Sparse Endmembers and Demixing

Matthew J. Hirn

Department of Mathematics and Applied Mathematics
Yale University

Applied Mathematics Seminar

Yale University

October 6, 2009

Acknowledgement

This talk is based on joint work with **Martin Ehler**, who is currently at the University of Maryland, College Park and the National Institutes of Health.

Figure: Martin Ehler

Outline

Outline

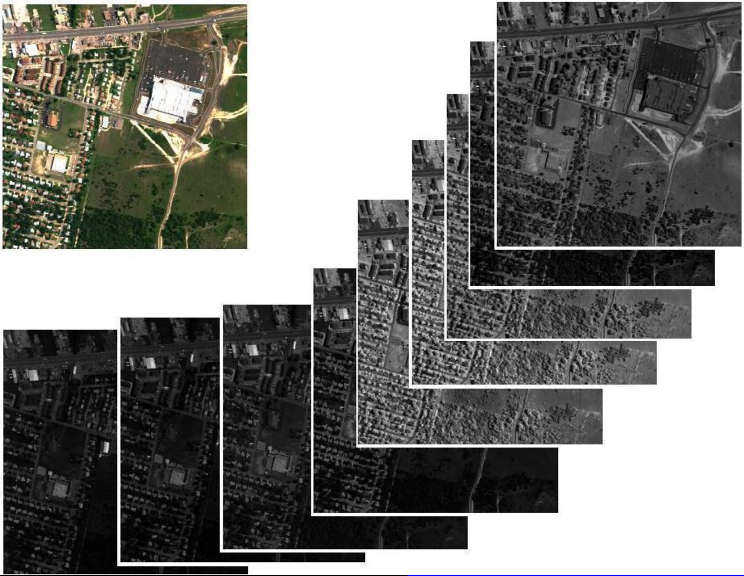
Color image

Red

Blue

Green

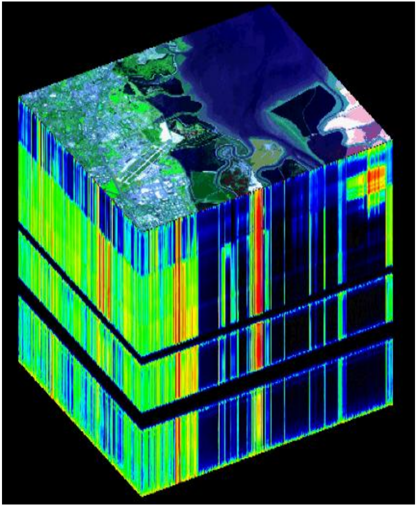
Hyperspectral imagery data



Hyperspectral camera in action

Figure: http://www.diamond-sensing.com/uploads/media/Hyperspectral.jpg

Hyperspectral data cube



Overview of hyperspectral imagery data

- Hyperspectral imagery (HSI) data is characterized by the narrowness and contiguous nature of the measurements.
- HSI data sets are spectrally overdetermined, and thus provide ample spectral information to distinguish between spectrally similar (but unique) materials.
- HSI data sets can be useful for the following purposes:
  - target detection
  - material classification
  - material identification
  - mapping details of surface properties

Overview of hyperspectral imagery data

- Hyperspectral imagery (HSI) data is characterized by the narrowness and contiguous nature of the measurements.
- HSI data sets are spectrally overdetermined, and thus provide ample spectral information to distinguish between spectrally similar (but unique) materials.
- HSI data sets can be useful for the following purposes:
  - target detection
  - material classification
  - **material identification**
  - mapping details of surface properties

Notations

- Assume our HSI data set is an $n_1 \times n_2 \times d$ cube.
  - n_1, n_2 spatial dimensions.
  - $n = n_1 n_2 =$ number of pixels.
  - d is the spectral dimension (so d wavelengths measured).
- d is usually large, e.g., $d > 100$.
- n is usually very large, e.g., $n = \mathcal{O}(10^5)$ or even $n = \mathcal{O}(10^6)$.
- Let $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ denote the pixel vectors of the HSI data cube in set form.
- Let $X = [x_1 \;\; x_2 \;\; \cdots \;\; x_n]$ be a $d \times n$ matrix where the columns x_i of X are the pixel vectors of the HSI data cube.

