

**High Dimensional Data Modeling Techniques for Detection
of Chemical Plumes and Anomalies in Hyperspectral Images
and Movies**

Journal:	<i>Journal of Selected Topics in Applied Earth Observations and Remote Sensing</i>
Manuscript ID	JSTARS-2015-00702
Manuscript type:	hyperspectral_remote_sensing_and_imaging_spectroscopy
Date Submitted by the Author:	15-Sep-2015
Complete List of Authors:	Wang, Yi; Syracuse University, Mathematics Chen, Guangliang; San Jose State University, Department of Mathematics and Statistics Maggioni, Mauro; Duke University, Electrical and Computer Engineering; Duke University, Computer Science
Keywords:	Remote sensing

High Dimensional Data Modeling Techniques for Detection of Chemical Plumes and Anomalies in Hyperspectral Images and Movies

Yi (Grace) Wang, Guangliang Chen, and Mauro Maggioni

Abstract—We discuss recent progress in techniques for modeling and analyzing hyperspectral images and movies, in particular for detecting plumes of both known and unknown chemicals. We discuss novel techniques for robust modeling of the background in a hyperspectral scene, and for detecting chemicals of known spectrum, we use partial least squares regression on a resampled training set to boost performance. For the detection of unknown chemicals we view the problem as an anomaly detection problem, and use novel estimators with low-sampled complexity for intrinsically low-dimensional data in high-dimensions that enable use to model the “normal” spectra and detect anomalies. We apply these algorithms to benchmark data sets made available by Lincoln Labs at the Automated Target Detection program co-funded by NSF, DTRA and NGA, and compare, when applicable, to current state of art algorithms, with favorable results.

Index Terms—Remote sensing, mixture models, robust modeling chemical plumes, automated detection.

I. INTRODUCTION

The ability to remotely sense and analyze chemical plumes has become increasingly important in many civilian and military applications. Hyperspectral imaging sensors that operate in the long-wave infrared (LWIR) part of the spectrum are particularly well suited for these chemical-sensing tasks, because their wide fields of view and high spectral resolutions allow many square kilometers to be imaged almost simultaneously and even optically thin chemical clouds to be detected [1]. Data collected by these sensors has the form of $m \times n \times p$ arrays, where m, n are the spatial dimensions and p is the spectral dimension. Under physically reasonable assumptions and simplifications [1], the following linear mixing model is obtained:

$$\mathbf{x} = \sum_{1 \leq i \leq N} g_i \mathbf{s}_i + \mathbf{v}, \quad (1)$$

where $\mathbf{x} \in \mathbb{R}^p$ represents the radiance spectrum in the scene, N the number of chemicals of interest, $\mathbf{s}_i \in \mathbb{R}^p$ the signature spectrum for the i -th chemical of interest, $g_i \geq 0$ the amount of chemical i , and \mathbf{v} the radiance spectrum of the background. In practice, it is usually assumed that $N \leq 3$. In this paper, due

to the data under consideration, we consider the case $N = 1$ (thus, $\mathbf{x} = g\mathbf{s} + \mathbf{v}$).

To effectively separate the chemical clouds from the background clutter, one needs to choose a proper model for the background radiation \mathbf{v} . Current approaches (e.g. [2], [3]) often represent the background by a single Gaussian distribution or subspace and then derive corresponding statistical estimators which assign detection scores to each spectrum (see the review [1]). These detection algorithms have shown their effectiveness in many applications, but there is room for improvement. In particular, background often consists of heterogeneous regions (such as sky, mountain, desert; see e.g. [1, Fig. 9]) which may require separate subspaces to better capture the complexity of the background.

Our contributions in this work are threefold: (a) we extend the single-subspace model [3] for modeling the background to a “mixture of subspaces” model; (b) we propose techniques to enhance the detectability of the chemical plume region, moving beyond classical least squares and using resampling techniques to boost detection; and (c) when detection of chemicals with unknown signature is of interest, we propose a flexible anomaly detection procedure which efficiently constructs an empirical model of “normal” spectra and flags anomalous ones according to likelihood assigned by the empirical model. In particular, (c) is achieved by applying an efficient multiscale transform [4] to the hyperspectral spectra of a training frame and then model the density of the transformed pixels for computing the likelihood of each spectrum in any testing frame: if the testing frame contains gas, then the gas region will be flagged as an anomaly if it is assigned low likelihood values.

We consider two different scenarios, depending on whether we are given only a single hyperspectral cube or a time series of them. In the former scenario, we use the available cube both for learning an empirical model for the background and for chemical detection, while in the second one we use the first few frames (assumed to be without chemical plume) for background modeling and any subsequent frame for testing for anomalies based on the learned background model. If the sensor is assumed stationary, one could use the temporal variation of the spectrum at each location to detect changes and detect anomalies: we do *not* assume that the sensor is stationary (albeit it is such in most of the data we consider) and therefore we do not use spatial coherence information across the images. This makes the problem harder, but the proposed solution applicable in a wider variety of settings.

Manuscript received September 15, 2015. The work of M. Maggioni was supported by the ATD program funded by NSF, NGA and DTRA, under award # 1222567, and by the DHS CAXSI program.

Y. Wang is with the Mathematics Department, Syracuse University, Syracuse, NY 13244, United States.

G. Chen is with Department of Mathematics and Statistics, San José State University, San José, CA 95192, United States.

M. Maggioni is with the Departments of Mathematics, Electrical and Computer Engineering, Computer Science, Duke University, Durham, NC 27708, United States.

II. PREVIOUS WORK FOR KNOWN TARGET SIGNATURE

Based on different assumptions on the background radiation \mathbf{v} , different detection algorithms have been proposed for the setting when the signature of the target plume is known. In this section we review some of the existing approaches by following the presentation of [1].

A. Gaussian Models

The simplest and most practical algorithm for chemical gas detection uses the linear model in (1) and assumes normally distributed background clutter with known covariance, that is,

$$\mathbf{x} = \mathbf{s}g + \mathbf{v}, \quad \mathbf{v} \sim \mathcal{N}(\mu, \sigma^2 \Sigma)$$

where the plume spectral signature \mathbf{s} is also assumed to be known. The coefficients g and σ^2 can be estimated by maximizing the likelihood from a given sample. The gas detection problem can be formalized as a hypothesis testing one:

$$H_0 : \text{Plume absent } (g = 0);$$

$$H_1 : \text{Plume present } (g > 0).$$

By applying the generalized likelihood ratio test (GLRT) one obtains the following detector:

$$T_{\text{NMF}}(\mathbf{x} | \mu, \Sigma, \mathbf{s}) = \frac{(\mathbf{s}^T \Sigma^{-1} \tilde{\mathbf{x}})^2}{(\mathbf{s}^T \Sigma^{-1} \mathbf{s})(\tilde{\mathbf{x}}^T \Sigma^{-1} \tilde{\mathbf{x}})}, \quad \tilde{\mathbf{x}} = \mathbf{x} - \mu \quad (2)$$

This is known as the normalized matched filter (NMF) and also the adaptive cosine or coherence estimator (ACE) [2]. Since the NMF requires estimation and inversion of the background covariance matrix Σ , they are approximated in the following way in order to avoid numerical instability and to obtain robust detector.

$$\hat{\Sigma} = \sum_{k=1}^p \lambda_k \mathbf{q}_k \mathbf{q}_k^T + \delta I, \quad \hat{\Sigma}^{-1} = \sum_{k=1}^p \frac{1}{\lambda_k + \delta} \mathbf{q}_k \mathbf{q}_k^T$$

where λ_k and \mathbf{q}_k ($k = 1, \dots, p$) are the eigenvalues and eigenvectors of the sample variance, δ is a regularization parameter and can be chosen as certain percentile of the values of $\{\lambda_k\}$. In this paper, we choose δ as the median of $\{\lambda_k\}$.

