JOURNAL OF APPLIED CRYSTALLOGRAPHY

ISSN: 1600-5767

journals.iucr.org/j

Tailored multivariate analysis for modulated enhanced diffraction

Rocco Caliandro, Pietro Guccione, Giovanni Nico, Goknur Tutuncu and Jonathan C. Hanson

J. Appl. Cryst. (2015). **48**, 1679–1691

JOURNAL OF APPLIED CRYSTALLOGRAPHY

ISSN 1600-5767

Received 24 June 2015
Accepted 11 September 2015

Edited by Th. Proffen, Oak Ridge National Laboratory, USA

Keywords: multivariate analysis; X-ray powder diffraction; modulated enhanced diffraction.

Supporting information: this article has supporting information at journals.iucr.org/j

Tailored multivariate analysis for modulated enhanced diffraction

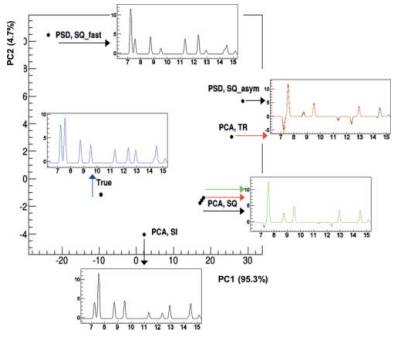
Rocco Caliandro,[a]* Pietro Guccione,[b] Giovanni Nico,[c] Goknur Tutuncu[d] and Jonathan C. Hanson[e]

[a]Institute of Crystallography, CNR, via Amendola 122/o, Bari, 70126, Italy, [b]Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, via Orabona 4, Bari 70125, Italy, [c]Istituto per le Applicazione del Calcolo 'Mauro Picone', CNR, via Amendola 122/o, Bari 70126, Italy, [d]NSLS II, Photon Science Division, Brookhaven National Laboratory, PO Box 5000, Upton, NY 11973-5000, USA, and [e]Chemistry Department, Brookhaven National Laboratory, PO Box 5000, Upton, NY 11973-5000, USA. *Correspondence e-mail: rocco.caliandro@ic.cnr.it

Modulated enhanced diffraction (MED) is a technique allowing the dynamic structural characterization of crystalline materials subjected to an external stimulus, which is particularly suited for *in situ* and *operando* structural investigations at synchrotron sources. Contributions from the (active) part of the crystal system that varies synchronously with the stimulus can be extracted by an offline analysis, which can only be applied in the case of periodic stimuli and linear system responses. In this paper a new decomposition approach based on multivariate analysis is proposed. The standard principal component analysis (PCA) is adapted to treat MED data: specific figures of merit based on their scores and loadings are found, and the directions of the principal components obtained by PCA are modified to maximize such figures of merit. As a result, a general method to decompose MED data, called optimum constrained components rotation (OCCR), is developed, which produces very precise results on simulated data, even in the case of nonperiodic stimuli and/or nonlinear responses. The multivariate analysis approach is able to supply in one shot both the diffraction pattern related to the active atoms (through the OCCR loadings) and the time dependence of the system response (through the OCCR scores). When applied to real data, OCCR was able to supply only the latter information, as the former was hindered by changes in abundances of different crystal phases, which occurred besides structural variations in the specific case considered. To develop a decomposition procedure able to cope with this combined effect represents the next challenge in MED analysis.

1. Introduction

Recently a technique has been proposed, which is able to capture structural features of a crystal system varying in phase with an external stimulus (Chernyshov *et al.*, 2011). This technique is called modulated enhanced diffraction (MED), and it is based on the joint analysis of a series of X-ray diffraction patterns, collected while varying the external stimulus (van Beek *et al.*, 2012). Thanks to the advent of modern synchrotron experimental setups, where higher X-ray brilliance is coupled with fast readout and low-noise area detectors, today complete diffraction patterns can be collected within minutes, even for poorly diffracting crystal systems. This has opened the possibility of *in situ* and *operando* crystallographic studies, where specific chemical or physical processes can be monitored. In this context, the classical outcome of the crystallographic analysis, *i.e.* a detailed structural knowledge of a given state of the crystal system, has been complemented by a dynamical characterization of the structural features of the system while its state is varying in time.



MED appears to be the perfect technique to respond to this new demand, and in fact new beamlines are being constructed, with the capability to perform MED measurements, *e.g.* the NSLS-II XPD (http://www.bnl.gov/ps/nsls2/beamlines/XPD.php), although few experiments have been performed so far (Caliandro *et al.*, 2012; van Beek *et al.*, 2012; Ferri *et al.*, 2013; Lu *et al.*, 2014; Ferri *et al.*, 2014; Palin *et al.*, 2015).

In its original formulation, this technique employed phase sensitive detection (PSD) for demodulating series of diffraction patterns and obtaining one that is representative of the part of the system varying with the stimulus (active part) (Caliandro *et al.*, 2012). Although the technique is very accurate and informative in ideal cases, PSD suffers from important limitations. Besides requiring separation between peaks in the case of powder patterns, basically it requires a periodic stimulus and a linear response of the crystal system. In other words, the response of the system must follow the same trend as the stimulus, and its variations in time should be symmetrical in time. This strongly limits the application of the MED technique, even considering the fact that the response of the system is unknown before doing the analysis, and often structural features underlying key properties in functional and engineering materials have a nonlinear behaviour.

In order to overcome such limitations, multivariate analysis has been applied to MED data (Milanesio *et al.*, 2014). The PSD demodulation has thus been replaced by a matrix decomposition aiming at extracting the part of the system response which varies in time. First applications of multivariate analysis showed the undoubted advantage of allowing MED analysis on systems whose response is not perfectly linear with respect to the stimulus, but, when applied to symmetric responses, the decomposition was more approximate than the PSD demodulation (Palin *et al.*, 2015).

In this paper we have investigated the MED theory to find proper conditions to be applied to multivariate methods in order to perform a more precise decomposition. In detail, we have modified the principal component decomposition to provide a new algorithm, called optimum constrained components rotation. With this algorithm the components extracted using the principal component technique are further subjected to a constrained optimization to refine the result. The objective functions found to solve the problem respond to conditions derived from the MED theory and so they result in a more precise solution than the one given by simple blind source decomposition. The derived conditions are also used as criteria for the assessment of the decomposition quality. The capabilities of the new algorithm have been demonstrated by using simulated and real data which otherwise would have not been demodulated effectively by the PSD.

2. The MED theory for powder diffraction

Suppose a set of X-ray powder diffraction (XPD) patterns is collected by changing an external parameter called the stimulus. The resulting data matrix can be described by the function $A(2\theta, t)$, which gives the measured profile as a function of the diffraction angle 2θ for each value of time t, *i.e.* for each value of the stimulus. $A(2\theta, t)$ can be written as

$$A(2\theta, t) = \sum_{\mathbf{h}} m_{\mathbf{h}} L_{\mathbf{h}} \left|F_{\mathbf{h}}(t)\right|^2 f(2\theta, 2\theta_{\mathbf{h}}, t) + b(2\theta, t), \qquad (1)$$

where $F_{\mathbf{h}}(t)$, $m_{\mathbf{h}}$ and $L_{\mathbf{h}}$ are, respectively, the structure factor, multiplicity and Lorentz–Polarization factor of the reflection $\mathbf{h}$, $f(2\theta, 2\theta_{\mathbf{h}})$ is a function describing the peak centred at $2\theta_{\mathbf{h}}$ (typically a Gaussian or Lorentzian shape), and $b(2\theta)$ is a function describing the background.

Suppose now that part of the crystal structure is changing with the stimulus, *i.e.* some (active) atoms vary their crystallographic parameters in phase with the variations of the stimulus, while the remaining (spectator) atoms are not affected by the stimulus. Then, the amplitude of each reflection can be parameterized as follows:

$$\begin{aligned}\left|\mathbf{F}_{\mathbf{h}}(t)\right|^2 &= \left|\left\langle\mathbf{F}_{\mathbf{h}\mathrm{A}}\right\rangle + \delta\mathbf{F}_{\mathbf{h}\mathrm{A}}(t) + \mathbf{F}_{\mathbf{h}\mathrm{S}}\right|^2 \\ &= \left|\delta\mathbf{F}_{\mathbf{h}\mathrm{A}}(t)\right|^2 + \left|\left\langle\mathbf{F}_{\mathbf{h}\mathrm{A}}\right\rangle + \mathbf{F}_{\mathbf{h}\mathrm{S}}\right|^2 \\ &\quad + 2\left|\delta\mathbf{F}_{\mathbf{h}\mathrm{A}}(t)\right|\left|\left\langle\mathbf{F}_{\mathbf{h}\mathrm{A}}\right\rangle + \mathbf{F}_{\mathbf{h}\mathrm{S}}\right|\cos(\varphi_{\mathrm{A}} - \varphi_{\mathrm{SA}}), \qquad (2)\end{aligned}$$

where subscripts A and S indicate active and spectator sublattices, respectively, the structure factor for active atoms $F_{\mathbf{h}\mathrm{A}}(t)$ has been written as the sum of an average value $\langle\mathbf{F}_{\mathbf{h}\mathrm{A}}\rangle$ and a time-dependent term $\delta\mathbf{F}_{\mathbf{h}\mathrm{A}}$, φ_{A} is the phase value of $\langle\mathbf{F}_{\mathbf{h}\mathrm{A}}\rangle$ and $\delta\mathbf{F}_{\mathbf{h}\mathrm{A}}$, and φ_{SA} is the phase value of $\langle\mathbf{F}_{\mathbf{h}\mathrm{A}}\rangle + \mathbf{F}_{\mathbf{h}\mathrm{S}}$. A vector representation of equation (2) is given in Fig. 1. It is worth noting our assumption that during the MED experiment the vector $\delta\mathbf{F}_{\mathbf{h}\mathrm{A}}$ only changes its modulus, while its phase φ_{A} remains constant: such an assumption is specific for the modulations that will be considered in the following. If we assume the same variations for all the active atoms, the corresponding variations of the structure factors can be written as

$$\delta\mathbf{F}_{\mathbf{h}\mathrm{A}} = \delta P(t)\left\langle\mathbf{F}_{\mathbf{h}\mathrm{A}}\right\rangle, \qquad (3)$$

where δP is $\delta n(t)/\langle n\rangle$ in the case of occupancy (n) variations, $[\delta f'(t) + i\delta f''(t)]/(\langle f'\rangle + i\langle f''\rangle)$ if the real (f') and imaginary

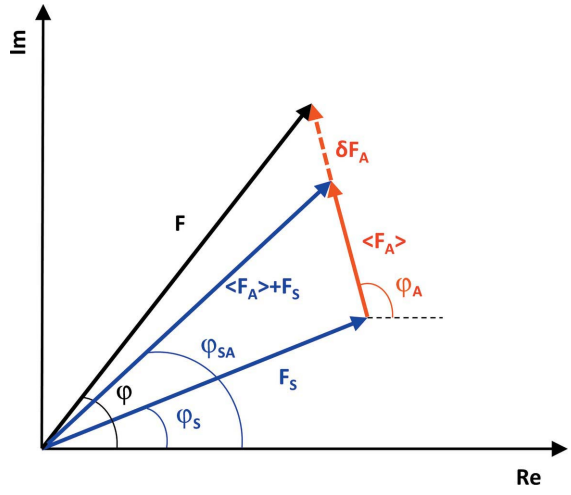


Figure 1
Vector representation of the structure factors for active (A) and spectator (S) atoms in a MED experiment.

