COMPARING THREE PCA-BASED METHODS FOR THE 3D VISUALIZATION OF IMAGING SPECTROSCOPY DATA

Alexander Broersen, Robert van Liere
CWI
Kruislaan 413, Amsterdam
The Netherlands
{a.broersen, robert.van.liere}@cwi.nl

Ron M.A. Heeren
FOM Institute for Atomic and Molecular Physics
Kruislaan 407, Amsterdam
The Netherlands
r.heeren@amolf.nl

ABSTRACT
In this paper we compare the quality of three different principle component analysis (PCA) based methods to generate transfer functions for the 3D visualization of imaging spectroscopy data. We discuss three criteria for judging the quality of features in these visualizations. These criteria are used to interpret visualizations of features in the brain of the snail Lymnaea Stagnalis. We show that the PCA method that uses model additional information, clearly results in superior visualizations.

KEYWORDS
Image processing and analysis, pattern analysis and recognition, transfer function, imaging spectroscopy, principal component analysis and multidimensional.

1. Introduction

Direct volume rendering is a well-known method for the visualization of three-dimensional volumetric datasets. In most volumetric datasets, each voxel contains a scalar value that represents the density of a material on that location. For visualization, the transfer function maps a color and opacity value to a scalar value. A volume renderer can draw the voxel data using the map specified in the transfer function. The challenge in designing transfer functions is to identify which structural properties are important and which relevant features in the data should be highlighted.

Imaging spectroscopy can be used to scan the structure of chemical elements on material surfaces. In contrast to a volume consisting of 3D points of scalar values, a spectral dataset consists of two spatial dimensions and mass to charge ratio in the third dimension. Each scalar value in the volume is interpreted as the intensity on a mass to charge ratio at a 2D position on the surface of a material. Material scientists often refer to a spectral volumetric dataset as a *multi-spectral data-cube*.

Since chemical elements have a unique and known spectral profile, scientists can use spectroscopy to investigate which elements are present on the surface of a material if their spectral profile can be extracted from the data-cube. Unfortunately, extracting a spectral profile is a difficult task. First, the intensity at each point in the volume consists of contributions of the mass to charge ratios of neighboring chemical elements at that position on the surface; i.e. the measured intensity at a voxel is a linear combination of mass to charge peaks. A robust extraction method will be needed to factor the linear combination of mass to charge ratios into the mass to charge of each chemical element. Second, spectra characterize themselves by different levels of scale in which peaks in the spectral profile can vary in order of magnitude. For example, consider Figure 1a. The sum of all spectral profiles in the data-cube is plotted, with on the x-axis the mass to charge ratio and on the y-axis the measured intensity. Figure 1b shows the spatial distribution of spectral peaks. The value of a pixel represents the sum of intensities at each mass to charge ratio at each position on the surface of the material. A color map is used to map intensity to a color. Figure 1 is an example of how scientists use two side-by-side views to analyze the data in the data-cube. One view is the spectral view; it shows the spectral profile at all mass to charge ratios. The second view is a spatial view; it shows the summation of the spectral profile at each position on the surface of the scanned material.

It is our goal to create a data analysis environment with one integrated 3D-view to gain insight into the spatial distribution of the spectra in the data-cube. In a previous paper [1] we have introduced a new application of the principal component analysis (PCA) to generate multidimensional transfer functions. PCA separates peaks in different uncorrelated spectral components and can simultaneously identify spatial patterns. The direct linkage between the resulting spectral and spatial components characterizes the approach. Figure 1c shows the first principal component linking spectral and spatial features in a multi-spectral data-cube.