Endmembers

Definition

Endmembers are a collection of a scene's constituent spectra. If $\mathcal{E} = \{e_i\}_{i=1}^{s} \subset \mathbb{R}^d$ are endmembers corresponding to a data set $\mathcal{X}$, then there is some representation of each $x_i \in \mathcal{X}$ in terms of the elements of $\mathcal{E}$.

- Let $E = [e_1 \;\; e_2 \;\; \cdots \;\; e_s]$ be a $d \times s$ matrix where the columns e_i of E are the endmembers.
- s is usually small, e.g., $s < d$.
- Many algorithms find the endmembers from within the data, so that $\mathcal{E} \subset \mathcal{X}$.
- One alternative is to find endmembers from a spectral library, $\mathcal{L}$, that can be used for multiple data sets.

Linear mixture model

- Given a data set $\mathcal{X}$ and corresponding endmembers $\mathcal{E}$, the linear mixture model states that:

$$x_i = \sum_{j=1}^{s} \alpha_{i,j} e_j + z_i, \quad \text{for all } x_i \in \mathcal{X}.$$

- $\alpha_{i,j} \geq 0$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, s$.
- $\sum_{j=1}^{s} \alpha_{i,j} = 1$ for all $i = 1, \ldots, n$.
- $z_i \in \mathbb{R}^d$ is a noise vector.

Visualization of the linear mixture model

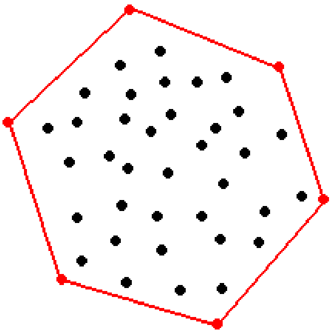


Figure: The linear mixture model

Examples of endmember algorithms

Some endmember algorithms are the following:

- N-FINDR [M. Winter]: Finds the simplex of maximal volume that contains the data set $\mathcal{X}$; the vertices of this simplex are the endmembers.
- SVDD [D. Tax and R. Duin]: Obtains a spherical shaped boundary around the data set $\mathcal{X}$; support vectors, or endmembers, are derived from this description.
- Pixel Purity Index [J. Boardman]: Repeatedly projects d–dimensional scatter plots onto a random unit vector; the extreme pixels in each projection are recorded and the total number of times each pixel is marked as extreme is noted.

Endmember coefficients

- After one finds an endmember set $\mathcal{E}$, the coefficients $\{\alpha_{i,j}\}_{i,j=1}^{n,s}$ must be computed.
- Two common ways of computing the coefficients are the following:

1. Minimum error:
$$\alpha_{i,\cdot} = \arg\min_{\tilde{\alpha}} \|x_i - E\tilde{\alpha}\|_{\ell^2}$$
2. Sparse: let $\tau_i > 0$,
$$\alpha_{i,\cdot} = \arg\min_{\tilde{\alpha}} \|x_i - E\tilde{\alpha}\|_{\ell^2}^2 + \tau_i \|\tilde{\alpha}\|_{\ell_1}$$

- Note when solving either minimization problem, $\tilde{\alpha}$ is subject to the constraints of the linear mixture model.

A look ahead

- Even if one uses the sparse coefficient model, the endmember algorithm itself does not necessarily select the endmembers with sparsity in mind!
- The endmember algorithm presented in the next section is based on the sparse coefficient model.
- We will be searching for endmembers as a subset of $\mathcal{X}$, i.e., we assume that $\mathcal{E} \subset \mathcal{X}$.

Outline

A simplistic model

- Assume the linear mixture model, and furthermore suppose that $z_i = 0$ for all $i = 1, \ldots, n$.
- Define an $n \times n$ weight matrix, $W = (w_{i,j})$, as follows.
- Let $c : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a cost function.
- If possible,

$$w_{i,\cdot} = \arg\min_{\tilde{w}} c(\tilde{w}), \quad \text{subject to:}$$

  - $\sum_{j=1}^{n} \tilde{w}_j x_j = x_i$
  - $\tilde{w}_j \geq 0$ for all $j = 1, \ldots, n$.
  - $\sum_{j=1}^{n} \tilde{w}_j = 1$
  - $\tilde{w}_i = 0$
- Otherwise, $w_{i,\cdot} = \delta_i$.
- We can extract the endmembers from the rows of W. Namely x_i is an endmember if its corresponding weight row satisfies $w_{i,\cdot} = \delta_i$.