(f'') parts of the atomic scattering factors vary, or $-\delta B(t)h^2$ in the case of small thermal factor (B) variations, so that $\delta B h^2 \ll 1$ (Caliandro *et al.*, 2013). In the general case, such variations can be expressed by

$$\delta P = |\delta P| g(t), \tag{4}$$

where $g(t)$ is a function that describes the time behaviour of the response of the crystal system to the stimulus variations. It is worth noticing that variations of the atomic coordinates can be hardly described by equation (3), and hence such variations cannot in general be recovered by MED. By relating equations (2), (3), and (4), equation (1) can be rewritten as

$$\begin{aligned} A(2\theta, t) = \sum_{\mathbf{h}} & m_{\mathbf{h}} L_{\mathbf{h}} f(2\theta, 2\theta_{\mathbf{h}}, t) \\ & \times \left\{ \left[|\delta P| \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle g(t) \right]^2 + \left| \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle + \mathbf{F}_{\mathbf{h}\mathrm{S}} \right|^2 \right. \\ & \left. + 2 |\delta P| \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle g(t) \left| \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle + \mathbf{F}_{\mathbf{h}\mathrm{S}} \right| \cos(\varphi_{\mathrm{A}} - \varphi_{\mathrm{SA}}) \right\} \\ & + b(2\theta, t), \end{aligned} \tag{5}$$

which means that the data matrix, once the background is subtracted, can be written as the sum of three terms, each having specific time dependence:

$$A(2\theta, t) - b(2\theta, t) = R(2\theta, t) g(t)^2 + S(2\theta, t) g(t) + T(2\theta, t). \tag{6}$$

Here

$$\begin{cases} R(2\theta, t) = \sum_{\mathbf{h}} m_{\mathbf{h}} L_{\mathbf{h}} f(2\theta, 2\theta_{\mathbf{h}}, t) |\delta P|^2 \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle^2, \\ S(2\theta, t) = \sum_{\mathbf{h}} m_{\mathbf{h}} L_{\mathbf{h}} f(2\theta, 2\theta_{\mathbf{h}}, t) 2 |\delta P| \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle \left| \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle + \mathbf{F}_{\mathbf{h}\mathrm{S}} \right| \\ \qquad \times \cos(\varphi_{\mathrm{A}} - \varphi_{\mathrm{SA}}), \\ T(2\theta, t) = \sum_{\mathbf{h}} m_{\mathbf{h}} L_{\mathbf{h}} f(2\theta, 2\theta_{\mathbf{h}}, t) \left| \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle + \mathbf{F}_{\mathbf{h}\mathrm{S}} \right|^2. \end{cases} \tag{7}$$

The first term of equation (7) has contributions only from active atoms, and $R(2\theta, t)$ can be interpreted as the set of XPD profiles that would be measured if the crystal system were constituted by active atoms only. The second term depends on both active and spectator atoms, so $S(2\theta, t)$ gives information about the interaction between the active and spectator sublattices. The third term has contribution from the part of the structure factors which does not vary with time.

In accordance with the hypothesis that the peak shape function and the peak position do not vary with time, *i.e.* the crystallite size, defectivity and cell parameters are not affected by the stimulus, equation (6) can be rewritten as

$$A(2\theta, t) - b(2\theta, t) = R(2\theta) g(t)^2 + S(2\theta) g(t) + T(2\theta) \tag{8}$$

and $R(2\theta)$ represents the diffraction intensities as determined by the averaged crystallographic parameters of the active atoms.

2.1. Demodulation by phase sensitive detection

If the stimulus is periodic with period T_{p}, one way to extract dynamic information from the set of measured patterns is to demodulate it through the following integral:

$$A_k(2\theta, \varphi) = \frac{2}{T_{\mathrm{p}}} \int_0^{T_{\mathrm{p}}} [A(2\theta, t) - b(2\theta, t)] \sin(k\omega t + \varphi) \, \mathrm{d}t. \tag{9}$$

It implements the concept of the lock-in amplifier, in which the (background-subtracted) response of the system $A(2\theta, t) - b(2\theta, t)$ is multiplied by a reference signal $\sin(k\omega t + \varphi)$, integrated over the period of the stimulus T_{p} and normalized by the factor of $2/T_{\mathrm{p}}$. The demodulated signal $A_k(2\theta, \varphi)$ only contains the components of $A(2\theta, t) - b(2\theta, t)$ varying at the same frequency as the reference signal. This technique is called phase sensitive detection (PSD) and is extensively used in spectroscopy to select the portions of the spectra that change under the influence of the periodic stimulus (see for example Urakawa & Baiker, 2006). Under the same hypotheses as equation (8) and by using equation (1), equation (9) can be rewritten as

$$A_k(2\theta, \varphi) = \sum_{\mathbf{h}} m_{\mathbf{h}} L_{\mathbf{h}} f(2\theta, 2\theta_{\mathbf{h}}) \frac{2}{T_{\mathrm{p}}} \int_0^{T_{\mathrm{p}}} \left| F_{\mathbf{h}}(t) \right|^2 \sin(k\omega t + \varphi) \, \mathrm{d}t. \tag{10}$$

Hereafter, we only consider the time-independent quantity

$$A_k(\mathbf{h}, \varphi) = \int_0^{T_{\mathrm{p}}} \left| F_{\mathbf{h}}(t) \right|^2 \sin(k\omega t + \varphi) \, \mathrm{d}t. \tag{11}$$

The demodulation integral in equation (9), applied to equation (2), extracts only the time-dependent terms and suppresses the time-independent contributions arising from $\mathbf{F}_{\mathbf{h}\mathrm{S}}$ and $\langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle$, leading to

$$\begin{aligned} A_k(\mathbf{h}, \varphi) = & \left[|\delta P| \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle \right]^2 \int_0^{T_{\mathrm{p}}} g^2(t) \sin(k\omega t + \varphi) \, \mathrm{d}t \\ & + \left[2 |\delta P| \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle \left| \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle + \mathbf{F}_{\mathbf{h}\mathrm{S}} \right| \cos(\varphi_{\mathrm{A}} - \varphi_{\mathrm{SA}}) \right] \\ & \times \int_0^{T_{\mathrm{p}}} g(t) \sin(k\omega t + \varphi) \, \mathrm{d}t. \end{aligned} \tag{12}$$

It has been demonstrated (Chernyshov *et al.*, 2011) that if the periodic function $g(t)$ can be expanded in a Taylor series with only odd (even) coefficients, *i.e.* it has an exact odd (even) symmetry in time, then $A_2(\mathbf{h}, \varphi)$ will be proportional to $[|\delta P| \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle]^2$, *i.e.* the signal demodulated at $k = 2$ will only contain contributions from active atoms. For example, if $g(t) = \sin(\omega t + \delta)$, then $A_2(\mathbf{h}, \varphi) = (T/4) \sin(2\delta - \varphi) \times [|\delta P| \langle \mathbf{F}_{\mathbf{h}\mathrm{A}} \rangle]^2$. As a consequence, by considering equations (7) and (10), the demodulated term $A_2(2\theta, \varphi)$ will be proportional to the above-defined function $R(2\theta)$. Therefore, PSD is able to recover the diffraction intensities as determined by the averaged crystallographic parameters of the active atoms, by using the fact that atomic variations occurring at the same frequency as the external stimulus are effectively varying with double this frequency in the diffraction intensities, as they are proportional to the square of the diffraction amplitudes. Following the above arguments, a figure of merit for PSD demodulation can be introduced, which measures the degree of symmetry in time of the response. It is defined as

$$\mathrm{FOM_{FFT}} = \frac{\left|\left\{\sum_{i=2}^{N/2}\left|\mathrm{FT}[g(t)]\right|\right\}_{\mathrm{even}} - \left\{\sum_{i=1}^{N/2}\left|\mathrm{FT}[g(t)]\right|\right\}_{\mathrm{odd}}\right|}{\left\{\sum_{i=2}^{N/2}\left|\mathrm{FT}[g(t)]\right|\right\}_{\mathrm{even}} + \left\{\sum_{i=1}^{N/2}\left|\mathrm{FT}[g(t)]\right|\right\}_{\mathrm{odd}}}, \quad (13)$$

where N is the number of measurements forming the profile $g(t)$, the symbol FT represents the (discrete) Fourier transform, and the summations refer to even and odd Fourier frequencies as indicated by the subscripts. The higher the even–odd frequency asymmetry in the power spectrum of $g(t)$, the higher the value of $\mathrm{FOM_{FFT}}$, and the higher the efficiency of the PSD demodulation, since the hypothesis underlying this approach is better satisfied. Conversely, low values of $\mathrm{FOM_{FFT}}$ characterize asymmetric or nonperiodic $g(t)$ functions.

2.2. Decomposition by multivariate analysis

Equation (8) represents a linear decomposition of the background-subtracted data matrix $A(2\theta, t) - b(2\theta, t)$ into the matrix $[R(2\theta), S(2\theta), T(2\theta)]$, which is defined for each 2θ value, and the matrix $[g(t)^2, g(t), 1]$, describing the time dependence of the data and defined for each value of the stimulus. Such decomposition can be conveniently obtained by a variety of multivariate methods. One of these methods is principal component analysis (PCA), which is a projection algorithm used to reduce the dimensionality of multivariate data.