In this paper we compare the quality of three different PCA methods to generate transfer functions for the 3D visualization of data-cubes. How can the quality of the 3D visualizations be compared? We have identified three important criteria for this. First, spatially correlated spectral features in the visualization should be

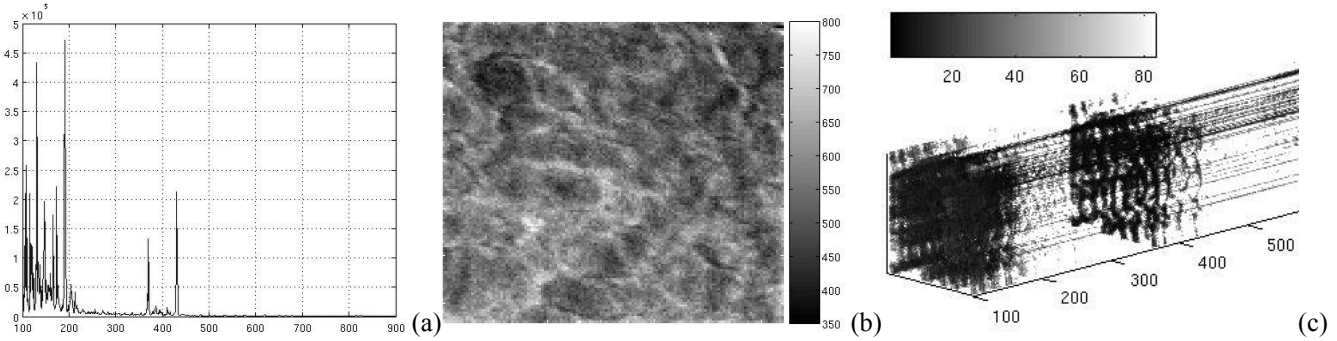

Figure 1: (a) A plot of the summation of all spectra in the data-cube with (b) the spatial distribution of spectral profiles with (c) the resulting the transfer function generated from the first principal component used to highlight spectral and spatial features.

distinguishable. As a rule of thumb, the higher the contrast between features the higher the quality of the visualization is. Second, these features should also be recognizable as bio-molecules in complex surfaces such as cells and tissue samples in the data-cube. For example, do these features represent a cell wall or a tissue, etc? If so, how well are the recognized spectral features correlated? Finally, are the spectral and spatial features distinct in different regions in the image? These criteria will be used in Section 4 to qualitatively compare the presented methods.

In the next section, we briefly discuss the techniques used by the three methods: PCA, PCA with VARIMAX rotation and PARAFAC. In section 3, we present a quantitative comparison of the methods. We do this by comparing the results of the methods with an a-priori known spectral data cube; i.e. a *ground truth*. Finally, in section 4, we discuss a qualitative comparison of the methods with a real world application.

2. Method

We briefly describe the three different PCA-based methods for generating transfer functions. Due to space constraints, we will not provide a detailed explanation of the mathematics and their implementation. We refer to the literature for a more thorough explanation of the characteristics of the methods. The methods we use are:

1. **PCA** In a previous paper [1], we extract spatial and spectral components using the well-known PCA method [2 and 3]. In our approach, we unfold a pre-processed λ by *x* by *y* data-cube in such a way that a 2D λ by *x*×*y* matrix *X* is constructed. The standard PCA model is used to compute a sorted list of principle components in an orthonormal matrix *P* (see Equation 1) using eigenvector decomposition.

$$Y = P \cdot X^T \qquad (1)$$

The first principle components in *P* describe those spatial regions in the data-cube with the greatest spectral variance. The original data-cube is projected using the principle components as bases results in a matrix *Y* with the spatial (Y_{images}) or spectral ($Y_{spectra}$) *score vectors.* Both these matrixes are extracted and combined to construct the transfer function.
One problem with this approach is the minimal contrast between different spectral peaks and spatial components. This results in less distinctive regions in the resulting volume-rendered data-cubes. Another problem is that the extracted score vectors can be negative, while it is known that all spectra are always positive.

