Efficient quantitative hyperspectral image unmixing method for large-scale Raman micro-spectroscopy data analysis

E. G. Lobanova, S. V. Lobanov

Abstract

Vibrational micro-spectroscopy is a powerful optical tool, providing a non-invasive label-free chemically specific imaging for many chemical and biomedical applications. However, hyperspectral image produced by Raman micro-spectroscopy typically consists of thousands discrete pixel points, each having individual Raman spectrum at thousand wavenumbers, and therefore requires appropriate image unmixing computational methods to retrieve non-negative spatial concentration and corresponding non-negative spectra of the image biochemical constituents. Here, we present a new efficient Quantitative Hyperspectral Image Unmixing (Q-HIU) method for large-scale Raman micro-spectroscopy data analysis. This method enables to simultaneously analyse multi-set Raman hyperspectral images in three steps: (i) Singular Value Decomposition with innovative Automatic Divisive Correlation which autonomously filters spatially and spectrally uncorrelated noise from data; (ii) a robust subtraction of fluorescent background from the data using a newly developed algorithm called Bottom Gaussian Fitting; (iii) an efficient Quantitative Unsupervised/Partially Supervised Non-negative Matrix Factorization method, which rigorously retrieves non-negative spatial concentration maps and spectral profiles of the samples’ biochemical constituents with no a priori information or when one or several samples’ constituents are known. As compared with state-of-the-art methods, our approach allows to achieve significantly more accurate results and efficient quantification with several orders of magnitude shorter computational time as verified on both artificial and real experimental data. We apply Q-HIU to the analysis of large-scale Raman hyperspectral images of human atherosclerotic aortic tissues and our results show a proof-of-principle for the proposed method to retrieve and quantify the biochemical composition of the tissues, consisting of both high and low concentrated compounds. Along with the established hallmarks of atherosclerosis including cholesterol/cholesterol ester, triglyceride and calcium hydroxyapatite crystals, our Q-HIU allowed to identify the significant accumulations of oxidatively modified lipids co-localizing with the atherosclerotic plaque lesions in the aortic tissues, possibly reflecting the persistent presence of inflammation and oxidative damage in these regions, which are in turn able to promote the disease pathology. For minor chemical components in the diseased tissues, our Q-HIU was able to detect the signatures of calcium hydroxyapatite and β-carotene with relative mean Raman concentrations as low as 0.09% and 0.04% from the original Raman intensity matrix with noise and fluorescent background contributions of 3% and 94%, respectively.

Keywords: Raman spectroscopy, Hyperspectral image analysis, Biochemical quantification,Baseline correction, Non-negative matrix factorization, Multivariate curve resolution

1. Introduction

Over the past few decades, the development in multivariate image analysis methods has paved the way to handle and interpret the biochemical information, received from spectroscopic images quite efficiently. For instance, spontaneous Raman micro-spectroscopy data have been treated using multivariate image reconstruction methods [1, 2], such as Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), and K-Means cluster Analysis (KMA) [3, 4, 5, 6, 7], utilizing spectral contrast originated from biochemical composition variations over sample image pixels in order to produce pseudo-color images revealing this contrast. However, in these methods the spatio-spectral information is sorted into components that do not represent individual biochemical species with physically meaningful spectra and concentration.

Raman data have also been analysed using Vertex Component Analysis (VCA) [8] and Multivariate Curve Resolution (MCR) [3, 9, 10, 11], which is also known as Non-negative Matrix Factorization (NMF) [12]. In contrast to VCA which requires the presence of pixels containing pure biochemical substances, MCR/NMF solves a general hyperspectral image unmixing problem without this restrict-
tion and therefore has broader chemical and biomedical applications.

In relation to Coherent Raman Scattering (CRS) microspectroscopy [13, 14], Coherent Anti-Stokes Raman scattering (CARS) and Stimulated Raman scattering (SRS) images have been analysed by PCA [15], cluster analysis based on spectral phasor approach [16], MCR [17] and NMF [18] both based on the Alternating Non-negativity-constrained Least Squares algorithm (ANLS). In contrast to PCA and HCA, MCR/NMF can provide a quantitative determination of the chemical composition. Importantly, the popular MATLAB realization of the MCR-ANLS algorithm [19] uses computationally inefficient realization of the non-negative least squares (NLS) algorithm, which can be several orders of magnitude slower than fast combinatorial NLS (FC-NLS) [20, 21]. Indeed, the standard NLS algorithm [22] is designed to calculate a non-negative right hand-side (RHS) vector. The multiple-RHS problem, used to quantify the chemical information in a hyperspectral image, can be reduced to the independent solution of problems for each RHS vector. However, the computation of this technique is time consuming, which becomes even worse for big data-sets. For instance, the MCR-ANLS approach implemented by [19] has running times of days on a standard PC for a typical Raman data-set (≈50×50×103 matrix) from a sample of unknown chemical composition, which requires in-depth analysis with the different number of components. For large volume data analysis with multi-set Raman images (≈104×104×104 matrix) it leads to excessively long running times of years. To that end, there is a need for faster chemometric approaches, which can be several orders of magnitude slower than fast combinatorial NLS (FC-NLS) [20, 21]. Indeed, the standard NLS algorithm [22] is designed to calculate a non-negative right hand-side (RHS) vector. The multiple-RHS problem, used to quantify the chemical information in a hyperspectral image, can be reduced to the independent solution of problems for each RHS vector. However, the computation of this technique is time consuming, which becomes even worse for big data-sets. For instance, the MCR-ANLS approach implemented by [19] has running times of days on a standard PC for a typical Raman data-set (≈50×50×103 matrix) from a sample of unknown chemical composition, which requires in-depth analysis with the different number of components. For large volume data analysis with multi-set Raman images (≈104×104×104 matrix) it leads to excessively long running times of years. To that end, there is a need for faster chemometric approaches, which can be several orders of magnitude slower than fast combinatorial NLS (FC-NLS) [20, 21]. Indeed, the standard NLS algorithm [22] is designed to calculate a non-negative right hand-side (RHS) vector. The multiple-RHS problem, used to quantify the chemical information in a hyperspectral image, can be reduced to the independent solution of problems for each RHS vector. However, the computation of this technique is time consuming, which becomes even worse for big data-sets. For instance, the MCR-ANLS approach implemented by [19] has running times of days on a standard PC for a typical Raman data-set (≈50×50×103 matrix) from a sample of unknown chemical composition, which requires in-depth analysis with the different number of components. For large volume data analysis with multi-set Raman images (≈104×104×104 matrix) it leads to excessively long running times of years. To that end, there is a need for faster chemometric approaches, which reduce running times to hours for large data-sets without compromising the accuracy.

The goal of our paper is to develop efficient Quantitative Unsupervised/Partially Supervised Hyperspectral Image Unmixing (Q-HIU) method for the analysis of large-scale Raman micro-spectroscopy data.