Observations on the previous model

- Notice that in the previous model we are representing each $x_i \in \mathcal{X}$ in terms of the dictionary, or finite frame,

$$\mathcal{X}^{(i)} = \mathcal{X} \backslash \{x_i\},$$

  subject to the constraints of the linear mixture model.
- If the endmember set $\mathcal{E} \subset \mathcal{X}$ is sparse enough in the dictionary $\mathcal{X}$ (and thus in each $\mathcal{X}^{(i)}$ as well), then we could set the cost function as

$$c(\tilde{w}) = \|\tilde{w}\|_{\ell^0} = |\text{supp}(\tilde{w})|,$$

  and expect that the support of each $w_{i,\cdot}$ lies within $\mathcal{E}$.
- Thus we could extract the endmembers from the columns of the weight matrix W as well! In particular, if $\text{supp}(w_{\cdot,i}) \neq \emptyset$, then $x_i \in \mathcal{E}$.

Another simplistic model

- In reality, the endmembers will not be quite so apparent.
- Assume only part of the linear mixture model: remove the convexity (i.e. the sum to one) constraint.
- Once again assume that $z_i = 0$ for all $i = 1, \ldots, n$.
- Suppose that $s < d \ll n$, which makes $\mathcal{E}$ sparse in $\mathcal{X}$.
- Define an $n \times n$ weight matrix $W = (w_{i,j})$ as follows:

$$w_{i,\cdot} = \arg\min_{\tilde{w}} \|\tilde{w}\|_{\ell^0}, \quad \text{subject to:}$$

  - $\sum_{j=1}^{n} \tilde{w}_j x_j = x_i$
  - $\tilde{w}_j \geq 0$ for all $j = 1, \ldots, n$.
  - $\tilde{w}_i = 0$

Observations on the second model

- Notice that even the endmembers $e_i \in \mathcal{E}$ will have such a representation in the dictionary $\mathcal{X} \backslash \{e_i\}$.
- However, this representation will be a misrepresentation!
- For each $x_i \notin \mathcal{E}$ though, the support of $w_{i,\cdot}$ will be contained in $\mathcal{E}$.
- Thus if the weight of the 'good' representations outweighs the 'bad' representations, then we will still extract the endmembers $\mathcal{E}$ from the columns of W.
- In particular, we know that for each $x_i \notin \mathcal{E}$, we have $\|w_{\cdot,i}\|_{\ell^0} \leq s$.
- For $x_i \in \mathcal{E}$ though, we know that $\|w_{\cdot,i}\|_{\ell^0} \leq n - s$.
- Also, due to the fact that $s \ll n$, we almost certainly have $\|w_{\cdot,i}\|_{\ell^0} > s$.
- Thus we could extract the endmembers by selecting the $x_i \in \mathcal{X}$ corresponding to the largest $\|w_{\cdot,i}\|_{\ell^0}$.

Synthetic weight matrix

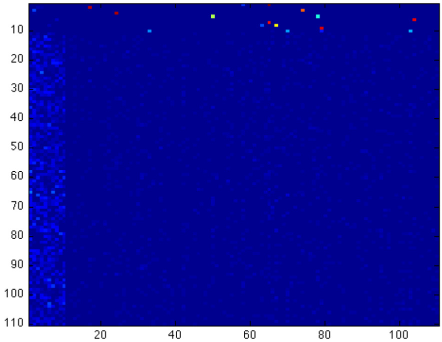


Figure: Weight matrix of synthetic data set

Adding back in the noise

- In practice of course, $z_i \neq 0$.
- Furthermore, the ℓ^0 pseudo–norm is computationally intensive, and so we turn to the ℓ^1 norm.
- In order to account for these two issues, we define our $n \times n$ weight matrix $W = (w_{i,j})$ as follows:

$$w_{i,\cdot} = \arg\min_{\tilde{w}} \|\tilde{w}\|_{\ell^1}, \quad \text{subject to:} \tag{1}$$