B. Subspace Model

Another category of approaches for chemical plume detection assumes that the background clutter can be well represented by a low-dimensional subspace. In that case, the signal model becomes

$$\mathbf{x} = \mathbf{s}g + \mu + B\mathbf{c} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

where μ, B represent a fixed point and a basis for the background subspace and ϵ is the additive noise. To find μ, B , principal component analysis (PCA) is applied to the background spectra. Similarly to before, the quantities g, \mathbf{c} and σ are estimated by maximizing the likelihood. Due to the assumption of normally distributed error ϵ , the maximum likelihood estimators (MLE) of g and \mathbf{c} may be computed by least squares.

The GLRT approach yields the following detector [3]:

$$T_{\text{NSS}}(\mathbf{x} | \mu, B, \mathbf{s}) = \frac{\|P_b^\perp \mathbf{x}\|^2}{\|P_{tb}^\perp \mathbf{x}\|^2} \quad (3)$$

where P_b^\perp and P_{tb}^\perp are projection matrices:

$$P_b^\perp = I - B(B^T B)^{-1} B^T;$$

$$P_{tb}^\perp = I - A(A^T A)^{-1} A^T, \quad A = [\mathbf{s} B].$$

Note that $\|P_b^\perp \mathbf{x}\|$ and $\|P_{tb}^\perp \mathbf{x}\|$ are respectively the orthogonal distances from the background and target-background subspaces.

III. MIXTURE MODELS AND ENHANCEMENT WITH KNOWN TARGET SIGNATURE

In this section, we describe both the mixture models and some techniques to enhance the detection. Our algorithm is presented at the end of this section.

A. Mixture Models

Gaussian mixture models $\mathbf{v} \sim \sum_j \pi_j \mathcal{N}(\mu_j, \Sigma_j)$, where \mathcal{N} represents the Normal distribution, may be used to capture complex backgrounds. We are particularly interested in the case where Σ_j is rank-deficient, and therefore $\mathcal{N}(\mu_j, \Sigma_j)$ is supported on an affine subspace spanned by $\text{cols}(B_j)$, the columns of a matrix B_j . More generally, in a *subspace mixture model* $\mathbf{v} \sim \sum_j \pi_j \mathcal{S}(\mu_j, B_j)$ where $\mathcal{S}(\mu_j, B_j)$ is a probability measure with mean μ_j and support contained in the subspace spanned by the columns of B_j . The model parameters π_j and Θ_j ((μ_j, Σ_j) or (μ_j, B_j)) in each case can be estimated by K -means, K -means-like subspace clustering algorithms (e.g. [5], [6], [7]), fast multiscale techniques [8], or Expectation-Maximization (EM) methods, through iterations between updating cluster assignments and model parameters. The signal \mathbf{x} is then assigned to the cluster that maximizes the estimation of $\pi_j p(\mathbf{x} | \Theta_j)$:

$$\hat{j} = \arg \max_j \pi_j p(\mathbf{x} | \Theta_j).$$

Given a target plume signature \mathbf{s} , the mixture versions of the two estimators, NMF (also known as ACE) and NSS, are given by

$$T_{\text{mixNMF}}(\mathbf{x} | \mathbf{s}, \{\pi_j, \Theta_j\}) = T_{\text{NMF}}(\mathbf{x} | \mathbf{s}, \mu_{\hat{j}}, \Sigma_{\hat{j}});$$

$$T_{\text{mixNSS}}(\mathbf{x} | \mathbf{s}, \{\pi_j, \Theta_j\}) = T_{\text{NSS}}(\mathbf{x} | \mathbf{s}, \mu_{\hat{j}}, B_{\hat{j}}). \quad (4)$$

Alternatively, for the subspace mixture model, we may simply use the coefficient g as the detector and solve for it by least squares. Specifically, letting \hat{j} be defined as above, it is easy to show that the least squares estimator \hat{g} of g is the first entry of

$$\hat{\beta} = (A_{\hat{j}}^T A_{\hat{j}})^{-1} A_{\hat{j}}^T (\mathbf{x} - \mu_{\hat{j}}), \quad A_{\hat{j}} = [\mathbf{s} B_{\hat{j}}].$$

This yields the mixture Linear Coefficient (LC) estimator

$$T_{\text{mixLC}}(\mathbf{x} | \mathbf{s}, \{\pi_j, \Theta_j\}) = \max\{\hat{g}, 0\}.$$

Finally, in section IV we consider a related but different family of multiscale models for background spectra, in the context of anomaly detection in hyperspectral movies, in the setting when the spectral signature \mathbf{s} of the chemical of interest is not given, nor the frames in the movies are assumed to be registered.

B. Enhancement Techniques

We propose a few enhancement techniques for background estimation to reduce the affects of the contamination:

1) *Outlier Removal*: We detect and remove a small fraction of the spectra that are dissimilar to the main part of the data, in terms of magnitude or connectivity (in this paper we simply classify those spectra with largest magnitudes along the spectral dimension as outliers).

2) *Resampling Enhancement*: This technique is relevant only when we are in Scenario (I), where we need to learn the background model and detect chemical plume region simultaneously. For this goal we utilize an iterative scheme. We first choose a few likely background spectra based on a reliable detection score (output of a detection algorithm, e.g. ACE), and then select their spatial neighbors as well, since adjacent pixels are very likely to be of the same category. Using these selected spectra, the background model parameters are re-estimated and the detection statistics are re-deduced accordingly.

Algorithm 1 Resampling Enhancement

Input: $\{\mathbf{x}_i\}_{i=1}^{mn} \subset \mathbb{R}^P$: spectra; $\{T_i\}_{i=1}^{mn} \subset \mathbb{R}$: detection score; $\mathbf{s} \in \mathbb{R}^P$: plume signature; $\tau_1 \in (0, 1)$: portion.

Output: $\{y_i\}_{i=1}^{mn} \subset \mathbb{R}$: enhanced detection score.

