Motivating Non-Negative Matrix Factorizations[*]

Stefan Wild, James Curry[†] and Anne Dougherty

1 Introduction

Given a vector space model encoding of a large data set, a usual starting point for data analysis is rank reduction [1]. However, standard rank reduction techniques such as the QR, Singular Value (SVD), and Semi-Discrete (SDD) decompositions and Principal Component Analysis (PCA) produce low rank bases which do not respect the non-negativity or structure of the original data. Non-negative Matrix Factorization (NMF) was suggested in 1997 by Lee and Seung (see [6] and [7]) to overcome this weakness without significantly increasing the error of the associated approximation. NMF has been typically applied to image and text data (see for example: house and facial images in [6], handwriting samples in [9]), but has also been used to deconstruct music tones [4]. The additive property resulting from the non-negativity constraints of NMF has been shown to result in bases that represent local *components* of the original data (i.e.- doors for houses, eyes for faces, curves of letters and notes in a chord). In this paper, we intend to motivate the application of NMF techniques (with noted corrections) to other types of data describing physical phenomena.

The contents of this paper are as follows. Section 2 details the NMF objective functions and update strategies used here and in practice. In Section 3 we illustrate both the error and resulting basis for text and image collections. We then turn to the specific example of NMF performed on remote sensing data in Section 4. We emphasize recently proposed NMF alterations and compare the output obtained with the remote sensing literature. We conclude in Section 5 by describing future work for this promising area.

[*] This work was initiated under NSF VIGRE Grant number DMS-9810751.

[†] Author to whom correspondence should be addressed: Department of Applied Mathematics, University of Colorado, UCB 526, Boulder, CO 80309-0526.

2 NMF Background and Theory

Given a non-negative matrix **X** of size $m \times n$, NMF algorithms seek to find non-negative factors **W** and **H** such that

$$X \approx \tilde{X} \equiv WH, \qquad \text{where} \quad W \in \Re^{m\times r} \quad \text{and} \quad H \in \Re^{r\times n}, \tag{1}$$

or equivalently, the columns $\{x_j\}_{j=1}^{n}$ are approximated such that

$$x_j \approx \tilde{x_j} = Wh_j, \qquad \text{where} \quad x_j \in \Re^{m} \quad \text{and} \quad h_j \in \Re^{r}. \tag{2}$$

For the class of full (nonsparse) matrices, this factorization provides a reduction in storage whenever the number of vectors, r, in the basis **W** is chosen such that $r < \frac{nm}{m+n}$. We emphasize here that the problem of choosing r for NMF is considerably more cloudy than looking at the decay of the magnitudes of the eigenvalues of the data, as is done in traditional rank reduction techniques. We have recently explored an efficient method of determining an r based on the error of the resulting approximation [10] and [11]. In practice, r is usually chosen such that $r \ll \min(m, n)$.

2.1 NMF Algorithms and Update Strategies

Using an approach similar to that used in Expectation-Maximization (EM) algorithms, Lee and Seung first introduced the NMF algorithms commonly used in practice [7]. In general, NMF algorithms seek to iteratively update the factorization based on a given objective function (distance measure). While each objective function could be minimized with several different iterative procedures, the update strategies given here are shown because of their implementation ease and because they have been proven to monotonically decrease their respective objective function. We acknowledge that other update strategies that monotonically decrease the objective functions here are conceivable. Further, for most conceivable objective functions, the lack of convexity in both factors **W** and **H**, means that we can, at best expect to achieve only local minima [7].

The first, and perhaps most natural, objective function is to minimize the (square of the) Euclidean distance between each column of **X** and its approximation $\tilde{X} = WH$. Using the Frobenius norm for matrices we have:

$$\Theta_E(W,H) \equiv \sum_{j=1}^{n} \|x_j - Wh_j\|_2^2 = \|X - WH\|_F^2 \equiv \sum_{i=1}^{m}\sum_{j=1}^{n} \left(X_{ij} - \sum_{l=1}^{r} W_{il}H_{lj} \right)^2 \tag{3}$$

