Journal of
Synchrotron Radiation
ISSN 1600-5775

Received 7 April 2014
Accepted 10 June 2014

Estimating the number of pure chemical components in a mixture by X-ray absorption spectroscopy

Alain Manceau,[a]* Matthew Marcus[b] and Thomas Lenoir[c]

[a]ISTerre, Université Grenoble Alpes and CNRS, F-38000 Grenoble, France, [b]Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, and [c]IFSTTAR, F-44344 Bouguenais, France. *E-mail: manceau@ujf-grenoble.fr

Principal component analysis (PCA) is a multivariate data analysis approach commonly used in X-ray absorption spectroscopy to estimate the number of pure compounds in multicomponent mixtures. This approach seeks to describe a large number of multicomponent spectra as weighted sums of a smaller number of component spectra. These component spectra are in turn considered to be linear combinations of the spectra from the actual species present in the system from which the experimental spectra were taken. The dimension of the experimental dataset is given by the number of meaningful abstract components, as estimated by the cascade or variance of the eigenvalues (EVs), the factor indicator function (IND), or the F-test on reduced EVs. It is shown on synthetic and real spectral mixtures that the performance of the IND and F-test critically depends on the amount of noise in the data, and may result in considerable underestimation or overestimation of the number of components even for a signal-to-noise (s/n) ratio of the order of 80 (σ = 20) in a XANES dataset. For a given s/n ratio, the accuracy of the component recovery from a random mixture depends on the size of the dataset and number of components, which is not known in advance, and deteriorates for larger datasets because the analysis picks up more noise components. The scree plot of the EVs for the components yields one or two values close to the significant number of components, but the result can be ambiguous and its uncertainty is unknown. A new estimator, NSS-stat, which includes the experimental error to XANES data analysis, is introduced and tested. It is shown that NSS-stat produces superior results compared with the three traditional forms of PCA-based component-number estimation. A graphical user-friendly interface for the calculation of EVs, IND, F-test and NSS-stat from a XANES dataset has been developed under LabVIEW for Windows and is supplied in the supporting information. Its possible application to EXAFS data is discussed, and several XANES and EXAFS datasets are also included for download.

Keywords: XANES; EXAFS; PCA; factor analysis; F-test.

1. Introduction

The advent of high-brilliance synchrotron radiation sources and improved X-ray optics has opened some unique possibilities for the micro- and nano-structural characterization of trace metals in heterogeneous organic and inorganic matrices with numerous applications in materials, biological and environmental science. With a mass sensitivity of about one part per hundred thousand (10 p.p.m. or 10 mg kg^{-1}) and a good sensitivity to chemical state and bonding environment, X-ray absorption spectroscopy (XAS) is one of the prime chemical and structural probes for the analysis of inhomogeneous samples. Ideally, one would want to record at least as many single-component XAS spectra as there are metal species in the heterogeneous matrix. However, this is rarely doable because the lack of uniformity often extends over a large range of spatial scales and every analyzed point has varying mixtures of two or more species of the same element. Because no two points generally are identical, inhomogeneous materials must be sampled at many points to be understood, with the accrued difficulty that the resultant spectrum at every point-of-interest (POI) is the weighted sum of the spectra of its component parts.

Multicomponent systems are analyzed traditionally in two steps using multivariate statistical analysis (Lochmüller & Reese, 1998; Malinowski, 2002; Brereton, 2003). In this context, the spectral dataset is analysed first by principal component analysis (PCA), also called factor analysis (Wasserman, 1997; Wasserman *et al.*, 1999; Ressler *et al.*, 2000; Frenkel *et al.*, 2002; Manceau *et al.*, 2002; Wang *et al.*, 2004). The main utility of PCA is to reduce the dimensionality of the dataset by approximating the spectra as weighted sums of a subset of linearly independent orthogonal eigenvectors, *i.e.* the principal components (PCs), that best explain the variability in the data. The pure component spectra (*i.e.* species identity) and their fractional contribution at each POI (*i.e.* species abundance) may be obtained subsequently by target transformation (Hopke, 1989; Isaure *et al.*, 2002; Struis *et al.*, 2004; Ryser *et al.*, 2005; Kirpichtchikova *et al.*, 2006; Manceau & Matynia, 2010; Donner *et al.*, 2011), iterative target factor analysis (ITFA) (Gampp *et al.*, 1986; Fernandez-Garcia *et al.*, 1995; Márquez-Alvarez *et al.*, 1997; Iglesias-Juez *et al.*, 2004; Anunziata *et al.*, 2011; Gemperline, 1984; Rossberg *et al.*, 2003) or multivariate curve resolution-alternating least squares (MCR-ALS) (de Juan & Tauler, 2006; Conti *et al.*, 2010; Jaumot *et al.*, 2005; Brugger, 2007; Testemale *et al.*, 2009) processing techniques applied to the PCA-reduced dataset.

Determining the number of statistically significant PCs that is necessary and sufficient to reconstruct all data within error is not straightforward. Three estimators are commonly used to differentiate PCs that contain statistically more variance than those that span noise: (i) the marginal decline of the eigenvalues from the PCA, which ranks PCs according to their importance in reproducing a dataset (% of variance), (ii) the semi-empirical indicator function IND (Malinowski, 2002) and (iii) Malinowski's F-test (Malinowski, 1988, 1990). According to the first estimator, a plot of the eigenvalues (λ_j) in descending order against the component number j, often called a scree plot (Cattell, 1966), shows a break at the number of meaningful PCs (x). The IND indicator is

$$\mathrm{IND}_j = \left(\frac{\sum_{k=j+1}^{m} \lambda_k}{p(m-j)^5} \right)^{1/2} \tag{1}$$

where m is the number of spectra in the dataset, $k \in [j+1, m]$ is the index to a particular component, and p is the number of data points for each spectrum. IND is calculated incrementally by varying the number of considered PCA components. The number of independent components (x) is the value of j for which IND is a minimum ($\mathrm{IND}_{\mathrm{min}}$). Malinowski's F-test is based on the observation that the following 'reduced eigenvalue' function REV is constant (a) for the secondary (*i.e.* non-significant) PCs (Malinowski, 1987),

$$\mathrm{REV}_j = \frac{\lambda_j}{(p-j+1)(m-j+1)} = a\sigma^2. \tag{2}$$

The test starts from the smallest eigenvalue λ_m and goes through the eigenvalues in increasing order until it finds the first significant one. Once one eigenvalue has been determined to be significant, all larger eigenvalues are also considered significant. The decision of what components correspond to the noise and what are the PCs is made on the basis of a Fisher test of the variance (σ^2) associated with eigenvalue λ_j and the summed variance associated with noise eigenvalues ($\lambda_{j+1}, \ldots, \lambda_m$). The jth component is accepted as a PC (jth = x) if the percentage of significance level (%SL) for the F-test is lower than some test level, generally 5% (Fernandez-Garcia *et al.*, 1995; Márquez-Alvarez *et al.*, 1997; Iglesias-Juez *et al.*, 2004; Anunziata *et al.*, 2011; Fay *et al.*, 1992; Menacherry *et al.*, 1997; Ruitenbeek *et al.*, 2000).

