A Brief Introduction to Multivariate Image Analysis (MIA)

Barry M. Wise* and Paul Geladi†
*Eigenvector Research, Inc.
†Umeå University

Introduction

Images have been used in the sciences for a long time and they are used increasingly. Large amounts of data representing complex systems can only be represented by visualization as images. Multivariate images arise from a surprising variety of sources. Some are images in the conventional sense (such as satellite data) while others are not (secondary ion mass spectroscopy, SIMS). Almost all physical units can be used to make images and multivariate images: temperature, gravitational field, impedance, magnetic field, electrical field, mass, wavelength, ultrasound wavelength, polarization, electron energy etc. A rough but practical subdivision of the fields of scientific imaging is in satellite imaging, medical (clinical) imaging and the microscopies. The simplest meaningful multivariate image has two pixel indices (*e.g.* width and height in the image plane) and a variable index, making up a three-way array. An important aspect in going from analog scenes or objects to digital images is resolution. Multivariate images have spatial, intensity, spectral and time (temporal) resolution. A typical older satellite image would have 512x512 pixels, in 7 wavelength bands and an intensity resolution of 256 gray levels. High spatial and intensity resolution is desirable and this makes the arrays rather large and the calculations slow.

The traditional field of univariate image analysis works in the spatial domain in 2D or 3D image arrays. When images become multivariate or multitemporal, the spectral or time domain become a higher priority than spatial considerations. When this is the case, the tools of Multivariate Image Analysis (MIA) become very useful.

Theory

Principal Component Analysis (PCA) is the workhorse of MIA. The key is the proper reorganization (matricizing) of the original 3-way or higher array. Unfolding is done so that each pixel (or voxel) becomes a single row in the analysis. Thus an image that is originally I by J pixels with K spectral channels is reshaped to form a two way array that is IxJ by K. PCA can then be performed on this matrix in the usual way. Mean centering is typically done and in some circumstances variance scaling may be used. After the PCA model is calculated, the scores, residuals and T^2 values can be folded back up to reform images. Loadings vectors can be interpreted in the usual way.

Computational Issues

Computational issues arise due to the shear size of the images. It is not uncommon for images to be 512 by 512 by 20. Stored as double precision arrays, such an image would take up 40 Megs of RAM. Often images are stored as unsigned 8 bit integers (*i.e.* 0 to 255), which reduces the storage space to 5 Megs for the array above. However, most math libraries and MATLAB do not define most mathematical operations (such as addition and multiplication!) for 8 bit integers. This often means that the data must be converted piecemeal to double precision as it is needed. Generally, algorithms must be written to take advantage of the smallest dimension in the problem, which is usually the spectral dimension (K). Scatter matrices can be constructed by converting the spectral channels to doubles pairwise. Covariance matrices are calculated from the scatter matrices given the

means of the spectral variables. Decompositions can then be performed on these matrices, recovering the spectral loadings.

Once the spectral loadings are in hand, scores for each pixel can be determined. It is often convenient to scale the scores back into the range of 0 to 255 and convert them back to unsigned 8 bit integers. This saves storage space and doesn't lose much information if the scores are to be observed as images. As will be discussed below, it also makes working in the score space easier.

Working in the Image Plane

Once scores have been calculated for each pixel, they can be folded back up to the original image dimensions (I by J) and displayed as pseudo color maps. This gives a graphic representation of the score value of each pixel as a function of position. Examination of the corresponding loadings gives information as to the original spectral variables which give rise to the variations captured in the scores plots. Residuals can also be displayed this way, as can Hotelling's T^2 and many other diagnostics.

Displays of any three of the scores, residuals, or T^2 values can be accomplished in the image plane by assigning each to be the red, green and blue values in the display of the image. If this is done for the first 3 PCs it is arguable that this is the most information about the data that can be displayed in a single image (a least to the non-colorblind). Areas with different attributes show up as different colors in the pseudo color image, often offering stark contrast to features that do not show up in any individual image or even any particular score image. Display of the residuals can be particularly useful with regard to identifying unique areas (such as minor amounts of contamination). The display and visual study of color images is informative, but it is also subjective. The human eye is not very linear in interpreting color differences and misjudgment of similarities or differences is a risk.

