EELS elemental mapping with unconventional methods
I. Theoretical basis: image analysis with multivariate statistics and entropy concepts

Pierre Trebbia *

Laboratoire de Physique des Solides, Bâtiment 510, F-91405 Orsay Cedex, France

and

Noël Bonnet

Unité INSERM 314 et Université de Reims, 21 rue Clément Ader, F-51100 Reims, France

Received 7 June 1990

Electron energy loss filtered images recorded within a transmission analytical electron microscope are now widely used for the mapping of the elemental distribution of a given atomic species in a specimen prepared as a thin film. Such an image processing may produce both valuable results and artifacts if a careful inspection of all the hypotheses needed by the calculation is not carried out. This paper presents some general statistical methods for a contrast information analysis of a noisy image data set. After a brief introduction of different concepts such as contrast, variance, information and entropy, two unconventional approaches for image analysis are explained: the relative entropy computed with respect to a pure random and signal-free image and the factorial analysis of correspondence (a branch of multivariate statistics). In the companion article (part II), these concepts are applied to real experiments and the results compared with those obtained with a conventional method. Although electron energy loss spectroscopy is the only technique considered here, these methods for image analysis can be applied to a wide variety of noisy data sets (spectra, images, ...) recorded from various sources (electrons, photons, ...).

1. Introduction

Analytical electron microscopy has been demonstrated, during this last decade, to be very efficient for solving specific problems such as near-trace-element concentrations, line profiles at high spatial resolution or time-resolved experiments [1]. Because the problems to be solved are more and more lying near the ultimate limits of the technique [2], there is a need for sophisticated data analysis and processing algorithms which can be run in real time on small computers like IBM PC's or Macintosh's. Within this context, it appears to be useful to have a look at well known statistical procedures which may be of some help for solving dedicated problems. Among these, the processing of energy-filtered images (electron energy loss spectroscopy technique) for the quantitative elemental mapping of a given atomic species at high spatial resolution is of major importance both in material sciences [3,4] (grain boundary segregation, in-situ chemical reaction, ...) and in biology or pathology [5–8] (localization of pre-stained molecules or macromolecules in a tissue, for example).

* Present address: Laboratoire d'Analyse des Solides Surfaces et Interfaces (LASSI), Université de Reims Champagne-Ardenne, B.P. 347, F-51062 Reims Cedex, France.

0304-3991/90/$03.50 © 1990 – Elsevier Science Publishers B.V. (North-Holland)

Whatever the specimen is, the images to be processed always contain three main contributions [9] as shown in fig. 1c: this energy-filtered image (125 eV) was recorded from a section of a guinea pig leukocyte. A cerium substrate was used for a chemical reaction with the alkaline phosphatase enzyme and the cerium mapping image (fig. 1d) enables biologists to determine the enzyme location. This cerium map was obtained by recording independent pictures of the same specimen at different energy losses (fig. 1a: 85 eV, 1b: 105 eV and 1c: 125 eV), the "cerium signal" appearing in EELS for energy losses slightly greater than the N_{45} excitation threshold (110 eV, see fig. 2a). Picture 1c is therefore the key picture in this image series. But besides the "useful cerium information" (which has been computed, extracted from fig. 1c and shown in fig. 1d), it is quite clear that

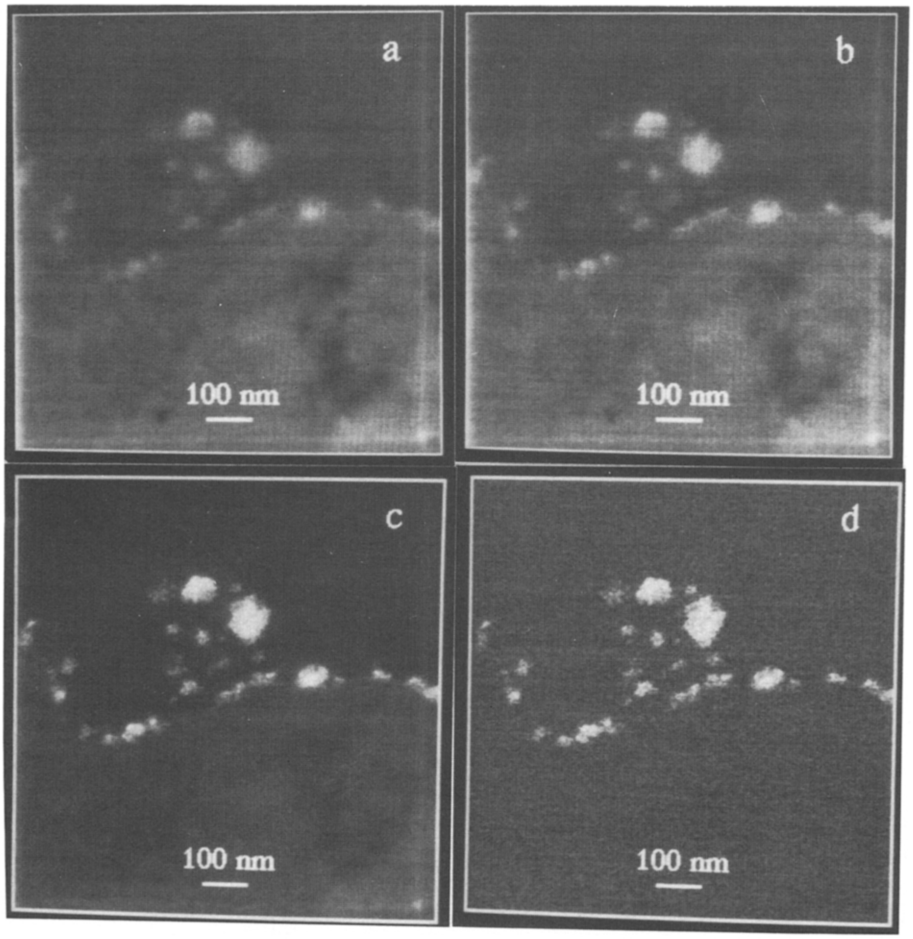

Fig. 1. Energy-filtered images of a thin section of a guinea pig leukocyte. A cerium substrate was used for a chemical reaction with the alkaline phosphatase enzyme. Specimen courtesy of Dr. C. Lechêne (Harvard Medical School, Boston). These images were recorded at different energy losses ((a) 85 eV, (b) 105 eV, (c) 125 eV) within an analytical field emission scanning transmission electron microscope fitted with a Gatan 607 electron spectrometer. Incident electron energy: 100 keV. Pixel size: 4 nm. Probe size: 2.3 nm. Total acquisition time: 197 s. Total dose received by the sample in the centre of the probe: 27 cb/cm^2. Picture (c) contains the signal related to the cerium N_{45} excitation but part of the contrast is also due to mass-thickness fluctuations; (a) and (b) show this latter contrast. Note, for example, the two contamination lines crossing near the bottom right corner. Using a power law model for the background (eq. (9)) and a Poisson distribution for the noise, the background is computed pixel by pixel [30] from (a) and (b), subtracted from (c) and displayed in (d). Image courtesy of C. Jeanguillaume, M. Tencé and P. Trebbia.