Although the three dimensionality estimators have been applied successfully in several studies (Fay *et al.*, 1992; Ruitenbeek *et al.*, 2000; Ciuparu *et al.*, 2005; Jalilehvand *et al.*, 2006; Kirpichtchikova *et al.*, 2006; Mah & Jalilehvand, 2008, 2010; Breynaert *et al.*, 2010; Manceau & Matynia, 2010; Wieland *et al.*, 2010), they frequently run into difficulties with over- or under-estimation of PCs (Manceau *et al.*, 2002; Panfili *et al.*, 2005; Sarret *et al.*, 2004; Márquez-Alvarez *et al.*, 1997; Conti *et al.*, 2010). Deviation from the real number of chemical species (r) notably occurs when the experimental noise is close to the variation in the data, the data normalized inconsistently from a spectrum to another, and some component species either minor, their fractional weight statistically constant in the dataset, or their spectral features indistinct (Fay *et al.*, 1992; Beauchemin *et al.*, 2002; Manceau *et al.*, 2002; Rossberg & Scheinost, 2005).

A situation commonly encountered in natural materials where multiple species are present in indeterminate number is shown in Fig. 1 and in Fig. S1 of the supporting information[1]. The example is a rhizospheric region from a paddy soil contaminated by a copper mine ($CuFeS_2$) discharge in Malaysia (Ali *et al.*, 2004). The soil matrix is a complex mixture of organic and inorganic Cu species which all three estimators fail to yield clear values for the number of significant components even with a large sampling of 99 independent µ-XANES spectra. Fig. 2 shows that the IND and F-test suggest that r = 31 and 14, respectively, whereas this number is expected to be ~5 based on elemental associations as seen from the X-ray fluorescence (XRF) map and visual inspection of the µ-XANES spectra (Fig. S1). The scree plot shows no large gap in eigenvalue at the jth = x followed by a plateau. However, the logarithmic first difference of the eigenvalues, $\Delta[\log(\lambda_{j+1}) - \log(\lambda_j)]$, shows two peaks at j = 4 and 6, which bracket the expected value of r = 5.

This article has two main goals: (i) to test the performance of the IND and F-test as a function of m and r in order to assess the robustness and improve the applicability of the two estimators to XAS, and (ii) to introduce a new estimator, NSS-stat, which resolves practical issues dealing with random noise. The accuracy of the IND and F-test is calculated first from the analysis of synthetic mixtures from theoretical XANES spectra. We show that the two estimators are most sensitive to data noise to the point where accuracy is decreased when the dataset is larger, thus when it contains theoretically more information about the mixture. The NSS-stat estimator is

[1] Supporting information for this paper is available from the IUCr electronic archives (Reference: HF5263).

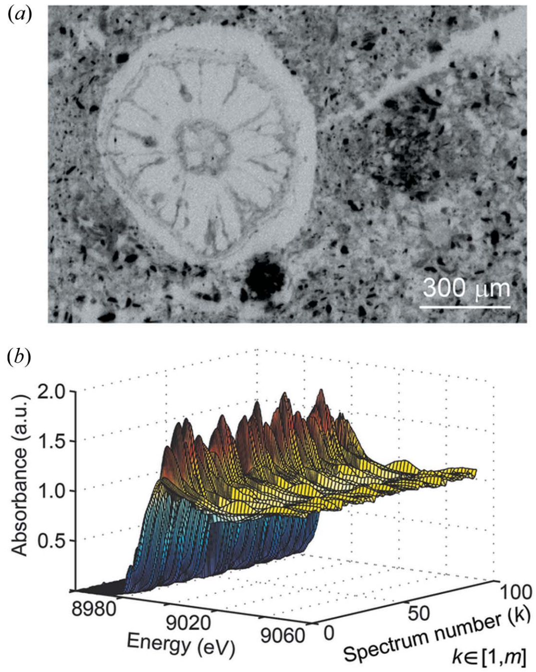

Figure 1
Dataset of 99 Cu *K*-edge XANES spectra collected on a thin section from a paddy soil. (*a*) Gray-scale micro X-ray fluorescence (XRF) map in negative contrast showing the inhomogeneous distribution of Cu which occurs in about six different organic and inorganic forms. Color maps are shown in Fig. S1 of the supporting information. Image size: 1.57 mm (H) × 1.04 mm (V); beam size and pixel size: 3.0 μm (H) × 3.0 μm (V). (*b*) Micro-XANES spectra collected at 99 spots having different compositions as seen from the XRF map. Data were recorded on beamline 10.3.2 at the ALS (Berkeley).

introduced next, and we show, also on theoretical mixtures, that it alleviates the intrusive noise problem. The performance of NSS-stat on the analysis of real data is discussed in the final section.