As an example, we'll consider data from SIMS. The sample surface is PMMA (polymethylmethacrylate)which has been exposed to deuterated polystyrene. The data consists of the SIMS spectra from 1 to 300 amu on a 64 by 64 grid. Thus, the image is unfolded to 4096 (64x64) by 300 and PCA is performed. A score image for the first PC is shown in Figure 1. The islands of deuterated polystyrene (low scores) are clearly visible against the background of PMMA (high scores). Inspection of the next two PCs (not shown) show little systematic variation across the image. The residual image, Figure 2, shows two small areas with large residuals. These areas are not representative of the remainder of the surface and may be minor amounts of contamination.

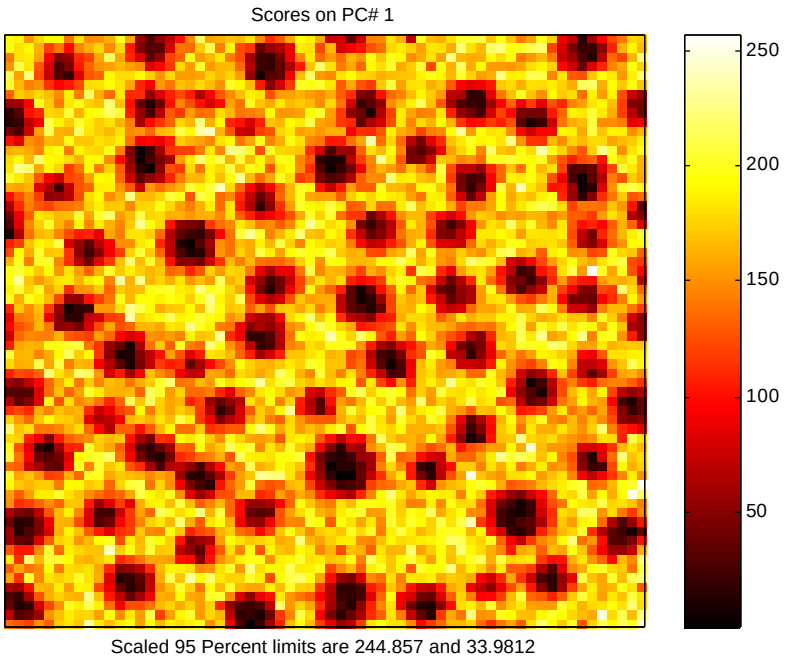

Figure 1. Image Scores on First PC

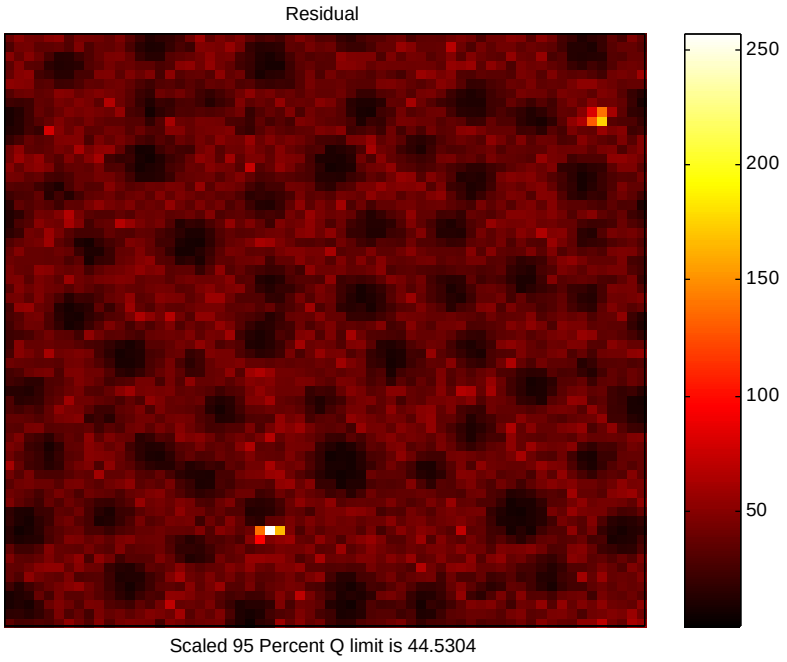

Figure 2. Residual Image of PS on PMMA Based on Two PC Model.