2. **PCA+VARIMAX** There are many variations for two-way bilinear decomposition similar to PCA [4]. One approach is to rotate the resulting principal components to obtain a better fit on the data without affecting the decomposition using the rotational ambiguity of PCA. The VARIMAX rotation proposed by Kaiser [5] is one of the most popular criteria for rotation. It can be applied as a post-processing step on extracted principal components. VARIMAX searches for an orthogonal rotation of the original components in so that the variance of the squared principal components is maximized. For each k^{th} principal component the objective function of Equation 2 is computed.

$$s_k^2 = \frac{f\sum_{i=1}^{\lambda}\left(\frac{x_{if}^2}{h_i^2}\right)^2 - \left(\sum_{i=1}^{\lambda}\frac{x_{if}^2}{h_i^2}\right)^2}{f^2} \qquad (2)$$

Where *f* is the number of principal components, λ is the number of spectral variables, x_{if} is the loading of spectral variable *i* on component *f*. and h_j^2 is the communality of the i^{th} spectral variable in *P* as defined in Equation 3.

$$h_i^2 = \sum_{j=1}^{f} P_{ji}^2 \quad (3)$$

The overall variance V in Equation 4 is being maximized using Equation 2 until the increase of V drops below a certain threshold (e.g. 10^{-6} in our examples).

$$V = \sum_{k=1}^{f} s_k^2 \quad (4)$$

In theory, the VARIMAX method can improve the contrast between spectral peaks, since rotating the principle component bases will result in sharper gradients in adjacent spectral peaks.

3. **PARAFAC** Another variation of PCA-like methods is the PARAFAC (PARAllel FACtors analysis) model of Harshman [6] which exact model was independently proposed by Carroll and Chang [7] as the CANDECOMP (CANonical DECOMPosition). Kiers [8] has shown that PARAFAC can be considered a constrained version of the two-way PCA. PARAFAC uses fewer degrees of freedom to fit the data on a simple model. We used the implementation of the algorithm described by Bro [9] to only put a non-negativity constraint on the decomposition in P and Y in Equation 1 t to improve interpretation of the scores. The convergence criterion for the algorithm was a relative change in fit of less than 10^{-6} in our examples.
   The advantage of the PARAFAC method as we use it is that the score vectors will always be positive. The resulting transfer function is therefore easier to interpret as only the most positive values in the score vectors instead of also the most negative ones have to be included in the opacity map.

We used similar PCA-based methods to be able to create a transfer function (as in [1]) that makes use of the link between spectral and spatial score vectors in Equation 5.

$$Y_{images} = \begin{bmatrix} i_1 \\ i_2 \\ \vdots \\ i_\lambda \end{bmatrix} \text{ and } Y_{spectra} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_{xy} \end{bmatrix} \quad (5)$$

Y_{images} is a λ by $x \times y$ size matrix and $Y_{spectra}$ a $x \times y$ by λ matrix. The transfer function that acts as a 3D opacity map is generated by adding each resulting spectral score vector in one dimension with the corresponding spatial score vector in the other dimension, for instance i_1 and s_1 result in an opacity map O_1 as shown in Equation 6.

$$o^1 = \begin{bmatrix} i_1 \\ i_1 \\ \vdots \\ i_1 \end{bmatrix} + \begin{bmatrix} s_1 \\ s_1 \\ \vdots \\ s_1 \end{bmatrix}^T \quad (6)$$

The resulting 3D array has the same size as the original unfolded data-cube and acts as an opacity map, with the most positive and most negative values assigned to the highest values for opacity. The 3D points with the highest or lowest contribution to the whole data-cube will be the most opaque in the resulting opacity map. Similarly, the opacity maps (O^2, O^3) of the combined second, third, etc. score vectors can be generated.

3. Quantitative comparison

A synthetic multispectral data-cube was created to be able to make a quantitative comparison between the three decomposition methods. Three different spectra including some overlap in the peaks were used to have a variety in the spectral and spatial dimensions. After this some different levels of Gaussian noise (mean: 0.000 and with a variance: between 0.0001 and 0.0500) were added to the whole data-cube to make it more realistic.