2. Theory and algorithms

2.1. Formulation of hyperspectral image unmixing problem

Forward scattering problem.

Suppose the examining specimens are a mixture of $N$ pure chemicals (components) with known individual Raman scattering cross-section spectra $S_i(\nu)$, $i = 1, 2, \ldots, N$ and known spatial distributions of concentration $C_i(r)$. Here, $\nu$ indicates Raman shift and $r = (x, y; r)$ is a radius vector drawn to the image pixel with coordinates $(x, y)$ of the $r^{th}$ hyperspectral image. For this example problem, one could measure Raman intensity matrix $I(\mathbf{r}; \nu)$, which is a linear combination of the individual spatially-resolved concentrations and corresponding pure spectra products of the samples’ constituents

$$I(\mathbf{r}; \nu) = \sum_{i=1}^{N} C_i(\mathbf{r}) S_i(\nu).$$

This equation represents the solution of the forward scattering problem, which is computationally straightforward.

Note, Eq. (1) can be also written in the matrix form as follows (see Supplementary material for details)

$$I = \mathbf{C}^T \mathbf{S},$$

where the superscript $T$ means matrix transpose.

Inverse scattering problem. A more challenging problem, which is usually the case of chemical and biomedical applications, is to retrieve the chemical composition of samples without prior knowledge of their constituents. This inverse scattering problem can be formulated in a following way. For a given Raman intensity $I(\mathbf{r}; \nu)$, one should factorize it into $N$ separate chemical components with individual Raman spectra $S_i(\nu)$ and spatially-resolved concentration profiles $C_i(\mathbf{r})$. In this paper, we also consider the case, where one or several Raman spectra $S_i(\nu)$ can be known.

It is important to note that since this problem retrieves unknown chemical composition from the Raman images, concentration of analytes can be only identified up to a constant factor, which can be determined in Raman experiment for each analyte individually. For this, one could measure Raman spectrum $S_i(\nu)$ of pure chemical compound (identified as a component from the Q-HIU analysis) with known chemical concentration $c_i$ and multiply its concentration profile $C_i(\mathbf{r})$ by a calibration factor $c_i \int S_i(\nu) d\nu / \int S_i^0(\nu) d\nu$, which gives the chemical concentration of analyte. Since calibration factors are unknown for the Raman datasets to be investigated in this paper, we will use the term ‘Raman concentration’ instead of ‘chemical concentration’ to avoid confusion.

2.2. Singular Value Decomposition with Automatic Divisive Correlation (SVD-ADC) for noise filtering

A standard formulation of SVD algorithm [23] applied to hyperspectral intensity matrix $I$ enables to project the data in a new spatio-spectral orthonormal basis consisting of scores $C_i(\mathbf{r})$ and loadings $S_i(\nu)$

$$I(\mathbf{r}; \nu) = \sum_{i \epsilon N} C_i(\mathbf{r}) \lambda_i S_i(\nu),$$

which can be sorted by their descending singular values $\lambda_i$ [24]. Here, $N = \{1, 2, \ldots, \text{rank}(I)\}$ and $C_i(\mathbf{r}) \cdot C_j(\mathbf{r}) = S_i(\nu) \cdot S_j(\nu) = \delta_{ij}$, where the dots indicate Euclidean scalar products (see Supplementary material for definition).

It is believed that physically meaningful components have significantly higher singular values compared to noise ones. However, an adequate threshold criterion, which rejects relatively small singular values and its corresponding...
singular vectors, to the best of our knowledge, does not exist in literature. Furthermore, singular vectors (scores and loadings) with small singular values can contain meaningful chemical information and, therefore, must not be discarded. Such meaningful components tend to have high values of autocorrelation function for small pixel shifts. For instance, if a chemical component with distinct Raman spectral profile (high autocorrelation) is localized in a few pixels, SVD might retrieve it as a singular component with small singular value and, therefore, it will potentially be filtered out. To this end, autocorrelation function appears to be more accurate for assessing singular components compared to singular values.

To address this problem, we propose SVD in a new formulation, which we call SVD with Automatic Divisive Correlation (SVD-ADC). This approach requires additional step once SVD is performed, i.e.: for each left and right singular vectors ($C_i(r)$ and $S_i(\nu)$, respectively) we calculate spectral $R^S_i$ and spatial $R^C_i$ autocorrelation coefficients at ”one” pixel shift

$$R^S_i = R\left[S_i(\nu), S_i^{\delta}(\nu)\right],$$
$$R^C_i = \max R\left[C_i(r), C_i^{\delta}(r)\right],$$

where $R[\cdots]$ denotes correlation coefficient (see Supplementary material for definition), $S_i^{\delta}(\nu)$ denotes loadings shifted by ”one” spectral point, i.e. $S_i^{\delta}(\nu_j) = S_i(\nu_{j+\delta})$, and $C_i^{\delta}(r)$ represents scores shifted by ”one” image pixel along $x$-$\delta = (\delta_x, 0)$ or $y$-axis $\delta = (0, \delta_y)$. Here, double quotation marks for the word ”one” imply that one pixel shift should be chosen only when the spectral/spatial resolution of the used instrumentation is smaller than the pixel size of a CCD camera. For example, in our analysis presented below we used $\delta = (\delta_x = \delta_y = 1$ for dataset 2 and $\delta = (\delta_x = 3, \delta_y = 1$ for dataset 3.

This approach enables to automatically filter the meaningful components from noise-dominated ones (e.g.: shot noise, read noise, residual noise after cosmic rays removal, and etc.) by plotting the spatial autocorrelation coefficients $R^C_i$ against spectral ones $R^S_i$ for each pair of singular vectors and then discarding components with mean autocorrelation coefficient $R_{\text{th}} = (R^C_i + R^S_i)/2$ lower than $R_{\text{th}} = 50\%$. This number was testified to adequately differentiate the meaningful components from the noise-dominated ones, which can be simply explained by the fact that an ideally noisy component exhibits 0% autocorrelation coefficient, whereas it is approximately 100% for an ideally meaningful component. After replacing the set $N$ in Eq. (3) by the set $\mathcal{N}_{\text{thr}} = \{i: R_i > R_{\text{thr}}\}$ and performing summation over this set, we find Raman intensity matrix with reduced/removed noise.

2.3. Bottom Gaussian Fitting (BGF) for background subtraction

Raman spectra have an inherent broad background originating from fluorescence and amorphous scattering from glass substrate and sample itself, and therefore baseline correction algorithms via polynomial functions of different orders [25, 26] and Asymmetric Least Squares (AsLS) method [27, 28] are used for background removal. However, polynomial baseline correction methods might be hardly reproducible and give incorrect conclusions [29], whereas AsLS is believed to provide relatively accurate and automated background removal from hyperspectral Raman data [30]. Importantly, the AsLS algorithm requires the optimization over the smoothing $\lambda$ and asymmetric $\mu$ parameters, allowing broad background to be higher than total measured intensity which leads to negative non-physical Raman signal.

