IMPROVED LOCAL SPECTRAL UNMIXING OF HYPERSPECTRAL DATA USING AN ALGORITHMIC REGULARIZATION PATH FOR COLLABORATIVE SPARSE REGRESSION

L. Drumetz\textsuperscript{a}, G. Tochon\textsuperscript{b}, M.A. Veganzones\textsuperscript{c}, J. Chanussot\textsuperscript{b,d}, C. Jutten\textsuperscript{a}

\textsuperscript{a}Université Grenoble Alpes, \textsuperscript{b}Grenoble Institute of Technology, \textsuperscript{c}CNRS, GIPSA-lab, Grenoble, France
\textsuperscript{d}Deparlement of Mathematics, University of California, Los Angeles (UCLA), USA

\textbf{ABSTRACT}

Local Spectral Unmixing (LSU) methods perform the unmixing of hyperspectral data locally in regions of the image. The endmembers and their abundances in each pixel are extracted region-wise, instead of globally to mitigate spectral variability effects, which are less severe locally. However, it requires the local estimation of the number of endmembers to use. Algorithms for intrinsic dimensionality (ID) estimation tend to overestimate the local ID, especially in small regions. The ID only provides an upper bound of the application and scale dependent number of endmembers, which leads to extract irrelevant signatures as local endmembers, associated with meaningless local abundances. We propose a method to select in each region the best subset of the locally extracted endmembers. Collaborative sparsity is used to detect spurious endmembers in each region and only keep the most influential ones. We compute an algorithmic regularization path for this problem, giving access to the sequence of successive active sets of endmembers when the regularization parameter is increased. Finally, we select the optimal set in the sense of the Bayesian Information Criterion (BIC), favoring models with a high likelihood, while penalizing those with too many endmembers. Results on real data show the interest of the proposed approach.

\textbf{Index Terms}— Hyperspectral imaging, local spectral unmixing, binary partition tree, collaborative sparsity, regularization path

\section{1. INTRODUCTION}

Spectral Unmixing (SU) is one of the most important applications in hyperspectral imaging for remote sensing \cite{1}. Because of the limited spatial resolution of hyperspectral images (HSI), observed pixels can account for the contribution of several materials present in the field of view of the sensor during the acquisition. SU is a source separation problem whose goal is to automatically identify the spectral signatures of the materials present in the observed scene (called endmembers) and then to estimate their proportions in each pixel (called abundances). Most SU algorithms assume a Linear Mixing Model (LMM), in which the observations are modeled as linear combinations of the endmembers’ spectra, weighted by the fractional abundances \cite{2}. The abundances are additionally subject to the abundance nonnegativity constraint (ANC) and a sum to one constraint (ASC). Therefore they lie in the probability simplex, and the data lie in another simplex whose vertices are the endmembers. The main two limitations of this approach have been identified as nonlinearity and spectral variability. Nonlinear mixing can happen locally when the incident light undergoes multiple reflections (e.g. in urban scenarios, tree canopies, or particulate media such as sand) \cite{3}. Spectral variability \cite{4,5} can affect the endmembers’ signatures locally depending on the geometry of the scene (topography and changing illumination conditions) or because of changes in the physico-chemical composition of the materials (e.g. soil moisture content) \cite{6,7,8}. Local Spectral Unmixing (LSU) is a technique in which the unmixing is performed in local regions of the image, instead of a whole HSI. The idea is that in smaller local regions, the effects of spectral variability are not so severe. Another motivation is that nonlinear effects are usually very localized in space. The key issue is then to choose the regions on which to perform LSU in an appropriate way. The first methods of the literature to perform the unmixing locally were using sliding windows \cite{9,10}. Segmentations of the HSI could also be used. The more recent work of \cite{11} uses Binary Partition Trees (BPT) \cite{12} to compute an unmixing-driven hierarchical segmentation of hyperspectral datasets. An optimal partition of this hierarchy in the sense of local reconstruction performance is computed, providing meaningful regions for LSU. However, one of the drawbacks of this approach is that the Intrinsic Dimensionality (ID) of each region has to be estimated with one of the algorithms of the literature \cite{13}, to choose the local number of endmembers to use. As shown in \cite{14}, the ID can be severely overestimated in small regions of the image, or for a large number of spectral bands. Besides, the ID only provides an upper bound of the number of endmembers to use in practice (even for large regions or full images), which is subjective, application and scale dependent. As a result, it often happens that irrelevant endmembers are extracted in many regions, and are associated to very sparse abundance maps. To avoid this, we propose an automatic method to identify and discard irrelevant endmembers in each region during the segmentation step. We make use of collaborative sparsity \cite{15}, allowing us to retain only the significant endmember contributions in the region. There is no need to tune any parameters, because we are able to compute a regularization path for the underlying optimization problem. The remainder of this paper is organized as follows: section 2 introduces in more detail the LSU scheme of \cite{11}. We present our contribution in section 3. Section 4 presents the results of the proposed approach on a real dataset. Finally, section 5 gathers some concluding remarks.