To better explain the multivariate approach, we will hereafter use a matrix notation, where the data matrix $A(2\theta, t) - b(2\theta, t)$ is written as the matrix $\mathbf{X}(m, n)$, of size $M \times N$, in which the columns run over the variables (2θ) and the rows over the diffraction profiles taken at different times (t). In PCA the data matrix is decomposed into a number of principal components (PCs) that maximize the explained variance in the data on each successive component, under the constraint of being orthogonal to the previous PCs:

$$\mathbf{X} = \mathbf{U}\mathbf{W}', \quad (14)$$

where the transformation is defined by a set of N-dimensional vectors of N loadings $\mathbf{W}(:,n)$ (this notation addresses the nth column vector of $\mathbf{W}$) that map each row vector of $\mathbf{X}$ to a new vector of principal component (or scores) $\mathbf{U}(:,n)$ ($\mathbf{U}$ has size $M \times N$). The loadings are calculated as the eigenvectors of the covariance matrix of the data, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$; the magnitude of the corresponding eigenvalues represents the variance of the data along the eigenvector directions (Wold *et al.*, 1987).

A relevant application of PCA is that, in many problems, the initial dimensionality of the data set, equal to the number of columns (N), can be reduced if the total variance can be approximated by the first (few) k components. The number k then represents the effective number of PCs used in the approximation:

$$\mathbf{X}(n, m) = \sum_{l=1}^{N}\mathbf{U}(n, l)\mathbf{W}(m, l) \cong \sum_{l=1}^{k}\mathbf{U}(n, l)\mathbf{W}(m, l). \quad (15)$$

The data matrix can be projected into a new coordinate system, given by the loadings. Plots based on this coordinate system are known as the score plots and can be used to reveal patterns in data. Score and loading vectors are then two alternative representations of the data matrix: the former characterize it in the sample space, the latter in the variable space.

When applied to the set of XPD patterns, PCA can be interpreted as sketched in Fig. 2. A powder diffraction profile (sample) can be seen as a data point of an N-dimensional space, where N is the number of 2θ values (variables), while the coordinates of the point in a reference system of this space are the values of intensity associated with each 2θ value. PCA can reduce the dimensionality of this representation, by using a reference system with only k orthogonal axes that represent the directions of maximum variability of the data. The coordinates of the data point in this new reference system are the scores, while the loadings are the coefficients which define the N directions with respect to the original reference system.

In the framework of the MED technique, PCA can supply an approximated decomposition, by projecting equation (8) in a space with $k = 2$. The approximation comes from the fact that, apart from their magnitude, the components in equation (8) might be correlated. The first principal component (PC1) is thus an approximation of the term $S(2\theta)g(t)$, because it is larger than the others, having contributions from all the atoms of the crystal system. The second principal component (PC2) is an approximation of the term $R(2\theta)g(t)^2$, while the third term is neglected, being constant in time (and represents the unexplained variance that is neglected in the approximation done using $k = 2$) (see Palin *et al.*, 2015). The matrix $[R(2\theta), S(2\theta)]$ is then approximated by the loadings of the first two principal components, *i.e.* $\mathbf{W}(:,1)$ and $\mathbf{W}(:,2)$, and the matrix $[g(t)^2, g(t)]$ by their corresponding scores, *i.e.* $\mathbf{U}(:,1)$ and $\mathbf{U}(:,2)$. As a matter of fact, $R(2\theta)$ is obtained as the loading associated with PC2, and $g(t)$ is obtained as the scores associated with PC1. Moreover, a condition should be verified between PC2 and PC1 scores: one should be proportional to the square of the other.

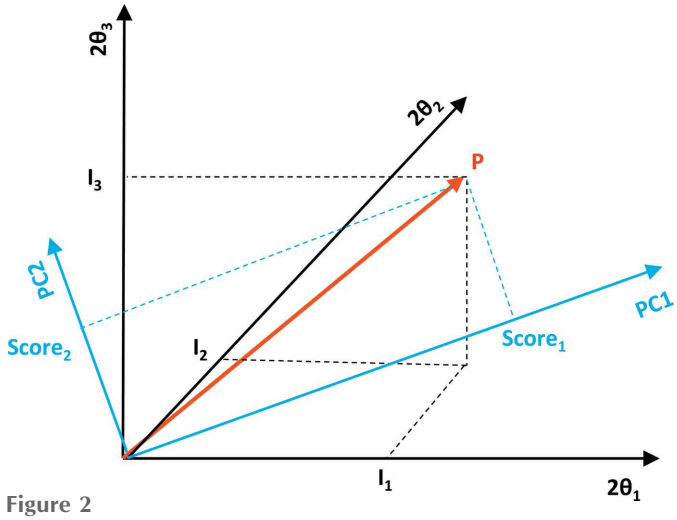


Figure 2
Schematic representation of a hypothetical powder diffraction profile (P) constituted by $N = 3$ intensity values (I_1, I_2, I_3) for respective 2θ values ($2\theta_1$, $2\theta_2$, $2\theta_3$). When projected in the space of the principal component directions PC1 and PC2, it can be described by only two values: $\mathrm{Score_1}$ and $\mathrm{Score_2}$.

It is worth noting that, if the hypothesis that the peak shape and position do not change with the stimulus does not hold, the PCA decomposition cannot be accomplished, because equation (8) is not valid and the dependence on 2θ and t variables cannot be separated.

3. PCA adapted to MED

As elaborated in §2, an optimal demodulation/decomposition of the MED signal would be characterized by the condition that $R(2\theta)$ is positive for each value of 2θ [see equation (7)]. For PCA, this condition applies to the loadings associated with PC2. In addition, the PC2 scores should be proportional to the square of the PC1 scores. In the previously presented geometrical view, these two conditions constrain the direction of PC2 to lie in the positive sector of the reference system in the N-dimensional space, and the arrangements of the data points in the PC2–PC1 plane, which should follow a parabolic trend. In the following, we will refer to these conditions through the following notation:

$$\mathbf{U}(:, 2) = \gamma\, \mathbf{U}(:, 1)^2 + \varepsilon, \tag{16}$$

$$\mathbf{W}(:, 2) > 0, \tag{17}$$

where for the loadings and scores we have used the notation of the previous section. We here remark that the proportionality between the second and the square of the first score is given up to a residual ε, which represents the limitations to further model the residual terms [as $T(2\theta, t)$ in equations (6) and (7)].

Previous reasoning suggests that application of PCA to MED data requires a number of constraints to be imposed [equations (16) and (17)]. Constrained PCA (CPCA) provides a common framework to introduce constraints on scores and loadings (Takane & Hunter, 2001). In the CPCA approach the data matrix is firstly modelled as the sum of different contributions, each one pertaining to a different kind of constraint (on samples or on variables), then decomposition in scores and loadings is achieved within each term $\mathbf{X} = \mathbf{GMH}' + \mathbf{BH}' + \mathbf{GC} + \mathbf{E}$. As explained by Takane & Hunter (2001), the first term in the CPCA model pertains to what can be explained by both **G** and **H**, the second term to what can be explained by **H** only, the third term to what can be explained by **G** only, and the last term is the residual of the model. Here **G** and **H** are the matrices including external information (the constraints) on the data samples and on the variables, respectively.

However, the conditions in equations (16) and (17) cannot be set as described above, since we *firstly* make a PCA decomposition, to find (without constraints) the different contributions in equation (8), and *then* we apply the constraints, *i.e.* we impose some conditions on the scores (the time profile) and loadings (the diffraction intensities) to satisfy the nature of the problem. The idea proposed here is then to modify the PCA by introducing such constraints, specific to the MED analysis. In PCA decomposition, it results that the scores are orthogonal to each other. Since this constraint is not required by the MED problem, and indeed its application could be detrimental, we allow the score axes to change their direction, by exploring the k-dimensional space (already reduced to the principal components) driven by a properly defined cost function. The idea is that we are able to detect the optimal rotated axes of a low-dimensional space (where data still have a meaningful representation) by minimizing an objective function, provided that the two conditions in equations (16) and (17) are satisfied. It is worth noting that, in this perspective, the axes will be no longer orthogonal.

To formalize the problem, we start the analysis by detailing the case $k = 2$. After a PCA, the (approximated) data matrix is in the form

$$\mathbf{X} \cong \mathbf{U}_{(k)}\mathbf{W}'_{(k)}, \tag{18}$$

where $\mathbf{U}_{(k)}$ and $\mathbf{W}_{(k)}$ are matrices of size $[M \times 2]$ and $[N \times 2]$, respectively. The subscript (k) refers to the first k columns of both the scores and the loadings matrices. Equivalently, equation (18) can be rewritten in the following way:

$$\mathbf{X} \cong \left(\mathbf{U}_{(k)}\mathbf{T}\right)\left(\mathbf{T}^{-1}\mathbf{W}'_{(k)}\right), \tag{19}$$

in which **T** is a $[k \times k]$ (in our case $[2 \times 2]$) matrix, very similar (but not equal) to a rotation matrix. If we put

$$\mathbf{T} = \begin{bmatrix} \cos\phi & \sin\psi \\ \sin\phi & \cos\psi \end{bmatrix}, \tag{20}$$

where ϕ and ψ are two independent parameters defining the change in direction of the axes, it is very simple to show that

$$\mathbf{T}^{-1} = \frac{1}{\cos(\phi + \psi)} \begin{bmatrix} \cos\psi & -\sin\psi \\ -\sin\phi & \cos\phi \end{bmatrix}. \tag{21}$$

The new scores are then

$$\hat{\mathbf{U}} = \left(\mathbf{U}_{(k)}\mathbf{T}\right) \Rightarrow \begin{cases} \hat{u}(m, 1) = u(m, 1)\cos\phi + u(m, 2)\sin\phi \\ \hat{u}(m, 2) = u(m, 1)\sin\psi + u(m, 2)\cos\psi \end{cases} \quad m = 1, \ldots, M \tag{22}$$

and the new loadings are

$$\hat{\mathbf{W}} = \left(\mathbf{T}^{-1}\mathbf{W}'_{(k)}\right) \Rightarrow \begin{cases} \hat{w}(n, 1) = \dfrac{1}{\cos(\phi + \psi)}[w(n, 1)\cos\psi - w(n, 2)\sin\psi] \\ \hat{w}(n, 2) = \dfrac{1}{\cos(\phi + \psi)}[-w(n, 1)\sin\phi + w(n, 2)\cos\varphi] \end{cases}$$

$$n = 1, \ldots, N. \tag{23}$$

The main properties of the matrix **T** are that (i) the variances associated with the first and second scores do not change in such a transformation (the columns of **T** have norm 1) and (ii) the changes in the directions of the two scores are independent ($\phi \neq \psi$).