To this end, we propose a new algorithm, Bottom Gaussian Fitting (BGF), allowing to autonomously subtract a complex curved background from Raman microspectroscopy data resulting in the accurate quantification of the Raman bands in the spectra at each image pixel. Compared to AsLS, our BGF benefits from only one intuitive parameter, which has units of Raman wavenumbers (cm$^{-1}$) and means minimum width of broad background features. The principles of this mathematical approach are explained in the following.

For the sake of brevity, we will omit in this section the first index in the Raman intensity matrix $I$ labelling image pixel, i.e. we will write $I_j$ instead of $I_{ij}$ meaning that the background removing procedure must be done independently for each image pixel $i$.

Let us split the Raman intensity $I$ into two non-negative parts

$$I = I^R + I^B,$$

where $I^R$ contains sharp resonance peaks representing Ra-
man bands, whereas $\mathbf{I}^B$ is a quasi-slowly varied background approaching the Raman intensity $\mathbf{I}$ from the bottom. Our aim is to find $\mathbf{I}^B$ as a quasi-supersposition of Gaussian functions with standard deviations (STD) larger than $\sigma$ \textit{the only one parameter} that will be used for background subtraction.

Let us consider some Raman shift $\nu_j$ and find maximum possible background intensity at this point $I_j^B$. The figure 1A illustrates a series of Gaussian functions (black, green, and red lines) with fixed STD $\sigma$ and varying amplitude $A$ and expectation value $\mu$, which can potentially be background fits for a Raman spectrum $\mathbf{I}$ shown in Fig. 1A by the blue line. As one can see from the figure, only the red solid line tangents the Raman spectrum $\mathbf{I}$, so that the representative Raman intensity $\mathbf{I}^R$ at the point $\nu_j$ is well quantified. Mathematically, since the background intensity $\mathbf{I}^B$ tangents the Raman intensity $\mathbf{I}$, it should be determined as a Gaussian function $G_{\sigma}(\nu_j; \nu_i, I_i, \nu_k, I_k)$ with STD $\sigma$, which passes through left-sided ($\nu_i, I_i$) and right-sided ($\nu_k, I_k$) points of the Raman intensity $\mathbf{I}$ relating to the considered Raman shift $\nu_j$. Furthermore, the tangent background fit $\mathbf{I}^B$ (red solid line) has minimum Raman intensity value at the point $\nu_j$ as can be visually observed from the Fig. 1A (compare the red solid line with the red doted and green lines). Thus, the bottom Gaussian fit can be defined as

$$I_j^B = \min_{i \leq j \leq k} G_{\sigma}(\nu_j; \nu_i, I_i, \nu_k, I_k),$$ (7)

where the Gaussian function

$$G_{\sigma}(\nu_j; \nu_i, I_i, \nu_k, I_k) = Ae^{-\frac{(\nu_j-\nu)^2}{2\sigma^2}}$$ (8)

has amplitude $A$ and expected value $\mu$, which are implicitly defined by the following equations

$$Ae^{-\frac{(\nu_j-\nu)^2}{2\sigma^2}} = I_i, \quad Ae^{-\frac{(\nu_j-\nu)^2}{2\sigma^2}} = I_k.$$ (9)

Computing Eq. (7) for all Raman shifts $\nu_j$, we find a quasi-slowly varying curve $\mathbf{I}^B$ fitting the Raman intensity $\mathbf{I}$ from the bottom. The difference between these two curves ($\mathbf{I}^R = \mathbf{I} - \mathbf{I}^B$) contains only sharp resonances or Raman bands.

Once the background-free Raman-like signal $\mathbf{I}^R$ is retrieved using BGF, we calculate the spatially-resolved spectral auto-correlation coefficients for this matrix at "one" pixel shift and then discard the spatial points, which have correlation values lower than $R_{\min} = 50\%$, from the further analysis. This optional additional step allows to eliminate contribution of spatial points, which are strongly affected by noise due to dominated background (orders of magnitude larger than the remaining signal).

Note that Eq. (7) requires two loops (over the indexes $i$ and $k$), which is apparently computationally time consuming. However, it can be simply reduced to one loop, which is utilized in our BGF method.

2.4. Efficient Quantitative Unsupervised/Partially Supervised Non-negative Matrix Factorization (Q-US/PS-NMF)

In a standard formulation, the NMF approach [31, 32, 33, 34] allows to decompose the spatially-resolved Raman intensity matrix $\mathbf{I}$ into a product Eq. (2) of non-negative spatial concentration maps $\mathbf{C}$ and corresponding non-negative spectra $\mathbf{S}$ with rows representing individual biochemical substances (components) of the samples’ composition. Both matrices $\mathbf{C}$ and $\mathbf{S}$ are unknown prior to NMF.

In this paper, we generalise NMF to operate as a partially supervised method, which we call Q-US/PS-NMF. This modality can be particularly useful when required to eliminate the contribution of wax residues to the spectral basis for paraffin-embedded samples. For this, the spectrum of paraffin-wax compound is fixed during the Q-US/PS-NMF procedure.

Mathematically, the problem can be formulated as follows. Suppose $N_k$ biochemical substances with Raman spectra $S_i(\nu), i = 1, \ldots, N_k$ are known and our goal is to determine the rest $N_u = N - N_k$ spectra $S_i(\nu), i = N_k + 1, \ldots, N$, as well as concentration maps for both known and unknown components $C_i(r), i = 1, 2, \ldots, N$. Denoting known/unknown parts of the spectral matrix $\mathbf{S}$ and corresponding parts of the concentration matrix $\mathbf{C}$ with letters $k/u$, we have

$$\mathbf{C} = \begin{pmatrix} C_k \\ C_u \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} S_k \\ S_u \end{pmatrix}.$$ (10)

Ideally, these matrices should satisfy Eq. (2). However, for real experimental data, the Raman intensity matrix $\mathbf{I}$ can be factorized only approximately with a residue $\mathbf{E}$, i.e. $\mathbf{I} - \mathbf{C}^T \mathbf{S} = \mathbf{E}$. A pair of matrices $\mathbf{C}$ and $\mathbf{S}$, which minimises Frobenius norm of the residue $\lVert \mathbf{E} \rVert_F$, represents the solution of the factorization problem Eq. (2). This is equivalent to the minimization of the following function

$$F(\mathbf{C}, \mathbf{S}) = \frac{1}{2} \lVert \mathbf{I} - \mathbf{C}^T \mathbf{S} \rVert_F^2$$ (11)

subject to $C_{ij} \geq 0$, $S_{ij} \geq 0$. The necessary condition for a minimum of Eq. (11) is that a variation of the factorization error $\delta F$ at the point $(\mathbf{C}, \mathbf{S})$ is non-negative, i.e.

$$\delta F = \sum_{i=1}^{N} \sum_{j=1}^{N_k} \frac{\partial F}{\partial C_{ij}} \delta C_{ij} + \sum_{i=N_k+1}^{N} \sum_{j=1}^{N} \frac{\partial F}{\partial S_{ij}} \delta S_{ij} \geq 0.$$ (12)

Note, since variations $\delta C_{ij}$ and $\delta S_{ij}$ are arbitrary, each term of this equation must be greater or equal to zero.