- 1: Sort T_i s.t., $T_{(1)} \leq \dots \leq T_{(mn)}$.
 - 2: Choose $\delta_1 = T_{(\lceil \tau_1 mn \rceil)}$. Let $A_1 := \{\mathbf{x}_i : T_i \leq \delta_1\}$.
 - 3: Let B be the union of A_1 and the 4 (above, below, left and right) spatial neighbors of $\mathbf{x}_i \in A_1$.
 - 4: Re-estimate model parameters Θ_j 's from B .
 - 5: Compute the statistic (detection score) based on the updated model.
-

3) *Partial Least Squares Regression (PLSR) Enhancement*: PLSR [9] reduces the dimensionality of the data in the way that the covariance between predictors and responses is maximized. The response is re-estimated on the reduced predictors. In our problem, PLSR is applied so that the radiance data is projected to a subspace that is most relevant to the detection score (output of a detection algorithm, or enhanced detection score resulted from resampling). A new and enhanced detection score is computed from the projected radiance data. In fact, we select a certain amount of spectra which are most likely to be background as well as chemical clouds and apply PLSR on the selections.

Algorithm 2 PLSR Enhancement

Input: $\{\mathbf{x}_i\}_{i=1}^{mn} \subset \mathbb{R}^P$: spectra; $\{T_i\}_{i=1}^{mn} \subset \mathbb{R}$: detection score; $\mathbf{s} \in \mathbb{R}^P$: plume signature; $\tau_2, \tau_3 \in (0, 1)$: portions.

Output: $\{y_i\}_{i=1}^{mn} \subset \mathbb{R}$: enhanced detection score.

- 1: Sort T_i s.t., $T_{(1)} \leq \dots \leq T_{(mn)}$.
 - 2: Choose $\delta_2 = T_{(\lceil \tau_2 mn \rceil)}$ and $\delta_3 = T_{(\lceil (1-\tau_3) mn \rceil)}$. Let $A_2 := \{\mathbf{x}_i : T_i \leq \delta_2\}$ and $A_3 := \{\mathbf{x}_i : T_i \geq \delta_3\}$.
 - 3: Apply PLSR on the selected pairs (\mathbf{x}_i, T_i) with $\mathbf{x}_i \in A_2 \cup A_3$ to obtain parameters $\beta \in \mathbb{R}^P$ and $\beta_0 \in \mathbb{R}$.
 - 4: $y_i = \beta^T \mathbf{x}_i + \beta_0, i = 1, \dots, mn$.
-

C. Algorithm for plume detection in hyperspectral images or movies

We propose a unifying algorithm (see Alg. 3) that can detect chemical plumes in both scenarios: in scenario 1 when we have only a single hyperspectral image, we incorporate the enhancement techniques proposed in the previous section to learn the background via mixture modeling; in the second scenario where we are given a time series of hyperspectral

cubes, we assume that the first few frames were collected before the chemical release so that we may use all their spectra for background modeling. Afterwards, we select a detector (from $T_{\text{mixNMF}}, T_{\text{mixNSS}}, T_{\text{mixLC}}$) and apply it to the given image(s).

Algorithm 3 Chemical plume detection through mixture background modeling in hyperspectral images or movies

Input: $\{\mathbf{x}_i^{(\ell)}\}_{i=1}^{mn}, 1 \leq \ell \leq L$: L hyperspectral frame(s); $\mathbf{s} \in \mathbb{R}^P$: plume signature, and detector (one of the $T_{\text{mixNMF}}, T_{\text{mixNSS}}, T_{\text{mixLC}}$).

Output: $\{T(\mathbf{x}_i^{(\ell)})\} \subset \mathbb{R}$: pixelwise detection score

- 1: Fit a mixture model $\{\Theta_j\}$ to the given cube (when $L = 1$) or the first few clean frames (when $L > 1$). Apply the enhancement techniques if relevant.
- 2: **for** each frame $\ell = 1, \dots, L$,
 - Assign spectra of the ℓ -th frame to nearest component models by maximizing the estimator of $\pi_j p(\mathbf{x}|\Theta_j)$
 - Evaluate the given detector at all pixels of the frame (within the corresponding clusters) to produce detection scores

end for

IV. ANOMALY DETECTION IN HYPERSPECTRAL MOVIES WITH UNKNOWN TARGET SIGNATURE

In this section we consider the problem of anomaly detection instead of that of detection of known chemicals. In particular, we are interested in the application to hyperspectral movies, where at unknown time and location a gas is released and the gas plume needs to be tracked. We think of the chemical plume region, once released, as a set of anomalous spectra, when compared against the background clutter, and thus we base detection on accurately estimating the probability measure modeling the space of the background pixels and computing the likelihood scores of every spectrum in any testing frame relative to the density model.

In the original paper on Geometric Multi-Resolution Analysis (GMRA) [4], an approach for approximating probability distributions in high-dimensions using the intrinsically low-dimensional GMRA structure was suggested, and those ideas were further developed in [8], [12]. In this case we are not given signatures of spectra to detect; instead we are given one or more hyperspectral scenes defined “normal” (a training set), and given a new hyperspectral scene we are interested in deciding if its spectra are normal or present “anomalies”. We model this problem as follows: we assume that there is an unknown probability measure μ in \mathbb{R}^P from which “normal” spectra are drawn. The training set $\mathbf{s} := \{\mathbf{s}_i\}_{i=1}^N \subseteq \mathbb{R}^P$ consisting of all the spectra in the training hyperspectral scenes is modeled as n i.i.d. samples from μ .¹ We use these n samples to learn a probability measure $\hat{\mu}_{\mathbf{s}}$ approximating, in a suitable sense, μ . Given a new scene, i.e. a new set of samples $\mathbf{s}^{\text{new}} = \{\mathbf{s}_i^{\text{new}}\}$, we could ask what is the probability of seeing $\mathbf{s}_i^{\text{new}}$ if it was sampled from μ , and call $\mathbf{s}_i^{\text{new}}$ an anomaly if this probability is below a certain threshold. Unfortunately this does not make sense since typically μ (and our estimator $\hat{\mu}_{\mathbf{s}}$) do not assign positive probability to any point. Often one then replaces probabilities by probability densities and associated likelihoods. An alternative is to replace the question above by evaluating the

¹Clearly, independence is a rather strong assumption, but could be relaxed with only rather minor technical difficulties to more general settings that accommodate mild dependencies.

probability (according to $\hat{\mu}_s$) of seeing a point within distance r from s_i^{new} , and decide whether s_i^{new} is an anomaly or not based on a threshold on such probability, i.e. we declare s_i^{new} an anomaly if $\mathbb{P}_{\hat{\mu}_s}(B_r(s_i^{\text{new}})) < \eta$. The values of η and r tune the sensitivity of the anomaly detection. We may choose them by first fixing η , then choosing r minimizing type I or type II error (or other similar criterion), or by choosing r to be smallest value such that $\mathbb{P}_{\hat{\mu}_s}(B_r(s_i^{\text{val}})) > \eta$ for all i 's, where $\{s_i^{\text{val}}\}$ is a validation data set (possibly extracted and excluded from a training set). Then as we vary η we obtain a ROC for the anomaly detector (assuming we know the ground truth).