2. Synthetic mixtures of theoretical XANES spectra

Synthetic mixtures were made with random combinations of seven theoretical spectra: elemental Cu (Suh *et al.*, 1988), tetragonal (Janosi, 1964) and hexagonal (Evans, 1979) Cu_2S, CuS (Evans & Konnert, 1976), $CuFeS_2$ (Kratz & Fuess, 1989; Hall & Stewart, 1973), Cu_2O (Wyckoff, 1978) and CuO (Åsbrink & Norrby, 1970; Massarotti *et al.*, 1998). The Cu *K*-edge XANES spectra were calculated from −20 eV below the edge (E_0) to +60 eV above with a 0.2 eV step using the *FDMNES* software (Joly, 2001) (Fig. S2). Multicomponent spectra of *r* spectra were calculated over the *r* interval [3, 7] by drawing randomly with a uniform distribution each component's weight using the Mersenne Twister algorithm (Matsumoto & Nishimura, 1988). Each combination of *r* spectra was sampled *m* times over the interval [10, 105] (Fig. S3). Then a Gaussian noise was added randomly to the synthetic mixtures with the probability distribution $P(s/n = 80, \sigma = 20)$ (Sayers, 2000). Finally, the average values ($\langle x \rangle$) and standard errors [$\sigma(x)$] of IND_{min}, the first λ for which %SL > 5%, and NSS-stat were calculated from 100 replications of the $r \times m$ matrix (Fig. S3).

3. Evaluation of the three common estimators

3.1. Eigenvalues

Fig. 3 shows two scree plots of λ_j for an artificial dataset of 100 spectra ($m = 100$) and seven components ($r = 7$), with and without added noise. Eigenvalues for the unnoised dataset drop abruptly from approximately 1.0 to 10^{-14} between the seventh and eighth PC. In theory, the eighth and higher eigenvalues should be 0, so the 10^{-14} value reflects numerical inaccuracies. In contrast, no break is observed for the noised dataset: the graph shows an 'elbow' at $j = 8$ and the first PC accounts for most of the variance. The PCA sees the noise as a continuum of independent components, as is generally the case for real data. Two maxima are observed on $\Delta[\log(\lambda)]$ at $j = 5$ and 7, consistent with results from the paddy soil data (Fig. 2*a*).

3.2. IND and F-test

Based on the mathematical definition of the IND and F-test estimators, and the scree plot for noise-free data which shows that $\lambda \simeq 0$ for $j > r$ (Fig. 3*b*), it comes that the two estimators yield $x = r$ and $\sigma = 0$ regardless of the size of the dataset when analyzing noise-free data (Fig. 4). In the presence of noise, the color-contour maps of $\langle x(\mathrm{IND}) \rangle$ and $\langle x(\text{F-test}) \rangle$ as a function of *r* and *m* show strong variation with the number of spectra (*m*). For both estimators, $\langle x \rangle$ increases with *m*, and more rapidly the higher the value of *r*. The $\langle x(\mathrm{IND}) \rangle$ value increases from 3 to 34 with *m* in the interval [10, 105], and $\langle x(\text{F-test}) \rangle$

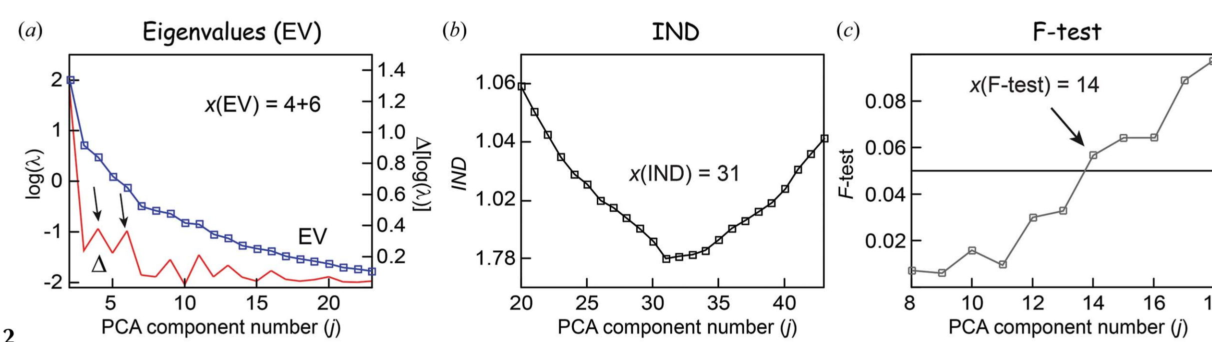

Figure 2
Estimation of the number of significant PCA components for the XANES dataset shown in Fig. 1. (*a*) Scree plot of the components and the associated eigenvalues (λ_j). (*b*) Malinowski's IND indicator (Malinowski, 1990). (*c*) Malinowski's F-test at 5% significance level (Malinowski, 1988).

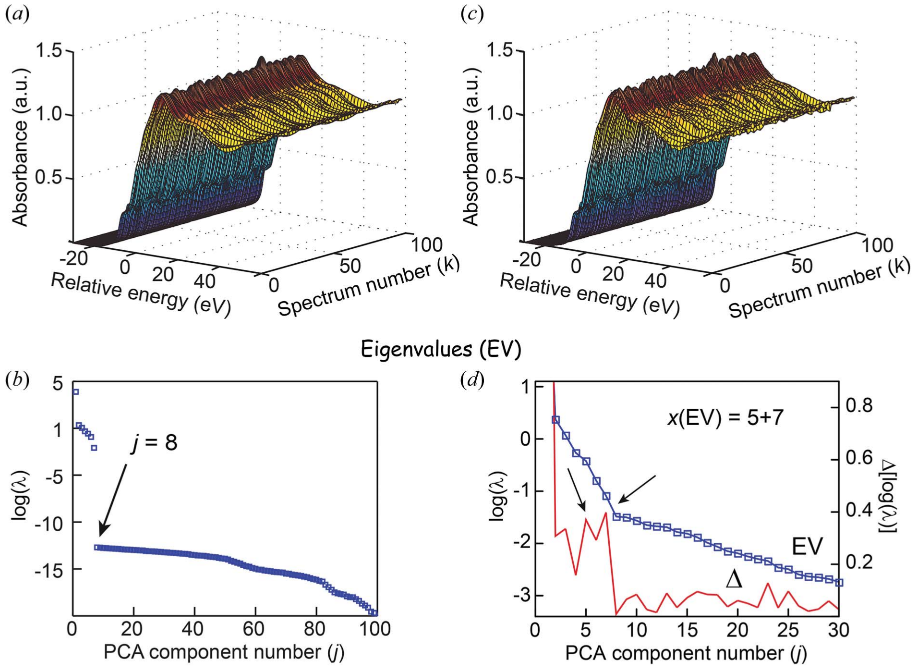

Figure 3
Effect of noise on the effective number of components (x) deduced from the decrease of eigenvalues. (a, b) 100 replications of a random mixture of seven pure components ($r = 7$, $m = 100$), and scree plot showing a clear separation between the 7 significant and 93 non-significant components. (c, d) Same theoretical set after addition of Gaussian noise with a s/n ratio of 80 and a standard error of 20. The scree plot shows an 'elbow' at $j = 8$, instead of nominally 7, and $\Delta[\log(\lambda)]$ shows two maxima at $j = 5$ and 7. For clarity, $j_{\max} = 30$.