Let us consider dependence of the factorization error $F$ on a variable $C_{ij}$ for fixed other variables. The graph of this function is parabola, which can have the minimum either at a point $C_{ij} > 0$ (blue line in Fig. 1B) or in the forbidden region $C_{ij} < 0$ (green line in Fig. 1B). In the former, variation $\delta C_{ij}$ (highlighted by red color in Fig. 1B) can be both positive and negative, which requires
the partial derivative $\partial F / \partial C_{ij}$ to be zero in order to satisfy Eq. (12). In the latter, the constrained minimum must be shifted from the region with negative values of $C_{ij}$ to the nearest possible point in the non-negative region, which is $C_{ij} = 0$, so the variation $\delta C_{ij}$ can have only positive values. Thus, the partial derivative $\partial F / \partial C_{ij}$ is not required to be zero and can take positive values. These requirements are called the Karush-Kuhn-Tucker conditions [22] and can be written in a following way

$$\frac{\partial F}{\partial C_{ij}} = 0 \text{ if } C_{ij} > 0, \quad \frac{\partial F}{\partial C_{ij}} \geq 0 \text{ if } C_{ij} = 0, \quad (13)$$

$$\frac{\partial F}{\partial S_{ij}} = 0 \text{ if } S_{ij} > 0, \quad \frac{\partial F}{\partial S_{ij}} \geq 0 \text{ if } S_{ij} = 0, \quad (14)$$

where the bottom formulas must be valid only for $i > N_k$, i.e. for unknown spectra.

Differentiating Eq. (11), we find partial derivatives

$$\frac{\partial F}{\partial C} = SS^T C - ST^T, \quad (15)$$

$$\frac{\partial F}{\partial S_u} = C_u C_u^T S_u - C_u I + C_u C_k^T S_k, \quad (16)$$

which lead to a non-linear problem after substitution them into Eqs. (13)-(14). This non-linear problem can be solved iteratively by alternative fixation of one matrix (C or $S_u$) and computation of another one ($S_u$ or C, respectively) using NLS algorithm [22]. Namely, alternately denoting $SS^T$ and $C_u C_u^T$ by A, $ST^T$ and $C_u I - C_u C_k^T S_k$ by B, unknown non-negative matrices C and $S_u$ by X, and independently considering each column $x$ and $b$ of the matrices $X = (x_1, x_2, \ldots)$ and $B = (b_1, b_2, \ldots)$, one find a solution of the NLS problem

$$x_p = A_{pp}^{-1} b_p \quad \text{subject to} \quad x_p > 0, \quad (17)$$
$$x_u = 0$$

Here, the indexes $p$ and $a$ denote projection of the vectors x, b and matrix A onto passive and active subspaces characterizing by two complementary sets $P$ and $A$, respectively:

$$P \cap A = \emptyset , \quad P \cup A = \{1, 2, \ldots\},$$
$$x_P = \{x_i : i \in P\}, \quad x_A = \{x_i : i \in A\},$$
$$A_pp = \{A_{ij} : i \in P, \ j \in P\}, \quad A_ap = \{A_{ij} : i \in A, \ j \in P\}$$

and similar for $B$.

The described procedure can be done independently for each column of the matrix X. However, working with the whole matrices rather than vectors is more computationally efficient [20, 21]. Indeed, one can group the columns that share the same index sets $P$ and compute the appropriate matrix inverse $A_{pp}^{-1}$ in Eq. (17) for each block. This procedure is apparently orders of magnitude faster than sequential computation of the matrix inverse for each individual column (see Secs. 4.2.3 and 4.3.4 for detail), which is used in the popular MATLAB realization of the MCR-ANLS algorithm by [19] (see Table 1 and Fig. S1 of the Supplementary material).

Note that Eqs. (15)-(16) can be generalized to the case of partially known spatial distributions of concentration, which might be relevant for other scientific applications. To this end, the term $S_u S_k^T C_k$ should be added to the right part of Eq. (15).

Furthermore, the Q-US/PS-NMF algorithm allows to quantitatively determine and compare pixel-by-pixel Raman concentrations of different biochemical components between samples by applying the following normalization conditions during final step of Q-US/PS-NMF analysis: the spectrum of each factorized component is normalized on const so that the mean spectral intensity for each component was the same and the sum of components’ mean Raman concentrations is set equal to one.

$$\int dv S_1(v) = \int dv S_2(v) = \ldots = \int dv S_N(v), \quad (18)$$

$$\frac{1}{N_p} \sum_{j=1}^{N} \sum_{i=1}^{N} C_i(r_j) = 1, \quad (19)$$

where $N_p$ is a total number of pixel points.

Overall, Q-HIU presented here enables to decompose large Raman multi-datasets, simultaneously analysed from a variety of samples, into individual chemical components with spatially-resolved concentration and spectral profiles, in a quantitative manner without prior knowledge of the chemical composition of samples or with partial knowledge of some constituents’ spectra.

3. Materials and methods

3.1. Software

Our Q-HIU software used to produce the results of this paper and containing three newly-developed functions: SVD-ADC, BGF, and Q-US/PS-NMF intended for the quantitative analysis of Raman microscopy images, is freely available from https://github.com/LobanovaEG-LobanovSV/Q-HIU.git.

3.2. Datasets

Dataset 1 consists of a superposition of Gaussian functions resembling Raman-like signal $I^{R}$ superimposed by a background $I^{B}$. The background has unitary amplitude $A^{B} = 1$, STD $\sigma^{B} = 400 \text{ cm}^{-1}$, and expected value $\mu^{B} = 1000 \text{ cm}^{-1}$. The amplitudes, STDs, and expected values for the ten Raman-like bands were randomly selected from the intervals (0, 1), (5, 20) cm$^{-1}$, and (370, 1783) cm$^{-1}$, respectively.

Dataset 2 consists of mixtures of Raman spectra of 8 pure chemical compounds (sphingomyelin, $\beta$-carotene, collagen, cholesteryl palmitate, elastin, iron transferrin, cholesterol, and iron (III) oxide) with simulated random spatial distribution of concentration (see Supplementary Figs. S31-S38). The dataset was superimposed by a white Gaussian noise and a background. The amplitude of the white Gaussian noise was chosen to have a dynamic range
of signal-to-noise ratio $\text{SNR} = (0.01, 10^{10})$. The background was modelled as a Gaussian function, which has a spectrum and simulated spatial distribution of concentration as shown in Supplementary Fig. S30.

