# Introduction to Hyperspectral and Multivariate Image Analysis and Principal Components Analysis for Multivariate Images

Neal B. Gallagher, Lyle Lawrence

Eigenvector Research, Inc., www.eigenvector.com

Dec., 2020

## Introduction

Hyperspectral imaging is a popular and useful tool for a broad range of sensing applications and the goal of this white paper is to introduce hyperspectral imaging in a manner assessable to those not yet familiar with the subject. Hyperspectral imaging is also referred to as multivariate imaging and a short introduction to multivariate image analysis using principal components analysis (PCA) is also provided. PCA is ubiquitous to multivariate analysis and is often used as an exploratory tool for hyperspectral images (HSI). Due to its favorable mathematical properties, PCA is also used as the framework for more advanced multivariate algorithms employed in the analysis of HSI. As a result, even a cursory understanding of PCA can provide insights into many algorithms used in multivariate image analysis (MIA). After an introduction to HSI and PCA, several examples of PCA applied to hyperspectral images are shown.

## Multivariate and Hyperspectral Imaging

Gray-scale images can be considered "univariate images" and typical uint8 (unsigned 8-bit integer) images have color values ranging from 0 to 255. These images provide spatial information and a vast array of tools exist for extracting spatial information from gray-scale images. Typical color images can be thought of as three univariate images combined into a red-green-blue (RGB) display and enhance the spatial statics significantly. Figure 1 demonstrates this concept by showing an example gray-scale on the left and a corresponding RGB true color image on the right. The example is intuitive and easily assessable by humans that see in color. However, what happens when the number of colors, $N$, increases beyond three? When $N > 3$, it is difficult to make a simple RBG representation unless only three colors are selected for display. Additionally, it is highly likely the three selected

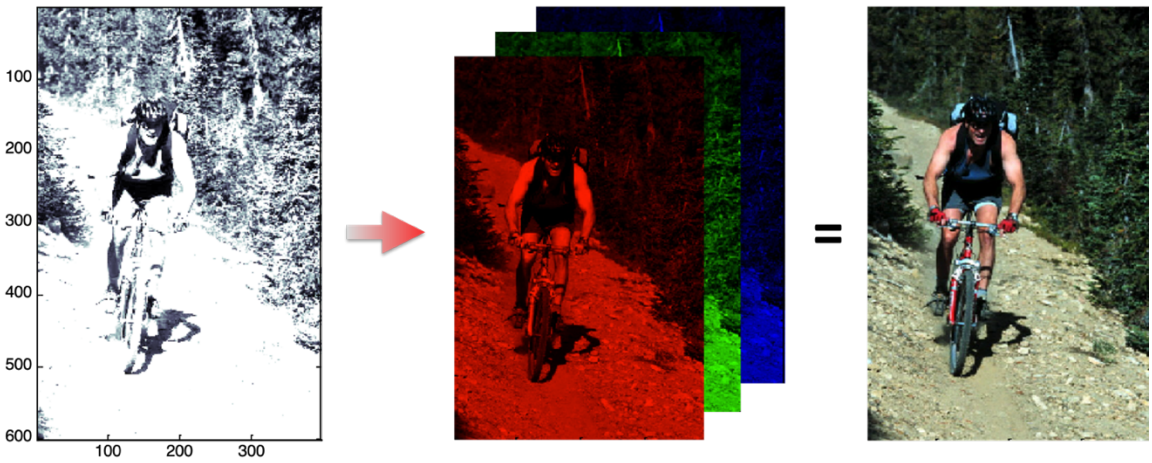

**Figure 1. Example gray-scale image (left), corresponding three univariate color images (middle), combined univariate color images to multivariate RGB image (right).**

colors will not match the natural true color scheme used in human vision resulting in a false color image. When $N > 1$, the image is no longer univariate and is considered multivariate, and when $N \gg 1$, the image is considered "multispectral" or "hyperspectral." The definition is loose, but images with $1 < N < 10's$ are often referred to as multispectral and for $N > 10's$ they are termed hyperspectral. It was shown above in an intuitive manner that the information content dramatically increased in Figure 1 as $N = 1 \rightarrow N = 3$. Consequently, it might be expected that the information content in a hyperspectral image can be very large and mathematical tools used for analysis of these images can be quite sophisticated.[1,2]

The variables in HSIs often correspond to signal measured in spectral channels such as absorbance or reflectance in infrared imaging or counts at specific mass channels in mass spectrometry, and this is where HSI gets "spectral" in its name. There are many types of spectroscopy used in HSI e.g., infrared, magnetic resonance imaging, Raman and UV-Vis. Anything that can be imaged for multiple variables can be considered a multivariate image and HSI can also be obtained for multiple images for a single sample. For example, confocal Raman imaging can be used to image samples at different depths and depth profiling using ToF-SIMS imaging for surface analysis can be obtained by successively sputtering layers from a sample after acquiring an image.