increases from 2.4 to 17 over the same m interval. $\sigma[x(\text{IND})]$ is independent of r, and increases uniformly with m; it is 0.6 for $m < 20$ and as high as 7 for $m = 100$. $\sigma[x(\text{F-test})]$ increases irregularly with both r and m from 0.5 to 2.6. One might expect that the standard error should decrease when the size of the dataset (m) increases, and the estimator values should be closer to the correct number of components (r), which is not the case. The results can be understood as follows: the noise is overwhelming the variation in the data, and as the estimate becomes more accurate for larger m values it picks up more noise components. Although the two estimators give the correct estimates with no noise, they both fail once noise is added. Simulations showed that the lower the s/n ratio then the higher the effective number of components (x). However, the F-test varies less with m than IND, which indicates a lesser sensitivity to noise. The $x(\text{IND}) = 31$ and $x(\text{F-test}) = 14$ values of the experimental XANES spectra (Fig. 2, Table 1) are consistent with the predicted better performance of the F-test.

When $m < \sim 30$, the effective number of components (x) may now be lower than the real number (r). This situation is observed typically when r approaches 7 and m approaches 10, *i.e.* when the number of species required to describe a set of spectra is similar to the number of spectra. In this case, data undersampling considerably stretches the error bars to the point of rendering an apparently good agreement between reality and prediction fortuitous.

In summary, the accuracy of estimation of the number of independent components in a mixture critically depends on the amount of noise in the data, which is not known in advance. The number of components recovered with IND and the F-test generally is underestimated when $m < \sim 20$ and $r > \sim 6$, and is always overestimated above $m \simeq 30$. Therefore, there exists a (r, m) range in which the two estimators come close to the truth; however, because we do not know the correct answer exactly, the result generally cannot be trusted.

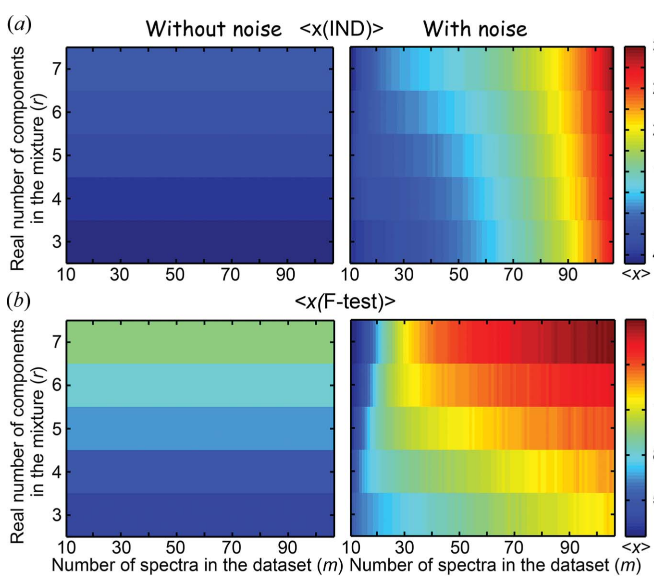

Figure 4
Effect of noise on the effective number of components (x) obtained by the IND and F-test estimators as a function of the number of pure components in the mixtures (r) and size of the dataset (m). (a) $\langle$IND$\rangle$. (b) $\langle$F-test$\rangle$. Mean values calculated from 100 replications of each (r, m) mixture (Fig. S3).

4. NSS-stat estimator

NSS is an abbreviation for Normalized Sum Squared difference. The idea of this estimator is to measure the degree to which a PCA fit to denoised spectra with x abstract components represents all original spectra in the dataset about equally well in comparison with the noise level of each spectrum. Keeping the same parameter definition as before, the PCA goodness-of-fit to the denoised spectra is given by

$$\text{NSS}_c^k(\text{denoised}) = \sum_{i=1}^{p}\left(\text{denoised}_i^k - \text{fit}_{i,c}^k\right)^2 \Big/ \sum_{i=1}^{p}\left(\text{denoised}_i^k\right)^2, \qquad (3)$$

Table 1
Summary of the PCA results.

All datasets are included for download in the supplementary information.

Dataset	r	m	λ	IND	F-test	NSS-stat	Γ	F_{cut}
Cu-XANES_th_3cp	3	100	3	28	12	3	2.0	0.4
Cu-XANES_th_7cp	7	100	5 + 7	29	18	7	2.0	0.4
Cu-XANES_th_3+7cp	3 + 7	70 + 30	4	29	18	4 + 7	2.0	0.4
Cu-XANES_Paddy-soil		99	4 + 6	14	31	4–5	2.0	0.4–0.7
Fe-XANES_Marcus_2014_Group1	5	209	6	31	18	7	1.5	0.3–0.4
Fe-XANES_Marcus_2014_Group2	4–5	126	6	15	12	6	1.5	0.5–1.0
Cu-EXAFS_Manceau_2010	3	27	3	3	3	4	0.3	3–5
Zn-EXAFS_vanDamme_2010	3	22	3	3	3	3	0.3	2.0
Zn-EXAFS_Isaure_2002	3	13	3	3	3	3	0.3	3–5

where $k \in [1, m]$ is the index to a particular spectrum, and $\mathrm{fit}_{i,c}^{k}$ is the fit to the kth spectrum using $c \in [1, m]$ PCA components, evaluated at the ith point.

