Title

Identifying Important Ions and Positions in Mass Spectrometry Imaging Data Using CUR Matrix Decompositions

Journal

Analytical Chemistry, 87(9)

ISSN

0003-2700

Authors

Yang, Jiyan

Rübel, Oliver

Prabhat

et al.

Publication Date

2015-05-05

DOI

10.1021/ac5040264

Peer reviewed

analytical chemistry

Article

pubs.acs.org/ac

Identifying Important Ions and Positions in Mass Spectrometry Imaging Data Using CUR Matrix Decompositions

Jiyan Yang,[†] Oliver Rübel,[‡] Prabhat,[‡] Michael W. Mahoney,[§] and Benjamin P. Bowen*[,∥]

[†]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, California 94305, United States

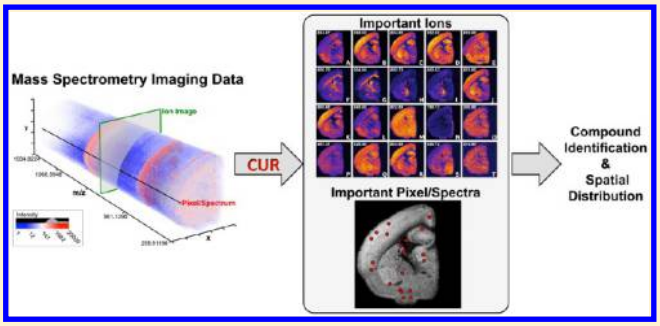
[‡]Computational Research Division, Lawrence Berkeley Lab, One Cyclotron Road, Berkeley, California 94720, United States

[§]International Computer Science Institute and Department of Statistics, University of California, Berkeley, California 94720, United States

[∥]Life Sciences Division, Lawrence Berkeley Lab, One Cyclotron Road, Berkeley, California 94720, United States

Supporting Information

ABSTRACT: Mass spectrometry imaging enables label-free, high-resolution spatial mapping of the chemical composition of complex, biological samples. Typical experiments require selecting ions and/or positions from the images: ions for fragmentation studies to identify keystone compounds and positions for follow up validation measurements using microdissection or other orthogonal techniques. Unfortunately, with modern imaging machines, these must be selected from an overwhelming amount of raw data. Existing techniques to reduce the volume of data, the most popular of which are principle component analysis and non-negative matrix factorization, have the disadvantage that they return difficult-to-interpret linear combinations of actual data elements. In this work, we show that CX and CUR matrix decompositions can be used directly to address this selection need. CX and CUR matrix decompositions use empirical statistical leverage scores of the input data to provide provably good low-rank approximations of the measured data that are expressed in terms of actual ions and actual positions, as opposed to difficult-to-interpret eigenions and eigenpositions. We show that this leads to effective prioritization of information for both ions and positions. In particular, important ions can be found either by using the leverage scores as a ranking function and using a deterministic greedy selection algorithm or by using the leverage scores as an importance sampling distribution and using a random sampling algorithm; however, selection of important positions from the original matrix performed significantly better when they were chosen with the random sampling algorithm. Also, we show that 20 ions or 40 locations can be used to reconstruct the original matrix to a tolerance of 17% error for a widely studied image of brain lipids; and we provide a scalable implementation of this method that is applicable for analysis of the raw data where there are often more than a million rows and/or columns, which is larger than SVD-based low-rank approximation methods can handle. These results introduce the concept of CX/CUR matrix factorizations to mass spectrometry imaging, describing their utility and illustrating principled algorithmic approaches to deal with the overwhelming amount of data generated by modern mass spectrometry imaging.



Recent advances in chemical imaging techniques have enabled detailed investigation of metabolic processes at length scales ranging from subcellular to centimeter resolution. One of the most promising chemical imaging techniques is mass spectrometry imaging (MSI).[1,2] Typically in MSI, a laser or ion beam is raster scanned across a surface. At each location, molecules are desorbed from the surface, often with the assistance of a matrix coating or specially prepared surface that enables the formation of gas phase ions. These ions are collected and analyzed by mass spectrometry.[3]

MSI presents many data analysis and interpretation challenges due to the size and complexity of the data. MSI acquires one or more mass spectra at each location. Each spectrum is digitized into 10^4 to 10^6 m/z bins. Depending on the sample and analysis technique, it is common to have tens of thousands of intense, sharp peaks at each location. Likewise, MSI data sets containing up to a million pixels are possible with existing technology. This results in a situation where each file is 10s to 100s of gigabytes, and careful analysis requires sophisticated computational tools, infrastructure, and algorithms to reduce the large volume of measured data into easier to interpret smaller blocks with the goal of prioritizing ions and positions according to their importance. The two most widely used techniques for this are principle component analysis (PCA) and non-negative matrix factorization (NMF).[4,5] These

Received: October 28, 2014
Accepted: March 31, 2015
Published: March 31, 2015

DOI: 10.1021/ac5040264
Anal. Chem. 2015, 87, 4658−4666

approaches express the original data in terms of concise but in general difficult-to-interpret components.[6−10]

In PCA and NMF, synthetic matrices are created from the original data such that these synthetic matrices can be combined to give a close approximation of the original data set. For example, by comparing the ions and locations with relatively large coefficients, one can quickly distinguish regions that have overall different spectra.[11] This approach can accelerate the interpretation of the large data sets generated by MSI by providing a manageable approximation that can be analyzed in a timely manner. Unfortunately, the synthetic coefficients are typically difficult to interpret: for example, eigenvectors are often not meaningful in terms of the physical processes of metabolism, sample preparation, and data collection; and in addition, it is not always clear whether a single ion is the distinguishing characteristic of a region or whether it is a complex combination of relative ion-intensities that distinguish regions.

In contrast, CUR and the related CX matrix decompositions are relatively new algorithmic approaches that allow scientists to provide a low-rank approximation of the measured data that is expressed in terms of actual data elements.[12,13] CX and CUR decompositions are provably almost as good as the low-rank approximation provided by the SVD, but instead of the blocks containing eigenions and eigenpositions, as they do with the SVD, the low rank approximation provided by CX/CUR is expressed in terms of actual rows and/or columns, i.e., actual ions and/or actual positions.

In this paper, CX/CUR matrix decompositions are applied to mass spectrometry imaging data sets and we show that this can lead to effective prioritization of information, both in terms of identifying important ions as well as in terms of identifying important positions. Previously, this approach has been applied to the study of gene expression and astronomy.[12,14,15] Here, we briefly introduce the concepts of CX/CUR matrix decompositions to the MSI literature, and we study in detail how they can be applied to identify (in a tractable manner for moderately large MSI data) important ions and locations in MSI data.