Hyperspectral imaging is most useful when measuring and investigating heterogeneous samples. The reason is that many pixels (1000-10,000's of spectra) are acquired simultaneously for an image and these can be used to characterize background clutter signal (signal not of interest). Anomalous signal that may be of interest can be found by comparing all the image spectra to the distribution of clutter spectra. Although signal from anomalies may be small and less frequently observed, the signal in a single pixel can be dominated by an anomaly. In contrast, many analytical procedures create homogeneous mixtures of a sample before measurements thus diluting minor signal making it much more difficult to detect and classify signal of interest. In addition to detecting anomalies, grouping pixels with similar signals results in image segmentation and the spatial statistics provided is useful for particle analysis, characterizing sample homogeneity, land use characterization, precision agriculture and change detection.

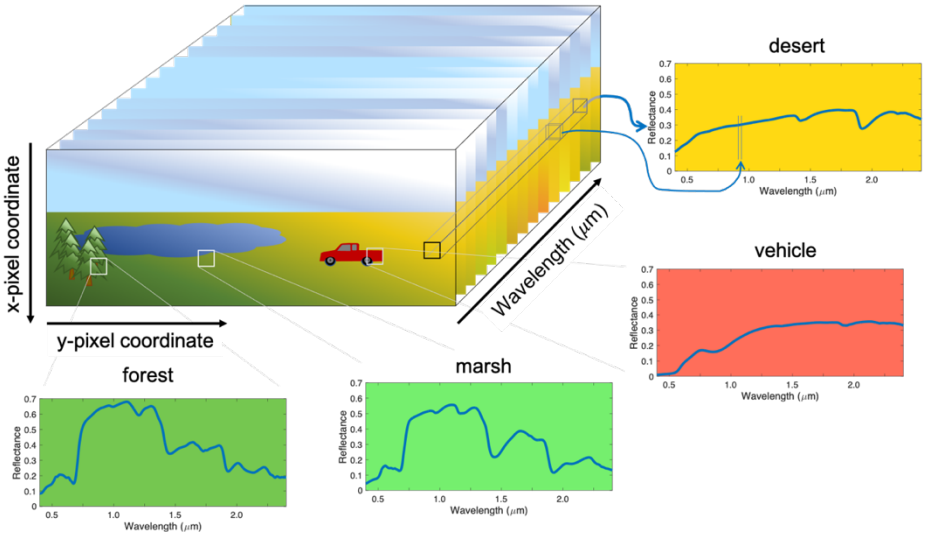

**Figure 2. Example application of remote hyperspectral imaging.**

HSI is used at scales from nanometer to astronomical. Imaging platforms are used for remote sensing from satellites and airborne aircraft, standoff sensing (Figure 2), and micro-imaging. Applications include surface analysis, medical imaging, chemical samples, cultural heritage and art, agriculture, environmental, earth observing, and many more.

A hyperspectral image (HSI) given by $\underline{\mathbf{X}}$ is size $M_\mathrm{x} \times M_\mathrm{y} \times N$ where $M_\mathrm{x} \times M_\mathrm{y}$ corresponds to the spatial dimensions and $N$ corresponds to the spectral dimension. Each image contains a total of $M = M_\mathrm{x}M_\mathrm{y}$ pixels and an individual pixel is $\mathbf{x}_m = [x_{m,1} \quad x_{m,2} \quad \cdots \quad x_{m,N}]^\mathrm{T}$ (Figure 3). Each pixel can be referenced by its subscript $\mathbf{x}_{m_\mathrm{x},m_\mathrm{y}}$ where $m_\mathrm{x} = 1, \ldots, M_\mathrm{x}$ and $m_\mathrm{y} = 1, \ldots, M_\mathrm{y}$, but it is often more convenient to use the linear index $\mathbf{x}_m$ where $m = 1, \ldots, M$. Correspondingly, each voxel can be referenced by its subscript $x_{m_\mathrm{x},m_\mathrm{y},n}$ where $n = 1, \ldots, N$ or by its linear index $x_{m,n}$. Although the subscript reference may seem more natural, many applications such as principal components analysis use the linear indexing reference. To understand this, the concept of matrix "matricizing" is introduced. Matricizing rearranges an $M_\mathrm{x} \times M_\mathrm{y} \times N$ image $\underline{\mathbf{X}}$ to a $M \times N$ matrix $\mathbf{X}$ as shown in Figure 4. Matricizing destroys the spatial structure but creates a matrix of measured spectra amenable to analysis by principal components analysis.

