Learning Sparse Codes for Hyperspectral Imagery

Adam S. Charles, Bruno A. Olshausen, and Christopher J. Rozell

Abstract

The spectral features in hyperspectral imagery (HSI) contain significant structure that, if properly characterized could enable more efficient data acquisition and improved data analysis. Because most pixels contain reflectances of just a few materials, we propose that a *sparse coding* model is well-matched to HSI data. Sparsity models consider each pixel as a combination of just a few elements from a larger dictionary, and this approach has proven effective in a wide range of applications. Furthermore, previous work has shown that optimal sparse coding dictionaries can be learned from a dataset with no other *a priori* information (in contrast to many HSI "endmember" discovery algorithms that assume the presence of pure spectra or side information). We modified an existing unsupervised learning approach and applied it to HSI data (with significant ground truth labeling) to learn an optimal sparse coding dictionary. Using this learned dictionary, we demonstrate three main findings: i) the sparse coding model learns spectral signatures of materials in the scene and locally approximates nonlinear manifolds for individual materials, ii) this learned dictionary can be used to infer HSI-resolution data with very high accuracy from simulated imagery collected at multispectral-level resolution, and iii) this learned dictionary improves the performance of a supervised classification algorithm, both in terms of the classifier complexity and generalization from very small training sets.

Index Terms

Remote Sensing, Hyperspectral Imagery, Sparse Coding, Dictionary Learning, Multisepctral Imagery, Deblurring, Inverse Problems, Material Classification

I. Introduction

Hyperspectral imagery (HSI) is a spectral imaging modality that obtains environmental and geographical information by imaging ground locations from airborne or spaceborne platforms. While multispectral imagery (MSI) acquires data over just a few (e.g., 3-10) irregularly spaced spectral bands, HSI typically uses hundreds of contiguous bands that are regularly spaced from infrared to ultraviolet. For example, the Worldview II MSI satellite [1] uses eight bands to represent the wavelengths from 0.435μm to 1.328μm, while typical HSI has approximately 60 bands over the same range in addition to many more bands at higher wavelengths. With spatial resolutions as

ASC and CJR are with the School of Electrical and Computer Engineering at the Georgia Institute of Technology. BAO is with the Helen Wills Neuroscience Institute and School of Optometry at the University of California, Berkeley. A preliminary version of portions of this work appeared in [18]. The authors are grateful to Charles Bachmann at the Naval Research Laboratory for generously providing the Smith Island HSI data set and the associated ground truth labels, as well as John Greer and Jack Culpepper for helpful discussions about this work.

low as 1m, the increased spectral resolution of HSI means that estimated ground reflectance data can be used to determine properties of the scene, including material classification, geologic feature identification, and environmental monitoring. A good overview of HSI and the associated sensors can be found in [36].

Exploiting HSI is often difficult due to the particular challenges of the remote sensing environment. For example, even "pure" pixels composed of a single material would have reflectance spectra that lie along a nonlinear manifold due to variations in illumination, view angle, material heterogeneity, scattering from the local scene geometry, and the presence of moisture [5], [36]. Additionally, pure pixels are essentially impossible to actually observe due to material mixtures within a pixel and scattering from adjacent areas [36]. One of the most common approaches to determining the material present in a given pixel $\boldsymbol{x}$ (called "spectral unmixing" [37]) is to use a linear mixture model such as

$$\boldsymbol{x} = \sum_{k=1}^{M} \boldsymbol{\phi}_k a_k + \epsilon, \tag{1}$$

where $\{\boldsymbol{\phi}_k\}$ is a dictionary of approximation elements, $\{a_k\}$ are the decomposition coefficients and ϵ is additive noise. Note that $\{\boldsymbol{x}, \boldsymbol{\phi}_k, \epsilon\} \in \mathbb{R}^N$, where N is the number of spectral bands and the vectors are indexed by λ (which is suppressed in our notation). When the dictionary represents spectral signatures of the various material components present in the scene, they are typically called "endmembers" and the resulting coefficients (assumed to sum to one) represent the material abundances in each pixel. The endmember vectors are conceptualized as forming a convex hull about the HSI data (e.g., see the red vectors in Figure 1). Such a decomposition is often used for detecting the presence of a material in the scene or classifying the materials present in a pixel. A number of methods have been proposed for determining endmembers, including algorithms which select endmembers from the data based on a measure of pixel purity [48] or the quality of the resulting convex cone [53], tools that assist in the manual selection of endmembers from the data [9], algorithms which optimize endmembers for linear filtering [12], methods based on finding convex cones using principal component analysis (PCA) or independent component analysis (ICA) decompositions [21], [24], [27], [32], iterative statistical methods that optimize the resulting convex cone [10], and iterative measures to select optimal endmember sets from larger potential sets [50]. However, these algorithms either rely on postulating candidate endmember sets for initialization [50], assume the existence of pure pixels in the scene [48], [53], attempt to encompass the data within a cone rather than directly represent the data variations [9], [10], [32], use orthogonal linear filters to attempt to separate out highly non-orthogonal spectra [12], or attempt to determine spectral statistics from decompositions in the spatial dimensions rather than the spectral dimension. [21], [24]. None of these methods attempt to directly *learn* from the spectral data a good representation of the low-dimensional, non-linear spectral variations inherent in HSI.

In addition to the difficulties determining the basic spectral components of an HSI dataset, there are many resource

costs (i.e., time, money, computation, availability of sensor platforms) that result from the high dimensionally of the data. During data acquisition, the high resolution of HSI data comes at the expense of sophisticated sensors that are costly and require relatively long scan times to get usable SNRs. After data acquisition, it is evident that reducing the dimensionality while retaining the exploitation value of the data would save significant computational and storage resources. If the higher-order statistics of the HSI data can be characterized, this information can be used to perform both dimensionality reduction of existing high-dimensional data and high-resolution inference from low-resolution data (collected from either a cheaper MSI sensor or a modified HSI sensor measuring coarse spectral resolution, thereby lowering scan times). One common approach to dimensionality reduction is PCA. However, the underlying Gaussian model in PCA means that it can only capture pairwise correlations in the data and not the higher-order (and non-Gaussian) statistics present in HSI data.

Following on developments in the computational neuroscience community, the signal processing community has recently employed signal models based on the notion of *sparsity* to characterize high-order statistical dependencies in data and yield state-of-the-art results in many signal and image processing algorithms [20]. Specifically, this approach models a noisy measurement vector $\boldsymbol{x}$ as being generated by a linear combination of just a few elements from the dictionary $\{\boldsymbol{\phi}_k\}$. This is the same model as in (1), but where the coefficients are calculated to have as few non-zero elements as possible. Much like PCA, sparse coding can be viewed as a type of dimensionality reduction where a high dimensional dataset is expressed in a lower dimensional space of active coefficients. However, while PCA calculates just a few principal components and uses essentially all of them to represent each pixel, sparse coding models typically employ a larger dictionary but use only a few of these elements to represent each pixel. When cast in terms of a probabilistic model, this sparsity constraint corresponds to a non-Gaussian prior that enables the model to capture higher order statistics in the data.

Due to the high spatial resolution of modern HSI sensors (resulting in just a few dominant materials in a pixel), sparsity models seem especially relevant for this sensing modality. In fact, initial research into sparsity models for spectral unmixing in HSI has shown promising results [29], [33]. While a sparse decomposition can be estimated for any dictionary, previous research [43] has shown that unsupervised learning techniques can be used in conjunction with an example dataset to iteratively learn a dictionary that admits optimally sparse coefficients (without requiring the dataset to contain any "pure" signals that correspond to a single dictionary element). These methods leverage the specific high-order statistics of the example dataset to find the underlying low-dimensional structure that is most efficient at representing the data.

In contrast to the typical endmember model described above, the sparse coding model does not assume that the data lie within the convex hull of the dictionary. Instead, the learned sparse coding dictionary elements will tend to look like the basic spectral signatures comprising the scene (early encouraging evidence of this can be found

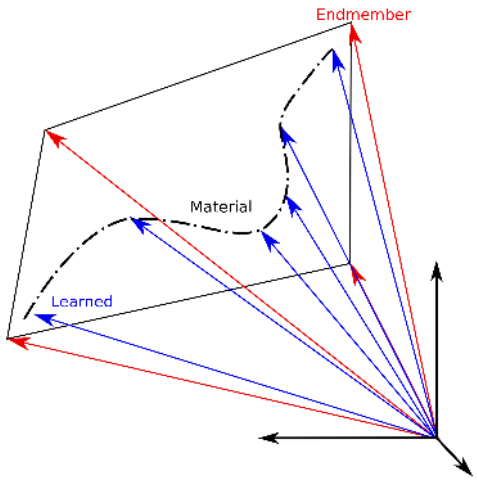

Fig. 1. Typical endmember analysis uses vectors that compose a convex hull around the data. In this stylized illustration, the data manifold is indicated by the dashed line and the red vectors represent the endmembers. In contrast, a learned dictionary for sparse coding attempts to learn a local approximation of the nonlinear data characteristics directly (indicated here by blue vectors).

in [28]). In fact, the sparse coding model may actually learn several dictionary elements to represent some types of materials, especially when that material spectra demonstrates highly nonlinear variations within the scene. Because of the sparsity constraint, one would expect these learned dictionaries to reflect the specific statistics of the HSI data by locally approximating these nonlinear data manifolds [45] (as illustrated in Figure 1, and in contrast to typical endmember models that form a convex hull containing the data).

We have modified the unsupervised learning approach described in [43] and applied it to HSI data to learn a dictionary that is optimized for sparse coding. Importantly, the HSI dataset used in this study has significant ground truth labeling of material classes making it possible to examine the characteristics of the learned dictionary relative to the data. Using this learned dictionary, we make three main contributions. First, we show that the sparse coding model learns meaningful dictionaries that correspond to known spectral signatures: they locally approximate nonlinear data manifolds for individual materials, and they convey information about environmental properties such as moisture content in a region. Second, we generate simulated imagery at MSI-level resolution and show that the learned HSI dictionaries and sparse coding model can be effectively used to infer HSI-resolution data with very high accuracy (even for data of the same region collected in a different season). Finally, we use ground truth labels for the HSI data to demonstrate that a sparse coding representation improves the performance of a supervised classification algorithm, both in terms of the classifier complexity (i.e., classification time) and the ability of the classifier to generalize from very small training sets.