The resulting spectral score vectors of the three methods are compared with the original spectra, our ground truth. For a quantitative analysis we use a widely used measure of error, the *Root Mean Squared Error* (RMSE, ε) similarly used in other analyses of correlated spectral data [10]. The absolute values of the resulting spectral component are compared with the synthetic one according to Equation 7. The number of spectra in the cube is represented by n and results in a ε for each method.

$$\varepsilon_{method} = \sqrt{\frac{\sum_{i=1}^{n} \left(|component_i| - synthetic_i \right)^2}{n}} \quad (7)$$

Each method is able to distinguish between the three different spectral components, while the other components clearly contain the added noise. An overview of ε of each method is shown in Table 1.

Method	PCA	PCA+ VARIMAX	PARAFAC
component1	0.0744	0.0813	0.0235
component2	0.0691	0.0581	0.0112
component3	0.0753	0.0629	0.0249
Total ε	**0.2190**	**0.2023**	**0.0597**

Table 1: The root mean squared error of the different components of each method.

This table clearly indicates that the PARAFAC decomposition results in the least amount of error. Also the VARIMAX rotation provides a better fit compared to the use of only a PCA without a rotational fit. To gain better insight of the influence of the added Gaussian noise, different levels of noise are introduced as shown in Table 2.

Method variance	PCA	PCA+ VARIMAX	PARAFAC
0.0001	0.2259	0.1983	0.0352
0.0010	0.2190	0.2023	0.0597
0.0100	0.2210	0.2038	0.1352
0.0500	0.2307	0.2219	0.1613

Table 2: The total root mean squared errors of each method with different levels of Gaussian noise.

Table 2 shows that even though noise levels are rising, the ε of the PARAFAC method still remains lower than the ε of the other two methods.

4. Qualitative comparison

Imaging mass spectrometry is a microscopic technique that is used to analyze the spatial organization of intact biomolecules in complex surfaces such as cells and tissue samples. It is particularly useful to directly visualize peptide and protein distributions in invertebrate or mammalian tissue. In the imaging MS data used here a 15 kV Indium primary ion beam is rastered over the surface of a cryosection of the cerebral ganglia of the freshwater snail Lymnaea Stagnalis. A data array of 256x256 x,y-coordinates, is generated with each position containing an entire mass spectrum. Each square pixel represents an area of approximately 500x500 nm Prior to the experiment the tissue surface has been covered with a thin layer of 2,5-dihydroxybenzoic (2,5-DHB) acid by electrospray deposition to enhance in the generation of intact biomolecular ions. The mass spectrometer used was a Time-of-Flight (ToF) mass spectrometer. High-resolution molecular ion maps have previously shown to provide insight in the spatial organization of various biomolecules in these brain sections [11]. Manual