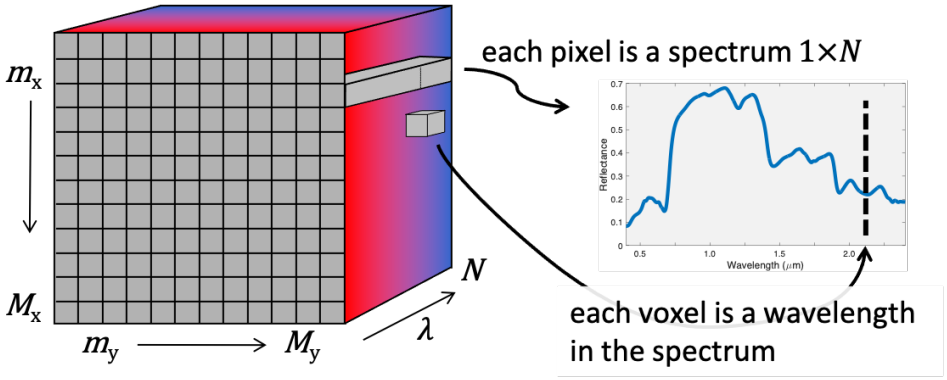


**Figure 3. Example of a hyperspectral image cube.**

## Principal Components Analysis (PCA) for Hyperspectral Images: Multivariate Image Analysis (MIA)

As stated in the introduction, principal components analysis (PCA) is ubiquitous to multivariate analysis and PCA can be traced back to at least the early twentieth century.[3,4] Additional descriptions of PCA and its applications can be found in [5] and [6] and in thousands of papers on the subject. PCA is especially useful when the number of variables $N$ in a data matrix $\mathbf{X}$ is large and there is redundancy, or correlation, between the variables. Spectroscopic measurements, like those used in HSI, often exhibit correlation (see the discussion in [7]). PCA finds linear combinations of the $N$ variables that captures correlations between variables and samples (pixels) to provide a smaller set of $K$ new variables called "loadings", i.e., loadings are linear combinations of the original variables. It is typical in HSI that $K \ll N$ resulting in significant data compression and improved signal-to-noise. Coefficients are also obtained for each pixel to describe how much each loading contributes to the signal in that pixel. The coefficients are called "scores" and are stored in a $M \times K$ matrix. In summary, PCA is a matrix decomposition performed on the $M \times N$ matricized data matrix $\mathbf{X}$, that reduces the size of the problem to a $M \times K$ scores matrix that

encompasses redundant information in the image. The scores capture the spatial statistics in the HSI and the final step reshapes the matricized scores image(s) back to a $M_x \times M_y \times K$ image cube that can be used for visualization. Several examples are shown below. Analysis was performed using Solo and MIA_Toolbox [Eigenvector Research, Inc., Manson, WA].

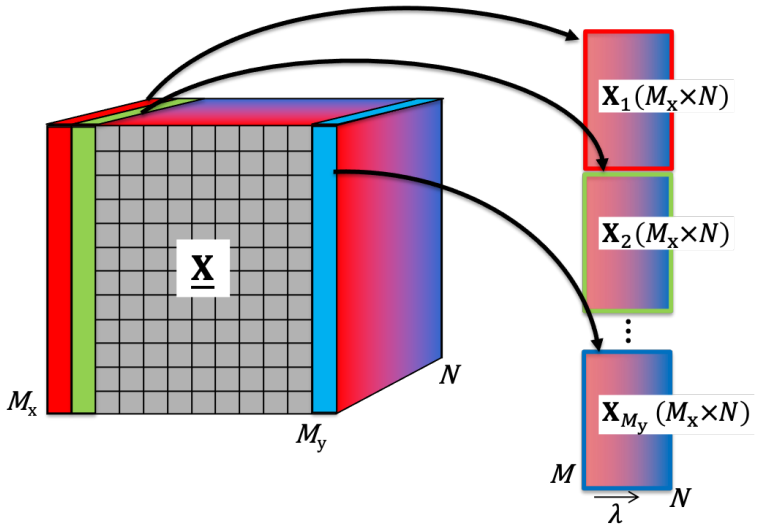


**Figure 4. Matricizing an $M_x \times M_y \times N$ image $\underline{\mathbf{X}}$ (left) to a $M \times N$ matrix $\mathbf{X}$ (right) with $M_x \times N$ sub-images $\mathbf{X}_{m_y}$ stacked vertically.**