  - $\|x_i - X\tilde{w}\|_{\ell^2} \leq \delta_i$,
  - $\tilde{w}_j \geq 0$ for all $j = 1, \ldots, n$,
  - $\tilde{w}_i = 0$.
- Note that (1) can be replaced with:

$$w_{i,\cdot} = \arg\min_{\tilde{w}} \|x_i - X\tilde{w}\|_{\ell^2}^2 + \lambda_i \|\tilde{w}\|_{\ell^1}, \quad \text{subject to:} \tag{2}$$

  - $\tilde{w}_j \geq 0$ for all $j = 1, \ldots, n$,
  - $\tilde{w}_i = 0$.

Comments on sparse endmember extraction

- We use the second formulation of the noisy minimization, usually setting
$$\lambda_i = (.01) \cdot (X^{(i)})^t x_i.$$
- λ_i controls the density of the weight matrix. Large values of λ_i will give less vectors in $\mathcal{X}$ as possible endmembers.
- In practice, if the non-negativity constraint is not enforced, the percentage of non-negative weights is around 0.01%. Therefore this constraint can usually be removed in order to speed up run time.
- There are (at least) two questions:

1. Are there any theoretical underpinnings to this approach?
2. How do we pick out the endmembers from W? In other words, which columns of W are the most significant?

Deterministic results

- Let Φ be a $d \times n$ dictionary.
- Let $x_0 \in \mathbb{R}^d$ be a signal that has a sparse representation $\alpha_0 \in \mathbb{R}^n$ in Φ, i.e. $x_0 = \Phi\alpha_0$.
- Suppose all we have observed though is $x = x_0 + z$, where $z \in \mathbb{R}^d$ is noise vector satisfying $\|z\|_{\ell^2} \leq \varepsilon$.
- Define $\hat{\alpha}_{\delta,\varepsilon}$ as

$$\hat{\alpha}_{\delta,\varepsilon} = \arg\min_{\tilde{\alpha}} \|\tilde{\alpha}\|_{\ell^1} \text{ subject to } \|x - \Phi\tilde{\alpha}\|_{\ell^2} \leq \delta.$$

Deterministic results continued

Theorem (Donoho, Elad, Temlyakov)

If Φ and α_0 satisfy certain sparsity conditions, then

$$\|\hat{\alpha}_{\delta,\varepsilon} - \alpha_0\|_{\ell^2} \leq C \cdot (\varepsilon + \delta).$$

Theorem (Donoho, Elad, Temlyakov)

If we exaggerate the noise level by setting $\delta = C' \cdot \varepsilon$, where C' is a particular constant dependent on Φ and α_0, then

$$\mathrm{supp}(\hat{\alpha}_{\delta,\varepsilon}) \subset \mathrm{supp}(\alpha_0).$$

Probabilistic results

- Again let Φ be a $d \times n$ dictionary.
- Let $x \in \mathbb{R}^d$ be our observed signal such that $x \approx \Phi\alpha$.
- Define $\hat{\alpha}_\varepsilon$ as

$$\hat{\alpha}_\varepsilon = \arg\min_{\tilde{\alpha}} \|\tilde{\alpha}\|_{\ell^1} \text{ subject to } \|x - \Phi\tilde{\alpha}\|_{\ell^2} \leq \varepsilon.$$

Theorem (Donoho)

There exists $\rho > 0$ and $C > 0$ so that for all large d, the overwhelming majority of all $d \times n$ matrices Φ have the following property: For each vector x admitting an approximation $\|x - \Phi\alpha_0\|_{\ell^2} \leq \varepsilon$, by some vector α_0 obeying $\|\alpha_0\|_{\ell^0} < \rho d$, then

$$\|\hat{\alpha}_\varepsilon - \alpha_0\|_{\ell^2} \leq C \cdot \varepsilon.$$

Endmember selection

Given the weight matrix W, we select the endmembers $\mathcal{E} \subset \mathcal{X}$ according to two criterion on the columns of W.

1. Support size - should be large.
2. Intensity per weight - should also be large.

The exact method of selection

We rank the columns of W according to the two criterion.

