Virtual Dimensionality of Hyperspectral Data: Use of Multiple Hypothesis Testing for Controlling Type-I Error

Vijayashekhar S S, *Student Member, IEEE*, Jignesh S. Bhatt, and Bhargab Chattopadhyay

***Abstract*—Estimating the number of materials present in a scene is the fundamental step in many hyperspectral remote sensing applications. The virtual dimensionality (VD) estimates the number of spectrally distinct materials in the hyperspectral data. The VD is generally considered as the number of signal sources under binary hypothesis, based on the Neyman–Pearson detection criteria. We observe that the hypothesis testing procedure used in many approaches is prone to inflated Type-I (false positive) error. This is due to carrying out the binary hypothesis test individually on each band image, i.e., more than 200 images in hyperspectral data. In this article, we propose multiple hypothesis testing to control the expected proportion of falsely rejected null hypotheses, i.e., false discovery rate (FDR), and in turn, improve the probability of better performance in estimating the VD. To this end, we employ Benjamini and Hochberg procedure that controls the FDR. We provide multiple hypothesis testing-based algorithms to estimate VD wherein the hypothesis can be formulated according to eigenanalysis, the target specified by statistical approach, and by geometric analysis. The efficacies of the proposed algorithms are evaluated by estimating the number of endmembers for the spectral unmixing application. We conduct experiments on four synthetic hyperspectral data sets at different noise levels as well as on two well-known real hyperspectral datasets. Time complexity and execution time are discussed to study the algorithmic aspects while sensitivity analyses of parameters are carried out for better performance analysis of the proposed approach. We found that the use of multiple hypothesis testing improves estimation of number of endmembers in hyperspectral data.**

***Index Terms*—False discovery rate (FDR), hyperspectral remote sensing, multiple hypothesis testing, Type-I error, virtual dimensionality (VD).**

I. Introduction

HYPERSPECTRAL imaging is an unprecedented remote sensing technology with an ambitious goal to develop/update the spectral library of materials on the Earth [1]. Hyperspectral sensors can vary from the visible (400 nm) to near-infrared (NIR) (780 nm) up to the shortwave-infrared bands (2500 nm) and provide a set of coregistered images (>200). Each image captures the reflection from a narrow contiguous (10 nm) wavelength range (spectral resolution). The set of high spectral resolution images forms a three-dimensional (3-D) hyperspectral data cube. Hence, all pixel vectors in the data cube contain contiguous spectra and can be used to identify/quantify materials in the scene with high precision. The materials present in the scene have their unique signature values (across the bands) called *endmembers* and corresponding fractional *abundances* that indicate the proportion of each endmember at every locations in the scene [2], [3].

Virtual dimensionality (VD) is defined as the number of spectrally distinct signatures present in the hyperspectral imagery [4]. The VD can be illustrated using the pigeon-hole principle, i.e., signal source and a spectral band are represented by pigeon and pigeon hole, respectively [5]. It is found that a far smaller number of signal sources are (generally) present than the number of spectral bands in a hyperspectral data cube. For instance, a portion of 224-band Cuprite mining site (USA) data has estimated to have upto 30 distinct materials [4], [6]–[9], and Urban data (USA) has four to five materials [10]–[12] spread over 210 bands. Further, many unknown materials may be uncovered with the help of finer spectral resolution of the data. On the other hand, it is often difficult to acquire the ground truth, especially in many physically inaccessible sites. Hence, estimating the VD from hyperspectral data is a challenging task.

We observe that majority of state-of-the-art methods for estimating VD rely on binary hypothesis testing procedure on every band images. However, the binary hypothesis testing procedure inherently inflates overall Type-I error rate [13]. Recently, we employed a multiple hypothesis (MH) testing procedure to reduce the expected proportion of falsely rejected null hypotheses, i.e., false discovery rate (FDR), while estimating the VD in hyperspectral data [14]. In this article, we extend our work [14] and show its effectiveness over three broad categories of approaches, i.e., eigenanalysis-based, target specified hypothesis-based testing approach, and geometry-based approaches. To this end, we choose representative state-of-the-art approaches in each of the category, and propose our MH testing algorithms for the each case. In particular, HFC [15] and NWHFC [4] from eigenanalysis based, ATGP-NPD [16] from target specified hypothesis testing approach based, and GENE-CH [6] and GENE-AH [6] from geometry-based approaches are chosen. We provide the motivation to use MH testing approach while