Scores images can be examined one-at-a-time as gray-scale images for each of the $K$ principal components (PCs) where $k = 1, \ldots, K$. This exploratory procedure is far more efficient than attempting to visualize all $N$ images at each wavelength one-at-a-time – especially when $N$ is large. Figure 5 (left) shows an example of PCA scores image for PC 2 ($k = 2$) for a pharmaceutical tablet obtained using infrared imaging. The yellow color corresponds to an active ingredient distributed as ~200-400 μm ellipses in the image. In the PCA decomposition, the first principal component (PC) is the combination of variables that captures the most signal in the image. The second PC captures the second most signal and successive PCs are ordered by the amount of signal described in the image. In Figure 5 (left), PC 2 captures approximately 19% of the variance, or information, in the image.

In contrast to images for a single PC, Figure 5 (right) shows an RGB false color image for the same HSI image for scores on PC 1, 3 and 4 combined (red, green, blue respectively) capturing approximately 72% of the total information in the image. Any combination of scores can be displayed in this fashion. In Figure 5 (right), two types of additives are shown as irregularly shaped ~200-400 μm particles in cyan and yellow-green. Figure 5 shows that the two additives were discriminated from the active ingredient even though their signal is overlapped in the image.

Figure 6 shows another chemical image for a drug bead embedded in epoxy using time-of-flight secondary ion mass spectroscopy (ToF-SIMS). The signal-to-noise is not high in an individual pixel but the averaging aspects of PCA provides the general signal trends such as the bead coating as a vertical purple blue band on the right-hand side, the active ingredient is are pink and orange particles embedded in a green colored excipient. The image in Figure 6 (left) shows PC 1 that captures nearly 27% of the information with the active ingredient as yellow “islands” and Figure

6 (right) shows the scores on PCs 2, 3 and 4 capturing a total of 8.7% of the signal in the image. This small

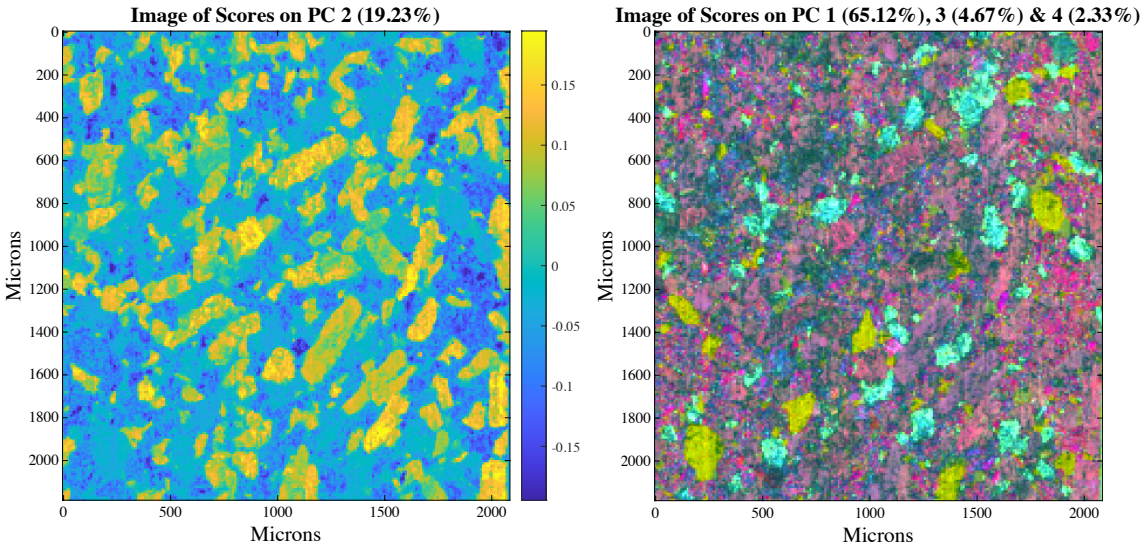


**Figure 5. Chemical image of a pharmaceutical tablet, 218 by 208 pixels measured in the infrared at 250 wavenumbers 1800 to 800 cm-1 with 4 cm-1 resolution. The data were mean-centered before PCA and the scores were scaled (contrasted) to lie between 0±2 standard deviations [this contrasting is called "autocontrasting"]. (left) Scores image on PC 2 showing an active ingredient in yellow. (right) RBG false color image of the scores on PC 1, 3 and 4 respectively showing additives in cyan and yellow-green. [Data from Agilent www.agilent.com.]**