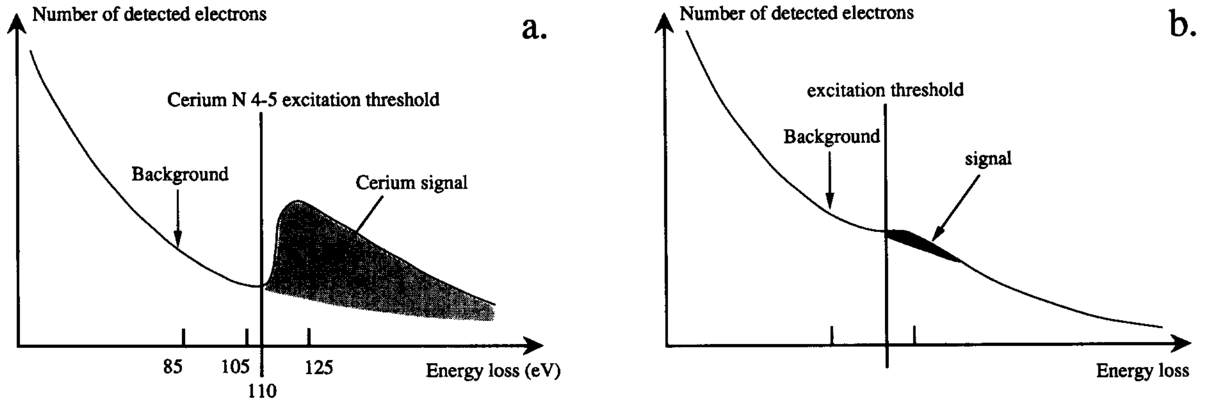

Fig. 2. Schematic representations of electron energy loss spectra. (a) As recorded from a cerium-rich area of fig. 1. Below 110 eV, the recorded intensity is only due to the background, whereas for energy losses above the N_{45} threshold (excitation of the 4d electrons of cerium atoms), a fraction of the measured intensity constitutes the "useful signal" and is used for the cerium map. This background may sometimes be modeled by a power law model (eq. (9)) and extrapolated below the "useful signal area". (b) Typical situation where the very low signal/background ratio would lead to a "negative elemental map" if an appropriate scaling of intensities measured on both sides of the excitation edge is not performed.

the contrast exhibited by fig. 1c also contains "useless information" which can be categorized into two components: the background and the noise.

What we call "background" mixes together all the information which is not the direct consequence of the excitation, by the primary 100 keV incident electron beam, of the N_{45} cerium atomic level; part of the contrast in fig. 1c is due, for example, to mass-thickness fluctuations and can be directly correlated to the contrast of figs. 1a and 1b recorded at energy losses below the N_{45} excitation threshold.

On the other hand, the noise component, which appears as a "snow effect" on the pictures, is merely due to the fact that these images have been recorded within a limited time period (about 1 minute per image; each image contains 256 × 256 picture elements (pixels), the counting time per pixel being 1 ms). The number of detected electrons per pixel is therefore limited (in the hundreds range) and pure statistical fluctuations in counting are expected to follow a Poisson distribution (variance = squared standard deviation ≃ mean).

The image-processing problem, which is not strictly restricted to the EELS elemental problem but is inherent to any imaging technique, can then be stated as the following: how to separate, in an image like the one shown in fig. 1c, the "useful" information (the cerium excitation signal) from the "useless" one (background and noise). In order to try to answer this question, we have to begin with the definitions of a few terms (contrast, variance, information, entropy, relative entropy) we are using throughout this paper.

2. A few definitions

2.1. Contrast and variance

Let two pixels of an image be (i, j), and X_i and X_j their respective contents, i.e. the number of detected electrons. The mean value $\langle X \rangle$ and the variance s^2 are defined as:

$$\langle X \rangle = (X_i + X_j)/2, \tag{1}$$

$$s^2 = (X_i^2 + X_j^2)/2 - \langle X \rangle^2 = ((X_i - X_j)/2)^2. \tag{2}$$

The standard deviation s is thus defined as:

$$s = |(X_i - X_j)/2|. \tag{3}$$

The contrast C between these two pixels is classically defined as [10]:

$$C = \frac{|X_i - X_j|}{X_i + X_j} = \frac{s}{\langle X \rangle}, \tag{4}$$

provided $\langle X \rangle \neq 0$; i.e. X_i and X_j both $\neq 0$, since none of them can be negative (counting numbers).

It turns out that the local contrast between two pixels is directly proportional to the standard deviation, which is the square root of the variance. The *analysis of the contrast* of an image can thus be carried out through the *analysis of the variance*.

2.2. Information

Since the pioneer work of Shannon [11–13], it is well known that the information $I(E)$ associated with an event E is defined as the cologarithm of the probability of that event occurring *before it happened*:

$$I(E) = -\log[P(E)]. \tag{5}$$

Information is a dimensionless number, the usual units being either the nat or the bit depending on the kind of the logarithm used (respectively neperian and base 2). It is always a positive number, taking a zero value only for the event that we know a priori will certainly occur. The consequence of this property is that information can only increase with new events occurring, the only way of reducing information being to forget (deliberately or not) the past.

2.3. Entropy

Let E be a random variable whose possible values are E_k $(k = 1, \ldots, N)$ and $P(E)$ the probability law ruling this variable. The entropy $S(P)$ of this probability law is defined as the mathematical expectation of the information conveyed by the N possible events [11–13]:

$$S(P) = \text{entropy} = \sum_k I(E_k)\, P(E_k)$$
$$= -\sum_k P(E_k) \log[P(E_k)]. \tag{6}$$

Entropy has several interesting properties:

- $S(P)$ is always positive or null.
- $S(P)$ is null only if among all the possible E_k values there is one whose probability is equal to 1, i.e. we are a priori certain that E will take the value E_k.
- $S(P)$ is maximum when all the E_k values have the same probability $1/N$ to occur: $S(P)$ is then equal to $\log(N)$ and increases with N.
- Entropy and information vary in reverse ways: the higher the information of a given event, the lower the a priori probability of that event, the less uniform the probability distribution, the higher the difference between the actual entropy value and its maximum, and therefore the lower the entropy value.