Working in Score Space

While working in the image plane can provide useful clues as to the nature of the data set at hand, many features of the data, particularly clustering of the pixels, show up only in score space. Here the scores for all the pixels on two of the loadings are plotted against each other. However, because of the very large number of points typically involved (512x512=262,144 points), simple scatter plots of the data often produce large "blobs" in the plots that lose information regarding the densities of points in a given area. For that reason, it is convenient to color code the plots in a way that gives the density of points in a given region. When the scores are stored as unsigned 8 bits this is particularly easy, one needs just find the number of pixels with scores in every element of a 256 by 256 array and display this as a pseudo color image. This image always shows outliers, dense clusters, sparse clusters and gradients between the clusters. The cluster size can be interpreted as standard deviation within the cluster. The clusters rarely have normally distributed ellipsoidal shapes, so the best way to delineate a cluster is by drawing a polygon around it.

The scores plots for our example problem are shown in Figure 3. The PC#2 versus PC#1 data is in the lower right quadrant, the PC#2 versus PC#3 is in the lower left, and PC#3 versus PC#1 in the upper right. A pseudo color image is in the upper left quadrant. The data appears to split into two groups in the lower right, a more concentrated group on the right and a more diffuse group on the left. A polygon has been drawn around the group on the left and the points within it have been highlighted in green. These points are also highlighted in the other score plots and on the original image. The ability to connect points in scores space with points on the image plane is critical to MIA.

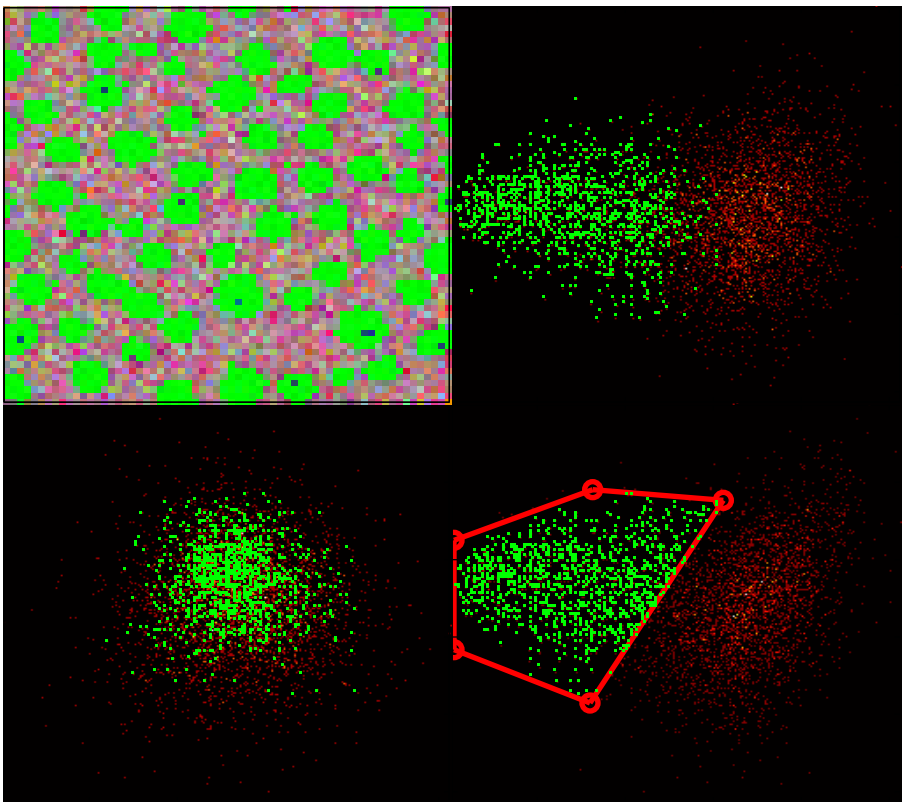

Figure 3. Working in Score Space Linked to Image Plane

Local modeling

Because of the size of the images, major expected constituents may swamp smaller more interesting ones, making them show up in higher and noisier components. It is therefore very useful to make local models for smaller parts of the images. The selection of these parts can be as rectangular subsets, but even better as subsets of an irregular shape selected in the score plots. The subset doesn't even have to be contiguous (see Geladi, 1995). In the future, when both high spatial and spectral resolution will be available, it may be necessary to develop local modeling both in space and in the spectral domain. In this case, each spatial subset may also benefit from the use of a specific spectral subset. The different methods of chemometrics will be useful both for selecting the spatial and spectral subsets and for analyzing the obtained subsets.