METHODS

Notation and Backgrounds. We start with some notation and basic linear algebra. For any $m \times n$ matrix A, consisting of m rows and n columns, we use a^i and a_j to denote the ith row and jth column of A, respectively. We also use a_{ji} to denote the jth element of the ith row of A. Suppose rank $(A) = r$. Let $A = U\Sigma V^T$ be the singular value decomposition (SVD) of A, where U and V are orthonormal matrices consisting of the left- and right-singular vectors and $\Sigma = \mathrm{Diag}\,(\sigma_1,...,\sigma_r)$ is a diagonal matrix containing the singular values. In particular, these satisfy $\sigma_1 \geq ... \geq \sigma_r \geq 0$, and this means that the columns of U and V are sorted by the order given by the singular values. Finally, we use $A^\dagger$ to denote the pseudoinverse of A.[16]

Leverage Scores and CX Decompositions. Given an $m \times n$ matrix A, the CX decomposition decomposes A into two matrices C and X, where C is an $m \times c$ matrix that consists of c actual columns of A, and X is a $c \times n$ matrix such that $A \approx CX$. (CUR decompositions can then be constructed by choosing rows from A to construct a matrix R by applying the CX decomposition to A^T.) That is, linear combinations of the columns of C can recover most of the "information" of the matrix A. A quantitative measurement of the closeness between CX and A is obtained by using the matrix Frobenius norm of the difference: if the residual error $\|A - CX\|_F$ is smaller, then CX provides a better quality approximation to A.

The construction of C follows the following two steps. First, compute (either exactly or approximately) the statistical leverage scores of the columns of A; and second, use those scores to select c columns from A. Once the matrix C is determined, the optimal matrix X that minimizes $\|A - CX\|_F$ can be computed by a least-squares approximation as $X = C^\dagger A$. In the following, we will elaborate more on the two steps of constructing C.

Given an $m \times n$ matrix A and a target rank parameter $k \geq 0$, for $j = 1, ..., n$, the jth leverage score can be defined as

$$l_j = \sum_{i=1}^{k} v_{ji}^2 \tag{1}$$

These scores $\{l_j\}_{i=1}^{n}$ can be interpreted as how much "leverage" or "influence" the jth column of A exerts on the best rank-k approximation to A.[12] To be more specific, recall that, for any matrix A, the best rank-k approximation of A is $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$. In other words, A_k gives the lowest possible error $\|A - B\|_F$ among all the rank-k matrix B. In fact, A_k can be viewed as the projection of A onto the top-k left singular space spanned by the columns of $(u_1...u_k)$. Since multiplying each column by the corresponding singular value does not alter the subspace, we can view $(\sigma_1 u_1...\sigma_k u_k)$ as a basis for this space. Then, for each column of A, we have that

$$a_j = \sum_{i=1}^{r} (\sigma_i u_i) v_{ji} \approx \sum_{i=1}^{k} (\sigma_i u_i) v_{ji}$$

That is, the jth column of A can be expressed as a linear combination of the basis of the top-k left singular space with v_{ji} as the coefficients. On the other hand, the scores $\{l_j\}_{j=1}^{n}$ equal to the diagonal elements of the projection matrix onto the top-k right singular subspace spanned by $(v_1...v_k)$, and thus these statistical leverage scores are a generalization of the diagonal elements of the "hat matrix" in regression diagnostics.[12] For $j = 1,...,n$, if we define the normalized leverage scores as

$$p_j = \frac{l_j}{\sum_{i=1}^{n} l_i} \tag{2}$$

and choose columns from A according to those normalized leverage scores, then the selected columns are able to reconstruct the matrix A nearly as well as A_k does.

To compute the normalized leverage scores exactly, i.e., using eqs 1 and 2, one needs to compute the full SVD. This takes $O(mn \times \min(m,n))$ time, which becomes inapplicable when dealing with data sets of even moderately large size. For completeness and as a control, we will use this naive method on a smaller data set, but to apply CX/CUR decompositions to larger data we will use the faster algorithms of Drineas et al.[17] These algorithms compute high-quality approximations to the normalized leverage scores of the input matrix, and the running time of these algorithms depends on the time to apply a random projection to the input matrix, which is much faster than computing the full (or even a truncated) SVD. We summarize the two ways of computing leverage scores of a given matrix as follows.

(a) ExactLev: Compute the normalized leverage scores exactly by using eqs 1 and 2.

(b) ApprLev: Compute approximations to the normalized leverage scores by using Algorithm 4 or Algorithm 5, proposed

DOI: 10.1021/ac5040264
Anal. Chem. 2015, 87, 4658−4666

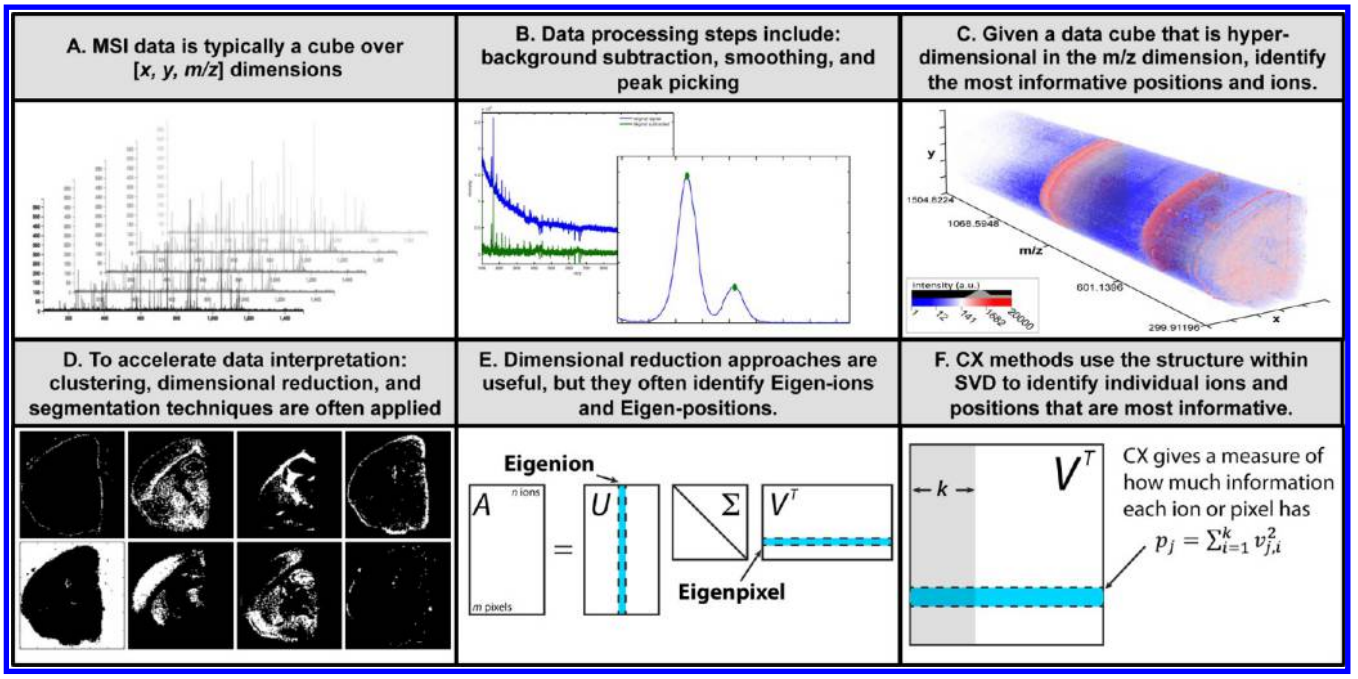


Figure 1. Mass spectrometry imaging collects one or more spectra at each location in a sample. Because of the scale and complexity of MSI data, computational tools are required to reach an understanding of the underlying physical processes. Panels A–D: A traditional processing workflow where raw data is cleaned and processed using traditional clustering and dimension reduction methods. Panel E: Multivariate statistics, such as PCA, yield informative combinations of ions and pixels, but they do not lend themselves to intuitive interpretation in terms of the biological processes generating the data. Panel F: In contrast, CX decomposition yields the most informative actual ions and actual positions instead of linear combinations of ions and positions.

by Drineas et al.;[17] we will refer to these as SPECTRALAPPRLEV and FROBENIUSAPPRLEV, respectively.