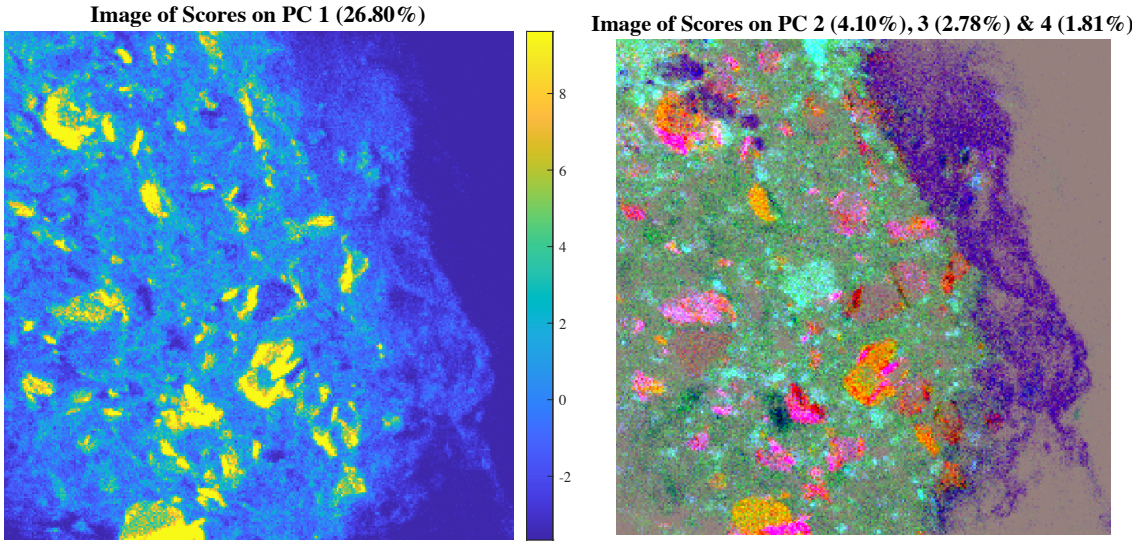


**Figure 6. Chemical image of a prednisolone sodium in microcrystalline cellulose and lactose, 256 by 256 pixels (250 by 250 µm$^2$) measured at 93 mass channels. The data were Poisson scaled (3% offset) and mean-centered before PCA. The scores were autocontrasted. (left) Scores image on PC 1 with active shown in yellow. (right) RBG false color image of the scores on PC 2, 3 and 4 respectively show the active ingredient is in pink and orange yellow, excipient in green and yellow and coating in purple blue as the vertical band on the right-hand side. (Data from Belu[9] and also described in [10].)**

percentage is not unusual for ToF-SIMS and a wide variety of data preprocessing approaches are often used to enhance signal of interest. In this example, the data were Poisson scaled [8] and mean-centered.

The chemical images in Figures 5 and 6 are examples of small spatial scales on the order of microns. In contrast, it is typical for cultural heritage and art research to study at centimeter or meter scales. Figure 7 shows PCA applied to an image of a small section of a 15th Century Palimpsest where older writing (horizontal script) was washed off and overwritten by more modern script (vertical). As might be imagined, the horizontal writing is generally more difficult to see in the image and a "de-cluttering" preprocessing was used to suppress signal from the modern writing and paper in an effort to enhance the ancient script.[11] The de-cluttering is based on generalized least squares weighting[12-14] and although the suppression isn't perfect, the procedure is useful as a synergistic image analysis tool and can be used in a wide variety of applications.[15,16]