For example, an image showing no contrast, no details, conveys no information and its entropy is maximum. Suppose that on a mid-summer night we take with a telescope a picture of a region of the sky where nothing can be seen. A priori, we expect the picture to be uniform. If, after the plate being processed, some contrast appears leading to the conclusion that a planet has been discovered, then:

- the information conveyed by this contrast is high;
- the entropy is lowered; the visible details in the picture have been "ordered" by an external agent (the real presence of this planet in that region of the sky);
- the a priori probability law for a next experiment must be changed; we may not forget what we learned.

2.4. Relative entropy

The entropy of an a priori probability law has been defined above within the context that one among the N possible events E_k will certainly occur; after the "experiment", the a posteriori probability that $E = E_k$ is 1 (E_k has occurred). Let us suppose now that the "experiment" does not deliver a deterministic value but a random variable. A relative entropy may then be defined as the *information associated with a change* of probability law: Let $P_1(E)$ be the a priori law ruling the event E and $P_2(E)$ the a posteriori probability law. The relative entropy of P_2 with respect to P_1 can be expressed as [14]:

$$Q(P_2/P_1) = \sum_k P_2(E_k)\log[P_2(E_k)/P_1(E_k)]. \tag{7}$$

When there exists k such that $P_2(E_k) = 1$ and $P_2(E_j) = 0$ for all $j \neq k$, then the relative entropy simply reduces to the classical definition of the information:

$$Q(P_2/P_1) \to I(E_k) = -\log[P_1(E_k)]. \tag{8}$$

On the other hand, if for all k, $P_2(E_k) = P_1(E_k)$, i.e. if there is no probability change, no information has been added by the experiment and $Q(P_2/P_1)$ is null.

2.5. Application of relative entropy to image analysis

When a given specimen is observed for the first time within an analytical electron microscope, we have a priori no idea about what the first image would look like: contrast amplitude, contrast location. Let us suppose that this image would be made of a collection of P pixels. Let N be the total number of electrons which will be detected in the whole image during the exposure time (N would be equal to the incident dose if all the incident electrons were to be detected, i.e. if the collection and detection efficiencies were equal to unity). If these two numbers N and P are the *only information* we know about the picture to be taken, then we have to guess that each pixel would collect a mean number N/P of electrons. Due to counting statistics, N/P is only a mean value, the actual number of collected electrons in a given pixel being a random variable following a Poisson distribution. Therefore, the only a priori image that we can imagine is a uniform one, containing absolutely no contrast, except for random noise due to Poisson statistics. The histogram of such an a priori image can be modeled (fig. 3) by a Gaussian distribution spreading over G grey levels (fixed by the dynamic of the detection unit) and whose characteristics (mean and standard deviation) only depend on the experimental set-up (N, P and G).

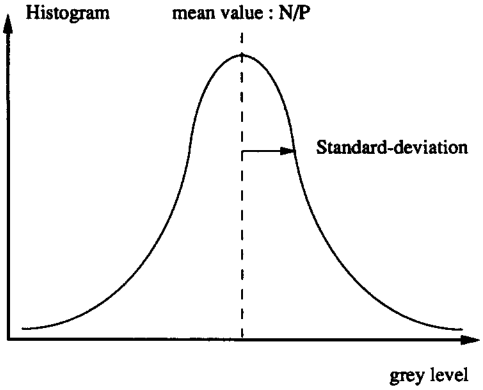

Fig. 3. Typical histogram of a fictitious pure random image taken from a perfectly transparent object. The random character of the counting measurement is the only cause for a spread distribution. The use of a Gaussian profile can be justified by the central limit theorem.

After an appropriate normalization, such a histogram can be interpreted as the *a priori probability* $P_1(g)$ to measure a given g grey value.

If now we insert our specimen into the microscope column and take a real picture within these experimental conditions (N, P and G), we can compute the histogram of that picture and, after normalization, use it as an *a posteriori probability law* $P_2(g)$ to measure a given grey value.

Using eq. (7) we are thus able to quantitatively determine the information contained in the change of the probability laws ($P_1 \to P_2$), i.e. the information conveyed by the only difference between these two laws, namely: "The specimen is inside the microscope column".

The practical use of this relative entropy estimation will be explained in more detail in the companion article. It is worth pointing out that we actually only need to record a single image of the specimen, giving the P_2 probability law, the statistics needed to build P_1 (mean grey value, standard deviation) being deduced from P_2 [15].

We can already state that this relative entropy measures the departure from a maximum-entropy image compatible with the experimental set-up. Provided the only a priori knowledge of the experiment (N, P, G, Poisson statistics) is correct, the relative entropy, as defined by eq. (7), must be positive even if some terms in the sum are negative. If it is found to be equal to zero, then we can conclude that we actually made the image of nothing. The higher the relative entropy, the higher the contrast induced by the specimen. Actually, the mathematical expectation of the relative entropy is equal to half the number of the grey levels really present in the image [16,17].

It is worth noting that the contrast we are dealing with now is a global contrast computed over the whole image and not the local contrast as defined by eq. (4). But we are free to select different reduced areas in the image and to compare the a priori and a posteriori histograms on these selected pixels. It is obvious, however, that a histogram has a physical meaning only if the number of sampled pixels is large enough ($P > G$).

3. General considerations about image processing

Strictly speaking, an image conveys twofold information:
- the contrast information as explained above, which is related to the intensity contained in each pixel, and
- the structural information; we have chosen pixels i and j among P pixels to convey a contrast information. In the absence of any knowledge about the structure of the image, we had the a priori probability $1/P$ to choose a particular pixel. For example, figs. 4a and 4b do convey the same global contrast information (their histograms are identical) but obviously do not convey the same structural information. It is clear, however, that a comparison of the contrast information on any selected reduced area would likely reveal that figs. 4a and 4b are different (their partial histograms would not be the same; see figs. 4c and 4d).

Any image processing which modifies either the image contrast or the image structure modifies, therefore, the global information conveyed by the image. This information change can be either useful (if the processing is well controlled and is supported by logical arguments) or useless in the reverse case. Moreover, any hypothesis needed by the processing and explicitly (or implicitly) used in the calculation also modifies the information content. Let us examine a few examples.