Actually, two figures of merit (*i.e.* cost functions for the minimization problem) have been defined as internal criteria to assess the quality of the MED decomposition, which implement conditions (16) and (17), respectively:

(1) The Pearson correlation coefficient between the second (rotated) score and the square of the first (rotated) score ($\langle\cdot\rangle$ stands for average on the m index):

$$\mathrm{FOM}_{\mathrm{scores}} = \frac{\left|\sum_{n=1}^{M}\left[\hat{u}(m,1)^2 - \langle\hat{u}(m,1)^2\rangle\right]\left[\hat{u}(m,2) - \langle\hat{u}(m,2)\rangle\right]\right|}{\left\{\sum_{n=1}^{M}\left[\hat{u}(m,1)^2 - \langle\hat{u}(m,1)^2\rangle\right]^2\right\}^{1/2}\left\{\sum_{n=1}^{N}\left[\hat{u}(m,2) - \langle\hat{u}(m,2)\rangle\right]^2\right\}^{1/2}}. \quad (24)$$

This figure of merit requires that the mean square of the residual ε in equation (16) is minimum, regardless of the proportional term γ. The absolute values in the numerator account for the sign ambiguity of PCA scores.

(2) The normalized difference between the positive and the negative parts of the area underlying the second loading:

$$\mathrm{FOM}_{\mathrm{loadings}} = \frac{\sum_{n=1}^{N}\left[\hat{w}(n,2) > \sigma_{\hat{w}}\right] - \left|\sum_{n=1}^{N}\left[\hat{w}(n,2) < -\sigma_{\hat{w}}\right]\right|}{\sum_{n=1}^{N}\left[\hat{w}(n,2) > \sigma_{\hat{w}}\right] + \left|\sum_{n=1}^{N}\left[\hat{w}(n,2) < -\sigma_{\hat{w}}\right]\right|}, \quad (25)$$

where $\hat{w}(n,2)$ is the intensity of the second loading at the angle $2\theta_n$, rotated by using equation (23), and $\sigma_{\hat{w}}$ is the standard deviation of $\hat{w}(n,2)$. Thus $\mathrm{FOM}_{\mathrm{loadings}}$ measures the positive–negative asymmetry of the second (rotated) loading; its definition is dictated by the fact that the overall sign of the PCA loadings $\hat{w}(n,2)$ is arbitrary.

Both the figures of merit have 1 as the highest and best value. The idea is to find the optimal combination of (ϕ, ψ) verifying equation (24) with the constraint in equation (22). Thus the cost function to be maximized is $\mathrm{FOM}_{\mathrm{scores}}$, while $\mathrm{FOM}_{\mathrm{loadings}}$ serves as a control for the convergence of the search procedure.

The method can be generalized to $k \neq 2$. In this case the PCA decomposition can be rewritten as

$$\mathbf{X} \cong \left(\mathbf{U}_{(k)}\mathbf{T}\right)\left\{\left[\mathbf{T}'(\mathbf{T}\mathbf{T}')^{-1}\right]\mathbf{W}'_{(k)}\right\}, \quad (26)$$

where the matrix $\mathbf{T}$ has now size $[k \times 2]$ and $[(\mathbf{T}'(\mathbf{T}\mathbf{T}')^{-1}]$ is the Moore–Penrose generalized inverse of $\mathbf{T}$. The degree of freedom in $\mathbf{T}$ is now $2(k-1)$, since we have

$$\mathbf{T} = \begin{bmatrix} \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \\ \vdots & \vdots \\ \alpha_k & \beta_k \end{bmatrix} \quad (27)$$

together with the constraints

$$\begin{cases} \sum_{i=1}^{k}\alpha_i^2 - 1 = 0 \\ \sum_{i=1}^{k}\beta_i^2 - 1 = 0 \end{cases} \quad (28)$$

to preserve the variance associated with the scores. The rotated scores are, in this case,

$$\hat{\mathbf{U}} = \left(\mathbf{U}_{(k)}\mathbf{T}\right) \Rightarrow \begin{cases} \hat{u}(m,1) = \sum_{i=1}^{k} u(m,i)\,\alpha_i \\ \hat{u}(m,2) = \sum_{i=1}^{k} u(m,i)\,\beta_i \end{cases} \quad m = 1, \ldots, M. \quad (29)$$

The general solution is hence given by the determination of the unknowns in $\mathbf{T}$, to be found by the solution of the following constrained optimization problem:

$$\left\{\mathbf{T}(\alpha_i, \beta_i)\right\}_{\mathrm{opt}} = \arg\max_{\mathbf{T}}\left\{\mathrm{FOM}\left[\mathbf{U}, \mathbf{W}, \mathbf{T}(\alpha_i, \beta_i)\right]\right\},$$
$$\text{s.t.}: \begin{cases} \sum_{i=1}^{k}\alpha_i^2 - 1 = 0 \\ \sum_{i=1}^{k}\beta_i^2 - 1 = 0 \end{cases} \quad (30)$$

where the figure of merit can be either the one in equation (24) or that in equation (25). In the case of $k = 2$ the previous problem is just an optimization problem (without constraints), since the form of the matrix $\mathbf{T}$ in equation (27) automatically includes the normalization.

This algorithm, which we call optimum constrained components rotation (OCCR), provides the optimal coefficient for the rotation of the scores. The rotation of the loadings is given by equation (19) for the case where $k = 2$ or by equation (26).

It is worth noting that while CPCA first decomposes the original data matrix into several components (external analysis) and then applies PCA to each components separately, or to some of the components combined (internal analysis), the proposed OCCR technique applies PCA to the data set and imposes the specific constraints of equations (16) and (17) by rotating each component separately, relaxing the orthogonal conditions between the scores.

Furthermore, it may be appropriate to remark here that differences exist also between OCCR and multivariate curve resolution (MCR), the latter being a multivariate analysis algorithm proposed especially in the chemometric research field. MCR is a data-driven method, *i.e.* it decomposes the data set with limited or no information on the system, and it intends to recover the pure response profiles (they can be spectra, pH profiles, time profiles) of the chemical constituents or species of an unresolved mixture obtained in chemical processes (see http://www.mcrals.info/). MCR (Lawton & Sylvestre, 1971; Sylvestre *et al.*, 1974), starting from PCA, tries to refine the solution by determining a decomposition into two matrices (the 'concentration' profiles $\mathbf{C}$ and the 'spectra' profiles $\mathbf{S}$ of individual components, corresponding to the scores and loadings in our case) that are both non-negative. The two matrices are found by solving a constrained least-squares problem and starting from the reduced data matrix $\mathbf{X}_{\mathrm{PCA}}$ achieved after the application of an initial PCA to original data:

$$\mathbf{X} = \mathbf{CS}' + \mathbf{E},$$
$$\mathbf{C}, \mathbf{S}\ \text{s.t.}: \begin{cases} \min\limits_{\hat{\mathbf{C}}} \left\| \mathbf{X}_{\mathrm{PCA}} - \hat{\mathbf{C}}\hat{\mathbf{S}}' \right\| \\ \min\limits_{\hat{\mathbf{S}}} \left\| \mathbf{X}_{\mathrm{PCA}} - \hat{\mathbf{C}}\hat{\mathbf{S}}' \right\| \end{cases} \tag{31}$$

This problem is exactly the alternating least-squares (ALS) algorithm applied in the non-negative matrix factorization (NNMF) algorithm (Lee & Seung, 2001; Berry *et al.*, 2007; Voronov *et al.*, 2014). The algorithms OCCR and MCR have similarities: they both start from an initial PCA to reduce the dimensionality and they both try to decompose the data matrix into the product of two matrices. However, they have basic differences:

(*a*) In OCCR the solution is given without solving a least-squares problem, as happens in MCR, since equation (30) is solved using a general optimization method and the unknowns [*i.e.* the elements of matrix **T** in equations (20) or (27)] are not linearly related to the figure of merit. As a consequence, MCR looks for solutions that are approximations of the initial matrix [$\mathbf{X}_{\mathrm{PCA}}$ in equation (31)], while in OCCR the solution is the optimization of a figure of merit properly designed for the MED problem.

(*b*) OCCR does not impose the constraint that both the matrices be positive, as MCR does.

We did not consider applying MCR (which is basically a constrained NNMF algorithm) to the MED problem for the following two reasons:

(1) The NNMF solution is very sensitive to initial conditions [*i.e.* the starting values given to matrices **C** and **S** to iteratively solve equation (31)], as demonstrated by the extensive literature on such problems (see for example Langville *et al.*, 2006).

(2) In the MED case, we do not have to require both the matrices of the decomposition to be positive. In detail, if $k = 2$, we need only the second loading to be positive, while the first loading, as well as the first score, can have both positive and negative parts, owing to the term $\cos(\varphi_{\mathrm{A}} - \varphi_{\mathrm{SA}})$ [see equation (7)]. Instead, the condition in equation (16) can be applied in a variant of the NNMF. [Many variants have been developed for NNMF and resumed by MCR. See for example Li *et al.* (2007).]

We tried to modify the ALS algorithm by relaxing the non-negativity condition (applied just for the second loading) and trying to impose the constraint in equation (16), but the results were not as successful as those achieved with OCCR.

4. Applications

To test the above concepts, we have chosen $CuFe_2O_4$, known as an inverse spinel, in its cubic form. The copper ions sit predominantly on octahedral cation sites and the iron atoms split between octahedral and tetrahedral ones (Fig. 3). It has been shown to be a catalyst for the water–gas shift (WGS) process, in which carbon monoxide reacts with water to produce carbon dioxide and molecular hydrogen. Several crystal phases are expected to play a role in the WGS reaction, and the configuration and properties of its active sites are still a matter of debate (Papavasiliou, 2004; Men *et al.*, 2004). To reproduce the working condition, $CuFe_2O_4$ was exposed *in situ* to a mixture of CO/H_2O, and time-resolved X-ray powder diffraction measurements were performed.

Previous temperature programmed reduction in CO studies showed that metallic Cu^0 and CuO crystal phases are formed in addition to the cubic $CuFe_2O_4$, with a mole fraction variability in the ranges [0, 0.5] and [0, 0.1], respectively. Moreover, the occupancy of the octahedral site in $CuFe_2O_4$ is expected to change with occupancy variations between 0.8 and 0.9 in a complex way (Estrella *et al.*, 2009). The calculated XPD profiles from the sole Cu atom and the whole $CuFe_2O_4$ crystal phase are compared in Fig. 3.

