Bioinformatics, 31(5), 2015, 753–760
doi: 10.1093/bioinformatics/btu734
Advance Access Publication Date: 5 November 2014
Original Paper

OXFORD

Bioimage informatics

An efficient TOF-SIMS image analysis with spatial correlation and alternating non–negativity-constrained least squares

Parham Aram[1,*], Lingli Shen[1], John A. Pugh[2], Seetharaman Vaidyanathan[2] and Visakan Kadirkamanathan[1]

[1]Department of Automatic Control and Systems Engineering and [2]Department of Chemical and Biological Engineering, University of Sheffield, Sheffield, UK

*To whom correspondence should be addressed.

Associate Editor: Robert F. Murphy

Received on July 20, 2014; revised on October 23, 2014; accepted on November 1, 2014

Abstract

Motivation: Advances in analytical instrumentation towards acquiring high-resolution images of mass spectrometry constantly demand efficient approaches for data analysis. This is particularly true of time-of-flight secondary ion mass spectrometry imaging where recent advances enable acquisition of high-resolution data in multiple dimensions. In many applications, the distribution of different species from a sampled surface is spatially continuous in nature and a model that incorporates the spatial correlation across the surface would be preferable to estimations at discrete spatial locations. A key challenge here is the capability to analyse the high-resolution multidimensional data to extract relevant information reliably and efficiently.

Results: We propose a framework based on alternating non–negativity-constrained least squares which accounts for the spatial correlation across the sample surface. The proposed method also decouples the computational complexity of the estimation procedure from the image resolution, which significantly reduces the processing time. We evaluate the performance of the algorithm with biochemical image datasets generated from mixture of metabolites.

Contact: p.aram@sheffield.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Imaging mass spectrometry (IMS) is an increasingly popular approach in the characterization of biological samples (Alexandrov and Bartels, 2013; Alexandrov and Kobarg, 2011). It is a powerful tool to image surface structure with a mass spectrum measured at each pixel. One IMS technique is time-of-flight secondary ion mass spectrometry (TOF-SIMS) with applications in areas as varied as hair care (Kempson and Skinner, 2005), medical implants (Viornery *et al.*, 2002) and drug delivery systems (Belu *et al.*, 2000). It is primarily applied in monitoring the distribution of targeted chemical species, but can also be employed for imaging surfaces in non-targeted analyses, when surface compositions are not well defined, in particular in the analysis of biological systems (Fletcher *et al.*, 2007; Vaidyanathan *et al.*, 2008). However, this latter extension poses several challenges with respect to data deconvolution. Multivariate analysis (MVA) methods have been used extensively to map large complex spatio-spectral TOF-SIMS data into factors of single components providing insights into the type and distribution of chemical species on a surface. Popular MVA techniques to describe spatio-spectral TOF-SIMS data include principal component analysis (PCA) (Jackson, 2005), maximum auto-correlation factor (MAF) (Larsen, 2002; Tyler, 2006) and multivariate curve resolution (MCR) (De Juan and Tauler, 2006).

The PCA is perhaps the most commonly used technique, either as a main analysis step (Wagner and Castner, 2001) or an initialization step

(Graham and Castner, 2012; Lee *et al.*, 2009) for MA approach. MAF is an alternative to PCA which is independent of pre-treatment scaling. The method is based on maximizing the autocorrelation between neighbouring pixels. A comparison of the two methods for TOF-SIMS data analysis is reported in Henderson *et al.* (2009), recommending the use of MAF when images with low spectral intensity are analysed.

The presence of negative peaks in the spectrum computed using PCA or MAF makes it more difficult to interpret the resulting factorization. MCR optimized by alternating least squares (MCR-ALS) (Esteban *et al.*, 2000; Garrido *et al.*, 2008) overcomes this problem by imposing a non-negativity constraint in an iterative optimization process (Kim and Park, 2007; Lawson and Hanson, 1974; Wang *et al.*, 2003). However, these methods can be computationally expensive when applied to large three-way datasets. The fast combinatorial non–negativity-constrained least squares (FC-NNLS) algorithm (Van Benthem and Keenan, 2004) is specifically designed to handle such datasets, allowing for significant gains in the speed and computational demands of the optimization process in ALS applications.

Despite a substantial performance increase in FC-NNLS, the estimation of loadings and scores when decomposing large TOF-SIMS data in an MCR-ALS algorithm can still be problematic. One limitation is that the complexity of the scores estimation is coupled with the number of image pixels. Therefore, an increase in the image resolution of the sample surface leads to increases in the uncertainty and computational demands of the score and loading estimates. This is becoming increasingly important as sophisticated TOF-SIMS images with higher spatial resolution are becoming more widespread (Hanrieder *et al.*, 2013; Kubicek *et al.*, 2014). Even for a currently typical TOF-SIMS dataset with 128×128 pixels and a spectral mass range up to 1000 Da, the implementation of ALS algorithm with three spectral basis functions contains the estimation of $128 \times 128 \times 3$ values in the scores estimation step. This emphasizes the benefit of decoupling the number of pixels in images and the number of parameters. In addition, the progress towards acquisition of image data in voxels as opposed to pixels confers further demand on the analysis time and efficiency of data processing (Fletcher and Vickerman, 2012; Fletcher *et al.*, 2007).

Additionally, in cases where the spatial distribution of compounds in an image is continuous such as chemical samples or samples from a biological cell or a cellular environment, the characteristics at proximate/distant regions on the surface are expected to be highly/weakly correlated. This requires the inclusion of the spatial correlation across the surface in the estimation of scores and loadings from TOF-SIMS data.

We propose an approach to MCR-ALS that incorporates a continuous-over-space formulation which accounts for spatial correlation across the sample surface. Additionally, the algorithm does not couple the complexity of the estimation procedure to the resolution of TOF-SIMS images, resulting in a significant reduction in the processing time. We exploit the method of a basis function decomposition of scores images which simplifies the estimation of individual pixel values into a set of weights. These weights scale the continuous basis functions and lie in a significantly lower dimensional space. We set out to examine the algorithm with biochemically relevant model image datasets that were generated from simple mixture of metabolites, as an example.

