# Hyperspectral Image Unmixing: Accounting For Wavelength Dependence[*]

Chia Chye Yee[†], Yves Atchadé[‡]

March 2, 2017

We introduce a method for hyperspectral unmixing that incorporates wavelength dependence in addition to spatial dependence. Spatial dependence is incorporated into the model using class labels on the pixels that is assigned using spectral clustering. Wavelength dependence is introduced by correlating the errors in the unmixing regression models. We propose a non-standard alternating direction method of multipliers (ADMM) algorithm to solve the resulting non-convex optimization problem that simultaneously recovers the abundances and the sparse precision matrices of the spectral signatures. Using data collected by the SpecTIR imaging sensor, we show that the proposed method outperforms several other well-established unmixing models.

***Keywords:*** Hyperspectral Unmixing, Classification, Spectral Clustering, Graphical Lasso, Wavelength Dependence

## 1 Introduction

Hyperspectral imaging has become ubiquitous due to recent advancement in imaging technology. However, the signal processing step remains a challenging task. Indeed, due to the limited spatial resolution of the sensors, each pixel in a hyperspectral image is typically a mixture of endmember spectral signatures. Hyperspectral unmixing is the task of identifying these mixtures ([2, 27]).

[*] This work is partially supported by the NSF grant DMS 1228164 and DMS 1513040

[†] C. C. Yee: University of Michigan, 1085 South University, Ann Arbor, 48109, MI, United States. E-mail address: chye@umich.edu

[‡] Y. F. Atchadé: University of Michigan, 1085 South University, Ann Arbor, 48109, MI, United States. E-mail address: yvesa@umich.edu

A standard approach for unmixing hyperspectral images is to fit pixel-by-pixel regressions of the observed spectral signature on a set of reference spectral signatures or endmembers ([32, 24, 12, 23, 7, 18, 34]). The set of reference endmembers is often referred to as a “spectral library” . The library can be constructed by spectral analysis of materials in a lab or via pure spectral extraction from the image. However this basic pixel-by-pixel approach ignores the spatial dependence between the observed signatures as well as the dependence between data observed at different wavelengths.

Several authors have recognized the importance of incorporating spatial dependence in the analysis of hyperspectral images ([30, 28, 22, 21, 34]). As shown in these works, accounting for spatial dependence improves the unmixing. Most of the literature makes use of discrete Markov random fields models to partition the images into homogeneous groups of pixels. Under the assumption that pixels in the same group are fairly homogeneous, it is possible to develop more parsimonious data generating models. Hence accounting for spatial dependence actually improves the performance of the unmixing task, but also improves the computational efficiency and interpretability of the model ([34, 30]).

In this work we introduce spatial dependence in the unmixing model using class labels on the pixels that is assigned using spectral clustering ([19]). All pixels with the same class assignment share the same regression parameters. Spectral clustering is a powerful clustering algorithm that has been documented to perform better than centroid-based clustering methods such as K-mean. However one difficulty with spectral clustering in the amount of floating point operations it entails, which grows as $\mathcal{O}(Lp^2 + p^3)$, where $p$ is the number of pixels, and $L$ is the number of wavelengths. This significantly limits the size of the images that can be effectively processed by spectral clustering. In order to circumvent this problem, we use the Nystrom algorithm ([13, 9]) to produce a low rank approximation of the adjacency matrix, resulting in a huge saving in the computation time.

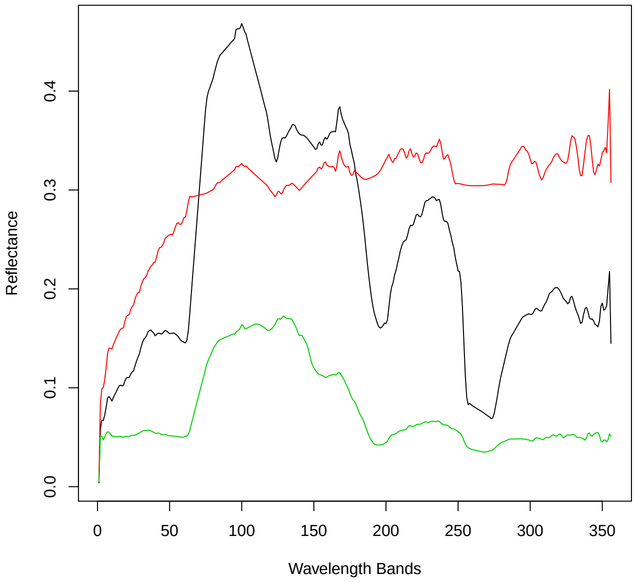


Figure 1: Spectral signatures of 3 randomly selected pixels from the Reno scene

As the discussion above shows, the spatial dependence inherent to hyperspectral images has been widely explored to improve the unmixing. However we are not aware of any statistical model in the literature that employs wavelength dependence in analyzing hyperspectral images, despite ample evidence that most hyperspectral data collected have wavelength dependence. For example Figure 1 shows the spectral signatures at three randomly selected pixels from an image collected by the SpecTIR imaging sensor over the city of Reno, Nevada. These spectral signatures are smooth, which suggests that contiguous wavelength bands are not independent of each other.

We propose a new hyperspectral unmixing framework that incorporate both spatial and wavelength dependence. As discussed above, the spatial dependence is incorporated into the model using class labels on the pixels that is assigned using spectral clustering. The wavelength dependence is introduced by assuming correlated error terms in the unmixing regression models. More precisely for each group of pixels, the joint distribution of the error terms is assumed to be a mean-zero Gaussian distribution parametrized by a possibly sparse precision matrix (inverse covariance matrix). Accounting for the fact that the number of wavelengths is potentially large, we propose to fit the model via a penalized maximum likelihood approach, with a sparsity promoting penalty. The proposed model leads to a non-convex optimization problems with several constraints and penalty (positivity and size constraints on the abundances, sparsity promoting penalty on the precision matrices). We propose to solve this problem using the alternating direction method of multipliers (ADMM) algorithm ([10, 5]). However the standard version of the ADMM algorithm is not tractable in our case. As a result, we develop a non-standard block-update version of the ADMM algorithm. We show that with an appropriately chosen step-size, the algorithm is stable in the sense that positive definiteness of the precision matrices are guaranteed throughout the iterations. Furthermore we show that the Lagrangian objective function of the problem is always non-increasing along the iterations.

We perform several numerical experiments with real and simulated data to investigate the behavior of the Nystrom approximation, and we formulate some practical guidelines. Using data collected by the SpecTIR imaging sensor, we compare the proposed model to several other well-established unmixing models. Using the Bayesian Information Criterion (BIC) as a goodness of fit criterion, we show that accounting for wavelength dependence significantly improves the unmixing model, compared to the alternative models considered.

The paper is organized as follows. **Section 2** contains the main methodological contributions. The main model is described in **Section 2**. The proposed ADMM algorithm is presented in **Section 2.1**, with the technical proofs postponed to **Section 5**. Spectral clustering and its approximation using the Nystrom method are presented in **Section 2.2** and **Section 2.3** respectively. The numerical experiments using both synthetic and real data are presented in **Section 3**. The paper ends with some concluding remarks in **Section 4**.

# 2 Unmixing with wavelength dependence

We begin with a presentation of the unmixing model. Our basic assumption is that the hyperspectral image is partitioned in $K$ groups (not necessarily contiguous) such that the pixels in each group represent similar materials, and their spectral signatures therefore have similar probability distributions. We impose this assumption in order to reduce the complexity of the model, and improve the scalability of the method. We show how to bring the data close to this model in Section 2.2 using spectral clustering. Under this assumption, we shall proceed to solve the unmixing problem on each group independently. Hence the model that follows applies to a given group $k$. However, for notational simplicity, we shall omit to explicitly write the dependence on the group in the modeling. Let $Y_1, \ldots, Y_N$ ($Y_i \in \mathbb{R}^L$) denote the spectral signatures observed on pixels $i = 1, \ldots, N$. To perform the unmixing, the library $X \in \mathbb{R}^{L \times R}$ of pure end-members (possibly tailored to the group) are recovered using one of several algorithms such as VCA [23] performed on the pixels within the group in question. We allow the library to include non-linear combinations of the pure end-members. Hence our methodology can be applied to both linear and non-linear (mainly bilinear) unmixing [24, 12]. We assume that the library matrix $X$ is full-rank column, which implies that $R \leq L$. Consistent with our assumption that the pixels in the same group represent the same material, we model the vectors $Y_1, \ldots, Y_N$ as independent and identically distributed random variables and

$$Y_i = X\beta + \epsilon_i \tag{1}$$

where $\epsilon_i \overset{i.i.d.}{\sim} \mathbf{N}(0, \Theta^{-1})$, for a precision matrix $\Theta$ and $\beta \in \mathbb{B}_+ \overset{\text{def}}{=} [0, B]^R$, for some constant $B > 0$. In other words, our working assumption is that the pixels in the same given group can be viewed as noisy representations of the same mixture of pure end-members with mixing coefficient $\beta$.

The constraint restricting all components of the mixing coefficient $\beta$ to be nonnegative is known as the positivity constraint, and is standard in hyperspectral unmixing literature [31, 18, 8, 17]. Here we also restrict the mixing coefficients to be upper bounded by the constant $B$. This is imposed in order to guarantee the stability of the algorithm that we propose to solve the unmixing problem. In practice this restriction is rarely an issue provided that $B$ is taken as a very large constant.

In the sequel we will use the notation $\mathcal{M}_L$ (resp. $\mathcal{M}_L^+$) to denote the space of symmetric (resp. symmetric positive definite ) $L \times L$ matrices. In most hyperspectral unmixing models, the precision matrix $\Theta$ is assumed to be a diagonal matrix, implying that there is no wavelength correlation. In practice, we frequently observe spectral signatures that are smooth as seen in **Figure 1**. This means the adjacent wavelength bands are correlated, implying that some of the off-diagonal values of the concentration matrix are non-zero under the Gaussian noise formulation as above. As we illustrate below a better solution to the unmixing problem is obtained by taking these correlations into account.

For $\beta \in \mathbb{B}_+$, we define

$$S(\beta) \overset{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} (Y_i - X\beta)(Y_i - X\beta)^T.$$

The negative log-likelihood of the model[1] is given by:

$$\ell(\beta, \Theta) = -\frac{1}{2} \log |\Theta| + \frac{1}{2} \mathsf{Tr}\left(\Theta S(\beta)\right), \quad \beta \in \mathbb{B}_+, \quad \Theta \in \mathcal{M}_L^+.$$

In the above display, $|\Theta|$ denotes the determinant of the matrix $\Theta$.

