A convex model for non-negative matrix factorization and dimensionality reduction on physical space

Ernie Esser

Joint work with Michael Möller, Stan Osher, Guillermo Sapiro and Jack Xin

University of California at Irvine

AI/ML Seminar

10-3-2011

*Supported by NSF DMS-0911277 and PRISM-0948247

Outline

- The general problem, our strategy and key assumptions
- Formulation of convex model for NMF
- Numerical optimization
- Application to hyperspectral images
- Comparison to existing methods
- An extended convex model
- Additional applications and future work

General NMF Problem

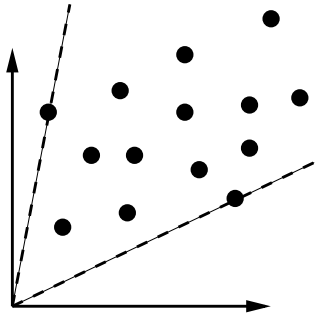
Nonnegative matrix factorization (NMF): Given nonnegative $X \in \mathbb{R}^{m \times d}$
Find nonnegative matrices $A \in \mathbb{R}^{m \times n}$ and $S \in \mathbb{R}^{n \times d}$ such that $X \approx AS$

- NMF is a very ill-posed problem
- Variational methods are typically nonconvex and involve estimating A and S alternatingly

Additional assumptions:

- Assume columns of dictionary A come from data X
- May also make additional assumptions about S (ie: sparsity)

Geometric interpretation: Find a small number of columns of X that span a cone containing most of the data



Example Application to Hyperspectral Images

Given data $X \in \mathbb{R}^{m \times d}$ (spectral signatures at each pixel)
find endmembers $A \in \mathbb{R}^{m \times n}$ and sparse abundance $S \in \mathbb{R}^{n \times d}$ such that

$$X \approx AS \qquad A \geq 0 \qquad S \geq 0$$

- $X = AS$ corresponds to the linear mixing model
- n - number of endmembers (columns of A)
- m - number of wavelengths
- d - number of pixels in image
- The non-blind demixing case is when A is known and we only need to solve for sparse abundance S
- We are considering the blind case where A must also be determined

Urban Hyperspectral Data

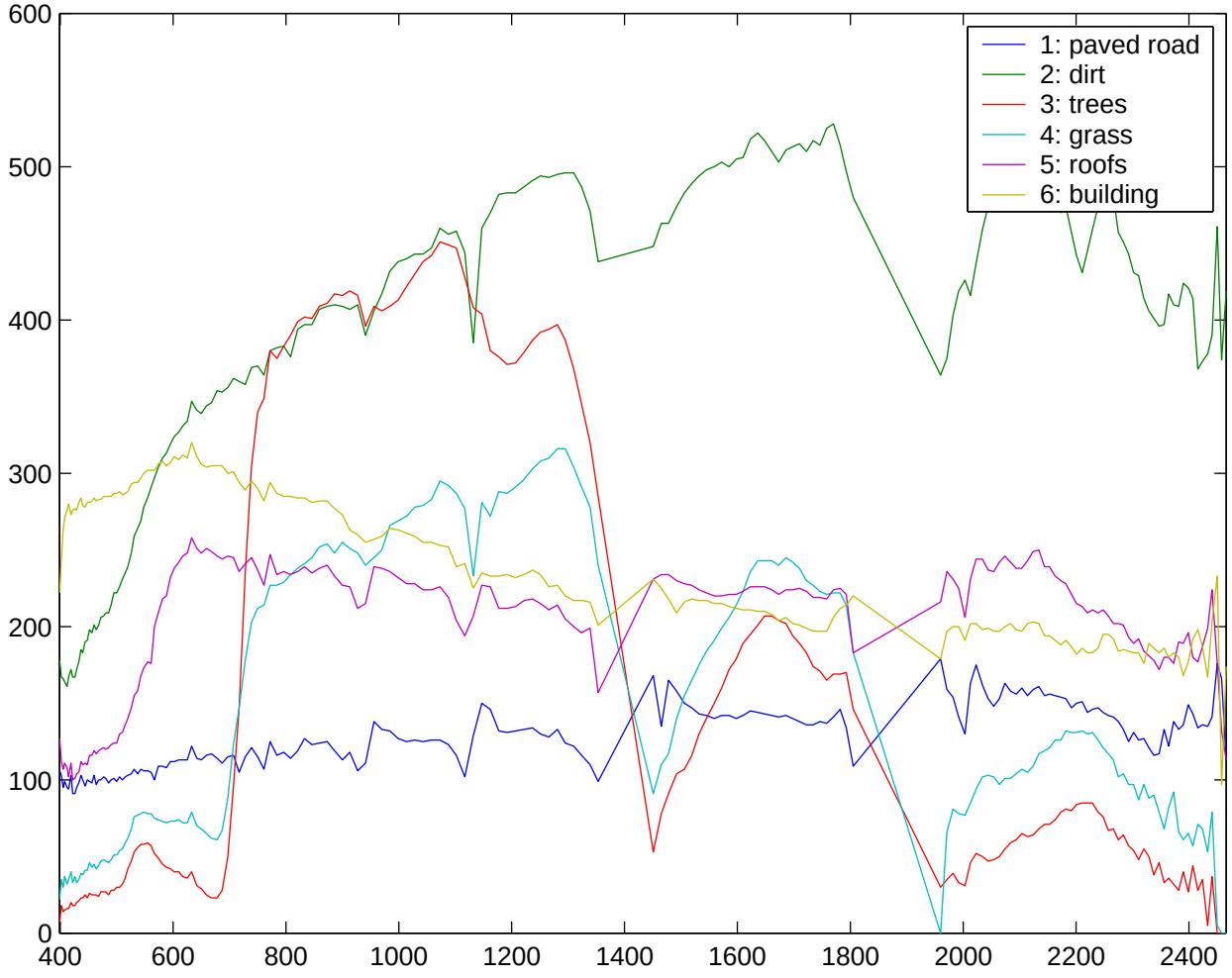
$m = 187 \qquad dr = dc = 307 \qquad d = 94249$

Thanks to Todd Wittman for the dataset.

US ARMY CORPS OF ENGINEERS, Urban hyperspectral dataset, http://www.tec.army.mil/Hypercurbe