2 Methods

2.1 Dataset description

Three metabolites, tyrosine, phenylalanine, and citric acid (all from Sigma Aldrich, UK) were used in the study. The metabolites were spotted on hexamethyldisilazane (HMDS) (Sigma Aldrich) coated silicon wafers (Compart Technology, UK), prepared as detailed elsewhere (Salim *et al.*, 2012). A focused spot of the sample (individual metabolites or metabolite mixture) was obtained, each spot containing about 125 pmole of the metabolite.

TOF-SIMS negative ion spectra and images were obtained using a SIMS V instrument (ION-TOF Inc., Germany). Fifty kiloelectron volt Bi_3^{++} was used as the primary ion source for the high current, bunched mode spectral acquisition, with a target current of 0.11 pA and 500 μm^2 field of view. The images collected contained 128×128 pixels. The vacuum in the analytical chamber was held at 10^{-9} mbar. Primary ion dose was kept below the static limit of 10^{13} primary ions/cm^2 to maintain the static SIMS status. Each spectrum was calibrated using hydrocarbon fragment peaks.

Spectral data up to m/z 200 was considered for analysis. Image data were exported in ASCII format to MATLAB where it was analysed. Only the intensities at 100 m/z data points were considered for the image analysis of all samples. This included the m/z corresponding to CN^- ($m/z = 26$) and the deprotonated metabolite ions $[M - H]^-$ for the three metabolites (as known signals).

Dataset contains measurements for three metabolites, tyrosine (T), phenylalanine (P) and citric acid (C), spotted separately as individual 'pure' species and as mixtures, one containing equimolar proportions of tyrosine and citric acid (TC) and another containing all three metabolites in equimolar proportions (TPC).

2.2 MCR model

The spatio-spectral TOF-SIMS data matrix can be described by the bilinear MCR model

$$\mathbf{Y}(f, \mathbf{s}) = \mathbf{W}(\mathbf{s})\mathbf{B}^\top(f) + \mathbf{E}(f, \mathbf{s}), \quad (1)$$

where f denotes the mass-to-charge ratio and $\mathbf{s}$ is the spatial location in the two-dimensional physical surface. The superscript $\top$ denotes the transpose operator. Each TOF-SIMS image with the dimension of l pixels by l' pixels is reshaped to form a column of the $p \times v$ data matrix, $\mathbf{Y}(\cdot)$, where $p = l \times l'$. The TOF-SIMS measurement is factored into a $p \times m$ scores matrix, $\mathbf{W}(\cdot)$, and a $v \times m$ loadings matrix, $\mathbf{B}(\cdot)$, comprising of m spectral basis vectors. The $p \times v$ matrix of residuals, $\mathbf{E}(\cdot)$, is added to account for unmodelled terms and experimental errors. This way spatio-spectral TOF-SIMS data are decomposed into spatial (scores) and spectral (loadings) data matrices which respectively provide information about the distribution and the identity of the species. Each element of $\mathbf{Y}$ at a given spatial location can be written as a sum of weighted loadings at a particular peak, where the weights are scores at that spatial location, that is,

$$\mathbf{Y}(f, \mathbf{s}) = \sum_{i=1}^{m} \mathbf{w}_i(\mathbf{s})\mathbf{b}_i(f) + \mathbf{E}(f, \mathbf{s}), \quad (2)$$

where $\mathbf{w}_i(\cdot)$ and $\mathbf{b}_i(\cdot)$ denote the ith weight and spectral basis vectors, respectively.

2.3 Estimation algorithm

A solution to the MCR model given in Equation (1) can be obtained by minimizing the following cost function

$$J(\mathbf{W}, \mathbf{B}) = \|\mathbf{Y} - \mathbf{W}\mathbf{B}^\top\|_F^2, \quad (3)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This optimization problem can be solved using the ALS algorithm (Paatero and Tapper, 1994). When initialized, at each iteration the estimated scores are used to update the loadings. The resulting estimates of the loadings are then

used when estimating new scores for the next iteration. The procedure stops when the convergence is achieved. The ALS algorithm is often combined with a non–negativity-constrained least squares to provide chemically meaningful solutions. In this case, the cost function (3) becomes

$$J(\mathbf{W},\mathbf{B}) = \|\mathbf{Y} - \mathbf{W}\mathbf{B}^\top\|_F^2 \text{ s.t. } \mathbf{W},\mathbf{B} > 0 \tag{4}$$

where $\mathbf{W},\mathbf{B} > 0$ indicates that all the elements of $\mathbf{W}$ and $\mathbf{B}$ are non-negative. Here, we use the FC-NNLS algorithm with ALS (Van Benthem and Keenan, 2004) which efficiently solves the non–negativity-constrained least squares problem.

The knowledge of the system's rank or the number of spectral basis vectors, m, is required prior to the ALS algorithm. Here, we use PCA and the scree test criterion as a guide to determine the number of spectral basis vectors. There are more sophisticated techniques that incorporate smoothing methods or subspace comparisons, in order to reduce the effect of measurement noise in the system's rank identification (Jiang *et al.*, 2004).

The stopping rule adopted is based on observing the Frobenius norms of the successive estimates of $\mathbf{W}$ matrices in Equation (4), that is,

$$||\mathbf{W}||_F^{(k)} - ||\mathbf{W}||_F^{(k-1)} < \rho, \tag{5}$$

where ρ is a threshold value.