Preprocessing

Not all images are ideal when they are digitized. They may be noisy, have missing pixels, have bad contrast etc. Fortunately there are the operations of univariate image analysis to correct for most of these unwanted properties. Noise removal, hole filling, and contrast improvement can all be done in an interactive manner by visual inspection. A special problem is non linearity. Many imaging techniques may produce a large range of intensity values. This range should be projected to the linear range of 0-255 for many applications by a combination of logarithmic transformation and rounding or truncation to the nearest integer. Another problem is that some imaging techniques give intensities in the reflectance mode and these may have to be transformed to absorbances and rescaled in order to better represent the underlying phenomena. A high intensity resolution to begin with is often crucial in avoiding rounding/truncation errors.

The use of preprocessing is especially important in satellite imaging. The satellite images that are made available have undergone quite some preprocessing for removing scanning errors. Additional preprocessing methods include texture filters, wavelets, Fourier analysis and angle measurement technique (AMT).

Classification

Once a basic MIA has been performed, it is easy to develop classification models which can be applied to the current image or future images. In the score space pixels which cluster can be selected for development of a separate PCA model (as in the SIMCA technique), or as input into other classification methods. Pixels can also be selected from the image plane. Once the classification models are developed areas on new images (or remaining parts of the calibration image) can be classified.

Regression

Images may be regressed against each other. An I by J by K image may be used as predictor (X) block for an I by J by M image (response variables, Y block). This is rare. More often, a smaller image or an image of lower spatial resolution is used as the Y-image. Even images of irregular shapes may be used as Y-images in regression. The regression model parameters allow a prediction image of the size I by J. The use of latent variable based calibration methods also leads to a residual of the X-block. This residual is used as a measure of the reliability of the predictions. In satellite images, regression is used for "ground truth" prediction of vegetation quality or quantity. In medical applications, discriminant analysis of known malignant and benign tissue may be formulated as a regression on a binary variable. In this way, an automated detector of malign tissue may be constructed that is useful for future images. An advantage of latent variable regression

methods is that the latent variables are images and can be studied as images or as scatter plots.

Other Extensions

The earliest uses of multivariate image analysis made use of the most obvious multivariate methods: PCA, ridge regression, principal components regression (PCR) and partial least squares regression (PLS). This was often dictated by limitations in storage space, memory and calculation capacity. Also the limited availability of wavelength bands (or other variables) often dictated a simple form of data analysis, but nothing prevents the use of more advanced methods, or hybrids. Curve resolution may be a good alternative to PCA if linear mixtures are studied. Also alternative supervised and unsupervised, interactive or automatic clustering methods are possible. For temporal imaging, the use of the principles of time series modeling is very promising. Positive matrix factorization, maximum likelihood models, parallel factor analysis (PARAFAC) on series of multivariate images etc. When both temporal and spectral data are available, the image array may be reorganized into a three-way array with pixel, time and wavelength as the ways. The pixel mode three-way loadings may be reorganized into images.

Hyperspectral Imaging

New technologies are emerging that make it very easy to get high spectral resolution. In satellite imaging, systems like AVIRIS (Airborne visual/infrared imaging spectrometer) with 224 wavelength bands are replacing the <20 band older satellite images. In optical clinical imaging, systems that combine the resolution of a single point infrared, fluorescence or Raman measurement with a moderate spatial resolution are becoming available. In electron and ion microscopy too, high energy or mass resolution are becoming easy and fast to obtain. The emergence of these systems, with a sometimes low (64x64) spatial resolution and a rather high (>200) spectral resolution is generating new ways of treating the data. More emphasis is going to the spectral aspect. With some techniques (Raman, infrared, fluorescence), it is easy to do wavelength selection and fall back to univariate imaging in one wavelength or an integral over a small wavelength band. Also band ratios are often made to good use.

Examples

The MIA paradigm will be demonstrated using two additional examples. The first is from SIMS of a PVA sample. The original data is 64 by 64 pixels with 300 mass channels. Only the positive SIMS spectra was considered. Figure 4 shows a number of false color images of the data at several different mass numbers. Some of the images appear to have some grouping of pixels with large numbers of ion counts on the image plane. The picture becomes much clearer when PCA is applied, as shown in Figure 5. Here a spot is clearly visible on the image surface. Inspection of the loadings (not shown) reveals some mass numbers which load positively, and others that load negatively. Thus, the spot is depleted in the ion fragments that load positively and enriched in those that load negatively compared to the remainder of the image. We would expect this to be meaningful to a competent mass spectroscopist . (Unfortunately, we’re not mass spectroscopists and thus won’t attempt to interpret this image chemically.)

