Automated Correlation and Classification of Secondary Ion Mass Spectrometry Images Using a *k*-Means Cluster Method

Andrew R. Konicek, Jonathan Lefman[†], and Christopher Szakal[*]

Surface and Microanalysis Science Division, National Institute of Standards and Technology,
100 Bureau Drive, Gaithersburg, MD 20899-8371, USA
cszakal@nist.gov
Ph: (301) 975-3816
Fax: (301) 417-1321

[*]Corresponding Author

[†]Present address: Engineer Research and Development Center, US Army Corps of Engineers
7701 Telegraph Rd.
Alexandria, VA 22315

Table of Contents for Supporting Information

Additional Algorithm Details

Input SIMS Images

SIMS images were prepared for the algorithm as described in the manuscript text, and imported into MATLAB along with a peak list file containing the mass identification for each of the *n* input SIMS images. The user has the option of applying a mask to select a subset of pixels that correspond to a region of interest. Benefits of the masking process include 1) a reduction in calculation time by 1 to 2 orders of magnitude because of a decrease in the number of pixels subject to analysis, and 2) exclusion of image regions that are not targeted for analysis, such as from a substrate or contamination. Some multivariate analysis (MVA) approaches may be dominated by these features without a masking step if the signal contributions are large relative to the sample region signals.

Algorithm Outputs

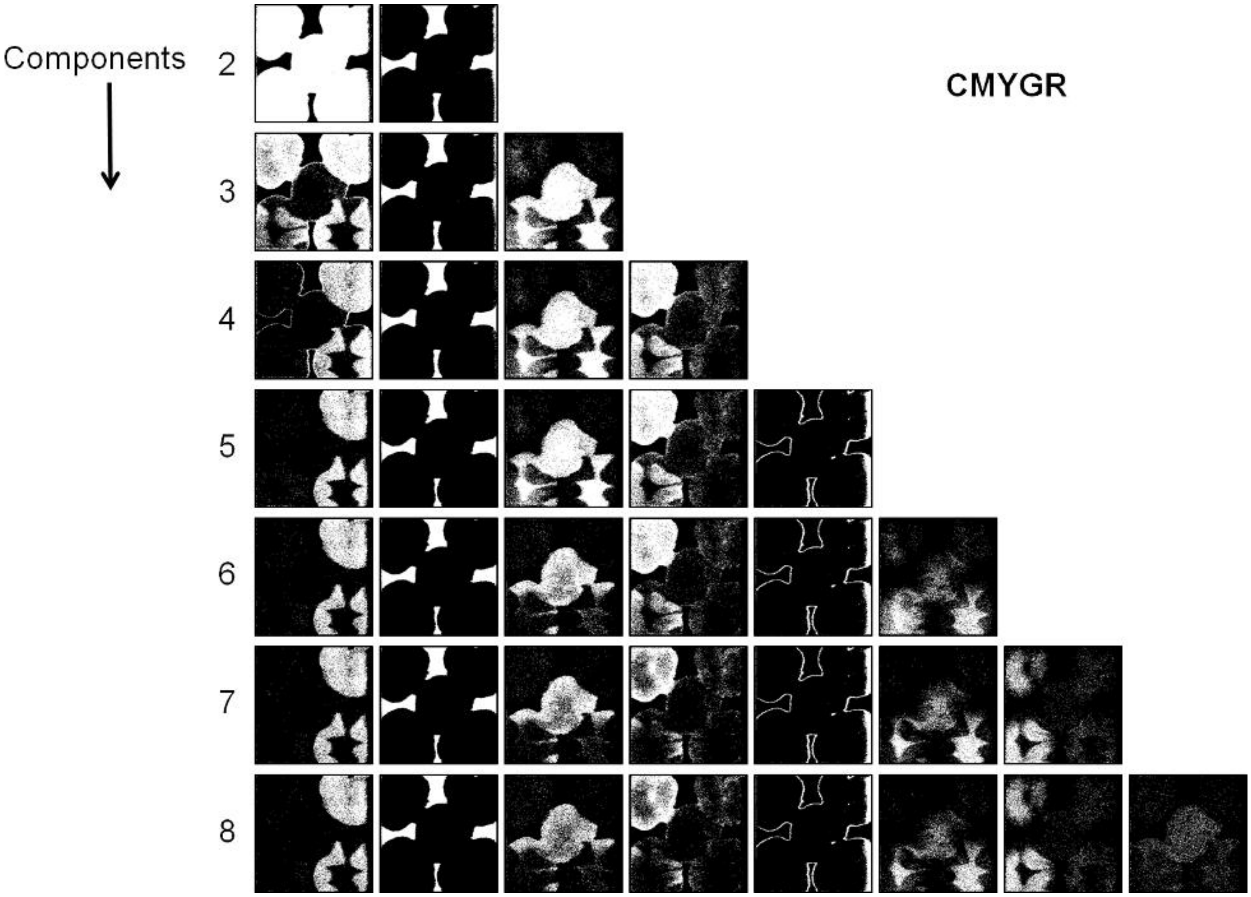
The algorithm produces a series of SIMS image and centroid image montages, with all SIMS images sorted by subclass and by correlation value. A spreadsheet file is generated containing the SIMS image number, SIMS image mass, class number, subclass label, and a list of correlation values to every centroid image created in that iteration. Mass spectra can be created for each class or subclass if desired.