Multi-dataset 3 consists of 3 large volume Raman images (see schematic of a hyperspectral data cube in Fig. 4A) of diseased aortic tissues, containing biochemical signatures of atherosclerotic plaque lesions (e.g.: foam cells/necrotic core, cholesterol crystals, calcification), and 3 Raman images of non-atherosclerotic age-matched control ones. In the following, we refer to atherosclerotic (A) and non-atherosclerotic control (C) samples as A1 (022,1000), A2 (025,1000), A3 (031,1005), and C1 (043,000), C2 (046,000), C3 (047,002), respectively (see [35] for details on tissue cohorts).

4. Results and discussion

In Sec. 2, we presented a novel efficient Q-HIU approach for large-scale Raman micro-spectroscopy data analysis, consisting of three methods: SVD-ADC, BGF, Q-US/PS-NMF, each of which represents main step in analysis. In the following, we demonstrate three examples of the Q-HIU analysis applied to both simulated and real experimental Raman micro-spectroscopy data.

4.1. Validation of BGF on artificial data (Dataset 1)

To show the efficiency of the developed algorithm, we give an example of the BGF procedure applied to an artificial spectrum displaying ten Raman bands (Fig. 1C). This input spectrum was chosen to be a superposition of Gaussian functions resembling Raman-like signal $\mathbf{I}^\text{AR}$ superimposed by a background $\mathbf{I}^\text{AB}$ with parameters as detailed in Sec. 3.2. As one can see from Fig. 1C, the background-free Raman-like signal $\mathbf{I}^\text{R}$ (black dashed line), resulted from background subtraction with the representative bottom Gaussian fit $\mathbf{I}^\text{B}$ of $\sigma = 300 \text{cm}^{-1}$ (red dashed line), is in a good agreement with the original Raman-like signal $\mathbf{I}^\text{AR}$ (green line).

4.2. Validation of Q-HIU on mixtures of pure biochemical references (Dataset 2)

4.2.1. Noise filtering: SVD-ADC

To show the efficiency of the SVD-ADC algorithm, we add white Gaussian noise $\mathbf{I}^\text{n}$ to the intensity matrix of biochemical mixtures of pure Raman spectra $\mathbf{I}^\text{mR}$ and apply our noise filtering procedure with no input parameters to the resulting matrix $\mathbf{I} = \mathbf{I}^\text{mR} + \mathbf{I}^\text{n}$. Fig. 2A shows dependence of noise removal factor (NRF) on SNR, where

$$\text{SNR} = \frac{\|\mathbf{I}^\text{mR}\|_F^2}{\|\mathbf{I}^n\|_F^2},$$

and NRF is defined as a relative difference between the denoised $\mathbf{I}^\text{R}$ and reference $\mathbf{I}^\text{mR}$ matrices normalized on noise power

$$\text{NRF} = \frac{\|\mathbf{I}^\text{R} - \mathbf{I}^\text{mR}\|_F^2}{\|\mathbf{I}^n\|_F^2}. $$

Figure 2: Noise removal efficiency of SVD-ADC applied to biochemical mixtures of pure Raman spectra with white Gaussian noise (dataset 2). A, Dependence of noise removal factor (NRF) on signal-to-noise ratio (SNR) for SVD-ADC (blue line) and MNF combined with S-G filtering (red line). B, Spatio-spectral autocorrelation coefficients map of singular vectors found from SVD-ADC for SNR = 10. The dashed diagonal line represents a decision line for mean autocorrelation coefficients $\rho$ at $\rho_{\text{thr}} = 50\%$, separating the coefficients above the line (circled cross signs - meaningful components) from those below it (cross signs - noise). C, Comparison of Raman spectrum after SVD-ADC procedure ($\mathbf{I}^\text{R}$, black dashed line) with the known reference ($\mathbf{I}^\text{mR}$, green line) for SNR = 10. The input Raman spectrum superimposed by white Gaussian noise is shown below ($\mathbf{I}^n$, blue line).
As can be seen from Fig. 2A, SVD-ADC allows to reduce noise from Raman signal by two orders of magnitude (NRF ≈ 0.01).

In order to illustrate the performance of SVD-ADC, we show the distribution of spatio-spectral autocorrelation coefficients for singular vectors found from SVD-ADC for SNR = 10 on Fig. 2B. Visual inspection verifies that the singular vectors representing noise are localised around the origin of coordinates, whereas the singular vectors attributed to meaningful components (circled cross signs) are in the upper-right half plane. Therefore, the dashed diagonal line with $R_{\text{thr}}$ = 50% separates the meaningful components from noise-dominated ones. Fig. 2C also confirms that the denoised Raman spectrum $I^R$ (black dashed line) is almost identical to the reference one $I^{mR}$ (green line), when the noise in the input Raman spectrum (blue line) is comparable with the reference Raman signal.

In order to show the accuracy of the proposed method, we compare our SVD-ADC with the Maximum Noise Fraction (MNF) method [36] combined with Savitzky-Golay (S-G) filtering [37]. For this, we applied MNF on the same hyperspectral dataset of known chemical mixtures superimposed by white Gaussian noise and retained $N$ eigenimages sorted according to decreasing SNR or reducing image quality. The rest eigenimages were denoised using S-G filtering with two smoothing parameters: polynomial order and frame length. In brief, we performed optimization over these 3 parameters for each SNR, which minimizes denoising error. As the final step, the calculation of inverse MNF on reduced noise eigenimages allowed to produce the resulting denoised data. Fig. 2A shows that both methods give similar NRF for noise-dominated data with SNR from the range (0.01, 1). However, when SNR is higher than 10, MNF reduces noise by only 80% (NRF = 20%), whereas SVD-ADC allows to achieve 98% of noise reduction (NRF = 2%). Moreover, the determination of optimal parameters for MNF and S-G filtering is possible only when the true solution is known. In contrast, our SVD-ADC method allows to autonomously reduce noise from the hyperspectral Raman data with no input parameters.

4.2.2 Background subtraction: BGF

To validate the accuracy of BGF, we compare it with AsLS algorithm [27, 28] using biochemical mixtures of pure Raman spectra $I^{mR}$ with Gaussian background. We investigated the dependence of background removal relative error

$$E = \frac{\max_{ij} |I_{ij} - I_{ij}^{mR}|}{\max_{ij} I_{ij}^{mR}}$$ (22)

on corresponding parameters of BGF ($\sigma$, blue line with crosses) and AsLS ($\lambda$ and fixed $p = 10^{-3}$, red line with pluses) and the results reveal that BGF allows to achieve relative error of 7%, whereas AsLS gives relative error of 64% (see Fig. 3A). Here, $I^R$ is a background-free Raman intensity matrix retrieved from BGF or AsLS. Examples of Raman spectra (blue lines) at two spatial pixels with low (B) and high (C) true background (red dashed lines) show that BGF works equally well for low/high background (black lines), whereas AsLS may produce the baseline (green lines), which significantly overfits the original Raman signal and also has non-physical negative values (see Fig. 3B-C). Note that Fig. 3B-C are shown for optimal parameters found from Fig. 3A. These are $\sigma = 449 \text{ cm}^{-1}$ for BGF and $\lambda = 223, p = 10^{-3}$ for AsLS.

