Micron 52–53 (2013) 49–56

Contents lists available at ScienceDirect

Micron

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

Multivariate statistical analysis as a tool for the segmentation of 3D spectral data

G. Lucas*, P. Burdet, M. Cantoni, C. Hébert

École Polytechnique Fédérale de Lausanne (EPFL), Interdisciplinary Centre for Electron Microscopy (CIME), Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 17 May 2013
Received in revised form 26 July 2013
Accepted 19 August 2013

Keywords:
Multivariate statistical analysis
Segmentation
Spectral image processing
3D EDX

ABSTRACT

Acquisition of three-dimensional (3D) spectral data is nowadays common using many different microanalytical techniques. In order to proceed to the 3D reconstruction, data processing is necessary not only to deal with noisy acquisitions but also to segment the data in term of chemical composition. In this article, we demonstrate the value of multivariate statistical analysis (MSA) methods for this purpose, allowing fast and reliable results. Using scanning electron microscopy (SEM) and energy-dispersive X-ray spectroscopy (EDX) coupled with a focused ion beam (FIB), a stack of spectrum images have been acquired on a sample produced by laser welding of a nickel–titanium wire and a stainless steel wire presenting a complex microstructure. These data have been analyzed using principal component analysis (PCA) and factor rotations. PCA allows to significantly improve the overall quality of the data, but produces abstract components. Here it is shown that rotated components can be used without prior knowledge of the sample to help the interpretation of the data, obtaining quickly qualitative mappings representative of elements or compounds found in the material. Such abundance maps can then be used to plot scatter diagrams and interactively identify the different domains in presence by defining clusters of voxels having similar compositions. Identified voxels are advantageously overlaid on secondary electron (SE) images with higher resolution in order to refine the segmentation. The 3D reconstruction can then be performed using available commercial softwares on the basis of the provided segmentation. To asses the quality of the segmentation, the results have been compared to an EDX quantification performed on the same data.

1. Introduction

Over the past few years, capabilities of microanalytical instruments have largely improved. Three-dimensional microanalysis has been recently extended to various techniques at different scales, covering many aspects of physics, biology or material science. Among these techniques able to determine the chemical composition of materials, we may cite time-of-flight secondary ion mass spectroscopy (ToF-SIMS) (Fletcher et al., 2007; Smentkowski et al., 2007), scanning electron microscopy (SEM) and energy dispersive X-ray spectroscopy (EDX) coupled with a focused ion beam (FIB) (Kotula et al., 2003b, 2006; Schaffer et al., 2007), electron energy loss spectroscopy (EELS) and energy filtered transmission electron microscopy (EFTEM) tomography (Weyland and Midgley, 2003; Gass et al., 2006), scanning transmission electron microscopy using energy dispersive X-ray spectroscopy (STEM-EDX) (Möbus et al., 2003; Yaguchi et al., 2004), or atom probe (AP) tomography (Cerezo et al., 1988; Blavette et al., 1993). All these techniques require a specific specimen preparation combined with the corresponding analytical detection chain. The accessible sample volumes varies between the different techniques from few tenths of μm^3 to few $Å^3$. However all these techniques have in common that they produce a huge amount of raw hyperspectral data. Nowadays the acquisition of raw datasets above 1 GByte is routine. Additional data processing steps are always required to retrieve relevant information prior to the tomographic reconstruction. The data processing and tomographic reconstruction algorithm also depends greatly on the method, in particular if it is a serial sectioning or an angular tomography technique. Most of the data processing applied to the data aims either to reduce noise and/or to segment the data on the basis of the spectral information.

The most popular multivariate statistical analysis (MSA) methods is probably the principal component analysis (PCA). It was invented more than a century ago by Pearson (1901). The usage of MSA emerged in chemometrics in the seventies with increasing computer power and became more and more popular in the eighties (Malinowski, 2002). It is only within the last two decades that PCA and related MSA methods started to be used in microscopy, with as a first application the treatment of hyperspectral data acquired in EELS (Trebbia and Bonnet, 1990; Bonnet et al., 1999).

* Corresponding author. Tel.: +41 21 693 48 27.
E-mail address: guillaume.lucas@epfl.ch (G. Lucas).

0968-4328/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.micron.2013.08.005

Its usage in EDX has followed shortly (Titchmarsh and Dumbill, 1996). MSA, in particular PCA and factor analysis, allow to improve the signal-to-noise ratio by reducing the dimensionality of the data and to simplify their interpretation.

In this article we will focus on how these methods can be used to improve the overall quality of 3D hyperspectral data. Transformed data can be more easily interpreted in terms of element or compound spatial distributions and used as a base for the segmentation and the analytical 3D reconstruction of the analyzed volume. A real case study has been considered here. A sample produced by laser welding of a nickel–titanium alloy wire and a stainless steel wire (Vannod et al., 2011). It presents a complex intermetallic microstructure that has been analyzed by FIB-EDX (Burdet et al., 2013). Previous studies from Kotula and co-workers on similar FIB-EDX data and the usage of MSA methods for the segmentation has been reported (Kotula et al., 2003b,a). In this work results are compared to a standard based EDX quantification performed on the same dataset (Burdet et al., 2013; Burdet, 2012). This allowed us to work on a already well characterized sample in order to demonstrate the reliability of the MSA methods for the segmentation of 3D hyperspectral data.

2. Methods

In social or life science, individuals and variables are clearly defined. For example in the case of a group of persons as individuals, and their weight, height, age, nationality, eating habits and so on as variables, the relationship between individuals and variables is unambiguous. On the contrary a microscopist who is dealing with spectral images can equally considered images and spectra. Pixels of an image can be seen as individuals and the different energy channels of a spectrum as variables, or inversely energy channels as individuals and pixels as variables. For this reason depending on the point of view, we will refer to spectral domain and spatial domain when applying operations on spectra and images, respectively.