Principal Component Analysis Discussion

In principal component analysis (PCA), the principal components (PCs) are calculated such that PC1 (principal component 1) describes the maximum variance of the data, PC2 the next highest variance, and so forth. A complication of such an approach is that the orthogonality of the PCs results in component spectra and images with features that do not physically correspond to the natural basis of the data, and are therefore not easily interpretable on their own.

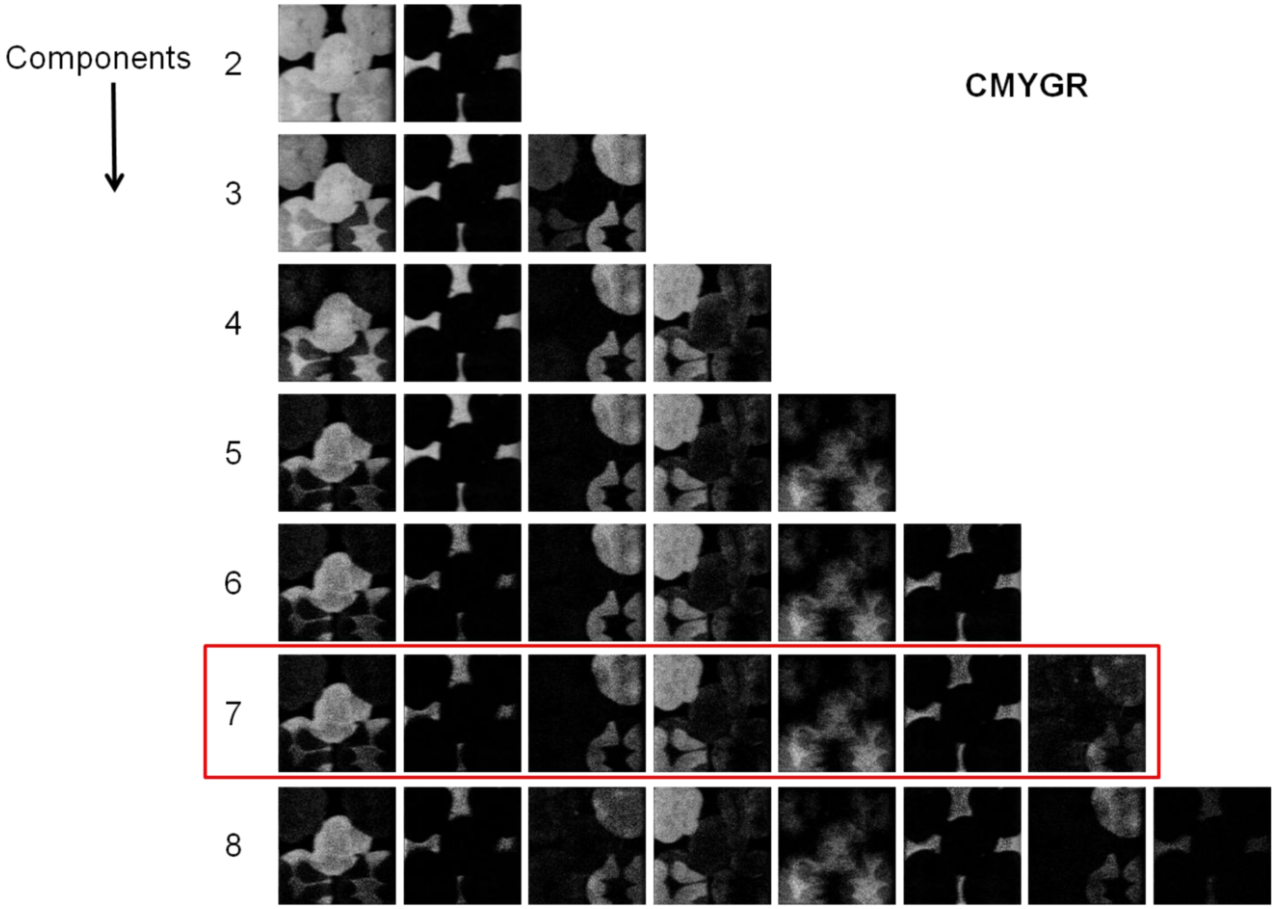
The CMYGR data were also analyzed by PCA using AXSIA, with results shown in Fig. S2. The “spatial simplicity” option for subsequent rotation of the data was chosen because it performs an analysis that is most comparable to the newly described *k*-means clustering algorithm and is the most appropriate PCA-based approach for identifying localized pixels in multispectral SIMS images. The procedure used normalized SIMS images, as is typical for PCA of ToF-SIMS data, but, for comparison, the spatial simplicity approach was performed for un-normalized SIMS images as well, with results provided in Fig. S3. The subsets of PCs retained by the analysis are rotated to be positive-definite (no spectral intensities less than zero) while maximizing the spatial clustering of pixels within the corresponding SIMS images. The analysis was performed by 1) keeping the number of output components fixed, with values ranging from two to eight, in order to mimic the first seven iterations of the *k*-means-based algorithm, and 2) allowing AXSIA to automatically determine the number of significant components, similar to how most PCA-based analyses are employed for SIMS data.

Fig. S1