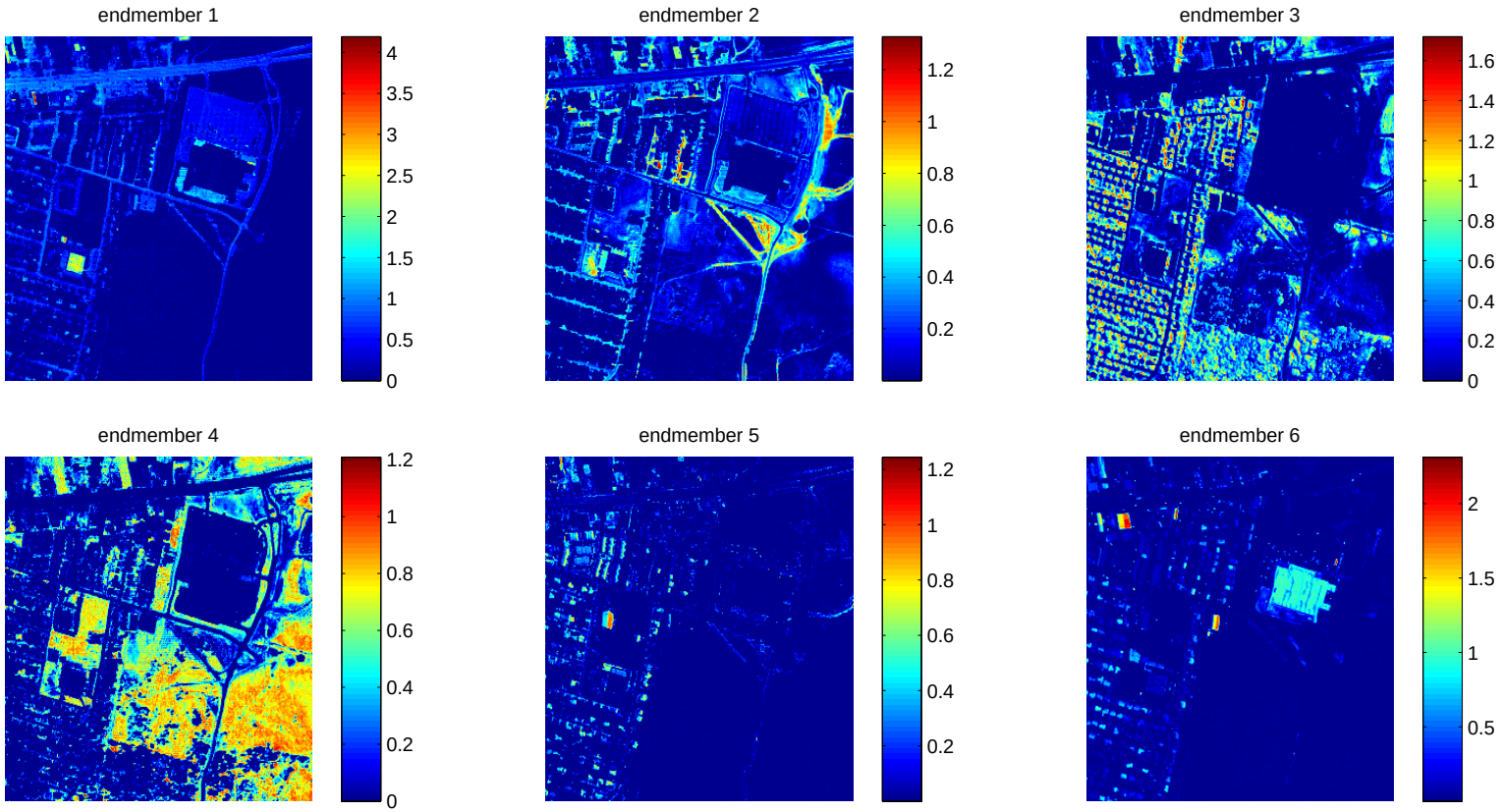
Hand Selected Endmembers

Choose 6 endmembers



Visualization of Abundance Matrix S

Rows of S



S was computed using nonnegative least squares

A. SZLAM, Z. GUO, AND S. OSHER, *A Split Bregman Method for Non-Negative Sparsity Penalized Least Squares with Applications to Hyperspectral Demixing*, 2010.

Interpretation of Dictionary Restriction

- In the context of hyperspectral images, the assumption that the endmembers can be found in the data is called the **pixel purity assumption**.
- This can more generally be interpreted as a **partial orthogonality assumption** on S
  ie: for each row i of S, there exists some column j such that $S_{i,j} > 0$ and $S_{k,j} = 0$ for $k \neq i$.
- For general NMF applications, this assumption guarantees that the columns of A are physically meaningful.

M. E. WINTER, *N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data*, 1999

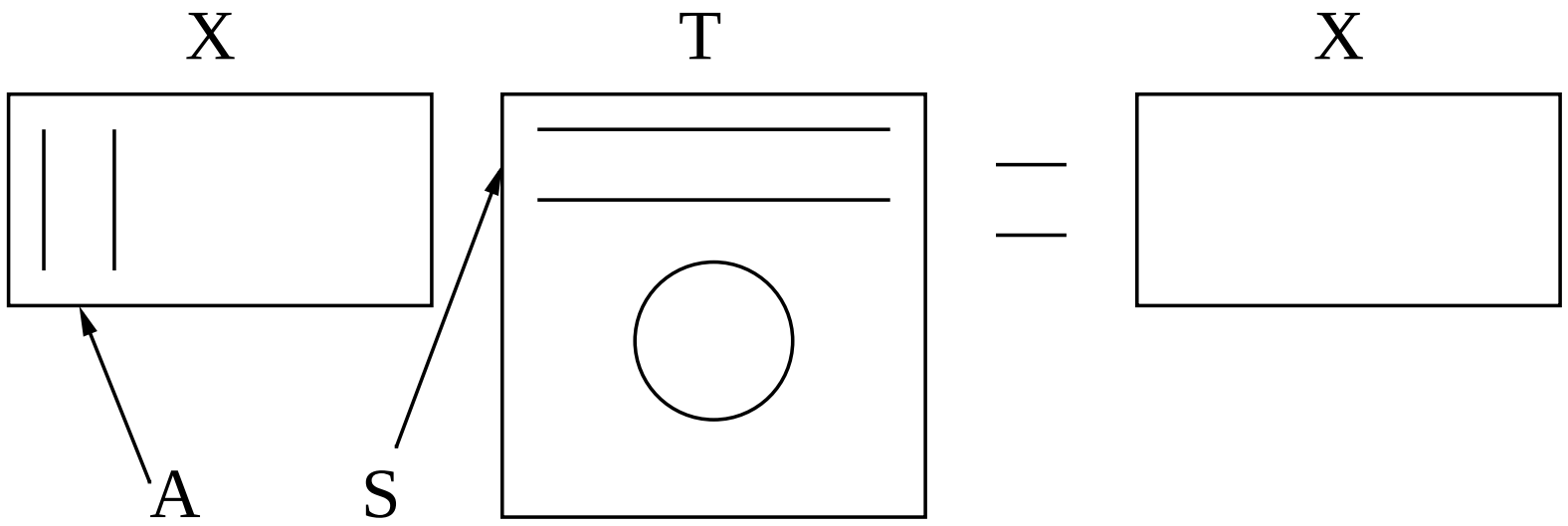
W. NAANAA AND J-M. NUZILLARD, *Blind source separation of positive and partially correlated data*, 2005

Our General Strategy

Let I index the columns of X that cannot be written as nonnegative linear combinations of the other columns. Any column X_j in X can be written as

$$X_j = \sum_{i \in I} X_i T_{i,j} \qquad \text{for } T_{i,j} \geq 0$$

Our Strategy: Find a nonnegative matrix T such that $XT = X$ and as many rows of T as possible are zero.



$XT = X$ with the index set I appearing first

Row Sparsity of T

There could be many nonnegative matrices T that satisfy $XT = X$, but we want to describe X using as few columns of X as possible. Consider instead

$$\min_{T \geq 0} \|T\|_{\text{row}-0} \qquad \text{such that} \qquad XT = X$$

where $\|T\|_{\text{row}-0}$ counts the number of nonzero rows of T.