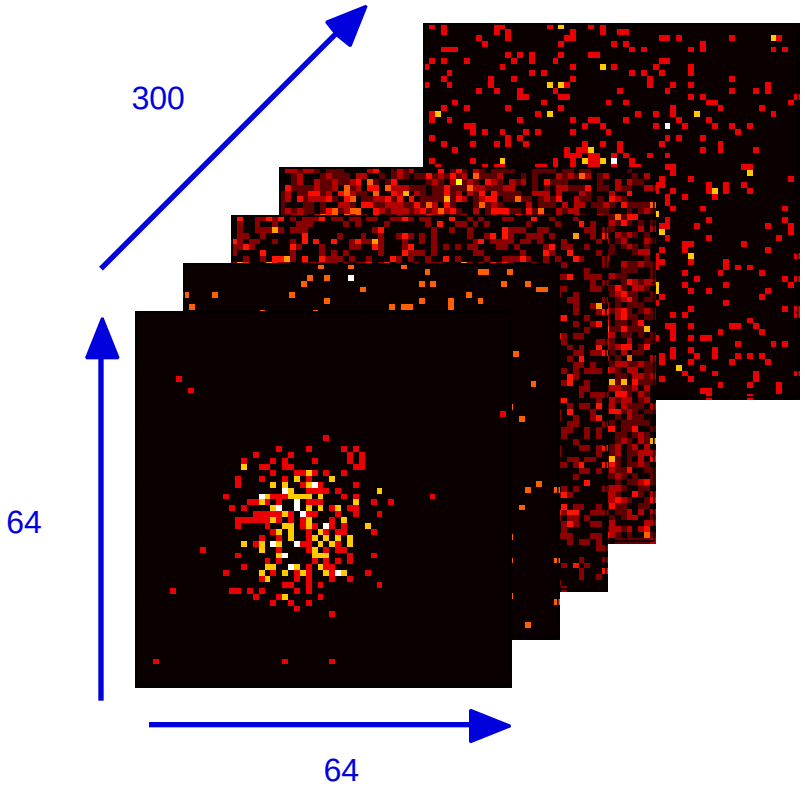

Figure 4. Slices of the SIMS Data of a PVA Sample.

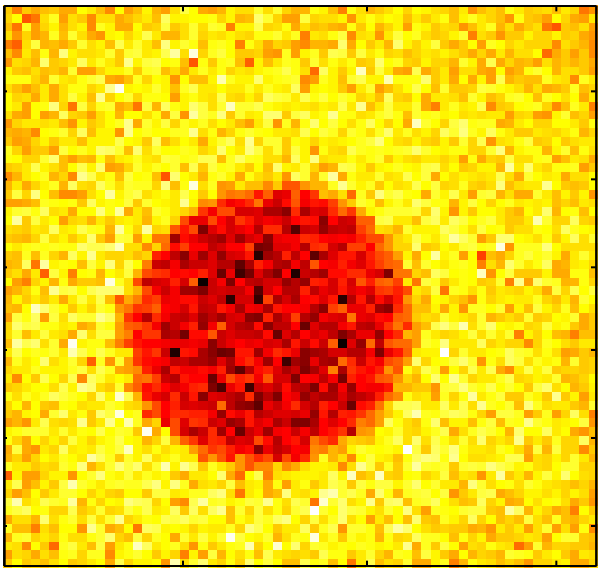

Figure 5. Scores on First PC of PVA Sample Showing Spot Clearly.

In our second example, PARAFAC will be used to analyze a series of multivariate images. PARAFAC cannot generally be used in MIA as most (single) images are not tri-linear. This is because the image plane is not easily decomposed as a summation over the outer product of pairs of factors. However, if the data consist of a sequence of multivariate images, the data can be unfolded in a way that should be approximately tri-linear. As an example we will consider a series of 5 images taken by the Eigenvector Research web camera on March 9, 2000. The first of the images is shown in Figure 6. The images are each 240 by 352 by 3 (the red, green and blue layers). Thus the total array is 240 by 352 by 3 by 5. This array can be unfolded to 84,480 (240x352) by 3 by 5 and analyzed with PARAFAC. The PARAFAC loadings in the time and color modes will describe the general changes of the image over the series while the loadings in the pixel dimension describe changes in specific pixels on the images.

Figure 6. Image 1 from Eigenvector Research Web Camera.