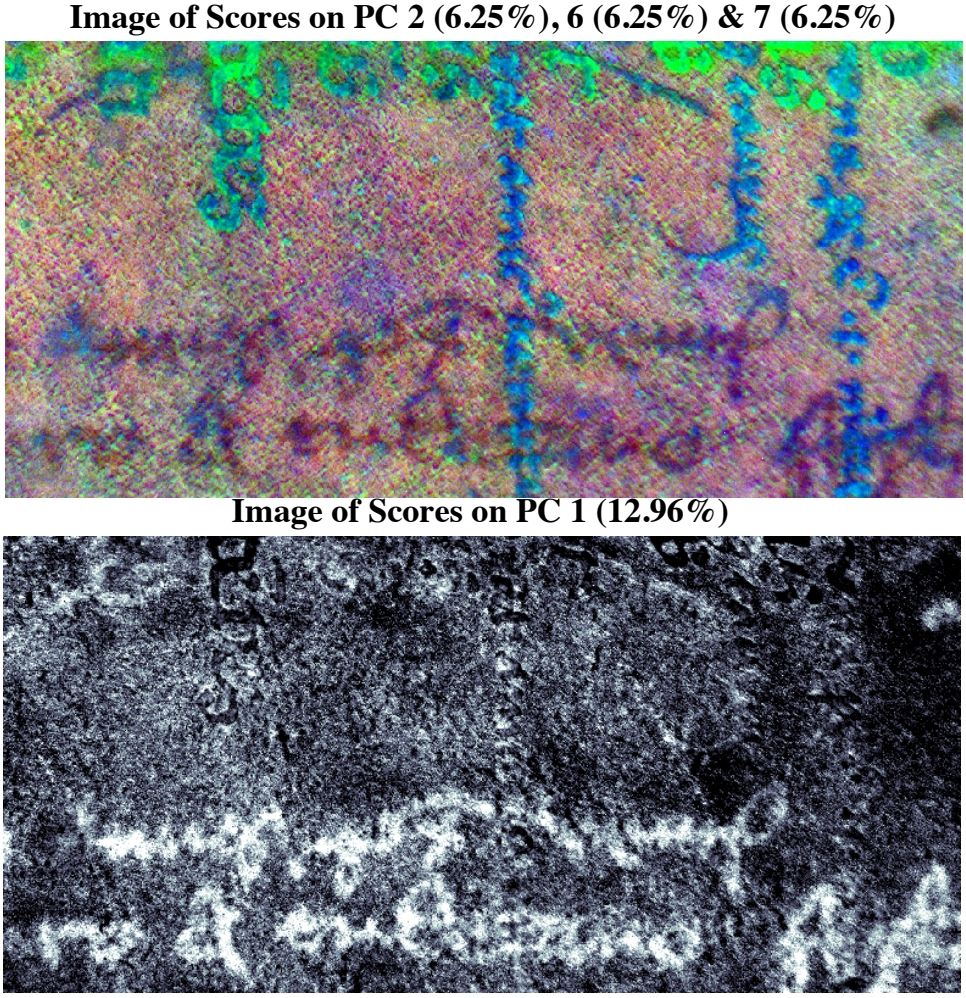


**Figure 7. Small section of a 15th Century Palimpsest 495 by 935 pixels 365-1000 nm, Library of Congress, Preservation Research and Testing Division.[11] (top) 1-Norm and GLSW suppression of paper and modern ink [verical script]. (bottom) GLSW suppression of paper ink with top paper pixels excluded.**

Standoff and remote sensing applications are at meters or kilometer distances where atmospheric correction might be employed. Figure 8 shows a mid-infrared image of mineral samples mounted on a 1.07 m × 1.22 m plywood board imaged at a distance of 14 m. The data are described in detail in Myers et al.[17] for a similar image on an aluminum board. The RGB image corresponds to PC 1, 3 and 4. PC 1 captures a large fraction of the variance while discrimination of the minerals is associated with PCs >1 that capture a much smaller fraction of the image's information. The objective of the original study was to determine if a library of laboratory measured spectra could be used to detect and classify each of the 24 samples in the image using target detection. The study showed that this was possible for many but not all of the minerals studied. However, with a few exceptions, it is encouraging that PCA exploratory analysis clearly discriminates the minerals quite well in Figure 8. PCA also discriminates the label under each sample and grain in the plywood.