- This can be interpreted as a minimal cone problem: finding the fewest columns of X that span a cone containing all the columns of X
- Columns of X that correspond to the nonzero rows of the minimizer T are the endmembers, or columns of A.

Methods for Finding Minimal Cone

A strategy used previously for blind source separation is to use convex programming to identify the minimal cone by looking for columns X_j of X inconsistent with

$$X_j = \sum_{q \neq j} \lambda_q X_q \qquad\qquad \lambda_q \geq 0$$

Our $XT = X$ approach will be more collaborative in the sense that it represents each columns of X as a nonnegative linear combination of the SAME small subset of columns of X.

Y. SUN AND J. XIN, *Unique solvability of under-determined sparse blind separation of nonnegative and overlapped data*, preprint, 2010.

W. NAANAA AND J-M. NUZILLARD, *Blind source separation of positive and partially correlated data*, Signal Processing 85, 2005, pp. 1711-1722.

Convex Relaxation

Two well-studied convex relaxations of $\|T\|_{\text{row}-0}$ are

- $\|T\|_{1,\infty} = \sum_i \max_j (T_{i,j})$
- $\|T\|_{1,2} = \sum_i \|T_i\|_2$, where T_i denotes the ith row of T

While both penalties can encourage row sparse matrices T, we choose to use the $l_{1,\infty}$ penalty because it is an exact relaxation under certain assumptions.

J. A. TROPP, *Algorithms for simultaneous sparse approximation: part II: Convex relaxation*, 2006.

J. MAIRAL, R. JENATTON, G. OBOZINSKI, AND F. BACH, *Network flow algorithms for structured sparsity*, 2010.

Exact Convex Relaxation

Suppose the columns of X are distinct and are normalized to have unit l_2 norm. Then the sets of minimizers of

$$\min_{T \geq 0} \|T\|_{\text{row-0}} \text{ such that } XT = X \tag{1}$$

and

$$\min_{T \geq 0} \|T\|_{1,\infty} \text{ such that } XT = X \tag{2}$$

are the same.

Proof of Exact Convex Relaxation

Recall that for any i in I, X_i can only be represented by itself. With the normalization constraint, that means $\max_j T_{i,j} = 1$ for i in I. Thus

$$\|T\|_{1,\infty} = \sum_{i=1}^{d} \max_j T_{i,j} \geq \sum_{i \in I} 1 = |I|$$

Equality is possible if and only if $T_{i,j} = 0$ for $i \neq I$. So an $l_{1,\infty}$ minimizer is a row-0 minimizer.

If $\hat{T}$ is a row-0 minimizer, then $\|\hat{T}\|_{1,\infty} = |I|$, so it is also an $l_{1,\infty}$ minimizer.

Noisy Case

If the data X is noisy, we can instead consider

$$\min_{T \geq 0} \|XT - X\|_F^2 + \alpha \|T\|_{1,\infty}$$

- For this regularization approach to be well posed, we would like the data to have sufficiently distinct columns so that we don't have noisy versions of the same endmember in the data. This will be dealt with in a data reduction step.

Data Reduction

There are two reasons to perform data reduction:

- There are too many variables in the $d \times d$ matrix T so solve for in most applications.
- We also want the endmember candidates to be sufficiently distinct for the convex relaxation to work well.

Data Reduction:

- Use clustering (we used k-means with farthest-first initialization) to cluster X and project centers onto data to get endmember candidates $Y \in \mathbb{R}^{m \times n_c}$
- Enforce angle constraint $\langle Y_i, Y_j \rangle < a_c$ to ensure distinct columns
- Optionally replace data as well by submatrix $X_s \in \mathbb{R}^{m \times d_s}$, $d_s \leq d$

The reduced problem is then to find a nonnegative row sparse $T \in \mathbb{R}^{n_c \times d_s}$ such that

$$YT \approx X_s$$

The proposed convex model for solving this will consist of two sparsity penalties and a data fidelity term.

Data Fidelity

To preserve data fidelity, we penalize

$$\frac{\beta}{2}\|(YT - X_s)C_w\|_F^2$$

where $C_w \in \mathbb{R}^{d_s \times d_s}$ is a diagonal matrix of weights reflecting the density of the original data.

For example:

- If $X_s = X$, let $C_w(j,j) = \frac{1}{d}$
- If $X_s = Y$, let $C_w(j,j) = \frac{\text{number of pixels in cluster } j}{d}$

Sparsity Penalties

In addition to a term of the form $\zeta\|T\|_{1,\infty}$ to encourage row sparse T, we also want additional sparsity of the nonzero rows of T, because we expect the data to usually be a mixture of only a few endmembers.

We use a weighted l_1 norm, which due to the nonnegativity constraint is simply a linear term:

$$\langle R_w \sigma C_w, T \rangle$$

- R_w is a diagonal matrix of row weights
  R_w is set to identity in all our experiments, but could be useful for example to encourage or discourage selection of endmembers in different regions
- $\sigma \in \mathbb{R}^{n_c \times d_s}$ has the same dimensions as T
  We want $\sigma_{i,j}$ to be small when the ith column of Y is similar to the jth column of X_s and large otherwise. We define σ according to

$$\sigma_{i,j} = \nu(1 - e^{\frac{-(1-(Y^T X_s)_{i,j})^2}{2h^2}})$$

Justification of Linear Sparsity Penalty

- In hyperspectral demixing applications, an l_1 penalty has been shown to improve sparsity of the abundance matrix S for unnormalized data.
- While clearly not helpful for l_1 normalized data, it still makes sense for l_2 normalized data, for which we expect $1 \leq \sum_i T_{i,j} \leq \sqrt{m} \; \forall \; j$. In this case the penalty will prefer data to be represented by nearby endmembers when possible.
- Having the least weight on the most similar data helps ensure a large entry in each nonzero row of T, which helps the $l_{1,\infty}$ term.

Z. Guo, T. Wittman, and S. Osher, *L1 Unmixing and its Application to Hyperspectral Image Enhancement*, 2009.

M. Berry, M. Browne, A. Langville, P. Pauca and R. Plemmons, *Algorithms and applications for approximate nonnegative matrix factorization*, 2007.

Overall Model

The overall proposed convex model is

$$\min_{T \geq 0} \zeta \sum_i \max_j (T_{i,j}) + \langle R_w \sigma C_w, T \rangle + \frac{\beta}{2} \|(YT - X_s) C_w\|_F^2$$

- Columns of Y have unit l_2 norm
- $Y \in \mathbb{R}^{m \times n_c}$
- $X_s \in \mathbb{R}^{m \times d_s}$
- $T, \sigma \in \mathbb{R}^{n_c \times d_s}$
- R_w and C_w are diagonal matrices of row and column weights

Refinement of Solution

Disadvantages of model:

- Due to convex nature, can't distinguish between identical or similar columns of Y
- Limited to selecting endmembers from Y

Advantages of model:

- Reliable when columns of Y are sufficiently distinct (true by construction)
- Despite reducing size of problem, solution already compares well to other methods
- Minimizer is an excellent initial guess for the alternating minimization approach to NMF, in which the refined endmembers can also be constrained to stay near the initialization.