In hyperspectral imaging, the number of spectral bands $L$ can be large (several hundreds). To improve the recovery of the precision matrix $\Theta$, we shall add a sparsity inducing penalty to recover sparse precision matrices $\Theta$. In general the structure of $\Theta$ depends on the material at hand. However, it seems a reasonable assumption that most wavelength bands are partially uncorrelated. Under this assumption we are led to the following optimization problem:

$$(\hat{\beta}, \hat{\Theta}) = \mathsf{argmin}_{\left(\beta \in \mathbb{R}^R,\ \Theta \in \mathcal{M}_L^+\right)} \left[ -\frac{1}{2} \log |\Theta| + \frac{1}{2} \mathsf{Tr}(\Theta S(\beta)) + \mathsf{I}_+(\beta) + \lambda \mathsf{Reg}(\Theta) \right], \qquad (2)$$

where

$$\mathsf{I}_+(\beta) = \begin{cases} 1 & \text{if } \beta \in \mathbb{B}_+ \\ +\infty & \text{otherwise} \end{cases}, \quad \text{and} \quad \mathsf{Reg}(\Theta) \overset{\text{def}}{=} \sum_{i,j} \left( \alpha |\Theta_{ij}| + \frac{1-\alpha}{2} \Theta_{ij}^2 \right).$$

[1] up to an additive constant that we ignore

The function $\mathsf{I}_+$ enforces the positivity constraint, and the regularization term $\mathsf{Reg}$ (a mixture of $\ell^1$ and $\ell^2$ regularization) is added to promote sparsity. The special case $\alpha = 1$, corresponds to the LASSO of [29] applied to the concentration matrix [11]. The addition of the second term corresponding to $\alpha \neq 1$ leads to the elastic net [36] to aid in terms of variable selection. The parameter $\lambda > 0$ controls the strength of the regularization. In practice the parameters $\alpha$ and $\lambda$ need to be tuned for a good behavior of the method.

It not hard to see that the optimization problem (2) has always at least one solution. Indeed, if we endow $\mathcal{M}_L$ with the Frobenius norm (denoted $\|\cdot\|_{\mathsf{F}}$), and $\mathbb{B}_+$ with the usual Euclidean norm $\|\cdot\|_2$, and since $\log|\Theta| \to -\infty$, as $\|\Theta\|_{\mathsf{F}} \to 0$, we have

$$\lim_{\|\Theta\|_{\mathsf{F}}+\|\beta\|_2\to 0} \left[-\frac{1}{2}\log|\Theta| + \frac{1}{2}\mathsf{Tr}(\Theta S(\beta)) + \mathsf{I}_+(\beta) + \lambda\mathsf{Reg}(\Theta)\right]$$

$$= \lim_{\|\Theta\|_{\mathsf{F}}+\|\beta\|_2\to +\infty} \left[-\frac{1}{2}\log|\Theta| + \frac{1}{2}\mathsf{Tr}(\Theta S(\beta)) + \mathsf{I}_+(\beta) + \lambda\mathsf{Reg}(\Theta)\right] = +\infty.$$

Therefore it suffices to solve Problem (2) on a sufficiently large compact ball. And since the objective function is continuous, it has at least one solution on any such balls.

Given a hyperspectral image partitioned into $K$ groups, and given a spectral library for each group, we solve $K$ optimization problems of the form (2) to find for each group, the mixing parameter $\hat{\beta}$ and the precision matrix $\hat{\Theta}$.

## 2.1 Computation

In this section we describe a practical algorithm to solve the optimization problem (2), and we show that the algorithm is stable, and converges to a stationary point of Problem (2). The algorithm that we propose is a modification of the well-known ADMM algorithm [10, 5]. One can set up the ADMM algorithm for Problem (2) by rewriting the problem equivalently as

$$\min\left[f(\Theta,\beta) + g(u)\right], \quad \Theta \in \mathcal{M}_L^+,\ \beta \in \mathbb{R}^R,\ u \in \mathbb{R}^R \ \text{ s.t. } \beta = u,$$

where

$$f(\Theta,\beta) = -\frac{1}{2}\log|\Theta| + \frac{1}{2}\mathsf{Tr}(\Theta S(\beta)) + \lambda\mathsf{Reg}(\Theta), \quad \text{and} \quad g(u) = \mathsf{I}_+(u).$$

The augmented Lagrangian for this problem is given by

$$\mathcal{L}(\Theta,\beta,u,q) = f(\Theta,\beta) + g(u) + \langle q, \beta - u\rangle + \frac{\rho_q}{2}\|\beta - u\|^2,$$

for a regularization parameter $\rho_q > 0$. Based on this augmented Lagrangian the ADMM algorithm for solving (2) takes the following form.

**Algorithm 2.1.** *Choose some initial value $(\Theta_0, \beta_0, u_0, q_0)$. For $k = 0, 1, \ldots$ repeat the following. Given $(\Theta_k, \beta_k, u_k, q_k)$:*

1. *Solve*
$$u_{k+1} = \textsf{\textit{Argmin}}_{u\in\mathbb{R}^R}\ \mathcal{L}(\Theta_k,\beta_k,u,q_k),$$

2. *Solve*
$$(\Theta_{k+1},\beta_{k+1}) = \textsf{\textit{Argmin}}_{(\Theta\in\mathcal{M}_L^+,\,\beta\in\mathbb{R}^R)}\ \mathcal{L}(\Theta,\beta,u_{k+1},q_k).$$

3. *Then update*
$$q_{k+1} = q_k + \rho_q\left(\beta_{k+1} - u_{k+1}\right).$$

Algorithm 2.1 is a standard ADMM algorithm as applied to the optimization (2). It allows a nice decoupling of the constraints on $\beta$. However the algorithm cannot be implemented because the optimization in Step 2 cannot be solved in closed form. We circumvent this difficulty by replacing Step 2 by an approximate block update. For $\theta\in\mathcal{M}_L$, $\delta>0$, define

$$\mathrm{Prox}_\delta(\theta) \overset{\text{def}}{=} \textsf{Argmin}\ _{u\in\mathcal{M}_L}\left[\textsf{Reg}(u) + \frac{1}{2\delta}\left\|u-\theta\right\|_{\textsf{F}}^2\right].$$

The matrix $\mathrm{Prox}_\delta(\theta)$ is straightforward to compute. Its $(i,j)$-th component is given by

$$\left(\mathrm{Prox}_\delta(\theta)\right)_{ij} = \begin{cases} 0 & \text{if } |\theta_{ij}| < \alpha\delta \\ \frac{\theta_{ij}-\alpha\delta}{1+(1-\alpha)\delta} & \text{if } \theta_{ij}\geq\alpha\delta \\ \frac{\theta_{ij}+\alpha\delta}{1+(1-\alpha)\delta} & \text{if } \theta_{ij}\leq-\alpha\delta \end{cases}$$

Similar let $\mathrm{Proj}_{\mathbb{B}_+}:\ \mathbb{R}^R\to\mathbb{B}_+$ denote the component-wise projection on $\mathbb{B}_+$. We are thus lead to the following algorithm.

**Algorithm 2.2** (Main algorithm)**.** *Choose some initial value* $(\Theta_0,\beta_0,u_0,q_0)$*. For* $k=0,1,\ldots$ *repeat the following. Given* $(\Theta_k,\beta_k,u_k,q_k)$*:*

1. *Solve*
$$u_{k+1} = \textsf{\textit{Argmin}}_{u\in\mathbb{R}^R}\ \mathcal{L}(\Theta_k,\beta_k,u,q_k),$$

2. *Compute*
$$\Theta_{k+1} = \mathrm{Prox}_{\lambda\delta}\left(\Theta_k - \delta\left(S(\beta_k)-\Theta_k^{-1}\right)\right),$$

3. *Solve*
$$\beta_{k+1} = \textsf{\textit{Argmin}}_{\beta\in\mathbb{R}^R}\ \mathcal{L}(\Theta_{k+1},\beta,u_{k+1},q_k).$$

4. *Then update*
$$q_{k+1} = q_k + \rho_q\left(\beta_{k+1} - u_{k+1}\right).$$

The optimization problems in Steps (1) and (3) can be solved explicitly. This yields the following practical version of the algorithm that can be easily implemented.

**Algorithm 2.3** (Main algorithm–Practical version)**.** *Choose some initial value* $(\Theta_0,\beta_0,u_0,q_0)$*. For* $k=0,1,\ldots$ *repeat the following. Given* $(\Theta_k,\beta_k,u_k,q_k)$*:*

*1. Compute*

$$u_{k+1} = \mathrm{Proj}_{\mathbb{B}_+}\left(\beta_k + \frac{1}{\rho_q} q_k\right),$$

*2. Compute*

$$\Theta_{k+1} = \mathrm{Prox}_{\lambda\delta}\left(\Theta_k - \delta\left(S(\beta_k) - \Theta_k^{-1}\right)\right),$$

*3. Compute*

$$\beta_{k+1} = \frac{1}{N}\left(X^T\Theta_k X + \frac{\rho_q}{N} I_R\right)^{-1}\left(X^T\Theta_k\left(\sum_{i=1}^N Y_i\right) + \rho_q u_{k+1} - q_k\right).$$

*4. Then update*

$$q_{k+1} = q_k + \rho_q\left(\beta_{k+1} - u_{k+1}\right).$$

One issue with the proposed algoirhm is that in Step (2) of Algorithm 2.2, when the inverse of $\Theta_k$ is taken there is no guarantee that the matrix $\Theta_k$ is non-singular. Another issue is that the objective function in Problem (2) is non-convex, although it is bi-convex. Hence it is not possible to provide a theoretical guarantee that the algorithm proposed above converges to a solution of (2). The theoretical analysis below addresses these two issues. We show that for a well-chosen step-size $\delta$, and a well-chosen initial value $\Theta_0$, all the matrix $\Theta_k$ produced by Algorithm 2.2 are in fact non-singular and the algorithm never fails. We also show that the sequence of Lagrangian values along the iterations of the algorithm is non-increasing. The derived results rely on more general results on gradient iterations for graphical lasso developed by [6], and a general analysis of the ADMM algorithm by [10].