2.4 Model decomposition

To facilitate the estimation procedure, we use a decomposition using a finite set of continuous basis functions to represent scores images. The proposed model accounts for the spatial correlation across the sample surface. This is preferable to estimations at discrete spatial locations where the distribution of species in the surface is continuous in nature. The decomposition is described by

$$\mathbf{w}_i(\mathbf{s}_p) \approx \sum_{j=1}^{n} \alpha_{ji}\phi_j(\mathbf{s}_p) \tag{6}$$

where $\phi(\cdot)$ are known basis functions scaled by unknown weights, α_{ji}, and n is the number of basis functions used to perform the decomposition. The basis functions used here are two-dimensional Gaussian functions given by

$$\phi(\mathbf{s}) = \exp\left(-\frac{(\mathbf{s}-\boldsymbol{\mu}_\phi)^\top(\mathbf{s}-\boldsymbol{\mu}_\phi)}{\sigma_\phi^2}\right), \tag{7}$$

where σ_ϕ and $\boldsymbol{\mu}_\phi$ are the basis function width and centre, respectively. These are standard basis functions often used to approximate continuous fields from sampled observations (Park and Sandberg, 1991). They satisfy all the required conditions for reconstruction functions (Scerri *et al.*, 2009) and have advantageous characteristics: (i) semi-compact support; (ii) dimensionally factorizable, allowing efficient computation in higher dimension and (iii) a closed form Fourier transform which is also Gaussian (Sanner and Slotine, 1992). The latter facilitates the determination of the width of the basis functions using spatial frequency analysis described in the following section.

Substituting Equation (6) into Equation (2), we obtain a continuous approximation which can be evaluated at $f = f_v$ and $\mathbf{s} = \mathbf{s}_p$ as

$$\mathbf{Y}(f_v,\mathbf{s}_p) = \sum_{i=1}^{m}\left[\sum_{j=1}^{n}\alpha_{ji}\phi_j(\mathbf{s}_p)\right]\mathbf{b}_i(f_v) + \mathbf{E}(f_v,\mathbf{s}_p). \tag{8}$$

This formulation accounts for spatial correlation across the sample surface via sum of weighted Gaussian basis functions. This can be written in a compact form as

$$\mathbf{Y} = \boldsymbol{\Phi}\mathbf{A}\mathbf{B}^\top + \mathbf{E}, \tag{9}$$

where $\mathbf{A}$ is an unknown $n \times m$ weight matrix and $\boldsymbol{\Phi}$ is a constant $p \times n$ matrix given by

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1(\mathbf{s}_1) & \phi_2(\mathbf{s}_1) & \phi_3(\mathbf{s}_1) & \dots & \phi_n(\mathbf{s}_1) \\ \phi_1(\mathbf{s}_2) & \phi_2(\mathbf{s}_2) & \phi_3(\mathbf{s}_2) & \dots & \phi_n(\mathbf{s}_2) \\ \phi_1(\mathbf{s}_3) & \phi_2(\mathbf{s}_3) & \phi_3(\mathbf{s}_3) & \dots & \phi_n(\mathbf{s}_3) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \phi_1(\mathbf{s}_p) & \phi_2(\mathbf{s}_p) & \phi_3(\mathbf{s}_p) & \dots & \phi_n(\mathbf{s}_p) \end{bmatrix}_{p\times n}. \tag{10}$$

This representation allows for a more efficient implementation of the ALS algorithm. The complexity of the scores estimation step in the optimization problem given in Equation (4) is directly linked to the resolution of TOF-SIMS images, limiting the applicability of the ALS algorithm to big datasets. However, in Equation (9), the scores matrix is decomposed into two parts: an unknown weight matrix $\mathbf{A}_{n\times m}$ and a constant matrix $\boldsymbol{\Phi}_{p\times n}$ depending only on basis functions. Therefore, the estimation problem of the scores matrix is simplified to the estimation of a matrix of weights with a significantly lower dimension. Similarly, the decomposition also simplifies the estimation of loadings matrix. This leads to decreases in uncertainty in the score and loading estimates and also improvement in the convergence property of the algorithm. An example of the decomposition of a score image using an equally spaced grid of 9×9 basis functions is shown in Figure 1.

Using the basis function representation, the cost function for the estimation of loadings, $\mathbf{B}$, becomes

$$J(\mathbf{A},\mathbf{B}) = \|\tilde{\mathbf{Y}} - \mathbf{A}\mathbf{B}^\top\|_F^2 \quad \text{s.t.} \quad \mathbf{B} > 0 \tag{11}$$

where

$$\tilde{\mathbf{Y}} = \boldsymbol{\Phi}^\dagger\mathbf{Y}, \tag{12}$$

$$\boldsymbol{\Phi}^\dagger = \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top \tag{13}$$

and A is fixed. The cost function for estimating the weight matrix, $\mathbf{A}$, is given by

$$J(\mathbf{A},\mathbf{B}) = \|\tilde{\mathbf{Y}}^\top - \mathbf{B}\mathbf{A}^\top\|_F^2 \quad \text{s.t.} \quad \mathbf{A} > 0 \tag{14}$$

where $\mathbf{B}$ is fixed. The matrix, $\tilde{\mathbf{Y}}$, can be calculated once and stored prior to the commencement of the algorithm. Here, the Frobenius

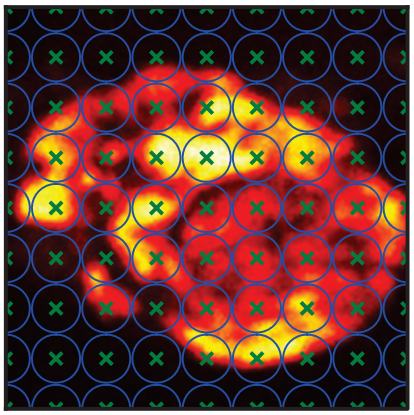

Fig. 1. Example of a basis decomposition. A 128×128 pixels TOF-SIMS image decomposed by a 9×9 grid of basis functions (shown by blue circles). The centre of each basis function is shown by a green cross (Color version of this figure is available at *Bioinformatics* online.)

norms of the successive estimates of **A** can be monitored to stop the algorithm.