The PCA NSS value *versus* spectrum number k (ordinate) and number of components c (abscissa) for the paddy soil dataset is shown in the upper center of Fig. 5. To analyze the PCA fit to the denoised spectra statistically, the sum-squared residuals calculated previously are normalized to the experimental error given by

$$\mathrm{NSS}^{k}(\mathrm{data}) = \sum_{i=1}^{p}\left(\mathrm{data}_{i}^{k} - \mathrm{denoised}_{i}^{k}\right)^{2} \Big/ \sum_{i=1}^{p}\left(\mathrm{data}_{i}^{k}\right)^{2}. \qquad (4)$$

The experimental error of each spectrum is represented in the lower center of Fig. 5 as a two-dimensional histogram of the residuals. Now, dividing NSS(denoised) by NSS(data) suppresses much of the x variability with the size of the dataset observed with IND and the F-test. The intensity graph of the noise-normalized NSS values (NSS-stat) as a function of k and c yields an estimate of the intrinsic dimensionality of the entire dataset at every spectrum in the dataset (upper right of Fig. 5). Then the global estimate (or an estimate over a particular subset) can be computed by averaging the local estimates over the whole dataset (or the subset in question). The plot in the lower right corner of Fig. 5 shows a scree plot of the standard deviation of NSS-stat. The logarithmic first difference of the eigenvalues produces a clear break at $c = 4$. The analysis of theoretical mixtures of known dimensionality showed that the real number of components is $r = x(\text{NSS-stat}) = c + 1$.

Because the denoising step is at the heart of the procedure, we will describe briefly the method we have used and is the one implemented in the PCA_Estimator.exe program included in the supporting information. The XANES abscissae in eV were transformed so that the minimum physically reasonable feature width, determined by the core-hole lifetime (Γ = 2 eV at the Cu K-edge, Table 1), and maximum path length for the EXAFS features in the extended XANES region (R_{max} = 5 Å) were roughly constant over the whole enrgy range. Then, a uniform Butterworth filter with adjustable cut-off frequency (F_{cut} parameter, Table 1) was applied

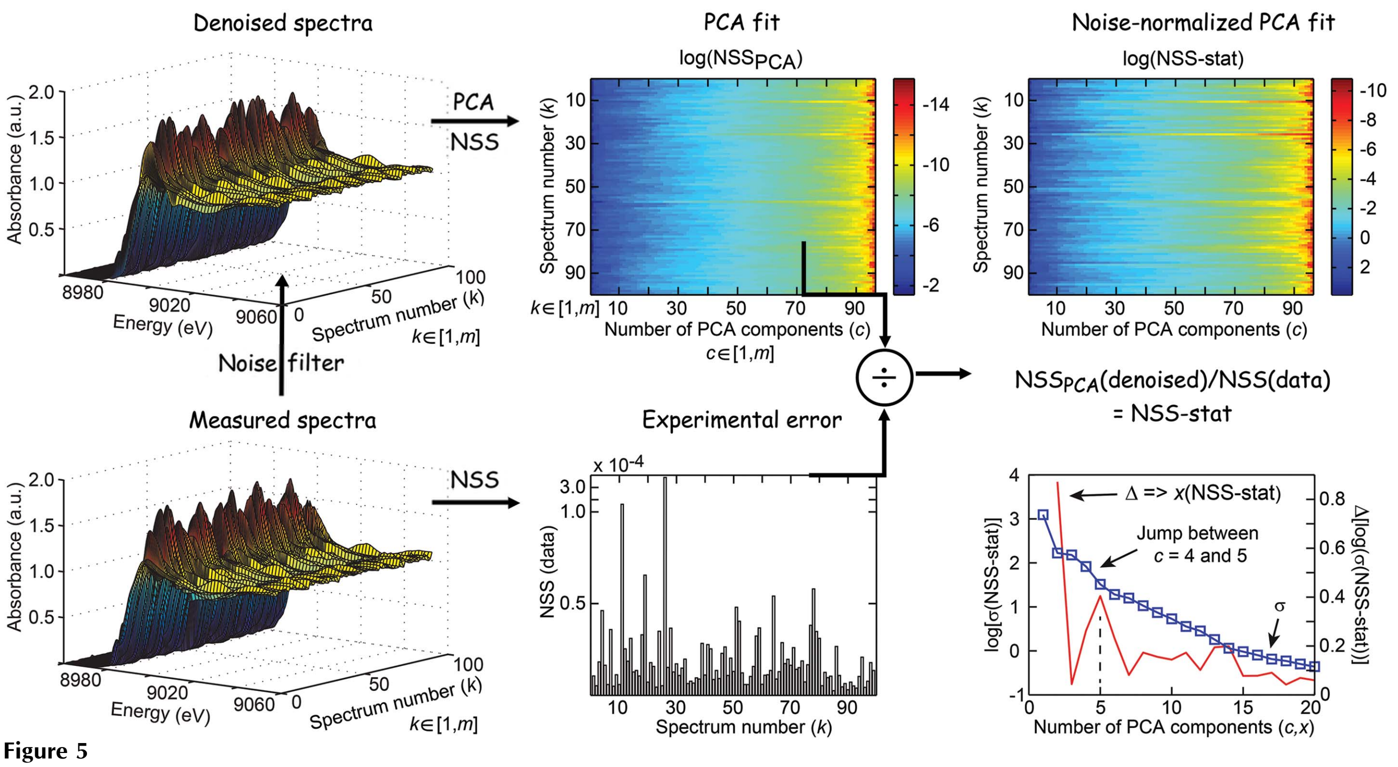

Figure 5
Flowchart definition of the NSS-stat estimator exemplified with the XANES dataset shown in Fig. 1. k is the spectrum number in the dataset ($k \in [1, m]$) and c is the number of PCA components used in the reconstruction ($c \in [1, m]$).

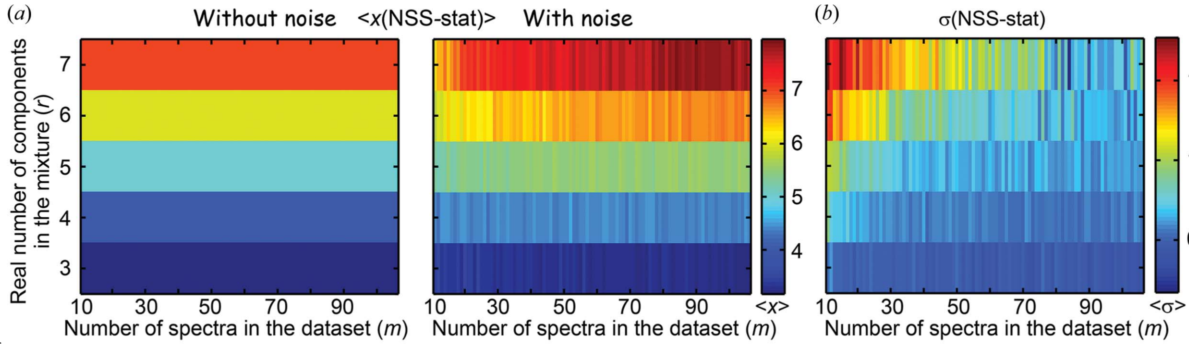

Figure 6
Evaluation of the NSS-stat estimator with the same set as in Fig. 4. Note the difference of the $\langle x \rangle$ maximum value, which decreases from $\sim$34 for IND to $\sim$17 for the F-test, and to $\sim$8 for NSS-stat.

and the coordinate transformation reversed. Details are given in the supporting information.