Model for alternating minimization refinement:

$$\min_{A\geq 0, S\geq 0, ||A_j - \tilde{A}_j||_2 < a_j} \frac{1}{2}||AS - X||_F^2 + \langle R_w \sigma, S\rangle$$

where $\tilde{A}$ are the endmembers selected by the convex model.

Numerics for Solving Convex Model

To solve

$$\min_{T\geq 0} \zeta \sum_i \max_j (T_{i,j}) + \langle R_w \sigma C_w, T\rangle + \frac{\beta}{2}\|(YT - X_s)C_w\|_F^2$$

first introduce as new variable Z and a Lagrange multiplier P for the constraint $Z = T$.

Then use the alternating direction method of multipliers (ADMM) to find a saddle point of

$$L_\delta(T, Z, P) = g_{\geq 0}(T) + \zeta \sum_i \max_j (T_{i,j}) + \langle R_w \sigma C_w, T\rangle$$
$$+ \frac{\beta}{2}\|(YZ - X_s)C_w\|_F^2 + \langle P, Z - T\rangle + \frac{\delta}{2}\|Z - T\|_F^2,$$

where $g_{\geq 0}$ is an indicator function for the $T \geq 0$ constraint and $\delta > 0$.

Application of ADMM

Initialize T^0 and P^0 and then iterate

$$Z^{k+1} = \arg\min_{Z} \langle P^k, Z\rangle + \frac{\beta}{2}\|(YZ - X)C_w\|_F^2 + \frac{\delta}{2}\|Z - T^k\|_F^2$$

$$T^{k+1} = \arg\min_{T} g_{\geq 0}(T) + \zeta \sum_i \|T_i\|_\infty + \langle R_w \sigma C_w, T\rangle - \langle P^k, T\rangle + \frac{\delta}{2}\|T - Z^{k+1}\|_F^2$$

$$P^{k+1} = P^k + \delta(Z^{k+1} - T^{k+1})$$

- This converges for any $\delta > 0$ if a saddle point exists.
- Each minimization step is straightforward to compute.

Moreau Decomposition

Let J be a closed proper convex function on $\mathbb{R}^n$, $f \in \mathbb{R}^m$ and $A \in \mathbb{R}^{n \times m}$. Then

$$f = \arg\min_u J(Au) + \frac{1}{2\alpha}\|u - f\|_2^2 + \alpha A^T \arg\min_p J^*(p) + \frac{\alpha}{2}\|A^T p - \frac{f}{\alpha}\|_2^2$$

where the Legendre transform of J is defined by

$$J^*(p) = \sup_w \langle w, p \rangle - J(w)$$

We can use this to simply the minimization subproblem for T.

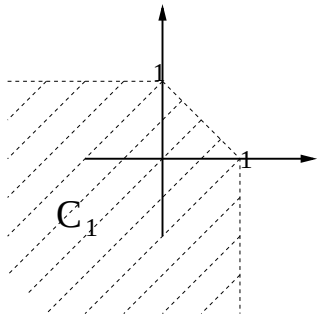
Solving for T Update

Note that the T update can be computed one row at a time.

Let $J(T_i) = g_{\geq 0}(T_i) + \zeta \sum_i \max_j(T_{i,j})$.

$$
\begin{aligned}
J^*(Q) &= \sup_{T_i} \langle Q, T_i \rangle - g_{\geq 0}(T_i) - \zeta \sum_i \max_j(T_{i,j}) \\
&= \sup_{T_i} (\max_j(T_{i,j}))(\| \max(Q, 0)\|_1 - \zeta) \\
&= \begin{cases} 0 & \text{if} \quad \| \max(Q, 0)\|_1 \leq \zeta \\ \infty & \text{otherwise} \end{cases}
\end{aligned}
$$

Let C_ζ denote the convex set $C_\zeta = \{Q \in \mathbb{R}^{d_s} : \| \max(Q, 0)\|_1 \leq \zeta\}$



The Explicit Update Formulas

Using the Moreau decomposition, we can write the T update in terms of an orthogonal projection onto the convex set

$$Z_j^{k+1} = (\beta Y^T Y C_{w_j} + \delta I)^{-1}(\beta Y^T X_j C_{w_j} + \delta T_j^k - P_j^k)$$

$$T^{k+1} = Z^{k+1} + \frac{P^k}{\delta} - \frac{R_w \sigma C_w}{\delta} - \Pi_{C_{\frac{\zeta}{\delta}}}(Z^{k+1} + \frac{P^k}{\delta} - \frac{R_w \sigma C_w}{\delta})$$

$$P^{k+1} = P^k + \delta(Z^{k+1} - T^{k+1})$$

Note: The projection for each row of the T update can be computed with complexity $O(d_s \log d_s)$

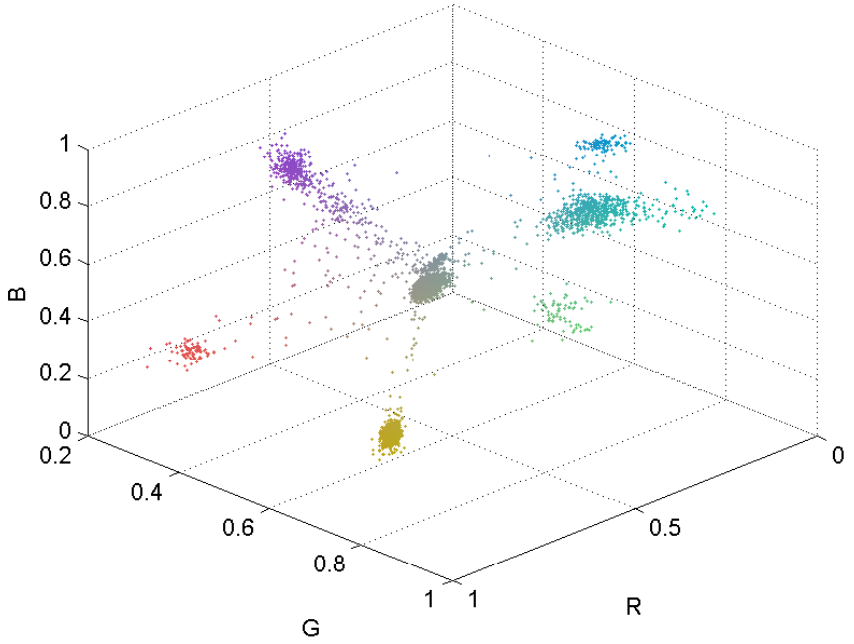
Refinement Steps

In a similar manner, ADMM can also be used to solve each of the convex subproblems involved in alternately minimizing