II. Background and Related Work

A. Methods

Given a pixel $\boldsymbol{x} \in \mathbb{R}^N$ and a fixed dictionary $\{\boldsymbol{\phi}_k\}$ with $\boldsymbol{\phi_k} \in \mathbb{R}^N$ for $k \in [1, \ldots, M]$, the goal of sparse coding is to find a set of coefficients that represent the data well using as few non-zero elements as possible. Written mathematically, the goal is to minimize an objective function that combines data fidelity and a sparsity-inducing penalty. A common choice is to use a regularized least-squares objective function such as,

$$J_\gamma\left(\boldsymbol{x}, \{a_k\}, \{\boldsymbol{\phi}_k\}\right) = \left\| \boldsymbol{x} - \sum_{k=1}^{M} \boldsymbol{\phi}_k a_k \right\|_2^2 + \gamma \sum_{k=1}^{M} |a_k|, \tag{2}$$

with mean-squared error as the data-fidelity term, the ℓ^1 norm (i.e., the sum of the coefficient magnitudes) as the sparsity inducing penalty, and γ a scalar parameter trading off between these two terms [19]. This objective is convex in the coefficients when the dictionary is fixed, meaning that solving $\{a_k\} = \arg\min_{\{a_k\}} J_\gamma\left(\boldsymbol{x}, \{a_k\}, \{\boldsymbol{\phi}_k\}\right)$ is a tractable optimization. Significant progress has recently been made developing solvers that are customized for this specific optimization program and run considerably faster than general purpose algorithms [15], [22], [30], [38]. This general approach is applicable directly to HSI with one small modification: we constrain the coefficients to be non-negative ($a_k \geq 0$) to maintain physical correspondence between the coefficients and the relative abundance of material spectra present in the scene. Due to its wide use in the community, its ability to enforce positive coefficients without a sum-to-one constraint, and established reputation for quick convergence, we use the specialized optimization package described in [38] to solve this constrained optimization and calculate sparse coefficients. While other solvers have been explored in the specific context of HSI [11], [46] that may be faster in some settings, many of these HSI-specific solvers include additional constraints which we do not employ (e.g. $\sum_k |a_k| = 1$). The framework we present here is largely agnostic to the specific solver as long as it returns accurate solutions, so other choices could be substituted if there were advantages for a given application. A detailed analysis of various algorithms to optimize (2) in the context of HSI unmixing is given in [33].

An alternate interpretation of the cost function in (2) is to consider the problem as Bayesian inference. With a white Gaussian noise distribution on ϵ, the likelihood on the data vector given the coefficients $p(\boldsymbol{x}|\{a_k\})$ is also Gaussian. We assume a Laplacian prior probability distribution on the coefficients $\{a_k\}$ because the high kurtosis of this distribution (i.e., the peakiness of the distribution around zero) encourages coefficients to be close to zero. Using Bayes' rule, the resulting (unnormalized) posterior is

$$\begin{aligned} p(\{a_k\}|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\{a_k\})p(\{a_k\}) \\ &\propto e^{\left(-\frac{1}{2\sigma_\epsilon^2}\|x - \sum_k \boldsymbol{\phi}_k a_k\|_2^2\right)} e^{\left(-\frac{\sqrt{2}}{\sigma_a}\sum_k |a_k|\right)}, \end{aligned}$$

where σ_a is the standard deviation on the prior over a_k (i.e., the signal energy) and σ_ϵ^2 is the noise variance. Taking the negative logarithm of the posterior results in the cost function (2). Thus the maximum a posteriori (MAP) estimate on the coefficients is equivalent to minimizing (2) with $\gamma = 2\sqrt{2}\sigma_\epsilon^2/\sigma_a$. This Bayesian formulation allows us to naturally extend the sparse approximation problem to more general observation models and inverse problems, such as the high resolution inference task described in Section III-B and similar inverse problems described in [55]. Note also that the sparse prior introduces a non-Gaussianity into the model that is critical for capturing the high-order data statistics. An approach such as PCA that fundamentally assumes a Gaussian data model can only learn from pairwise correlations in the data, and is therefore unable to capture the higher-order statistics.

To learn an optimal dictionary for sparse coding, we would like to minimize the cost function $J_\gamma\left(\boldsymbol{x}, \{a_k\}, \{\boldsymbol{\phi}_k\}\right)$ with respect to both the coefficients and the dictionary. Unfortunately, this objective function is not jointly convex in the coefficients and the dictionary, making global minimization prohibitive. The work presented in [43], [44] takes a typical approach to this type of problem, using a variational method that alternates between minimizing (2) with respect to the coefficients given the current dictionary, then taking a gradient descent step over the dictionary elements given the calculated coefficients. Specifically, after inferring coefficients for a randomly selected batch of data, the dictionary learning proceeds by descending the gradient of $J_\gamma\left(\boldsymbol{x}, \{a_k\}, \{\boldsymbol{\phi}_k\}\right)$ with respect to $\{\boldsymbol{\phi}_k\}$, resulting in the learning rule:

$$\boldsymbol{\phi}_l \leftarrow \boldsymbol{\phi}_l + \mu \left\langle a_l \left(\boldsymbol{x} - \sum_{k=1}^{M} \boldsymbol{\phi}_k a_k \right) \right\rangle, \tag{3}$$

where μ is the step-size of the update (possibly varying with iteration number) and $\langle \cdot \rangle$ indicates the average over the current batch of data. This approach has the drawback that a trivial solution of enlarging the norm of the dictionary elements can always produce smaller total energy in the coefficients, which we avoid by renormalizing the dictionary elements to have unit norm after each learning step. Returning briefly to the Bayesian interpretation detailed above, it is worth noting that the cost function $J_\gamma\left(\boldsymbol{x}, \{a_k\}, \{\boldsymbol{\phi}_k\}\right)$ may be viewed as an approximation to the negative log-likelihood of the model [44]. Therefore, gradient descent on $J_\gamma\left(\boldsymbol{x}, \{a_k\}, \{\boldsymbol{\phi}_k\}\right)$ is tantamount to maximizing the log-likelihood of the model. As with the coefficient optimization, other algorithms have been proposed for the learning step that could be substituted for this steepest decent approach. In particular, many other methods (including the recently proposed K-SVD) use second order information in the learning step to reduce the number of learning iterations required for convergence (though this may come at the cost of increasing the batch size per iteration to get better estimates for the update step) [2], [3].

The results in [43], [44] demonstrate that this unsupervised approach can start with an unstructured random dictionary and recover known sparse structure in simulated datasets, as well as uncover unknown sparse structure in complex signal families such as natural images. We again adopt this general approach with a small modification:

we constrain the dictionary elements to be non-negative ($\boldsymbol{\phi}_k \geq 0$) to maintain physical correspondence with spectral reflectances. To be concrete, the dictionary learning method we use is specified in Algorithm 1, and we determine convergence visually by when the dictionary elements stopped adapting. In our experience, most of the dictionary elements were well-converged by 1000 iterations of the learning step (approximately 50 minutes of computation on an 8-core Intel Xeon E5420 with 14GB of DDR3 RAM). Some dictionary elements corresponding to less prominent materials (that are randomly selected less often during learning) seem to require 10,000-20,000 learning iterations to converge (approximately 10-15 hours on the same machine). We often conservatively let the algorithm run for 20,000 to 80,000 iterations at a smaller step size to assure good convergence. The increasing prevalence of parallel architectures in multi-core CPUs and graphics processing units should provide increasing opportunities to speed up this type of unsupervised learning approach.

Algorithm 1 Sparse coding dictionary learning algorithm of [44], modified for HSI.

Set $\gamma = 0.01$
Set $\mu = 10$
Initialize $\{\boldsymbol{\phi}_k\}$ to random positive values
repeat
  for $i = 1$ to 200 **do**
    Choose HSI pixel $\boldsymbol{x}$ uniformly at random
    $\{a_k\} = \arg\min J\left(\{a_k\}, \{\boldsymbol{\phi}_k\}\right)$ s.t. $a_k \geq 0$
    $\Delta\boldsymbol{\phi}_l(i) = a_l \left(\boldsymbol{x} - \sum_{k=1}^{M} \boldsymbol{\phi}_k a_k\right)$
  end for
  $\boldsymbol{\phi}_l \leftarrow \left[\boldsymbol{\phi}_l + \frac{\mu}{200} \sum_i \Delta\boldsymbol{\phi}_l(i)\right]_+$
  $\mu \leftarrow 0.995\mu$
until $\{\boldsymbol{\phi}_k\}$ converges

Finally, we note that the proposed approach can have local minima or non-unique solutions in at least two respects, especially in the case of HSI. First, though the coefficient optimization using an ℓ^1 sparsity penalty is convex, the ideal ℓ^0 sparse solution may not be unique when the one-sided coherence of the dictionary $\max_{i \neq j} |\langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle| / \|\boldsymbol{\phi}_i\|_2^2$ is large [13]. Second, though there are few analytic guarantees about the performance of dictionary learning algorithms, recent results indicate that the ideal dictionary is more likely to be a local solution to the optimization presented here when the coherence of the dictionary is also low [26]. Since many materials have spectral signatures with high correlation in some bands, typical HSI dictionary databases have coherence values very close to unity [33], and we observe similar values in our learned dictionaries. Despite not being favorable for the technical results described above regarding coefficient inference and dictionary learning, the inferred coefficients and learned dictionaries appear to be robust and useful in the applications described here. Indeed, it is likely in these cases that despite there being many local solutions (and a unique minima perhaps even not existing), many of the suboptimal solutions are also quite good and useful in applications. In particular, we have repeated the dictionary learning experiments

described in this paper many times (with different random initial conditions), with no significant changes in the qualitative nature of the dictionary or the performance in the tasks highlighted in Section III. This also corresponds to the results in [33] showing that despite the near-unity coherence in a standard hyperspectral endmember dictionary, these dictionaries can yield good sparse representations useful in spectral unmixing applications.

B. Hyperspectral dataset and learned dictionaries

In this paper we apply the dictionary learning method described in Algorithm 1 to learn a 44-element dictionary for a HSI scene of Smith Island, VA. This scene has 113 usable spectral bands (ranging from 0.44–2.486μm) acquired by the PROBE2 sensor on October 18, 2001.[1] The data has a spatial resolution of approximately 4.5m and was postprocessed to estimate the ground reflectance. Of the 490,000 pixels in the dataset, 2700 pixels are tagged with ground truth labels drawn from 22 categories. These categories include specific plant species and vegetation communities common to wetlands, and were determined by *in situ* observations made with differential GPS aided field studies during October 8–12, 2001. More information about the HSI dataset and the ground truth labels can be found in [4], [6]–[8]. The size of the dictionary (44 elements) was made to ensure that there were multiple elements available for each of the 22 known material classes in this particular dataset. The number 44 represented a compromise between smaller dictionaries that didn't perform as well on the tasks described in Section III (especially the local manifold approximation), and larger dictionaries that presented more difficulty getting all of the elements to converge in the learning.[2] In general, determining the optimal number of dictionary elements to learn for a dataset is an open question and could be a valuable future research direction.

We cross-validated the results of this paper in two ways. First, 10,000 randomly selected pixels were excluded from the dataset before the dictionary learning so that they could be used in testing. Second, we also have available data from another HSI collection of the same geographic region using the same sensor on August 22, 2001. While this is close enough in time to assume that there are no major geologic changes in the scene, this data does come from a different season where the vegetation and atmospheric characteristics are potentially different, resulting in different statistics from the data used in the learning process. We use this dataset specifically to assess the potential negative effects of mismatch between the statistics of the training and testing datasets when performing signal processing applications using the learned dictionary.

[1]Smith Island is a barrier island that is part of the Virginia Coast Reserve Long-Term Ecological Research Project. For more details, see http://www.vcrlter.virginia.edu. This dataset was generously provided by Charles Bachmann at the Naval Research Laboratory.

[2]While performance in the signal processing tasks we tested did improve with larger dictionaries, we note that the performance difference was often relatively minor when using 22 element dictionaries and this size would likely be sufficiently for this dataset in many applications.