The performance of the NSS-stat estimator was assessed with theoretical mixtures over the whole r interval [3, 7] and m interval [10, 105], similarly to the IND and F-test (Fig. 6). Results show that NSS-stat gives the correct estimates to within 30% for $m < \sim 50$ and 20% above $m = 50$. In contrast to the IND and F-test, NSS-stat estimates vary very little across replications (small standard deviation), and the precision now increases as the size of the dataset increases, as expected.

The theoretical datasets used heretofore do not represent well all possible real situations because every component is present in every spectrum with a random probability of 0 to 1. In real data, there may exist a certain number of spectra from spots which have a certain set of components, and some other spots in which other components are included (Marcus & Lam, 2014). A theoretical example is shown in Fig. 7 for a binary set of ($r = 3, m = 70$) + ($r = 7, m = 30$) and P(s/n = 80, σ = 20). The main and outlier components yield two maxima at x(NSS-stat) = 4 and 7, with the highest peak at x(NSS-stat) = 7. The two populations give a single peak at x = 4 on the first difference of the log of EVs. The IND and F-test, which are by definition insensitive to coexisting groups of mixtures with variable proportions of a different number of species, give x(IND) = 29 and x(F-test) = 18 (Fig. 1 and Table 1).

5. Application to experimental XANES and EXAFS spectra

As demonstrated above, NSS-stat is a useful estimator for data analysis of multicomponent XANES spectra. It can help decide with some statistical significance how many abstract PCA components to use in PCA-based component-number estimation of a large dataset. The analysis of the datasets (Table 1) supplied in the supporting information with the four estimators confirms that NSS-stat outperforms the three traditional estimators. It also leads to the following commentaries.

(i) As noise level increases all estimators begin to suffer, but NSS-stat is, as expected, intrinsically less sensitive to noise than the three others.

(ii) It takes obviously less noise for the estimates to break down when the spectra are similar than when they are different (Levina *et al.*, 2007); still NSS-stat provides a better separation between significant and non-significant PCs.

(iii) The IND estimates of XANES data are systematically, and sometimes greatly, in excess of the actual values. The main reason for this difference is the higher sensitivity of this estimator to non-statistical noise (Malinowski, 1990; Maschmeyer *et al.*, 1995), which is always present in XANES spectra to some degree as a result of normalization uncertainties.

(iv) Any statistical method, such as the one introduced here, is limited in that in reality it only measures the dimensionality of the dataset, not the actual number of components in the sample. Consider a system which is actually a ternary but for which all the sampled points lie along a line in the ternary diagram. In that case the data are correctly described in terms of one variable, which can be related to the fraction of an end-member whose composition is represented by one end of the line. This situation, which occurs when the proportion of one component is fixed, has been described in natural materials (Manceau *et al.*, 2002). On the other hand, consider for example a set of data representing a binary mixture but affected by over-absorption, which causes a non-linear distortion of the spectra. In that case the data lie along a one-dimensional curve in a two-dimensional region, and a linear method such as PCA will report a dimensionality of at least 2. Similarly, the spectra of a binary solid-solution series may not be expressible as linear combinations of the end-member spectra, as may be the case when a Mn-XANES dataset contains Jahn-Teller Mn^{3+} cations (Manceau *et al.*, 2012), so again the dimensionality will be greater than 1.

The PCA-based NSS statistical method described here, and in more detail in the supporting information, has been developed and tested for normalized XANES spectra, including the EXAFS features in the extended XANES region, not for $k^n\chi(k)$-type EXAFS data. We have not yet developed similar methods for EXAFS which we consider as trustworthy as the one we described for XANES. One issue is that the s/n of an EXAFS spectrum tends to be strongly non-uniform because the counting time and amplitude change with k. A possible approach would be to take $1 + k^n\chi(k)$, convert to energy space the EXAFS data normalized to unit step, and analyze them as if they were XANES. Alternatively, one could leave them in k-space and filter with a uniform filter, since in most cases the lifetime broadening effect is small in the