Since PCA used for MVA of time-of-flight secondary ion mass spectrometry (ToF-SIMS) data is typically pre-normalized to avoid the influence of Poisson noise statistics, a test was conducted to determine the influence of SIMS image normalization on the *k*-means algorithm results. The above plot of the iteratively generated *k*-means centroid images for the CMYGR test sample was created after the SIMS images were normalized. The normalization method was the same method that is typically employed for Poisson noise correction prior to performing PCA. Every SIMS image in the data cube was normalized by the mean SIMS image, and then subsequently every spectrum (at every pixel) was normalized by the mean spectrum. The centroid image shapes are mostly similar to those shown in Fig. 4 of the associated manuscript. One noticeable difference is that the artifact class is not created until the five-centroid iteration, while the four-centroid iteration has already separated the three ink classes and the paper class. Also, there are no singleton centroid image classes created (like the iterations with six and eight centroids in Fig. 4), where instead each new centroid iteration in Fig. S1 seems to better define new spatial features that are common to groups of SIMS images. However, the total numbers of SIMS images correlated with each centroid image were quite similar to those for the un-normalized SIMS images, and any improvement in definition is most likely due to the noise leveling provided by the normalization. These results suggest that there are few, if any, benefits to pre-processing or normalizing the SIMS images prior to a *k*-means clustering analysis as the correlation and classification of the SIMS images to the centroid images produce nearly identical chemical information.