The key problem here is to efficiently construct an estimator $\hat{\mu}_s$: this is a challenging task, since μ is in \mathbb{R}^p , with p typically large ($p > 100$ is common). This is accomplished by using techniques based on GMRA, originally suggested in [4], and further developed and analyzed in [8], [13]. We have no space to describe the details of these constructions. At a high level, one can think of GMRA as a principled and efficient way of encoding a data set that, while high-dimensional, is intrinsically low-dimensional, at different levels of accuracy, by first constructing a multiscale tree decomposition of the data, and then in each node of the tree, corresponding to a portion of the data, construct a data-driven low-dimensional projection of the data, and further compressing these projections by encoding the difference between the projection at one node and those of children nodes in the tree [4]. The results in [13] guarantee that under suitable geometric assumptions about the data – essentially the data should be close to a(n unknown) low-dimensional manifold – GMRA run on the data will efficiently construct sparse representations of the data. In order to extend this construction to the estimation of probability measure, at each node of the GMRA tree we estimate a simple probability measure (e.g. a low-rank Gaussian) and combine these measures appropriately to construct a probability measure $\hat{\mu}_s$ approximating μ . The results in [8], [12] guarantee that – under suitable geometric assumptions on the data and on the regularity of μ_s – the estimator $\hat{\mu}_s$ gets increasingly closer to μ (in Wasserstein distance) with high probability as N increases, and with a rate that depends only on the intrinsic dimension d of the data and not on the ambient dimension p . Furthermore, these constructions are yielded by efficient algorithms, having complexity essentially linear in the training data size pN , with a constant depending essentially on the intrinsic dimension of the data d . In the setting of hyperspectral images considered here, $N = mn$ is the number of pixels in an image. The intrinsic dimension of the data, measured by Multiscale SVD [14], [15], [16] is often very small in hyperspectral data, between 1 and 5. Moreover, these algorithms [8]–[12] easily allow to quickly (in $O_d(\log N)$) incorporate new samples in an online fashion, in both the GMRA construction and the estimator $\hat{\mu}_s$. Finally, the underlying GMRA construction, simultaneously to the above, yields a dictionary that sparsifies the data (in our case the spectra) [13], [16], enabling the compression of the data. These representations also lead to a variation of Compressed Sensing where the data model is that of a nonlinear manifold, with extremely efficient inversion algorithms that do not require the solution of high-dimensional convex optimization problems [8].

Algorithm 4 Multiscale-transform based Density Estimation

Input: Data set X_n

Output: Multiscale densities $\{\hat{\mu}_{j,k}\}_{j \geq 0, k \in \Lambda_j}$

- 1: Apply GMRA to the training data to obtain a multiscale dictionary $\{c_{j,k}, \Phi_{j,k}, w_{j,k}, \Psi_{j,k}\}_{j \geq 0, k \in \Lambda_j}$
- 2: For each $j > 0, k \in \Lambda_j$, apply the Geometric Wavelet Transform to the data in each $C_{j,k}$ and obtain low dimensional coefficients
- 3: Apply a density estimator (e.g. KDE [10]) to each set of geometric scaling and wavelet coefficients corresponding to each $C_{j,k}$, and obtain density estimates $\hat{\mu}_{j,k}$; also record $\hat{\pi}_{j,k}$, the empirical probability of a point belonging to $C_{j,k}$.
- 4: By testing on a validation set, select an optimal scale j^* and return the mixture model $\sum_k \hat{\pi}_{j^*,k} \hat{\mu}_{j^*,k}$.

V. EXPERIMENTAL RESULTS

In this section we consider both synthetic data sets and data sets at the repository at Colorado State and made available through the ATD program (*Algorithms for Threat Detection*), co-sponsored by NSF, DTRA and NGA.

A. Synthetic data sets

1) Gaussian mixture models

Description. We use synthetic data set to realize the Gaussian mixture model. We take the MIT Lincoln Lab Challenge Data (see details in the next section) to obtain the mean spectra for three regions – sky, mountain and ground (denoted by μ_1 , μ_2 and μ_3 respectively) and the target plume signature s . These spectra consist of 68 measurements at different values of wavelength and are shown in Figure 1.

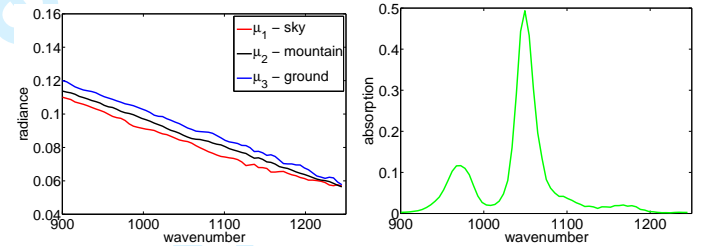


Fig. 1: Left: the mean spectra of sky, mountain and ground. Right: the absorption of the chemical plume.

Moreover, 5,000, 5,000 and 4,000 spectra are generated i.i.d. from Gaussian distributions $\mathcal{N}(\mu_1, \Sigma_1)$, $\mathcal{N}(\mu_2, \Sigma_2)$ and $\mathcal{N}(\mu_3, \Sigma_3)$ respectively for the three regions, where $\Sigma_i = \text{diag}\{\sigma_{i,1}^2, \dots, \sigma_{i,68}^2\}$, $i = 1, 2, 3$ and all $\sigma_{i,j}$'s are drawn i.i.d. from the uniform distribution on $[0.002a, 0.008a]$ with $a = \max(\mu_3)$. Finally, 1,000 spectra on the ground with chemical plume are generated from $gs + v$, where $v \sim \mathcal{N}(\mu_3, \Sigma_3)$ and $g \sim \mathcal{N}(-0.01, 0.001)$. Top left of Figure 2 displays 5 sample spectra for each of the groups – sky, mountain, ground and plume.

Task. Detect the plume and compare the Gaussian mixture model with the single Gaussian model by the detection.

Technique. We compute the ACE (or NMF) statistics (detection scores) using the ground truth. More precisely, we compute T_{mixNMF} by (4) using the truth values of μ_i and Σ_i with $i = 1, 2, 3$ and T_{NMF} by (2) using $\mu = (\mu_1 + \mu_2 + \mu_3)/3$ and $\Sigma = (\Sigma_1 + \Sigma_2 + \Sigma_3)/3$ for all spectra.

Results. Ten sample ACE statistics each from the 4 groups are demonstrated on the top right of Figure 2 for both the

single Gaussian and mixture Gaussian models. In addition, the summary of these statistics are shown as boxplot at the bottom of Figure 2, with the single Gaussian on the left and the mixture Gaussian on the right. The boxes are colored by groups of sky, mountain, ground and plume. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers. These figures show that when using three Gaussians instead of a single Gaussian, the ACE statistics of the plumes are greatly larger than those of the other regions. This makes the plumes more separable and thus more detectable and is also consistent with our understanding about the ACE algorithm, i.e., the spectra with plume are of a smaller angle from the plume signature and thus have larger values of the ACE statistics.