Then, with these normalized leverage scores at hand, one can select columns from A either by viewing p_j's as an importance sampling distribution over the columns and randomly sampling columns according to it or by viewing p_j's as a ranking function and greedily selecting the columns with highest scores.

(a) RANDCOLSELECT: Select c columns from A, each of which is randomly sampled according to the normalized leverage scores $\{p_j\}_{j=1}^{n}$.

(b) DETERCOLSELECT: Select the c columns of A corresponding to the largest c normalized leverage scores p_j's.

Finally, our main algorithm CX DECOMPOSITION is the following. It takes as input an $m \times n$ matrix, A, a rank parameter, k, and desired number of columns c as inputs.

(1) Compute the leverage scores by either ExactLev or ApprLev.

(2) Select c columns from A according to RandColSelect or DeterColSelect.

(3) Let $X = C^{\dagger}A$.

It has been shown by Drineas et al. that if RANDCOLSELECT is used and the sampling size $c = \boldsymbol{O}(k \log k/\epsilon^2)$, then with probability at least 0.99, the output of CX DECOMPOSITION, C, X will satisfy

$$\|A - CX\|_F \leq (1 + \epsilon)\|A - A_k\|_F \tag{3}$$

where A_k is the best rank-k approximation to A.[13] A freely available implementation of CUR decomposition in the R programming language is available and provides an excellent reference.[18]

As is illustrated in Figure 1, the CX decomposition uses the leverage score structure within SVD to find actual rows and actual columns of an MSI matrix that are most informative. In each computation that will be described in the next section, after having specified which scheme is used to compute the leverage scores, i.e., EXACTLEV or APPRLEV, we will, respectively, use randomized CX decomposition and deterministic CX decomposition to denote the algorithm CX DECOMPOSITION with RANDCOLSELECT or DETERCOLSELECT scheme.

RESULTS AND DISCUSSION

Data and Approach. In the following we use two data sets to demonstrate the utility of CX decompositions for MSI. These two data sets are publicly available on the OpenMSI Web gateway, and they are selected from two diverse acquisition modalities, including one NIMS image of the left coronal hemisphere of a mouse brain acquired using a time-of-flight (TOF) mass analyzer and one MSI data set of a lung acquired using an Orbitrap mass analyzer.[19−22] These files are previously described elsewhere and were chosen because of the commonality of brain-lipid images and the large number of m/z bins generated by Orbitrap detectors, respectively. To illustrate the utility of CX, we focus initially on results obtained from the NIMS image of a coronal brain section. For the analyses described for the NIMS brain image, the data were processed using peak-finding. The peak-finding identifies the most intense ions and integrates the peaks, so that each peak is represented by a single image, rather than a series of images spanning a range of m/z values. Using this approach, the original data is reduced from O 100 000 m/z values to the most intense ions. The size of the brain section data set is $(122 \times 120 \times 1926)$. The (i,j,l)th value of the matrix represents the intensity of the ion with the lth m/z value at position (i,j) in a (122×120) regular lattice which discretizes physical space. To compute the CX decomposition and select ions and spectra, we reshape the

DOI: 10.1021/ac5040264
Anal. Chem. 2015, 87, 4658−4666

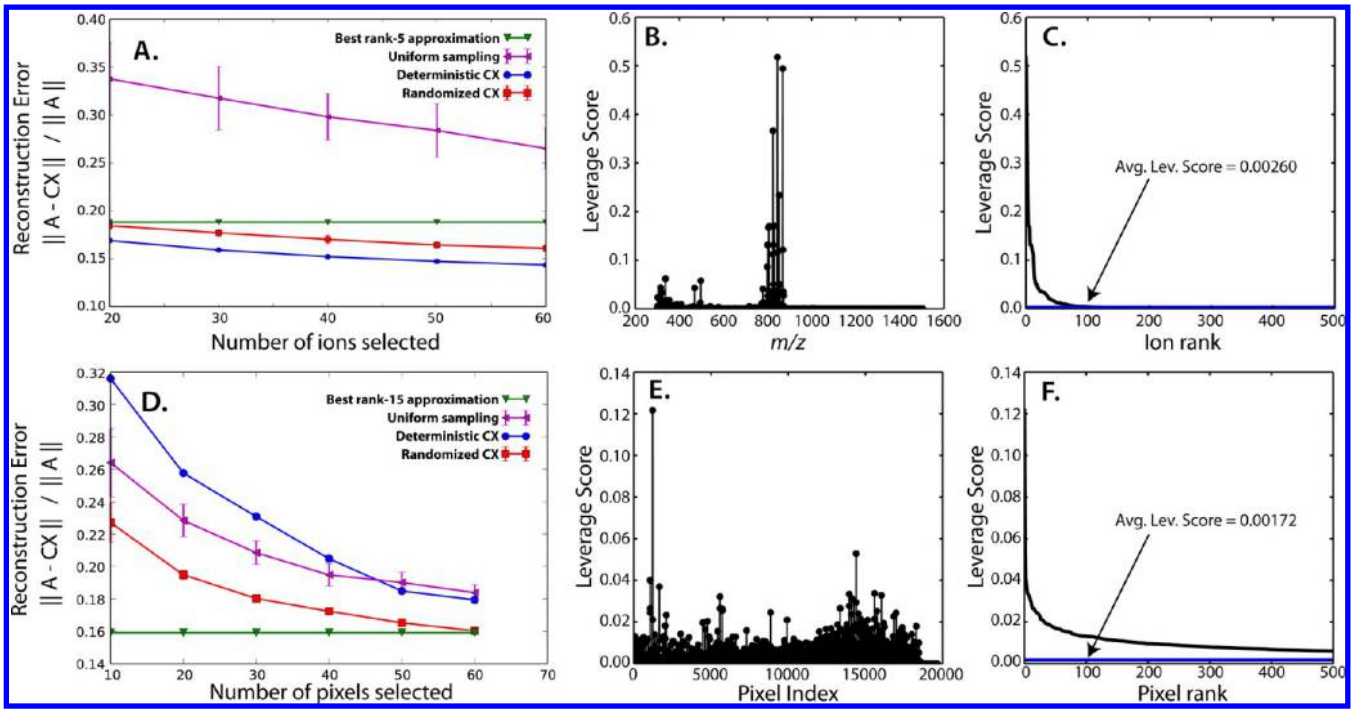


Figure 2. Analysis of the reconstruction error used to determine the most appropriate CX-based schemes and settings for selection of ions and locations/spectra on the brain data set. Panels A and D: Reconstruction error of the CX decomposition for selection of ions (panel A) and locations (panel D) using randomized and deterministic selection schemes with a varying parameter c. Panels B and E: Distribution of leverage scores of A and A^T, relative to the best rank-k space, respectively. Panels C and F: Sorted distribution of the leverage scores of A and A^T, respectively. The blue horizontal line denotes the mean/average leverage score. Because of the fairly nonuniform shape of the leverage score distribution for ions, deterministic CX selection outperforms randomized CX sampling for ions. In contrast, pixel selection is best achieved by randomized CX sampling, since the leverage score distribution for pixels is much more uniform.