The published variation of the octahedral site occupancy suggested that the system could be a good test of the MED technique. Model calculations showed that a 20% variation in the occupancy of the octahedral site could be extracted from demodulation of cyclic model data (Dooryhee *et al.*, 2014; Tutuncu *et al.*, 2015). These model data did not include any of the other phases that are present in the experimental powder pattern and did not model the cell dimension changes in $CuFe_2O_4$ that occur as the Cu is removed. However, the PSD demodulation of experimental data to determine the change of octahedral Cu^{2+} did not yield the expected trend. This could be attributed to the peak overlap of $CuFe_2O_4$ with the Cu^0 and CuO powder patterns, the peak shifts, and a nonlinear response of XPD data to a square wave cyclic pulse. In addition, the experimental octahedral Cu^{2+} occupancies may have been overestimated in the published refinement, because there is a high correlation between the weight fraction and the occupancy.

The above problems make this system a challenging application of the multivariate analysis here proposed. To tackle the complex real data we preliminarily applied the algorithms to simulated data, where the expected variations were

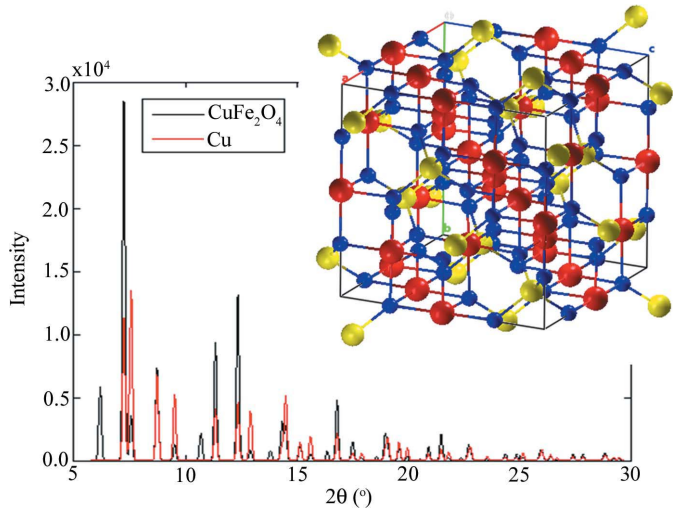


Figure 3
Crystal structure of $CuFe_2O_4$, with Cu atoms in red, Fe atoms in yellow and O atoms in blue, and calculated X-ray diffraction patterns (λ = 0.3196 Å) of the entire structure (black) and of the Cu atoms only (red).

implemented with an increasing level of complexity. The successful application of these techniques to real data is also presented.

4.1. Experimental data collection

X-ray data were collected at the beamline X7B of the National Synchrotron Light Source (NSLS) at Brookhaven National Laboratory with an X-ray energy of 38 keV (0.3196 Å) and 0.5 × 0.5 mm beam size. A large two-dimensional Perkin Elmer area detector (2048 × 2048 pixels and 200 × 200 µm pixel size) was mounted orthogonal to the beam path, 400 mm downstream from the sample. Raw two-dimensional data were azimuthally integrated and converted to one-dimensional intensity *versus* 2θ by using the *FIT2D* program (Hammersley *et al.*, 1996). Lanthanum hexaboride was measured as standard material to calibrate the sample and detector geometry, including the sample-to-detector distance. Several periods of measurements were performed, with CO flowed in the first half-period, and O_2 in the second one, thus reproducing a square-wave stimulus. The temperature was kept constant at 508 K.

4.1.1. Data generation and analysis. Simulated data representing occupancy variations of the Cu atom in $CuFe_2O_4$ were generated by using the program *GSAS* (Larson & Von Dreele, 1995; Toby, 2001). Occupancy values beyond realistic limits were used, for extensive tests of our algorithms. Experimental and simulated XPD profiles were processed by the programs *2DMED* (Tutuncu *et al.*, 2015) to implement the PSD demodulation and *RootProf* (Caliandro & Belviso, 2014) to implement the PCA decomposition, and by a MATLAB (The MathWorks Inc., Natick, MA, USA) script to implement the OCCR method. A Rietveld analysis (Rietveld, 1969) was performed on real data by using the program *QUANTO* (Altomare *et al.*, 2001); crystal structures have been displayed by the viewer *JAV* (Burla *et al.*, 2012).

4.2. Results on simulated data

Our first goal was to check the ability of the multivariate decomposition to detect and quantify occupancy variations for different kinds of system response and compare its performance with that of the PSD demodulation. Different patterns of occupancy variations for the Cu atom within the $CuFe_2O_4$ crystal phase were implemented, starting from regular ones, resembling triangular, sinusoidal or square waves, and then increasing their complexity, by introducing fast, slow or asymmetric decay in the square waves. These represent the profile of the function $g(t)$ introduced in §2. Notably, the scores of the first principal component obtained by OCCR are perfectly superimposed on the occupancy values used in the simulations (Fig. 4). Since the scores are in arbitrary units, in Fig. 4 they have been rescaled as follows:

$$\text{scores}' = \langle \text{occ} \rangle + \sigma_{\text{occ}} \frac{\text{scores} - \langle \text{scores} \rangle}{\sigma_{\text{scores}}}, \qquad (32)$$

where $\langle \text{occ} \rangle$ and σ_{occ} are, respectively, the mean value and the standard deviation of the occupancy values put in the simulations, and $\langle \text{scores} \rangle$ and σ_{scores} are the same quantities for the OCCR scores values. The scaling proposed in equation (32) is simply one of the possible methods (remove the score mean, scale to the occupancy dynamics, add the occupancy mean) to obtain a satisfying visual comparison between the score of the first principal component obtained by OCCR and the actual occupancy values. It works so nicely because the two profiles have very similar shape (the degree of similarity, without any scaling, can be inferred by the value of the Pearson correlation coefficient as in Table 1).

Regarding the loadings, the case of sinusoidal variation of occupancy (SI) is shown as an example in Fig. 5. §§2 and 3 show that the loadings of the second principal component represent the XRD diffraction pattern related to the active part of the crystal system, *i.e.* $R(2\theta)$ in the notation of equation (7), and in fact those obtained after OCCR (Fig. 5, full line) coincide with the calculated diffraction pattern from the Cu atom (Fig. 3, red line). Conversely, the loadings from PCA have negative peaks which hinder their use in phasing procedures aiming at finding the positions of the active atoms.

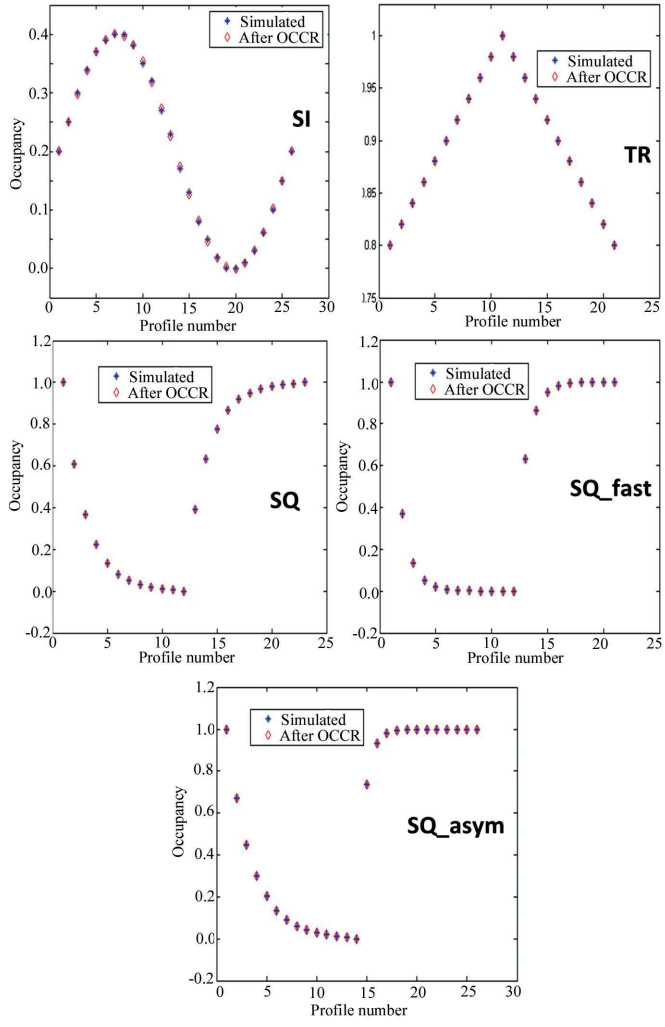


Figure 4
Values of the occupancy of the Cu atom set in the simulations (blue) and scores of the first principal component (red) obtained after the OCCR procedure and rescaled according to equation (32).

Table 1
Results of the MED analysis on simulations where the occupancy of the Cu atom is varied according to various functions (system response).

PCA, OCCR and PSD are the values of the Pearson correlation coefficient between the calculated XRD profile of the Cu atom and those obtained, respectively, by PCA or OCCR decomposition, or by PSD demodulation. FOM_{FFT} is the figure of merit for PSD demodulation, as defined in equation (13). The intervals spanned by the occupancy values are reported in square brackets.

System response	Acronym	PCA	OCCR	PSD	FOM_{FFT}
Sinusoidal [0 0.4]	SI	0.940	1.000	1.000	1.00
Triangular [0.8 1]	TR	0.478	1.000	1.000	0.94
Square, slow decay [0 1]	SQ	0.695	1.000	1.000	0.79
Square, fast decay [0 1]	SQ_fast	0.704	1.000	0.860	0.70
Square, asymmetric decay [0 1]	SQ_asym	0.684	1.000	0.316	0.64
Sinusoidal [0.8 1]	–	0.521	1.000	1.000	1.00
Ramp [0.8 1]	–	0.609	1.000	0.788	0.24
Ramp [0 1]	–	0.919	1.000	0.215	0.15

The overall results obtained for simulated data are summarized in Table 1, where the Pearson correlation coefficients between the calculated XRD profiles of the Cu atom and the profiles obtained by PCA and OCCR demodulation (corresponding to their PC2 loadings), and those obtained by PSD demodulation, are compared. Other simulated system responses have been added with respect to those reported in Fig. 4, to explore other system responses (ramp) and/or different ranges spanned by occupancy values. It can be noted that PCA can only approximate the true $R(2\theta)$ profiles, as the correlation coefficient can be much lower than 1. As an example, the value of 0.478 is obtained for the case of TR, shown in Fig. 5, where most of the low-angle peaks of the true profile are completely missed by the PCA decomposition. It is worth noting that the PCA results are related in an unpredictable way to the system response. For example, the correlation coefficient is 0.940 if the occupancy is varied as a sinusoid between 0.0 and 0.4, or 0.521 if it is varied in the same way between 0.8 and 1.0. The way PCA behaves is basically related to the criterion that its components be orthogonal (*i.e.* uncorrelated), while the time-dependent terms present in the MED signal are strongly correlated [see equation (8)].