The lower bound of zero for $\Theta_E(W,H)$ will only be attained when a strict equality $X = WH$ is obtained. Seung and Lee have chosen to balance algorithm complexity and convergence speed by using the following update procedure:

$$H_{aj} \leftarrow H_{aj} \frac{[W^T X]_{aj}}{[W^T WH]_{aj}}, \tag{4}$$

$$W_{ia} \leftarrow W_{ia} \frac{[XH^T]_{ia}}{[WHH^T]_{ia}}, \tag{5}$$

where $[\,\cdot\,]_{ij}$ indicates that the noted divisions and multiplications are computed element by element. To aid convergence, we use the newest factors of **W** and **H** available. From an iterative standpoint, it may be helpful to write the update above as:

$$H_{aj}^{(t+1)} = H_{aj}^{(t)} Q_E\left(W^{(t)}, X^T, H^{(t)}\right)_{aj} \tag{6}$$

$$W_{ia}^{(t+1)} = W_{ia}^{(t)} Q_E\left(X^T, H^{(t+1)}, W^{(t)}\right)_{ia} \tag{7}$$

When written this way, it is evident that the update consists of multiplying the current factors by a measure of the quality of the current approximation:

$$Q_E(A, B, C) \equiv \frac{[A^T B^T]_{ia}}{[AC \diamond C^T]_{ia}}, \tag{8}$$

where $AC \diamond C^T$ signifies the "correct" right or left multiplication of AC by C^T. Under these updates, the Euclidean distance objective function Θ_E has been proven [7] to be monotonically decreasing: $\Theta_E\big(W^{(t+1)}, H^{(t+1)}\big) \leq \Theta_E\big(W^{(t)}, H^{(t)}\big)$.

The second objective function that is commonly used in practice is called the divergence, or entropy, measure:

$$\Theta_D(W, H) \equiv \mathrm{Div}(X \| WH) \equiv \sum_{i=1}^{m} \sum_{j=1}^{n} \left(X_{ij} \log \frac{X_{ij}}{\sum_{l=1}^{r} W_{il} H_{lj}} - X_{ij} + [WH]_{ij} \right) \tag{9}$$

The objective function $\Theta_D(W, H)$ is not a distance measure because it is not symmetric in both **X** and **WH**. Further motivation behind this objective function can be seen when the columns of **X** and the columns of the approximation **WH** sum to 1, in which case $\Theta_D(W, H)$ reduces to the Kullback-Leibler information measure used in probability theory. This objective function is related to the likelihood of generating the columns in **X** from the basis **W** and encoding coefficients **H**.

Again, this objective function equals its lower bound of zero only when we have strict equality, $X = WH$. To balance complexity and speed, the following iterative updates are commonly used:

$$H_{aj}^{(t+1)} = H_{aj}^{(t)} Q_D\left(\left[W^{T\,(t)}\right]_{ai}, \frac{X_{ij}}{[W^{(t)} H^{(t)}]_{ij}}\right)_{aj} \tag{10}$$

$$W_{ia}^{(t+1)} = W_{ia}^{(t)} Q_D\left(\frac{X_{ij}}{[W^{(t)} H^{(t+1)}]_{ij}}, \left[H^{T\,(t+1)}\right]_{ja}\right)_{ia} \tag{11}$$

$$W_{ia}^{(t+1)} \leftarrow \frac{W_{ia}^{(t+1)}}{\sum_j W_{ja}^{(t+1)}} \tag{12}$$

where $Q_D(A, B)_{ij} \equiv \sum_k A_{ik} B_{kj} = AB$ and the subscripts again indicate element by element division or multiplication. Lee and Seung have also proven [7] that this update monotonically decreases the objective function Θ_D.

A recently proposed refinement of NMF is a slight variation of the Divergence Algorithm detailed above which seeks to impose additional constraints on the spatial locality of the features of a data set:

$$\Theta_L(W,H) \equiv \sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}\log\frac{X_{ij}}{[WH]_{ij}} - X_{ij} + [WH]_{ij} + \alpha U_{ij}\right) - \beta\sum_{i} V_{ii}, \quad (13)$$

where α, $\beta > 0$ are some constants, $U = W^T W$ and $V = HH^T$. We refer the reader to [8] for further justification of the objective function $\Theta_L(W,H)$. A set of update rules that minimize this objective function are:

$$H_{aj}^{(t+1)} = \sqrt{H_{aj}^{(t)} Q_D\left(\left[W^{T\,(t)}\right]_{ai}, \frac{X_{ij}}{[W^{(t)}H^{(t)}]_{ij}}\right)_{aj}} \quad (14)$$

$$W_{ia}^{(t+1)} = W_{ia}^{(t)} Q_D\left(\frac{X_{ij}}{[W^{(t)}H^{(t+1)}]_{ij}}, \left[H^{T\,(t+1)}\right]_{ja}\right)_{ia} \quad (15)$$

$$W_{ia}^{(t+1)} \leftarrow \frac{W_{ia}^{(t+1)}}{\sum_j W_{ja}^{(t+1)}} \quad (16)$$