\section{2. BINARY PARTITION TREE BASED LSU}

In this section, we briefly summarize the BPT-based LSU approach \cite{11}, which is the basis of the present work. The construction of a BPT is conceptually simple: starting from an initial partition of the image (typically an oversegmented partition, obtained using a segmentation algorithm), the two most similar adjacent regions are iteratively merged, building a tree structure. An example is shown in Fig. 1 (a). To obtain an unmixing driven process, we model each region by the set of endmembers extracted by any endmember extraction algorithm (EEA) of the literature (here the Vertex Component Analysis (VCA) \cite{16}). This implies to estimate the ID of each region prior to the endmember extraction. One can for instance use the Random Matrix Theory (RMT) based algorithm of \cite{17}, with the...
We denote by $X$ matrices, proposed in [18]. The abundance maps of each region are also estimated, assuming the LMM holds locally, in each region $R$:

$$X \in R \times N$$

where $X$ is the hyperspectral pixels of the region ($L$ and $N$ are the number of spectral bands, and the number of pixels in the region, respectively). $S \in R \times d$ is the local endmember matrix, extracted by VCA, and $A \in R \times d \times N$ is the abundance matrix, estimated using the classical Fully Constrained Least Squares Algorithm (FCLS) [19]. $d$ is the estimated ID in region $R$. $E \in R \times N$ is an additive noise, usually assumed to be Gaussian. The BPT is a hierarchical representation of the regions of the image, and in order to get a single partition out of all those possible using the BPT structure, the tree has to be pruned, as shown in Fig. 1 (b). Here, we are able to recover an optimal partition using an optimization process on the hierarchy defined by the BPT. We obtain a partition with a desired number of regions which minimizes an energy based on the reconstruction errors in each region (see [11] for details). In this sense, this segmentation is optimal in terms of SU performance. The local endmembers and abundances can then be recovered for each region of this partition.

3. PROPOSED REGION MODEL

3.1. Motivation

The BPT based LSU scheme described in the previous section has proved useful, but due to the ID overestimation issue in local regions, and to the fact that the ID does not always match the expected number of endmembers in an image, the extraction of spurious endmembers is frequent. To show this, we built a BPT on the Washington DC mall dataset (shown in Fig. 2 (a)) acquired by the HYDICE sensor, in the visible and near infrared, with a spatial resolution of 2.8m. The initial segmentation was obtained using a mean shift clustering algorithm [20], giving an initial partition with 5760 regions. From the BPT, we show in Fig. 2 (b) a plot of the estimated ID as a function of the region size (before the pruning). The estimated ID seems relatively high, up to 70 for the largest regions of the BPT, and regularly over 10 for small regions, (even for regions of 100 pixels or less). To provide more evidence for this phenomenon, we compare visually in Fig. 3 two regions of the optimal segmentation obtained by keeping around 500 regions. Although a visual inspection can be misleading, we do not expect more than three, perhaps four endmembers in the region on the left, and one or two on the region on the right. However, the estimated IDs were respectively of 20 and 9. The local endmembers extracted by VCA are shown in the bottom row. They clearly show that many signatures are very similar and are probably associated to the same materials. For instance, on the left there are at least 6 signatures with a very low spectrum, all associated to shadowed areas of the region. For the rooftop region, only 3 signatures are significantly different from one another in terms of spectral distance (neglecting scaling effects of the signatures). Therefore, most of the abundance maps of these two regions are very sparse (sometimes only non negligible in extremely small regions or isolated pixels), or only there to fit the noise. This means that most of them do not correspond to significant instances of the same materials, and are then not meaningful in terms of spectral variability. However, because they are given the same weight as legitimate endmembers in the region model, they can affect the whole BPT construction and pruning. To get better interpretable results in local regions, these dummy endmembers should be discarded in the BPT construction process.