three-dimensional MSI data cube into a two-dimensional (14640 × 1926) matrix A, where each row of A corresponds to the spectrum of a pixel in the image, and where each column of A corresponds to the intensities of an ion over all pixels, describing the distribution of the ion in physical space. For finding informative ions and pixels, we perform CX DECOMPOSITION with exact computations for leverage scores, i.e., EXACTLEV, on A and A^T, respectively. In each case, for clarity, we only report the results with a fixed small value of the rank parameter k. Varying in a range of small values does not have a large effect on the reconstruction errors. This behavior may indicate that the information that the corresponding top-k singular spaces contain does not vary a lot as k varies in this range.

Finding Important Ions. Figure 2A shows the reconstruction errors $\|A - CX\|_F/\|A\|_F$ using CX decomposition for selection of c = 20, 30, 40, 50, 60 ions, using a rank parameter k = 5 and using both randomized and deterministic CX decompositions. For completeness, we also show the reconstruction errors using uniform sampling for varying numbers of selected ions and that of the optimal rank-k approximation of A. Figure 2B,C shows the distribution of the leverage scores of A, relative to the best rank-k space, and their relative magnitudes. Figure 3 then presents the spatial distributions of the 20 most important ions selected using deterministic CX decomposition with k = 5 and c = 20.

The selection of important ions from the brain data set (Figure 2A) shows clearly that using deterministic CX decomposition will lead to a smaller error than using randomized CX decomposition with the same parameters. The reason for this behavior lies in distribution of the leverage scores for the ions, as shown in Figure 2B,C. These leverage scores are very nonuniform: a few dozen leverage scores are much larger, e.g., 50 times larger than the average score. Hence, since the leverage scores are highly nonuniform, the corresponding ions can be considered as very informative in reconstructing the matrix, and keeping the ions with the top leverage scores leads to a good basis. The randomized CX decomposition carries a large variance, for the values of the parameters used here, since in many trials it failed to select those important ions, and thus it resulted in a large error. Not surprisingly, uniformly selecting columns do not give particularly meaningful results, i.e., many irrelevant ions were chosen and informative ions were not chosen.

As for the absolute magnitude of the error, we use that of the best rank-k approximation of A, i.e., A_k, as a reference scale suggested by eq 3. In Figure2A,D, we can see that the reconstruction error of the CX decomposition is close to that of A_k. In some cases, CX decomposition can even produce a lower error. This is because the matrix CX returned by CX DECOMPOSITION is a rank-c matrix with $c > k$. It is possible to choose X to be a rank-k matrix; see section 4.3 in Drineas et al. for detailed construction.[13]

Finding Important Pixels/Spectra. Similar to Figure 2A–C, Figure 2D–F provide an overview of the reconstruction errors and the distribution and magnitude of the leverage scores, relative to the best rank-k approximation, for the application of CX decomposition to A^T for selection of pixel. In Figure 4, we illustrate the application of both randomized and deterministic CX decompositions, with k = 15 and c = 20, on A^T for finding informative pixels. The first subplot (Figure 4A) shows the result returned by the deterministic CX decom-