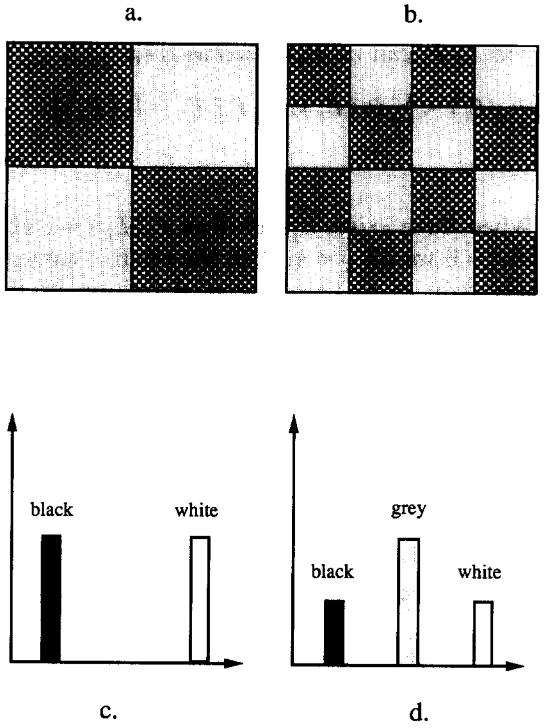

Fig. 4. Schematic illustration of two images (a and b) conveying the same global contrast information as measured by the global relative entropy, and displaying quite different structural information. But if the analysis is performed over a reduced area, the local contrast information is no longer identical: histograms (c) and (d) respectively refer to the top left quarter of (a) and (b) and exhibit strong differences.

3.1. Noise filtering

Usually, when a "freshly recorded" data set (spectrum, image, ...) is found to be very noisy, one has the temptation to "filter the noise" by applying a convolution product with a normalized window function whose finite width is adjusted rather empirically. The signal-to-noise ratio (SNR) is enhanced indeed, but such a treatment induces three major consequences:
- the noise is not fully removed because its power spectrum is generally very large and the frequencial components which are superimposed over the signal components cannot be isolated;
- the signal is also changed (for example, the intensity of a sharp peak is lowered even if the SNR is increased);

– the additional information due to the contrast enhancement is inevitably associated with a decrease of the structural information: the structure has been smoothed and the lateral resolution (spatial resolution in an image) lowered.

Some authors have suggested using optimized non-linear filtering [18] in order to minimize these consequences. Even if it appears that these procedures are very attractive for a given class of data analysis, it seems unlikely that they might fully escape the three inconveniences listed above: whatever the sophistication of the algorithm, it will always contain some kind of "smoothing" and hence some information transfer.

3.2. *Adding hypotheses*

Actually, there is no way to increase the global information of an image, except for injecting new information from "outside hypotheses". Depending on the validity of these hypotheses, the processing would thus give either valuable results or artifacts.

For example, Fourier filtering in the reciprocal space may induce severe consequences: when setting the characteristics of the frequency filter (low-pass, band-pass, high-pass) and its parameters (cut-off frequency, bandwiths, ...), the user has to make some hypothesis about the signal spectrum and the noise spectrum. Whatever the filter configuration, one has to keep in mind the injection of the following information: "my data have been filtered". As already demonstrated by several authors, forgetting this would lead to erroneous conclusions (see, for example, fig. 5 in ref. [19]).

The problem is even more subtle when the hypotheses are implicit, i.e. not explicitly stated. For example, the first idea for increasing the SNR would be to increase the counting time of the experiment. Since the number of detected electrons obeys a Poisson law, increasing the experiment time by a factor F would increase the SNR by a factor $\sqrt{F}$. But one has to check whether the data are stationary or, in other words, that the specimen is not changing with time (contamination, radiation damage, drift, ...).

Another example of implicit hypothesis consists in the superposition (after appropriate alignment: translation and/or rotation) of different areas of a single low-dose image containing multiple replicas of the same object. Here again, piling up these sub-images would increase the SNR with no dose-cost, but one has to carefully inspect the assumption that these different sub-images can be related to a single object (i.e. is the word "replica" reliable?) [20].

Since it is impossible to increase the total information conveyed by an experiment without explicit or implicit hypotheses, one must never forget that "good processing" would never be used as a substitute for "good experiments": the more the information collected by the experiment, the better the result. In any case, one has to check, with appropriate statistical tests (the so-called hypothesis tests), that the hypotheses needed by the processing algorithm may be accepted (see for example refs. [20–22]).

4. Image processing for EELS elemental mapping

Going back to the problem of background removal in an energy-filtered image (fig. 1c), we would like to briefly summarize the different procedures already known. A detailed critical analysis of the most often used algorithms can be found elsewhere [23].

According to the general behavior of the EELS spectrum shown in fig. 2a, one may first think about a simple subtraction between two images recorded on both sides of the excitation threshold (say fig. 1c minus fig. 1b). This kind of approach has been extensively used by several authors [24–28] and has been demonstrated [23] to induce possible artifacts; as can be seen, the background variation with energy loss is rather important and image 1b must be scaled to an appropriate value before being subtracted. In the case of a very poor signal/background ratio (fig. 2b), neglecting this scale factor would even lead to "negative images"! The problem is therefore to find an appropriate scaling factor or, in other words, to find some relation between intensities recorded in the pure background area (i.e. at energy losses smaller than the excitation threshold). Furthermore, this scaling factor cannot generally be considered as constant over the whole field of view because it is by itself a function of the specimen content.

The first possible approach lies in finding a mathematical expression for a background model. The best estimate of the different parameters needed by such a model would be obtained through an iterative hypothesis test (does that value for the parameter better fit the experiment than this one?). Once the maximum likelihood of the model is reached, then the model is extrapolated up to the energy loss corresponding to the "signal position" (125 eV in the case of fig. 1) and subtracted.

This maximum-likelihood search can be performed either on the whole image at once or pixel by pixel. The model is said to be global in the former case and local in the latter. As was demonstrated by Bonnet et al. [23], the local algorithm is to be preferred, provided the model is simple enough to be run P times (P being the number of pixels in each image) within a reasonable time period and at a sufficient statistical confidence level.