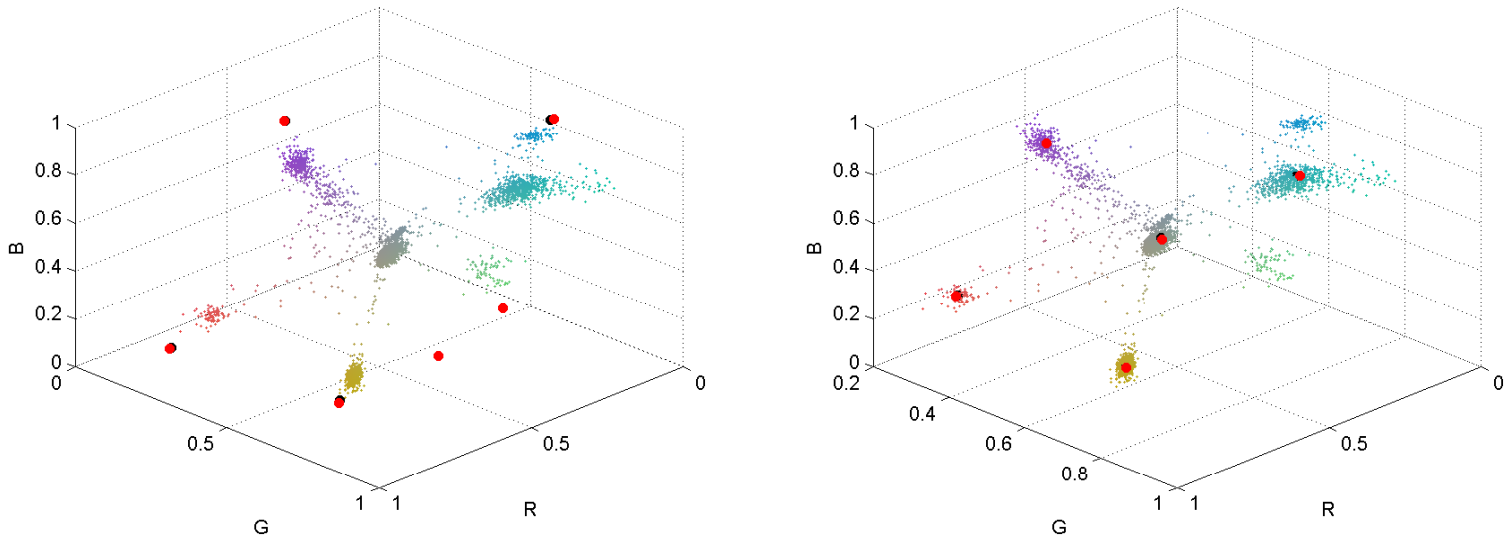
$$\min_{A\geq 0, S\geq 0, ||A_j - \tilde{A}_j||_2 < a_j} \frac{1}{2}\|AS - X||_F^2 + \langle R_w \sigma, S\rangle$$

to either determine S or refine A beyond the estimate from the convex model.

RGB Test For Visualization



Minimizers (w/o and with sparsity)



Parameters:

$$\zeta = 1$$

$$\beta = 250$$

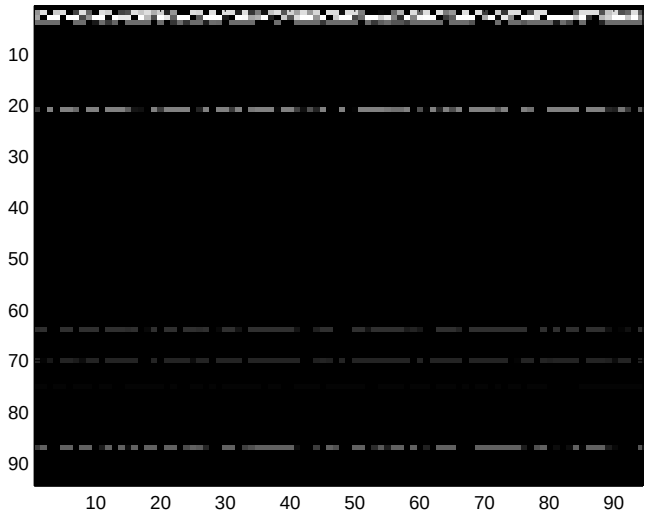
$$h = 1 - \cos(.042)$$

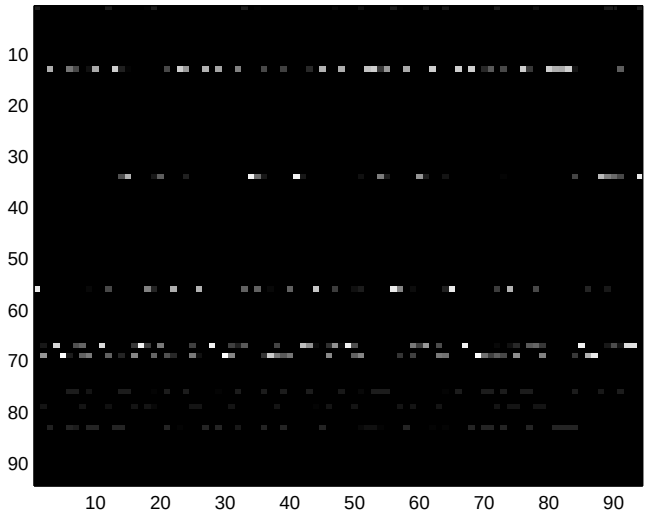
$$\nu = 50$$

$$a_c = .999$$

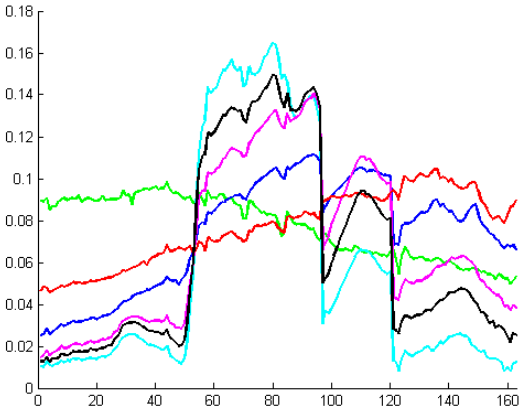
$$n_c < 150$$

Abundance Matrix Comparison

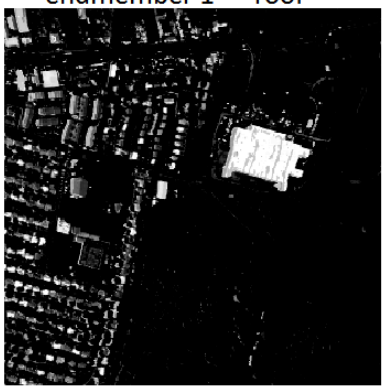




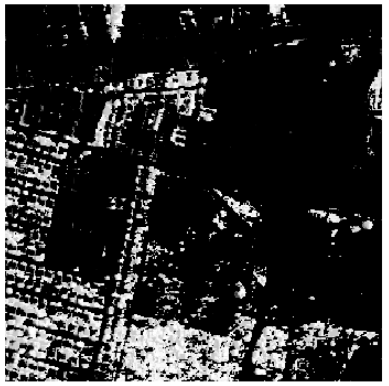
Application to Urban Hyperspectral image



endmember 1 = "roof"



endmember 2 = "trees"

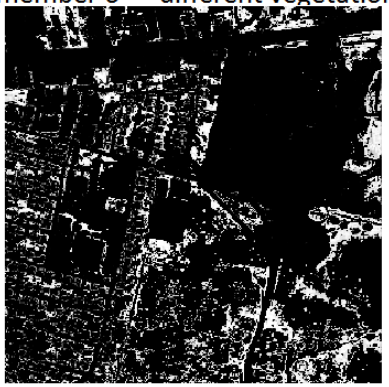


endmember 3 = "grass"

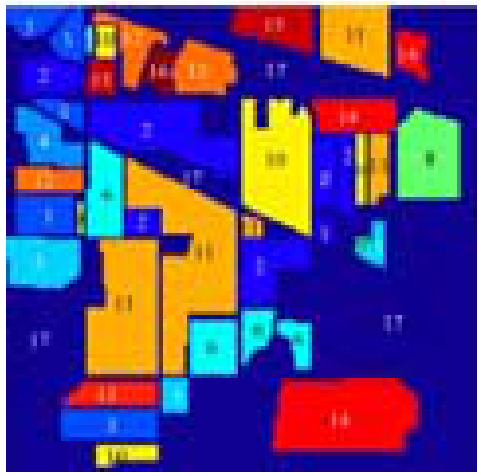
endmember 4 = "soil/dirt"

endmember 5 = "road"

endmember 6 = "different vegetation"



Supervised Endmember Detection

Indian Pines Dataset: 

Extract 9 endmembers averaged over ground truth

- 50 data points for each endmember
- 30 for each combination of 2
- 10 for each combination of 3
- 30 for mixtures of all

Add zero mean Gaussian noise with standard deviation .006 and normalize data.