Fig. S2

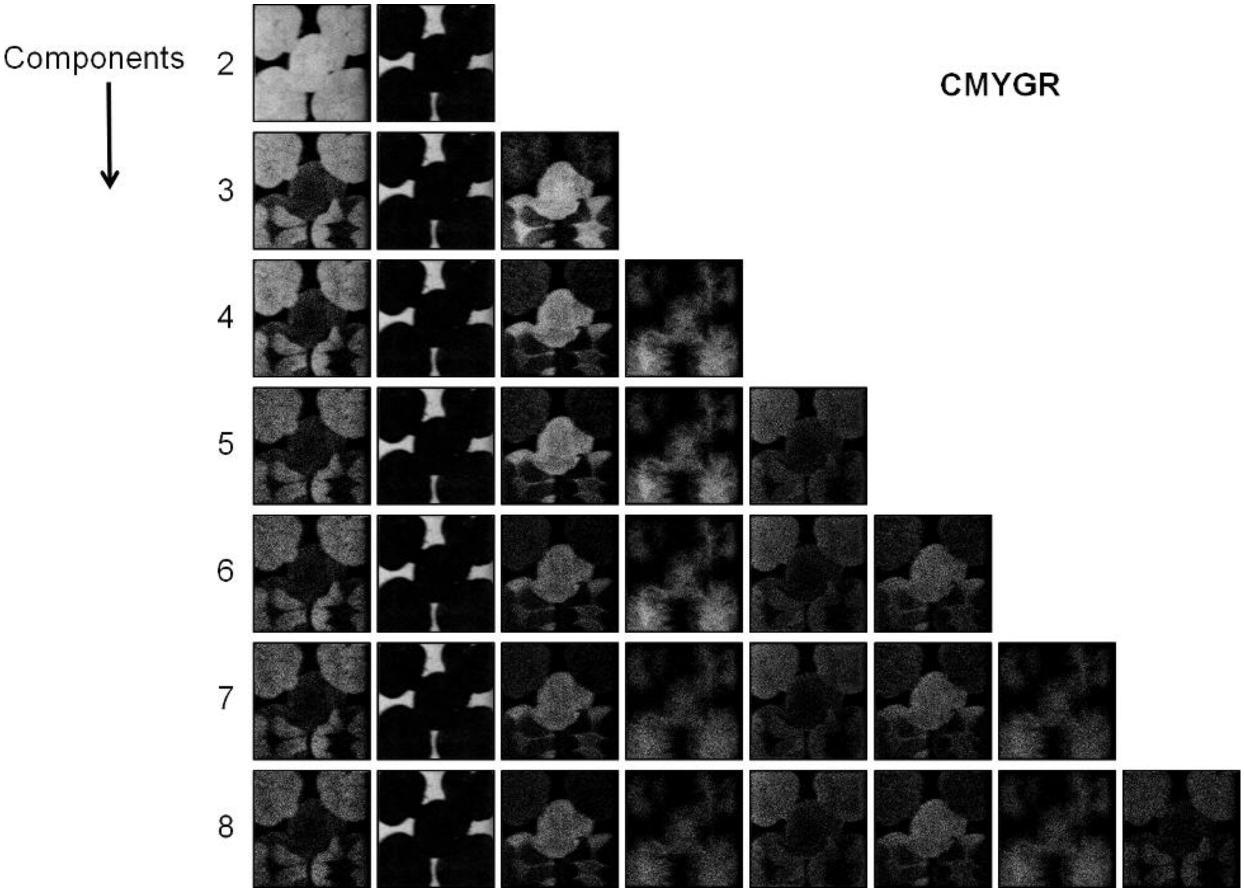


Above are plots of the scores images from a PCA with spatial simplicity approach in order to compare with the *k*-means results in Fig. 4. The component rows were generated by fixing the number of PCs used in each row from two (2) to eight (8) for the CMYGR test sample (shown in Fig. 1a,d,g). The row with seven PCs (highlighted in red) was the set determined automatically by AXSIA as the optimal number of components. Comparing Fig. S2 with Fig. 4, both methods seem to identify differences that could be assigned to potential topographic effects, as illustrated by the split of the yellow ink images into two complementary shapes. The split is observed in the seven-centroid iteration in Fig. 4 and the five component output from Fig. S2, representing one for the yellow ink by itself and one for yellow ink that has been printed on top of another ink. The most noticeable difference between the two outputs is that the PCA results have differing gray levels in each scores image, whereas the *k*-means centroid images are binary. This is because the PCA scores images are the weights by which the loadings spectra are multiplied to attempt to reconstruct the original spectra. [As a reminder from the manuscript text, all of the scores images in Fig. S2 (and their corresponding loading vectors) have been rotated in vector space to eliminate negative values]. PCA does not identify the artifact class directly, but every score image has zero or very low intensity in the pixels where the artifact occurs. This is in contrast to *k*-means, which creates a separate centroid image that identifies the artifact class directly. It makes sense that they are not identified with PCA, as they would not contribute largely to the variance.

There is generally good shape agreement between the PCA output (Fig. S2) and the *k*-means output (Fig. 4), illustrating the ability of the new method to provide similar results to PCA with spatial simplicity. However, PCA has automatically identified seven PCs, which are believed to represent an overclassification of the data. Knowing the number of inks used, and assuming a component for the paper, one might have assumed there would be only four components and restricted the PCA output as such. If additional PCs are added to the spatial simplicity input, the meaning of the output scores images and loading spectra can be compromised.

Changing the number of included PCs affects the results from the spatial simplicity rotation, since the rotation only operates on the number of components selected from PCA. This is why PCs in Fig. S2 change depending on the number of components. By only relying on the static output (*i.e.*, the output generated by the automatic determination of the number of significant components), the user is left to believe in MVA components when they may just reflect the most pronounced features of the data set, and could overlook relevant information.