It should be noted that basis decomposition technique results into smoother estimates of scores images which might blur sharp boundaries and details. However, if high resolution scores images are required, an extra step can be added to the estimation algorithm. The final estimate of **B**, obtained from the iterative optimization of Equations (11) and (14), can be used in a single run of the FC-NNLS algorithm to minimize

$$J(\mathbf{W},\mathbf{B}) = \|\mathbf{Y} - \mathbf{W}\mathbf{B}^{\top}\|_F^2 \quad \text{s.t.} \quad \mathbf{W} > 0, \tag{15}$$

and **B** is fixed. This way a detailed estimate of the scores images is obtained, while the loadings are still calculated from the reduced continuous representation of scores images. The complete estimation framework is given in Algorithm 1. The minimization of Equations (11), (14) and (15) can be implemented using the MATLAB m-file fcnnls provided by Van Benthem and Keenan (2004). Note, FC-NNLS uses the solution of the overwriting method (Gallagher *et al.*, 2004) to initialize the algorithm.

2.5 Spatial frequency analysis

The spatial frequency response of the ion image can be used to specify the width and spacing of the basis functions. The spatial cutoff frequency of the image, ν_c, can be found by calculating its power spectral density. This can be then used to determine the distance between basis functions such that Shannon's sampling theorem is satisfied

$$\Delta_\phi \leq \frac{1}{2\rho\boldsymbol{\nu}_c}, \tag{16}$$

where $\rho \in \mathbb{R}\geq 1$ is an oversampling parameter (Sanner and Slotine, 1992). The width of the basis functions is also governed by the spatial cutoff frequency of scores images. For an attenuation of 3 dB at $\boldsymbol{\nu}_c$, the width of Gaussian's basis functions should be set to (Freestone *et al.*, 2011)

$$\sigma_\phi = \frac{1}{\pi\boldsymbol{\nu}_c}\sqrt{\frac{\ln 2}{2}}. \tag{17}$$

The number of basis functions can then be specified by dividing the spatial region of interest into Δ_ϕ intervals. From reciprocal role of $\boldsymbol{\nu}_c$ in Equations (16) and (17), it follows that a high cutoff frequency results into a representation which comprises a large number of basis functions with narrow widths. This results into a computationally more complex estimation procedure as a higher number of weights is required to be estimated. Therefore, a compromise should be made between the accuracy and the computational demands of the estimation algorithm. The number of basis functions can be reduced by limiting the spatial bandwidth of the approximated images to a lower value.

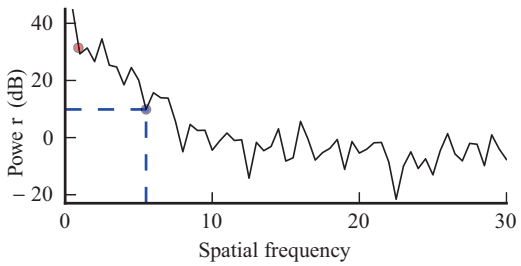

Fig. 2. Spatial frequency analysis. Cross-section of the spatial frequency response along the *x*-axis for TPC mixture

3 Results

The proposed algorithm was first implemented on pure species to extract discriminatory information from the estimated scores and loadings. The extracted information was then used to perform peak assignment in the computed spectra of TC and TPC mixtures. During this process, the spectral information was summarized into three major peaks that can differentiate different species. This information was also used to examine replicate measurements of each dataset to further evaluate the performance of the algorithm. The total ion images for each dataset are depicted in Supplementary Figure S1. TOF-SIMS images were mapped onto $[-1\,1]$ in both x and y directions and therefore the spatial aspects of the model can be considered arbitrary. Cross-section of the spatial frequency response along the x-axis for TPC mixture is shown in Figure 2. From this figure, it can be seen that to capture the full spatial bandwidth, ν_c should be set to a high value (≈ 5.5, shown by blue dashed line),

Algorithm 1 Scores and loadings estimation

1. Decomposition:
-determine m using PCA,
-define basis function centres μ using Equation (16),
-define basis function widths σ_ϕ using Equation (17),
-construct $\tilde{\mathbf{Y}}$ using Equations (10), (12) and (13).
2. Initialization:
-initialize the weight matrix, $\mathbf{A}_0$, as a random dense matrix.
3. scores and loadings estimation:
-define stopping condition threshold ρ,
-set $k=1$,
while $||\mathbf{A}_k - \mathbf{A}_{k-1}||_F > \rho$
-update the loadings, $\mathbf{B}_{k-1}$, using FC-NNLS and Equation (11),
-update the weight matrix, $\mathbf{A}_k$, using FC-NNLS and Equation (14),
-set $k = k + 1$,
end while
-update the loadings, $\mathbf{B}_{k-1}$, using FC-NNLS and Equation (11).
4. Estimation of high resolution scores matrices:
-calculate **W** from Equation (15) and the final estimate of **B**.

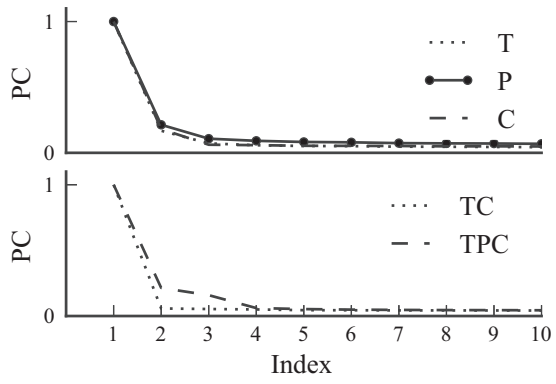

Fig. 3. PCA analysis. The normalized scree plots for the first 10 components are shown