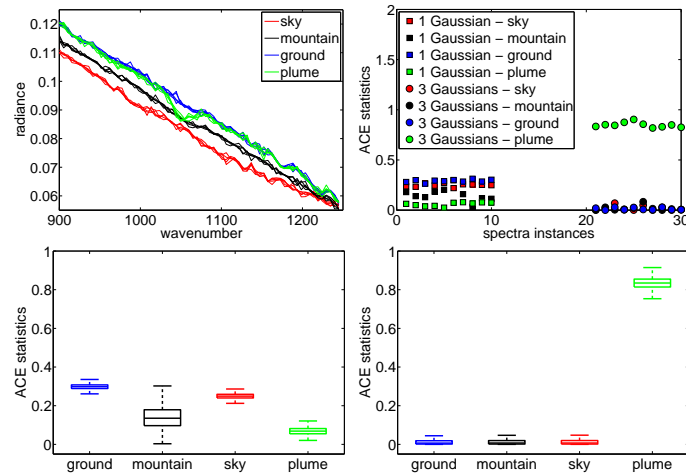


Fig. 2: Top left: radiance against wavenumber of 5 samples each from the groups of sky, mountain, ground and plume. Top right: the ACE scores of 10 spectra samples computed by 1 Gaussian and 3 Gaussians respectively. Bottom: the boxplot of the ACE scores of using 1 Gaussian (left) and 3 Gaussians (right).

2) Subspace mixture models

Description. We use synthetic data set to realize the subspace mixture model. 5,000, 5,000 and 4,000 spectra are generated i.i.d. from $c\mu_1 + \epsilon$, $c\mu_2 + \epsilon$ and $c\mu_3 + \epsilon$ respectively for sky, mountain and ground regions, where $c \sim \mathcal{N}(1, 0.01)$, $\epsilon \sim \mathcal{N}(0, \Sigma_\epsilon)$, $\Sigma_\epsilon = \text{diag}\{\sigma_{\epsilon,1}^2, \dots, \sigma_{\epsilon,68}^2\}$ and $\sigma_{\epsilon,j}$'s are drawn i.i.d. from $\mathcal{N}(0, 0.005 \max(\mu_3))$. Then 1,000 spectra on the ground with plume are generated as $gs + c\mu_3 + \epsilon$ with $g \sim \mathcal{N}(-0.01, 0.001)$. Top left of Figure 3 displays 5 sample spectra for each of the groups - sky, mountain, ground and plume.

Task. Detect the plume and compare the subspace mixture model with the single subspace model by the detection.

Technique. We compute the NSS statistics (detection scores) using the ground truth. More precisely, we compute T_{mixNSS} by (4) letting $B_i = \mu_i$, $i = 1, 2, 3$ and T_{NSS} by (3) using $B = (\mu_1 + \mu_2 + \mu_3)/3$.

Results. Ten sample NSS statistics each from the 4 groups are demonstrated on the top right of Figure 3 for both the single subspace and mixture subspace models. In addition, the summary of these statistics are shown as boxplot at the bottom of Figure 3, with the single subspace on the left and the mixture subspace on the right. From these figures, we know

that the separation between the plume and other regions greatly improves when using the mixture subspace model.

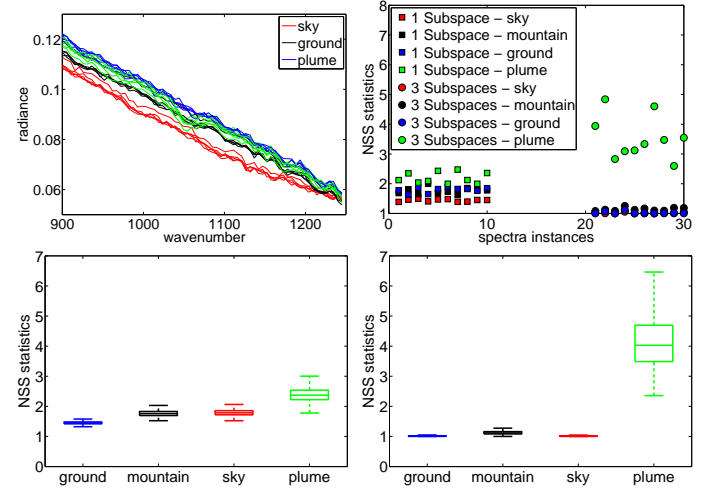


Fig. 3: Top left: radiance against wavenumber of 5 samples each from the groups of sky, mountain, ground and plume. Top right: the NSS scores of 10 samples computed by 1 subspace and 3 subspaces respectively. Bottom: the boxplot of the NSS scores of using 1 Gaussian (left) and 3 Gaussians (right).

3) Partial least square regression

Description. We generate 10,000 spectra from $\mu + be$, where $\mu = (\mu_1 + \mu_2 + \mu_3)/3$, $b = \max \mu$ and $e_j \sim \text{Poisson}(0.005)$, i.i.d., $j = 1, \dots, 68$. Then 1,000 spectra are generated from $\mu + be + gs$ with $g \sim \mathcal{N}(-0.01, 0.001)$. 10 sample spectra of each class (scene and plume) are shown in Figure 4.

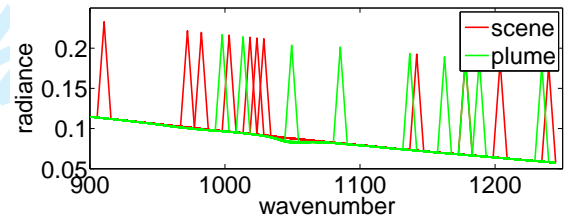


Fig. 4: Background and plume spectra contaminated by Poisson noise.

Task. Detect the plume and compare the performances before and after PLSR.

Technique. We compute the ACE statistics as described in Section II-A. Then we apply Algorithm 2 with $\tau_2 = \tau_3 = 0.2$ to the computed ACE scores.

Results. The top of Figure 5 displays the 50 sample ACE scores on the left and the corresponding enhanced results after applying PLSR on the right. Their statistical summary are demonstrated as boxplots at the bottom of Figure 5 with ACE on the left and PLSR on the right. These figures demonstrate that the PLSR procedure greatly improves the separation of the detection scores between scene and plume.