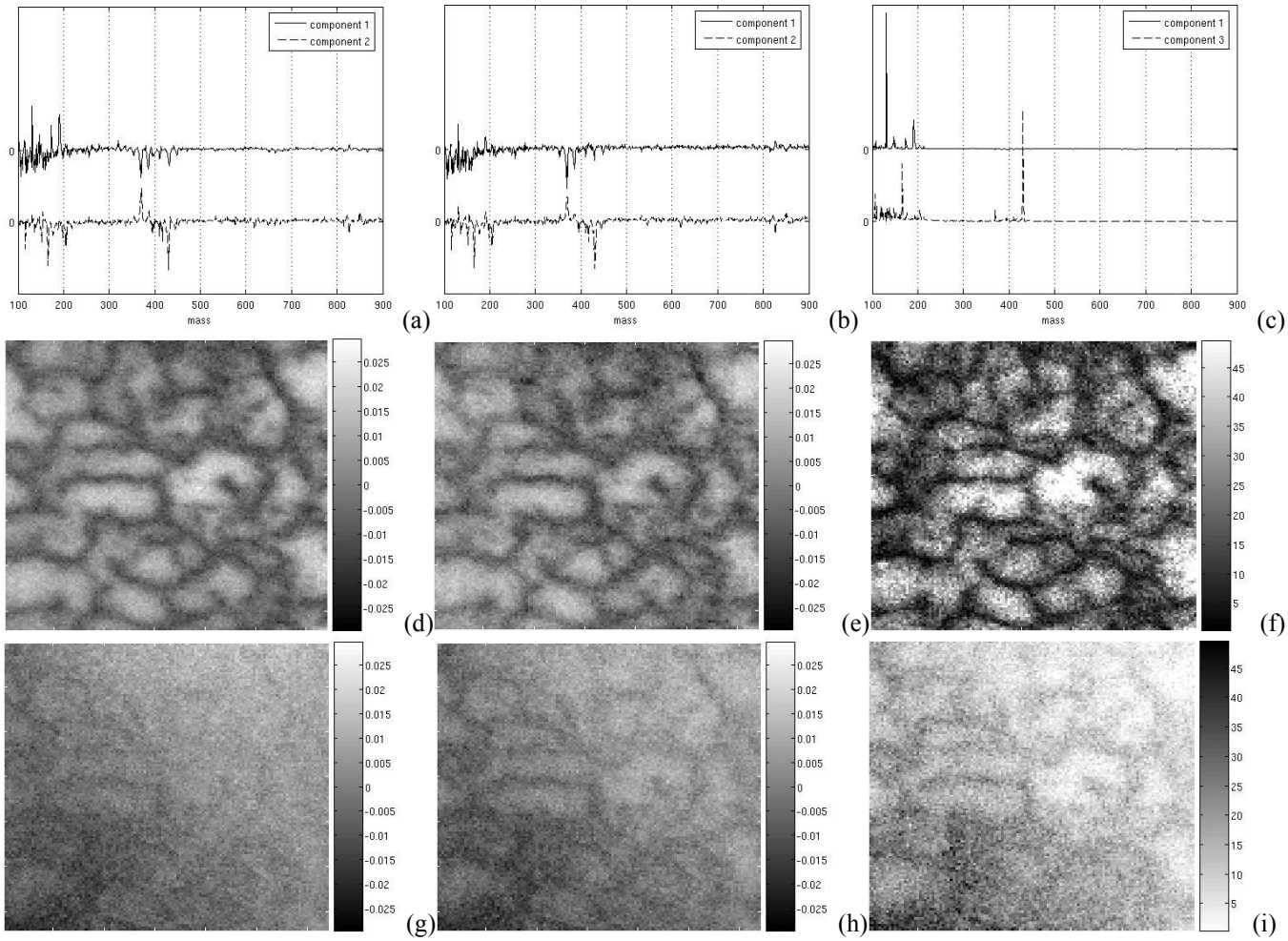

Figure 2: Two spectral (a) and two spatial (d-g) components derived using PCA. Two spectral (b) and two spatial (e-h) components derived with PCA and VARIMAX rotation. Two spectral (c) and two spatial (f-i) components derived with PARAFAC.