1. First sort the columns of W according to their ℓ^0 pseudo-norm, $\|w_{\cdot,i}\|_{\ell^0}$. The larger the support, the better the rank (i.e., the more important that column is).
   - Columns with empty support are automatically discarded at this step.
2. Similarly, sort the columns according to the value of $\|w_{\cdot,i}\|_{\ell^2} / \|w_{\cdot,i}\|_{\ell^0}$. The larger the intensity per weight, the better the rank in this ordering.
3. Combine the two rankings to form a final ordering on the columns of W. The s highest ranked columns in this ordering correspond to the s endmembers in $\mathcal{E}$.

Outline

Small subset of Urban

Figure: Small subset of Urban

- 50×50 pixels.
- 161 spectral dimensions.

Weight matrix of the Urban subset

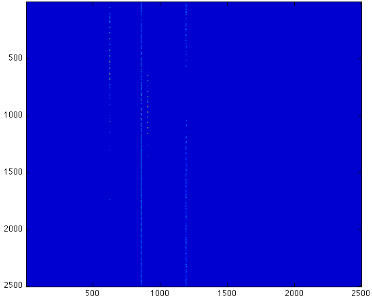
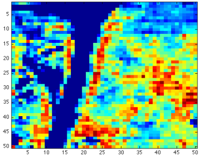


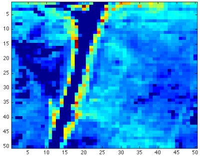
Figure: Weight matrix of the Urban subset

- Number of nonzero columns: 31.
- Number of columns $w_{\cdot,i}$ such that $\|w_{\cdot,i}\|_{\ell^0} \geq 10$: 18.

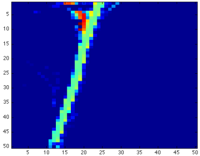
Weight columns of the Urban subset



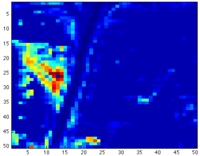
(a) W column 1



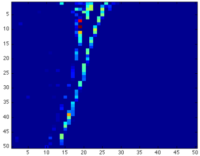
(b) W column 2



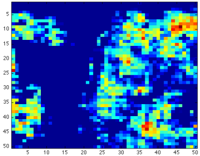
(c) W column 3



(d) W column 4

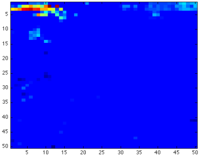


(e) W column 5

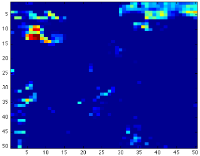


(f) W column 6

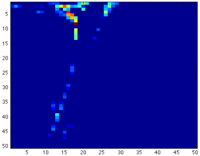
Weight columns of the Urban subset



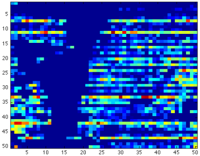
(g) W column 7



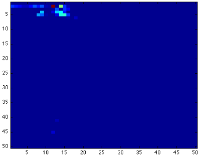
(h) W column 8



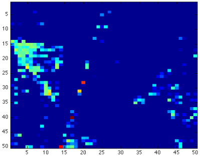
(i) W column 9



(j) W column 10



(k) W column 11



(l) W column 12

Endmembers of the Urban subset

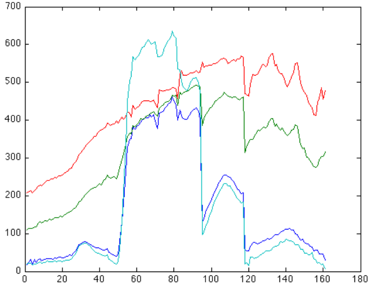
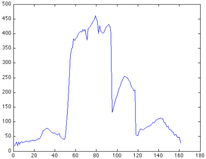
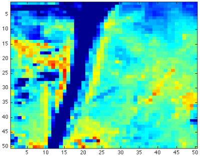