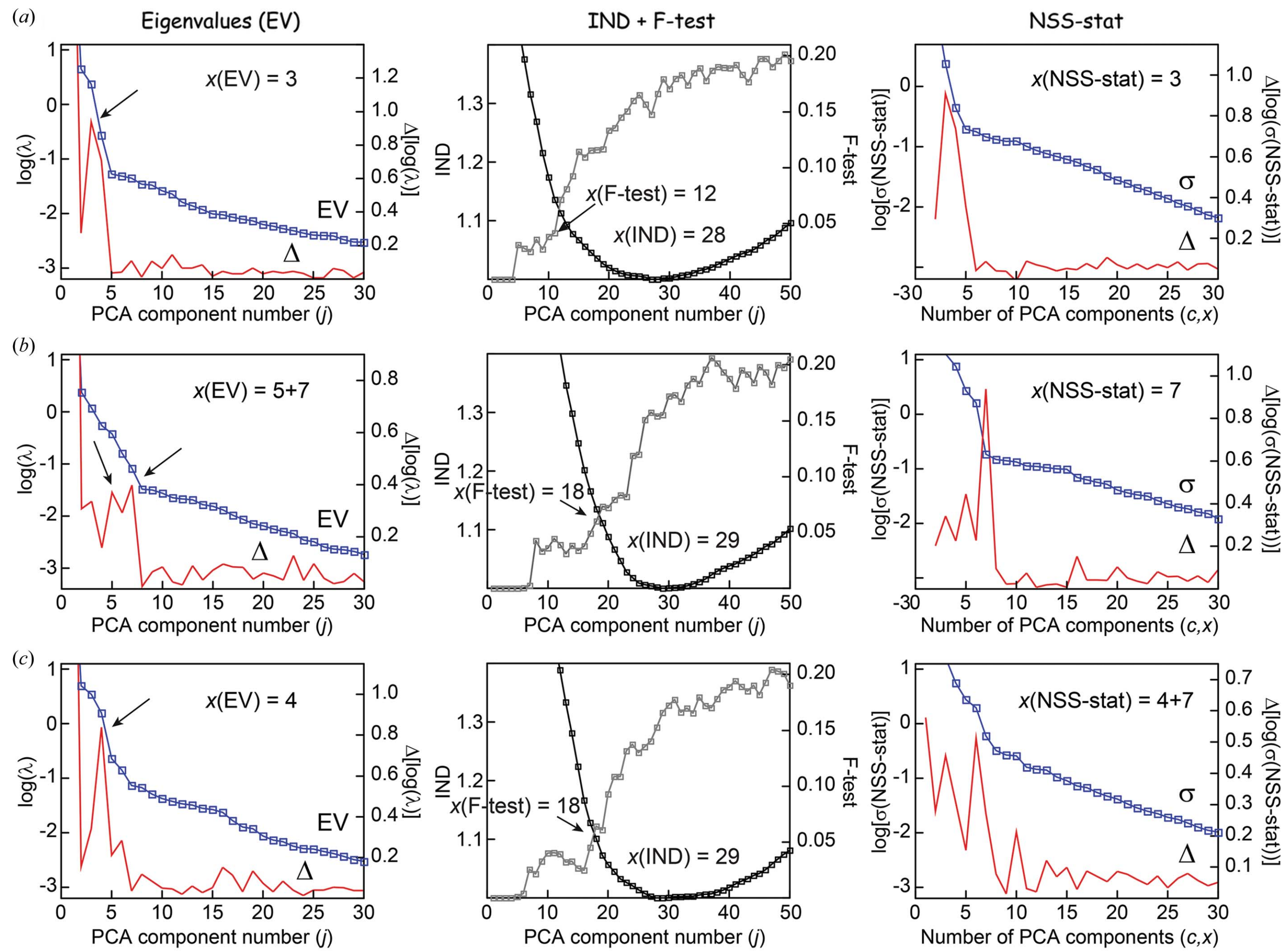

Figure 7
Comparative performance of the four PCA-based component-number estimators for 100 random amounts of three components (*a*) and seven components (*b*), and for a set composed of 70 mixtures of three components and 30 mixtures of seven components (*c*).

EXAFS region (Table 1). One may argue that the higher shells, which contribute to the narrowest features, are more attenuated at high *k* than the lower shells, due to disorder effects. In that case, and also because EXAFS data tend to be noisier at the high end, one might want to filter more severely (lower cut-off) at the high end. The amount of this effect would be an adjustable parameter. Three EXAFS datasets were, however, included in the supplementary information because the NSS method, although theoretically unsound, appeared to work in practice and yield the correct number of components (Table 1). Furthermore, the three traditional estimators included in the visualization program are general and can be applied to any type of data from a chemical mixture.

We thank Phoebe J. Lam for sharing her Fe-XANES datasets. The Advanced Light Source is supported by the Director, Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under Contract No. DE-AC02-05CH11231.

References

Ali, M. F., Heng, L. Y., Ratnam, W., Nais, J. & Ripin, R. (2004). *Bull. Environ. Contam. Toxicol.* **73**, 535–542.

Anunziata, O. A., Beltramone, A. R., Martinez, M. L., Giovanetti, L. J., Requejo, F. G. & Lede, E. (2011). *Appl. Catal. A*, **397**, 22–26.

Åsbrink, S. & Norrby, L.-J. (1970). *Acta Cryst.* B**26**, 8–15.

Beauchemin, S., Hesterberg, D. A. & Beauchemin, M. (2002). *Soil Sci. Soc. Am. J.* **66**, 83–91.

Brereton, R. G. (2003). *Chemometrics: Data Analysis for the Laboratory and Chemical Plants.* Chichester: John Wiley and Sons.

Breynaert, E., Scheinost, A. C., Dom, D., Rossberg, A., Vancluysen, J., Gobechiya, E., Kirschhock, C. E. A. & Maes, A. (2010). *Environ. Sci. Technol.* **44**, 6649–6655.

Brugger, J. (2007). *Comput. Geosci.* **33**, 248–261.

Cattell, R. B. (1966). *Multiv. Behav. Res.* **1**, 245–276.

Ciuparu, D., Haider, P., Fernandez-Garcia, M., Chen, Y., Lim, S., Haller, G. L. & Lisa Pfefferle, L. (2005). *J. Phys. Chem. B*, **109**, 16332–16339.

Conti, P., Zamponi, S., Giorgetti, M., Berrettoni, M. & Smyrl, W. H. (2010). *Anal. Chem.* **82**, 3629–3635.

Donner, E., Howard, D. L., de Jonge, M. D., Paterson, D., Cheah, M. H., Naidu, R. & Lombi, E. (2011). *Environ. Sci. Technol.* **45**, 7249–7257.

Evans, H. T. (1979). *Z. Kristallogr.* **150**, 299–320.