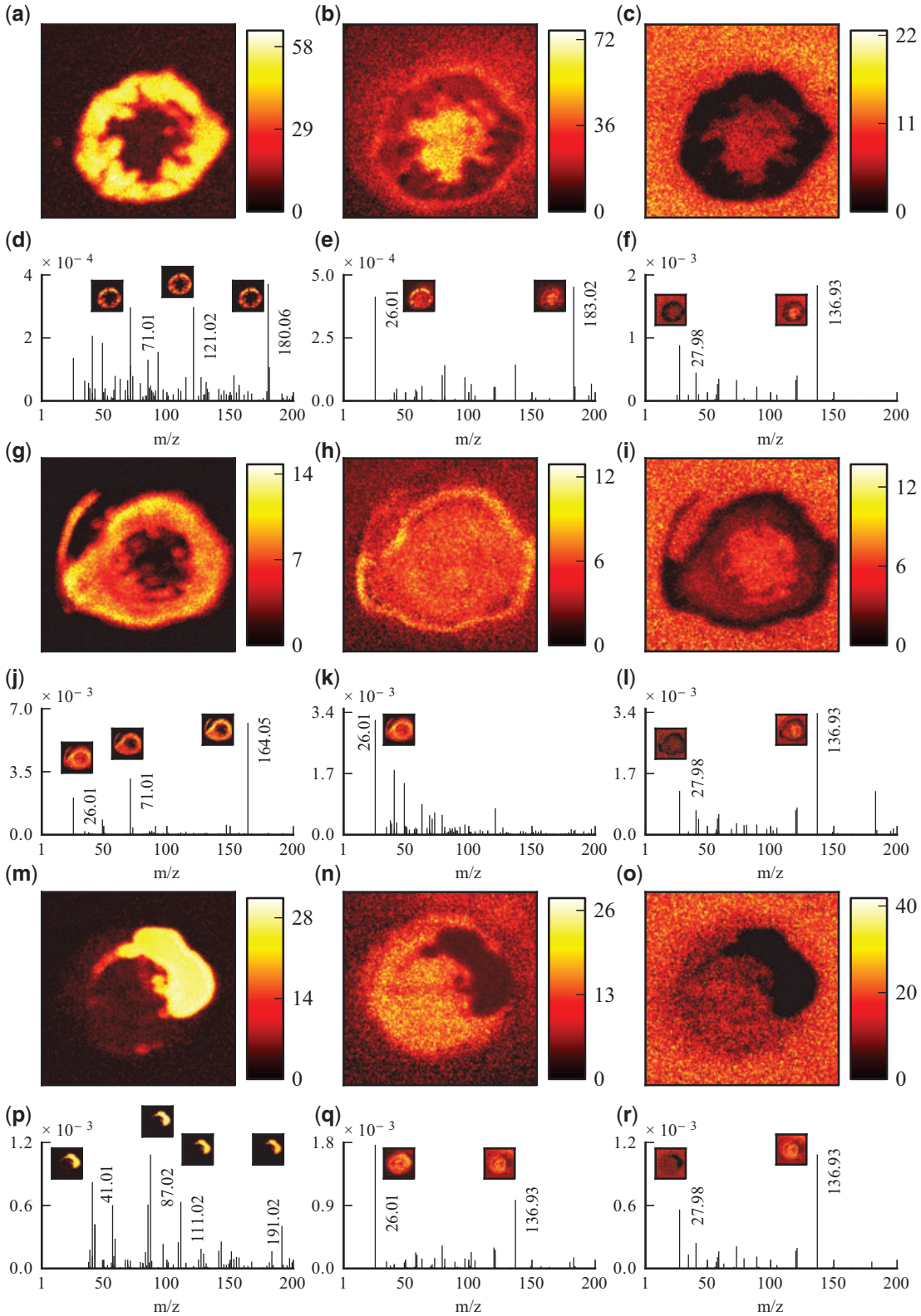

Fig. 4. Score and loading estimates using ALS with non-negativity constraint for individual components. (**a–f**) Scores and loadings for component *T*; (**g–l**) scores and loadings for component *P*; (**m–r**) scores and loadings for component *C*

resulting into a high number of basis functions (a grid of 44×44 with $\rho = 2$). Alternatively, the configuration can be chosen to represent the reconstructed images up to a specified bandwidth to limit the number of basis functions and hence computational demand of the estimation algorithm. We set the cutoff frequency to $\nu_c = 0.85$, giving approximately 10 dB attenuation from the maximum power spectrum. This is shown by a red circle in Figure 2. In order to account for the slow roll-off in the frequency response of Gaussian

basis functions, the oversampling parameter of $\rho = 2$ was chosen. Using these values in Equations (16) and (17) yielded the distance between adjacent basis functions' centres $\Delta_\phi = 0.3$ and the width of basis functions $\sigma_\phi = 0.22$. Given the spatial domain of interest, a grid of 9×9 equally spaced basis functions can be used to satisfy Shannon's sampling criterion. With this arrangement, the number of unknown parameters was reduced from $m \times 16\,384$ to $m \times 81$.

The pre-processing stage involved normalizing the data to the total ion counts. This was followed by Poisson scaling (Keenan and Kotula, 2004) for the system's rank analysis.

An initial guess of the number of spectral basis vectors, m, for each Poisson-scaled dataset was obtained using PCA and the scree test. The scree plot for the first few principal components is shown in Figure 3. The results suggested three spectral basis vectors for each of T, P and C elements and two and three spectral basis vectors for TC and TPC mixtures, respectively.

The proposed algorithm was then applied on TOF-SIMS datasets to estimate loadings and the corresponding scores images. In the estimation procedure, we set $m = 3$ for all pure and mixed species despite the suggested rank of two for TC compound in the PCA analysis. The results for scores and loadings estimation for T, P and C components are shown in Figure 4. The m/z number for dominant peaks are illustrated in the loadings plots where the inset figures show the corresponding ion images in the TOF-SIMS dataset. Figure 4d shows large peaks (sorted in order of decreasing magnitudes) at $m/z = 180.06$, 121.02 and 71.01 for the first factor of component T. For component P, the first factor comprises of large peaks at $m/z = 164.05$, 71.01 and 26.01 (Fig. 4j). For component C, significant peaks for the first factor are located at $m/z = 87.02$, 41.01, 111.02 and 191.02 (Fig. 4p). As can be seen, the $[M - H]^-$ signals for each of the metabolite predominates, but fragment ion contributions also exist.