C. Related work

Prior work in using unsupervised methods to learn HSI material spectra has used some algorithms that are very related to our present approach. For example, ICA can be viewed as finding linear filters that give high sparsity, and prior work [23], [27], [51] demonstrates that ICA can be effective at determining a range of spectral signatures from preprocessed data. Other approaches also based on Bayesian inference (but not necessarily a sparsity-inducing prior) [41] have been used to learn HSI dictionaries, but this approach has trouble including information from large datasets and often uses ICA as a preprocessing stage to reduce the number of pixels to analyze. The technique most closely related to our current approach is blind source separation based on non-negative matrix factorization (NMF) [35], [47]. While not explicitly incorporating sparsity constraints, results using NMF have been shown to exhibit sparse behavior [31]. In the NMF setup, the sparsity level of the decomposition is difficult to control [31] and previous work in [35] mitigates this by adding an explicit sparsity inducing term. Additionally, the above mentioned approaches all retain the sum-to-one constraint, which we drop due to the variable power in the pixels throughout the scene.

In addition to these results on unsupervised learning, as well as additional encouraging prior work on using sparsity models for spectral unmixing [29], [33] and learning dictionaries that resemble material spectra [28], Castrodad et al. [16] have explored using a sparsity model and learned dictionaries to improve supervised classification performance on HSI data.[3] In Section III-C we will explore the advantages of using sparse coefficients from a learned dictionary in an off-the-shelf classification algorithm. In [16], the authors use labeled data to learn a separate dictionary for each class and classify data by determining which of these candidate dictionaries best describes an unknown pixel (defined by having the minimum value for the objective function in equation (2)). This approach is customized to the classification problem, and we expect the classification performance would outperform the general approach we describe in Section III-C. In contrast, the approach in [16] requires a more computationally expensive learning process (due to the multiple dictionaries), requires labeled data before the learning process, and generates a dictionary that is tailored to the classification task and may not generalize as well to other tasks.

Zhou et al. [55] have explored using a sparsity model and learned dictionaries to effectively solve inverse problems in HSI. In Section III-B we will explore the ability of sparse coefficients from a learned dictionary to infer high resolution spectral data from low resolution imagery by formulating the task as a linear inverse problem. In [55], the authors show that when removing substantial amounts of data from an HSI datacube, a learned dictionary can be used to exploit the correlation structure present in each band to infer the missing data and reconstruct the spatial image associated with each band. This inpainting task is a very similar inverse problem to the one we

[3]The authors in [16] use a different learning algorithm (K-SVD [2]) from our gradient approach, but it is attempting to achieve the same goal of learning an optimal dictionary for sparse approximation.

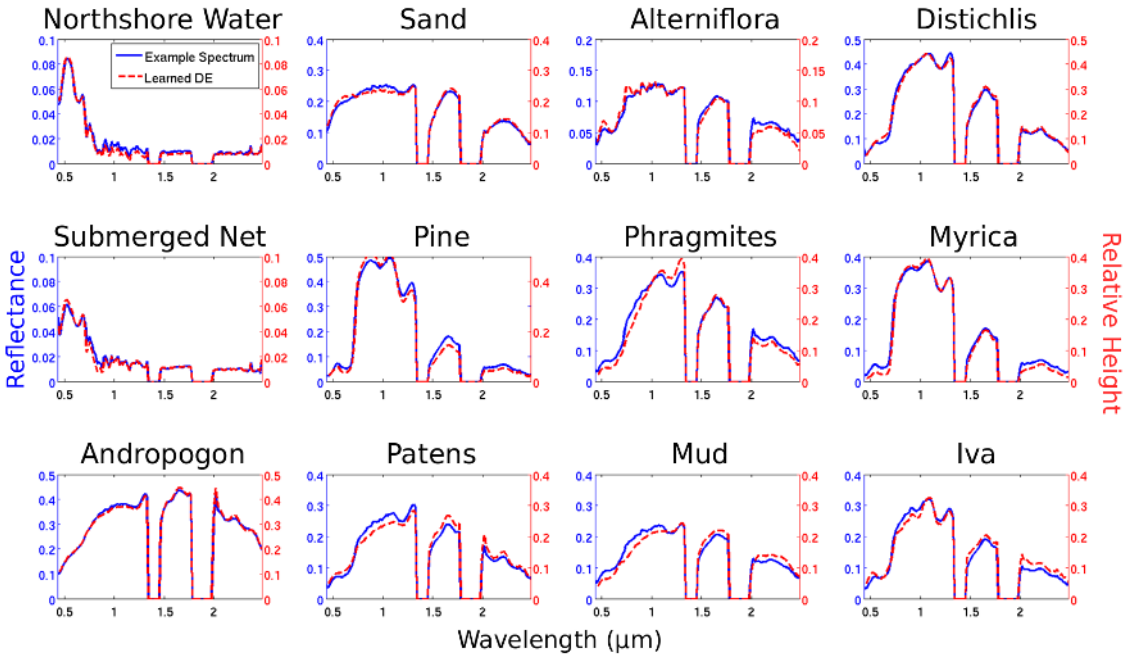

Fig. 2. Example spectra for materials in the labeled classes of the Smith Island dataset and the learned dictionary element (DE) that is the closest match for each example. The two obvious gaps in the spectra are bands removed from consideration in the original dataset due to the interactions with the atmosphere in these regions.

examine in Section III-B, differing primarily in the type of measurement operator used in the model (i.e., blurring vs. subsampling) and the dimension of the data used in the learning and reconstruction (i.e., spectral vs. spatial).

III. Main Results

A. *Learned Dictionary Functions*

While the learning procedure described in Algorithm 1 adapts the dictionary to the high-order statistics of the HSI data, there are no constraints added that ensure the resulting dictionary elements will correspond to physical spectra or be informative about material properties in the scene. To examine the properties of the learned dictionary, examples elements are plotted in Figure 2. It is clear that these dictionary elements not only have the general appearance of spectral reflectances, they also match the spectral signatures of many of the materials that are known to be in the scene. Using the ground truth labels from the Smith Island dataset (which denote the dominant material present in the pixel), Figure 2 shows an example spectral signature from a class along with the dictionary element that has the largest coefficient in the sparse decomposition of that pixel. Despite being given no *a priori* information about the data beyond the sparsity model (i.e., without being given the class labels and corresponding pixels), the algorithm learns spectral shapes that correspond to a number of component material spectra present in the image. These learned dictionaries cover a wide variety of distinct material classes for which we have ground truth labels,

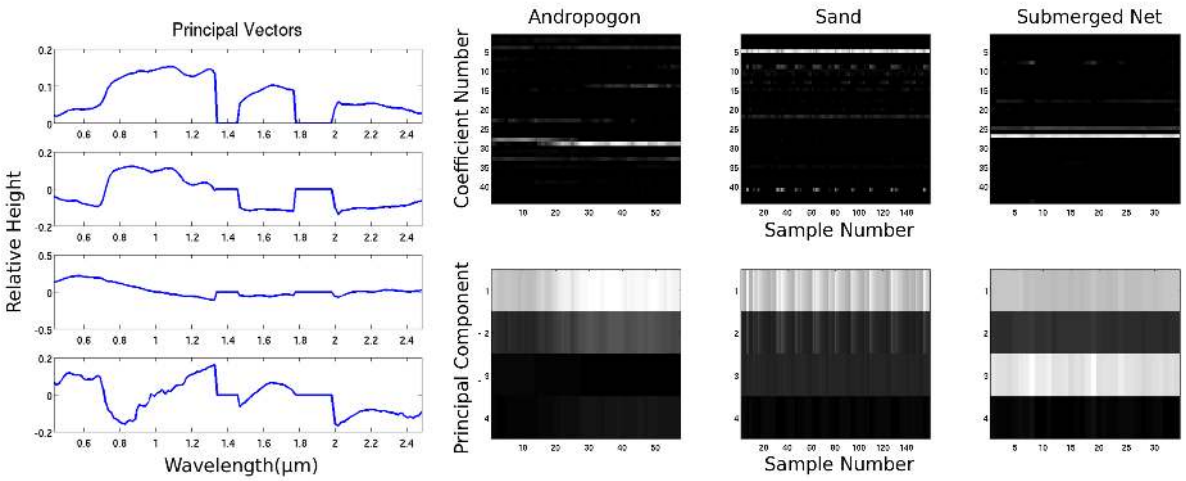

Fig. 3. (Left) The top four principal components for the Smith Island dataset (capturing 99.9% of the variance). In contrast to the learned dictionary elements in Figure 2, only one of the principle components looks generally like a spectral signature. (Right) PCA and sparse coding coefficients representing every sample of the data from three of the labeled classes ("Andropogon", "Sand", and "Submerged Net"). The brightness at each pixel represents the intensity of a given coefficient for a specific pixel. Note that PCA uses many of the same coefficients for different materials (e.g., coefficient 1 is always used), while sparse coding tends to select distinct coefficients for the different materials.

including "Pine", "Water", "Mud" and "Distichlis", as well as very similar spectra, such as "Water" and "Submerged Net" or "Pine trees" and "Iva".

In contrast, Figure 3 shows the first four principal components found through PCA analysis on the same HSI dataset, which is sufficient to capture 99.9% of the variance in the data. While the first principal component does have some similarity to a general vegetation spectrum, the other spectral components do not correspond to physically meaningful spectral features. Figure 3 also shows the comparison between the decomposition coefficient in the sparsity model and PCA for all pixels in four of the labeled classes. The raster plots show that while the sparse decomposition and the principal components both only need a few coefficients to represent the data, the sparse decomposition chooses different coefficients for different spectral shapes (i.e., the material information is encoded in the selection of active coefficients) whereas PCA uses the same four vectors to represent nearly all of the data. This comparison illustrates that the learned dictionary under the sparsity model has a much closer correspondence to the individual spectral characteristics found in the dataset than PCA, indicating that this representation may have many advantages for tasks such as classification.

While it is clear that the dictionary elements are learning spectral elements present in the scene, this representation will be most meaningful if there is consistency in the way environmental features are represented. In other words, when looking across the scene, do the sparse decompositions change in a way that reflects the changes in the underlying geologic features? We extracted a row of pixels from the Smith Island dataset, starting inland and ending in the water off the coast of the island. The selected row of pixels is shown in red in Figure 4, superimposed on