Comparison of Methods

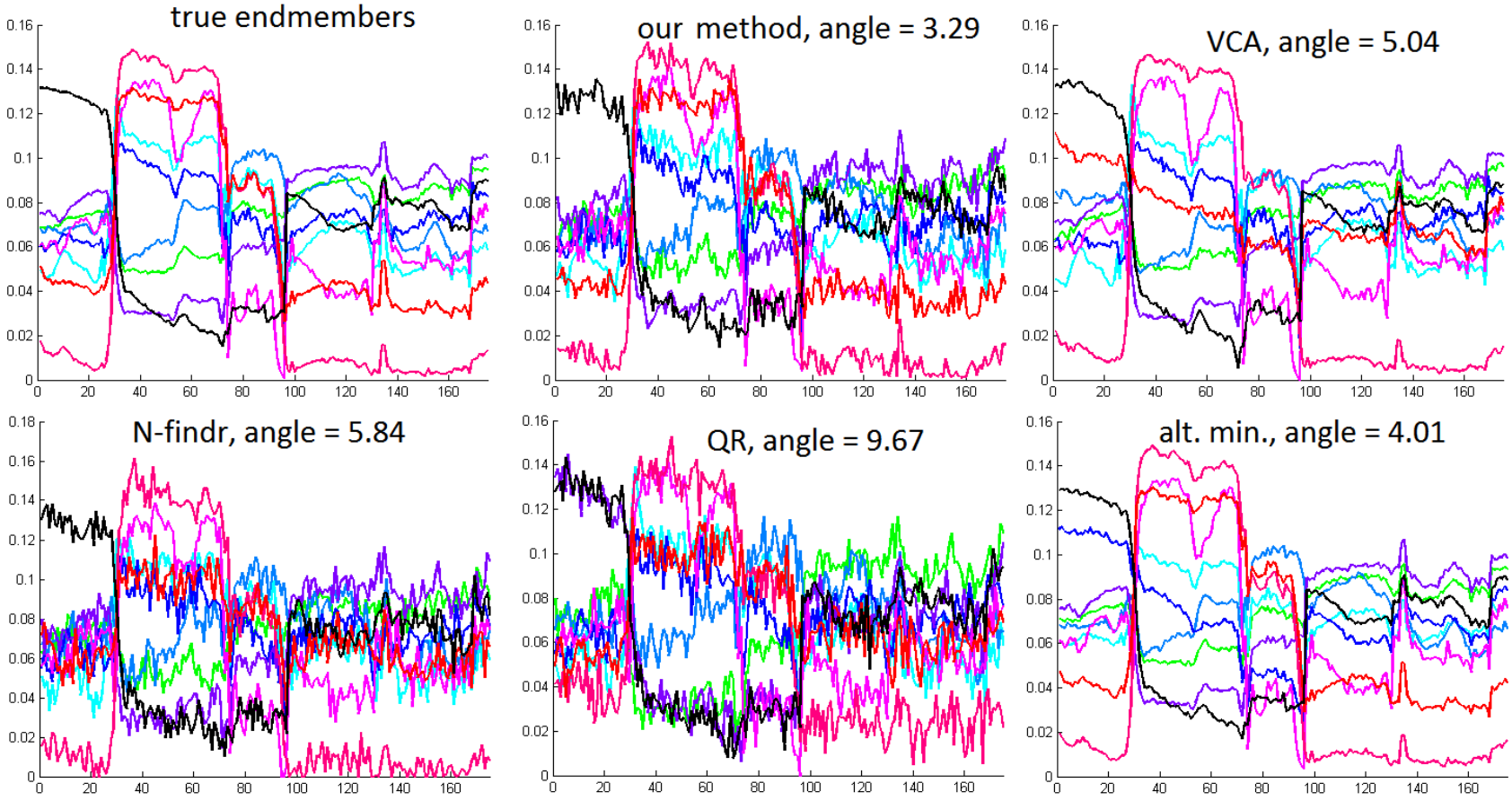
Method	**Evaluation on 15 test runs**		
	Avg. α	Min. α	Max. α
Ours refined	**3.37**	3.30	**3.42**
Ours without refinement	3.93	3.84	4.01
VCA	4.76	1.78	6.95
N-findr	10.19	7.12	13.79
QR	9.87	4.71	12.74
Alt. Min.	4.50	**1.76**	8.17

Comparison of different endmember finding methods by the angle of deviation from the true endmember vectors

J.M.P. NASCIMENTO AND J.M. BIOUCAS-DIAS, *Blind Hyperspectral Unmixing*, 2007.

T.F. CHAN AND P.C. HANSEN, *Some Applications of the Rank Revealing QR Factorization*, 1992

Comparison of Recovered Endmembers



An Extended Model

Goals of extended model:

- Take into account normalization of data to better distinguish between noise and outliers
- Allow better control over the number of selected endmembers

Instead of $\|(YT - X_s)C_w\|_F^2$, require

$$YT - X_s = V - X_s \operatorname{diag}(e)$$

where

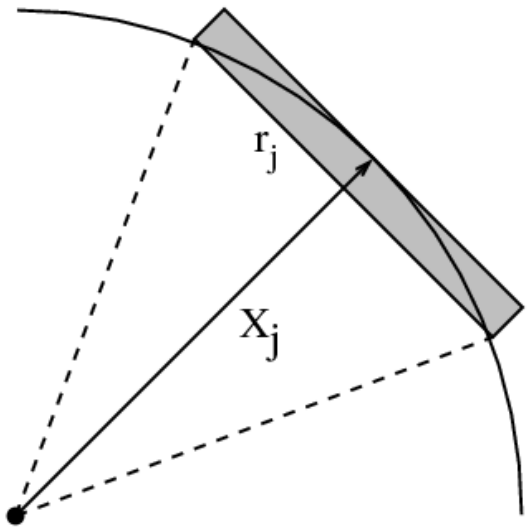
- $T \in \mathbb{R}^{n_c \times d_s}$ should be row sparse as before
- $V \in \mathbb{R}^{m \times d_s}$ models small noise
- $e \in \mathbb{R}^{d_s}$ indicates outlier data

V and e will be restricted to convex sets, and the rest of the model will remain the same as before.

Constraint on V

To encourage the columns of YT to be normalized, prevent V from having large components in the direction of X_s.