We define

$$\mathcal{L}_k = \mathcal{L}(\Theta_k, \beta_k, u_k, q_k), \quad k \geq 0.$$

And we set $\mu \overset{\text{def}}{=} \sup_{\beta\in\mathbb{B}_+} \|S(\beta)\|_2 + \alpha\lambda L$,

$$\pi_\star \overset{\text{def}}{=} \frac{\sqrt{\mu^2 + 4(1-\alpha)\lambda} - \mu}{2(1-\alpha)\lambda}, \quad \text{and} \quad \Pi_\star \overset{\text{def}}{=} \frac{\sqrt{\alpha^2\lambda^2L^2 + 4(1-\alpha)\lambda} + \alpha\lambda L}{2(1-\alpha)\lambda}.$$

Given $0 < a < b \leq \infty$, we denote $\mathcal{M}_L^+(a,b)$ the set of all symmetric positive definite matrices $A$ such that $\lambda_{\mathsf{min}}(A) \geq a$, and $\lambda_{\mathsf{max}}(A) \leq b$, where $\lambda_{\mathsf{min}}(A)$ (resp. $\lambda_{\mathsf{max}}(A)$) denotes the smallest (resp. largest) eigenvalue of $A$.

**Theorem 2.1.** *Suppose that $\Theta_0 \in \mathcal{M}_L^+(\pi_\star, \Pi_\star)$, and $\delta \in (0, \pi_\star^2]$. Then for all $k \geq 0$, $\Theta_k \in \mathcal{M}_L^+(\pi_\star, \Pi_\star)$. Furthermore if the sequence $\{\beta_k,\ k \geq 0\}$ remains bounded and $\rho_q$ is taken sufficiently large, then the sequence $\{\mathcal{L}_k,\ k \geq 0\}$ is non-increasing and converges to a limit $\mathcal{L}_\star$, as $k \to \infty$.*

*Proof.* See Section 5. □

[31].

## 2.2 Spectral Clustering

Hyperspectral images are dense in terms of spectral wavelength resolution. Due to the complexity involved in modeling the data generating mechanism, we assumed that for a given image, the pixels in the image can be partitioned into groups that are fairly homogeneous in terms of their observed spectral signature. Within a homogeneous group, the observed spectral signatures are assumed to share the same mixing composition and library in addition to wavelength correlation structure as developed in Section 2. This assumption greatly simplifies the problem of unmixing by reducing the number of parameters estimated. In addition to the model simplification, the construction of the spectral library via VCA[23] within each group is also made easier due to the reduced number spectral signatures considered for each within group unmixing. In

Although the partitioning can be done using central grouping methods like k-means [20], the method does not work well for pixels that are similar but are not in spatial proximity[2]. A more appealing alternative is the use of pair-wise affinities: spectral clustering [19].

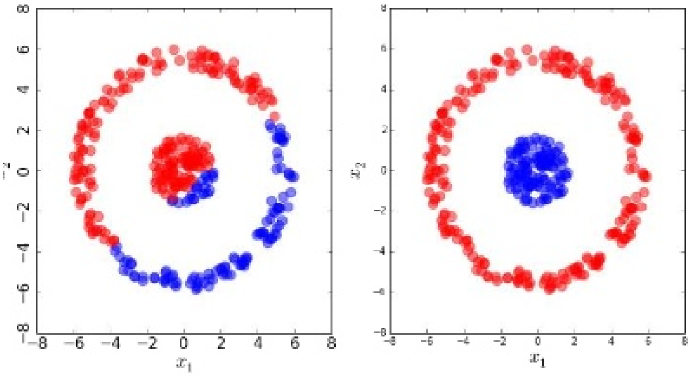


Figure 2: A demonstration of k-means compared to Spectral Clustering applied on the same dataset[1].

Spectral clustering[3] is an empirical approach to cluster pixels using eigenvectors of matrices derived from data associated with the pixels [25]. The method involves the use of an adjacency/weight matrix constructed via a distance kernel applied to the pixels. Intuitively, the method seeks to group pixels in such a way that the weights between pixels belonging to the same group (intra-group weights) are large while the weights between pixels belonging to different groups (inter-group weights) are small. This means the pixels within the same group are more similar to each other while the pixels belonging to different groups are dissimilar to each other. Suppose there are $P$ pixels in the image, the method begins by building an adjacency matrix $W \in \mathbb{R}^{P \times P}$ using the following kernel:

$$W_{ij} = \exp\left\{-\frac{\|Y_i - Y_j\|_2^2}{L}\right\}$$

[2] refer to Figure 2 for an illustration.

[3] see for instance [19] for a good review

where $Y_i \in \mathbb{R}^L$ is the spectral signature observed at pixel $i$. The kernel used here only depends on the distance between the spectral signatures. The kernel may be altered to incorporate various other measures of similarity between[4]. But for the purpose of this paper we will concentrate on the $\ell_2$ distance between the spectral signatures. Once the adjacency matrix has been constructed, we proceed to build the Laplacian matrix $L$:

$$D = diag\left(\sum_j W_{ij}\right)$$
$$L = D^{-1/2} W D^{-1/2}$$

where $L$ is the normalized Laplacian. Once this has been constructed, we apply the eigen-decomposition to $L$. It is important to note that there are alternate ways to define the Laplacian. In the present formulation, the leading eigenvalues are most pertinent, while the formulation as in [19] will result in the case where the smallest eigenvalues are most meaningful to spectral clustering. The number of groups $G$ is determined by examining the largest eigenvalues. Suppose the eigenvalues of $L$ are arranged in descending order $\{e_{(P)}, \ldots e_{(1)}\}$, the number of groups is:

$$G = \max\{P - i : e_{(i)} \geq \tau, i = 1, \ldots P\}$$

Where $\tau$ is the threshold for setting the number of groups. At this point, it is important to note that spectral clustering is a supervised approach to clustering. The threshold $\tau$ is dependent on the observed data and setting this requires a certain amount of judgement regarding the number of groups relative to the number of pixels in the image.

Once $G$ has been established, we collect $G$ eigenvectors associated with the $G$ largest eigenvalues into the matrix $V \in \mathbb{R}^{P \times G}$. We then apply $K$-means clustering on the $P$ $G$-dimensional row-vectors. The resulting classification is assigned to the pixels.

Spectral clustering is known to be an extremely effective classification algorithm. **Figure 3** is an example of a classification plot recovered from a synthetic dataset.

[4] such as geographical distance, or similarity along some important covariates

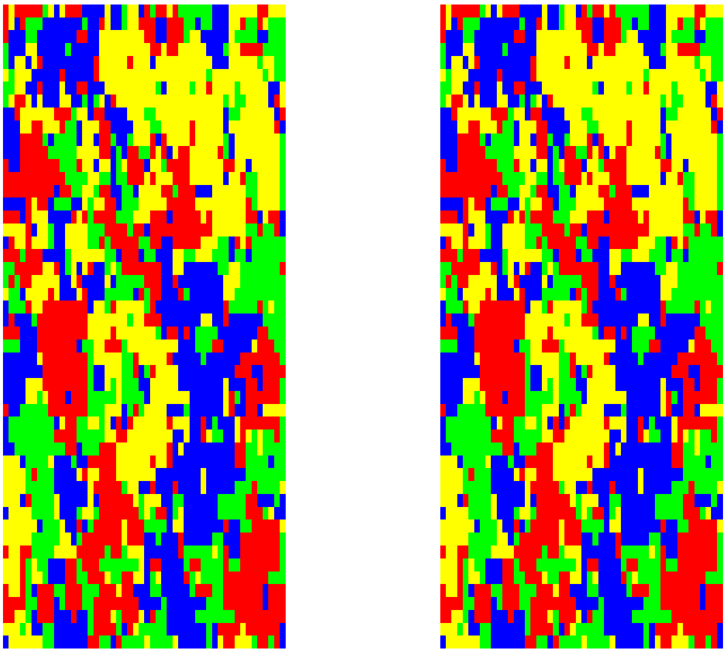


Figure 3: Classification plots of the ground truth (left) and the recovered (right) classes for synthetic data using exact spectral clustering.

## 2.3 Spectral Clustering via the Nystrom Method

One difficulty with spectral clustering in large image analysis is that the computation and storage of the adjacency matrix $W$ scales exponentially with the number of pixels. For instance, on a small image with $p = 10^4$ pixels, the computation of $W$ requires $\mathcal{O}(Lp^2) = \mathcal{O}(L \times 10^8)$ operations and the memory to store $W$ is $O(p^2) = O(10^8)$ values. The exponential growth of memory and computation costs associated with spectral clustering limits the size of the image that can be effectively partitioned. In order to circumvent this problem, we use a low rank approximation to the adjacency matrix based on the Nystrom algorithm ([26, 16, 14, 13, 9]). As we will illustrate in this section, the Nystrom method dramatically reduces the storage and computational burden for spectral clustering while incurring an acceptable loss of fidelity in the recreation of the image partition. This is especially apparent in Table 1. There is a tradeoff between the fidelity of the partition and the number of ranks used in the adjacency matrix. The choice of the number of ranks has to be balanced with the computational and storage burden incurred.

Suppose that we randomly select $n \ll P$ pixels and partition the adjacency matrix as:

$$W = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

Where $A \in \mathbb{R}^{n \times n}$ is the adjacency matrix computed from the $n$ sampled pixels, $B \in \mathbb{R}^{n \times (P-n)}$ consists of the adjacency computed from the $n$ sampled points with respect to the $P-n$ non-sampled points, and $C \in \mathbb{R}^{(P-n)\times(P-n)}$ is the adjacency between the non-sampled points. Under this formulation, the approximate eigenvectors $\hat{U}$ for W via the Nystrom method takes the following form:

$$\hat{U} = \begin{bmatrix} U \\ B^T U \Lambda^{-1} \end{bmatrix}$$

Where $U$ contains orthogonal eigenvectors of $A$ and $\Lambda$ is diagonal containing eigenvalues of $A$. Therefore, $A = U\Lambda U^T$. The low rank representation/approximation of $W$ takes the following form:

$$\begin{aligned} \hat{W} &= \hat{U}\Lambda\hat{U}^T \\ &= \begin{bmatrix} A \\ B^T \end{bmatrix} A^+ \begin{bmatrix} A & B \end{bmatrix} \end{aligned}$$

