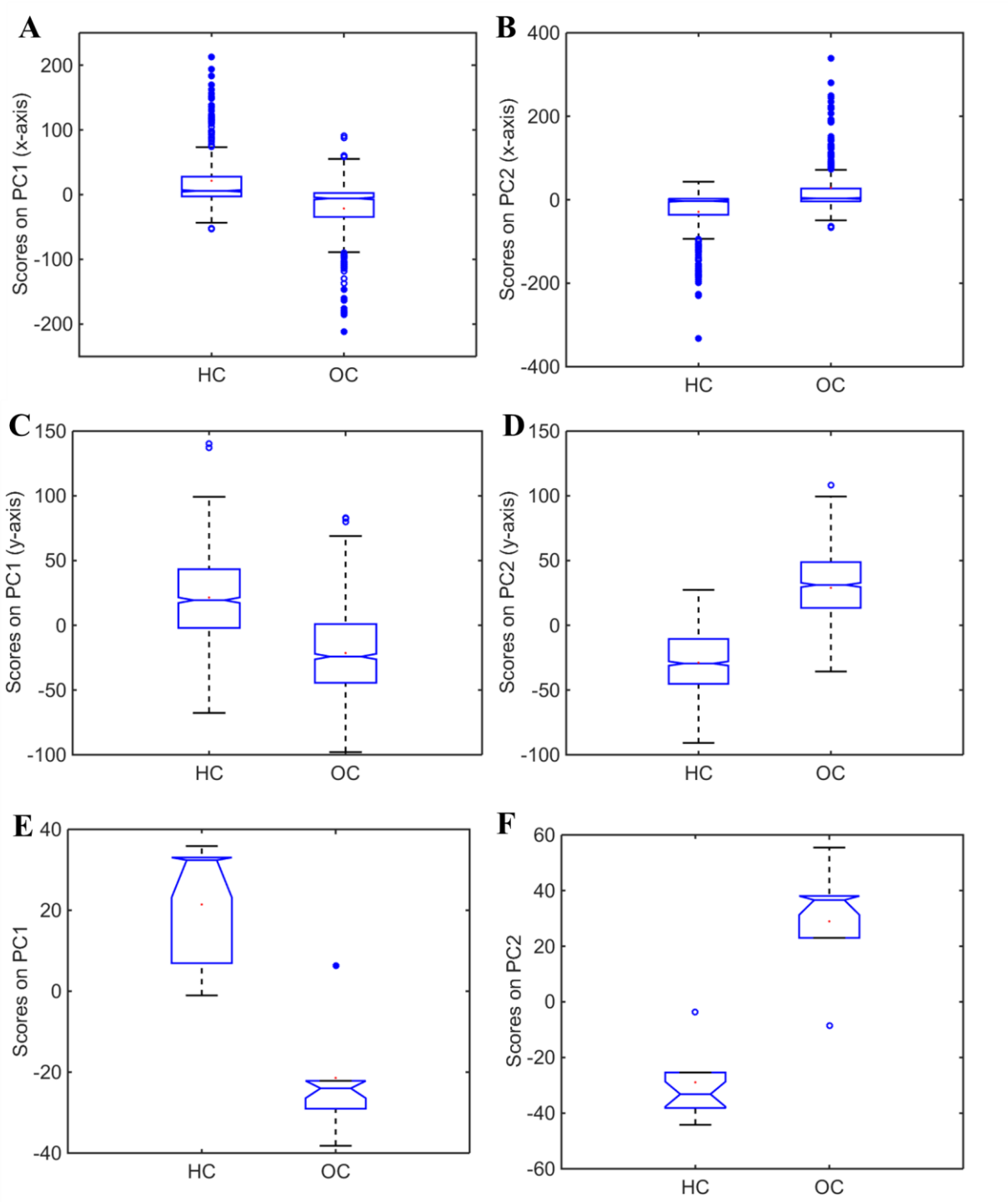
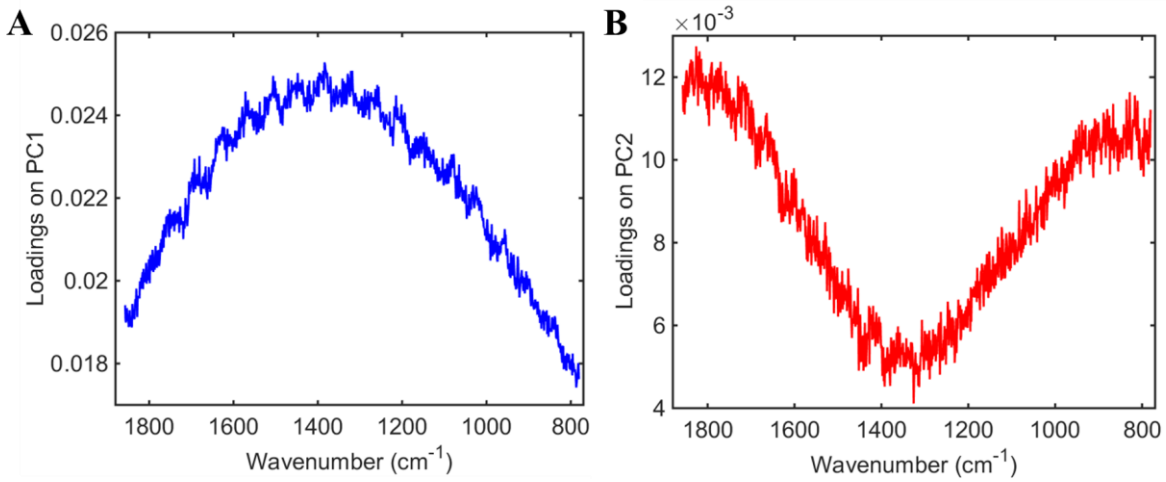

A
Scores on PC1 (x-axis)
200
100
0
-100
-200
HC
OC
B
Scores on PC2 (x-axis)
400
200
0
-200
-400
HC
OC
C
Scores on PC1 (y-axis)
150
100
50
0
-50
-100
HC
OC
D
Scores on PC2 (y-axis)
150
100
50
0
-50
-100
HC
OC
E
Scores on PC1
40
20
0
-20
-40
HC
OC
F
Scores on PC2
60
40
20
0
-20
-40
-60
HC
OC

Figure 6


A
0.026
0.024
0.022
0.02
0.018
Loadings on PC1
1800
1600
1400
1200
1000
800
Wavenumber (cm-1)
B
×10-3
12
10
8
6
4
Loadings on PC2
1800
1600
1400
1200
1000
800
Wavenumber (cm-1)