Concerning this point of a simple expression for the background, Egerton [29] suggested that a power law function may fit, at least on a reduced energy range, the decreasing intensity:

$$I_{\text{backgr}} = A\, \Delta E^{-R}, \tag{9}$$

where ΔE is the energy loss value, and A and R two empirical parameters to be adjusted from the measurement. Since there are two parameters to be calculated, at least two equations (i.e. independent measurements) are needed; two equations would give a single solution (A, R) whereas three or more equations would account for the random character of the measurement, that is, to estimate the maximum likelihood of what should be (A, R).

In the algorithm described by Jeanguillaume et al. [30], the model is found to be simple enough and the two parameters (A, R) are estimated for each pixel from two (or more) images recorded before the excitation edge onset (figs. 1a and 1b). The background to be removed from the "key picture" (fig. 1c) is estimated via eq. (9).

This processing algorithm for EELS elemental mapping has been proved to be superior to any other [23] and was successfully applied to various problems.

But there exist several practical situations where no mathematical model can be validly used for the description of the background. For example, the case of the phosphorus mapping in thin biological sections is severe (L_{23} edge at 132 eV): due to the tail of the very intense peaks located in EELS spectra near 20 or 30 eV, and to rather important modulations on that tail, any hypothesis test running on a power law model tends to reject such an empirical model. One could then be tempted to increase the number of parameters in the model. This temptation has two serious drawbacks:

- One may not forget the physical meaning of the model under process: if Egerton's model may be derived from what we know about the behavior of the inelastic scattering cross sections involved in the electron–matter interaction, the use of any other empirical model may be questioned.
- Everyone is aware that using higher and higher degree polynomial expressions does, to some extent, increase the "goodness of a fit", but also increases the number of possible nodes in the function; in other words, such an extended model could be used for *interpolation* purposes but would certainly be very inaccurate for *extrapolation*. And because the EELS background cannot be estimated on both sides of the excitation threshold (see fig. 2), extrapolation is the very problem we are dealing with.

In order to cure this dilemma, we suggest a two-step procedure: before doing any image *processing*, one has first to make a close inspection of the available data information, that is, to perform an image *analysis*; and, in a second step, to take into account the collected information for deciding which process has to be done under which hypothesis.

We have thus to begin with the analysis of the information contained in the original data set of the unprocessed images (figs. 1a–1c, for example) and to classify this information hierarchically: We will analyze the contrast of all the unprocessed images measuring some distance between pixel intensities (we know from eq. (4) that the data variance is a pertinent parameter) and, using unbiased statistical tools, build a new image containing only a fraction of the whole information. If we were able to identify this fraction of information as the one which is "useful", then we may feel justified that we have built an image which is likely the one we needed.

This procedure (first analyze, then process) does not require any mathematical model for the background behavior. The only question we try to answer is the following: Is it possible to make a distinction between the data variance due to the real presence, in some pixels, of the signal we look for, and the data variance induced by both the background and the noise?

This is a typical question that the factorial analysis of correspondence (FAC), a branch of multivariate statistics, can answer.

5. Factorial analysis of correspondence

Among all the pioneers of multivariate statistics, the work performed by Pearson [31] and Spearman [32] is acknowledged as being of major importance, and the theoretical foundations of FAC were established by Benzecri [33,34]. Applications of multivariate statistics in electron microscopy are rather recent [35–39]. In previous papers [40,41], one of us already suggested that FAC would be suited for the specific problem of EELS elemental mapping. In the companion article [15] we shall illustrate the power of such an analysis. Here, in order to clear up any possible feeling that some "magic trick" is applied, we would like to explain the different steps involved in a factorial analysis of correspondence.

5.1. The aim of FAC

Let X_{ij} be a matrix made of two kinds of variables: C columns noted i ($i = 1, \ldots, C$) and P lines noted j ($j = 1, \ldots, P$). Let us suppose $C < P$. This matrix can be seen either as a description of the C coordinates (the columns) of P variables in $\mathbb{R}^C$ or as a description of the P coordinates (the lines) of C variables in $\mathbb{R}^P$.

For our purpose (elemental mapping with EELS filtered images), the C columns can be seen as C filtered images, each of them containing P pixels. Each one of these C images is therefore treated as a vector with P components.

In the most general case, there is likely some kind of dependence between some lines and/or some columns. In such a case, the information conveyed by the CP values X_{ij} is partially redundant. The goal of FAC is to remove this redundant information and to classify in a hierarchic order the remaining information: from the original data set X_{ij} (i.e. lines and columns), one shall try to reduce the dimension of the original vectorial space ($\mathbb{R}^C$ or $\mathbb{R}^P$) by building Q eigenvectors with the following properties:

- they are calculated as a linear combination of the original variables (lines or columns),
- they are orthogonal to each other and they generate a vectorial space $\mathbb{R}^Q$,
- they are normalized to unity, and
- the axes supported by them will describe most of the information conveyed by X_{ij}.

These eigenvectors are successively calculated in order that the associated axes will "pass through the variables as closely as possible"; that is, by minimizing the distance between the variables and the axes or, equivalently, by maximizing the coordinates of the variables on the axes (see fig. 5). These axes would thus build a Q-dimensional "skeleton" of the X_{ij} data cloud.

This procedure (looking for such eigenvectors) is also the aim of another multivariate statistical analysis: principal component analysis (PCA).

5.2. Main differences between FAC and PCA

Basically, PCA makes a clear distinction between the variables: lines are not treated as the columns, and the metric used to measure the distance between two points is Euclidean. On the other hand, FAC keeps a complete symmetry between the two kinds of variables and uses a different metric called a χ^2 metric.

Therefore, in order to be able to use the PCA technique in FAC, one has to make a transforma-

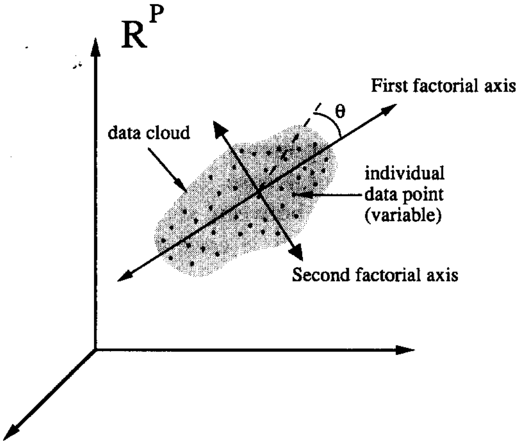

Fig. 5. Schematic representation of the cloud of the C variables in $\mathbf{R}^P$. Each variable (an energy-filtered image) is defined by P coordinates (the P pixel intensities). In order to measure some distance between these coordinates, FAC tries to find the Q orthogonal axes which would build the skeleton of this cloud: these axes are computed with the constraint that the distances between the variables and the axes are minima.

tion of the original data set X_{ij} and build a new matrix Y_{ij} such that:
- the formulation of Y_{ij} is symmetrical with respect to i and j, and
- the χ^2 metric used on X_{ij} is equivalent to a Euclidean metric on Y_{ij}.