DOI: 10.1021/ac5040264
Anal. Chem. 2015, 87, 4658–4666

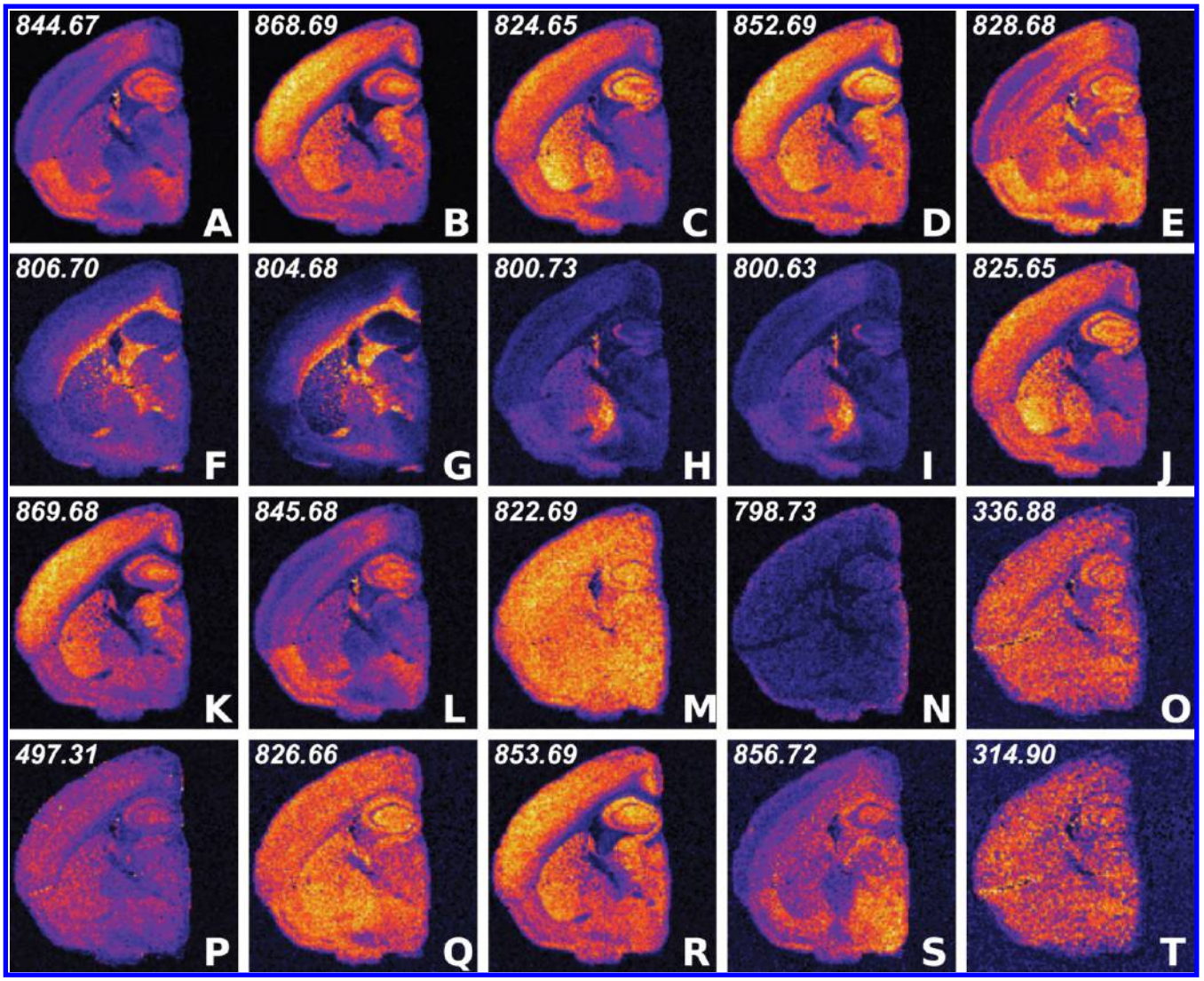


Figure 3. Ion-intensity visualization of the 20 most important ions selected via deterministic CX decomposition with $k = 5$ and $c = 20$ on brain data set. The distribution of leverage scores is presented in Figure 2B. Some of these ions map to distinct regions in the brain. Particular regions of the cortex, pons, and corpus collosum stand out as distinct anatomically identifiable regions. Also in the list are likely background ions and contaminants from the embedding material. Of the 20 ions, little redundancy is present, pointing to the effectiveness of the CX approach for information prioritization.

position, meaning the pixels with the top leverage scores are greedily selected and plotted. The remaining subplots in Figure 4B−F we show the results returned by running randomized CX decomposition in five independent trials.

In contrast with the selection of ions, deterministic CX decomposition results in larger reconstruction errors than randomized CX decomposition (Figure 2D). Also, the pixels selected using CX tend to be more localized in specific regions of the images, rather than selecting characteristic pixels from different physical components of the sample images. The reason for this behavior lies in the distribution of the leverage scores for the pixels, as shown in Figure 2E,F. These leverage scores are fairly uniform: most of them are less than 20 times the average. Also, there are many more pixels than ions, and thus we can consider the distribution of leverage scores to be fairly uniform. Furthermore, since each row in A represents a pixel in the image, many rows will contain a similar spectrum. Similar locations tend to "split up" the leverage scores, resulting in smaller values for the score at each location. Importantly, applying random sampling here may still be able to identify pixels from the important regions (i.e., those with high total leverage scores), even when the value of any of its single pixel is small.

Comparison with Established Factorization Methods. As is mentioned above, non-negative matrix factorization (NMF) has been widely applied in the MSI literature. Like principle component analysis (PCA), NMF factors the MSI data into two matrices whose product serves as a low rank approximation to the original matrix. Because of the positive values in the coefficients, the factored data from NMF has a more meaningful appearance and is often preferred by experimentalists. Shown in Figure 5 is a three-component visualization of the brain data set using NMF. In each of the three components, an image and a spectrum are shown. The images corresponding to spatial-component coefficients guide the identification of regions characterized by a component, and the spectra corresponding to the ion-component coefficients

DOI: 10.1021/ac5040264
Anal. Chem. 2015, 87, 4658−4666

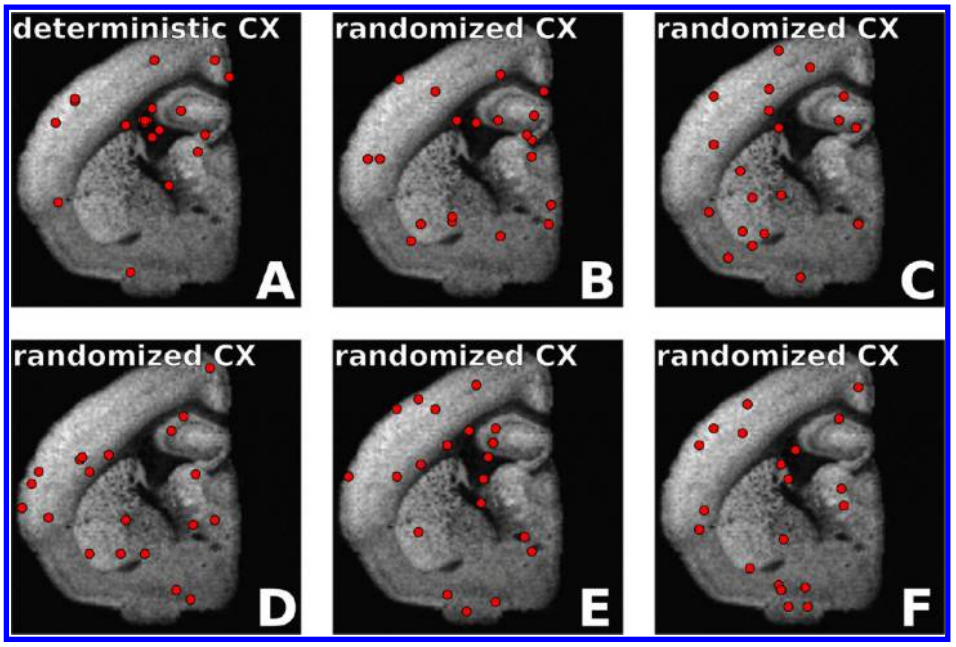


Figure 4. Visualization of the selection of important pixels using CX decompositions on the brain data set. All visualizations show a gray scale image of a selected ion as context, and the 20 locations selected using the CX decomposition with $k = 15$ and $c = 20$ are highlighted via red circles. Panel A shows the result of using the deterministic CX decomposition. With this approach, the algorithm selects locations clustered around a few regions. In comparison, panels B−F show the results from five independent trials using the randomized CX decomposition. Because of the uniformity in leverage scores for pixels, the randomized selection outperforms the deterministic approach for comprehensive sampling of important locations. The distribution of leverage scores is presented in Figure 2E.