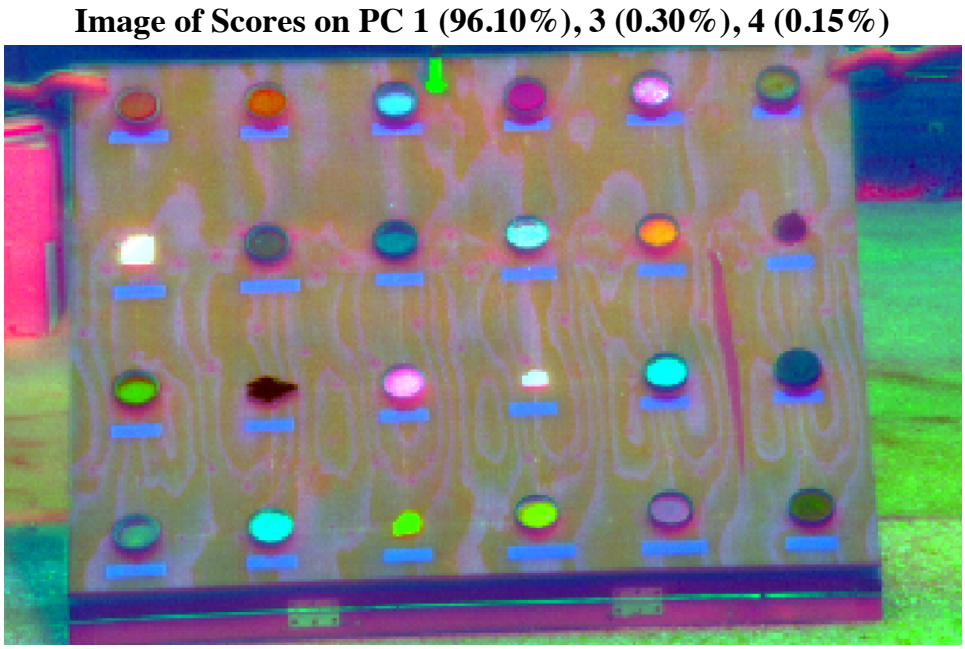


**Figure 8. Standoff image 320 by 200 pixels of mineral samples mounted on a plywood board at 14 m in the 1300 to 850 $cm^{-1}$ (7.7 to 11.8 µm) range at 4 cm−1 resolution acquired on a Telops imager [www.telops.com]. The data Poisson-scaled, 1-mean-centered prior to PCA.**

Figure 9 shows the last example measured from a Landsat 8 satellite. This is a multi-spectral image measured at 8 bands from a distance of about 705 km above the earth. The image is of Lake Chelan in north central Washington, USA. The lake is over 80 km long and 450 m deep. Figure 9 shows a false color RGB image for scores on PC 1, 2 and 3. Because color order is arbitrary it is a coincidence that PC 2 corresponds to the green color associated with vegetation near valley bottoms, and PC 3 corresponds to the blue color associated with water signal in the image. Bright yellow-green in the southeast of the image corresponds to agricultural areas (orchards and vineyards) and lawn. The discrimination of different land characteristics in Figure 9 demonstrates how Landsat images can be used for land use monitoring and management.

## Conclusions

Hyperspectral images contain enormous amounts of information and are being used in a broad range of applications. Principal components analysis (PCA) is a useful tool from multivariate

analysis that can be used to explore hyperspectral images. There are many more imaging methodologies and tools available for analysis of hyperspectral images. However, one of the biggest challenges in evaluating the performance of HSI analysis algorithms is establishing ground truth because of potential registration mismatch in the spatial domain and mixed pixels can be difficult to unequivocally quantify and classify. New methods and uses for HSI are being developed and it is anticipated that the field of hyperspectral imaging will grow for many years to come. For additional information into the growing field of hyperspectral imaging and multivariate image analysis, the interested reader is encouraged to explore references [1] and [2] and view webinars at https://eigenvector.com/resources/webinars/.

**Image of Scores on PC 1 (60.43%) & 2 (27.90%) & 3 (10.95%)**

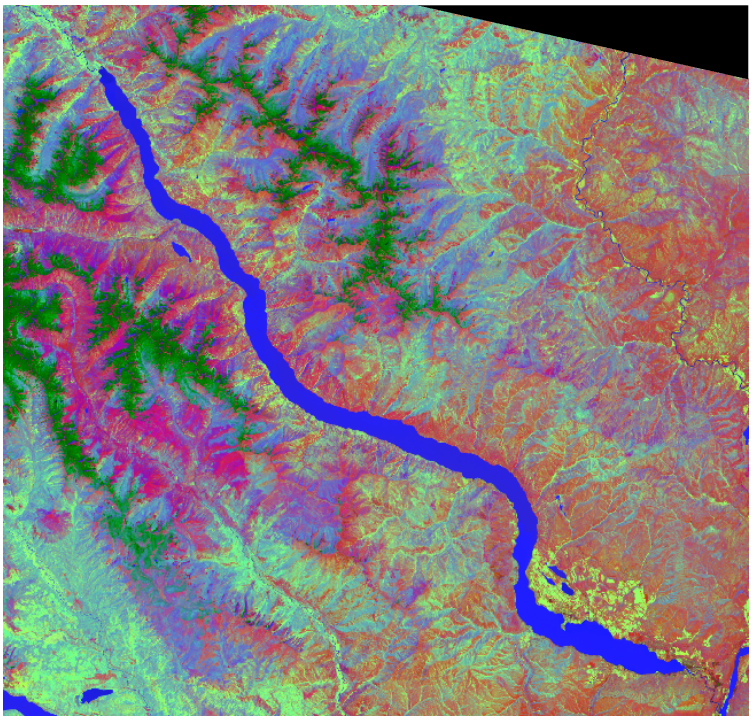

**Figure 9. Remotely sensed image 2050 by 2150 pixels measured at eight bands (Bands 1 to 7 plus Band 9) and a 30 m spatial resolution. The data were Poisson-scaled, normalized to 1-norm and mean-centered prior to PCA and the scores were autocontrasted [a similar image without Poisson-scaling can be found athttps://www.impopen.com/vi-toc/V_IASIM-2018].**