Peaks at $m/z = 26.01$, $m/z = 27.98$ and $m/z = 136.93$ are common between second and the third factors of all components. Therefore, these m/z values do not contain discriminatory information and can be considered noise in the system. In fact, the corresponding scores images of the second and third factors show noisy structures. Also, the peak at $m/z = 71.01$ presents in the first factors

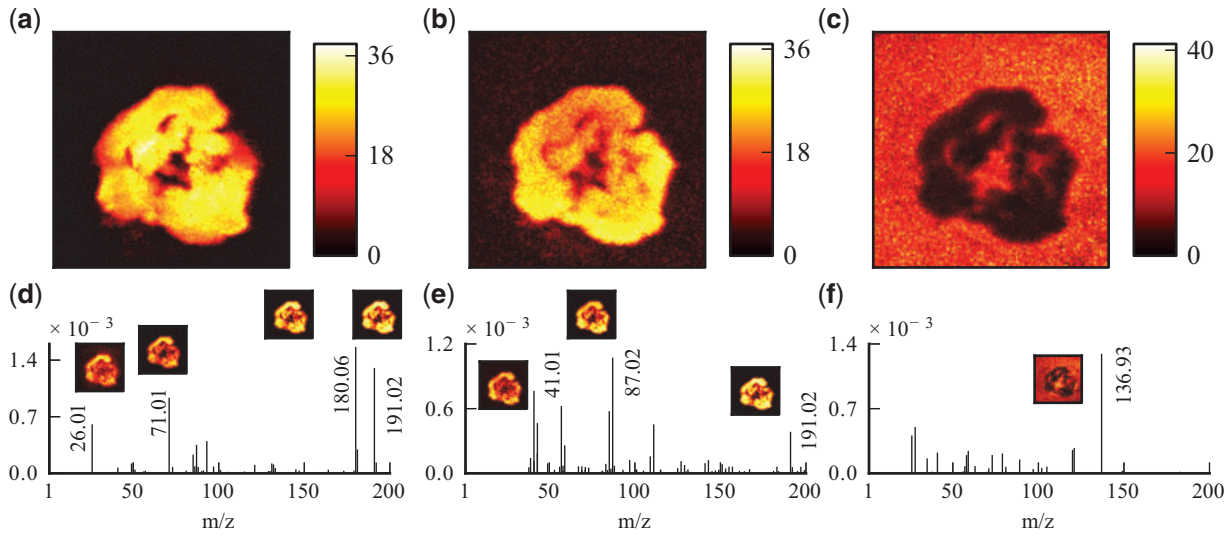

Fig. 5. Score and loading estimates using ALS with non-negativity constraint for TC mixture. (**a–c**) Score estimates; (**d–f**) loading estimates

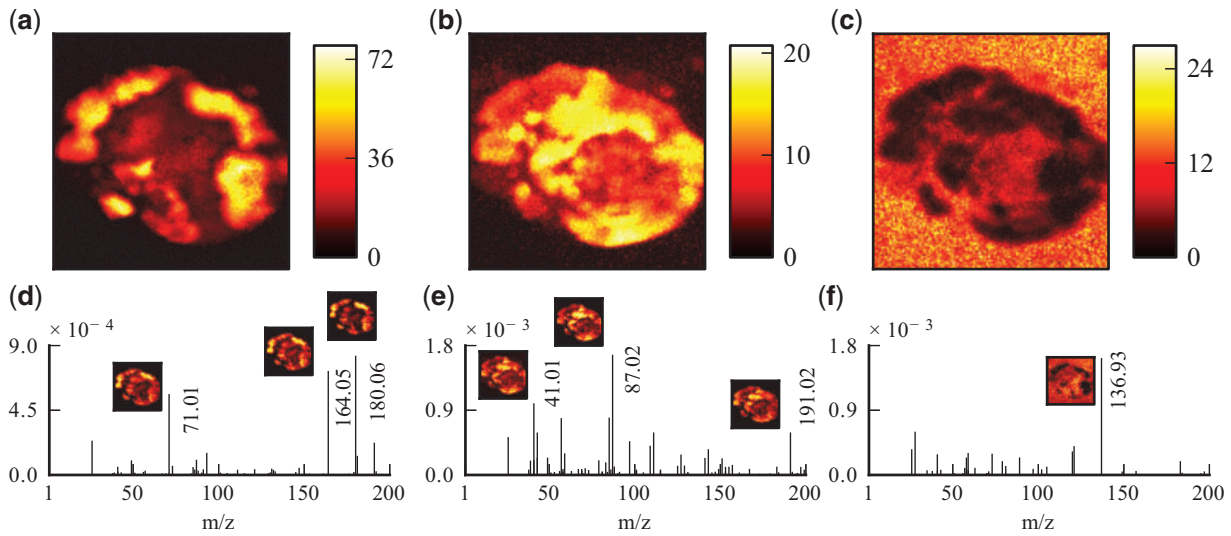

Fig. 6. Score and loading estimates using ALS with non-negativity constraint for TPC mixture. (**a–c**) Score estimates; (**d–f**) loading estimates

of *T* (Fig. 4d) and *P* (Fig. 4j) and cannot be used to distinguish these species. From this analysis, the characteristic peak for component *P* can be identified as $m/z = 164.05$, which would be the deprotonated metabolite ion $[M - H]^-$.

The result of the algorithm for TC mixture is illustrated in Figure 5. Similar to the previous example, we associate the peak at $m/z = 136.93$ in Figure 5f to the existence of the noise in the system. The peak at $m/z = 180.06$ of the first factor in component *T* can be also identified in Figure 5d with its corresponding scores image illustrated in Figure 5a. The same observation can be made for the component *C*, where three peaks of the first factor, that is, $m/z = 41.01$, 87.02 and 191.02, can be identified (Fig. 5e). However, the peaks at $m/z = 41.01$, 87.02 and 111.02 can be putatively identified as fragment ions $CHCO^-$, $C_4H_7O_2^-$ and $[M - H - CO_2 - 2H_2O]^-$, respectively. The algorithm performed well in this case where both *T* and *C* species were successfully identified. The analysis suggested $m/z = 180.06$ and $m/z = 191.02$ as characteristic peaks for *T* and *C*, respectively.