PARAFAC models with 1, 2 and 3 factors were developed. The two factor model, capturing 99.6% of the sum of squares, was selected. The PARAFAC model can be interrogated to determine the "trends" in the series of images. The loadings from the color and time modes are shown in Figure 7. The color dimension shows one factor that is high in blue relative to red and green. This factor is associated with an almost constant factor in the time domain. In the pixel mode, Figure 8, this factor separates the foreground, which is mostly grass, from the mountains and sky, which appear blue. The second factor is highest in red and green in the color domain and is associated with a generally decreasing time factor. This factor separates different types of vegetation from each other, such as the bushes from the lawn, as shown in Figure 9. This is most apparent in the plot of pixel mode loadings, which we will refer to here as scores, shown in Figure 10. Here the different clusters of the scores are associated with different elements of the images. Mountains, sky, front lawn, mowed field, unmowed field, shrubs, rocks, orchard, and man-made structures (road, fence and some buildings) all have their own cluster as indicated.

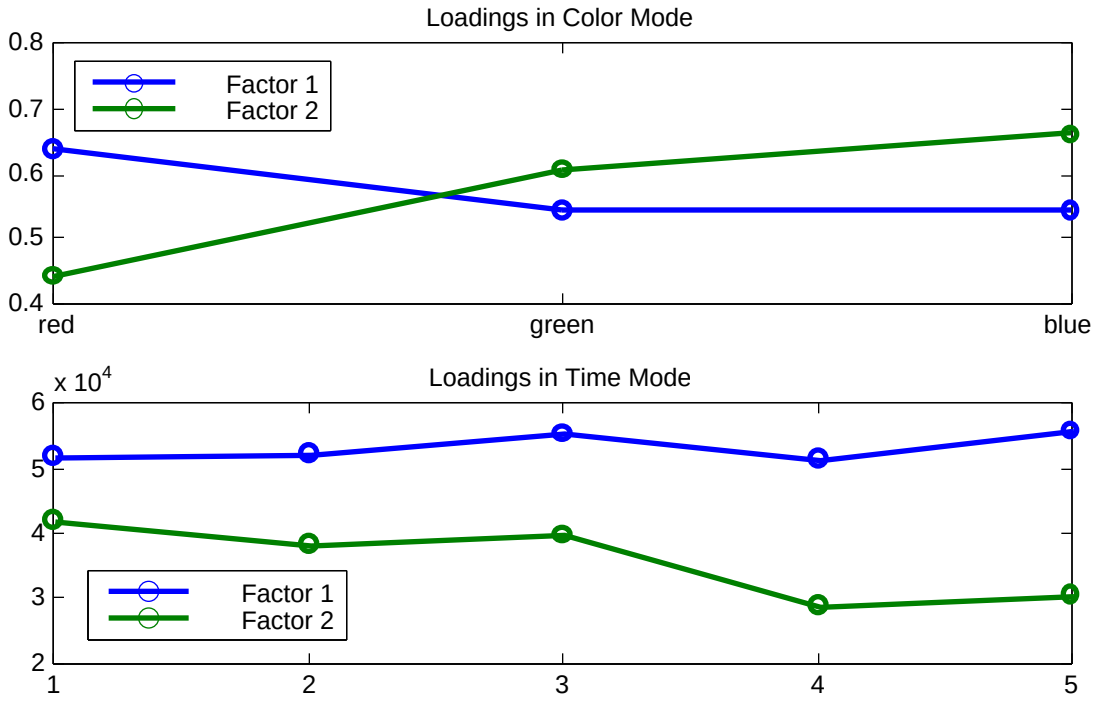

Figure 7. PARAFAC Loadings from Color and Time Modes of Image Sequence.