## References


1. *Hyperspectral Imaging*, Vol 32 (Data Handling in Science and Technology) 1st Ed, J. Manuel Amigo editor, Elsevier 2019. ISBN: 9780444639776.
2. *Techniques and Applications of Hyperspectral Image Analysis*, Grahn, H.F., Geladi, P., Eds. John Wiley & Sons: West Sussex, England, 2007. ISBN: 978-0-470-01087-7.
3. Jackson, J.E., *A User's Guide to Principal Components*, John Wiley & Sons, New York, NY, 1991. ISBN: 9780471622673.

4. Person, K., "On Lines and Planes of Closest Fit to Points in Space," *Philosophical Magazine*, **2**, 559–572 (1901).

5. Bro, R., Smilde, A.K., "Principal Components Analysis," *Anal. Methods*, **6**(9), 2812–2831 (2104). DOI: 10.1039/c3ay41907j.

6. Wise, B.M., Gallagher, N.B., "The Process Chemometrics Approach to Chemical Process Monitoring and Fault Detection," *J. Proc. Cont.*, **6**(6), 329–348 (1996). https://doi.org/10.1016/0959-1524(96)00009-1

7. Chatterjee, S., Singh, B., Diwan, A., Lee, Z.R., Engelhard, M.H., Terry, J., Tolley, H.D., Gallagher, N.B., Linford, M., "A Perspective on Two Established Chemometrics Tools: PCA and MCR, and Introduction of a New One: Pattern Recognition Entropy (PRE), As Applied to XPS and ToF-SIMS Depth Profiles of Organic and Inorganic Materials," *Appl. Surf. Sci.*, **433**, 994–1017 (2018), doi: 10.1016/j.apsusc.2017.09.210.

8. Keenan, M.R., "Multivariate Analysis of Spectral Images Composed of Count Data," 89–126, in [2].

9. Belu, A.M., Davies, M.C., Newton, J.M., Patel, N., "TOF-SIMS Characterization and Imaging of Controlled-Release Drug Delivery Systems," *Anal. Chem.*, **72**(22) 5625–5638 (2000). https://doi.org/10.1021/ac000450+

10. Gallagher, N.B., Shaver, J.M., Martin, E.B., Morris, J., Wise, B.M. and Windig, W., "Curve resolution for images with applications to TOF-SIMS and Raman", *Chemometr. Intell. Lab.*, **73**(1), 105–117 (2003). https://doi.org/10.1016/j.chemolab.2004.04.003.

11. Dahlberg, D., Wilson, M., France, F., Gallagher, N., "Revealing the Hidden Writing of a 15th Century Palimpsest Using Hyperspectral Imaging Analyzed by Principal Component Analysis and Generalized Least Squares Weighting," EAS15, Somerset, NJ, Nov 16-18 (2015).

12. Martens, H., Høy, M., Wise, B.M., Bro, R., Brockhoff, P.B., "Pre-whitening of data by covariance-weighted pre-processing," *J. Chemometr.*, **17**(3), 153–165 (2003).

13. Gallagher, N.B., "Classical Least Squares for Detection and Classification," 231–246, in [1].

14. Gallagher, N.B., "Detection, Classification and Quantification in Hyperspectral Images using Classical Least Squares Models," 181–201, in [2].

15. Blake, T.A., Kelly, J.F., Gallagher, N.B, Gassman, P.L., Johnson, T.J., "Passive Detection of Solid Explosives in Mid-IR Hyperspectral Images," *Anal. Bioanal. Chem.*, **395**(2), 337-348 (2009). https://doi.org/10.1007/s00216-009-2907-5.

16. Gallagher, N.B., Shaver, J.M., Bishop, R., Roginski, R.T., Wise, B.M., "Decompositions with Maximum Signal Factors," *J. Chemometr.*, **28**(8), 663-671 (2014), DOI: 10.1002/cem.2634.

17. Myers, T.L., Johnson, T.J., Gallagher, N.B., Bernacki, B.E., Beiswenger, T.N., Szecsody, J.E. Tonkyn, R.G., Ashley M. Oeck, A.M., Su,Y.-F., Danby, T.O., "Hyperspectral Imaging of Minerals in the Longwave Infrared: The Use of Laboratory Directional-Hemispherical Reference Measurements for Field Exploration Data," *J Appl Remote Sens*, **13**(3), 034527 (2019). DOI: 10.1117/1.JRS.13.034527.