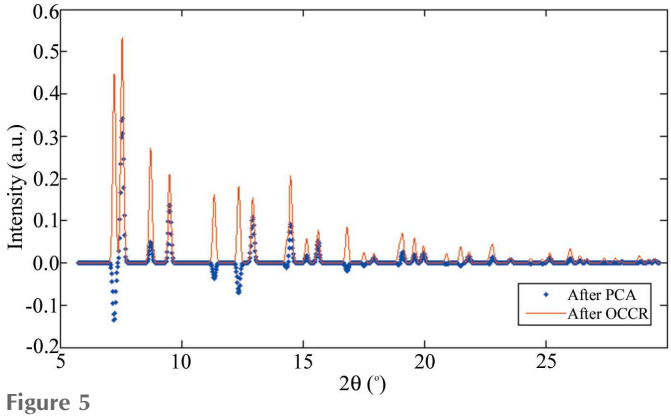


Figure 5
X-ray diffraction patterns obtained after PCA (blue dots) and OCCR (red line) decomposition applied to the simulation with triangular variation of the Cu occupancy (TR). The calculated profile from the Cu atom coincides with that after OCCR decomposition.

Moreover, most of the data variance is accounted for by PC1, while PC2 is only a small fraction (see Table S1 in the supporting information), so the direction of the PC2 eigenvector is not well defined. OCCR overcomes both problems by using the MED relations (16) and (17) to explore the space of the solutions with non-orthogonal components. In fact OCCR is able to perfectly reproduce the true $R(2\theta)$ profiles in all the considered cases, *i.e.* it is able to properly change the intensities of the PCA profiles to match the XRD profile of the active atom, as shown in Fig. 5. Notably, the performance is independent of the system response and of the spanned range of values. The PSD demodulation produces the same results as OCCR for periodic responses at definite symmetry in time. If the response is made asymmetrical (SQ_asym) or nonperiodic (ramp), the correlation coefficient of PSD drops significantly (nonperiodic signals are demodulated by assuming that they continue periodically). This behaviour is expected from the PSD theory (see §2.1), and in fact the trend of the correlation coefficient for PSD reproduces that of the FOM_{FFT} values. Specifically, a perfect PSD demodulation occurs for $FOM_{FFT} > 0.80$, while a dramatic decrease of PSD efficiency occurs for $FOM_{FFT} < 0.70$, where OCCR gives much better results. Thus OCCR is the only option to treat MED data when $FOM_{FFT} < 0.70$.

A comprehensive analysis of the $R(2\theta)$ profiles assessed by the different methods is given in Fig. 6, where they have been classified by PCA. Each point in the scatter plot of the PC2 *versus* PC1 scores represents a decomposed or demodulated

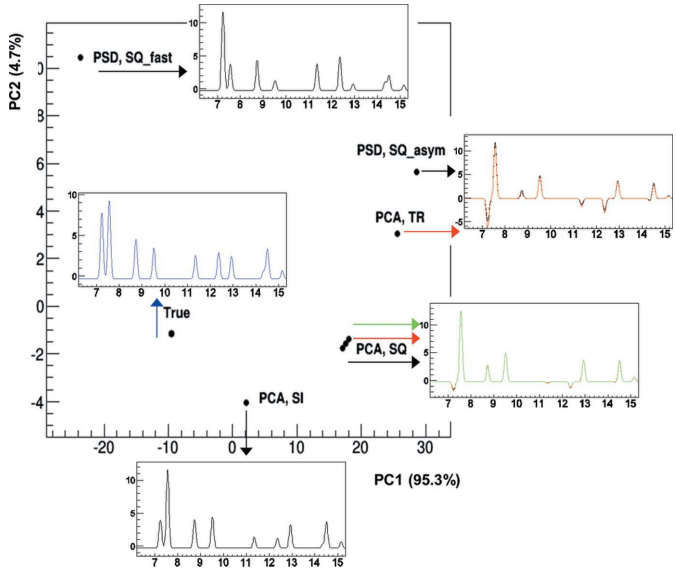


Figure 6
Results of the MED analysis on simulated profiles. The calculated profile from the Cu atom (True) is compared with profiles demodulated by PSD and decomposed by PCA, through a PCA analysis. They are all represented as points in the scatter plot of the second (PC2) *versus* the first (PC1) principal component. The data variance explained by the PC is reported in parentheses. X-ray powder diffraction profiles corresponding to each point or cluster of points are shown, zoomed in the 2θ range [6.0°, 15.5°] and by using lines of different colour. All the profiles obtained by OCCR and those obtained by PSD for simulations SI, TR and SQ coincide with the True profile.

Table 2
Results of the MED analysis on simulations where the occupancy of the Cu atom is varied according to various functions (system response).

$CORR_{scores}$ is the Pearson correlation coefficient between the occupancy values used in the simulation and the scores of the first principal component; FOM_{scores} and $FOM_{loadings}$ are the values of the figures of merit defined by equations (23) and (24), respectively. All the values refer to the PCA decomposition. Corresponding values after OCCR are all equal to 1.0.

System response	$CORR_{scores}$	FOM_{scores}	$FOM_{loadings}$
Sinusoidal [0 0.4]	0.999	0.993	1.000
Triangular [0.8 1]	1.000	0.994	0.540
Square, slow decay [0 1]	0.994	0.956	0.876
Square, fast decay [0 1]	0.997	0.860	0.891
Square, asymmetric decay [0 1]	0.996	0.945	0.858
Sinusoidal [0.8 1]	0.998	0.986	0.602
Ramp [0.8 1]	1.000	0.996	0.437
Ramp [0 1]	0.996	0.950	1.000

profile, and its separation from the calculated profile (True) signifies the differences between the profiles. To appreciate the distance between points, one should take into account the amount of data variance explained by PC1 and PC2, reported in Fig. 6: separations along the PC1 axis are much more relevant than those along the PC2 axis. The low-angle part of the profiles associated with single points or groups of points is also shown in Fig. 6. All the profiles obtained by OCCR have their representative points perfectly superimposed on the 'True' point, and they all coincide with the profile shown in the corresponding profile panel. The same holds for PSD demodulated profiles calculated for simulations with symmetric response (SI, TR, SQ), while those obtained from simulations with asymmetric response (SQ_asym and SQ_fast) are very far from the 'True' point, and the corresponding panels show peaks with negative (SQ_asym) or anomalous (SQ_fast) intensities. PCA decomposition represents an intermediate result: none of its points overlap with the 'True' one, but they are closer to it than the PSD demodulated profiles from simulations with asymmetric responses. In particular, TR has a profile that is close to that obtained by PSD applied on SQ_asym; SQ, SQ_fast and SQ_asym have similar profiles, with very small negative peaks, and SI has a profile close to the True one, with no negative peaks.

Further insights into the performance of the multivariate analysis can be gained by inspection of Table 2, where $CORR_{scores}$ is the Pearson correlation coefficient between the implemented $g(t)$ values and the PC1 scores, and FOM_{scores} and $FOM_{loadings}$ are defined in §3, equations (24) and (25), respectively. The values in Table 2 refer to PCA decomposition, while the corresponding values after OCCR are all 1.0. Thus the first evidence is that OCCR is able to optimize both the loadings and the scores output of the PCA, by supplying precise estimates of the $R(2\theta)$ profile and the $g(t)$ trend. The second evidence is that the PCA can estimate the scores of the first two principal components much better than the corresponding loadings. In fact, both $CORR_{scores}$, expressing the agreement of the PC1 scores with the $g(t)$ trend, and FOM_{scores}, assessing the expected relation between PC1 and PC2 scores, have values higher than 0.860 for all simulated cases. $FOM_{loadings}$ can instead be as low as 0.437, as the presence of negative peaks in the PC2 loadings deteriorate the agreement with the expected XPD profile (see Fig. 5).

An example of the errors that can be made in estimating $g(t)$ is given in Fig. 7(a), where the scores obtained by PCA and OCCR for the simulation ramp [0 1] are reported: the right trend is obtained only after OCCR, while PCA gives a good approximation, with $CORR_{scores}$ = 0.996. It is worth noting that for PCA a better estimate for $g(t)$ is obtained if a subset of measurements is taken. In fact, ramp [0.8 1] has $CORR_{scores}$ = 1; however this occurs at the expense of the estimate of $R(2\theta)$, as $FOM_{loadings}$ drops from 1.00 to 0.44 (see Table 2). The correlation between PC1 and PC2 scores for the same simulations is shown in Fig. 7(b): a clear parabolic shape indicates that the MED condition given by equation (17) is satisfied. It is only approximated after PCA decomposition and is exactly followed after OCCR decomposition. In fact, from the geometrical point of view, equation (16) corresponds to a parabola having the axis parallel to the Y axis and passing through x = 0. It is worth remarking that the case shown in

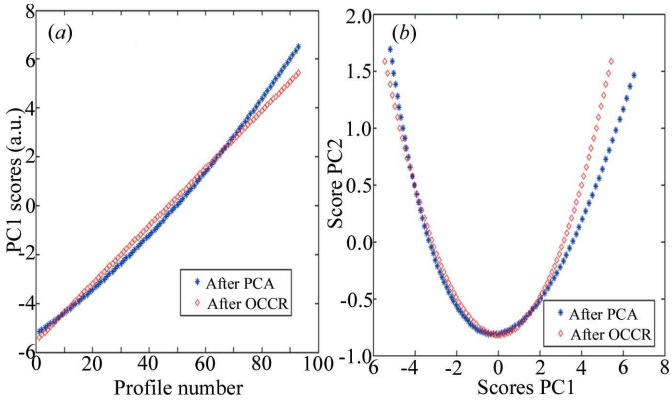


Figure 7
Scores of the first principal component (PC1) (a) and scatter plot of the second (PC2) *versus* the first (PC1) principal component scores (b) obtained after PCA and OCCR procedures for a simulation where the occupancy of the Cu atoms varies linearly between 0 and 1. Such an asymmetric response would not be treatable by PSD.