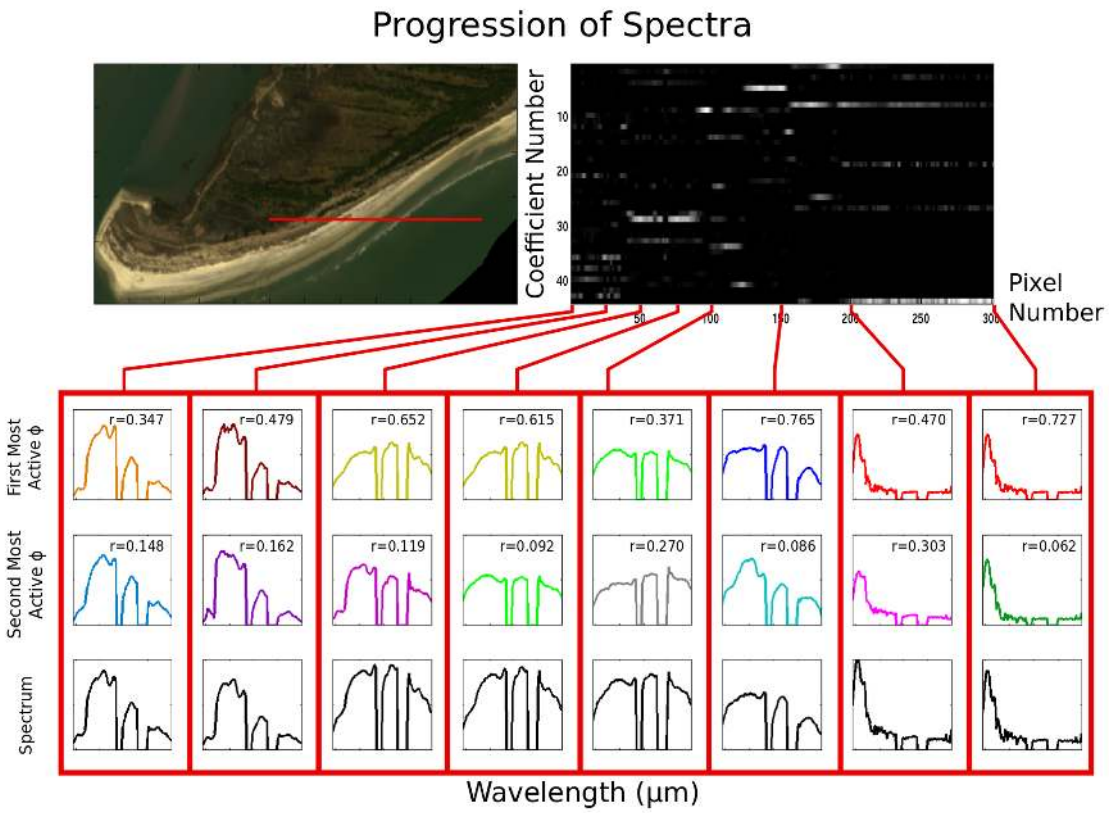

Fig. 4. Progression of sparse coding coefficients from a row of contiguous pixels in the Smith Island dataset. (Upper left) The red line indicates a row of 300 pixels selected for analysis. These pixels (numbered left to right) represent a progression from an inland region to the water off the east coast of the island. (Upper right) The sparse coding coefficients for the row of pixels is shown, where the brightness of a pixel indicates the intensity of each coefficient for each pixel. Note that many of the same coefficients are often active in the same geographic regions, and the progression from one type of element to another (e.g., sand to water) can be seen by different coefficients dominating the decomposition. (Bottom) The spectra for pixels 1, 25, 50, 75, 100, 150, 200 and 300 are shown in the bottom row (in black), along with the two most active dictionary elements in the top two rows (color coded). The fractional abundance for each dictionary element in each pixel is given by $r = |a_i|/\|\mathbf{a}\|_1$. Note that many of the same dictionary elements can be seen dominating the decomposition in regions with similar material composition.

a magnified RGB rendering of that portion of the island. Figure 4 shows the coefficient decomposition of each pixel, as well as the measured spectrum and the two most active dictionary elements at various locations along the row. Included with each of the two most active dictionary elements is the fractional abundance $r = |a_i|/\|\mathbf{a}\|_1$ of that dictionary element in the decomposition. This row starts with mostly vegetation spectra for the first 75 pixels, changing to sand-like spectra by the the 100^{th} pixel and eventually to water spectra by the 160^{th} pixel.

We highlight two important properties of the coefficient decompositions over the pixel progression in the raster plot in Figure 4. First, the sparse coefficients are relatively consistent over contiguous spatial ranges, with the same small sets of coefficients generally dominating the decomposition over small contiguous regions. While this is evident in the regions dominated by sand and water, there are also repeated dictionary elements across several spatial locations in the regions dominated by vegetation (which we would expect to have much more variability over pixels with 4m resolution). Second, some slowly changing geologic properties are actually observable in the

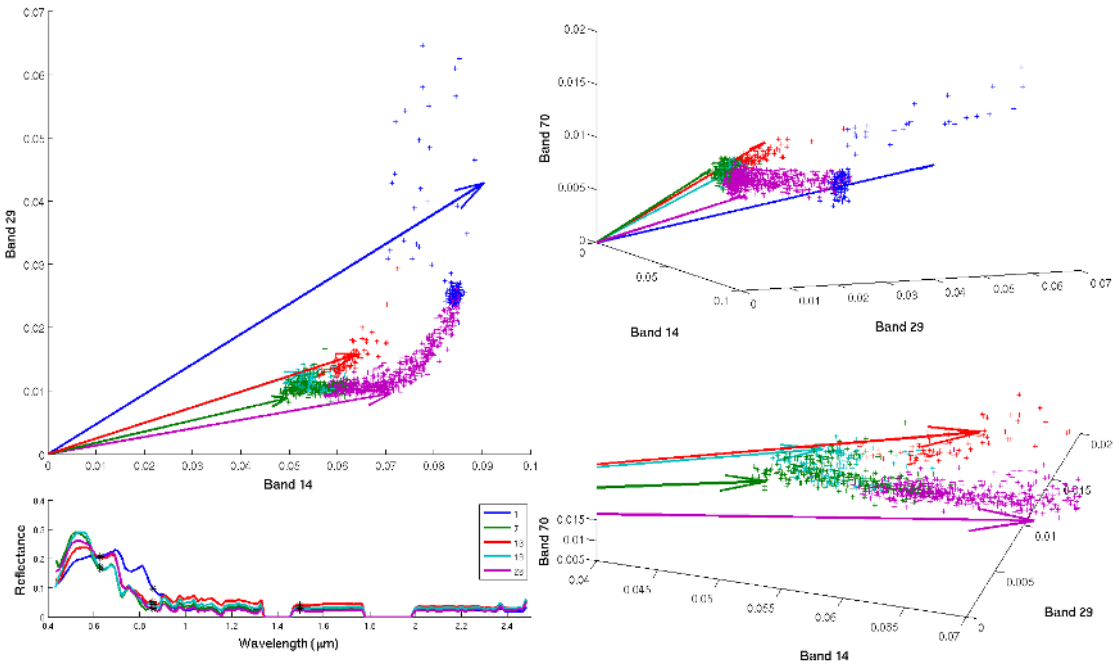

Fig. 5. The nonlinear structure of water pixels is locally approximated by the learned dictionary. The plots in the upper left, upper right and bottom right all show the spectra water pixels (selected from a contiguous region) projected onto three spectral bands (14,29,70). Even in three dimensions, it is clear that the data live on a nonlinear manifold, and there is clear structure in the variability. The vectors represent the projection onto the same three bands of five learned dictionary elements. The points representing water pixels are color coded to indicate which dictionary element has the largest value when inferring the sparse coefficients, showing that contiguous values on the manifold are coded using the same dictionary element.

gradual onset and offset of specific dictionary elements in the decomposition. One prominent example of this is the slow change from dictionary element 8 to dictionary element 44 over the span of water moving away from the shoreline, indicating the slow fading of shallow water to deep water (which have different spectral characteristics and are represented by different dictionary elements). Another example of this is the rise of dictionary element 9 from the second most active to the most active element from pixels 75 and 100, indicating the slowly increasing presence of a particular vegetation characteristic in this region.

In addition to the spectral matches shown in Figure 2 and the spatial coefficient variations shown in Figure 4, another important aspect of the learned dictionary is to examine how it represents the nonlinear variations within a particular material class [5]. For example, Figure 5 shows full spectral signatures for a patch of water off the coast of Smith Island, as well as spectral bands 14, 29 and 70 (0.6278, 0.8572 and 1.4962 μm) from three different view angles to show the geometry of these points in 3-D spectral space.[4] Despite being one material class ("water"), it is evident even in these few bands that the measured spectrum lies on a nonlinear manifold. Superimposed on the 3-D spectral plots are five of the learned dictionary elements projected onto these same three spectral bands. The measured spectra are color coded to indicate which of these five learned dictionary elements are dominant

[4]These are the same spectral bands and approximately the same region highlighted in Figure 1 of [5].

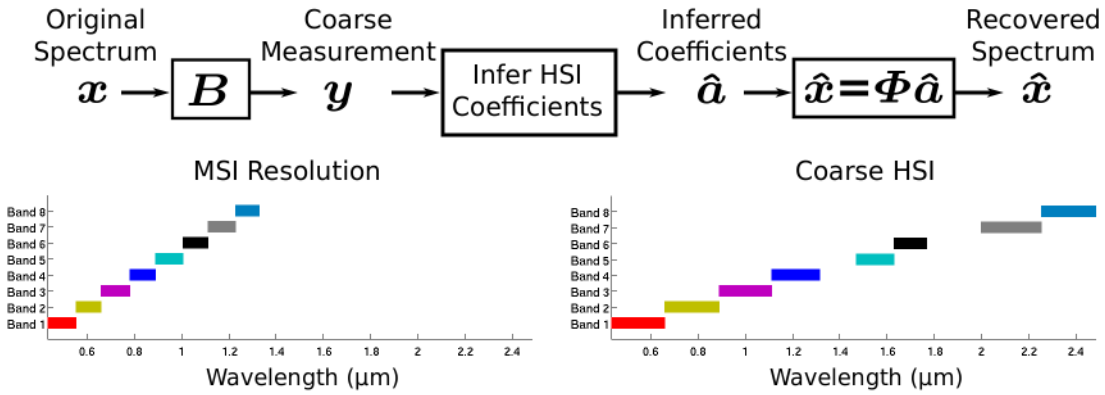

Fig. 6. Reconstructing spectra with HSI resolution from measurements with MSI-level resolution. (Top) A schematic of the process for simulating low resolution spectral data and performing recovery. The matrix $\boldsymbol{B}$ characterizes the measurement process (i.e., the sensitivity function of the sensor), simulating the aggregation of high resolution spectral information into low resolution spectral bands. (Bottom left) A diagram indicating the sensitivity function for MSI resolution measurements, where 113 HSI bands are collapsed into 8 equally spaced measurement bands over the lowest wavelengths (approximately matching the spectral bands reported by the Worldview II MSI sensor). Note that no information is measured from the highest wavelength region. (Bottom right) A diagram indicating the sensitivity function for coarse HSI measurements, where 113 HSI bands are collapsed into 8 nearly equally spaced measurement bands across the whole HSI spectrum.

in their sparse decomposition. The contiguity of this color coding over small manifold regions demonstrates that rather than containing the measured spectra in a convex hull, the learned dictionaries are essentially forming a local linear approximation to this manifold. So, despite being a linear data model, the dictionary learns multiple elements that capture the nonlinear spectral variations by locally approximating the manifold structure in a meaningful way. In our experiments with other endmember extraction algorithms such as [53], the learned sparse dictionary does appear to produce a representation that more closely tracks the nonlinear variations in the data points (e.g., produces a smaller relative MSE between the data and the dictionary elements) compared to a method restricted to finding a convex cone around the data. A more detailed characterization of the differences between various linear models at representing nonlinear material variations would be a valuable direction for future research.

B. Reconstructing HSI-resolution from MSI-resolution data

As discussed earlier, while the high spectral resolution of HSI is valuable, acquiring data at this resolution comes at a cost. In terrestrial remote sensing, hyperspectral imagers are relatively rare instruments, and it would be much more resource efficient to perform most spectral imaging at MSI-level resolution. Data at this resolution could either be gathered by actual MSI sensors, or by HSI sensors modified to decrease their spectral resolution (which could potentially decrease scan times). The question we consider here is whether a dictionary learned on an HSI training set could be used to accurately infer high resolution spectra from subsequent data collected at MSI-level spectral resolution.