4.2.3 Q-US/PS-NMF unmixing procedure

To validate the accuracy of Q-US/PS-NMF, we compare it with the MCR-ALS algorithm [19] and VCA [38] using biochemical mixtures of 8 pure Raman spectra (32×32×804 data matrix). The results reveal that VCA is unable to retrieve the original chemical components and produces the relative factorization error of 70%. Also, our Q-US/PS-NMF shows fast convergence to the true solution and allows to achieve relative decomposition error of 2.6%, 0.7%, and 0.5% for 1, 5, and 10 minutes of standard PC running times, respectively, which are about 20 times accurate compared to MCR-ALS (see Table 1).

Table 1: Comparison of relative factorization error of Q-US/PS-NMF, MCR-ALS, and VCA for the 32×32×804 data matrix (dataset 2) performed on Intel Core i5-2537M

<table>
<thead>
<tr>
<th>CPU time (min)</th>
<th>Q-US/PS-NMF</th>
<th>MCR-ALS</th>
<th>VCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>70%</td>
<td>70%</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.6%</td>
<td>46%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.7%</td>
<td>11%</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.5%</td>
<td>11%</td>
<td></td>
</tr>
</tbody>
</table>

(0.4 sec)

4.3 Validation of Q-HIU on real data (Dataset 3)

In this section, we apply the Q-HIU method to the large-scale analysis of real experimental Raman microscopy images of human thoracic aortic tissues, which were sourced from [39] (see description of the data in Sec.3.2), and compare its performance with existing
state-of-the-art methods. The Q-HIU analysis was performed simultaneously on all Raman images, enabling to show the computational efficiency of the developed approach. The resulting hyperspectral Raman multi-image was a \( \approx 6 \times 10^2 \times 10^3 \times 10^3 \) matrix, where the first, second, third, and fourth dimensions represent the numbers of participating samples, spatial pixels in \( x \)- and \( y \)-directions, and spectral points, respectively.

Importantly, the new developed Q-US/PS-NMF method in a partially supervised formulation also presented in this paper is already used by us and co-authors in the quantitative analysis of large-scale label-free Raman hyperspectral images from biomedical paraffin-embedded samples in Alzheimer’s disease [40]. In that paper, wax component was fixed during the Q-US/PS-NMF procedure in order to avoid the spectral contamination of the biochemical components, characteristic for human brain tissue, by wax residues.

4.3.1. Noise filtering: SVD-ADC

As well-documented in the literature, conventional SVD procedure for spectral image processing allows to separate signal from noise by rejecting small singular values and their corresponding singular vectors from the further analysis. The SVD-ADC approach, proposed in this paper, utilises autocorrelation coefficients of spatial and spectral singular vectors for noise filtering and allows to autonomously remove noise from Raman data with no input parameters. In particular, for each pair of singular vectors we calculate spatial and spectral autocorrelation coefficients and discard the singular vectors, which have a mean autocorrelation coefficient smaller than 50%.

Fig. 4D shows the autocorrelation coefficients map for singular vectors received from the SVD-ADC analysis of 6 Raman images. The dashed diagonal line represents a decision line at \( R_{\text{thr}} = 50\% \) cut-off, that separates the coefficients above the line (meaningful components) from those below it (noise). Note that read-noise and shot noise are localized around zero, whereas physically meaningful components show high mean autocorrelation.

Importantly, the new approximating matrix, received after SVD on given data, is compressed (or has significantly reduced-rank), therefore it is important to carefully choose the cut-off value and avoid rejection of meaningful components. For our data example, SVD-ADC with a 50% cut-off of mean autocorrelation coefficients has found the approximating matrix, consisting of 19 singular components. To verify this number, we show the spatial distributions and corresponding spectra for two singular vectors with mean autocorrelation coefficients, which are 20% lower and higher than 50% value (Fig. 4E,F). For a pair of discarded singular vectors, corresponding to 30% mean autocorrelation, spatial maps do not show any recognised pattern and spectrum resembles noise, whereas for a pair of meaningful singular vectors with 70% mean autocorrelation, spatial maps show some distinct features and spectrum exhibits several clear Raman bands.

4.3.2. Background subtraction: BGF

Fig. 4B,C shows an example of our BGF background subtraction algorithm applied to the Raman data, found from SVD-ADC as detailed in the previous section. The Raman spectra at two spatial pixels (high and low background) from the selected hyperspectral image before (black line) and after (green line) background removal with the representative bottom Gaussian fit (red dashed line) are shown.

4.3.3. Q-US/PS-NMF unmixing procedure and interpretation of results

After SVD-ADC filtering and BGF background subtraction, we analysed the Raman data using Q-US/PS-NMF in the fingerprint (370 – 1783 cm\(^{-1}\)) region. The Q-US/PS-NMF analysis found that the given 6 Raman images were optimally described using 11 separate components (C). To validate this component number selection, we performed Q-US/PS-NMF with the numbers of components varied from 1 to 18, and then investigated the evolution of resulting component spectra (see Fig. 5D), which includes the formation of new biochemical components (red numbers in circles) and their splitting/mixing. The degree of component similarity between two analyses with subsequent numbers of components is indicated by solid (strong), dashed (medium), and dotted (weak) lines. For example, the formation of \( C_1 \) and \( C_{10} \) in 11-component analysis, which will be later attributed to actin and hydroxyapatite, respectively, was a result of splitting of \( C_9 \) in 10-component analysis. Note the components are sorted by their mean Raman concentration, from largest to smallest. The dependence of the relative factorization error on the number of components was also inspected. The results (see Fig. 5 of the Supplementary material) show a gradual decrease of this error when increasing the number of components, and therefore imply that the factorization error can not be used to determine the optimal number of components in the Q-US/PS-NMF data analysis.