Where $A^+$ is the pseudo-inverse of $A$. In most cases, if $A$ is symmetric and invertible, this coincides with its inverse, $A^+ = A^{-1}$ .From the formulation above, the low-rank representation of $W$ via the Nystrom method approximates the matrix $C$ as $C \approx B^T A^+ B$. Suppose we define $A^{1/2}$ as the symmetric positive definite square root of the matrix $A$, $Z = A + (A^+)^{1/2} B B^T (A^+)^{1/2}$, and $Z \in \mathbb{R}^{n \times n}$ is diagonalized as $U_Z \Lambda_Z U_Z^T$. The matrix $V$ defined as:

$$V = \begin{bmatrix} A \\ B^T \end{bmatrix} (A^+)^{1/2} U_Z \Lambda_Z^{-1/2}$$

will be the matrix of eigenvectors which will diagonalize $\hat{W} = V \Lambda_Z V^T$. Suppose we wish to perform spectral clustering that solves the regularized optimization of the min-cut problem[19], we need the row sums of $\hat{W}$ which can be computed as

$$\begin{aligned} \mathbf{d} &= \hat{W}\mathbf{1} \\ &= \begin{bmatrix} A\mathbf{1}_n + B\mathbf{1}_{(P-n)} \\ B^T\mathbf{1}_n + B^T A^+ B \mathbf{1}_{(P-n)} \end{bmatrix} \end{aligned}$$

We then “normalize” the matrix by replacing the entries of $A$ and $B$ with:

$$\tilde{A}_{ij} \leftarrow \frac{A_{ij}}{\sqrt{d_i d_j}} \qquad \forall i, j = 1, \dots n$$

$$\tilde{B}_{ij} \leftarrow \frac{B_{ij}}{\sqrt{d_i d_j}} \qquad \forall i = 1, \dots n, j = n+1, \dots, P$$

After renormalization we may use the renormalized version of the eigenvectors:

$$\tilde{V} = \begin{bmatrix} \tilde{A} \\ \tilde{B}^T \end{bmatrix} (\tilde{A}^{+})^{1/2} U_{\tilde{Z}} \Lambda_{\tilde{Z}}^{-1/2}$$

Where $\tilde{Z} = \tilde{A} + (\tilde{A}^{+})^{1/2} \tilde{B} \tilde{B}^T (\tilde{A}^{+})^{1/2}$. The number of groups $\tilde{G}$ is determined by examining the leading eigenvalues given by the diagonal values of $\Lambda_{\tilde{S}}$. Suppose once more we have eigenvalues arranged in descending order $\{\tilde{e}_{(n)}, \ldots \tilde{e}_{(1)}\}$, the number of groups is:

$$\tilde{G} = \max\{n - i : \tilde{e}_{(i)} \geq \tilde{\tau}, i = 1, \ldots n\} \tag{3}$$

Note that the low-rank approximation only requires the storage of matrix $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times (P-n)}$ and the diagonalization of the matrix $S \in \mathbb{R}^{n \times n}$ which is significantly less complex than the storage and the diagonalization of the matrix $W \in \mathbb{R}^{P \times P}$. Similar to the group number determination in exact spectral clustering, $\tilde{\tau}$ is dependent on the observed data and setting it requires a judgmental call regarding the number of groups relative to the number of pixels in the image. Once $\tilde{G}$ has been established, we collect $\tilde{G}$ eigenvectors associated with the $\tilde{G}$ largest eigenvalues into matrix $\tilde{E} \in \mathbb{R}^{P \times \tilde{G}}$. We then apply $K$-means clustering on the $P$ $\tilde{G}$-dimensional row-vectors.

In order to explore the behavior of the Nystrom method we conducted an experiment with synthetic data containing 600 × 600 pixels[5] with 7 groups. For each of the group, we generated a random $\beta$ with number of non-zeros set at 15% of its entries. A library $X$ is chosen from the group of libraries extracted from the Reno scene analyzed below. In addition, each library is augmented with bilinear combinations resulting in each library having 20 columns (ie. $X \in \mathbb{R}^{356 \times 20}$). Each pixel in a given group will have the following observed spectral signature:

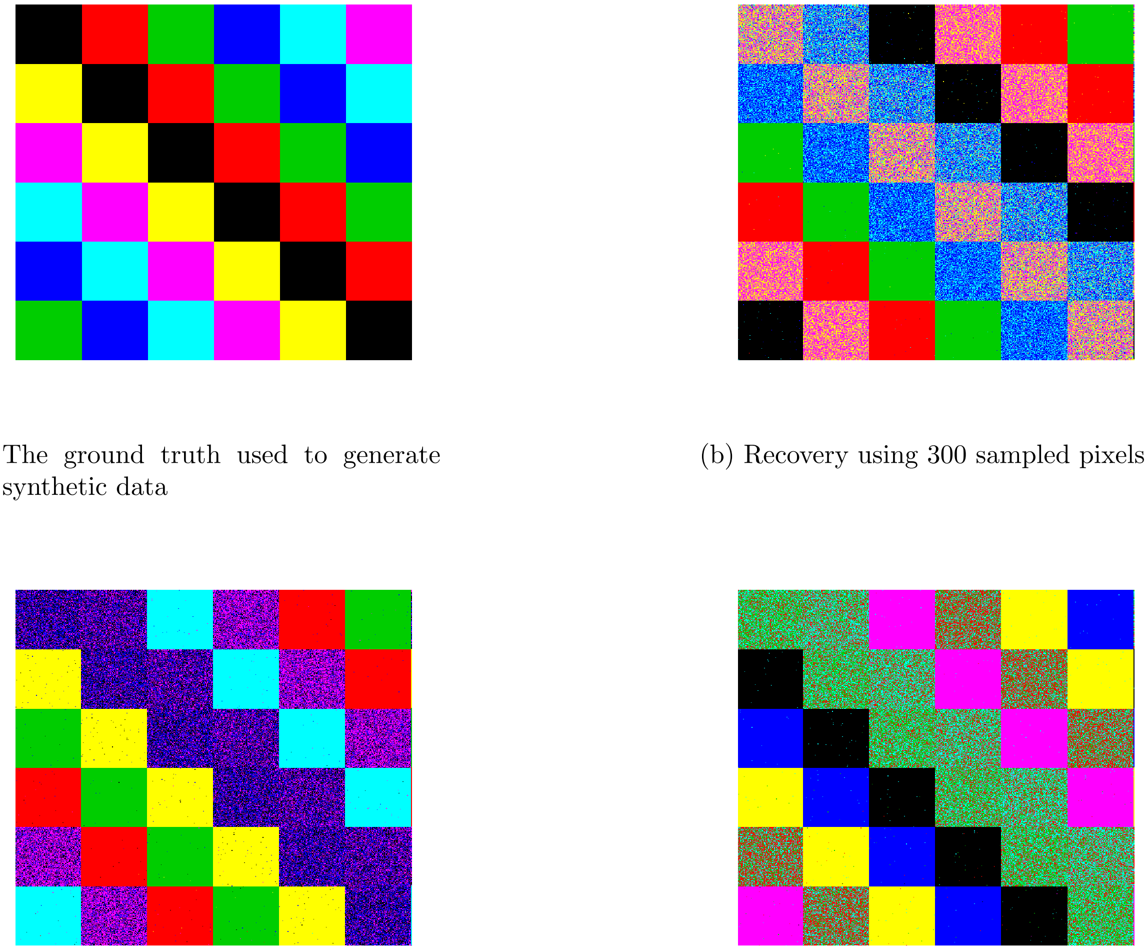
$$Y_i = X\beta + \epsilon_i \qquad\qquad \epsilon_i \sim N(0, 0.125^2 I_{356})$$

We investigate how well spectral clustering based on the Nystrom method can recover the true classes. In our experiments we tested the low-rank approximation with different number of sampled pixels used. The performance of the Nystrom-based spectral clustering depends crucially on $n$, the number of sampled pixels. We experimented with $n \in \{100, 200, 300, 400, 500, 600, 700, 800\}$ and **Table 1** contains the average misclassification rate for these experiments over 20 repetitions. As reference, we included classification plots for some interesting cases of the recovery using Nystrom method in **Figure 4**. It is important to note that increasing the number of pixels sampled increases the computational burden in terms of memory and operations by the order of $\mathcal{O}(n^2)$ while results of the experiments with synthetic data does not show dramatic improvement in the classification rate. Based on these results, we chose $n = 300$ pixels in our application of the proposed method in real data simulations in order to strike a balance between accuracy and computational complexity.

[5] which is comparable to the number of pixels in the Gulf Wetlands image analyzed below

| # sampled | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 |
|---|---|---|---|---|---|---|---|---|
| Error rate | 0.3008 | 0.2700 | 0.2759 | 0.2701 | 0.2557 | 0.2398 | 0.2466 | 0.1972 |

Table 1: Misclassification rates using the low-rank approximation to spectral clustering.



(a) The ground truth used to generate synthetic data

(b) Recovery using 300 sampled pixels

(c) Recovery using 600 sampled pixels

(d) Recovery using 700 sampled pixels

Figure 4: Classification for the ground truth and the recovered classification for different sampled pixels.

# 3 Numerical experiments

This section contains the simulation studies conducted using the proposed methodology and the results of the simulation studies. The simulation will start off with a controlled experiment using synthetic data set which has known parameters $\Theta$ and $\beta$ from which the synthetic spectral signatures are generated.

In the real data experiments, we generated the library $X$ using **VCA** [23]. **VCA** is an iterative algorithm that extracts endmembers from observed spectral signature in the image. This method assumes that there is at least one pure pixel in the image (pixel group). The algorithm takes the following form:

1. Begin by select a random pixel from the image as an endmember in the library. The first endmember can be taken from the pixel with the largest spectral signature.
2. Find the space that is orthogonal to the column space of the library.
3. From the remaining pixels, the pixel with the largest orthogonal projection is added into the library as an endmember.
4. Repeat steps 2 and 3 until an appropriate number of endmembers are included in the library. It is obvious that the maximum number of endmembers cannot exceed the number of pixels in the image.