3.2. Collaborative sparsity in LSU

In order to solve this issue, we propose a modified region model for the BPT construction which eliminates the spurious endmembers extracted in a given region. To select the endmembers which should be discarded in the unmixing process, we want to force the ones whose abundance maps are already very sparse or very low in most pixels to be zero everywhere, for each region. To do that, we are going to replace the abundance estimation step within a region with FCLS by a collaborative sparse unmixing step [15, 21]. Indeed, the mixed $E_{2,1}$ norm [22] $\sum_{p=1}^{N} \|a_{R,p}\|_2$, where $a_{R,p}$ is the $p^{th}$ row of $A_R$ used in this type of sparse regression problems encourages row-wise sparsity in the abundance matrix. This means that some endmembers in the region will have zero abundance maps for all the pixels of the region. Then we will only have to discard those from the set of local endmembers. Unmixing the region with collaborative sparsity boils down to solving the following convex optimization problem:

$$\arg \min_{A} \frac{1}{2} \|X - S A\|_F^2 + \lambda \sum_{p=1}^{N} \|a_{R,p}\|_2 + \mathcal{I}_{\Delta_{R}}(A_R),$$

where $\mathcal{I}_{\Delta_{R}}$ is the indicator function of the probability simplex (0 inside the simplex, $+\infty$ otherwise), applied independently to each column of $A_R$, and $\lambda$ is a regularization parameter. This problem can be relatively easily solved using proximal methods, such as...
use this, we introduce split variables to decouple the different terms in the optimization. We then rewrite problem (2) in the equivalent formulation using linear constraints, which is suitable for the ADMM:

$$\arg \min_{\mathbf{X}_R} \frac{1}{2} \left\| \mathbf{X}_R - \mathbf{S}_R \mathbf{A}_R \right\|_F^2 + \lambda_R \left\| \mathbf{U}_R \right\|_{2,1} + \mathcal{I}_{\Delta_{d,R}} \left( \mathbf{V}_R \right)$$

s.t. $\mathbf{U}_R = \mathbf{A}_R, \mathbf{V}_R = \mathbf{A}_R$.

(3)

The ADMM then minimizes an augmented Lagrangian w.r.t. to all the blocks of variables involved, alternatively and iteratively, before updating the Lagrange multipliers (called $\mathbf{C}_R$ and $\mathbf{D}_R$ in Algorithm 1) in a so-called dual update.

However, there are two problems with this approach. The first is that since the linear constraints of the ADMM are only satisfied asymptotically, in practice the entries of the supposedly discarded rows of the abundance matrix are often not exactly zero, but very small values. Then an arbitrary thresholding step is required to eliminate endmembers with a small contribution [15, 21]. The second is that to obtain the appropriate sparsity level, the regularization parameter $\lambda_R$ needs to be tuned in every region. A grid search over a set of parameters in each region would be very computationally costly and would require a criterion to select the best run of the algorithm. We will see that we can find solutions for both issues.

3.3. Obtaining an algorithmic regularization path

In order to tackle both the regularization parameter issue and the inexact sparsity of the collaborative sparse regression at once, we would like to obtain the regularization path of the solution, as a function of $\lambda_R$. Regularization paths can sometimes be computed easily, for instance on the LASSO (for Least Absolute Shrinkage Selection Operator) problem [24]. However, for more complex problems, such as ours, there is no way, to our knowledge, to obtain this regularization path easily. A convenient workaround for this is to compute a so-called ADMM algorithmic regularization path, introduced in [25]. This approach is able to use the ADMM to quickly approximate the sequence of active supports of the variable of interest, when the regularization parameter increases, for certain sparsity regularized least squares problems. Even though there are as of today no theoretical guarantees on the efficiency of this algorithm, it was experimentally shown to be able to efficiently approximate the true sequence of active sets on several problems [25], including the LASSO. Here, we propose to extend this algorithm to collaborative sparsity. Since exactly solving the optimization problem for a large number of regularization parameters would be too time consuming, we are more interested in finding the active set of endmembers when the weight of the sparsity term increases w.r.t. this of the data fit term. The idea is, for each region involved in the construction of the BPT, to find a sequence of endmember matrices, whose number of endmembers are decreasing from $d_R$ to zero (when the model is fully sparse). Each new matrix contains the same endmembers as the previous one, except for one, which is the next endmember to be discarded when the weight of the sparsity term gets more important. To do that, we modify the ADMM in order to quickly obtain the support of the regularization path, for each region. An iteration of the ADMM is carried out for a very small value of the regularization parameter (guaranteeing a fully dense solution). Then, the variables obtained after this iteration are used as a warm start for another iteration with a new slightly higher regularization parameter. By repeating this for several iterations with higher and higher regularization parameters, the split variable $\mathbf{U}_R$, which undergoes a block soft thresholding (the proximal operator of the $L_{2,1}$ norm [23]) becomes increasingly sparse. Since we are using warm starts, and because regularization parameters vary slowly, even if the ADMM is not fully converged at each iteration, the support of the active set is encoded in $\mathbf{U}_R$, often in one iteration only, long before this active set is propagated to $\mathbf{A}_R$ (this will be the case only at convergence, when the constraints of problem (3) are satisfied). With these modifications, we obtain Algorithm 1. The notation $\left\| \mathbf{U}_R \right\|_{2,0}$ denotes the number of nonzero rows of the matrix $\mathbf{U}_R$, and $\rho$ is the barrier parameter of the ADMM. soft denotes the block soft thresholding operator with scale parameter in index, and $\text{proj}_{\Delta_{d,R}}$ denotes the projection on the probability simplex, performed with the algorithm of [26]. These operators are applied columnwise. Here, we are using a geometric progression for $\gamma_R$ (we changed the notation of this regularization parameter, because we do not completely solve the optimization problem (3), whose common ratio is $t$). This value should be small to approximate the active sets of the regularization path well enough. The regularization space can be explored very quickly since the algorithm provides at most $d_R$ endmember subsets of the full endmember set extracted by VCA, that need to be tested after this process. In practice we chose $\gamma_R = 10^{-4}$ and $t = 1.04$, which allows $\gamma_R$ to sweep from $10^{-4}$ to $5.10^{-6}$ in 500 iterations (less than what is usually required in practice to reach a fully sparse model).