The results for the score and loading estimates for TPC mixture are depicted in Figure 6. The algorithm is able to identify all the species, *T* ($m/z = 180.06$), *P* ($m/z = 164.05$) and *C* ($m/z = 191.02$), as is shown in Figure 6(d–f). Again, the fragment ions at $m/z = 41.01$ and $m/z = 87.02$ are also present in the spectrum shown in Figure 6e. The algorithm is unable to separate the distribution of *T* ($m/z = 180.06$) and *P* ($m/z = 164.05$) as is shown in Figure 6a. We attribute this to the peak at $m/z = 71.01$, which is common between *T* and *P* species, making it difficult to separate the two types. However, the algorithm can almost isolate the distribution of *C* component from the mixture. Note a small peak at $m/z = 191.02$ still exists in Figure 6d.

From the above analysis, the important peaks required to identify *T*, *P* and *C* components are $m/z = 180.06$, 164.05 and 191.02, respectively. There are also strong peaks at $m/z = 121.02$ and $m/z = 183.02$ for *T* (Fig. 4d and e, respectively) and at $m/z = 111.02$ for *C* (Fig. 4p); however, these peaks are not present in TC and TPC mixtures. It should be noted that the peaks at $m/z = 41.01$ and 87.02 (Fig. 4p) are also present in TC and TPC spectra. Although these peaks can provide discriminatory information for the component C, they are fragment ions of the process. The results using appropriate number of spectral basis vectors, that is, $m = 1$ for *T*, *P* and *C* datasets and $m = 2$ for TC dataset, are also provided in Supplementary Figure S2, showing successful identification of different components in each case.

The extracted information was also used to perform similar analysis on replicate images of *T*, *P*, *C*, TC and TPC to further examine the performance of the algorithm on discerning different species. In this case, we used the correct number of spectral basis vectors, that is, $m = 1$, 2 and 3, respectively, for each pure replicate measurements, TC replicate measurements and TPC replicate measurements. The results are given in Supplementary Data, confirming the ability of the proposed algorithm to identify different components. In the case of pure species, the characteristic peak for the component *C* at $m/z = 191.02$ is dominated by fragment ions, that is, $m/z = 41.01$ and 87.02 (bottom panels of Supplementary Fig. S3c–f). However, this peak can be clearly identified from TC and TPC mixtures as is shown in Supplementary Figures S4 and S5.

In each experiment, the algorithm was allowed to run for a maximum number of 100 iterations, although typically the change in the weights matrix Frobenius norm dropped below 10^{-5} after less than 20 iterations.

We compared ion image estimates from the full model and the reduced model with TOF-SIMS data using the mean (over spectra) of the root mean square error (MRMSE) over space. The average MRMSE of the image estimates over three replicate measurements are shown in Table 1. The slight increase in the error is due to the low-pass action of the basis functions which attenuates the high frequency details of the scores images during the iterative algorithm.

A comparison between the convergence of the algorithms when reduced and full models were used is also shown in Figure 7. Note that for 128×128 pixels images and $m = 3$, each iteration of the algorithm takes 65.3 ms for the full model. This is 7.8 ms when the reduced model is used.

We also generated images using random numbers with different pixel densities to compare the computation time for the full model (FC-NNLS) and the reduced model (Algorithm 1 with $n = 81$) using different number of spectral basis vectors. The computation time is approximately 100 times faster when the reduced model is used as is shown in Figure 8.

Table 1. Model error for different experiments

Component	Full model	Reduced model ($n = 81$)
T	1.68×10^{-3}	1.83×10^{-3}
P	1.78×10^{-3}	1.96×10^{-3}
C	1.70×10^{-3}	1.77×10^{-3}
TC	1.36×10^{-3}	1.44×10^{-3}
TPC	1.30×10^{-3}	1.38×10^{-3}

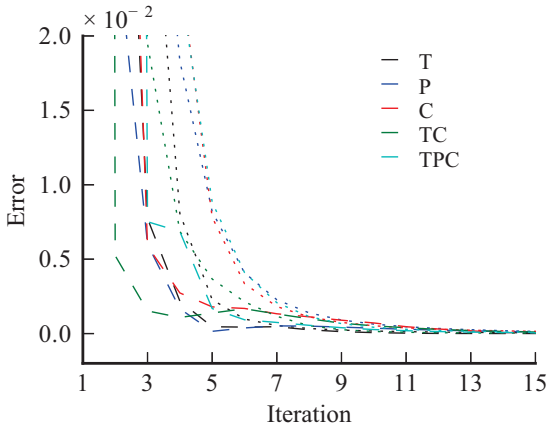

Fig. 7. Convergence of the estimation algorithm. The change in the Frobenius norm of the error, *E* (Equations (1) and (9)). The full and reduced models are shown by dotted and dashed lines, respectively

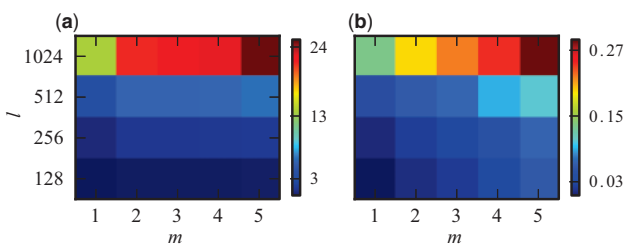

Fig. 8. Computation time in seconds for one step of the algorithm. (**a**) Full model; (**b**) reduced model using Algorithm 1 ($n = 81$). In each case, 1000 random images of the size $l \times l$ pixels are used

4 Discussion

This article has presented a novel framework for solving the alternating non–negativity-constrained least squares using TOF-SIMS measurement. The novel and key developments of the article include a continuous-over-space formulation which accounts for the spatial correlation across the sample surface and decoupling the computational complexity of the estimation procedure from the number of pixels in the ion images. This results in a significant reduction in the processing time that will translate favourably when high-resolution images are analysed and in the analyses of three-dimensional images, where voxels of information require processing.