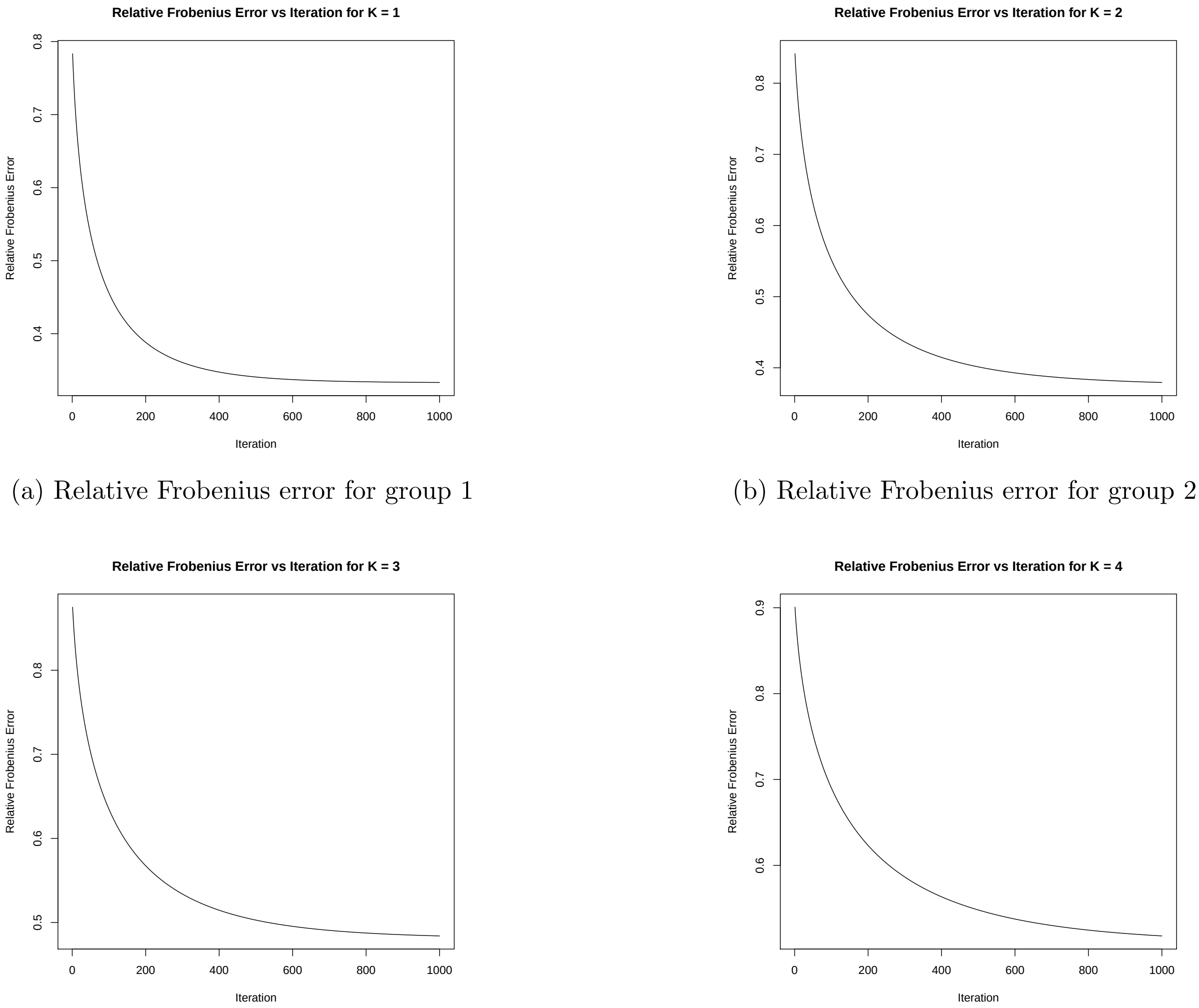
From our observation of the recovered endmembers from **VCA**, only the first few iterations of **VCA** are needed. We augmented the library with bilinear combinations of the original spectral signatures.

## 3.1 Synthetic Data

We document the synthetic and real data simulation results in this section. The synthetic data involves a $50 \times 50$ image consisting of 4 classes. Spectral clustering does a good job of separating the pixels correctly as evidenced by **Figure 3**. Within each class, we implemented the algorithm described in **Algorithm 2.3** in order to recover the abundances and the structure of the wavelength dependence. We measure the accuracy of the recovered concentration matrix $\Theta$ within each group using the relative Frobenius error which is defined as

$$\text{RelFrob}_\Theta(\hat{\Theta}) = \frac{\|\Theta - \hat{\Theta}\|^2}{\|\Theta\|^2}$$

Where $\Theta$ is the concentration matrix used to generate the synthetic wavelength dependence in the synthetic data and $\hat{\Theta}$ is the recovered concentration matrix. **Figures 5a - 5d** shows how the relative Frobenius error of the concentration matrix decreases as a function of the number of iterations. This shows that iterating the updates $\Theta_k$ are converging. The relative $l_2$ error for the abundances for all 4 groups converge very quickly as evidenced by **Figures 6a - 6d**.



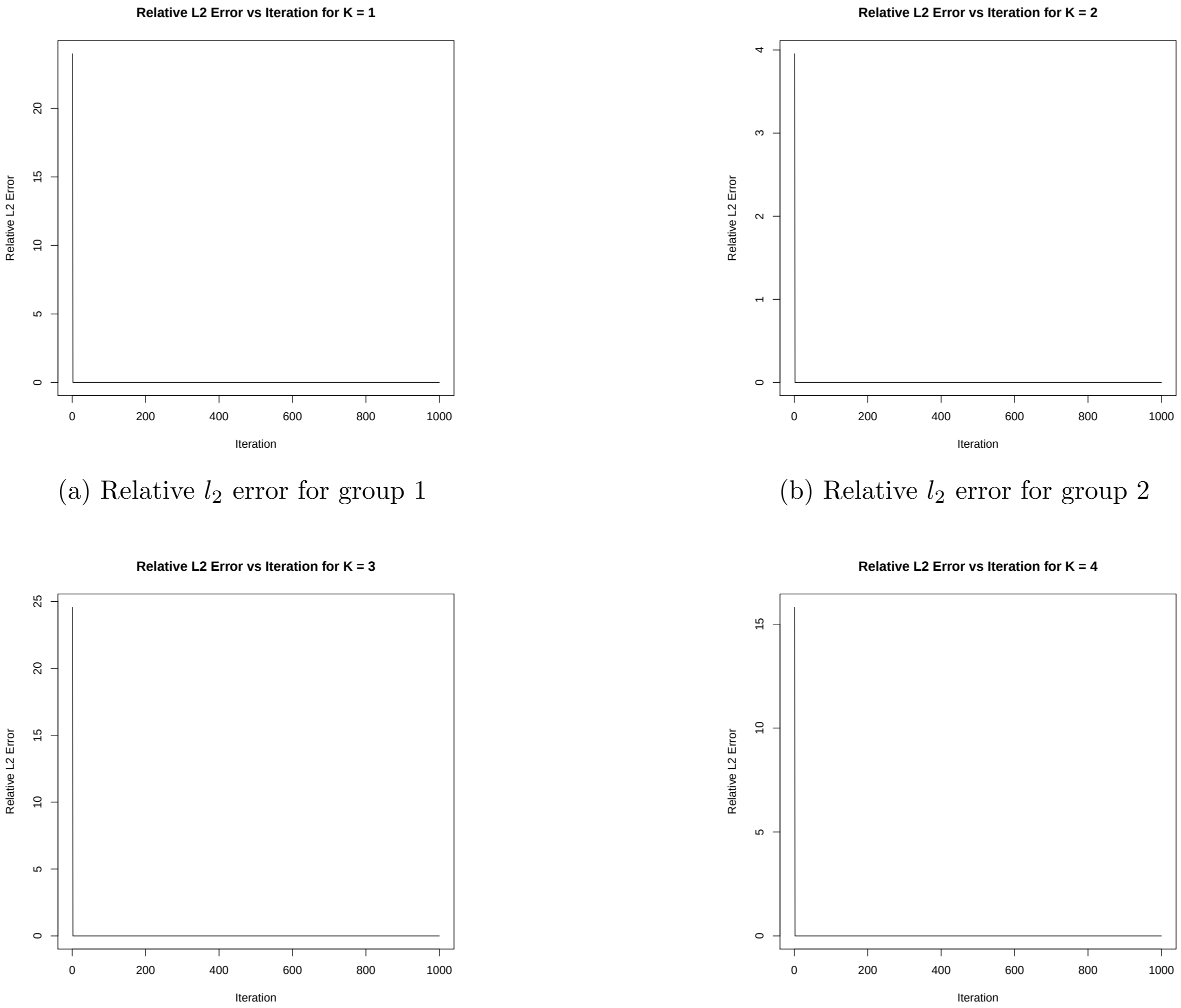
(a) Relative Frobenius error for group 1

(b) Relative Frobenius error for group 2

(c) Relative Frobenius error for group 3

(d) Relative Frobenius error for group 4

Figure 5: Relative Frobenius error of the estimated wavelength dependence as a function of iterations. This provides empirical proof that the algorithm produces a sequence of Θs that converges



(a) Relative $l_2$ error for group 1

(b) Relative $l_2$ error for group 2

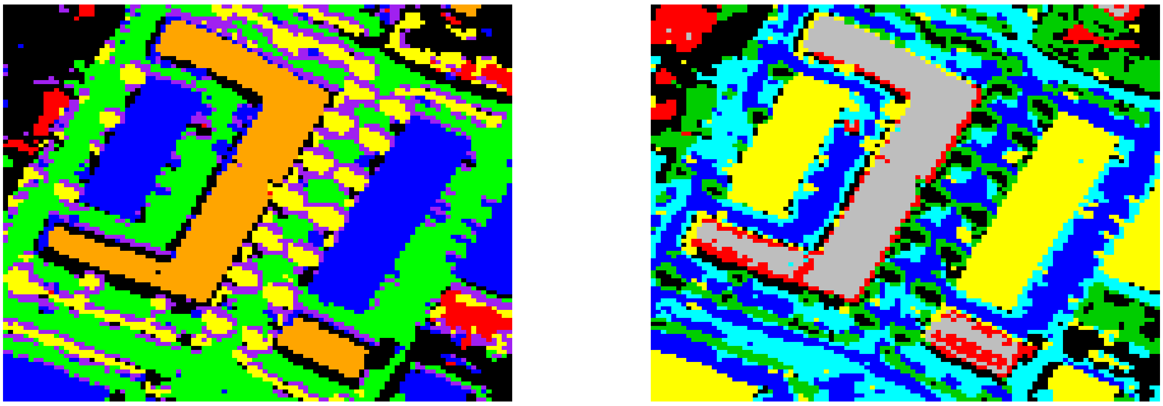
(c) Relative $l_2$ error for group 3

(d) Relative $l_2$ error for group 4

Figure 6: Relative $l_2$ error of the estimated abundances as a function of iterations. This provides empirical proof that the algorithm produces a sequence of $\beta$s that converges.

## 3.2 Reno Scene

We applied the proposed methodology to a scene from Reno obtained from [4]. The scene is an urban area from Reno, Nevada contains $600 \times 320$ pixels. The scene contains buildings, roads, parking lots, and a river. This scene was chosen because of the distinctive features of the urban environment enables some form of "eyeball" validation of the classification plots using images taken on the visible spectrum. We performed exact spectral clustering on the top-left $100 \times 100$ subset of the scene in the exploratory analysis. The classification recovered from the exact spectral clustering can be seen in **Figure 7** (a).