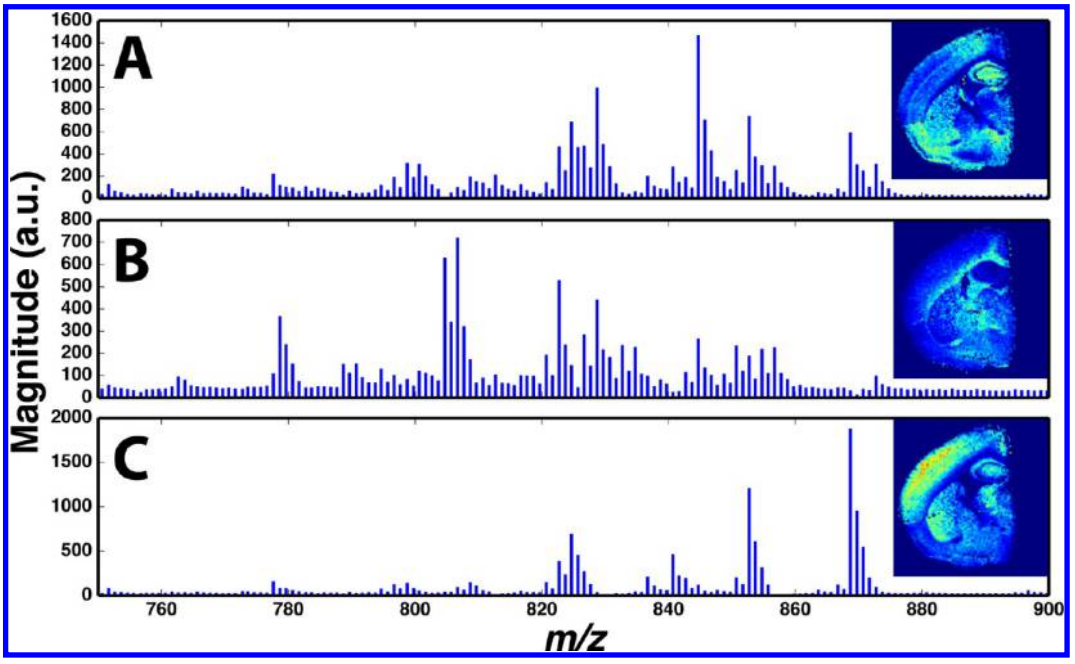


Figure 5. Visualization of three components returned by using NMF on the brain data set. They are shown in the three panels, respectively, each of which shows the image corresponding to the spatial-component coefficients and the spectrum corresponding to given ion-component coefficients. Many of the informative ions identified by CX in Figure 3 have large-magnitude coefficients in the spectra corresponding to given ion-components. For the NMF approach, however, the relative importance of each ion and pixel corresponding to each component is not provided.

show what a characteristic spectrum could look like for those regions.

In comparison to NMF retrieving characteristic spectra that describe a linear combination of measured spectra, CUR and CX methods retrieve individual spectra from specific locations. Likewise, in comparison to retrieving overall images, the CUR and CX methods retrieve images of specific ions. Thus, CUR and CX methods allow the reconstruction of the original data set using a limited set of spectra from specific locations and specific ions. On the other hand, with NMF and PCA, the factorization produces matrices containing weighted coefficients for all ions and all locations. Consequently, it is hard with NMF and PCA to tell the significance of specific ions or pixels given the components. In fact, all the ions and locations are

DOI: 10.1021/ac5040264
Anal. Chem. 2015, 87, 4658−4666

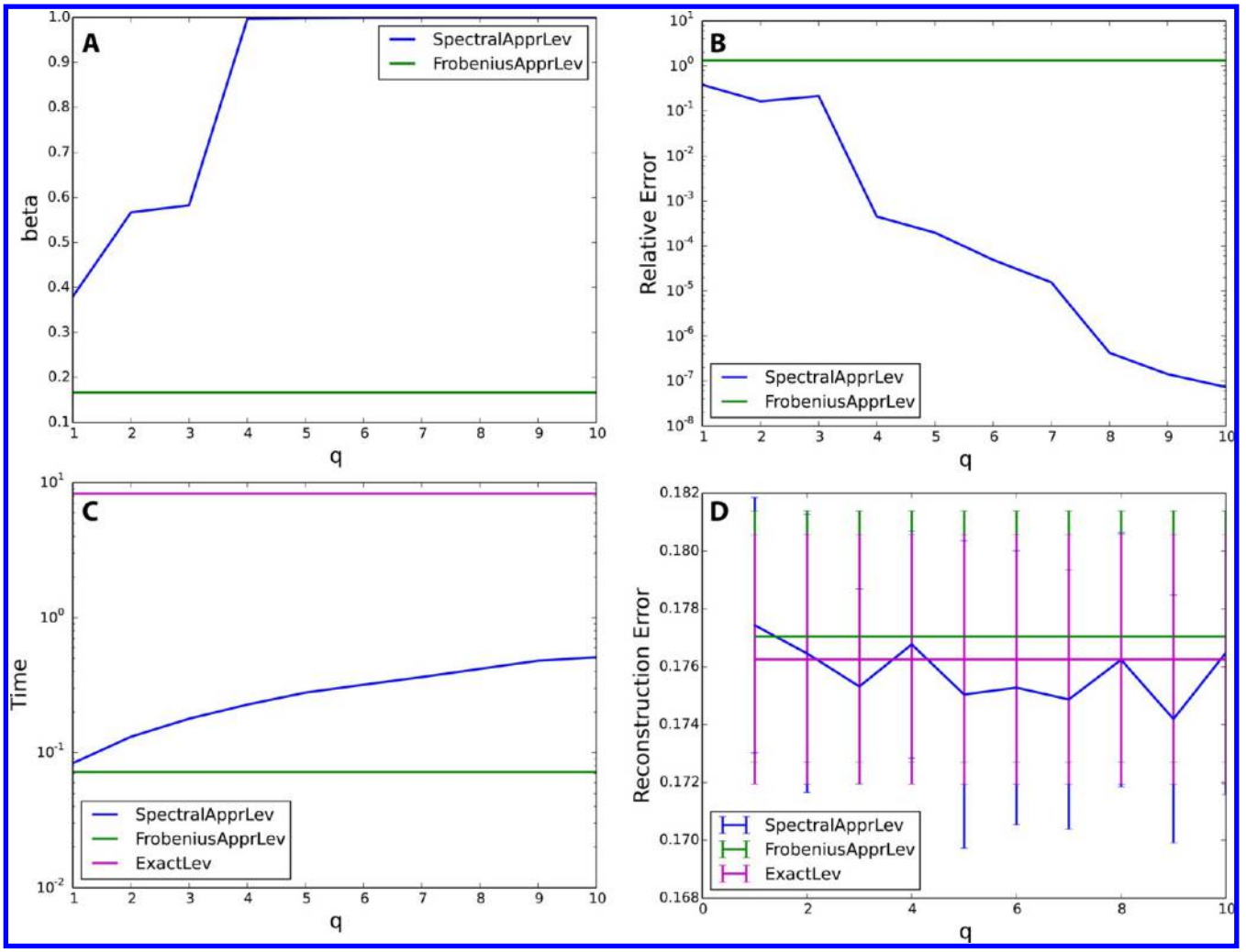


Figure 6. Quality of the normalized leverage scores using APPRLEV on the Brain data set. Both algorithm SPECTRALAPPRLEV and algorithm FROBENIUSAPPRLEV are used. Above, p_i and $\hat{p}_i$ denote the exact normalized leverage scores and the approximate normalized leverage scores, respectively; and p and $\hat{p}_i$ are vector in $\mathbb{R}^n$, the n-dimensional Euclidean space, with elements p_i and $\hat{p}_i$, respectively. Panel A shows the approximation quality of the normalized leverage scores $\beta = \min_i \{\hat{p}_i/p_i\}$. Panel B shows the L_2 distance between exact and approximate normalized leverage scores, i.e., $\|\hat{p} - p\|/\|p\|$. Panel C shows the running time, and panel D shows the reconstruction error of randomized CX decomposition.

combined in a linear model, and it is their combination that facilitates the recreation of the original data set. Used together, NMF and CUR/CX methods have the potential to be synergistic. The leverage scores computed by the CUR and CX methods can provide a measure of how informative the high intensity coefficients in the various NMF components are.