7 components were found to be spectrally assigned to actin (\( C_1 \), \( R^2 = 85\% \)); elastin (\( C_3 \), \( R^2 = 84\% \)); a mixture of phospholipids and triglycerides (\( C_4 \), \( R^2 = 93\% \)) at a ratio of about 2:1 (Fig. 5A); a mixture of cholesterol, cholesteryl esters with saturated fatty acid chains, and oxidized all-trans retinol also known as vitamin A (\( C_5 \), \( R^2 = 94\% \)) at ratios of about 6:5:1 (Fig. 5B); collagen (\( C_6 \), \( R^2 = 83\% \)); hydroxyapatite (\( C_{10} \), \( R^2 = 93\% \)); a mixture of \( \beta \)-carotene and cholesterol (\( C_{11} \), \( R^2 = 88\% \)) at a ratio of about 3:1 (see the component spectra in Figs. S6-S16 of the Supplementary material). Here, the contributions of biochemical reference components into each component spectra was determined using NLS fitting algorithm. The degree of similarity between each component spectra and its biochemical model was indicated by a correlation coefficient (\( R^2 \)). The model uses the Raman spectra of the chemical species of analytical standards [35, 41, 42]. Additionally, \( C_2 \) was found to be consistent with the Raman spectrum of water (\( R^2 = 94\% \)) using an ID expert tool of Bio-Rad’s
Figure 4: A, Schematic of a hyperspectral data cube. The arrows mark two pixels, Raman spectra of which are shown on the panels B and C. B-C: The input Raman spectrum (blue line) before SVD-ADC and the corresponding noise-ﬁltered Raman spectrum (black line) after this procedure. The results of background subtraction using BGF is shown by the green line: the bottom Gaussian ﬁt (red dashed line) is subtracted from the noise-ﬁltered Raman spectrum, resulting in background-free Raman spectrum. D, As Fig. 2B. E-F, The spectra and spatial distributions of two singular vectors with the mean autocorrelation coefﬁcients of (50±20)% as indicated on the panel D. On both panels, 6 Raman images are labelled according to the sample source. Scale bars are 100 μm.
Figure 5: A-C, Component spectra (blue lines) and spatial distributions of Raman concentration of C4, C5, and C8 found from Q-US/PS-NMF of 6 Raman images from atherosclerotic aortic tissue regions (A1, A2, A3) and non-atherosclerotic controls (C1, C2, C3). The chemical attribution of components is based on the comparison with analytical standard Raman spectra of pure chemical species. Fits are shown by red dashed lines. Partial contributions (green, magenta, and cyan lines) to each component spectra (blue lines) are also indicated in the inset. Scale bars are 100 µm. D, Evolution of component spectra resulting from Q-US/PS-NMF with the numbers of components varied from 1 to 18, showing the appearance of new biochemical components (red numbers in circles) and their splitting/mixing. The degree of component similarity between two analysis with subsequent numbers of components is indicated by solid (strong), dashed (medium), and dotted (weak) lines.
with triglyceride and phosphocholine in the foam cells. This oxidized linoleic acid component shows one order of magnitude increase in concentration for the atherosclerotic plaque lesions compared to the atherosclerotic plaque-negative and non-atherosclerotic control regions. Altogether, this observation indicates the interplay between oxidized linoleic acid and cholesterol/triglyceride-rich components in the aortic tissue, which might be one of the possible molecular mechanisms involved in the evolution of the atherosclerotic plaque. Two independent studies, which used several different animal models of atherosclerosis, also found that oxidized fatty acids in the diet increase fatty streak lesion formation in the aortic tissues [46, 47]. Furthermore, our quantitative analysis reveals the significant levels of oxidative damage components strongly correlated with the atherosclerotic lesions in the aorta and therefore might indicate the persistent presence of inflammation known to be implicated in atherogenesis [48]. In particular, the components predominantly assigned to oxidized triglyceride ($C_4$), cholesterol ($C_5$) and linoleic acid ($C_4$) and accumulated in the atherosclerotic lesions of the aortic tissues show about 20, 60, 12-fold increase in their maximum concentrations, respectively, compared to the non-atherosclerotic control regions (for comparison we use the samples A1-A3 and C3). The previous studies also support the oxidation hypothesis of atherogenesis, suggesting that the products of lipid peroxidation in the aorta stimulate the innate immune system response, which triggers the uncontrolled development of foam cells, resulting in the initiation and progression of the atherosclerotic lesion [49, 50].

Also, the final results received from the quantitative Q-HIU analysis reveals the ability of the proposed method to detect both major and minor chemical components of atherosclerotic and non-atherosclerotic tissues. Table 2 shows that Q-HIU allows to identify low concentrated chemical compounds down to 0.04% from the original Raman data matrix characterised by fluorescent background of 94% and noise of 3%.

Altogether, we have demonstrated that the new developed Q-HIU method presented in this paper is capable to quantitatively decompose the spatially-resolved Raman spectra of human atherosclerotic aortic tissues into a number of individual biochemical components with relative mean Raman concentration down to 1%. These are structural proteins (actin, elastin, collagen), pathological lipid aggregates (cholesterol/cholesterol ester and triglyceride crystals), and minerals, which are characteristic biomarkers of atherosclerosis. Furthermore, our quantitative analysis of Raman micro-spectroscopy images allowed to identify and quantify oxidatively modified lipids (oxidative stress and inflammation bio-marker) found to be correlated/co-localized with the atherosclerotic lesions in the aorta, implying the use of the Q-HIU method in the quantitative characterization of human aortic tissues and potentially in the diagnosis of atherosclerosis.
Table 2: Relative mean Raman concentrations of noise, background and chemical components retrieved from the Q-HIU analysis of the biomedical Raman dataset of atherosclerotic tissues and non-atherosclerotic controls.

<table>
<thead>
<tr>
<th></th>
<th>Normalization 1*</th>
<th>Normalization 2**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AS</td>
<td>non-AS</td>
</tr>
<tr>
<td>noise background</td>
<td>94%</td>
<td>93%</td>
</tr>
<tr>
<td>$C_1$</td>
<td>0.6%</td>
<td>3%</td>
</tr>
<tr>
<td>$C_2$</td>
<td>0.7%</td>
<td>1%</td>
</tr>
<tr>
<td>$C_3$</td>
<td>1%</td>
<td>0.6%</td>
</tr>
<tr>
<td>$C_4$</td>
<td>1%</td>
<td>0.2%</td>
</tr>
<tr>
<td>$C_5$</td>
<td>0.8%</td>
<td>0.3%</td>
</tr>
<tr>
<td>$C_6$</td>
<td>0.5%</td>
<td>0.7%</td>
</tr>
<tr>
<td>$C_7$</td>
<td>0.3%</td>
<td>0.7%</td>
</tr>
<tr>
<td>$C_8$</td>
<td>0.6%</td>
<td>0.3%</td>
</tr>
<tr>
<td>$C_9$</td>
<td>0.2%</td>
<td>0.2%</td>
</tr>
<tr>
<td>$C_{10}$</td>
<td>0.09%</td>
<td>0.2%</td>
</tr>
<tr>
<td>$C_{11}$</td>
<td>0.04%</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

AS = atherosclerotic aortic tissue regions
* Relative mean Raman concentrations normalized on the mean original Raman intensity matrix (see Supplementary material for definition)
** Relative mean Raman concentrations of chemical components normalized according to Eqs. (18)-(19)

4.3.4. Comparison of Q-US/PS-NMF with existing methods

In this section, we will compare the performance of our Q-US/PS-NMF with the state-of-the-art methods such as MCR, FC-NMF and VCA using the same biomedical dataset of atherosclerotic aortic tissues as in the previous section. The results show that our realisation of NMF algorithm, Q-US/PS-NMF, which is implemented in a fast combinatorial ANLS framework, has exponential convergence to a local minimum with 200 and three times faster computational performance compared to MCR-ANLS [19] (see Fig. S1 of the Supplementary Material) and FC-NMF [21] (see Fig. S2 of the Supplementary material), respectively.