(a) (b)

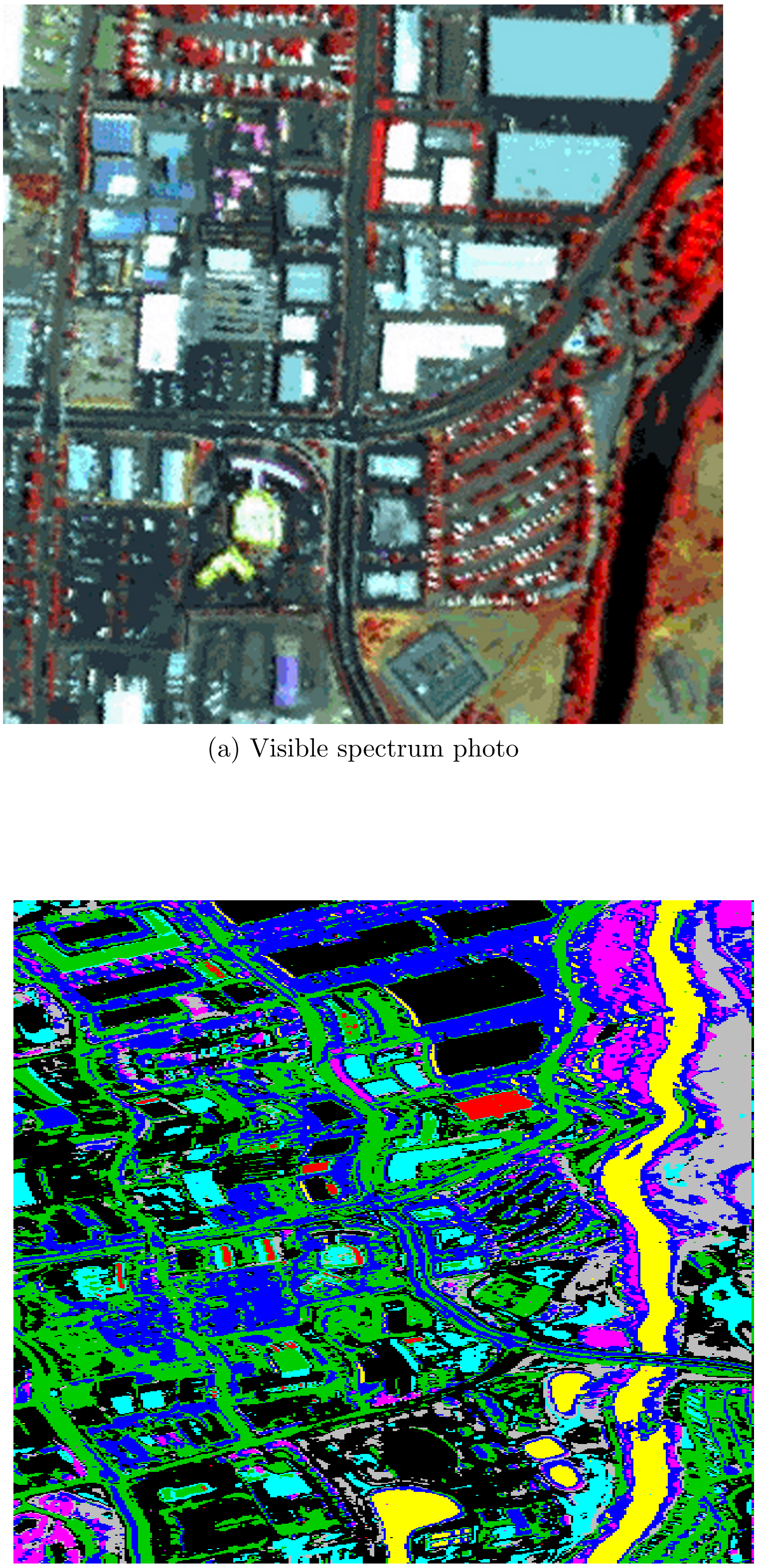
Figure 7: Classification plot of a subset of the Reno data using spectral clustering. (a) Exact spectral clustering, (b) Nystrom approximation.

We performed the approximate spectral clustering via the Nystrom method as outlined in **Section 2.3** on the whole image. In the approximation, we sampled $n = 300$ pixels from $600 \times 320$ pixels and performed Nystrom method-based spectral clustering resulting $G = 14$ distinct groups in the image.

**Figure 7** (b) shows the top-left $100 \times 100$ subset of the classification plot retrieved by the Nystrom method-based spectral clustering. This image is to be compared with **Figure 7** (a). From these two classification plots, we can see that there is very little loss in fidelity in utilizing the low-rank approximation, since the classification plot for the approximate spectral clustering is still able to capture the building and the parking lot covered in the classification plot recovered from exact spectral clustering. This exercise confirms that the loss in fidelity in utilizing a low-rank representation of $W$ for spectral clustering is minimal. Also note that there appears to be attenuation based on the "wavy" appearance of the roads and rivers in the image. This may be an artifact of image processing which converts the radiance data from the remote sensor into reflectance data via the radiance to reflectance equation in [3].

**Figure 8** (b) shows the full classification plot retrieved from the Nystrom-based spectral clustering, and **Figure 8** (a) shows a subset of the Reno scene image in the visible spectrum, which can be distinctly recognized in the middle part of **Figure 8** (b).

After classification, we performed **VCA** within each class to recover endmembers for the library. In this exercise, we extracted 5 endmembers from the each group of pixels and augmented the 5 endmembers with bilinear combinations resulting in a library containing 20 endmembers (columns). Once we have the library $X$, we applied the updates outlined in **Algorithm 2.3** .



(a) Visible spectrum photo

(b) Recovered classification plot

Figure 8: Complete classification of the Reno scene using Nystrom method.

In order to achieve $\frac{10}{L\times L}$ level of sparsity for $\Theta$, we can set $\alpha \approx 0.9$ with $\lambda \propto \sqrt{\frac{L}{|G_i|}}$ where $|G_i|$ is the number of pixels within group $G_i$. For the Reno scene we set $\alpha = 0.5$ and $\lambda = c\sqrt{\frac{L}{|G_i|}}$. The constant $c \in \{0, 0.025, \ldots, 0.25\}$ is found by grid search evaluating a model selection criterion.

Selecting the penalization level requires a metric that measures the fit of the model while penalizing for model complexity. For most cases, the Bayesian Information Criterion (**BIC**) as defined below is adequate

$$BIC_\lambda(\hat{\Theta}, \hat{\beta}) = -2 \log L(\hat{\Theta}, \hat{\beta}) + R|\hat{\Theta}|_0 \log |G_i|$$

Where $|G_i|$ is the number of pixels in group $i$. However, in the estimation of the concentration matrix it is often that we run into cases where the number of parameters[6] is growing with the sample size which violates one of the assumption required for **BIC** consistency[15]. In lieu of the regular **BIC** as a criterion for selecting $c$, we used the extended **BIC**[15] which allows for the growth of the number of non-zero entries in the concentration matrix. The extended **BIC** takes the following form for group $G_i$ :

$$B\tilde{I}C_\lambda(\hat{\Theta}, \hat{\beta}) = -2 \log L(\hat{\Theta}, \hat{\beta}) + R|\hat{\Theta}|_0 \log |G_i| + 4|\hat{\Theta}|_0 \lambda \log L \tag{4}$$

Note the additional penalization term in the extended **BIC** which means the extended **BIC** is more punitive towards complex models. The resultant **BIC**s are compared for different values of $c \in \{0, 0.025, \ldots, 0.25\}$ and the one with the lowest value is chosen as the optimal $\lambda^\star$. The estimates are then computed using $\lambda^\star$. The resultant estimates of $\hat{\Theta}$ and $\hat{\beta}$ are taken as the ideal estimates.

In order to gain insight into the improvement spatial dependence and wavelength dependence brings to our model, we compare our proposed model with several other unmixing models listed below:

1. Each group has one mixture parameter $\beta$ shared by all pixels in that grou, but there is no wavelength dependence. This model accounts for spatial dependence (using the same Nystrom-based spectral clustering described above) while leaving out wavelength dependence. Hereforth we would refer to this as Model 1.

2. Each pixel in each group has a mixture parameter $\beta$ without accounting for wavelength dependence. This model is similar to Model 1 in terms of accounting for spatial dependence but has more mixture parameters $\beta$ which results in a better fit residual-wise. Hereforth we would refer to this as Model 2.

3. Each pixel in the image has a mixture parameter $\beta$ without accounting for wavelength dependence. This model does not account for spatial and wavelength dependence. Essentially, this model performs pixel level unmixing. The library used in this model is extracted from the whole image rather than at the group level. Hereforth we would refer to this as Model 3.

[6] number of non-zeros in the concentration matrix

In order to compare the proposed to the 3 models listed above, we calculated the regular BIC for the 3 models listed and compared them to the extended BIC (eBIC) of the proposed model. The regular BIC for the challenger models:

$$BIC(\hat{\beta}) = -2\log L(\hat{\beta}) + R\log(\text{number of pixels}) \tag{5}$$