B. MIT Lincoln Lab Challenge Data

Description. We use hyperspectral images available at the repository at Colorado State, collected by the MIT Lincoln Lab, and made available through the ATD program. This data set consists of four individual hyperspectral images, two of which are for released chemical plume and two for embedded plume. In each of the four, available measurements include a radiance

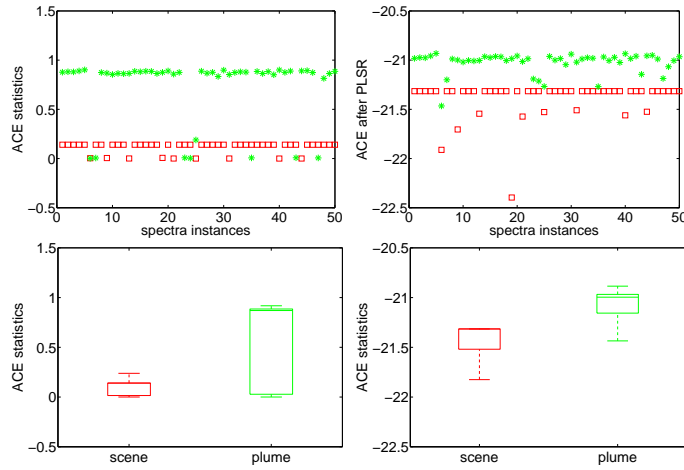


Fig. 5: Top: 10 sample spectra of each class (scene and plume); Middle left: ACE statistics of 50 sample spectra of each class; Middle right: ACE statistics after PLSR of the same 50 sample spectra of each class; Bottom left: the boxplot of the ACE statistics; Bottom right: the boxplot of the ACE statistics with PLSR.

data cube, its matrix form, the absorption coefficient spectrum of the chemical plume of interest and plume present mask. The plume needs to be detected based on a single data cube. Data cubes are about of the size $200 \times 300 \times 100$, where 200×300 is the spatial size and 100 is the spectral dimension.

Task. Detect the chemical plume.

Technique. We apply the three detection algorithms (NMF, NSS, LC), their corresponding mixture models (mixNMF, mixNSS, mixLC), and the two enhancement techniques (resampling and PLSR) as described in Section II and III. The dimension for the subspaces is chosen to be 2 by looking at where the singular values start to flatten. The parameters for the enhancement techniques are $\tau_1 = 0.2$, $\tau_2 = \tau_3 = 0.15$.

Results. The receiver operating characteristic (ROC) curves and the detection maps are shown in Figure 6 for one of the embedded data cube (also the most difficult one). Comparisons are made between single and mixture models on the left and between mixture models and those followed by resampling and then PLSR (the combination of these two enhancement techniques are denoted as *-eh*). From the left figure, we see that using mixture models for background can improve the detection results. Among the three algorithms, mixture LC works the best. In the right figure, when adding the enhancement procedures, both mixNMF and mixNSS improve, but mixLC does not. Overall, mixNMF with enhancement outperforms others.

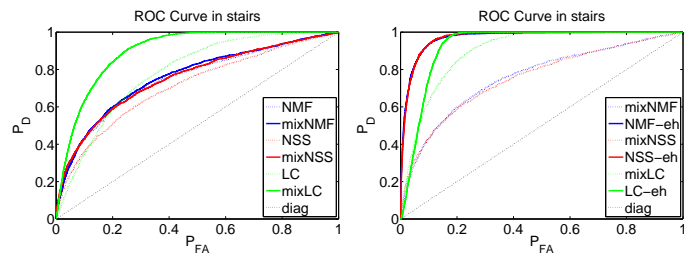


Fig. 6: ROC curves and detection maps. The left is comparison for single and mixture models and the right for mixture models before and after enhancement (denoted as *-eh*).

Furthermore, we demonstrate the detection maps of the intermediate steps of the best performed method mixNMF-

eh, namely mixNMF, mixNMF with resampling (mixNMF-rs), mixNMF with resampling twice (mixNMF-rs2), mixNMF with resampling twice and followed by PLSR (mixNMF-rs2-plsr). These results are shown with those of the NMF on the original data (NMF-outliers) and on the data with outliers removed (NMF). All other methods stated in this section precede the outlier removal first, except NMF-outliers. Figure 7 shows that the mixture model and the various enhancement steps (resampling and PLSR) improve the results gradually; the plume region becomes more and more separable over these steps.

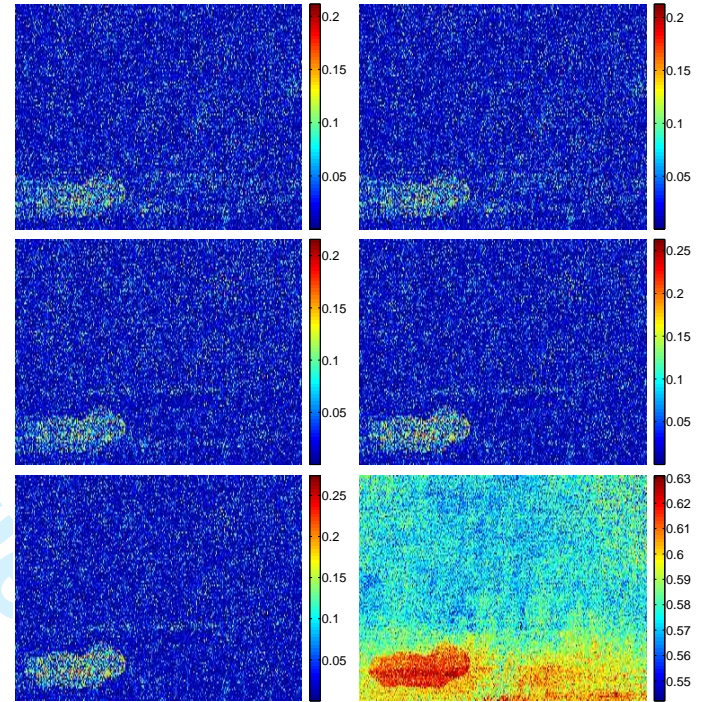


Fig. 7: From top to bottom and from left to right: the detections map of NMF-outlier, MNF, mixNMF, mixNMF-rs, mixNMF-rs2 and mixNMF-rs2-plsr.

C. Fabry-Perot Interferometer Sensor Data

Description. This dataset consists of five time series of radiance data, collected using an imaging Spectroradiometer that operates in the 8 – 11 micron range under five different combinations of the kind of chemical material (TEP, or MeS, or GAA), the release amount (75 kg/burst or 150), and the sensor used. To generate each sequence, a predetermined quantity of simulant is released into the troposphere by using explosives (i.e. a burst release). Successive data cubes are collected, beginning before the event and terminating when it is deemed that additional data collected will not augment scientific value. As a result, each hyperspectral time series contains several hundred frames of hyperspectral images, which all have 256×256 pixels along 20 spectral bands.

Task. Detect and track the chemical plume.

Technique. We first checked the time derivatives of the spectra (along the sequence) but did not observe anything useful (this indicates that the data is particularly challenging). We then applied Alg. 3 with the mixLC estimator and compared our results with ACE [2].