Figure: Endmembers of the Urban subset

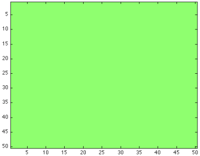
Coefficients of the Urban subset



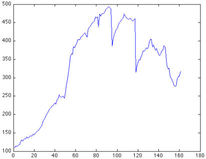
(a) Endmember 1



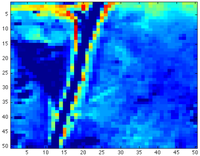
(b) ℓ^2 coefficients



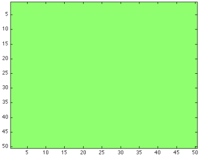
(c) $\ell^2 - \ell^1$ coefficients



(d) Endmember 2

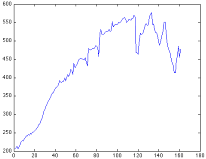


(e) ℓ^2 coefficients

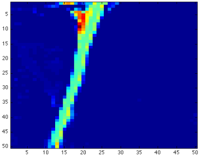


(f) $\ell^2 - \ell^1$ coefficients

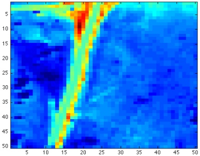
Coefficients of the Urban subset



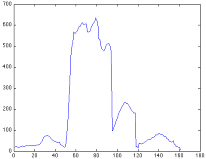
(g) Endmember 3



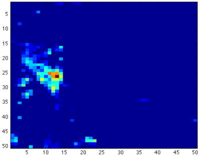
(h) ℓ^2 coefficients



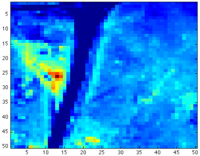
(i) $\ell^2 - \ell^1$ coefficients



(j) Endmember 4



(k) ℓ^2 coefficients



(l) $\ell^2 - \ell^1$ coefficients

Sparsity statistics for the Urban subset

The average ℓ^0 norm per pixel of each coefficient set:

- ℓ^2 coefficients: 1.9044.
- $\ell^2 - \ell^1$ coefficients: 1.9044.

Note though that the ℓ^2 coefficients though do a better job of putting different materials with different endmembers!

Large data sets

- For large data sets, e.g. $n \geq 10^4$, computing the weight matrix W may be too time intensive.
- In order to get around this problem, we sample the data set uniformly at random; call this sample $\mathcal{Y} \subset \mathcal{X}$.
- We then compute the weight matrix for $\mathcal{Y}$, and in turn select the endmembers from $\mathcal{Y}$ as well.
- The coefficients for the whole data set $\mathcal{X}$ are then computed from these endmembers.

Urban

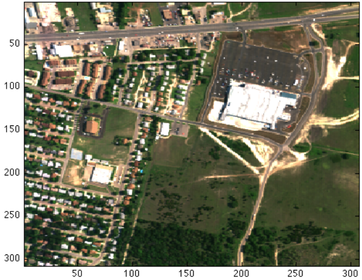


Figure: Urban: http://www.agc.army.mil/Hypercube/index.html

- 307×307 pixels.
- 161 spectral dimensions.
- Sample size: 4000 pixels

Weight matrix of Urban

The weight matrix, W, of Urban had the following statistics:

- Number of nonzero columns: 90.
- Number of columns $w_{\cdot,i}$ such that $\|w_{\cdot,i}\|_{\ell^0} \geq 10$: 50.

Endmembers of Urban

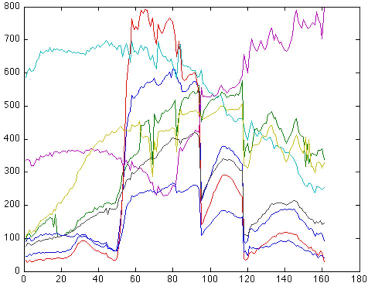
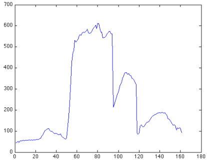
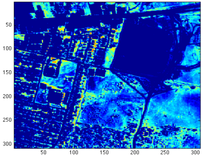