The structure of this LNMF update for **W** is identical to that of the Divergence Algorithm update, using the same quality function and differing only in the coefficient matrix **H** used. The update for **H** now uses an element by element square root to satisfy the three additional constraints in [8].

All three of these update strategies are $\mathcal{O}(mnr)$ at each iteration and are usually initialized with random factors $W^{(0)}$ and $H^{(0)}$. We have detailed additional seeding strategies and properties/ implementation modifications in [10] and [11].

3 Standard NMF Results

Using this representation, we see that the left factor **W** contains a basis used for the linear approximation of **X**. The right factor **H** is a coefficient matrix used to *add* up combinations of the basis vectors in **W**. The non-negative constraint on **W** allows us to visualize the basis columns in the same manner as the columns in the original data matrix. This is the first benefit of NMF versus alternative factorizations like the SVD where the basis vectors contain negative components that prevent similar visualization. The non-negativity constraints on both **W** and **H** do not come without an increase in both computational cost and approximation error. However, the error of NMF is competitive with the best low rank approximation obtained by the SVD, especially when compared to the (ordered) truncated QR. Figure 1 shows the Frobenius error $\|X - \tilde{X}\|_F$ for each of these three techniques for the standard rank 3889 Classic3 text data set.

The second, and usually desired, benefit of NMF is the structure of the resulting basis. For the text and image applications typically used, this basis will be r conceptual (or representative) documents/images stored in the columns of **W** that can sum up to (approximately) reconstruct the original document/image collection.

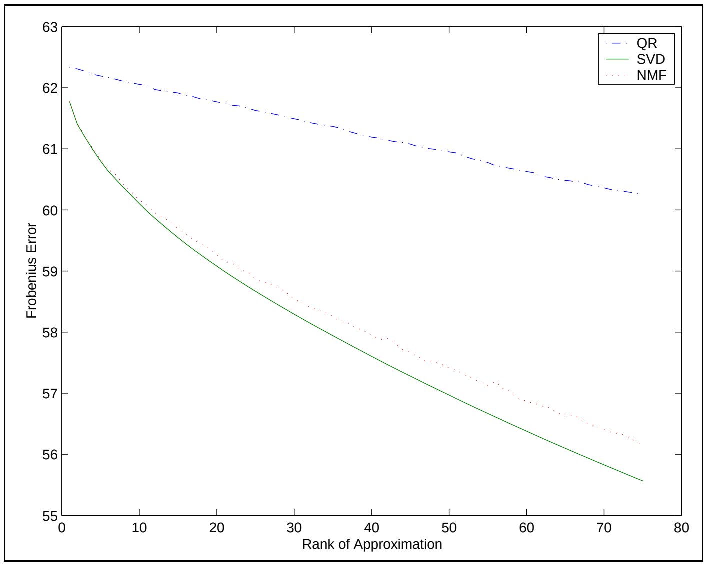


Figure 1. *Frobenius error of low rank approximations to Classic3 data set.*

When coupled with the constraint on the coefficient matrix **H**, we often arrive at a basis consisting of interesting local features. For example, when the Local NMF Algorithm is applied to a collection of 382 facial images, we obtain the basis (consisting of eyes, lips, etc.) shown in Figure 2. The special structure illustrated in this figure is usually the motivation behind standard NMF applications.