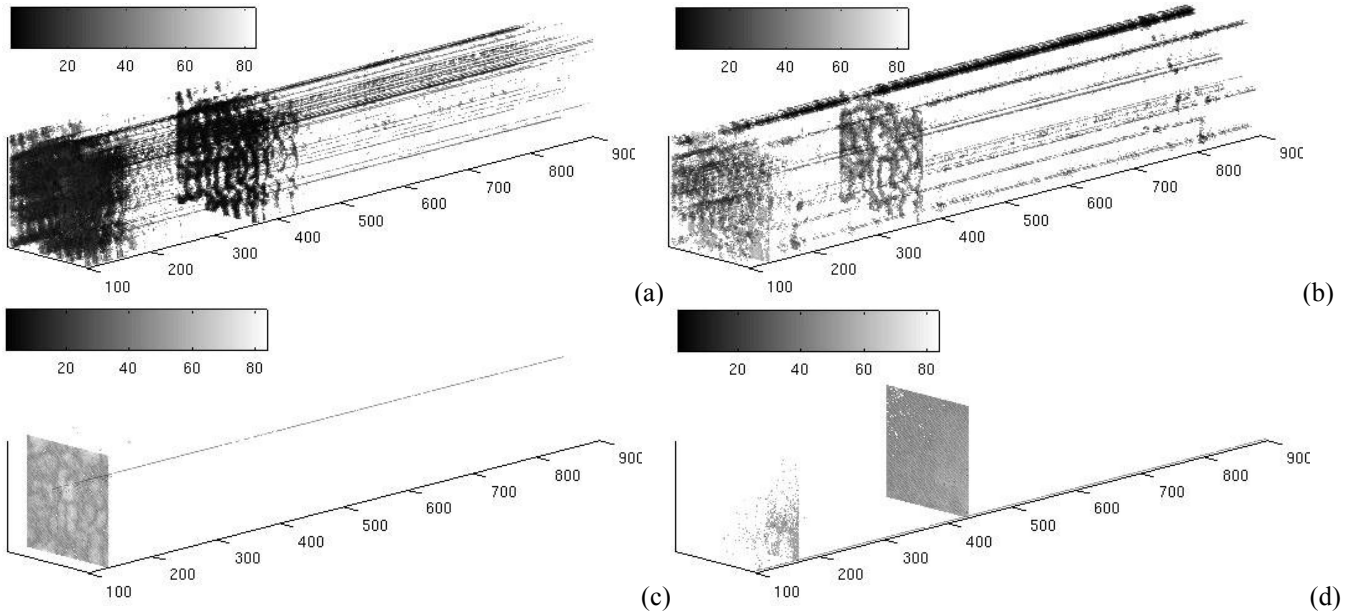

Figure 3: The resulting multispectral data-cubes with a transfer function from (a) PCA, (b) PCA with VARIMAX rotation and (c-d) two different components from PARAFAC.

interpretation of these types of datasets is a time consuming procedure where either for spectral peaks of interest image are created or the spectra of interesting spatial features in images are interrogated. In order to identify spatially correlated spectral data (often attributed to a specific compound) of statistical analysis tools are called for. Here, we qualitatively examine the results of the three different multivariate statistical analysis algorithms applied to a single MS image dataset of the brain of Lymnaea Stagnalis.

4.1 PCA

The reconstructed score vectors in Figure 2a display two separated components where the first component contains predominantly spectral features that are clearly correlated to the applied matrix 2,5-DHB as positive peaks. Intermixed with this compound negative spectral features of cholesterol (368/385) are also observed. As in regular spectra the matrix peaks usually constitute the base peak in the spectrum. This method seems to under represent the spectral intensities. The spatial features are distinctively related to the areas in between individual cells that seem to indicate a stronger matrix signal is found there.

The second component found is again a mixture of cholesterol, but now positive peaks and a peak at 425 m/z that previously has been attributed to APGWamide. It also contains some higher mass lipid molecules around m/z 815. The spatial features are barely recognizable. The multispectral data-cube in Figure 3a also shows that many different spectral features display a certain amount of spatial correlation. This makes it difficult to identify the individual features from these two principal components.

4.2 PCA with VARIMAX rotation

The second method, PCA with VARIMAX rotation shows similar spectral features in its components but judging from Figure 3b an improved spatial correlation is found. This is also obvious from the improvement in quality of the image in Figure 2h. A better feature contrast is found, but the individual components are not fully separated.