Data: $\mathbf{X}_R, \mathbf{S}_R$

Result: The sequence of $\mathbf{U}_R^i$, $i = 0, \ldots, i_{\text{max}}$

Initialize $\mathbf{A}_R^0$ and choose $\gamma_R$ and $t > 0$:

while $\left\| \mathbf{U}_R^i \right\|_{2,0} \neq 0$ do

$$\gamma_R \leftarrow \gamma_R^t$$

$$\mathbf{U}_R^i \leftarrow \text{soft}_{\gamma_R^t / t} (\mathbf{A}_R^{i-1} - \mathbf{C}_R^{i-1})$$

$$\mathbf{A}_R^i \leftarrow (S_R^i \mathbf{S}_R + 2 \rho \mathbf{D}_R^{i-1})^{-1} (S_R^i \mathbf{X}_R + \rho (\mathbf{U}_R^i + \mathbf{V}_R^i + \mathbf{C}_R^i + \mathbf{D}_R^i))$$

$$\mathbf{V}_R^i \leftarrow \text{proj}_{\Delta_{d,R}} (\mathbf{A}_R^i - \mathbf{D}_R^i)$$

$$\mathbf{C}_R^i \leftarrow \mathbf{C}_R^{i-1} + \mathbf{U}_R^i - \mathbf{A}_R$$

$$\mathbf{D}_R^i \leftarrow \mathbf{D}_R^{i-1} + \mathbf{V}_R^i - \mathbf{A}_R^i$$

$i \leftarrow i + 1$

end

Algorithm 1: Algorithmic regularization path for problem (3).

3.4. Selecting the best model

Using the active sets, we can store a sequence of $d_R$ sparser and sparser candidate endmember matrices (denoted as $\mathbf{S}_k$). The last step is to select the optimal active set in the sense of some criterion. We used the Bayesian Information Criterion (BIC) [27], which helps choosing from a set of candidate models, by favoring those with an important likelihood, and penalizing those with a high number of parameters. This criterion assumes that the noise is spectrally and spatially white, a strong but still widely used assumption. A candidate model $\mathbf{M}_i$ is made of one of the $\mathbf{S}_k$ and the corresponding estimated abundances with FCLSU. For our problem, the BIC writes [28]:

$$BIC_i = \ln(L) P_i + L \ln \left( \frac{\left\| \mathbf{X}_R - \mathbf{S}_k \mathbf{A}_k \right\|_F^2}{L} \right),$$

where $P_i$ is the number of endmembers in $\mathbf{S}_k \in \mathbb{R}^{L \times K}$, $\mathbf{A}_k$ is the abundance matrix estimated by FCLSU using the data and the endmember matrix $\mathbf{S}_k$. The best model is simply the one minimizing the BIC value. In our case, to alleviate the computational load, we perform these FCLSU steps from the smallest endmember matrices to the bigger ones, and stop when the BIC value has increased for three consecutive iterations, to avoid performing numerous useless abundance estimation methods in each region. The flowchart of the proposed region model is shown in Fig. 4.

4. RESULTS

To assess the impact of the proposed modifications to the region model, we built two BPTs on the Washington Mall dataset, one with-
We have proposed a new region model for local spectral unmixing (LSU) based on binary partition trees (BPT). This model still compares the local endmembers extracted in each region, but is able to cope with the extraction of spurious endmembers due to local ID estimation. We get rid of those endmembers by using collaborative sparsity, and avoid parameter tuning in each region by computing an algorithmic regularization path for the resulting optimization problem. We are then able to select the best sparse model by using the Bayesian Information Criterion (BIC). The results show that the proposed modifications allow to eliminate the redundant information in each region, without penalizing the unmixing performance. Future work will include an extension of the proposed method for the unmixing of complete images rather than regions for LSU. In combination with the ideas of [21], a completely blind and parameter free algorithm for simultaneous SU and ID estimation can be designed.
6. REFERENCES