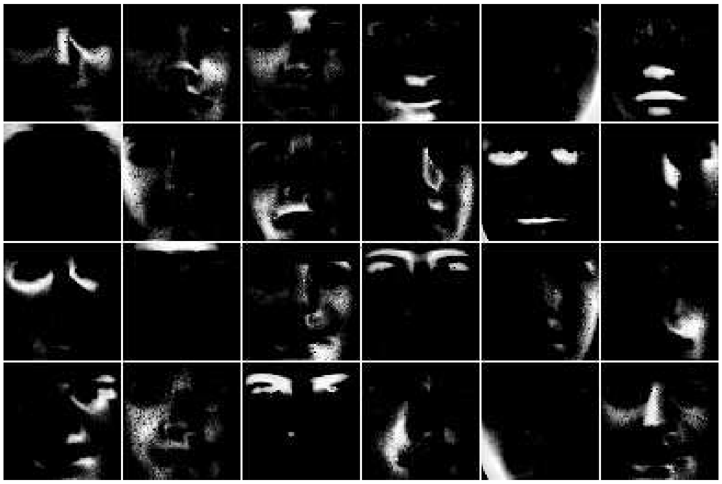


Figure 2. *Local NMF Algorithm: basis faces ($r = 24$).*

4 NMF and AVIRIS Data

We now introduce a data set intended to illustrate the robustness of NMF techniques. It is our hope that the reader be inspired by this application of NMF to a physical data set and to emphasize that the suggested methods not be limited to text and image data.

The AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) first began operations aboard a NASA research craft in 1987. It has since become a standard instrument used for remote sensing of the Earth, collecting spectral radiance data for characterization of the Earth's surface and atmosphere. AVIRIS data is often used in the fields of oceanography, environmental science, snow hydrology, geology, volcanology, soil and land management, atmospheric and aerosol studies, agriculture, and limnology. For each location flown over, reflectance data (used for quantitative characterization of surface features) for 224 contiguous spectral channels (bands) is recorded. As a result, a 3-dimensional cube of data is generated for each area analyzed by AVIRIS.[1]

One area that has received extensive study (see, for example [2] or [3]) with AVIRIS flights is Cuprite, NV. Throughout this location are scattered several small orebodies and the various geological and mineral features of the area are well documented from groundmapping. In this paper, we will consider the 200×200 pixel by pixel square region near Cuprite in Figure 4 (a).

Some preprocessing of the data needed to be performed for the methods described in this paper. First, we have followed standard AVIRIS techniques in removing the bands that correspond to the regions where water vapor is so strongly absorbing that few photons return to the AVIRIS sensor (quantified by a very low signal-to-noise ratio) [3]. Some additional atmospheric noise may still be present for the remaining bands, as is shown in the sample (preprocessed) profile in Figure 3 (a). Including all 40,000 pixel locations results in a 198×40000 band by pixel matrix $\mathbf{A}$ containing non-negative reflectance data.[2]

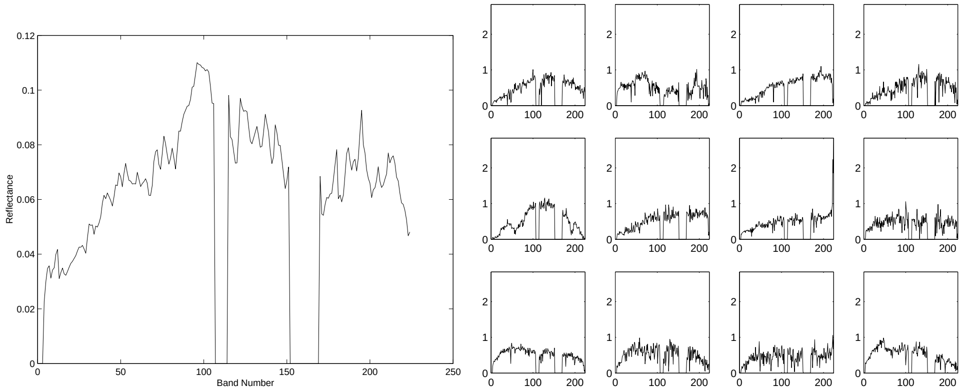
4.1 Feature Extraction Using Random NMF Initialization

Due to the limitations of AVIRIS imagery, each location (pixel) actually consists of a roughly 20 meter by 20 meter square. It is our goal to improve this resolution by doing "sub-pixel" feature extraction. Based on a location's spectral profile, we would like to determine what primary physical components exist within the 400 square meter area that the profile represents. To determine these components we will perform NMF on the 40,000 locations in the data set $\mathbf{A}$. Using this many locations, we hope to obtain the $r = 12$ components that could best be *added* together to reconstruct each location's profile as closely as possible.