In this basic paradigm, assume that we start with a learned dictionary that has been adapted to the specific structure of the desired HSI data. This could arise from earlier HSI of the scene being imaged, or imaging from

other geographic regions with similar environmental features (and therefore similar statistics). For the new data acquired at MSI-level resolution, we assume for a first approximation that each band is a linear combination of some group of spectral bands in the underlying true HSI data. Specifically, we model the MSI-resolution data as

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x} + \epsilon = \boldsymbol{B}\sum_{k=1}^{M} \boldsymbol{\phi}_k a_k + \epsilon, \tag{4}$$

where $\boldsymbol{y} \in \mathbb{R}^L$ $(L < N)$ is the new coarse resolution data and $\boldsymbol{B}$ can be thought of as an $(L \times N)$ "blurring" matrix that bins the spectral bands of the desired HSI data. While $\boldsymbol{B}$ could be any matrix describing the sensitivity function of the imager acquiring the MSI-resolution data, we will consider $\boldsymbol{B}$ that simply adds spectral bands over a contiguous range.

This measurement paradigm fits nicely into the well-known framework of Bayesian inference (or equivalently, linear inverse problems in image processing). Essentially, given the wealth of information about the statistics of the HSI we would like to obtain, Bayesian inference allows one to optimally answer the question of what underlying HSI data $\boldsymbol{x}$ is most likely given the observed MSI-resolution data $\boldsymbol{y}$. Specifically, given the new model in (4), the likelihood of the data $\boldsymbol{y}$ given the coefficients $\{a_k\}$ is now the Gaussian distribution

$$p(\boldsymbol{y}|\{a_k\}) \propto e^{-\frac{1}{2\sigma_\epsilon^2}\|\boldsymbol{y} - B\sum_k \boldsymbol{\phi}_k a_k\|_2^2}.$$

We can again use an independent Laplacian prior on the sparse coefficients $\{a_k\}$, and write the posterior distribution using exactly the same simplifications as before. The optimal MAP estimate of the sparse coefficients given the observed data $\boldsymbol{y}$ is therefore given by optimizing the following objective function with respect to the coefficients:

$$\widetilde{J}_\gamma\left(\boldsymbol{y}, \{a_k\}, \{\boldsymbol{\phi}_k\}\right) = \left\| \boldsymbol{y} - \boldsymbol{B}\sum_{k=1}^{M} \boldsymbol{\phi}_k a_k \right\|_2^2 + \gamma \sum_k |a_k|. \tag{5}$$

This optimization program is very similar to (2) (and can be solved by the same software packages), but incorporates the measurement process described by $\boldsymbol{B}$ into the inference. Given the estimated sparse coefficients, the HSI vector $\boldsymbol{x}$ is reconstructed according to (1): $\hat{\boldsymbol{x}} = \sum_k \boldsymbol{\phi}_k \hat{a}_k$. The full workflow is shown schematically in Figure 6. We note that many linear inverse problems are formulated in a similar way depending on the choice of $\boldsymbol{B}$, including inpainting missing data such as the application considered by [55].

For proof-of-concept simulations we generated simulated data with MSI-level resolution from pixels that were not used in the training dataset, and perform the inference process described above to estimate the high-resolution spectra from the low-resolution measurements. In the first set of simulations, the matrix $\boldsymbol{B}$ (illustrated in Figure 6) generates simulated data with 8 equally spaced bands covering the entire spectral range of the HSI data. This $\boldsymbol{B}$ is

TABLE I
RELATIVE RECOVERY ERROR FOR HSI SPECTRA FROM COARSE HSI MEASUREMENTS (FULL SPECTRUM). RESULTS ARE REPORTED FOR TESTING DATA COLLECTED ON THE SAME DAY (SD) AS THE TRAINING DATA USED TO LEARN THE DICTIONARY, AS WELL AS RESULTS FOR TESTING DATA COLLECTED ON A DIFFERENT DAY (DD).

	Mean Error (SD)	Median Error (SD)	Mean Error (DD)	Median Error (DD)
44 Learned DE	$8.249\text{x}10^{-4}$	$4.911\text{x}10^{-4}$	$7.054\text{x}10^{-3}$	$6.005\text{x}10^{-3}$
44 Exemplar DE	$6.280\text{x}10^{-3}$	$2.709\text{x}10^{-3}$	$1.493\text{x}10^{-2}$	$1.105\text{x}10^{-2}$
44 Random DE	$4.143\text{x}10^{-1}$	$4.524\text{x}10^{-1}$	$3.965\text{x}10^{-1}$	$4.165\text{x}10^{-1}$

TABLE II
RELATIVE RECOVERY ERROR FOR HSI SPECTRA FROM MSI MEASUREMENTS (NO MEASUREMENTS FROM HIGHEST WAVELENGTHS). RESULTS ARE REPORTED FOR TESTING DATA COLLECTED ON THE SAME DAY (SD) AS THE TRAINING DATA USED TO LEARN THE DICTIONARY, AS WELL AS RESULTS FOR TESTING DATA COLLECTED ON A DIFFERENT DAY (DD).

	Mean Error (SD)	Median Error (SD)	Mean Error (DD)	Median Error (DD)
44 Learned DE	$1.271\text{x}10^{-2}$	$1.791\text{x}10^{-3}$	$2.456\text{x}10^{-2}$	$1.219\text{x}10^{-2}$
44 Exemplar DE	$1.132\text{x}10^{-2}$	$5.552\text{x}10^{-3}$	$2.225\text{x}10^{-2}$	$2.135\text{x}10^{-2}$
44 Random DE	$7.845\text{x}10^{-1}$	$8.974\text{x}10^{-1}$	$7.775\text{x}10^{-1}$	$9.946\text{x}10^{-1}$

intended to model a hyperspectral imager collecting spectral data with an order of magnitude less spectral resolution than the original data. We used two testing datasets in this simulation: the 10,000 pixels from the October 2001 scan of Smith Island that were withheld from the learning process, and 10,000 randomly selected pixels from the August 2001 scan of the same geographic region. By using HSI collected on a different date we can examine the effects of using a dictionary that was learned on data with different statistics than the data we are trying to reconstruct (due to different vegetation characteristics in the different seasons and different atmospheric conditions present on the different days).

We infer the sparse coefficients in the HSI dictionary given the simulated MSI-resolution data by minimizing the objective function in Equation (5) as described above. For comparison purposes and to determine the value of the learning process in the reconstruction, we repeated this recovery process with a 44-element dictionary of random values (i.e., the initialization conditions for the dictionary learning) and with an exemplar dictionary formed by taking two random spectral signatures from each class in the original labeled HSI data (for a total of 44 dictionary elements). Figures 7 and 8 show examples of the original HSI, the simulated coarse resolution data, the estimated sparsity coefficients in the learned dictionary, and the subsequent recovered HSI data for the test datasets collected on the same date (SD) and a different date (DD) as the training dataset. The set of examples shown in the figure span the range of the most favorable and least favorable reconstructions.

Table I reports the average relative MSE for the reconstructions, calculated as

$$e_i = \frac{\|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i\|_2^2}{\|\boldsymbol{x}_i\|_2^2}. \tag{6}$$

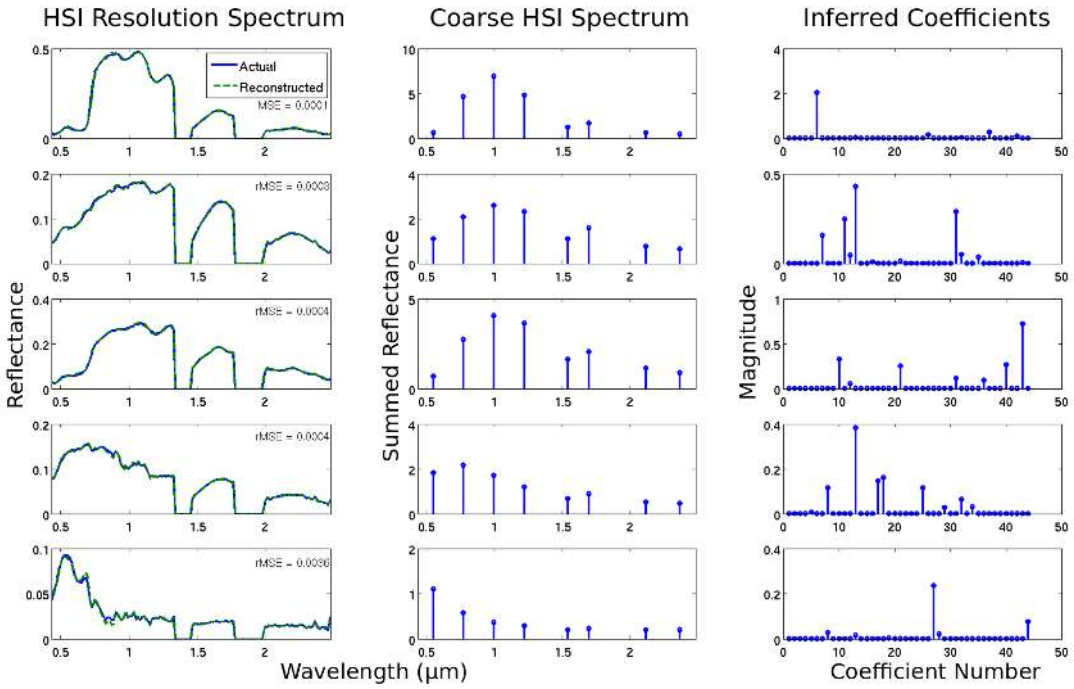

Fig. 7. Reconstructing HSI data from simulated coarse HSI measurements using training and testing data collected on the same date. Plots show original HSI spectrum in blue (113 bands), simulated coarse HSI spectrum (8 bands), inferred sparse coefficients, and reconstructed HSI spectrum in green. Examples were selected to illustrate a range of recovery performance, from examples of the best recovery on top to examples of the worst recovery on the bottom.

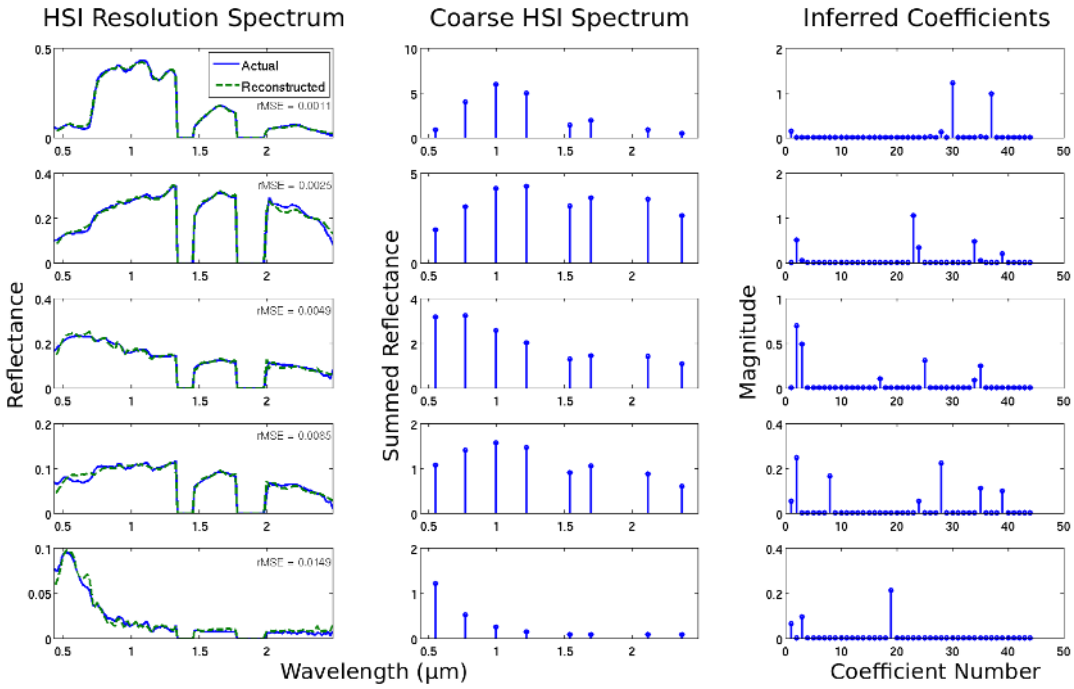

Fig. 8. Reconstructing HSI data from simulated coarse HSI measurements using training and testing data collected on different dates (in different seasons). Plots show original HSI spectrum in blue (113 bands), simulated coarse HSI spectrum (8 bands), inferred sparse coefficients, and reconstructed HSI spectrum in green. Examples were selected to illustrate a range of recovery performance, from examples of the best recovery on top to examples of the worst recovery on the bottom.