If such a Y_{ij} matrix is built, FAC is then merely reduced to a twin PCA on the lines and on the columns of Y_{ij}.

5.3. Building the Y_{ij} matrix

This will be done in three steps:

(a) Usually, X_{ij} values are related to counting numbers. In order to make valuable comparisons between lines and/or columns, one has to normalize these numbers; they are divided by partial summations over lines (or columns): Let be

$$X_i = \sum_j X_{ij}, \quad X_j = \sum_i X_{ij}, \quad X = \sum_i X_i = \sum_j X_j, \tag{10}$$

$$X'_{ij} = X_{ij}/X_i \tag{11}$$

(in the case of a column normalization).

Because $\sum_j X'_{ij} = 1$, the number of independent variables (the lines) is reduced: $P-1$ instead of P. Therefore the dimension of the X'_{ij} vectorial space is no longer P. This reduction in the space dimension will give a trivial eigenvalue equal to 1 (see below).

(b) We transform now X'_{ij} into X''_{ij} in such a way that a χ^2 distance between two X'_{ij} will be equal to a Euclidean distance between the relevant X''_{ij}:

$$d^2_{ii'} \overset{\chi^2}{=} \sum_j X(X'_{ij} - X'_{i'j})^2/X_j. \tag{12}$$

To make it Euclidean (sum of squared differences), one has to transform

$$X''_{ij} = X'_{ij}\sqrt{X/X_j} = \frac{X_{ij}\sqrt{X}}{X_i\sqrt{X_j}} \tag{13}$$

and

$$d^2_{ii'} \overset{\text{Euclidean}}{=} \sum_j (X''_{ij} - X''_{i'j})^2. \tag{14}$$

(c) The last transformation insures the symmetry between the two kinds of variables; each X''_{ij} is weighted by the factor $\sqrt{X_i/X}$:

$$Y_{ij} = X_{ij}/\sqrt{X_i X_j}. \tag{15}$$

5.4. Building the variance–covariance S matrix

In order to compose the variance–covariance matrix containing the "contrast information", the Y_{ij} matrix has to be multiplied by the transposed Y^t_{ji} one. The eigenvalues and eigenvectors of the resultant matrix S will be then computed. It can be shown that the two matrices Y^tY and YY^t have the same number of non-zero eigenvalues and that these eigenvalues are identical. One has thus rather to choose the product which builds the matrix S with the smallest number of terms. In our case, since C is assumed to be smaller than P:

$$Y^tY \to S_{ii'} = \sum_j (Y_{ij}Y_{i'j}) = \frac{1}{\sqrt{X_i X_{i'}}} \sum_j \frac{X_{ij}X_{i'j}}{X_j}. \tag{16}$$

Due to commutation properties, $S_{ii'} = S_{i'i}$ and the variance–covariance matrix S is symmetrical.

5.5. Properties of the S matrix

(a) It can be shown that all the eigenvalues of S are positive and less than or equal to 1.

(b) The Y_{ij} matrix on which PCA is performed is not a matrix with centered values; the matrix with centered values would be:

$$Y'_{ij} = \frac{X_{ij} - X_i X_j}{\sqrt{X_i X_j}}. \tag{17}$$

But it can be shown that, provided the eigenvector associated with the trivial eigenvalue 1 is not taken into account, a PCA performed on the non-centered matrix Y_{ij} gives the same results as if it was performed on the centered one.

(c) The sum of the eigenvalues of the matrix $S_{ii'}$ (among them one is trivial and is equal to 1) is equal to the trace of that matrix.

Let us order the eigenvalues by decreasing order λ_α ($\alpha = 1, \ldots, C$):

$$\lambda_\alpha \leqslant \lambda_{\alpha-1} \quad \text{and} \quad \lambda_1 = 1, \tag{18}$$

$$\sum_1^C \lambda_\alpha = 1 + \sum_2^C \lambda_\alpha = \sum_1^C S_{ii} = \text{trace}(S), \tag{19}$$

$$\sum_1^C \lambda_\alpha = \sum_i \left(\frac{1}{X_i} \sum_j \frac{X_{ij}^2}{X_j} \right). \tag{20}$$

(d) Let S_λ be the sum of the non-trivial eigenvalues:

$$S_\lambda = \sum_2^C \lambda_\alpha. \tag{21}$$

It can be shown that the random variable XS_λ obeys a χ^2 distribution law with $\nu = (C-1)(P-1)$ degrees of freedom under the hypothesis H that the lines and the columns of X_{ij} are statistically independent. This can be a good test to check whether some information in the X_{ij} is redundant (giving zero eigenvalues).

5.6. Finding eigenvalues and eigenvectors

Whatever the PCA performed (over the lines or the columns), the non-trivial and non-zero eigenvalues are noted λ_α with α ranging from 2 to the minimum of the two numbers C or P.

Depending on the PCA performed, the relevant eigenvectors are expressed as linear combinations of the variables: if the variables are the lines, the eigenvectors are noted: $u_\alpha = \{u_{\alpha j}\}$; if the variables are the columns, the eigenvectors are noted: $v_\alpha = \{v_{\alpha i}\}$. Since the diagonalization is done on the $S_{ii'}$ matrix, one obtains: (a) the values of the λ_α, (b) the C coordinates $v_{\alpha i}$ of the associated eigenvectors.

The P coordinates $u_{\alpha j}$ of the u_α eigenvector can be deduced from $v_{\alpha i}$ by the relation

$$u_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_i Y_{ij} v_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_i \frac{X_{ij}}{\sqrt{X_i X_j}} v_{\alpha i}. \tag{22}$$

As already noticed, the non-trivial λ_α values are reorganized by decreasing value and, *from now, in order to keep clear the notations, we merely omit the first trivial eigenvalue* and start the λ_α with $\lambda_1 \neq 1$ (that is, α is now ranging from 1 to $\min(C, P) - 1$).