To demonstrate the new methodology, the proposed algorithm was evaluated using image data from simple mixture of metabolites. Although the algorithm performed well, validating the framework on more complex test datasets is still required. Such datasets can include multiple components with similar mass spectrum but different intensities, patterned samples or images with higher pixel densities.

Acknowledgements

The authors thank Dr Claire Hurley for help with the TOF-SIMS data acquisition. This work was supported by the Engineering and Physical Sciences Research Council [EP/H00453X/1, EP/IO3453X/1].

Conflict of interest: none declared.

References

Alexandrov,T. and Bartels,A. (2013) Testing for presence of known and unknown molecules in imaging mass spectrometry. *Bioinformatics*, **29**, 2335–2342.

Alexandrov,T. and Kobarg,J.H. (2011) Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics*, **27**, i230–i238.

Belu,A.M. *et al*. (2000) TOF-SIMS characterization and imaging of controlled-release drug delivery systems. *Anal. Chem.*, **72**, 5625–5638.

De Juan,A. and Tauler,R. (2006) Multivariate curve resolution (MCR) from 2000: progress in concepts and applications. *Crit. Rev. Anal. Chem.*, **36**, 163–176.

Esteban,M. *et al*. (2000) Multivariate curve resolution with alternating least squares optimisation: a soft-modelling approach to metal complexation studies by voltammetric techniques. *Trends Anal. Chem.*, **19**, 49–61.

Fletcher,J.S. and Vickerman,J.C. (2012) Secondary ion mass spectrometry: characterizing complex samples in two and three dimensions. *Anal. Chem.*, **85**, 610–639.

Fletcher,J.S. *et al*. (2007) TOF-SIMS 3D biomolecular imaging of *Xenopus laevis* oocytes using buckminsterfullerene (C60) primary ions. *Anal. Chem.*, **79**, 2199–2206.

Freestone,D. *et al*. (2011) A data-driven framework for neural field modeling. *Neuroimage*, **56**, 1043–1058.

Gallagher,N.B. *et al*. (2004) Curve resolution for multivariate images with applications to TOF-SIMS and Raman. *Chemometr. Intell. Lab. Syst.*, **73**, 105–117.

Garrido,M. *et al*. (2008) Multivariate curve resolution—alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes. *Anal. Bioanal. Chem.*, **390**, 2059–2066.

Graham,D.J. and Castner,D.G. (2012) Multivariate analysis of ToF-SIMS data from multicomponent systems: the why, when, and how. *Biointerphases*, **7**, 1–12.

Hanrieder,J. *et al*. (2013) Time-of-flight secondary ion mass spectrometry based molecular histology of human spinal cord tissue and motor neurons. *Anal. Chem.*, **85**, 8741–8748.

Henderson,A. *et al*. (2009) A comparison of PCA and MAF for ToF-SIMS image interpretation. *Surf. Interface Anal.*, **41**, 666–674.

Jackson,J.E. (2005) *A User's Guide to Principal Components*. Vol. 587. John Wiley & Sons, New York.

Jiang,J.-H. *et al*. (2004) Principles and methodologies in self-modeling curve resolution. *Chemometr. Intell. Lab. Syst.*, **71**, 1–12.

Keenan,M.R. and Kotula,P.G. (2004) Accounting for poisson noise in the multivariate analysis of ToF-SIMS spectrum images. *Surf. Interface Anal.*, **36**, 203–212.

Kempson,I.M. and Skinner,W.M. (2005) ToF-SIMS analysis of elemental distributions in human hair. *Sci. Total Environ.*, **338**, 213–227.

Kim,H. and Park,H. (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, **23**, 1495–1502.

Kubicek,M. *et al*. (2014) A novel ToF-SIMS operation mode for sub 100 nm lateral resolution: application and performance. *Appl. Surf. Sci.*, **289**, 407–416.

Larsen,R. (2002) Decomposition using maximum autocorrelation factors. *J. Chemometr.*, **16**, 427–435.

Lawson,C.L. and Hanson,R.J. (1974) *Solving Least Squares Problems*. Vol. 161. SIAM.

Lee,J. *et al*. (2009) Multivariate image analysis strategies for ToF-SIMS images with topography. *Surf. Interface Anal.*, **41**, 653–665.

Paatero,P. and Tapper,U. (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, **5**, 111–126.

Park,J. and Sandberg,I.W. (1991) Universal approximation using radial-basis-function networks. *Neural Comput.*, **3**, 246–257.

Salim,M. *et al*. (2012) A solvation-based screening approach for metabolite arrays. *Analyst*, **137**, 2350–2356.

Sanner,R.M. and Slotine,J.-J. (1992) Gaussian networks for direct adaptive control. *IEEE Trans. Neural Netw.*, **3**, 837–863.

Scerri,K. *et al*. (2009) Estimation and model selection for an IDE-based spatio-temporal model. *IEEE Trans. Signal Process*, **57**, 482–492.

Tyler,B.J. (2006) Multivariate statistical image processing for molecular specific imaging in organic and bio-systems. *Appl. Surf. Sci.*, **252**, 6875–6882.

Vaidyanathan,S. *et al*. (2008) Subsurface biomolecular imaging of *Streptomyces coelicolor* using secondary ion mass spectrometry. *Anal. Chem.*, **80**, 1942–1951.

Van Benthem,M.H. and Keenan,M.R. (2004) Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. *J. Chemometr.*, **18**, 441–450.

Viornery,C. *et al*. (2002) Surface modification of titanium with phosphonic acid to improve bone bonding: characterization by XPS and ToF-SIMS. *Langmuir*, **18**, 2582–2589.

Wagner,M. and Castner,D.G. (2001) Characterization of adsorbed protein films by time-of-flight secondary ion mass spectrometry with principal component analysis. *Langmuir*, **17**, 4649–4660.

Wang,J.-H. *et al*. (2003) Application of modified alternating least squares regression to spectroscopic image analysis. *Anal. Chim. Acta*, **476**, 93–109.