Comparison with the VCA method reveals that the relative factorization error in the 11 components’ factorization analysis using this method was 11% versus 5% for our Q-US/PS-NMF. Also, the VCA was not able to identify cholesterol component, which is a known pathological biomarker of atherosclerosis, as well as linoleic acid. Furthermore, the $R^2$ values for the majority of component spectra found from the VCA are significantly lower compared to that for Q-US/PS-NMF (see Table 3 for detail). Overall, compared to VCA, Q-US/PS-NMF allows to identify additional chemical components which are characteristic bio-markers of the diseased tissues.

5. Conclusion

To conclude, we present a new efficient Quantitative Hyperspectral Image Unmixing (Q-HIU) method, allowing to autonomously analyse large-scale Raman micro-spectroscopy data with minimum input parameters (only two) and high accuracy (see an overview flowchart of Q-HIU analysis in Fig. 6). This in-house developed method integrates three consecutive steps of data analysis, called Singular Value Decomposition with innovative Automatic Divisive Correlation, Bottom Gaussian Fitting, and an efficient Quantitative Unsupervised/Partially Supervised Non-negative Matrix Factorization, which is capable to operate as a partially supervised method, when one or several samples' constituents are known. This is implemented by fixing the spectra of the known compounds in the approximating factorization expansion of the data matrix.

Compared with other existing hyperspectral image processing methods, Q-HIU shows significant improvement in accuracy as verified on both simulated and large volume Raman images of atherosclerotic tissues. The experimental results found from Q-HIU reveal a proof-of-principle for the biomolecular characterisation and quantitative imaging of individual biochemical components in the samples of various complexity (e.g.: the retrieval of individual component spectra and spatially-resolved concentration profiles of cholesterol/cholesteryl ester, oxidized linoleic acid, calcium hydroxyapatite, β-carotene crystals with relative mean Raman concentration down to 0.04% in atherosclerotic plaques compared to 0.1% in non-atherosclerotic controls from the original Raman intensity matrix with strong fluorescent background of 94% and noise of 3%). The great advantage of Q-HIU to directly work with matrices allows to achieve two orders of magnitude acceleration in computational speed as compared with the popular realization of MCR-ANLS.

Acknowledgements

The authors acknowledge S.G. Tikhodeev, N.A. Gippius, K. Triantafillou, P. Borri, W. Langbein, and F. Masia for useful discussions.

Funding

This work has been supported by the Cardiff University College of Biomedical and Life Sciences under the International Scholarship for PhD student and RFBR Project No. 16-29-03283.

Appendix A. Supplementary material

Supplementary material contains vector notations, definitions of Euclidean scalar product, correlation coefficient, and normalization on the mean original intensity matrix, as well as presents comparison of Q-US/PS-NMF...
with other software, convergence of Q-US/PS-NMF to local/global minima, results of the Q-HIU and VCA analysis, Raman spectra and concentration profiles of the dataset 2.

Table 3: Comparison of $R^2$ values for chemical components retrieved from Q-US/PS-NMF and VCA of the same Raman images as in Fig. 5

<table>
<thead>
<tr>
<th>Chemical component attribution</th>
<th>$R^2$ for component number $C_i$ for 11 components</th>
<th>$R^2$ for component number $C_i$ for 11 components</th>
<th>$R^2$ for component number $C_i$ for 9 components</th>
</tr>
</thead>
<tbody>
<tr>
<td>actin</td>
<td>85% ($C_1$)</td>
<td>77% ($C_{10}$)</td>
<td>not found</td>
</tr>
<tr>
<td>water</td>
<td>94% ($C_2$)</td>
<td>94% ($C_9$)</td>
<td>not found</td>
</tr>
<tr>
<td>elastin</td>
<td>84% ($C_3$)</td>
<td>90% ($C_1$), 83% ($C_9$)</td>
<td>89% ($C_7$)</td>
</tr>
<tr>
<td>phospholipids + triglycerides</td>
<td>93% ($C_4$)</td>
<td>91% ($C_4$)</td>
<td>86% ($C_7$)</td>
</tr>
<tr>
<td>cholesteryl esters</td>
<td>87% ($C_5$)</td>
<td>91% ($C_5$)</td>
<td>92% ($C_5$)</td>
</tr>
<tr>
<td>+ vitamin A</td>
<td>89% ($C_6$)</td>
<td>92% ($C_5$)</td>
<td>93% ($C_6$)</td>
</tr>
<tr>
<td>+ cholesterol</td>
<td>93% ($C_6$)</td>
<td>not found</td>
<td>not found</td>
</tr>
<tr>
<td>collagen</td>
<td>83% ($C_7$)</td>
<td>68% ($C_1$), 92% ($C_9$), 70% ($C_9$)</td>
<td>91% ($C_1$)</td>
</tr>
<tr>
<td>magnetite</td>
<td>75% ($C_8$)</td>
<td>not found</td>
<td>not found</td>
</tr>
<tr>
<td>linoleic acid</td>
<td>90% ($C_9$)</td>
<td>not found</td>
<td>not found</td>
</tr>
<tr>
<td>triglycerides + arachidonic acid</td>
<td>61% ($C_9$)</td>
<td>not found</td>
<td>not found</td>
</tr>
<tr>
<td>hydroxyapatite</td>
<td>93% ($C_{10}$)</td>
<td>66% ($C_7$)</td>
<td>not found</td>
</tr>
<tr>
<td>$\beta$-carotene</td>
<td>80% ($C_{10}$)</td>
<td>68% ($C_{11}$)</td>
<td>68% ($C_6$)</td>
</tr>
<tr>
<td>+ cholesterol</td>
<td>88% ($C_{11}$)</td>
<td>not found</td>
<td>not found</td>
</tr>
</tbody>
</table>

Calculations of $R^2$ values were based on correlation of each component spectra with the pure biochemical references. * indicates presence of water in the component.

Hyperspectral Raman multi-image I

Pre-processing
- cosmic ray removal, wavenumber calibration, etc.
  (not described in this paper)

- Noise filtering
  - Required parameters: no
  - Optional parameters: $R_{th}$, $\delta$, $\beta$

- Background removal
  - BGF
  - Required parameters: $\sigma$ (cm$^{-1}$)
  - Optional parameters: $R_{bg}$, $\delta$

Are there any points with spectral autocorrelation less than $R_{th}$?

- Yes
- No

Remove these points

- Image unmixing
  - Q-US/PS-NMF
  - Are there any pure chemical components with known spectral profiles in the Raman images?

- Yes
- No

Unsupervised
- Required parameters: $N$

Partially supervised
- Required parameters: $N, S$

Raman concentration

Figure 6: Flowchart of Q-HIU analysis.