Each one of these eigenvectors is supported by a factorial axis with arbitrary orientation. Each axis α "explains" a part C_α of the total variance (i.e. of the total information):

$$C_\alpha = \lambda_\alpha / \sum_\alpha \lambda_\alpha. \tag{23}$$

With increasing α, the coefficients C_α are decreasing: the highest-order factorial axes do support less and less variance. In our case ($C < P$), there is a maximum of $C - 1$ factorial axes. In the absence of any further information, one should expect that each axis would explain a mean variance:

$$C_{\text{mean}} = 1/(C-1). \tag{24}$$

Therefore, any axis α with $C_\alpha \geqslant C_{\text{mean}}$ may be considered as conveying an important part of the total information. One could, in a first step, neglect all the β factorial axes such that

$$C_\beta \ll C_{\text{mean}} \tag{25}$$

and

$$\sum_{\alpha \neq \beta} C_\alpha \cong 100\%. \tag{26}$$

Let Q be the number of factorial axes which have been stated as "important". Then,

$$\sum_{1}^{Q} C_\alpha \leqslant 100\%. \tag{27}$$

And the vectorial space supporting "the important part of the total information" will be fully described by the first Q eigenvectors u_α (or v_α).

5.7. Coordinates of the variables on the factorial axes

For a full description of each one of the initial variables X_{ij}, one has to calculate the coordinates of the points representing the variables on the axes:
(a) the value of the projection of the ith column on axis α is

$$P_{\alpha i} = \sqrt{\frac{X\lambda_\alpha}{X_i}}\, v_{\alpha i}; \tag{28}$$

(b) the value of the projection of the jth line on axis α is

$$Q_{\alpha j} = \sqrt{\frac{X\lambda_\alpha}{X_j}}\, u_{\alpha j} = \frac{\sqrt{X}}{X_j} \sum_i \frac{X_{ij} v_{\alpha i}}{\sqrt{X_i}}; \tag{29}$$

(c) these projections can be deduced from each other by the following relations:

$$P_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_j \frac{X_{ij} Q_{\alpha j}}{X_i}, \tag{30}$$

$$Q_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_i \frac{X_{ij} P_{\alpha i}}{X_j}. \tag{31}$$

5.8. Help for the interpretation of the physical meaning of the axes

Up to now, we have only performed a variance analysis of the original data set X_{ij}. FAC analysis gives us a hierarchical classification with respect to the different variance magnitudes. One has now to interpret this classification using our preknowledge of the physical properties of the different images.

(a) The absolute contribution AC_α of a variable (i or j) to axis α measures the part that this variable plays in the explanation of the variance supported by the axis.

Since λ_α is the variance of all the variables along axis α, and since $(P_{\alpha i})^2 X_i / X$ is the variance of column i along axis α, the ratio $\mathrm{AC}_{\alpha i}$ measures the absolute contribution of variable i to axis α:

$$\mathrm{AC}_{\alpha i} = \left[(P_{\alpha i})^2 X_i \right] / (X\lambda_\alpha) \tag{32}$$

and

$$\mathrm{AC}_{\alpha j} = \left[(Q_{\alpha j})^2 X_j \right] / (X\lambda_\alpha). \tag{33}$$

One can notice that

$$\sum_i \mathrm{AC}_{\alpha i} = \sum_j \mathrm{AC}_{\alpha j} = 1. \tag{34}$$

(b) The relative contribution RC_α of a given variable (i or j) to axis α is equal to $(\cos\theta)^2$ where θ is the angle between the factorial axis and the line joining the variable to the gravity center of all the variables (see fig. 5). $\mathrm{RC}_{\alpha i}$ is thus an indication of the deformation induced when variable i is projected onto axis α. When $(\cos\theta)^2$ is nearly equal to 1, the deformation is weak and the variable really owns the exclusive property of this axis.

$$\mathrm{RC}_{\alpha i} = \left[X (P_{\alpha i})^2 \right] \Big/ \left[\sum_j X_j \left(\frac{XX_{ij}}{X_i X_j} - 1 \right)^2 \right] \tag{35}$$

and

$$\mathrm{RC}_{\alpha j} = \left[X (Q_{\alpha j})^2 \right] \Big/ \left[\sum_i X_i \left(\frac{XX_{ij}}{X_i X_j} - 1 \right)^2 \right]. \tag{36}$$

In order to interpret the physical meaning of the axes, that is, the usefulness of the information conveyed by them, one has thus to look for the variables which have both the highest values RC_α and AC_α. The meaning of the axis is then deduced from the physical meaning of these selected variables.

5.9. Reconstruction of the initial data set X_{ij}

If all the factorial axes are taken into account, one can *exactly* reconstruct the initial values X_{ij} by the following formula:

$$X_{ij} = \frac{X_i X_j}{X}\left(1 + \sum_1^{C-1} \frac{P_{\alpha i} Q_{\alpha j}}{\sqrt{\lambda_\alpha}}\right), \tag{37}$$

in the case $C < P$.

But if one only considers the Q first factorial axes, one can *approximately* reconstruct the initial values by the partial sum

$$X_{ij} \simeq \frac{X_i X_j}{X}\left(1 + \sum_1^{Q} \frac{P_{\alpha i} Q_{\alpha j}}{\sqrt{\lambda_\alpha}}\right). \tag{38}$$

These approximate X_{ij} values will contain a fraction

$$\sum_1^Q \lambda_\alpha \Big/ \sum_1^{C-1} \lambda_\alpha$$

of the initial variance. Such a partial reconstruction is equivalent to an information compression, the compression being "useful" only if the axes which have been discarded have been stated as conveying "useless" information [42].

6. Conclusions

The purpose of this paper is to introduce a few concepts, well known for people dealing with statistics, but rather underemployed in the analytical electron microscopy field. We have shown how various notions such as contrast, variance, information and signal/noise ratio are actually deeply nested. Concerning the problem of image processing (within the scope of the elemental mapping via electron energy loss spectroscopy technique), we have made a close inspection of how any processing induces an a priori agreement of explicit or implicit hypotheses, the information conveyed by these hypotheses being injected into the global (contrast and structural) information carried by the experiment. If these hypotheses are not well controlled, artifacts may be generated.

We suggest, therefore, handling the data through a two-step procedure: first analyze, then process. We have explained that there exist two analyzing tools which may greatly help the analyst in deciding which processing algorithm has to be preferred:

- The estimation of a relative entropy, computed with respect to a fictitious pure random, signal-free image, "mind-recorded" under exactly the same experimental conditions as the actual images under analysis. It is an unbiased tool for a quantitative estimation of the contrast image. This tool can be used, either on the whole image or only on a selected area, for a direct control of the contrast information variations after any step of the processing algorithm.
- The factorial analysis of correspondence can be run without any a priori hypothesis about the background and/or the noise; it merely analyzes the variance between the data (pixel intensities) of an image collection treated as a single set. The variance magnitudes are hierarchically classified. The final judgment of which part of the variance is "useful" is left to the analyst's responsibility, but several tools (absolute and relative contributions of the variables to the factorial axes) may help him in his decision.