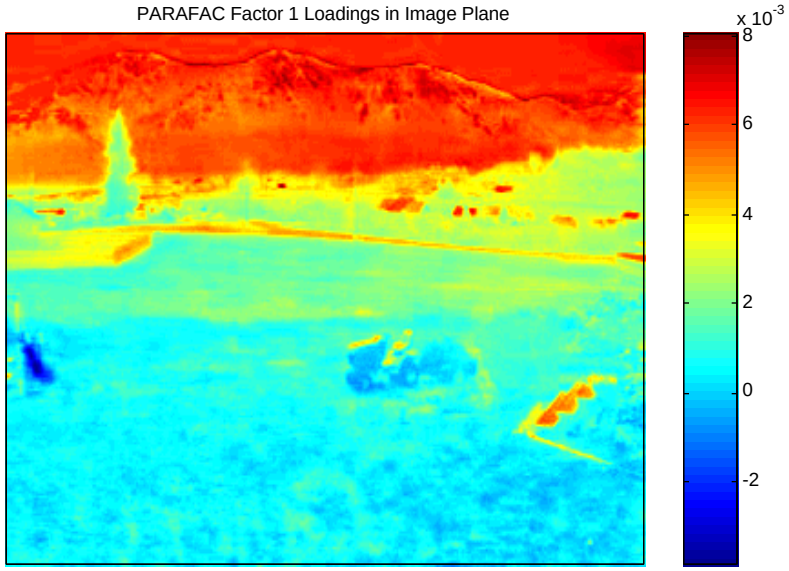

Figure 8. PARAFAC Model Factor 1 Loadings in Image Plane

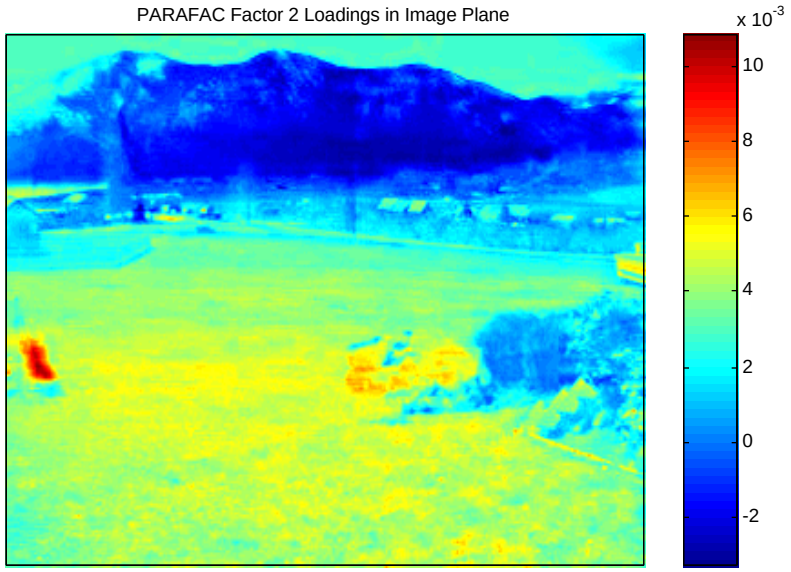

Figure 9. PARAFAC Model Factor 2 Loadings in Image Plane

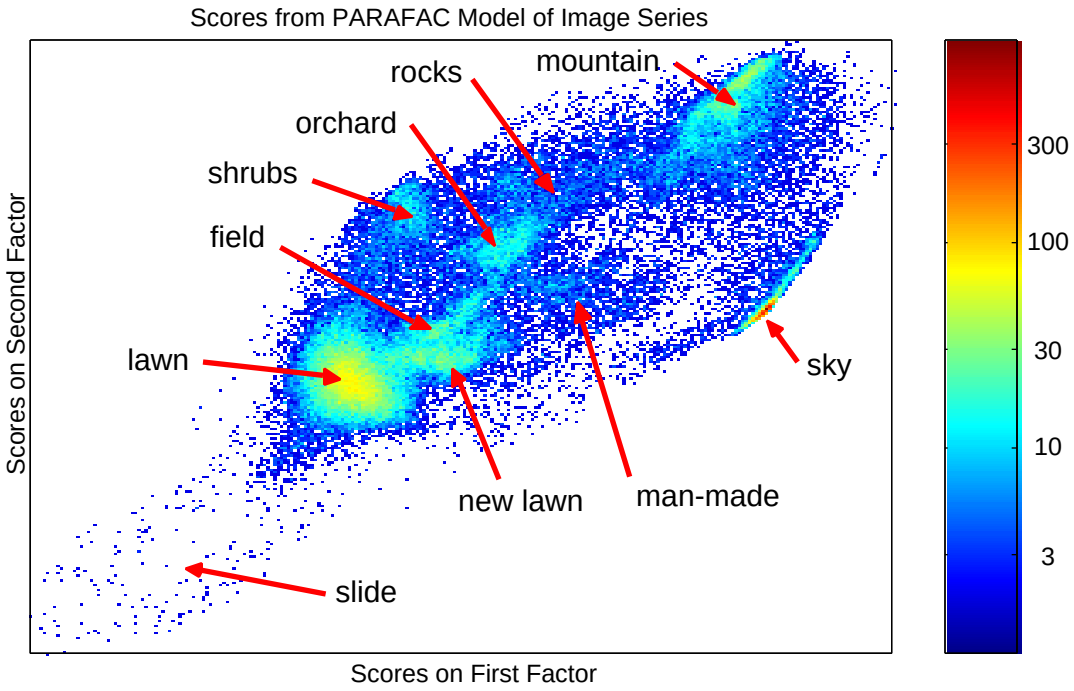

Figure 10. Scores on First Two PARAFAC Factors in Pixel Mode.