In our case the $m \times n$ data matrix $\mathbf{D}$ is organized in the following manner: the columns represent all the pixels of the image (or all voxels in the case of a 3D stack) and the rows represent the energy channels of the spectra. Thus we have m pixels/voxels and n energy channels. PCA first finds a linear combination of variables maximizing the variance in the least-square sense. This combination is called the principal component. Then it removes this variance and finds a second linear combination which maximizes the remaining variance, and so on. The PCA decomposition corresponds to the following matrix factorization:

$$\mathbf{D} = \mathbf{T}\mathbf{P}^T, \tag{1}$$

where $\mathbf{T}$ is an orthogonal basis of vectors called the score matrix and $\mathbf{P}$ is an orthonormal basis of vectors, called loading matrix. The columns of the score $\mathbf{T}$ can be seen as the pixels, expressed using a new basis of variables. The weights of each original variable, i.e. energy channel, in the new variables are given by the columns of the loading $\mathbf{P}$. Despite these weights reveal the influence of a given energy channel on the pixels of the spectral image, this new representation of the data is said to be "abstract" as it as no direct physical meaning. The product $\mathbf{P}^T\mathbf{P}$ yields to the identity matrix and the product $\mathbf{T}^T\mathbf{T}$ to a diagonal matrix whose elements are the eigenvalues of the covariance matrix $\mathbf{D}^T\mathbf{D}$. Each component, referring to a score vector and its corresponding loading vector, is associated to an eigenvalue that is directly related to the amount of variance explained by the component. Hence the first components are the vectors explaining the best the variance of the data. Usually by selecting the p components having the highest eigenvalues, it is possible to construct a model $\tilde{\mathrm{D}}$ of the data containing most of the valuable information:

$$\mathbf{D} \simeq \tilde{\mathrm{D}} = \tilde{\mathrm{T}}\tilde{\mathrm{P}}^T, \tag{2}$$

where $\tilde{\mathrm{T}}$ is here a $m \times p$ matrix and $\tilde{\mathrm{P}}$ is a $n \times p$ (for the sake of simplification the tilde notation will be omitted in the following of the article). The advantage of PCA is thus to reduce the dimensionality of data since $p \ll n$. It is used to efficiently reduce the noise in the data and the resulting model is a compression of the data without much loss of information.

In practice in the case of spectra recorded by EDX and any data generated by counting events, it is common to scale the data to take into account the Poisson statistic (Keenan and Kotula, 2004), so that the uncertainties in data are more uniform in the weighted space (Cochran and Horne, 1977). Indeed PCA assumes that data are normally distributed and a scaling step is often applied to satisfy this assumption. For time-of-flight secondary ion mass spectrometry with high counting rate, a scaling following a binomial model is more appropriate (Keenan et al., 2008). Weighted PCA decomposition also tends to be more stable and robust to outliers, and usually leads to more compressed data.

There are several methods to perform the PCA decomposition. The most straightforward is obviously to diagonalize the covariance matrix but it is rarely used because it not numerically efficient to compute it directly. If only few components are needed, the non-linear iterative partial least squares (NIPALS) algorithm (Wold, 1966, 1975) can be used, as it avoids the calculation of the covariance matrix. However the iteration time increases steadily for each consecutive eigenvalue so that it can only be used to find a limited number of components. This method has also an higher numerical accuracy. The method of choice is actually the singular value decomposition (SVD) which is a widely used technique in linear algebra. Efficient and numerically stable algorithm implementations are available freely (Anderson et al., 1999; Openblas, 2012) or in commercial packages. SVD decompose the data matrix into three matrix factors:

$$\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{3}$$

where the orthonormal $m \times m$ $\mathbf{U}$ and $n \times n$ $\mathbf{V}$ matrices are called respectively left and right singular vectors. The diagonal matrix $\mathbf{\Sigma}$ matrix is known as the singular value matrix. $\mathbf{\Sigma}$ is uniquely determined when its diagonal entries are listed in descending order, though the matrices $\mathbf{U}$ and $\mathbf{V}$ are not. They might be defined up to sign only in the case all singular values are non-degenerated. SVD provides a natural way to compute the PCA, because it is directly related to the eigenvalue decomposition of the covariance matrix. It can be easily shown that $\mathbf{U}$ and $\mathbf{V}$ are respectively eigenvectors of $\mathbf{D}\mathbf{D}^T$ and $\mathbf{D}^T\mathbf{D}$.

The main problem of the PCA is the abstract nature of the scores and loadings, i.e. they have no direct physical meaning. They are linear combinations of the original images and spectra and are only a representation of the data in a space maximizing the variance. As a consequence, they are difficult to interpret. One would expect components that could be directly interpretable in terms of pure element or pure compound spectra $\mathbf{S}$ and abundance maps $\mathbf{A}$, providing the relationship $\mathbf{D} = \mathbf{A}\mathbf{S}^T$. In terms of linear algebra this means to find matrices that are sparse, i.e. with many elements being equal or close to zero (Smentkowski et al., 2009). One way to achieve this goal is to perform rotations of the factors $\mathbf{U}$ or $\mathbf{V}$ depending on if we want to simplify the spatial or the spectral domain respectively (Keenan, 2009). These rotations preserve the orthogonality constraints on the domain they are applied but relax these constraints on the other domain. There are many methods to obtain rotation matrices. They can be classified in two categories: orthogonal rotations or oblique rotations. Orthogonal rotations preserve