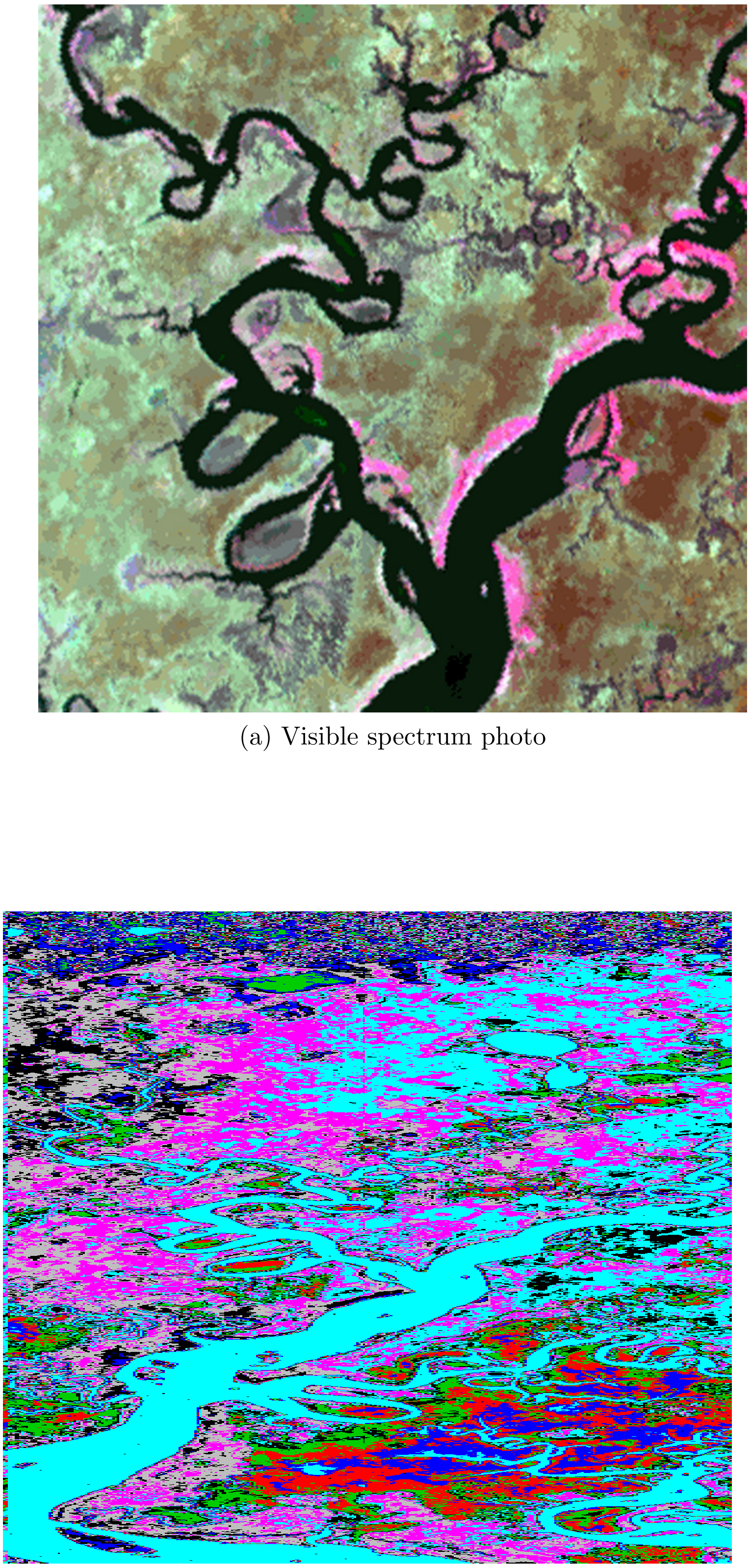
Note that the extended BIC is just the regular **BIC** with additional penalties for non-zero values for the concentration matrix. **Table 2** documents the **BIC** and e**BIC** computed for the models as a comparison. The reason we chose **BIC** and e**BIC** as a comparison is to provide a meaningful way to compare model fit for models with different number of parameters. As evidenced from the results, even with the extra penalization term in e**BIC**, the model with the best fit is the one that incorporates spatial and wavelength dependence.

| Model | Proposed | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|
| BIC and eBIC | 52,455,479 | 125,789,217 | 166,595,863 | 172,547,769 |

Table 2: The BIC for the 3 challenger models and extended BIC for the proposed model applied to the Reno scene.

## 3.3 Gulf Wetlands (Suwannee River) Scene

Besides the Reno scene, we also examined a scene from the Suwannee River obtained from [4]. The scene contains a river delta, wetlands, and plants indigenous to swamp lands. The image contains 1200 × 320 pixels. We performed the approximate spectral clustering using the low-rank approximation of $W$ using 300 sampled pixels and recovered $G = 13$ groups from the scene. **Figure 9** contains the classification plot and the image on the visible spectrum. From the classification plot, we can tell that approximate spectral clustering is able to clearly distinguish the river from the wetlands overgrown with indigenous plants.



(a) Visible spectrum photo

(b) Recovered classification plot

Figure 9: Complete classification of the Gulf Wetlands (Suwannee River) scene using Nystrom method.

In addition, we also performed similar comparisons via the BIC/eBIC relative to the 3 challenger models as described in **Section 3.2**. **Table 3** documents the resultant BIC/eBIC computed for the models as a comparison between the models. In this case, the improvement in model fit is more dramatic when compared to the Reno scene.

| Model | Proposed | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|
| BIC and eBIC | 99,419,358 | 254,110,811 | 338,084,541 | 352,836,042 |

Table 3: The BIC for the 3 challenger models and extended BIC for the proposed model applied to the Gulf Wetlands (Suwannee River) scene.

# 4 Conclusion

We have proposed a novel model for hyperspectral unmixing and classification that incorporates both spatial and wavelength dependence. Spectral clustering combined with the Nystrom method for fast computation is employed for clustering the image, and a non-convex optimization algorithm is proposed to fit the model. We give some theoretical guarantee that the proposed algorithm is stable and consistently lower the Lagrangian objective function along the iterations. The empirical evidence from the synthetic data simulation suggests that the algorithm converges to the problem solution. The real data analysis results from the Reno and Suwannee scenes demonstrate that the incorporation of spatial and wavelength dependence significantly improves model fit when compared with several standard unmixing models. One promising avenue for future is the extension of the model to handle hyperspectral images acquired over time for change detection.

# 5 Proof of Theorem 2.1

For the proof we need the following results. Lemma 5.1 is a special case of Lemma 2 of [6], and Lemma 5.2 is a special case of Lemma 12 and Lemma 14 of [6].

**Lemma 5.1.** *Suppose that $\delta \in (0, \pi_\star^2]$. For $\Theta \in \mathcal{M}_L^+(\pi_\star, \Pi_\star)$, and for $\beta \in \mathbb{B}_+$, we set $\bar{\Theta} = \mathrm{Prox}_\delta\left(\Theta - \delta\left(S(\beta) - \Theta^{-1}\right)\right)$. Then $\bar{\Theta} \in \mathcal{M}_L^+(\pi_\star, \Pi_\star)$.*

**Lemma 5.2.** *For some arbitrary symmetric matrix $S \in \mathbb{R}^{L\times L}$, we define*

$$h(\Theta) \overset{\text{def}}{=} -\log|\Theta| + \mathit{Tr}(\Theta S), \qquad \textit{and} \qquad f(\Theta) \overset{\text{def}}{=} h(\Theta) + \lambda \mathit{Reg}(\Theta)\,, \quad \Theta \in \mathcal{M}_L^+.$$

*Fix $0 < \pi < \Pi \leq \infty$.*

1. *For $\Theta_1, \Theta_0 \in \mathcal{M}_p^+(\pi, \Pi)$, we have*

$$h(\Theta_0) + \left\langle S - \Theta_0^{-1}, \Theta_1 - \Theta_0 \right\rangle + \frac{1}{2\Pi^2} \left\| \Theta_1 - \Theta_0 \right\|_{\mathsf{F}} \leq h(\Theta_1)$$
$$\leq h(\Theta_0) + \left\langle S - \Theta_0^{-1}, \Theta_1 - \Theta_0 \right\rangle + \frac{1}{2\pi^2} \left\| \Theta_1 - \Theta_0 \right\|_{\mathsf{F}}.$$

2. *Let $\delta \in (0, \pi^2]$, and $\Theta, \bar{\Theta}, \Theta_0 \in \mathcal{M}_p^+(\pi, \Pi)$. Suppose that*

$$\bar{\Theta} = \mathrm{Prox}_\delta \left( \Theta - \delta(S - \Theta^{-1}) \right),$$

*then*

$$2\delta \left( f(\bar{\Theta}) - f(\Theta_0) \right) + \left\| \bar{\Theta} - \Theta_0 \right\|_{\mathsf{F}}^2 \leq \left( 1 - \frac{\gamma}{\Pi^2} \right) \left\| \Theta - \Theta_0 \right\|_{\mathsf{F}}^2.$$

Since $\mathcal{L}_k = \mathcal{L}(\Theta_k, \beta_k, u_k, q_k)$, we have

$$\begin{aligned}
\mathcal{L}_{k+1} - \mathcal{L}_k \;=\; & \mathcal{L}(\Theta_{k+1}, \beta_{k+1}, u_{k+1}, q_{k+1}) - \mathcal{L}(\Theta_{k+1}, \beta_{k+1}, u_{k+1}, q_k) \\
& + \mathcal{L}(\Theta_{k+1}, \beta_{k+1}, u_{k+1}, q_k) - \mathcal{L}(\Theta_{k+1}, \beta_k, u_{k+1}, q_k) \\
& + \mathcal{L}(\Theta_{k+1}, \beta_k, u_{k+1}, q_k) - \mathcal{L}(\Theta_k, \beta_k, u_{k+1}, q_k) \\
& + \mathcal{L}(\Theta_k, \beta_k, u_{k+1}, q_k) - \mathcal{L}(\Theta_k, \beta_k, u_k, q_k)
\end{aligned}$$

We have

$$\begin{aligned}
\mathcal{L}(\Theta_{k+1}, \beta_{k+1}, u_{k+1}, q_{k+1}) - \mathcal{L}(\Theta_{k+1}, \beta_{k+1}, u_{k+1}, q_k) &= \langle q_{k+1} - q_k, \beta_{k+1} - u_{k+1} \rangle \\
&= \frac{1}{\rho_q} \| q_{k+1} - q_k \|_2^2. \quad (6)
\end{aligned}$$

We can write the second term as

$$\mathcal{L}(\Theta_{k+1}, \beta_{k+1}, u_{k+1}, q_k) - \mathcal{L}(\Theta_{k+1}, \beta_k, u_{k+1}, q_k) = \Psi(\beta_{k+1}) - \Psi(\beta_k),$$

where

$$\Psi(\beta) \overset{\text{def}}{=} \frac{1}{2} \mathsf{Tr}(S(\beta)\Theta_{k+1}) + \langle q_k, \beta - u_{k+1} \rangle + \frac{\rho_q}{2} \| \beta - u_{k+1} \|_2^2.$$

By the choice of $\beta_{k+1}$ in Step (3), we have $\nabla \Psi(\beta_{k+1}) = 0$. Using this and a Taylor expansion of $\Psi$ around $\beta_{k+1}$, we have

$$\begin{aligned}
\Psi(\beta_{k+1}) - \Psi(\beta_k) &= -\left( \beta_k - \beta_{k+1} \right)' \left( \rho I_R + N X' \Theta_{k+1} X \right) \left( \beta_k - \beta_{k+1} \right) \\
&\leq -\left( \rho_q + N \| X' \Theta_{k+1} X \|_2 \right) \| \beta_{k+1} - \beta_k \|^2. \quad (7)
\end{aligned}$$