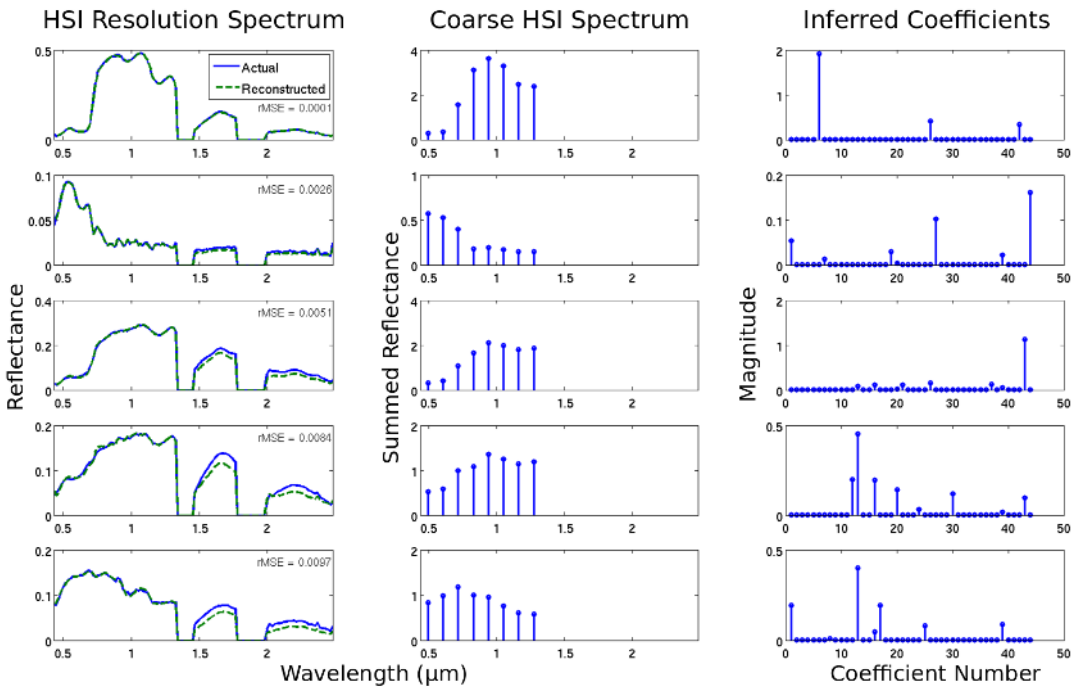

Fig. 9. Reconstructing HSI data from simulated MSI measurements using training and testing data collected on the same date. Plots show original HSI spectrum in blue (113 bands), simulated coarse HSI spectrum (8 bands), inferred sparse coefficients, and reconstructed HSI spectrum in green. Examples were selected to illustrate a range of recovery performance, from examples of the best recovery on top to examples of the worst recovery on the bottom.

The aggregate results as well as the specific plotted examples demonstrate that the HSI-resolution data is recovered with less than 0.09% relative MSE for the SD testing set and less than 0.71% relative MSE on the DD testing set. While the reconstruction is worse on the DD dataset because of the mismatch in the training and testing statistics, the reconstructions are still very good overall and often capture even fine detail in the HSI spectra. Also note that the learned dictionary is performing significantly better than both the exemplar dictionary (which was chosen using oracle knowledge of the classes to ensure good coverage of the various materials) and the random dictionary (indicating the value of the learning process).

In the second set of simulations, the matrix $\boldsymbol{B}$ (illustrated in Figure 6) generates simulated data with 8 equally spaced bands excluding the highest wavelength regions. This $\boldsymbol{B}$ is intended to model a multispectral imager, and we selected the bands to approximately match the reported bands of the WorldView II multispectral sensor. We used the same SD and DD testing datasets in the simulation, with Figures 9 and 10 showing example reconstructions and Table II reporting average reconstruction results. While the overall performance does suffer compared to the previous experiment when the whole spectral range was measured, the HSI spectra are again recovered with low error overall: less than 1.28% for the SD dataset, and less than 2.47% error for the DD dataset. As expected from the previous simulation, the lower wavelengths can be reconstructed very well. As might be expected because no data was collected in the higher wavelength range, the recovery in these spectral bands can suffer from higher

errors even despite getting the general shape correct. Table II also shows that overall, both the learned and exemplar dictionaries have approximately the same mean relative error in this setting. However, the distribution of the relative errors over the test pixels is more tightly peaked about the origin for the learned dictionary, with a median relative error approximately a third of that for the exemplar dictionary. This indicates that while most test pixels were recovered better with the learned dictionary, there were a minority of pixels that suffered more egregious errors than seen with the exemplar dictionary.

Though the results of the high-resolution reconstructions given above are very encouraging, as with any engineering application it is important to characterize what causes variations in the performance. Figure 11 shows a more detailed analysis of the errors for the worst performing case in the above simulations: using simulated MSI data with a dictionary that was learned on data taken on a different date from the test data. This analysis quantifies the observation that the better the model is at fitting the data, the better we expect the resulting algorithm to perform. Specifically, we group the pixels in the test dataset into three groups based on the (normalized) sparsity of their resulting inferred coefficients (i.e., how well the data point is fit by a sparsity model) measured by $\|\boldsymbol{a}\|_1/\|\boldsymbol{a}\|_2$. The clear trend is that the performance in this task is strongly dependent on how amenable that pixel is to admitting a sparse decomposition. Fortunately, only a small fraction of the data (less than 9%) falls into the worst performing category. Currently we have not found any quantitative correlations between material classes and model fit, but anecdotally we observe that classes such as pine trees and water appear prevalent among the pixels with the lowest rMSE in the reconstruction task, and classes such as mixed vegetation and mud are more prevalent in the outliers that have higher rMSE. Of course, an interesting topic of future study would be to understand more precisely how to modify the model to improve the fit with the current outliers (and subsequently the performance on the current task).

We note that there are many other linear inverse problems that may be of interest, including other methods for reducing data acquisition resources. For example, in the field of compressed sensing [40], a sparsity model is also assumed and data is measured by using a coded aperture that forms each measurement by taking a (generally random) linear combination of the input data. In this case, the original data is recovered by solving the same optimization problem as in (5). Indeed, similar acquisition strategies have already been implemented in novel HSI sensors [25], [52]. Looking carefully, the only difference between the compressed sensing strategy and the approach presented above is in the choice of $\boldsymbol{B}$. The "blurring" choice of $\boldsymbol{B}$ in our experiments should actually result in a more difficult reconstruction problem than when $\boldsymbol{B}$ is chosen to be a random matrix because the introduction of randomness will tend to improve the conditioning of the acquisition operator. We have performed similar simulations to the ones above (not shown) using $\boldsymbol{B}$ drawn randomly and independently from a Bernoulli distribution, and the

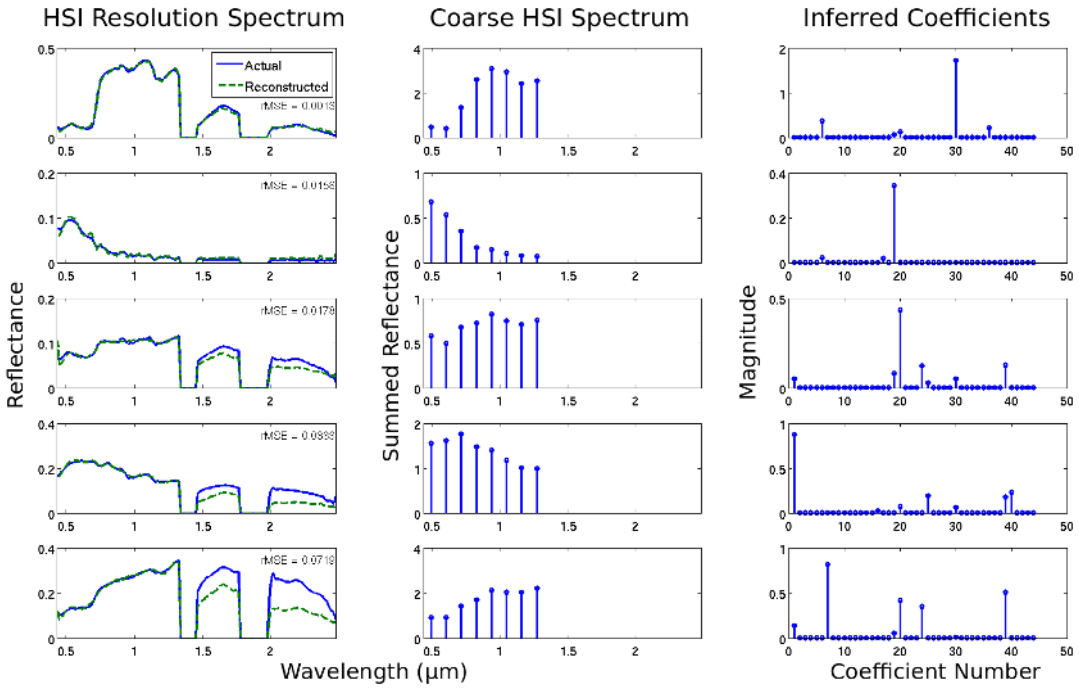

Fig. 10. Reconstructing HSI data from simulated MSI measurements using training and testing data collected on different dates (in different seasons). Plots show original HSI spectrum in blue (113 bands), simulated coarse HSI spectrum (8 bands), inferred sparse coefficients, and reconstructed HSI spectrum in green. Examples were selected to illustrate a range of recovery performance, from examples of the best recovery on top to examples of the worst recovery on the bottom.