Scalability of the CX Algorithm. Here, we investigate the quality of the approximation of the leverage scores using APPLEV, by which we mean one of the two algorithms, Algorithms 4 and 5 of Drineas et al.,[17] that can be used to approximate quickly the leverage scores of the Brain data set. We call them SPECTRALAPPRLEV and FROBENIUSAPPRLEV, respectively, since the returned approximate leverage scores are a good approximation to those of a matrix that is close to A_k, when measured in spectral norm and Frobenius norm, respectively.[17]

Our evaluation is conducted in two parts. First, we evaluate these algorithms for approximating leverage scores on the Brain data set where we know the ground truth, i.e., which are small enough that we can compute the exact scores with the full SVD. Second, we apply these algorithms on the raw lung data set on which EXACTLEV cannot be performed, and we check if the outputs are still meaningful in MSI applications.

For the Brain data set, we evaluate the quality of approximation of the ion leverage scores with $k = 5$. For SPECTRALAPPRLEV, there is a parameter q that indicates the number of power iteration steps to do within the algorithm. In general, the larger q is, the more accurate the resulting approximation will be. In Figure 6A we present the value of $\beta = \min_{1<i<n}\{\hat{p}_i/p_i\}$, where p_is and $\hat{p}_i$s are the exact and the approximate normalized leverage scores. In Figure 6B, the Euclidean distance between the approximate leverage scores and the exact ones, i.e., $\|p - \hat{p}\|/\|p\|$ where $p = (p_1 \ldots p_n)$ and $\hat{p} = (\hat{p}_1 \ldots \hat{p}_n)$. In Figure 6C, we show the running time of SPECTRALAPPRLEV and FROBENIUSAPPRLEV compared to that of using EXACTLEV. Lastly, in Figure 6D, we present the corresponding CX reconstruction errors by using randomized CX decomposition. In all the figures, the mean value among 10 independent trials is reported. In Figure 6D, the standard deviation is also reported.

As we can clearly see, using both SPECTRALAPPRLEV or FROBENIUSAPPRLEV can retain a fairly high accuracy in approximating the leverage scores, while they run orders of

DOI: 10.1021/ac5040264
Anal. Chem. 2015, 87, 4658−4666

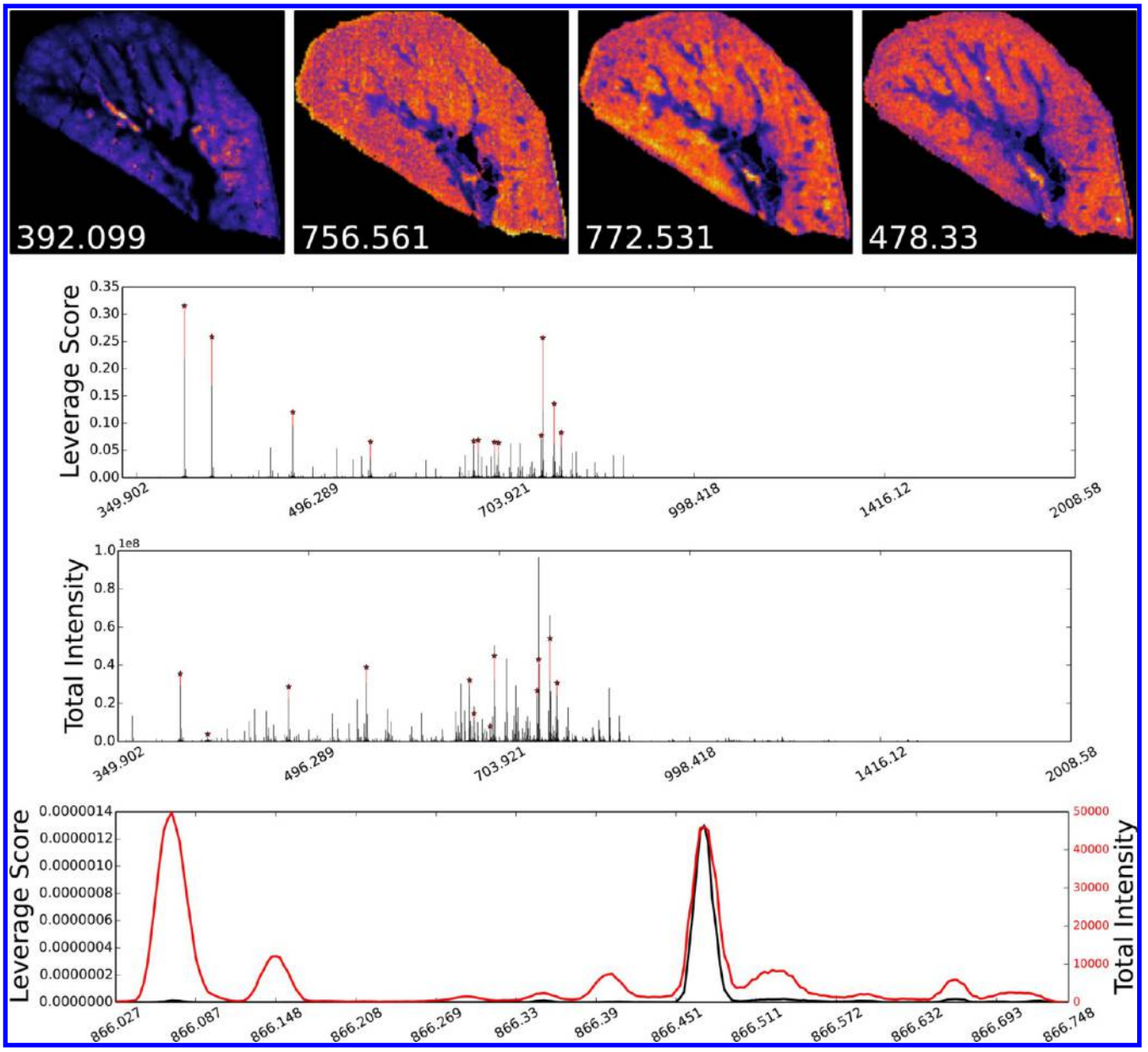


Figure 7. Quality of the normalized leverage scores using APPRLEV on the lung data set. Algorithm SPECTRALAPPRLEV with $q = 5$ is used. In panel A, we select four ions that are the most representative from the 30 most important ions returned by running deterministic CX decomposition with SPECTRALAPPRLEV. In panel B, we plot the approximate normalized leverage scores versus the m/z value. The ions with the highest leverage scores are marked by red stars. Note, for a group of ions with similar m/z values and high leverage scores, only the one with the highest leverage score is plotted. In panel C, the total sensitivities are plotted. The same ions marked in panel B are marked. In panel D, a zoom-in version of panel B,C when the m/z value is ranging from 866.02 to 866.75 is shown. The black and red curves are the leverage scores and sensitivities, respectively.

magnitude faster than the exact computation via the full SVD (and also faster, but relatively less faster, than more sophisticated computations via thin or truncated SVDs). Since leverage scores are used to identify the most influential or important ions/pixels, and since approximate leverage scores still identify these ions/pixels, little quality is lost by using the much faster approximate leverage scores.