Fig. S3



Even though the typical PCA analysis of ToF-SIMS data involves a pre-normalization step, the above plot displays scores images from a PCA with spatial simplicity approach for the CMYGR test sample *without* first normalizing the data in order to eliminate the influences of Poisson noise statistics. The displayed image sets were determined by fixing the number of PCs used in each step from two (2) to eight (8). As illustrated in Fig. S3, the PCA output does not fully characterize the data set for any number of PCs when the SIMS images are not normalized. Here, the effect of not normalizing the data creates obvious differences from the pre-normalized PCA-based output shown in Fig. S2. While cyan and magenta areas are separated from the yellow area in the three component output, there is never any subsequent splitting of the cyan and magenta areas into distinct components. The first component remains for all subsequent outputs, and is actually further split into three similar score images by the eight-component set. The paper score image remains constant over all sets, while the yellow score image from the three component-set is eventually broken into four components by the eight-component set. These results emphasize that the PCA with spatial simplicity approach requires pre-normalization of SIMS image data to meaningfully characterize key components. Also of note is that the automated PCA output produced 12 PCs for this un-normalized set of SIMS images, which would incorrectly inform the user of the true dimensionality of the SIMS image data. The comparisons between Fig. S1, Fig. S2, Fig. S3, and Fig. 4 illustrate that the mathematical differences between PCA with spatial simplicity and the new *k*-means clustering algorithm are important to highlight. Poisson noise dramatically alters the PCA with spatial simplicity results in such a way that the SIMS images are not fully identified with respect to our *a priori* knowledge of the test sample. However, such noise does not appear to alter the ability of *k*-means to classify the SIMS images based on individual components and any artifacts. We completely understand that such a comparison is contrived since it is generally accepted that PCA requires data normalization prior to analysis. However, we believe that the smallest possible amount of data manipulation to achieve accurate analysis is a worthy goal, and the newly created *k*-means clustering algorithm for mass spectral image correlation and classification does not appear to require any SIMS image pre-processing, even though there is an arguable improvement to the classification process with normalization.