Using the standard random initializations for $\mathbf{W}$ and $\mathbf{H}$, 300 iterations of the Euclidean distance NMF algorithm were performed to obtain the spectral basis in

[1] The authors are grateful to B. Kindel, E. Johnson and A.F.H. Goetz from the *Center for the Study of Earth from Space* at the Univ. of Colorado for valuable conversations on AVIRIS data.

[2] We emphasize that reflectance is a continuous function of wavelength but that each feature (band number) corresponds to a discrete sampling of a particular location's spectral profile.



(a) Sample preprocessed spectral profile.

(b) NMF basis (random initialization).

Figure 3.

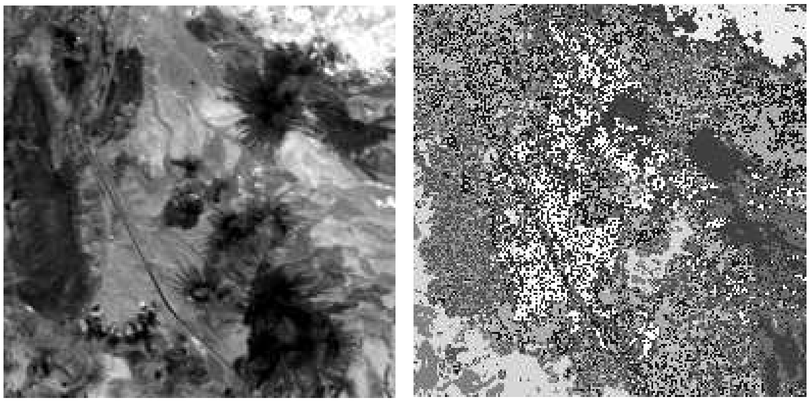
Figure 3 (b). While the 12 basis profiles shown here retain the non-negativity of the reflectance data, they no longer maintain any of the spectral continuity that we hoped to obtain. This is not to say that the basis profiles found will not sum up to reconstruct the original profiles. The basis profiles found are just significantly corrupted with noise and do not maintain the clear structure that would allow us to determine if they correspond to other known surfaces (i.e. sand, vegetation, etc.).

4.2 Feature Extraction Using Centroid NMF Initialization

In an effort to remove some of this noise, we now apply our Spherical K-Means initialization proposed in [10] and [11] to the data set $\mathbf{A}$.[3] Figure 4 (b) graphically shows the clusterings obtained. In this figure, we are confined to a palette of 12 colors, each corresponding to a different cluster – each of the 40,000 pixels have the shading that represents the cluster they belong to. For example, all pixels that are nearly white (like those in the body near the upper right hand corner of Figure 4 (b)) were determined to be most similar by Spherical K-Means. Compare this picture with the one in Figure 4 (a) which corresponds to the image of the data set $\mathbf{A}$ for one particular spectral band. Here one can clearly see that some of the original structures, such as large bodies of ore and the roadbed running through the valley in the center of the image, were returned as a result of clustering.

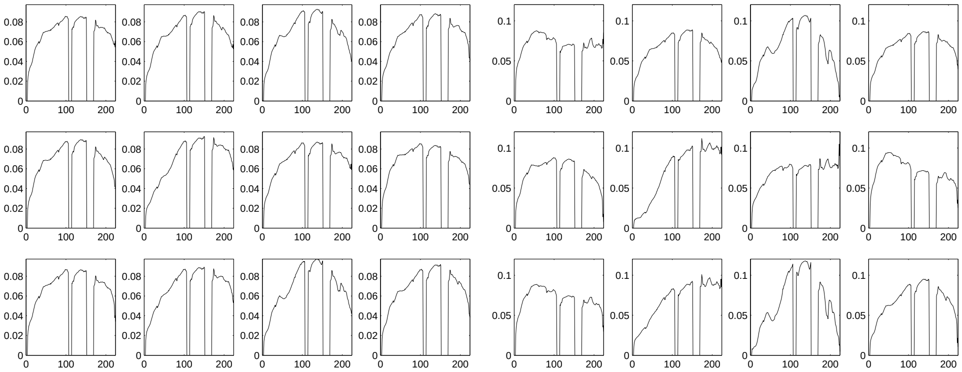
Figure 5 (a) shows the 12 centroids that result after approximately 50 iterations of Spherical K-Means clustering. Since these 12 profiles represent "average" profiles representing similar locations, we do expect to see continuity maintained. Further, these profiles are much smoother than even the original AVIRIS spectral

[3] It is not the intent of this paper to provide a theoretical foundation for structured initialization or Spherical K-Means techniques. The interested reader is referred to [10] and [11].