The residual in the pixel model over all of the images collectively is shown in Figure 11. The image has been "thresholded" so that a few pixels with very high residuals do not take up most of the color map. The largest residuals are associated with elements of the image that changed from frame to frame. The lawn tractor and trailer is obvious in the foreground, it is only in the last frame. A car moving along the road can also be seen (in image 2 only), along with the lake, which changed considerably due to breezes in the first two frames.

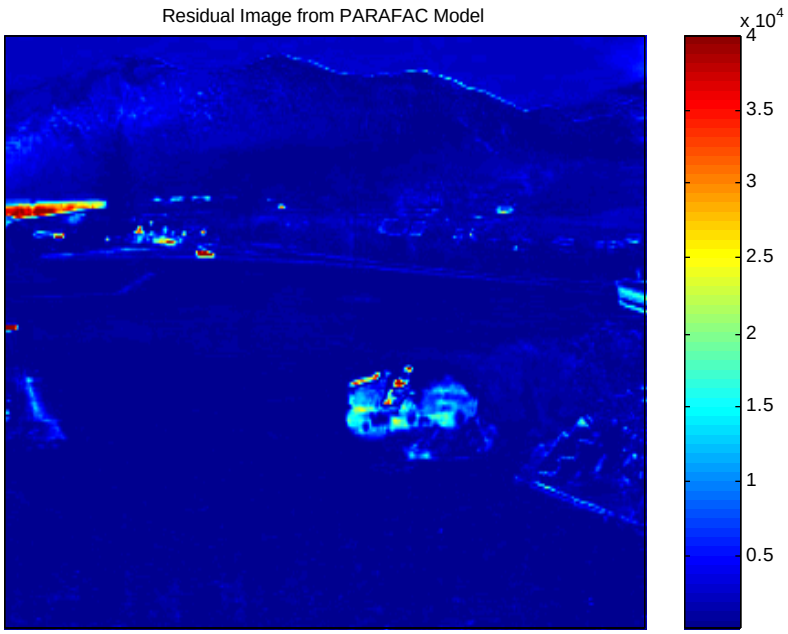

Figure 9. Residual Image from Two Factor PARAFAC Model Thresholded to 40,000.

The analysis of this series of images can also be done with PCA if the images are unfolded in the pixel modes and time and color modes to produce a matrix that is 84,480 (240x352) by 15 (3x5). Results (not shown) are similar. The main advantage of the PARAFAC model is that the time and color modes are not convolved, easing interpretation. The main disadvantage of the PARAFAC model is that it takes much longer to calculate than the PCA model (about an hour versus 18 seconds).

Conclusions

Multivariate images are a rich source of information but present unique challenges due to their structure and abundance of data. Techniques based on PCA can be used to gain insight into their overall structure and the relationships of the parts of the image. Additional techniques, such as PARAFAC, can be used on series of images.

Acknowledgment

All results shown in this manuscript were generated with MATLAB and PLS_Toolbox. The authors would like to thank Anna Belu of Physical Electronics for providing the SIMS data on samples provided by the Garcia Center for Polymers at Engineered Interfaces at Stoney Brook.

Literature

For more information on MIA we suggest the following references:

P. Geladi and H. Grahn, *Multivariate Image Analysis*, Wiley, Chichester, 1996.

P. Geladi and H. Grahn,, "Multivariate Image Analysis", in *Encyclopedia of Analytical Chemistry*, Wiley, Chichester, in press, 2000.

P. Geladi, "Sampling and local models for multivariate image analysis", *Microchimica Acta*, **120**, pps. 211-230, 1995.

A. Kriete ed., *Visualization in Biomedical Microscopies*, VCH, Weinheim, 1992.

F. Toselli and J. Bodechtel eds, *Imaging Spectroscopy: Fundamentals and Applications*, ECSC, EEC, EAEC, Brussels, 1992.

B. M. Wise and P. Geladi, "Analysis of a Series of Images with PCA and PARAFAC," presented at *TRICAP 2000: Three-way Methods in Chemistry and Psychology*, Fåborg, Denmark, July, 2000.