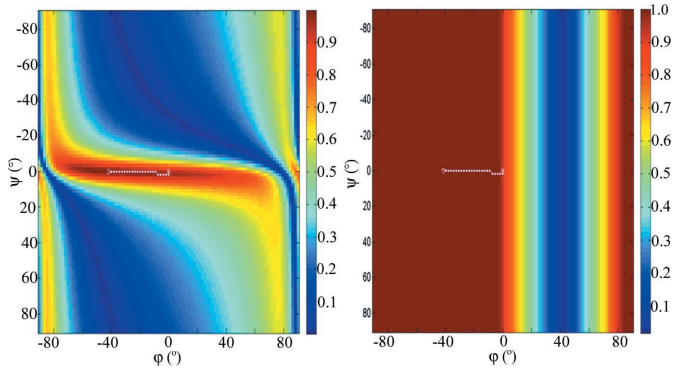


Figure 8
Contour plots of FOM_{scores} (left) and $FOM_{loadings}$ (right) as a function of the rotation parameters ϕ and ψ for the simulation SQ. The paths visited during the OCCR procedure are highlighted by white dots. The starting angles are (0, 0).

Fig. 7 could not be accomplished by PSD, owing to its nonperiodic response.

The landscapes of FOM_{scores} and $FOM_{loadings}$ in the space of the rotation parameters ϕ and ψ, and the actual path followed during the OCCR minimization procedure, are shown in Fig. 8 for the SQ simulation. Similar plots for the other simulations are given as supporting information. Interestingly, it results that the FOM_{scores} landscape has a stronger dependence on the ψ parameter, while the $FOM_{loadings}$ landscape only depends on the ϕ parameter. This can be explained by the fact that FOM_{scores} is mainly influenced by changes in the orientation of the smaller PC2 axis with respect to the larger PC1 axis. Therefore its dependence on $\hat{U}(:, 2)$, which in turn depends on the ψ rotation angle [equation (22)], is stronger than that on $\hat{U}(:, 1)$, while $FOM_{loadings}$ depends only on $\hat{W}(:, 2)$ [equation (25)], which in turn depends only on the parameter ϕ [equation (23)].

The dependence of the multivariate decomposition on sampling has been studied for the simulation ramp [0 1], by removing 2θ values uniformly and gradually and comparing the resulting PC1 scores with the $g(t)$ profiles (Fig. 9*a*). Similarly, measurements have been sampled and the resulting PC2 loadings have been compared with the $R(2\theta)$ profiles (Fig. 9*b*). The results indicate that OCCR maintains its performance even on decreasing the 2θ sampling by a factor of ten, or reducing the number of measurements to the minimum allowed for the decomposition, *i.e.* three. Also, a worse result is obtained for PCA, which is affected by sampling when the 2θ sampling is decreased by a factor of four or when using less than eight measurements.

4.3. Results on experimental data

When the OCCR procedure was applied to real data measurements, a trend of the PC1 scores (Fig. 10*a*) was obtained, which is very similar to that obtained with the simulated data SQ_asym (Fig. 10*b*). Thus the response of the system for real data resembles that of a square wave with asymmetric decay, like that implemented in SQ_asym. The FT analysis of the OCCR scores suggests a more pronounced asymmetry for real data, as FOM_{FFT} is 0.38, compared to the value of 0.64 for SQ_asym.

The obtained time dependence of the system response can be readily interpreted on the basis of the experimental conditions and on a (static) Rietveld analysis carried out on each measured profile, separately. Data can be clearly divided into two half-periods:

(*a*) During measurements from 1 to 91 the fluxed CO produces a decrease in the amount of the $CuFe_2O_4$ phase, which is always the dominant phase, and a corresponding increase in the amount of the Cu^0 phase. These trends are smooth and highly correlated. The CuO phase is only present in the first 4–5 measurements.

(*b*) At measurement 92 a drastic change of experimental conditions occurs, since CO flow is closed and O_2 starts being fluxed. As a consequence, the Cu^0 phase undergoes a sharp decrease and a CuO phase an abrupt increase.

(*c*) During measurements from 93 to 180 the amounts of the $CuFe_2O_4$ and CuO phases remain nearly constant, while the Cu^0 phase is present in negligible amount.

However, it should be noted that the Rietveld analysis cannot precisely disentangle the changes in the relative abundances of different crystal phases from the variation of the occupancy of the Cu atom occupying the octahedral site of $CuFe_2O_4$. This is due to the strong overlap between the peaks of the dominant $CuFe_2O_4$ phase and those of the minority Cu^0 and CuO phases, the high correlation between refinement parameters (the octahedral occupancy correlates by −0.82 with the scale factor of the $CuFe_2O_4$ phase, and by 0.67 with the thermal factor associated with the octahedral site), and the concurrent presence of Cu^{2+} and Fe^{2+} ions in the octahedral site. As an example, the results of the Rietveld analysis of two measurements in the first and second half-periods are shown

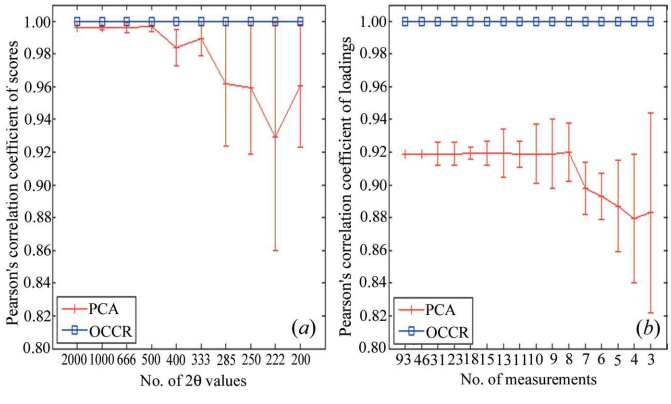


Figure 9
Sensitivity to sample size. (*a*) Person's correlation coefficient between the loadings of the second principal component and the calculated Cu pattern as a function of the number of uniformly sampled measurements; (*b*) Person's correlation coefficient between the scores of the first principal component and the occupancy values used in simulations as a function of the number of uniformly sampled 2θ values. Error bars account for different ways to sample the corresponding number of 2θ values or number of measurements.

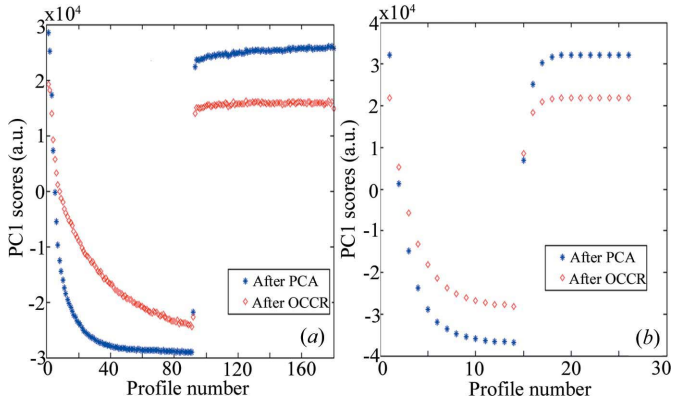


Figure 10
Scores of the first principal component (PC1) obtained after PCA and OCCR procedures applied to real data (*a*) and simulated data with squared asymmetric response (*b*).

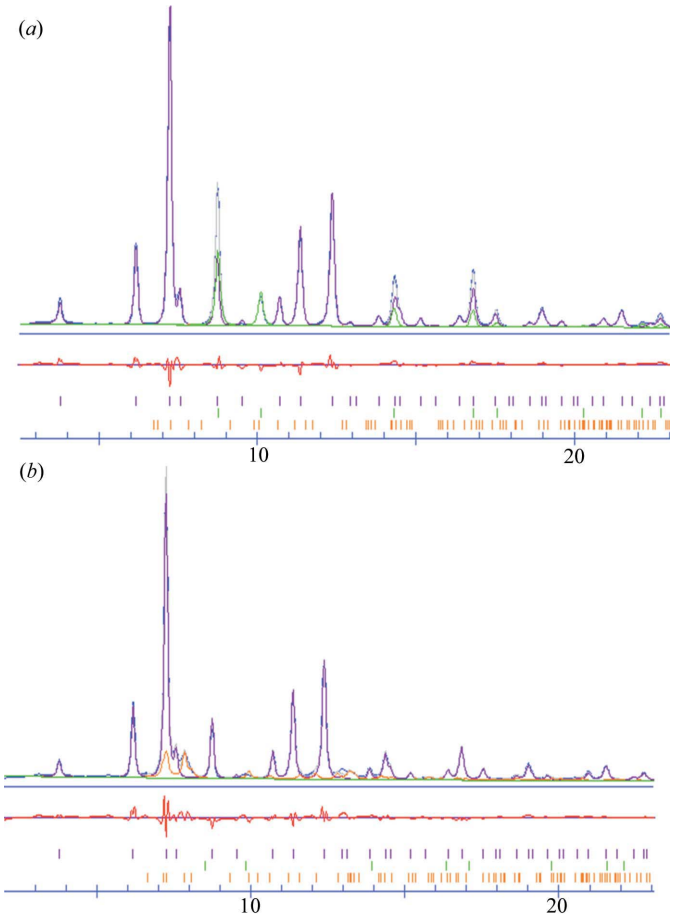


Figure 11
Results of the Rietveld analysis applied on measurements No. 5 (*a*) and No. 160 (*b*). Observed (grey), calculated (blue) and difference (red) profiles are shown. Profiles and reflection positions of $CuFe_2O_4$ (violet), Cu^0 (green) and CuO (yellow) phases are also shown.

in Fig. 11, where contributions of the different phases are highlighted.

The presence of structural changes in the $CuFe_2O_4$ phase is confirmed by the (dynamic) MED analysis, considering the correlation between PC2 and PC1 scores. The parabolic trend followed by real data (Fig. 12*a*) is similar to the trend followed by simulated SQ_asym data (Fig. 12*b*), suggesting that the MED relation of equation (16) is valid also for real data, although approximated. The corresponding values of FOM_{scores} are 0.698 after PCA and 0.982 after OCCR, to be compared with 0.945 and 1.000, respectively, for SQ_asym (see Table 2).