(a) Original (One band) (b) Spherical K-Means clustering

Figure 4. *AVIRIS data set representations (*$r = k = 12$*).*



(a) Spherical K-Means centroids. (b) NMF basis profiles.

Figure 5. *Centroid initialization (a), and the resulting NMF basis (b).*

profiles – in grouping similar profiles together and taking an average we have substantially reduced the noise inherent in AVIRIS data.

We will again use the Euclidean distance NMF algorithm with $r = 12$ and the same random initialization for $\mathbf{H}$. However, we will now use the centroid profiles shown here to seed the basis $\mathbf{W}$. Using this Spherical K-Means initialization, 300 iterations of the Euclidean distance NMF algorithm yielded the basis profiles shown in Figure 5 (b). These basis profiles should be compared to those obtained using a random initialization of $\mathbf{W}$ in Figure 3. These new profiles are much smoother (preserve continuity better) than those obtained earlier. The reader should also remark that these profiles are not the same as the centroid profiles of Figure 5. The profiles in Figure 5 (b) may resemble their associated centroids because they were originally initialized with them, however, the two sets of profiles differ by their emphasis of certain features (marked by peaks and valleys in the spectral profiles).

4.3 Unmixing and Related Work

We now emphasize how the initialization-factorization approach with Spherical K-Means and NMF done here differs from previous work done on AVIRIS images. With the increasingly wide spread use of AVIRIS data several researchers have begun to actively look into the problem of "unmixing" – overcoming the limited AVIRIS pixel resolution by determining what smaller sub-elements make up a pixel. In the literature (again, see [2] and [3]), the spectral profiles of these sub-elements are called "endmembers" and they are usually chosen from a library of known profiles taken from ground samples of different types of vegetation, minerals, metals, etc. For each endmember, the resulting abundance plot is generated, where each pixel of the plot signifies the abundance of that endmember in the pixel. The abundance of an endmember is necessarily non-negative. Consequently, the non-negative least squares (NNLS) technique in [5] offers a very convenient way of obtaining the best non-negative least squares solution y to $\|Ey - x\|^2$, where E is the library matrix whose columns are each endmembers.

Our work presents a new approach to this problem because we do not assume that a library of endmembers is already in place. Ours is an unsupervised method of actually deriving the r endmembers that best represent the given data when added together (with unequal weights). The resulting endmembers are stored in the basis matrix $\mathbf{W}$ and we do not require additional NNLS computations because each abundance plot is already stored in the rows of the coefficient matrix $\mathbf{H}$. Further, both initializations discussed above essentially brought the rank 196 matrix $\mathbf{A}$ down to a rank 12 factorization with a relative Frobenius error of less than 2.5%.

One shortcoming of our method is that the columns of $\mathbf{H}$ do not sum to 1. Ideally, we would like the i-th element of column h_j to correspond to the percentage of pixel j that is made up of endmember i. Currently, each element in a column of $\mathbf{H}$ corresponds to the abundance of endmember i when viewed relative to all other endmembers. To illustrate the concept of an abundance plot, we have shown the left factor $\mathbf{H}$ that resulted from 300 NMF iterations in Figure 6.

At first glance, some of these 12 abundance plots (each corresponding to the abundance of one of the 12 endmembers) may appear to be somewhat random. However, these abundance plots correspond to endmembers which are found throughout the image and resemble other endmembers associated with similar abundance plots. Further, these plots are distinct because of the ore body in the lower left which contains (or lacks) these endmembers. Even more interesting is the appearance of features such as the roadbed through the center of the image. For example, in looking at the second row of Figure 6, we see that the first endmember in the row shows the roadbed with dark pixels while the last endmember in the row shows the roadbed with light pixels. Here we may conclude that the roadbed is significantly made up of the last endmember while lacking in the first.