Figure: Endmembers of Urban

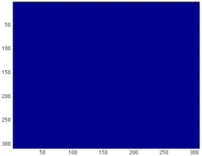
Coefficients of Urban



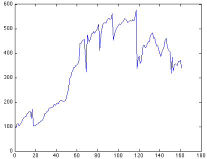
(a) Endmember 1



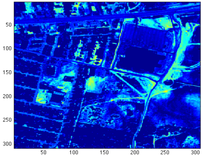
(b) ℓ^2 coefficients



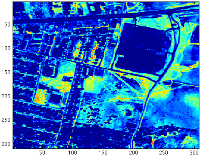
(c) $\ell^2 - \ell^1$ coefficients



(d) Endmember 2

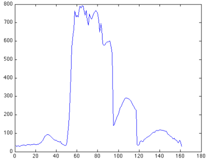


(e) ℓ^2 coefficients

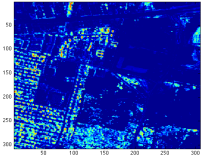


(f) $\ell^2 - \ell^1$ coefficients

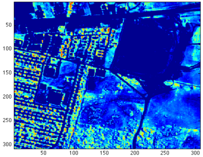
Coefficients of Urban



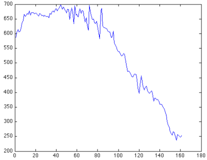
(g) Endmember 3



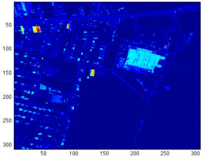
(h) ℓ^2 coefficients



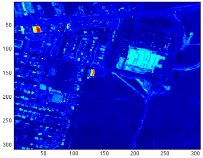
(i) $\ell^2 - \ell^1$ coefficients



(j) Endmember 4

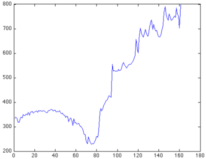


(k) ℓ^2 coefficients

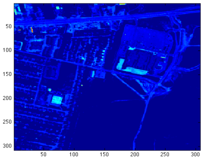


(l) $\ell^2 - \ell^1$ coefficients

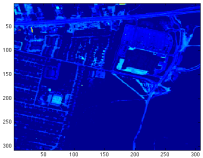
Coefficients of Urban



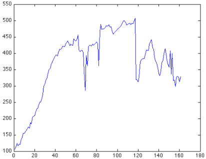
(m) Endmember 5



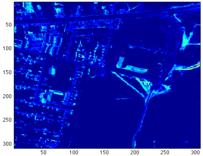
(n) ℓ^2 coefficients



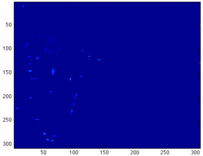
(o) $\ell^2 - \ell^1$ coefficients



(p) Endmember 6

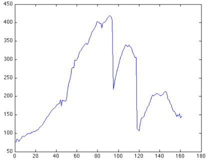


(q) ℓ^2 coefficients

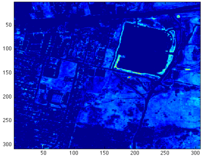


(r) $\ell^2 - \ell^1$ coefficients

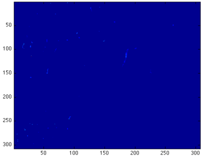
Coefficients of Urban



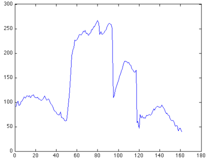
(s) Endmember 7



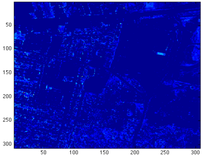
(t) ℓ^2 coefficients



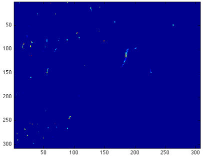
(u) $\ell^2 - \ell^1$ coefficients



(v) Endmember 8



(w) ℓ^2 coefficients



(x) $\ell^2 - \ell^1$ coefficients

Sparsity statistics for Urban

The average ℓ^0 norm per pixel of each coefficient set:

- ℓ^2 coefficients: 3.8503.
- $\ell^2 - \ell^1$ coefficients: 2.7217.

Smith

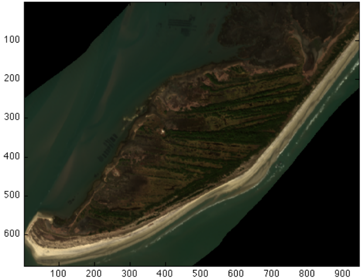


Figure: Smith

- 679×944 pixels (497182 nonzero pixels).
- 110 spectral dimensions.
- Sample size: 5000 pixels.