Beside the encouraging results on scores, those on loadings were not as good. The OCCR procedure was found to decrease the value of $FOM_{loadings}$, which goes from 0.449 after PCA to 0.370, contrary to the SQ_asym simulation, where it increases from 0.858 to 1.000 (see Table 2). Similar results are obtained if subsets of measurement are considered. We envisaged three reasons for such a failure:

(1) The tailored multivariate analysis presented in §3 is only valid if a MED signal is considered, which can be originated by occupancy, scattering factor or thermal factor variations of a subset of atoms, according to equations (3) and (4). If,

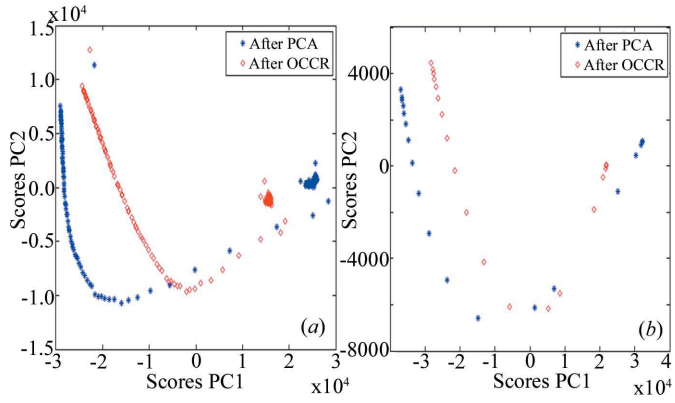


Figure 12
Scatter plot of the second (PC2) *versus* the first (PC1) principal component scores obtained after PCA and OCCR procedures applied to real data (*a*) and simulated data with squared asymmetric response (*b*).

however, the relative abundance of diverse crystal phases changes with time, besides structural variations of one of the phases, the decomposition of equation (8) is no longer valid, and PCA and/or OCCR will produce unpredictable results.

(2) The time dependence of variations of phase abundances are highly correlated with those of the occupancy variations. For this reason, they cannot be empirically assigned to different principal components by PCA, but more likely they will be mixed in the same component, definitely hindering the MED decomposition of its scores.

(3) The XPD peaks influenced by occupancy variations of the octahedral site significantly overlap with those related to the $CuFe_2O_4$ phase, making it difficult to differentiate the principal components through their loadings.

For such reasons, the new OCCR approach was able to extract the time dependence of the overall (structural and quantitative) variations in experimental data but, similarly to the PSD approach, not to extract the signal related to changes in the octahedral site occupancy.

5. Conclusions

In this paper, we took a step forward in the application of the MED technique. In its original formulation, it could only be applied to crystal systems that produce a linear response to an external stimulus. More specifically, the response had to cover a full period and be perfectly symmetric in the time domain. This strict requirement, which in practice narrows the scope of MED, does not come from the technique itself but from the PSD algorithm used to demodulate the MED signal. Driven by the need to extend the applicability of MED to real-world data, we faced the problem by replacing the PSD demodulation by a more flexible approach, based on multivariate analysis. As a first step, the standard formulation of PCA has been applied for the purpose: MED data have been decomposed into principal components, which are expected to contain the same information as the harmonic components coming from PSD. Unfortunately, this decomposition is only approximate, and we realized that PCA can produce better

results than PSD only for systems with nonlinear response, while its results are worse than PSD if linear-response systems are involved. A way to prevent this deadlock was to modify the PCA approach, by adapting it to MED. We then found relations between scores and loadings of the principal components which must be met in the case of perfect decomposition of MED data, put these in mathematical form by using appropriate figures of merit, and developed an algorithm, OCCR, to modify the principal components in order to fulfil the MED requirements. We have thus recovered the gap between PCA and PSD, so that the multivariate analysis is now able to produce perfect decomposition of MED data for systems with both linear and nonlinear responses. Another advantage of the OCCR approach is that its loadings represent the diffraction pattern due to the active sublattice, thus allowing chemical selectivity in X-ray diffraction, and, in addition, its scores represent the system response in the time domain, thus allowing a dynamic characterization of the crystal system. At the same time, however, real-world data can be more complex than expected. In particular, in our experiment the structural effect related to a nonlinear response to the external stimulus (gas flowing) is added to variations in abundances of different crystal phases within a period of application of the stimulus. The formation and destruction of other crystal phases in addition to that experiencing structural variations is an effect that was not foreseen in the original MED theory, and additional effort is needed to account for such phenomena while treating MED data. As an additional drawback, we realized that structural and quantitative changes were highly correlated, in both the time and reciprocal-space (peak overlapping) domains. When we applied our OCCR approach to such real data, we were only able to characterize the time dependence of the system response through the OCCR scores, while the information from its loadings was lost. The further extension of the multivariate analysis approach to MED data to treat such cases is being developed. Finally, the proposed OCCR approach gives an elegant solution of the inverse problem of MED (Chernyshov *et al.* 2011); this could be developed to a new tool to probe the kinetic response of complex systems, as an alternative to Fourier analysis.

Acknowledgements

Use of the National Synchrotron Light Source, Brookhaven National Laboratory, was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under contract No. DE-AC02-98CH10886. This research has been partially supported by the short-term mobility CNR program. Caterina Chiarella is acknowledged for technical support.

References

Altomare, A., Burla, M. C., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G., Polidori, G. & Rizzi, R. (2001). *J. Appl. Cryst.* **34**, 392–397.
Beek, W. van, Emerich, H., Urakawa, A., Palin, L., Milanesio, M., Caliandro, R., Viterbo, D. & Chernyshov, D. (2012). *J. Appl. Cryst.* **45**, 738–747.
Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. & Plemmons, R. J. (2007). *Comput. Stat. Data Anal.* **52**, 155–173.
Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Mallamo, M., Mazzone, A., Polidori, G. & Spagna, R. (2012). *J. Appl. Cryst.* **45**, 357–361.
Caliandro, R. & Belviso, D. B. (2014). *J. Appl. Cryst.* **47**, 1087–1096.
Caliandro, R., Chernyshov, D., Emerich, H., Milanesio, M., Palin, L., Urakawa, A., van Beek, W. & Viterbo, D. (2012). *J. Appl. Cryst.* **45**, 458–470.
Caliandro, R., Chernyshov, D., Emerich, H., Milanesio, M., Palin, L., Urakawa, A., van Beek, W. & Viterbo, D. (2013). *Neutron and Synchrotron Sources: Role in Crystallography – Small Angle Scattering, Supramolecular Assemblies, Emerging Characterization Facilities and Tools, and Chemical Crystallography*, Transactions of the Symposium Held at the 2013 American Crystallographic Association Annual Meeting, Honolulu, HI, 20–24 July 2013. http://www.amercrystalassn.org/2013-transactions_toc.
Chernyshov, D., van Beek, W., Emerich, H., Milanesio, M., Urakawa, A., Viterbo, D., Palin, L. & Caliandro, R. (2011). *Acta Cryst.* A**67**, 327–335.
Dooryhee, E., Yakovenko, A., Hanson, J., Ghose, S., Rodriguez, J. & Senanayake, S. (2014). *Acta Cryst.* A**70**, C1175.
Estrella, M., Barrio, L., Zhou, G., Wang, X., Wang, Q., Wen, W., Hanson, J. C., Frenkel, A. I. & Rodriguez, J. A. (2009). *J. Phys. Chem. C*, **113**, 14411–14417.
Ferri, D., Newton, M. A., Di Michiel, M., Chiarello, G. L., Yoon, S., Lu, Y. & Andrieux, J. (2014). *Angew. Chem. Int. Ed.* **53**, 8890–8894.
Ferri, D., Newton, M. A., Di Michiel, M., Yoon, S., Chiarello, G. L., Marchionni, V., Matam, S. K. M. H. A., Aguirre, M. H., Weidenkaff, A., Wen, F. & Gieshoff, J. (2013). *Phys. Chem. Chem. Phys.* **15**, 8629–8639.
Hammersley, A. P., Svensson, S. O., Hanfland, M., Fitch, A. N. & Hausermann, D. (1996). *High. Pressure Res.* **14**, 235–248.
Langville, A. N., Meyer, C. D. & Albright, R. (2006). *Initializations for the Nonnegative Matrix Factorization*. Philadelphia: KDD.
Larson, A. C. & Von Dreele, R. B. (1995). *GSAS* (*General Structural Analysis System*). Report LAUR 86-748. Los Alamos National Laboratory, New Mexico, USA.
Lawton, W. H. & Sylvestre, E. A. (1971). *Technometrics*, **13**, 617–633.
Lee, D. D. & Seung, H. S. (2001). *Adv. Neural Inf. Process. Syst.* **13**, 556–562.
Li, H., Adal, T., Wang, W., Emge, D., Cichocki, A. & Cichocki, A. (2007). *J. VLSI Sign. Process.* **48**, 83–97.
Lu, Y., Keav, S., Marchionni, V., Chiarello, G. L., Pappacena, A., Di Michiel, M., Newton, M. A., Weidenkaff, A. & Ferri, D. (2014). *Catal. Sci. Technol.* **4**, 2919–2931.
Men, Y., Gnaser, H., Zapf, R., Hessel, V., Ziegler, C. & Kolb, G. (2004). *Appl. Catal. Gen.* **277**, 83–90.
Milanesio, M., Palin, L., Viterbo, D., Caliandro, R., Urakawa, A., van Beek, W. & Chernyshov, D. (2014). *Acta Cryst.* A**70**, C1471.
Palin, L., Caliandro, R., Viterbo, D. & Milanesio, M. (2015). *Phys. Chem. Chem. Phys.* **17**, 17480–17493.
Papavasiliou, J. (2004). *Catal. Commun.* **5**, 231–235.
Rietveld, H. M. (1969). *J. Appl. Cryst.* **2**, 65–71.
Sylvestre, E. A., Lawton, W. H. & Maggio, M. S. (1974). *Technometrics*, **16**, 353–368.
Takane, Y. & Hunter, M. A. (2001). *Appl. Algebra Eng. Commun. Comput.* **12**, 391–419.
Toby, B. H. (2001). *J. Appl. Cryst.* **34**, 210–213.
Tutuncu, G., Yakovenko, A., Hanson, J., Ghose, S. & Dooryhee, E. (2015). *J. Appl. Cryst.* In preparation.
Urakawa, A., Bürgi, T. & Baiker, A. (2006). *J. Chem. Phys.* **324**, 653–658.
Voronov, A., Urakawa, A., van Beek, W., Tsakoumis, N. E., Emerich, H. & Rønning, M. (2014). *Anal. Chim. Acta*, **840**, 20–27.
Wold, S., Esbensen, K. & Geladi, P. (1987). *Chemom. Intell. Lab. Syst.* **2**, 37–52.