Constrain each column V_j to a hockey puck shaped disk D_j as pictured below.



This is simply a box constraint in cylindrical coordinates and is easy to project onto.

Constraint on e

We might expect some outliers in the data, namely some columns of X that are not well approximated by nonnegative linear combinations of other columns of X, but that we still don't wish to include as endmembers.

Model outlier error by $-X_s \operatorname{diag}(e)$

- Non-outlier case: want $e_j \approx 0$
- Outlier case: want $e_j \approx 1$, in which case regularization on T encourages corresponding column of T to be small

Restrict e to the convex set $E = \{e : e \geq 0,\ \sum_j (C_w e)_j \leq \gamma\}$

γ can be roughly interpreted as the fraction of allowed outliers

For outliers, $\|YT_j\|_2$ should be small. Otherwise it should be close to 1.

Proposed Extended Convex Model

$$\min_{T\geq 0, V_j \in D_j, e\in E} \zeta \sum_i \max_j (T_{i,j}) + \langle R_w \sigma C_w, T \rangle \qquad (3)$$

$$\text{such that} \qquad YT - X_s = V - X_s \operatorname{diag}(e).$$

The structure of this model is closely related to the robust PCA model proposed by Candes, Li, Ma and Wright even though it has a different noise model and uses $l_{1,\infty}$ regularization instead of the nuclear norm.

E. CANDES, X. LI, Y. MA, AND J. WRIGHT, *Robust principal component analysis*, 2009.

Numerical Optimization

It's convenient to use a variant of ADMM which allows the objective function to be split into more than two parts. A method proposed by He, Tao and Yuan works well here and is again based on finding a saddle point of the augmented Lagrangian

$$\begin{aligned} L_\delta(Z,T,V,e,P_1,P_2) =& g_{\geq 0}(T) + g_D(V) + g_E(e) \\ &+ \zeta \sum_i \max_j (T_{i,j}) + \langle R_w \sigma C_w, T \rangle \\ &+ \langle P_1, Z - T \rangle \\ &+ \langle P_2, YZ - V - X_s + X_s \operatorname{diag}(e) \rangle \\ &+ \frac{\delta}{2} \|Z - T\|_F^2 \\ &+ \frac{\delta}{2} \|YZ - V - X_s + X_s \operatorname{diag}(e)\|_F^2, \end{aligned}$$

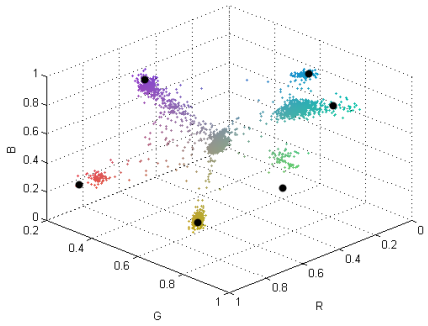
where g_D and g_E denote indicator functions for the $V \in D$ and $e \in E$ constraints.