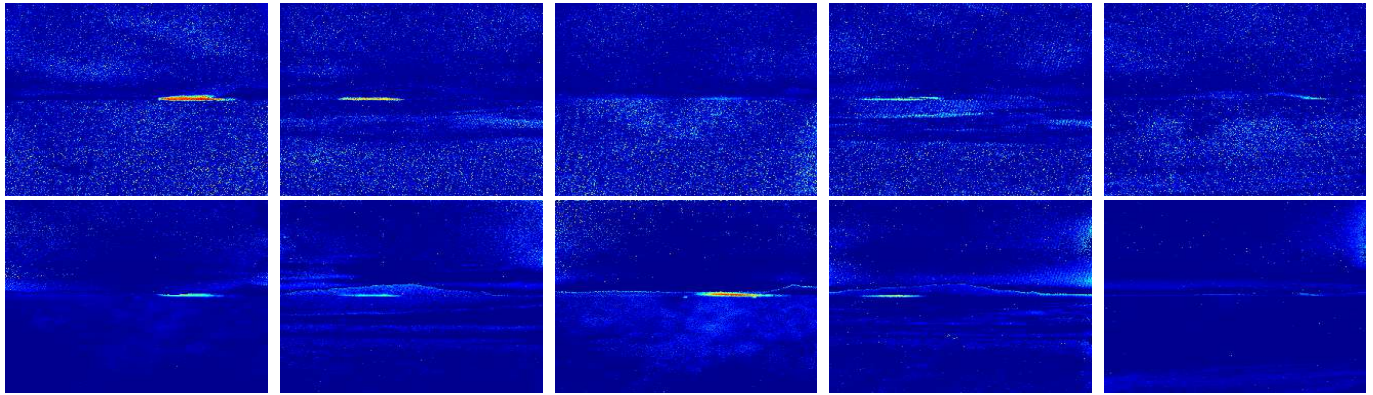


Fig. 8: Detection results by ACE (top row) and by T_{mixLC} (bottom row) on a fixed but arbitrarily selected frame in each of the five hyperspectral movies of the Fabry-Perot data set. Note that (1) our detection map is consistently cleaner than ACE, and (2) our algorithm significantly outperformed ACE in the third movie.

Result. The comparison is shown in Figure 8. It is obvious that the detection map by our algorithm is always much cleaner than that by ACE. Furthermore, in one case, our algorithm was able to clearly locate the plume while ACE could not (see Figure 8, third row). In this experiment, we used the 5th frame for background modeling (via a union of lines) and the 35th frame for testing (for all five movies); future work will utilize multiple clean frames for joint background modeling.

D. Anomaly detection in hyperspectral movies

Description. We use hyperspectral movies available at the repository at Colorado State, collected by Johns Hopkins Applied Physics Lab, and made available through the NSF ATD program. These movies are taken with an FTIR based long wave infrared sensor, recording a variety of releases of known chemicals in a gaseous, liquid and gaseous state. They have frames consisting of $128 \times 320 \times 120$ hyperspectral cubes, with one collected approximately every 8 seconds. The data consists of a desert scene where an unknown (to us) chemical is released at an unknown location at an unknown time (after the beginning of the movie), growing into a chemical plume.

Task. Detect and track the chemical plume.

Technique. We detect the spectra in the chemical plume as an anomaly with respect to an empirical model for the background. We use the first two frames of the movies (where we are told that no chemical plume is present) as our training set s , and use the techniques detailed in section IV to construct $\hat{\mu}_s$, and detect anomalies.

Results. An example of anomaly detection is visualized in Figure 9. In the context of these movies, the intrinsic dimension of the data d appears to be very low, typically $d \leq 5$, as measured by Multiscale SVD [14]-[15]. Therefore the number of samples required in order to learn $\hat{\mu}_s$ is expected to be low, even if the high-dimensional space R^p with $p = 120$. In practice, with non-optimized Matlab code, the construction of $\hat{\mu}_s$, using the first frame as a training set (a much smaller set would be more than sufficient) takes about a minute, and the evaluation of $\hat{\mu}_s$ at all the spectra of a new frame takes a few seconds. The anomaly detection requires the choice of a threshold, but not knowing the ground truth we cannot study ROC curves. However the choice of threshold did not seem to affect the results much, especially in the detection of the “core” part of the chemical plume, albeit it did affect the

detection of (presumably less dense) regions of the plume: see Figure 9, where we chose on purpose a particularly conservative threshold.

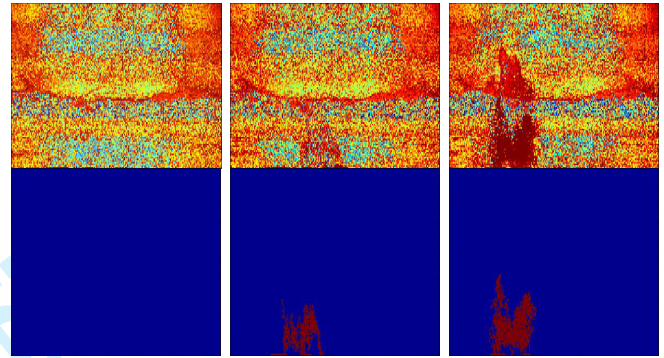


Fig. 9: Anomaly detection of a chemical plume of an unknown chemical appearing at an unknown time and location in a hyperspectral movie. No spatial registration among the frames is assumed. The chemical plume detected as anomaly is seen to evolve in time and space. Top row: log-likelihood score according to empirical GMRA-based model, with darker red meaning lower log-likelihood, and therefore higher probability of being an anomaly. Bottom row: thresholded version of the images on the left, with a non-optimized, conservative threshold, showing detection of the chemical plume. Ground truth was not provided for this data set.

VI. COMPUTATIONAL COMPLEXITY

We discuss the issue of computational complexity. While this is sometimes not of fundamental importance when the data to be analyzed is not large and/or the analysis is to be performed off-line, efficient algorithms are extremely important when data needs to be collected and analyze quickly in real time.

In the mixture models discussed in section III, a segmentation of the spatial pixels is involved. If we use [6] to estimate the subspaces modeling background, then the computational complexity is $O(c_1 mn(dp + c_2 + \log(mn)))$, where d is the intrinsic dimension of the subspaces, c_1 and c_2 are two parameters of the algorithm (typical choices for our case are e.g., 10 and 20 respectively). Algorithm [7] has similar cost and performance. The number of operations for computing the estimators T_{mixNMF} , T_{mixNSS} , T_{mixLC} (also described in section III) is $O(p^2)$, $O((p + d)d^2)$ and, respectively, $O((p + d)d^2)$. As for the enhancement techniques, resampling requires a sequential sorting algorithm, with cost $O(mn \log(mn))$. The complexity of a typical PLSR algorithm requires $O((8pmn +$

$l^2 + 2pl)(l + 1))$ operations, where l is the dimension chosen for PLSR (typically $O(1)$, independently of p).

The cost of the GMRA-based algorithms described in section IV is: $O(C_d mnp \log(mn)d^2)$, where C_d is a constant that depends exponentially in the intrinsic dimension d of the subspaces, for constructing the GMRA; of $O(mnd^2)$ for constructing the estimator $\hat{\mu}_s$ using low-rank Gaussians or KDE, and for evaluating it at new points, but this depends on the density estimator used; $O(mnp \log(mn))$ for computing the coefficients of new data needed to evaluate the likelihood.