Weight matrix of Smith

The weight matrix, W, of Smith had the following statistics:

- Number of nonzero columns: 23.
- Number of columns $w_{\cdot,i}$ such that $\|w_{\cdot,i}\|_{\ell^0} \geq 10$: 16.

Endmembers of Smith

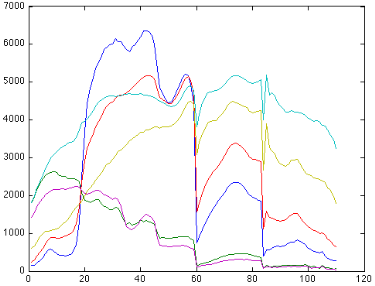
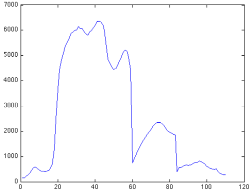
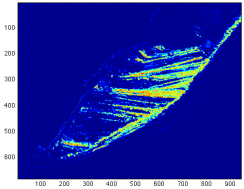


Figure: Endmembers of Smith

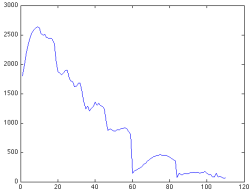
Coefficients of Smith



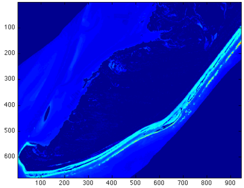
(a) Endmember 1



(b) ℓ^2 coefficients

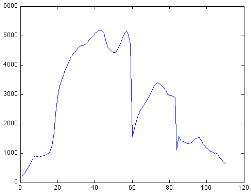


(c) Endmember 2

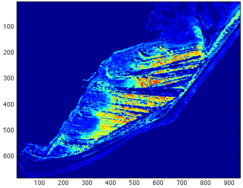


(d) ℓ^2 coefficients

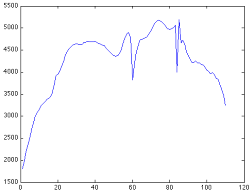
Coefficients of Smith



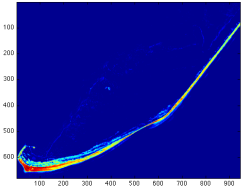
(e) Endmember 3



(f) ℓ^2 coefficients

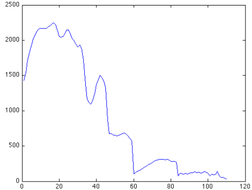


(g) Endmember 4

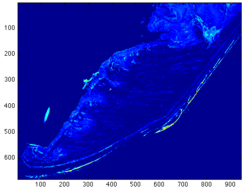


(h) ℓ^2 coefficients

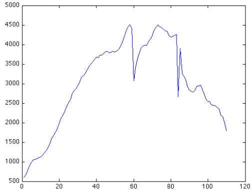
Coefficients of Smith



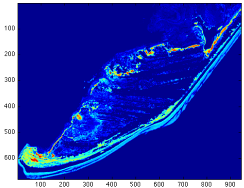
(i) Endmember 5



(j) ℓ^2 coefficients



(k) Endmember 6



(l) ℓ^2 coefficients

Sparsity statistics for Smith

- The average ℓ^0 norm per pixel of the ℓ^2 coefficients: 2.5063.

Extension to spectral libraries

- We can easily extend this method to search for endmembers from a spectral library, $\mathcal{L}$.
- In fact, by computing the weight matrix, W, with dictionary

$$\Phi = \mathcal{X}^{(i)} \cup \mathcal{L},$$

  we can search simultaneously for endmembers from both the given data set and the spectral library.
- Note that we would only compute weights for $x_i \in \mathcal{X}$.

For the future

Things we are working on:

- Developing more intricate and realistic models in which it is possible to obtain provable results.
- Consider smarter sampling methods that ideally would:
  - Reduce randomness in the endmember output.
  - Facilitate further gains in sparsity when computing the weight matrix, W.
- Continue to run trials on both real and synthetic data sets, and in particular, branch out to biomedical imaging.
- Systematically compare with other endmember methods.

Thank you!

Thank you for your time!