For the third term we have

$$\mathcal{L}(\Theta_{k+1}, \beta_k, u_{k+1}, q_k) - \mathcal{L}(\Theta_k, \beta_k, u_{k+1}, q_k) = f(\Theta_{k+1}, \beta_k) - f(\Theta_k, \beta_k).$$

We know from Lemma 5.1 that the sequence $\{\Theta_k,\ k \geq 0\}$ remains in the set $\mathcal{M}_L^+(\pi_\star, \Pi_\star)$, and $\Theta_{k+1} = \mathrm{Prox}_\delta\left(\Theta_k - \delta\left(S(\beta_k) - \Theta_k^{-1}\right)\right)$. Hence, by Lemma 5.2-(2), we have

$$\begin{aligned}\mathcal{L}(\Theta_{k+1}, \beta_k, u_{k+1}, q_k) - \mathcal{L}(\Theta_k, \beta_k, u_{k+1}, q_k) &= f(\Theta_{k+1}, \beta_k) - f(\Theta_k, \beta_k) \\ &\leq -\frac{1}{2\delta}\left\|\Theta_{k+1} - \Theta_k\right\|_{\mathsf{F}}^2. \end{aligned} \tag{8}$$

For the last term, we note that the function $u \mapsto \mathcal{L}(\Theta_k, \beta_k, u, q_k)$ is strongly convex with strong convexity parameter $\rho_q$, and since $u_{k+1}$ minimizes this function we have

$$\mathcal{L}(\Theta_k, \beta_k, u_{k+1}, q_k) - \mathcal{L}(\Theta_k, \beta_k, u_k, q_k) \leq -\frac{\rho_q}{2}\|u_{k+1} - u_k\|_2^2. \tag{9}$$

By putting together the bounds in (6-9), we get

$$\begin{aligned}\mathcal{L}_{k+1} - \mathcal{L}_k \leq \frac{1}{\rho_q}\|q_{k+1} - q_k\|_2^2 - \left(\rho_q + \|X'\Theta_{k+1}X\|_2\right)\|\beta_{k+1} - \beta_k\|^2 \\ - \frac{1}{2\delta}\left\|\Theta_{k+1} - \Theta_k\right\|_{\mathsf{F}}^2 - \frac{\rho_q}{2}\|u_{k+1} - u_k\|_2^2. \end{aligned} \tag{10}$$

The optimality condition of Step (3) of the Algorithm 2.2 is

$$\frac{\partial \mathcal{L}(\Theta_k, \beta, u_{k+1}, q_k)}{\partial \beta}|_{\beta_{k+1}} = q_k - X'\Theta_{k+1}\sum_{i=1}^N (Y_i - X\beta_{k+1}) + \rho_q\left(\beta_{k+1} - u_{k+1}\right) = 0.$$

And since $\rho_q(\beta_{k+1} - u_{k+1}) = q_{k+1} - q_k$, this optimality condition gives

$$\begin{aligned} q_{k+1} &= X'\Theta_{k+1}\sum_{i=1}^N \left(Y_i - X\beta_{k+1}\right) \\ &= X'\left(\Theta_{k+1} - \Theta_k\right)\sum_{i=1}^N \left(Y_i - X\beta_{k+1}\right) + X'\Theta_k \sum_{i=1}^N \left(Y_i - X\beta_{k+1}\right). \end{aligned}$$

Hence

$$q_{k+1} - q_k = X'\left(\Theta_{k+1} - \Theta_k\right)\sum_{i=1}^N \left(Y_i - X\beta_{k+1}\right) + NX'\Theta_k X(\beta_k - \beta_{k+1}).$$

Hence

$$\begin{aligned}\|q_{k+1} - q_k\|_2^2 \leq 2\|X\|_2^2 \sup_{\beta \in \mathbb{B}_+}\left\|\sum_{i=1}^N \left(Y_i - X\beta_{k+1}\right)\right\|_2^2 \|\Theta_{k+1} - \Theta_k\|_{\mathsf{F}}^2 \\ + 2N^2\|X'\Theta_k X\|_2^2\|\beta_{k+1} - \beta_k\|_2^2. \end{aligned}$$

Setting $c_0 \overset{\text{def}}{=} 2\|X\|_2^2 \sup_{\beta \in \mathbb{B}_+}\left\|\sum_{i=1}^N \left(Y_i - X\beta_{k+1}\right)\right\|_2^2$, and using this last inequality in (10), we get

$$\mathcal{L}_{k+1} - \mathcal{L}_k \leq -\left(\rho_q + N\|X'\Theta_{k+1}X\|_2 - \frac{2}{\rho_q}N^2\|X'\Theta_k X\|_2^2\right)\|\beta_{k+1} - \beta_k\|^2$$
$$-\left(\frac{1}{2\delta} - \frac{c_0}{\rho_q}\right)\|\Theta_{k+1} - \Theta_k\|_{\mathsf{F}}^2 - \frac{\rho_q}{2}\|u_{k+1} - u_k\|_2^2. \quad (11)$$

Therefore for $\rho_q > 0$ large enough so that $\frac{1}{2\delta} > \frac{c_0}{\rho_q}$ the conclusion of the theorem readily follows.

□

# References

[1] Example of spectral clustering compared to k-means. http://pictures.netne.net/open-source-use-image-spectral-graph.html.

[2] NASA Visible Earth Website. http://visibleearth.nasa.gov/.

[3] SpecTIR Website. https://eo1.usgs.gov/faq/question?id=21.

[4] USGS Website. http://www.spectir.com/free-data-samples/.

[5] M. Afonso, J. Bioucas-Dias, and M. A. T. Figueiredo. A fast algorithm for the constrained formulation of compressive image reconstruction and other linear inverse problems. In IEEE International Conf. on Acoustics, Speech, and Signal Processing - ICASSP, pages –, March 2010.

[6] Y. F. Atchadé, R. Mazumder, and J. Chen. Scalable Computation of Regularized Precision Matrices via Stochastic Optimization. ArXiv e-prints, September 2015.

[7] A.m. Baldridge, S.j. Hook, C.i. Grove, and G. Rivera. The aster spectral library version 2.0. Remote Sensing of Environment, 113(4):711715, 2009.

[8] J. Bioucas-Dias and M. A. T. Figueiredo. Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing. In IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, volume 1, pages –, June 2010.

[9] Djallel Bouneffouf and Inanc Birol. Sampling with minimum sum of squared similarities for nyström-based large scale spectral clustering. In Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, pages 2313–2319. AAAI Press, 2015.

[10] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn., 3(1):1–122, January 2011.

[11] Tony Cai, Weidong Liu, and Xi Luo. A constrained 1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association, 106(494):594–607, 2011.

[12] Nicolas Dobigeon, Jean-Yves Tourneret, Cedric Richard, Jose Carlos M. Bermudez, Stephen McLaughlin, and Alfred Hero. Nonlinear unmixing of hyperspectral images. IEEE Signal Processing Magazine, pages 82–94, 2013.

[13] Petros Drineas and Michael W. Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. J. Mach. Learn. Res., 6:2153–2175, December 2005.

[14] Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik. Spectral grouping using the nyström method. IEEE Trans. Pattern Anal. Mach. Intell., 26(2):214–225, January 2004.

[15] Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, Advances in Neural Information Processing Systems 23, pages 604–612. Curran Associates, Inc., 2010.

[16] Michael A. Golberg. Numerical solution of integral equations. Plenum Press, 1990.

[17] M. Iordache, J. Bioucas-Dias, and A. Plaza. Hyperspectral unmixing with sparse group lasso (accepted). In IEEE International Geoscience and Remote Sensing Symp.- IGARSS, volume 1, pages 1–4, July 2011.

[18] M. D. Iordache, J. M. Bioucas-Dias, and A. Plaza. Sparse unmixing of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, 49(6):2014–2039, June 2011.

[19] Ulrike Luxburg. A tutorial on spectral clustering. Statistics and Computing, 17(4):395–416, December 2007.

[20] J. MacQueen. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[21] G. Moser and S. B. Serpico. Combining Support Vector Machines and Markov Random Fields in an Integrated Framework for Contextual Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 51(5):2734–2752, May 2013.

[22] G. Moser, S.B. Serpico, and J.A. Benediktsson. Land-cover mapping by markov modeling of spatial-contextual information in very-high-resolution remote sensing images. Proceedings of the IEEE, 101(3):631–651, March 2013.

[23] J. M. P. Nascimento and J.M. Bioucas-Dias. Vertex Component Analysis: A Fast Algorithm to Unmix Hyperspectral Data. IEEE Trans. on Geosci and Remote Sens., 43:898–910, 2005.

[24] Jose M. P. Nascimento and Jose M. Bioucas-Dias. Nonlinear Mixture Model for Hyperspectral Unmixing. Proc. SPIE, 7477:74770I–74770I–8, 2009.

[25] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, Advances in Neural Information Processing Systems 14, pages 849–856. MIT Press, 2002.

[26] H. J. Reinhardt. Analysis of approximation methods for differential and integral equations. Springer-Verlag, 1985.

[27] Robert A. Schowengerdt. Remote Sensing, Third Edition: Models and Methods for Image Processing. Academic Press, Inc., Orlando, FL, USA, 2006.

[28] Yuliya Tarabalka, Mathieu Fauvel, Jocelyn Chanussot, and Jn Atli Benediktsson. Svm and mrf- based method for accurate classification of hyperspectral images. IEEE Geoscience and Remote Sensing Letters, pages 640–736, 2010.

[29] R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B, 58:267–288, 1996.

[30] Olivier Eches Nicolas Dobigeon Jean-Yves Tourneret. Enhancing Hyperspectral Image Unmixing with Spatial Correlations with Spatial Correlations. IEEE Transactions on Geoscience and Remote Sensing, pages 4239–4247, 2011.

[31] R Warren and S Osher. Hyperspectral unmixing by the alternating direction method of multipliers. Inverse Problems and Imaging, 14(3):917–933, 2015.

[32] Michael E. Winter. N-FINDR: An Algorithm for Fast Autonomous Spectral Endmember Determination in Hyperspectral Data. Proc. SPIE, Imaging Spectrometry V, 3753:266–277, 1999.

[33] Chia Chye Yee and Yves Atchade. On the Sparse Bayesian Learning of Linear Models. Communication in Statistics: Publication to Appear, 2015.

[34] Chia Chye Yee and Yves Atchade. Simultaneous Unmixing and Classification of Hyperspectral Images. In Progress, 2016.

[35] A. Zare and P. Gader. Sparsity promoting iterated constrained endmember detection in hyperspectral imagery. IEEE Geoscience and Remote Sensing Letters, 4(3):446–450, July 2007.

[36] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B, 67:301–320, 2005.