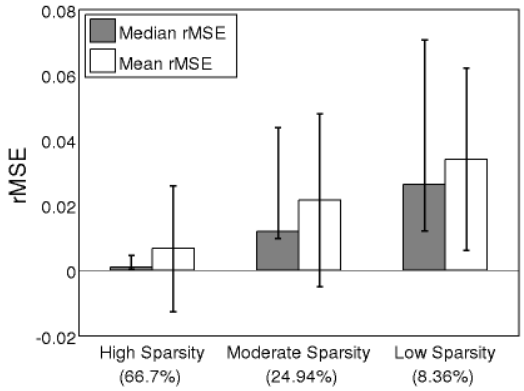

Fig. 11. The reconstruction errors when inferring HSI-resolution data from simulated MSI measurements are closely related to the normalized sparsity $\|\boldsymbol{a}\|_1/\|\boldsymbol{a}\|_2$ of the coefficients. The mean and median errors are shown for 3 categories measuring the sparsity model fit: High Sparsity represents an excellent model fit (normalized sparsity is between 1 and 1.92), Moderate Sparsity represents a good model fit (normalized sparsity is between 1.92 and 2.3), and Low Sparsity represents only a fair model fit (normalized sparsity is above 2.3). The data shown is for the reconstructed pixels in the worst performing scenario in our simulations (test pixels from the August 2001 dataset and dictionaries learned from the October 2001 dataset), and the percentage of pixels falling into each category are displayed below the category labels. The error bars of the mean rMSE represent the standard deviation and the error bars on the median rMSE represent the 25^{th} and 75^{th} percentiles. The differences between these two indicates that the reconstruction errors are tightly packed for the data points with low normalized sparsity with a few outliers, and spread out for points with higher normalized sparsity.

results indicate that recovery with similar accuracy is also possible when using this learned dictionary.

C. Supervised classification

Clearly one of the most important HSI applications is classifying the dominant materials present in a pixel [4], [7], [39]. Because sparse coding is a highly nonlinear operation that appears to capture different spectral features by using different dictionary elements (and not just changing the coefficient values on those elements), we suspect that performing classification on the sparse coefficients can improve HSI classification performance compared to classification on the raw data (or other dimensionality reduced representations such as PCA). Intuition for this approach comes from the well-known idea in machine learning that expanding a data representation with a highly nonlinear kernel can serve to separate the data classes and make classification easier (especially with a simple linear classifier). Indeed, several researchers have reported that sparse coding in highly overcomplete learned dictionaries (which is a highly nonlinear mapping) does improve classification performance [34], [49].

To gain further intuition, consider a very simple classifier based on finding the maximum sparse coefficient for each pixel in the scene. This sparse decomposition with one coefficient can be thought of as a type of vector quantization (VQ) [42], and the coefficient index can be used as a rough determination of the class of the pixel. Figure 12 shows a segment of the Smith Island dataset, where each pixel is independently unmixed and colored according to the index of the maximum sparse coefficient representing that pixel.[5] Relevant environmental features such as tree lines and sandbanks are clearly distinct, indicating a correlation between the most active dictionary element and the material in the image. Additionally, variations within a class can be captured by different coefficients. For example, different water characteristics are clearly visible, including depth changes due to sandbars (the orange stripes in the left side of the image) and areas with submerged nets (the red stripes offshore by the sandbanks).

While the simple demonstration in Figure 12 is an encouraging illustration, this approach clearly going to underperform compared to a classification scheme that includes information from all of the coefficients simultaneously. To demonstrate the utility of sparse coefficient representations using learned dictionaries for classification, we performed several classification tests on the Smith Island dataset using Support Vector Machines (SVMs) and verifying the results with ground truth labels. SVMs [14] are a widely used supervised learning technique capable of performing multi-class classification. Specifically, we use the C-SVM algorithm (implemented in the freely available `libsvm` package [17]) with a linear kernel.

There are two potential factors to consider when performing supervised classification: overall performance (i.e., classification error) and classifier complexity. While classification error on a test dataset is an obvious performance metric of interest, classifier complexity is also an important aspect to consider. For a fixed performance rate, less

[5]The colors in Figure 12 are assigned to give as much visual distinction as possible between elements that are physically adjacent, but have no other meaning.

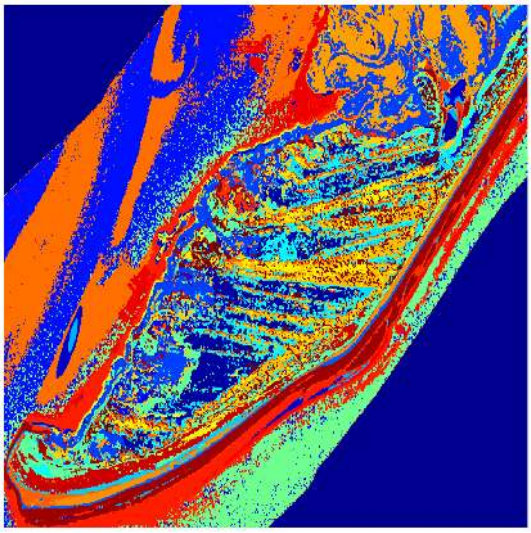

Fig. 12. Vector Quantization classification of the scene. The color in each pixel indicates which dictionary element had the largest coefficient value in the sparse code for that pixel. Distinct shapes consistent with known material structures from the ground truth data (e.g., sand bars and tree lines) can be easily seen.

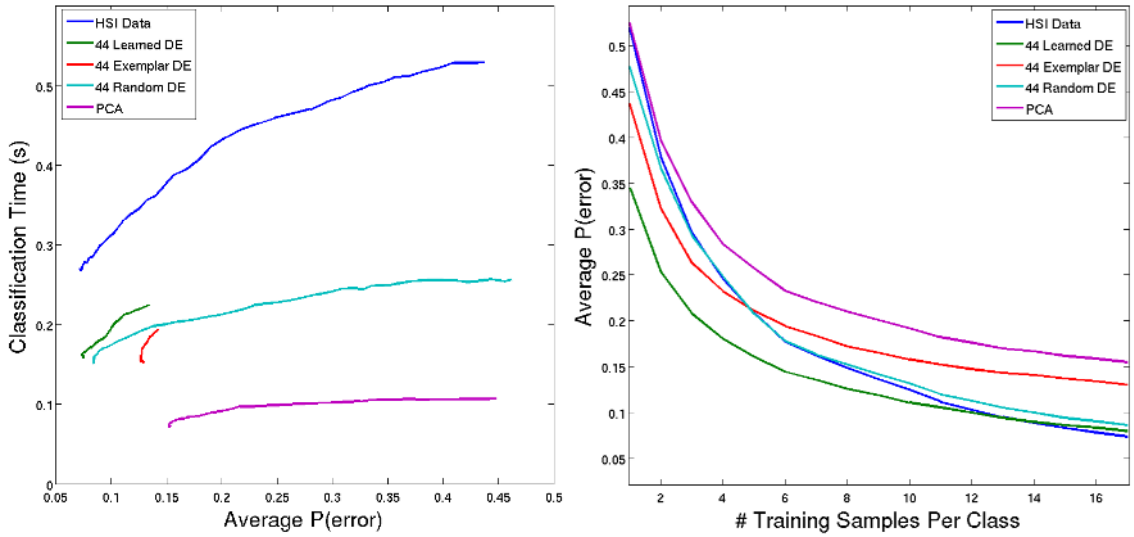

Fig. 13. Classification on 22 material classes in the Smith Island dataset. (Left) Average classification error plotted as a function of average classification time (as a proxy for classifier complexity) as the complexity parameter of the SVM is varied. Using coefficients from a sparse code in a learned dictionary as input to the SVM performs essentially as well as using the raw data, but with a classifier 30% less complex. (Right) Average classification error as a function of the training dataset size for each class. The power of the lower complexity classifier is demonstrated in the ability to generalize better, with sparse coefficients in the learned dictionary clearly showing better performance for the very small training sets.

complex classifiers take less computation time (which is important in large datasets), and are typically less prone to over-fitting during the training (which may lead to better generalization beyond the training data). With linear kernels, the only parameter of the C-SVM algorithm is the cost variable C which controls the complexity of the classifier by changing the cost of the wrongly classified points in the training process. We sweep C over a range

of values from 1 to 10000 and observe the probability of error and classification time using the raw HSI data, reduced dimensionality data using PCA, and sparse coefficient representations for the learned, exemplar and random dictionaries discussed earlier. For each value of C, we performed 20 trials where each trial consists of selecting a subset of 17 pixels from the labeled data for each of the 22 classes to train a new SVM classifier and then testing the classification performance on the remaining labeled data withheld from the SVM training.[6] We average over all trials and all 22 classes to find the average classification error and average classification time (as a proxy for classifier complexity). Figure 13 shows the changes in classification time and probability of classification error.

There are three interesting things to note about the results in Figure 13. First, while the raw data achieved the lowest overall error for the range of C tested (P(error)=0.0721), the sparse coefficients in the learned dictionary are nearly as good (P(error)=0.0736) using a much simpler classifier that operates $\sim 30\%$ faster than the SVM on the raw data.[7] Second, while PCA reduces the classification time farther then the other approaches due to its extremely low dimensionality (4 principle components), it performs significantly worse than the raw data or the sparse coefficients. Third, using sparse coefficients in the random dictionary surprisingly performs better (P(error)=0.0838) than sparse coefficients in the exemplar dictionary (P(error)=0.1262), despite having no apparent relevance to material spectra in the scene. While this is counter-intuitive, other recent results have shown that projection onto random dictionaries can be a way to preserve information useful for classification [34], [54], and it is likely that these dictionaries cover the signal space better than random pixels drawn from the labeled classes to form the exemplar dictionaries. Despite this, the coefficients of the learned dictionary do perform better than the random dictionary, demonstrating the value of the learning process. Finally, we should note that while we only display average classification errors, there is a wide variety in the per-class classification errors classes (i.e., some classes are inherently very challenging to distinguish because of their similar spectral features [5]). In our observations (not plotted), the relative difficulty of these classes in the classification task is roughly the same in the different data representations.

As mentioned earlier, one advantage of using classifiers with less complexity is that they may generalize better from the training data, especially when the training dataset is very small. We test the generalization ability of the SVM classification approach described above by repeating the experiment with variable sizes for the training dataset, in the extreme case using only one training pixel per class. We performed and evaluated this simulation in largely the same manner as described above, fixing $C = 10,000$ to achieve the lowest classification error and conservatively using 50 trials (i.e., random selections of training data for calculating a new SVM) to mitigate the

[6]We choose a training set size of 17 because we want the same amount of training data per class, and the smallest class has 18 labeled samples (leaving one testing pixel for the cross-validation). Average classification performance can be improved significantly on this dataset when larger training samples are used (but at the expense of consistent training set sizes per class).

[7]We note that in other simulations (not shown), the best classification performance of the SVN does not improve when using a nonlinear kernel such as a radial basis function (though the complexity obviously increases compared to the linear kernel). This indicates that linear decision boundaries are nearly optimal for this particular dataset, and little advantage is gained from a nonlinear mapping of the decision boundary. While in general we would hope to see lower possible classification error when using sparse coefficients, it appears that nonlinear mappings simply do not add much value to the decision boundaries for this particular dataset.

increased result sensitivity due to the low training set size. Figure 13 plots the results, showing that the sparse coefficients in the learned dictionary do in fact generalize better than the other methods, outperforming the other data representations for very small training set sizes (less than 12 training pixels from the total ground truth data).

IV. Conclusions

In this work we have shown that a sparse coding model and the dictionary learning approach described in [44] (with minor modifications) can yield valuable representations of HSI data using no *a priori* information about the dataset. The learned dictionary elements resemble many of the spectra corresponding to known material properties in the scene, and the sparse decomposition of the HSI data using this dictionary shows that the variations in the surface properties are often sensibly represented. In particular, in contrast with a typical endmember approach that seeks to contain the HSI data in a convex hull, this learned dictionary captures nonlinear material variations directly by forming a locally linear approximation to the manifolds observed within a material class.