4.3 PARAFAC

The PARAFAC approach offers a spectral view that is more similar to the spectral view the mass spectrometrists are used to. In addition, the relative intensities and signal-to-noise ratio in the two components "spectra" are as would be expected from these types of measurements. More importantly a much better separation between the matrix-related peaks and the cellular peaks is obtained. This also results in better contrast in the feature images facilitating an easier localization of the compounds. The smaller peaks around 815 are not clearly visible on this scale, but are maybe incorporated in other component spectra. In Figure 4c and 4d it becomes clear that the compound separation and localization has significantly improved.

5. Conclusion

In this paper we have compared the quality of three different PCA-based methods to generate transfer functions for the 3D visualization of imaging spectroscopy data. For this we used the PCA, PCA with VARIMAX rotation and PARAFAC method. We compared the methods quantitatively and qualitatively.

For the quantitative comparison, we used a RMSE metric to compare the methods with *ground truth* spectra under various noise conditions. For the qualitative comparison, we used three criteria to judge the quality of features in the resulting visualizations. These criteria were applied to interpret the visualizations of features in the brain of the snail Lymnaea Stagnalis.

This study shows that the PARAFAC method is clearly superior to the other methods. PARAFAC results in features that are more clearly recognizable than the other two methods (see Figure 3). The reason for this is that PARAFAC uses some model information, while PCA does not. The VARIMAX rotation uses a post-processing fitting to maximize the variance of the components which results in more contrast rich images and spectra.

We learn from the synthetic data case that although the root mean squared error becomes larger with higher noise levels, the PARAFAC method still produces the most distinctive results. We expect that these trends are similar in the real life application. The implication is that more noisy samples will still result in good visualizations.

6. Acknowledgements

We thank Drs. A.F.M. Altelaar from the Institute for Atomic and Molecular Physics (AMOLF) who prepared and supplied the spectral dataset that was used in the presented example.
This work was carried out in the context of the Virtual Laboratory for e-Science project (www.vl-e.nl). This project is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ).

References:

[1] A. Broersen & R. van Liere, Transfer functions for imaging spectroscopy data using principal component analysis, *Proc. Eurographics / IEEE VGTC Symposium on Visualization*, Leeds, UK, 2005, 117-123.
[2] M. Wall, A. Rechtsteiner, L. Rocha, Singular value decomposition and principal component analysis, *A Practical Approach to Microarray Data Analysis*, 2003, 91-109.
[3] P. Lasch, W. Wäsche, W. McCarthy & D. Naumann, Imaging of human colon carcinoma thin sections by ft-ir microspectroscopy, *Infrared Spectroscopy: New Tool in Medicine* 3257, 1998, 187-197.
[4] M.E. Timmerman, *Component analysis of multisubject multivariate longitudinal data* (Groningen: University Library Groningen, 2001).
[5] H.F. Kaiser, The varimax criterion for analytic rotation in factor analysis, *Psychometrika, 23*, 1958, 187-200.
[6] R.A. Harshman, Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis. *UCLA Working Papers in Phonetics (16)*, 1970, 1-84.
[7] J.D. Carroll & J.J. Chang, Analysis of Individual Differences in Multidimensional scaling via an N-way generalization of Eckart-Young decomposition, *Psychometrika 35*, 1970, 283-319.
[8] H.A.L. Kiers, Hierarchical relations among three-way methods, *Psychometrika 56*(3), 1991, 449-470.
[9] R. Bro, PARAFAC, tutorial and applications, *Chemometrics and Intelligent Laboratory Systems, 38*(2), 1997, 149-171.
[10]I. Scarminio & M. Kubista, Analysis of Correlated Spectral Data, *Analytical Chemistry, 65*(4), 1993, 409-418.
[11]L.A. McDonnell, S.R. Piersma, A.F.M. Altelaar, T.H. Mize, P.D.E.M. Verhaert, J. van Minnen & R.M.A. Heeren, Matrix-enhanced secondary ion mass spectrometry imaging of brain tissue, *Journal of Mass Spectrometry 40*, 2005, 160-168.