Finally, we consider a moderately large data set on which performing the full SVD exactly will take hours to finish. In particular, we present the result on the raw lung data before peak-finding, which has a size approximately 20k by 500k. We apply SPECTRALAPPRLEV, with $k = 15$ and with $q = 5$, to compute the approximate leverage scores of the raw lung data set.

As no peak-finding was done on the raw data set, some ions with high leverage scores have similar m/z values. In Figure 7A, we present the spatial distributions of the four most representative ions selected from different groups. In Figure 6B,C, the approximate leverage scores and the total sensitivities versus the m/z values are plotted, respectively. In addition, a "zoom-in" version of the above two plots, overlaid on each other, on ions with m/z values in the range between 866.02 and 866.75 is shown in Figure 7D.

Since the exact leverage scores are unavailable, we are not able to evaluate the accuracy of the approximation of the leverage scores, but the convergence results from Figure 6 suggest these scores are reliable. In addition, the results suggest that the ion at $m/z = 392$ (a drug administered to the tissue)

DOI: 10.1021/ac5040264
Anal. Chem. 2015, 87, 4658−4666

was identified as the highest leverage ion, and ions specific to regions of the lung were also identified. That the administered drug was identified as the highest importance ion could be significant for pharmacokinetics/pharmacology and could also be a marker to accelerate identification of degradation products or byproducts that are of unexpected/unpredetermined m/z values.

What is most significant in this approach is the lack of reliance on peak-finding. By applying scalable factorization approaches like CX and CUR to raw, profile spectra, a multitude of previously ignored features can be considered. As can be seen in Figure 7D, the zoomed in portion of the leverage score overlaid with the total intensity spectra shows a large number of recognizable features with high intensity. Strikingly, only one of these features has a high leverage score. This prioritization allows accelerated interpretation of results by pointing a researcher toward which ions might be most informative in a mathematically more objective manner.

CONCLUSION

In this work, we have introduced CX and CUR factorizations as a new concept to mass spectrometry imaging. We have also demonstrated that using this approach can lead to prioritization of specific ions and locations. The algorithms described here give a step-by-step method for these factorization methods to be applied as an alternative strategy to the PCA, NMF, and related clustering-based approaches that are currently widely used. By using CX factorizations, the empirical statistical leverage scores are used to represent the measured data in terms of a smaller number of actual ions and actual locations. This leads to an easier to interpret low-rank approximation of the original data than PCA-based methods that construct eigenions and eigenpositions. In addition, we have shown here the specific ranking methods for identifying important ions differs from that of selecting important pixels. By considering the distribution of leverage scores a probability distribution, a random-sampling algorithm can yield the best selection of important locations. In comparison, ions can be selected greedily by taking those with the highest leverage scores. This difference is due to the uniformity of the leverage score for locations, i.e., many pixels can represent similar information content and thus no particular pixels have particularly large leverage. In the case of ions, the leverage scores are much more nonuniform, and thus a small number of ions gives very unique images. Lastly, because MSI is generating ever larger and more complex data sets, we use a scalable implementation of this algorithm that is suitable for more large-scale data sets.

ASSOCIATED CONTENT

Supporting Information

Binary mask images for the work presented. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bpbowen@lbl.gov.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Partial support for this work was provided by the Defense Advanced Research Projects Agency.

REFERENCES

(1) Caprioli, R. M.; Farmer, T. B.; Gile, J. *Anal. Chem.* **1997**, *69*, 4751−4760.
(2) McDonnell, L. A.; Heeren, R. M. A. *Mass Spectrom Rev.* **2007**, *26*, 606−643.
(3) Chughtai, K.; Heeren, R. M. A. *Chem. Rev.* **2010**, *110*, 3237−3277.
(4) Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer Science & Business Media: New York, 2002.
(5) Lee, D. D.; Seung, H. S. *Nature* **1999**, *401*, 788−791.
(6) Jones, E. A.; Deininger, S.-O.; Hogendoorn, P. C. W.; Deelder, A. M.; McDonnell, L. A. *J. Proteomics* **2012**, *75*, 4962−4989.
(7) Alexandrov, T. *BMC Bioinf.* **2012**, *13* (Suppl 16), S11.
(8) Reindl, W.; Bowen, B. P.; Balamotis, M. A.; Green, J. E.; Northen, T. R. *Integr. Biol.* **2011**, *3*, 460−467.
(9) Klinkert, I.; McDonnell, L. A.; Luxembourg, S. L.; Altelaar, A. F. M.; Amstalden, E. R.; Piersma, S. R.; Heeren, R. M. A. *Rev. Sci. Instrum.* **2007**, *78*, 053716.
(10) Lee, D. Y.; Platt, V.; Bowen, B.; Louie, K.; Canaria, C. A.; McMurray, C. T.; Northen, T. *Integr. Biol.* **2012**, *4*, 693.
(11) Jones, E. A.; van Remoortere, A.; van Zeijl, R. J. M.; Hogendoorn, P. C. W.; Bovée, J. V. M. G.; Deelder, A. M.; McDonnell, L. A. *PLoS One* **2011**, *6*, e24913.
(12) Mahoney, M. W.; Drineas, P. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 697−702.
(13) Drineas, P.; Mahoney, M. W.; Muthukrishnan, S. *Siam J. Matrix Anal. Appl.* **2008**, *30*, 844−881.
(14) Paschou, P.; Ziv, E.; Burchard, E. G.; Choudhry, S.; Rodriguez-Cintron, W.; Mahoney, M. W.; Drineas, P. *PLoS Genet.* **2007**, *3*, 1672−1686.
(15) Yip, C.-W.; Mahoney, M. W.; Szalay, A. S.; Csabai, I.; Budavari, T.; Wyse, R. F. G.; Dobos, L. *Astron. J.* **2014**, *147*, 110.
(16) Moore, E. H. *Bull. Am. Math. Soc.* **1920**, *26*, 385−397.
(17) Drineas, P.; Magdon-Ismail, M.; Mahoney, M. W.; Woodruff, D. P. *J. Mach. Learn. Res.* **2012**, *13*, 3475−3506.
(18) Bodor, A.; Csabai, I.; Mahoney, M. W.; Solymosi, N. *BMC Bioinf.* **2012**, *13*, 103.
(19) Louie, K. B.; Bowen, B. P.; Cheng, X.; Berleman, J. E.; Chakraborty, R.; Deutschbauer, A.; Arkin, A.; Northen, T. R. *Anal. Chem.* **2013**, *85*, 10856−10862.
(20) Rübel, O.; Greiner, A.; Cholia, S.; Louie, K.; Bethel, E. W.; Northen, T. R.; Bowen, B. P. *Anal. Chem.* **2013**, *85*, 10354−10361.
(21) Balamotis, M. A.; Tamberg, N.; Woo, Y. J.; Li, J.; Davy, B.; Kohwi-Shigematsu, T.; Kohwi, Y. *Mol. Cell. Biol.* **2012**, *32*, 333−347.
(22) Marko-Varga, G.; Fehniger, T. E.; Rezeli, M.; Döme, B.; Laurell, T.; Végvári, Á. *J. Proteomics* **2011**, *74*, 982−992.

DOI: 10.1021/ac5040264
Anal. Chem. 2015, 87, 4658−4666