In the companion article [15] these concepts are demonstrated on practical examples. The results are compared with those obtained by conventional procedures.

Acknowledgements

We are grateful to Drs. J. Cazaux, C. Colliex, C. Daly, D. Jeulin, P. Hannequin, C. Mory, D. Taupin and M. van Heel for numerous and lively discussions. We also acknowledge Dr. C. Lechêne

(Harvard Medical School) for providing the guinea pig specimen.

References

[1] P. Trebbia, Present state of the art in electron energy loss spectroscopy, in: Proc. 12th Int. Congr. on X-Ray Optics and Microanalysis, Cracow, Poland, 28 August - 1 September 1989 (1990) p. 489.

[2] C. Mory and C. Colliex, Ultramicroscopy 28 (1989) 339.

[3] C. Colliex, J.L. Maurice and D. Ugarte, Ultramicroscopy 29 (1989) 31.

[4] C. Colliex, D. Ugarte, Z.L. Wang, M. Gasgnier and V. Paul-Boncour, Surf. Interface Anal. 12 (1988) 3.

[5] P. Trebbia, C. Jeanguillaume and M. Walls, EELS: instrumentation, theory, practice and applications, in: Electron Microscope Imaging and Analysis for Biologists, Eds. P.W. Hawkes and U. Valdrè (Academic Press, London, 1990), in press.

[6] C. Jeanguillaume, Scanning Microsc. Int. 1 (1987) 437.

[7] C. Colliex, Inst. Phys. Conf. Ser. 93 (1988) 567.

[8] C. Colliex, C. Jeanguillaume, C. Mory and M. Tencé, Progress in electron energy loss spectroscopic imaging and analysing biological specimens with a field emission scanning transmission electron microscope, in: Electron Probe Microanalysis, Eds. K. Zierold and H.K. Hagler, Vol. 4 of Springer Series in Biophysics (Springer, Berlin, 1989).

[9] P. Trebbia, Scanning 12 (1990) 237.

[10] M. Jourlin, J.C. Pinoli and R. Zeboudj, J. Microscopy 156 (1989) 33.

[11] C.E. Shannon, Bell Syst. Tech. J. 27 (1948) 379, 623.

[12] C.E. Shannon and W. Weaver, The Mathematical Theory of Communication (University of Illinois Press, Urbana, 1949).

[13] E.T. Jaynes, Phys. Rev. 106 (1957) 620.

[14] S. Kullback, Information Theory and Statistics (Wiley, New York, 1959).

[15] P. Trebbia and C. Mory, Ultramicroscopy 34 (1990) 179.

[16] G.Y. Fan, Scanning Microsc. Suppl. 2 (1988) 157.

[17] G.Y. Fan, PhD Dissertation, Arizona State University, 1987.

[18] C. Daly, D. Jeulin and C. Lajaunie, Application of Multivariate Kriging to the Processing of Noisy Images, in: Proc. 3rd Int. Congr. of Geostatistics (Reidel, Dordrecht, 1988).

[19] D. van Dyck and W. Coene, J. Microsc. Spectrosc. Electr. 13 (1988) 463.

[20] M. Unser, A.C. Steven and B.L. Trus, Ultramicroscopy 18 (1986) 337.

[21] P. Trebbia, Ultramicroscopy 24 (1988) 399.

[22] P. Trebbia and T. Manoubi, Ultramicroscopy 28 (1989) 266.

[23] N. Bonnet, C. Colliex, C. Mory and M. Tencé, Scanning Microsc. Suppl. 2 (1988) 351.

[24] F.P. Ottensmeyer and J.W. Andrew, J. Ultrastruct. Res. 72 (1980) 336.

[25] A.P. Korn, P. Spitnik-Elson, D. Elson and P. Ottensmeyer, Eur. J. Cell. Biol. 31 (1983) 334.

[26] D.P. Bazett-Jones and M.L. Brown, in: Proc. 46th Annu. EMSA Meeting, Ed. G.W. Bailey (San Francisco Press, San Francisco, 1988) p. 172.

[27] J.J. Godleski, R.C. Stearns and E.J. Millet, in: Proc. 47th Annu. EMSA Meeting, Ed. G.W. Bailey (San Francisco Press, San Francisco, 1989) p. 404.

[28] H. Lehmann, A. Kramer, D. Schulz and W. Probst, Ultramicroscopy 32 (1990) 26.

[29] R. Egerton, Phil. Mag. 31 (1975) 199.

[30] C. Jeanguillaume, C. Colliex and P. Trebbia, Ultramicroscopy 3 (1978) 237.

[31] K. Pearson, Phil. Mag. 2 (1901) 559.

[32] C. Spearman, Am. J. Psychol. 15 (1904) 72, 201.

[33] J.P. Benzecri, in: Methodologies of Pattern Recognition, Ed. S. Watanabe (Academic Press, New York, 1969).

[34] J.P. Benzecri, L'Analyse des Données (Dunod, Paris, 1978).

[35] M. van Heel and J. Frank, in: Pattern Recognition in Practice, Eds. E.S. Gelsema and L.N. Kanal (North-Holland, Amsterdam, 1980) p. 235.

[36] M. van Heel and J. Frank, Ultramicroscopy 6 (1981) 187.

[37] J. Frank and M. van Heel, J. Mol. Biol. 161 (1982) 134.

[38] M. Prutton, M.M. El Gomati and C.G. Walker, Inst. Phys. Conf. Ser. 90 (1987) 1.

[39] M. Unser, B.L. Trus and A.C. Steven, Ultramicroscopy 30 (1989) 299;
M. van Heel, Optik 82 (1989) 114;
M. Schatz and M. van Heel, Ultramicroscopy 32 (1990) 255;
L. Borland and M. van Heel, J. Opt. Soc. Am. A 7 (1990) 601.

[40] P. Hannequin and N. Bonnet, Optik 81 (1988) 6.

[41] N. Bonnet and P. Hannequin, Ultramicroscopy 28 (1989) 248.

[42] J.P. Bretaudière and J. Frank, J. Microscopy 144 (1986) 1.