Evans, H. T. & Konnert, J. A. (1976). *Am. Miner.* **61**, 996–1000.
Fay, M. J., Proctor, A., Hoffmann, D. P., Houalla, M. & Hercules, D. M. (1992). *Microchim. Acta*, **109**, 281–293.
Fernandez-Garcia, M., Marquez Alvarez, C. & Haller, G. L. (1995). *J. Phys. Chem.* **99**, 12565–12569.
Frenkel, A. I., Kleifeld, O., Wasserman, S. R. & Sagi, I. (2002). *J. Chem. Phys.* **116**, 9449.
Gampp, H., Maeder, M., Meyer, C. J. & Zuberbühler, A. D. (1986). *Talanta*, **33**, 943–951.
Gemperline, P. J. (1984). *J. Chem. Inform. Comput. Sci.* **24**, 206–212.
Hall, S. R. & Stewart, J. M. (1973). *Acta Cryst.* B**29**, 579–585.
Hopke, P. K. (1989). *Chemometr. Intell. Lab. Syst.* **6**, 7–19.
Iglesias-Juez, A., Martinez-Arias, A., Hungria, A. B., Anderson, J. A., Conesa, J. C., Soria, J. & Fernandez-Garcia, M. (2004). *Appl. Catal. A*, **259**, 207–220.
Isaure, M. P., Laboudigue, A., Manceau, A., Sarret, G., Tiffreau, C., Trocellier, P., Hazemann, J. L. & Chateigner, D. (2002). *Geochim. Cosmochim. Acta*, **66**, 1549–1567.
Jalilehvand, F., Leung, B. O., Izadifard, M. & Damian, E. (2006). *Inorg. Chem.* **45**, 66–73.
Janosi, A. (1964). *Acta Cryst.* **17**, 311–312.
Jaumot, J., Gargallo, R., de Juan, A. & Tauler, R. (2005). *Chemom. Intell. Lab. Syst.* **76**, 101–110.
Joly, Y. (2001). *Phys. Rev. B*, **63**, 125120.
Juan, A. de & Tauler, R. (2006). *Crit. Rev. Anal. Chem.* **36**, 163–176.
Kirpichtchikova, T. A., Manceau, A., Spadini, L., Panfili, F., Marcus, M. A. & Jacquet, T. (2006). *Geochim. Cosmochim. Acta*, **70**, 2163–2190.
Kratz, T. & Fuess, H. (1989). *Z. Kristallogr.* **186**, 167–169.
Levina, E., Wagaman, A. S., Callender, A. F., Mandair, G. S. & Morris, M. D. (2007). *J. Chemom.* **21**, 24–34.
Lochmüller, C. H. & Reese, C. E. (1998). *Crit. Rev. Anal. Chem.* **28**, 21–49.
Mah, V. & Jalilehvand, F. (2008). *J. Biol. Inorg. Chem.* **13**, 541–553.
Mah, V. & Jalilehvand, F. (2010). *Chem. Res. Toxicol.* **23**, 1815–1823.
Malinowski, E. R. (1987). *J. Chemom.* **1**, 33–40.
Malinowski, E. R. (1988). *J. Chemom.* **3**, 49–60.
Malinowski, E. R. (1990). *J. Chemom.* **4**, 102.
Malinowski, E. R. (2002). *Factor Analysis in Chemistry*, 3rd ed. New York: John Wiley and Sons.
Manceau, A., Marcus, M. A. & Grangeon, S. (2012). *Am. Miner.* **97**, 816–827.
Manceau, A., Marcus, M. A. & Tamura, N. (2002). *Rev. Mineral. Geochem.* **49**, 341–428.
Manceau, A. & Matynia, A. (2010). *Geochim. Cosmochim. Acta*, **74**, 2556–2580.
Marcus, M. A. & Lam, P. J. (2014). *Environ. Chem.* **11**, 10–17.
Márquez-Alvarez, C., Rodríguez-Ramos, I., Guerrero-Ruiz, A., Haller, G. L. & Fernández-García, M. (1997). *J. Am. Chem. Soc.* **119**, 2905–2914.
Maschmeyer, T., Rey, F., Sankar, G. & Thomas, J. M. (1995). *Nature (London)*, **378**, 159–162.
Massarotti, V., Capsoni, D., Bini, M., Altomare, A. & Moliterni, A. G. G. (1998). *Z. Kristallogr.* **213**, 259–265.
Matsumoto, M. & Nishimura, T. (1988). *ACM Trans. Model. Comput. Simul.* **8**, 3–30.
Menacherry, P. V., Fernandez-Garcia, M. & Haller, G. L. (1997). *J. Catal.* **166**, 75–88.
Panfili, F., Manceau, A., Sarret, G., Spadini, L., Kirpichtchikova, T., Bert, V., Laboudigue, A., Marcus, M. A., Ahamdach, N. & Libert, M. F. (2005). *Geochim. Cosmochim. Acta*, **69**, 2265–2284.
Ressler, T., Wong, J., Roos, J. & Smith, I. L. (2000). *Environ. Sci. Technol.* **34**, 950–958.
Rossberg, A., Reich, T. & Bernhard, G. (2003). *Anal. Bioanal. Chem.* **376**, 631–638.
Rossberg, A. & Scheinost, A. C. (2005). *Anal. Bioanal. Chem.* **383**, 56–66.
Ruitenbeek, M., van Dillen, A. J., de Groot, F. M. F., Wachs, I. E., Geus, J. W. & Koningsberger, D. C. (2000). *Top. Catal.* **10**, 241–254.
Ryser, A. L., Strawn, D. G., Marcus, M. A., Johnson-Maynard, J. L., Gunter, M. E. & Möller, G. (2005). *Geochem. Trans.* **6**, 1–10.
Sarret, G., Balesdent, J., Bouziri, L., Garnier, J. M., Marcus, M. A., Geoffroy, N., Panfili, F. & Manceau, A. (2004). *Environ. Sci. Technol.* **38**, 2792–2801.
Sayers, D. E. (2000). *Error Reporting Recommendations: A Report of the Standards and Criteria Committee*, http://ixs.iit.edu/subcommittee_reports/sc/err-rep.pdf.
Struis, R. P. W. J., Ludwig, C., Lutz, H. & Scheidegger, A. M. (2004). *Environ. Sci. Technol.* **38**, 3760–3767.
Suh, I. K., Ohta, H. & Waseda, Y. (1988). *J. Mater. Sci.* **23**, 757–760.
Testemale, D., Brugger, J., Liu, W., Etschmann, B. & Hazemann, J. L. (2009). *Chem. Geol.* **264**, 295–310.
Wang, X. Q., Hanson, J. C., Frenkel, A. I., Kim, J. Y. & Rodriguez, J. A. (2004). *J. Phys. Chem. B*, **108**, 13667–13673.
Wasserman, S. R. (1997). *J. Phys. IV*, **7**, 203–205.
Wasserman, S. R., Allen, P. G., Shuh, D. K., Bucher, J. J. & Edelstein, N. M. (1999). *J. Synchrotron Rad.* **6**, 284–286.
Wieland, E., Dähn, R., Vespa, M. & Lothenbach, B. (2010). *Cem. Conc. Res.* **40**, 885–891.
Wyckoff, R. W. G. (1978). *Crystal Structures.* New York: Interscience.