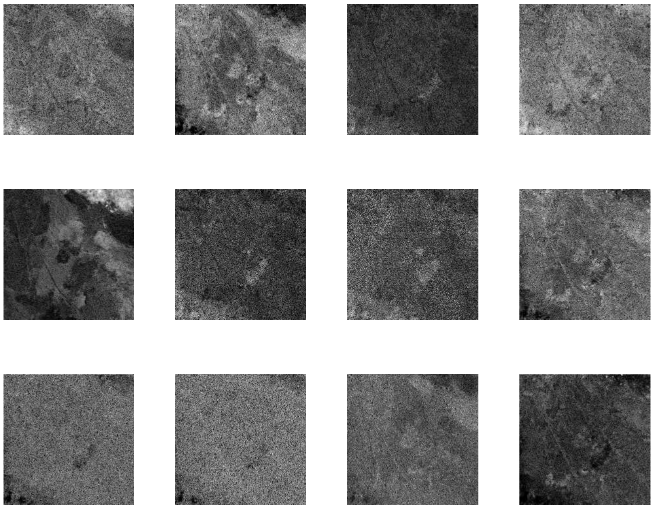


Figure 6. *NMF coefficient matrix* ***H****.*

5 Summary and Future Work

In this paper we have shown preliminary results for computing the (specially seeded) Non-Negative Matrix Factorization on AVIRIS remote sensing data. We have found a basis whose profiles emphasize local spectral features and use these profiles to "unmix" the low resolution data. We emphasize here that the spectral profiles (endmembers) that we have obtained may not make complete sense to the reader. In general, we are more familiar with human faces and so we can recognizes noses and ears when they appear in a face basis as in Figure 2. However, an expert in remote sensing could likewise immediately identify the features corresponding with the spectral profile of water and could consequently make some skilled interpretation of the basis obtained. In this way, efficient Non-negative Matrix Factorization techniques may be applied to data from a variety of fields and yield both a basis and a coefficient matrix whose interpretation must then ultimately come from an expert in that field.

Much work remains to be done in the area of Non-negative Matrix Factorizations. We have made some contributions, especially regarding structured initialization of the factors **W** and **H**, in [10] and [11]. However, in Section 2 we noted that a particular factorization is determined by the objective function which it seeks to minimize (maximize). Further, each objective function may have many iterative update strategies which monotonically decrease its value, and some updates will be inherently better (in convergence, overcoming local extrema, etc.) than others. The application of NMF techniques to non-traditional types of data underscores the problem of determining a "suitable" objective function constructed specifically for that type of data. The continued discovery and exploration of these objective functions will sufficiently refine NMF techniques for their application in practice.

Bibliography

[1] M.W. Berry and M. Browne, *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, SIAM Press (1999).

[2] J.W. Boardman, F.A. Kruse and R.O. Green, *Mapping Target Signatures Via Partial Unmixing of AVIRIS Data*, In AVIRIS Airborne Geoscience Workshop Proceedings (1995).

[3] A.F.H. Goetz and B. Kindel, *Comparison of Unmixing Results Derived from AVIRIS, High and Low Resolution, and HYDICE Images at Cuprite, NV*, In AVIRIS Airborne Geoscience Workshop Proceedings (1999).

[4] T. Kawamoto, K. Hotta, T. Mishima, J. Fujiki, M. Tanaka and T. Kurita, *Estimation of Single Tones from Chord Sounds Using Non-Negative Matrix Factorization*, Neural Network World, Vol. 3, (July 2000), pp. 429–436.

[5] C.L. Lawson and R.J. Hanson, *Solving Least Squares Problems*, Prentice–Hall, Englewood Cliffs (1974).

[6] D.D. Lee and H.S. Seung, *Learning the parts of objects by Non-Negative Matrix Factorization*, Nature, Vol. 401 (October 1999).

[7] D.D. Lee and H.S. Seung, *Algorithms for Non-negative Matrix Factorization*, In Advances in Neural Information Processing Systems 13, (2000).

[8] S.Z. Li, X.W. Hou and H.J. Zhang, *Learning Spatially Localized, Parts-Based Representation*, In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, (December 2001).

[9] L.K. Saul and D.D. Lee, *Multiplicative Updates for Classification by Mixture Models*, In Advances in Neural Information Processing Systems 14 (2002).

[10] S.M. Wild, *Seeding Non-Negative Matrix Factorizations with the Spherical K-Means Clustering*, Thesis for the Department of Applied Mathematics, University of Colorado (April 2003).

[11] S.M. Wild, J.H. Curry and A.M. Dougherty, *Improving Non-Negative Matrix Factorizations Through Structured Initialization*. In preparation (2003).