The crucial point in the above, is that all the algorithms we discuss run in time proportional to (up to logarithmic factors) the size mnp of the hyperspectral data cube, with constants that depend on the intrinsic dimension d of the data.

VII. CONCLUSION

We have presented several ideas aimed at improving the current state-of-art in several tasks related to the analysis of hyperspectral images, in particular for background modeling, gas plume detection and anomaly detection. We discussed the application of these algorithms to a variety of data sets, with state-of-art or better results (when ground truth was available). The proposed techniques are diverse, but are mostly motivated by the observation that hyperspectral data is often noisy but intrinsically low-dimensional, allowing one to use ideas from the dimension reduction and manifold learning algorithms originally considered in view of machine learning applications. In particular we use techniques for approximating data by mixtures of distributions on low-dimensional subspaces [5], [6], [7], first with a small, fixed number of subspaces for background modeling, and then with more complex, multiscale mixtures of subspaces using GMRA [4] and its extensions to the estimate of probability measures in high-dimensions.

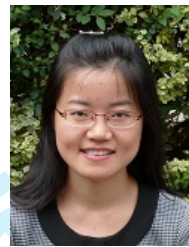
ACKNOWLEDGMENT

The authors thank Dimitris G. Manolakis of Massachusetts Institute of Technology for sharing with them a manuscript [1] and for his correspondence answering several questions.

REFERENCES

- [1] D.G. Manolakis, S.E. Golowich and R.S. DiPietro. Long-Wave Infrared Hyperspectral Remote Sensing of Chemical Clouds: A focus on signal processing approaches. *IEEE Signal Processing Magazine*, vol. 31, no. 4, pp. 120–141, 2014.
- [2] S. Kraut, L. Scharf, and L. T. McWhorter. Adaptive subspace detectors. *IEEE Transactions on Signal Processing*, vol. 49, no. 1, pp. 1–16, 2001.
- [3] D. Manolakis, C. Siracusa and G. Shaw. Hyperspectral subpixel target detection using the linear mixing model. *IEEE Trans. on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1392–1409, 2001.
- [4] W.K. Allard, G. Chen, and M. Maggioni. “Multiscale geometric methods for data sets II: Geometric multiscale analysis”. *Applied and Computational Harmonic Analysis (ACHA)*, vol. 32, no. 3, pp. 435–462, 2012. Available online September 2011.
- [5] G. Chen and G. Lerman. Spectral Curvature Clustering (SCC). *International Journal of Computer Vision*, vol. 81, no. 3, pp. 317–330, 2009.
- [6] T. Zhang, A. Szlam, Y. Wang and G. Lerman. Hybrid Linear Modeling via Local Best-fit Flats. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1927–1934, 2010.
- [7] G. Chen and M. Maggioni. Multiscale Geometric and Spectral Analysis of Plane Arrangements. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2825–2832, 2011.
- [8] G. Chen, M. Iwen, S. Chin and M. Maggioni. A Fast Multiscale Framework for Data in High Dimensions: Measure Estimation, Anomaly Detection, and Compressive Measurements. *Visual Communications and Image Processing (VCIP)*, 2012 IEEE, pp. 1–6, 2012.

- [9] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning: data mining, inference and prediction. *Springer*, pp. 66–68, 2009.
- [10] A. Ihler and M. Mandel, Kernel density estimation toolbox. Available at <http://www.ics.uci.edu/~ihler/code/kde.html>, 2003.
- [11] G. Chen, M. Maggioni and Y.G. Wang “High-dimensional data-modeling techniques for detection of chemical plumes and anomalies in hyperspectral images and movies”. *Proc. NATO-SET conf.*, 2014.
- [12] M. Maggioni. “Geometric Estimation of Probability Measures in High Dimensions”. *IEEE Asilomar Conference on Signals, Systems and Computers*, 2013.
- [13] M. Maggioni, S. Minsker and N. Strawn. “Dictionary Learning and Non-Asymptotic Bounds for the Geometric Multi-Resolution Analysis”. *arXiv*, to appear in *Jour. Mach. Learn. Res.*, 2014.
- [14] A. V. Little, Y.-M. Jung, and M. Maggioni. “Multiscale Estimation of Intrinsic Dimensionality of Data Sets”. *Proc. A.A.A.I.*, 2009.
- [15] A. V. Little. “Estimating the Intrinsic Dimension of High-Dimensional Data Sets: A Multiscale, Geometric Approach”. *PhD Thesis, Duke University*, Apr. 2011.
- [16] A. V. Little, M. Maggioni, and L. Rosasco. “Multiscale Geometric Methods for Data Sets I: Multiscale SVD, Noise and Curvature”. MIT report MIT-CSAIL-TR-2012-029/CBCL-310, <http://dspace.mit.edu/handle/1721.1/72597>, Sep. 2012; to appear in *Appl. Comp. Harm. Anal.*, 2015.
- [17] M. Maggioni. “Multiscale Geometric Measure Estimation”. In preparation, 2015.



Her research interests lie in applied harmonic analysis, machine learning, signal and image processing, as well as applications to real data.

Yi (Grace) Wang received her Diploma in Mathematics from Huazhong University of Science and Technology, Wuhan, China, 2005. From 2006 to 2012, she was with the University of Minnesota, where she received her M.S. in Statistics and Ph.D. in Mathematics. Between 2012 and 2015, she was a postdoctoral researcher at the Statistical and Applied Mathematical Sciences Institute (SAMSI) and a visiting assistant professor at the Mathematics Department, Duke University. In 2015, she joined the Mathematics Department, Syracuse University as an Assistant Professor.



Guangliang Chen received his Bachelor's degree from the University of Science and Technology of China, Hefei in 2003, and a Ph.D. degree in Applied Math from the University of Minnesota Twin Cities in 2009. Afterwards, he held visiting faculty positions at Duke University (2009–2013) and Claremont McKenna College (2013–2014). Since August 2014, he has been with San José State University as an Assistant Professor of Statistics. His research interests include high dimensional data modeling, machine learning, and signal and image processing.



Mauro Maggioni received his Laurea cum laude in pure Mathematics at the Università degli Studi di Milano in 1999, and the Ph.D. in Mathematics from the Washington University, St. Louis, in 2002. He was a Gibbs Assistant Professor in Mathematics at Yale University, till 2006 when he moved to Duke University as an Assistant Professor of Mathematics and Computer Science. He became Professor of Mathematics, Electrical and Computer Engineering and Computer Science at Duke University in 2013. He was awarded the V. Popov Prize in approximation theory in 2007, a Sloan fellowship in 2008, a NSF CAREER award in 2009, and was nominated Fellow of the A.M.S. in 2013. His work focuses on high-dimensional data analysis, probability, approximation theory, harmonic analysis, machine learning, graph theory, signal processing and control theory.