B. He, M. Tao, and X. Yuan, *A splitting method for separate convex programming with linking linear constraints*, 2010.

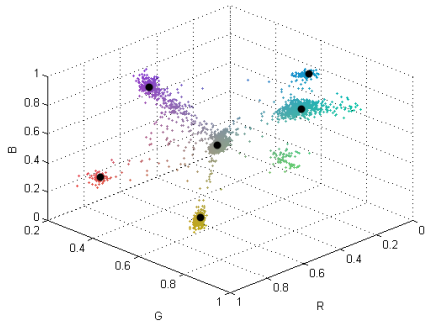
Effect of Extended Model

original image

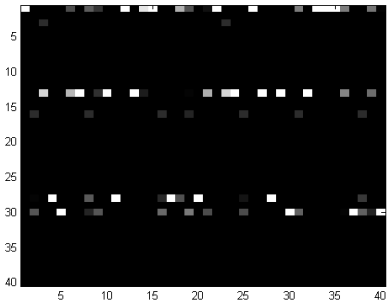
$\nu = 0, \gamma = 0$



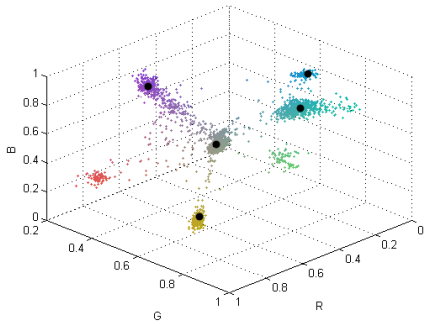
$\nu = 50, \gamma = .005$



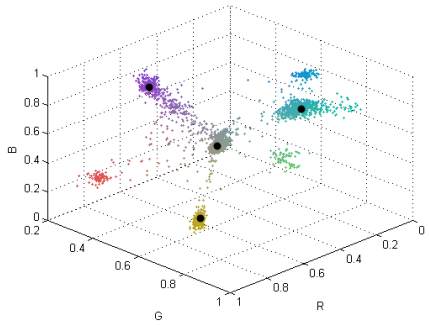
T for $\nu = 50, \gamma = .005$



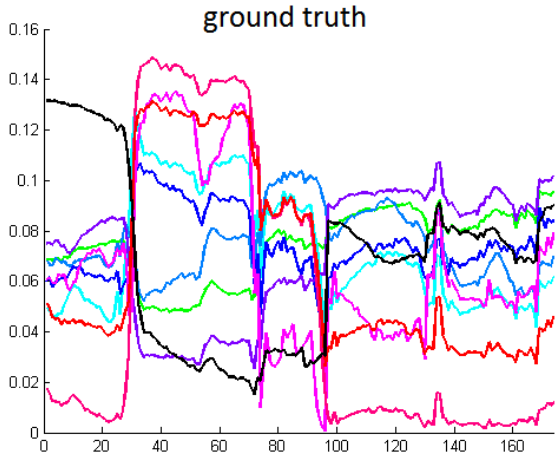
$\nu = 50, \gamma = .01$

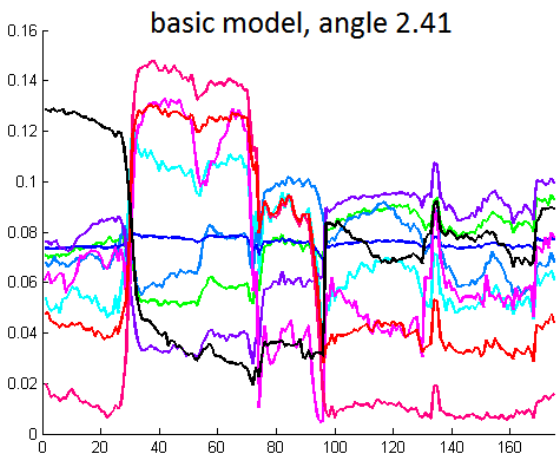


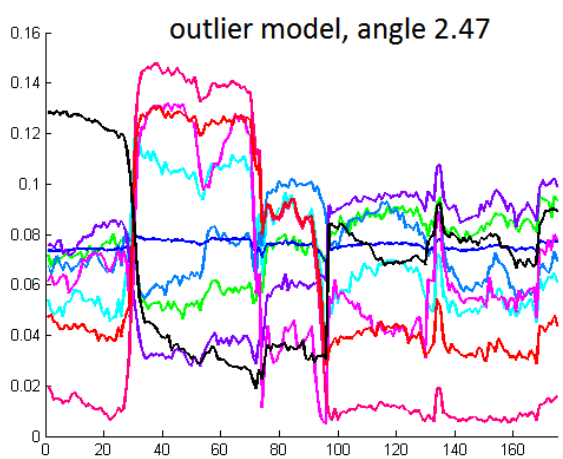
$\nu = 50, \gamma = .1$



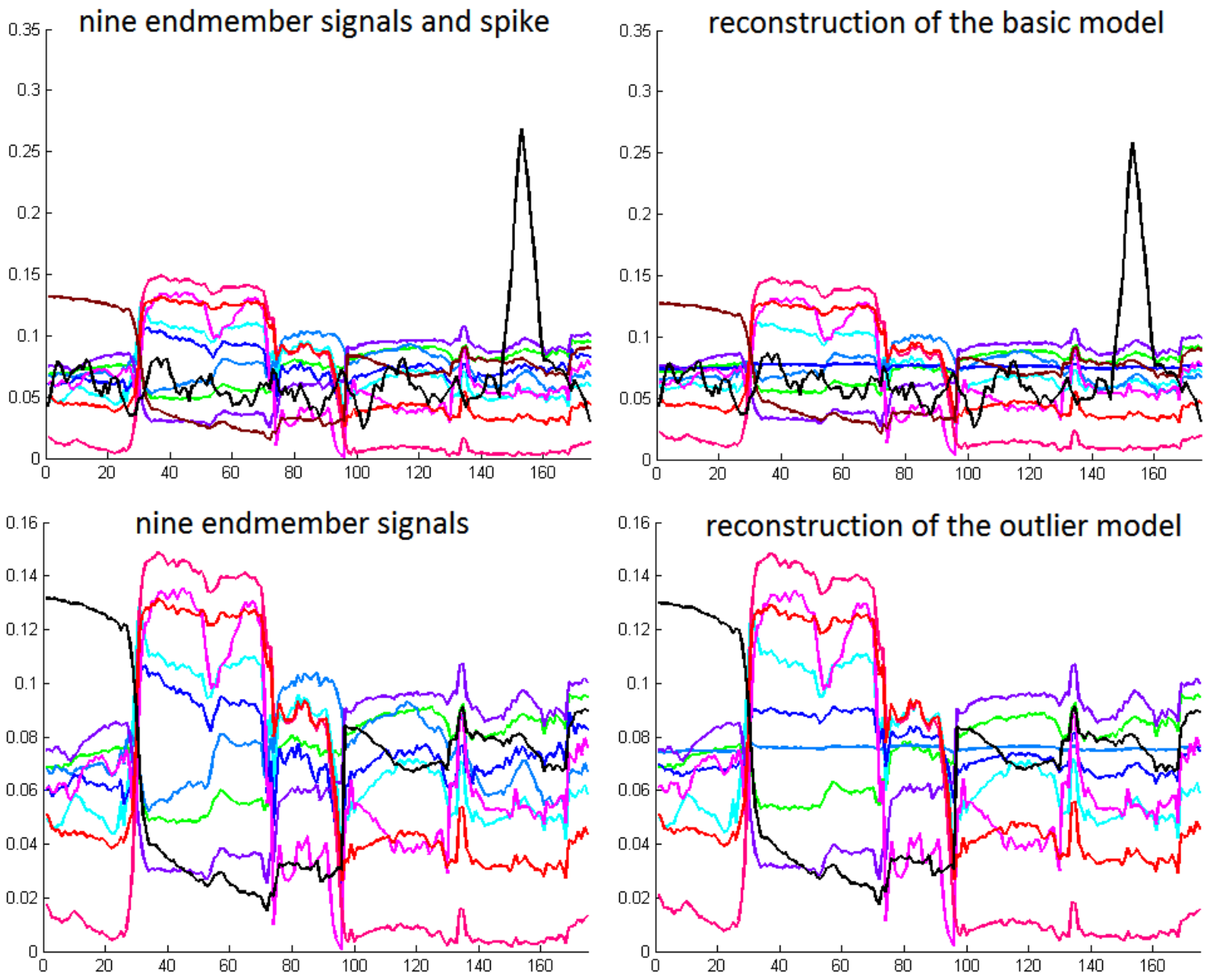
Comparison of Models On Indian Pines Data



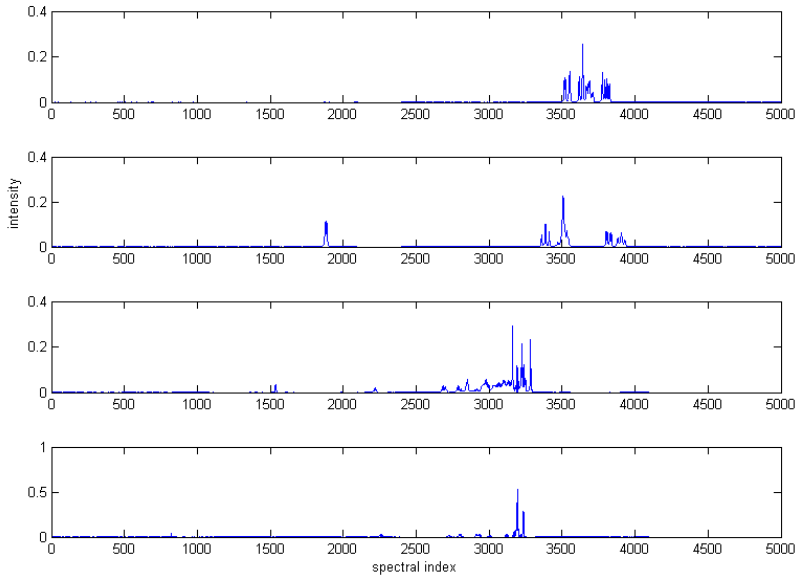




Comparison On Indian Pines Data with Outliers



Application to Blind Source Separation



- Generate $X_0 = A_0 S_0$ from known sources S_0 (from Naanaa and Nuzillard) and synthetic A_0
- Normalize nonzero columns of X_0 and use extended model to recover A in $X = AS$
- We can recover S_0 by $S_0 = \max(0, A^{-1} X_0)$ if possible, or by solving an optimization problem.

BSS Results

Recovered A after permutation is:

$$A = \begin{bmatrix} .3267 & .6524 & .3327 & .4933 \\ .3180 & .3300 & .6544 & .5110 \\ .6228 & .1757 & .1658 & .4836 \\ .6358 & .6593 & .6585 & .5114 \end{bmatrix}$$

and the synthetic A_0 used to generate X_0 is

$$A_0 = \begin{bmatrix} .3162 & .6576 & .3288 & .5000 \\ .3162 & .3288 & .6576 & .5000 \\ .6325 & .1644 & .1644 & .5000 \\ .6325 & .6576 & .6576 & .5000 \end{bmatrix}$$

Other Applications and Future Work

- It may be interesting to try these methods on other dimension reduction and dictionary learning applications where NMF is used and where the restriction of the dictionary to the data is desirable, like text mining or music classification.
- An interesting challenge for future work is to extend these approaches to problems for which the pixel purity (or partial orthogonality) assumption is almost but not exactly satisfied.