The learned dictionaries capture many high-order statistics of the data they are learned from, and this representation showes advantages in applications relevant for remote sensing scenarios. For example, when coupled with a linear inverse problem, this learned dictionary demonstrated that HSI-resolution spectra could be recovered with remarkable fidelity from (simulated) spectra collected with just MSI-level resolution. This performance is only possible because the learned dictionaries are capable of effectively capturing the high level of statistical dependencies inherent in HSI data. Furthermore, encouraging results show that the performance on this task is still very good when there is some mismatch in the statistics because the training and testing data was collected at different times (i.e., a different season of the year, with different characteristics in the vegetation and the atmosphere). While this reconstruction problem was intended to mimic a realistic and useful data acquisition scenario, we note that this linear inverse problem framework captures many problems of interest (including other acquisition models such as those in compressed sensing [40]). Finally, we showed that the sparse coefficients from this learned dictionary form a useful representation for performing classification compared to the raw data, yielding classifiers with less complexity that generalize better when the training dataset size is very small.

From these results we can conclude that the sparse coding model is a potentially valuable approach to analyzing HSI data, and the learned dictionaries for this model form a meaningful representation of the high-order statistics in the HSI data. While this approach shares the same linear model as the common endmember approach for spectral unmixing, the different philosophy of representing the data variations directly appears to have value both in the general understanding of the data and in specific applications. We believe that this exploration (along with the other related results in [16], [28], [29], [55]) demonstrates that more extensive exploration of the utility of this model in HSI is warranted, and improvements in many specific applications are likely. In the future, in addition to more thorough application of these ideas to other datasets, it will be valuable to explore the utility

of including increasingly complex models in in the learning process. For example, there may be potential benefits to learning much larger dictionaries than those shown in this work, learning joint spectral-spatial dictionaries, learning dictionaries customized for specific applications (such as in [16]), and learning dictionaries that attempt to explicitly capture features such as correlations between pixels and nonlinear variations within material classes.

REFERENCES

[1] "The benefits of the 8-spectral bands of worldview-2," Mar 2010, available Online at http://www.digitalglobe.com/downloads/spacecraft/WorldView-2_8-Band_Applications_Whitepaper.pdf.

[2] M. Aharon, M. Elad, A. Bruckstein, and Y. Katz, "K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations," *IEEE Proceedings - Special Issue on Applications of Compressive Sensing & Sparse Representation*.

[3] M. Aharon, M. Elad, and A. Bruckstein.

[4] C. M. Bachmann, "Improving the performance of classifiers in high-dimensional remote sensing applications: An adaptive resampling strategy for error-prone exemplars (ARESEPE)," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 9, pp. 2101–2112, Sept 2003.

[5] C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina, "Exploiting manifold geometry in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 441–454, 2005.

[6] C. M. Bachmann, M. H. Bettenhausen, R. A. Fusina, T. F. Donato, A. L. Russ, J. W. Burke, G. M. Lamela, W. J. Rhea, and B. R. Truitt, "A credit assignment approach to fusing classifiers of multiseasonal hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 11, pp. 2488–2499, Nov 2003.

[7] C. M. Bachmann, T. F. Donato, G. M. Lamela, W. J. Rhea, M. H. Bettenhausen, R. A. Fusina, K. R. D. Bois, J. H. Porter, and B. R. Truitt, "Automatic classification of land cover on smith island, va, using hymap imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 10, pp. 2313–2330, Oct 2002.

[8] C. M. Bachmann, T. F. Donato, G. Lamela, J. Rhea, M. Bettenhausen, R. A. Fusina, K. D. Bois, J. Porter, and B. Truitt, "Automatic classification of land cover on Smith Island, VA, using HyMAP imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 10, pp. 2313–2330, 2002.

[9] A. Bateson and B. Curtiss, "A method for manual endmember selection and spectral unmixing," *Remote Sens. Environ.*, vol. 55, no. 3, pp. 229–243, Mar 1996.

[10] M. Berman, H. Kiiveri, R. Lagerstrom, A. Ernst, R. Dunne, and J. F. Huntington, "ICE: A statistical approach to itentifying endmembers in hyperspectral images," *IEEE Transaction on Geoscience and Remote Sensing*, vol. 42, no. 10, pp. 2085–2095.

[11] J. Bioucas-Dias and M. Figueiredo, "Alternating direction algorithms for constrained sparse regression application to hyperspectral unmixing," *2nd IEEE GRSS Workshop on Hyperspectral Image and Signal Processing -WHISPERS'2010*, raykjavik, Iceland.

[12] J. Bowles, P. J. Palmadesso, J. A. Antoniadas, M. M. Baumback, and J. L. Rickard, "Use of filter vectors in hyperspectral data analysis," *Proceedings of SPIE, Infrared Spaceborne Remote Sensing III*, pp. 148–157, 1995.

[13] A. M. Bruckstein, M. Elad, and M. Zibulevsky, "On the uniqueness of nonegative sparse solutions to underdetermined sparse solutions to underdetermined systems of equations," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 4813–4820, Nov 2008.

[14] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[15] E. Candes and J. Romberg, "ℓ^1-Magic: Recovery of sparse signals via convex programming," 2005, http://www.acm.caltech.edu/l1magic/.

[16] A. Castrodad, Z. Xing, J. Greer, E. Bosch, L. Carin, and G. Sapiro, "Discriminative sparse representations in hyperspectral imagery," *IMA Preprint Series #2319*, Mar 2010.

[17] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[18] A. S. Charles, B. A. Olshausen, and C. J. Rozell, "Sparse coding for spectral signatures in hyperspectral images," *Asilomar Conference on Signals, Systems and Computers*, 2010.

[19] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[20] M. Elad, M. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *IEEE Proceedings - Special Issue on Applications of Compressive Sensing & Sparse Representation*, Oct 2008.

[21] S. Erard, P. Drossart, and G. Piccioni, "Multivariate analysis of visible and infrared thermal imaging spectrometer (virtis) venus express nightside and limb observations," *Journal of Geophysical Research*, vol. 114, 2009.

[22] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, 2007.

[23] O. Forni, S. M. Clegg, R. C. Wiens, and S. M. anc O. Gasnault, "Multivariate analysis of ChemCam first calibration samples," *40th Lunar and Planetary Science Conference*, vol. 1523, 2009.

[24] O. Forni, F. Poulet, J.-P. Bibring, S. Erard, C. Gomez, Y. Langevin, B. Gondet, and Omega Team, "Component separation of omega spectra with ica," *Lunar and Planetary Science Technical Report*, vol. 1623, 2005.

[25] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz, "Single-shot compressive spectral imaging with dual-disperser architecture," *Optics Express*, vol. 15, no. 21, pp. 14 013–14 026, Oct 2007.

[26] Q. Geng, H. Wang, and J. Wright, "On the local correctness of ℓ^1-minimization for dictionary learning," no. arXiv:1101.5672v1, 2011, http://arxiv4.library.cornell.edu/pdf/1101.5672v1.

[27] C. Gomez, H. L. Borgne, P. Allemand, C. Delacourt, and P. Ledru, "N-FindR method versus independent component analysis for lithological identification in hyperspectral imagery," *International Journal of Remote Sensing*, vol. 28, no. 23, pp. 5315–5338, Jun 2007.

[28] J. Greer, "Sparse demixing," *IEEE Proceedings of SPIE*, vol. 7695, May 2010.

[29] Z. Guo, T. Wittman, and S. Osher, "L1 unmixing and its application to hyperspectral image enhancement," *Proc. SPIE Conference on Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV*, orlando, Florida.

[30] E. T. Hale, W. Yin, and Y. Zhang, "A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing," Rice University Department of Computational and Applied Mathematics, Tech. Rep., Jul 2007.

[31] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, no. 9, pp. 1457–1469, Nov 2004.

[32] A. Ifarraguerri and C.-I. Chang, "Multispectral and hyperspectral image analysis with convex cones," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 2, pp. 756–770, Mar 1999.

[33] M.-D. Iordache, J. M. B. Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, 2010.

[34] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" *IEEE Proceedings International Conference on Computer Vision (ICCV'09)*, 2009.

[35] S. Jia and Y. Qian, "Constrained nonnegative matrix factorization for hyperspectral imaging," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 1, pp. 161–173, Jan 2009.

[36] J. P. Kerkes and J. R. Schott, "Hyperspectral imaging systems," in *Hyperspectral Data Exploitation: Theory and Applications*, C.-I. Chang, Ed. John Wiley & Sons, Inc., 2007, pp. 19–45.

[37] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Proceesing Magazine*, vol. 19, no. 1, pp. 44–57.

[38] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large scale l1-regularized least squares," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, Dec 2007.

[39] D. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Proceesings on Signal Processing*, vol. 19, no. 1, pp. 29–43, Jan 2002.

[40] R. Maraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, Jul 2007.

[41] S. Moussaoui, H. Hauksdottir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, S. Doute, and J. A. Benediksson, "On the decomposition of mars hyperspectral data by ica and bayesian positive source separation," *Neurocomputing*, vol. 71, no. 10-12, pp. 2194–2208, Jun 2008.

[42] K. L. Oehler and R. M. Gray, "Combining image compression and classification using vector quantization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, May 1995.

[43] B. A. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 13, pp. 607–609, Jun 1996.

[44] ——, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[45] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs," *Current Opinion in Neurobiology*, vol. 14, pp. 481–487, 2004.

[46] S. Osher, B. D. Y. Mao, and W. Yin, "Fast linearized bregman iteration for compressive sensing and sparse denoising," *Advances in Neural Information Processing Systems*, pp. 505–512, 2008.

[47] V. P. Pauca, J. Piper, and R. J. Plemmons, "Nonnegative matrix factorization for spectral data analysis," *Journal of Linear Algebra and its Applications*, vol. 416, no. 1, pp. 29–47, 2006.

[48] A. Plaza, P. Martinez, R. Perez, and J. Plaza, "Spatial/spectral endmember extraction by multidimensional morphological operations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 9, pp. 2025–2041, Sep 2002.

[49] R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng, "Self-taught learning: Transfer learning from unlabled data," *IEEE Proceedings International Conference on Computer Vision (ICCV'07)*, 2007.

[50] D. Rogge, B. Rivard, J. Zhang, and J. Feng, "Iterative spectral unmixing for optimizing per-pixel endmember sets," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 12, Dec 2006.

[51] T. Wachtler, T. Lee, and T. J. Sejnowski, "Chromatic structure od natural scenes," *Journal of the Optical Society of America*, vol. 18, no. 1, pp. 65–77, Jan 2001.

[52] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Computational Optical Sensing and Imaging, Applied Optics*, vol. 47, no. 10, pp. B44–B51, Apr 2008.

[53] M. Winter, "N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data," *IEEE Porceedings of SPIE*, vol. 3753, no. 5, 1999.

[54] A. Y. Yang, J. Wright, Y. Ma, and S. Sastry, "Feature selection in face recognition:a sparse representation perspective," University of California, Berkeley, Tech. Rep. UCB/EECS-2007-99, 2007.

[55] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, D. Dunson, G. Sapiro, and L. Carin, "Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images," *IMA Preprint Series #2307*, Apr 2010.