Manuscript received February 7, 2020; revised April 6, 2020; accepted April 23, 2020. Date of publication May 28, 2020; date of current version June 16, 2020. This work was supported in part by the Indian Institute of Information Technology Vadodara, and in part by the Indian Institute of Management Vishakapatnam. *(Corresponding author: Jignesh S. Bhatt.)*

Vijayashekhar S S and Jignesh S. Bhatt are with the Indian Institute of Information Technology Vadodara, Government Engineering College, Gandhinagar 382028, India (e-mail: 201671003@iiitvadodara.ac.in; jignesh.bhatt@iiitvadodara.ac.in).

Bhargab Chattopadhyay is with the Department of Decision Sciences, Indian Institute of Management Vishakapatnam, Visakhapatnam 530003, India (e-mail: bhargab@iimv.ac.in).

Digital Object Identifier 10.1109/JSTARS.2020.2991170

estimating the VD as well as discuss how the proposed MH testing algorithms reduce the overall Type-I error. Considering spectral unmixing application, we have validated our algorithms on four different synthetic hyperspectral datasets as well as on the two well-known real hyperspectral datasets, i.e., Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Cuprite, and HYperspectral Digital Imagery Collection Experiment (HYDICE) Urban.

The main contribution of this article is the needed statistical correction in the state-of-the-art approaches in order to control the FDR and in turn avoid false estimation of VD in the hyperspectral data. To the best of our knowledge, this is the first attempt in the remote sensing community for controlling the FDR while estimating VD. It is much needed since many applications in hyperspectral data rely on correctness of this dimensionality for further analysis as well as generating various remote sensing data products, say generating material (abundance) maps using the spectral unmixing. To this end, we propose MH testing that effectively avoids Type-1 error and restricts the expected proportion of falsely rejected null hypotheses to a level of significance.

The rest of the article is organized as follows. In Section II, we begin by providing a brief literature review on virtual dimensionality. The problem formulation, motivation, and the three proposed algorithms are then discussed in Section III. Section IV validates our approaches by conducting experiments on synthetic and real hyperspectral datasets. We discuss time complexity (Big-O), execution time, and parameter sensitivity analysis to better assess the performance of the proposed approach. Finally, Section V concludes this article.

II. Literature Review

Earlier, Malinowski's error theory was popular for estimating the data dimensionality by means of a factor analysis [17]. Malinowski derived an empirical indicator function (EIF) that determines the number of factors in a data matrix and finds a threshold that can separate the first and secondary eigenvalues of data. Recently, a similar idea to the Malinowski's error theory is explored in random matrix theory (RMT) [18] that also uses eigenvalues decomposition to estimate data dimensionality for hyperspectral images. The only difference between EIF and RMT is how they interpret the two sets of eigenvalues, i.e., first and secondary eigenvalue set for EIF [17] while signal and noise eigenvalues for RMT [18]. However, such approaches require a judicious selection of threshold to determine the dimensionality of data.

Binary hypothesis tests have been traditionally used in statistical signal processing and communications to address the problem of signal detection in the presence of noise. The null and alternative hypotheses are typically formed representing signal presence and absence, respectively. Inspired by this, researchers in hyperspectral imaging have devised various algorithms to estimate the VD using binary hypothesis formulation. The VD is typically estimated as signal sources under binary hypotheses characterized according to the following broad aspects, viz., eigenanalysis with eigenvalues [4], [15], eigenvectors [7], the target specified by statistical analysis [19], and by the geometric analysis [6].

Several techniques have been developed to estimate VD based on formulating the binary hypothesis. A very first detection technique called Harsanyi-Farrand-Chang (HFC) [15] is developed by formulating binary hypothesis testing to separate signal eigenvalues from noise eigenvalues as a detection problem, unlike [17], [18]. The HFC method casts the signal/noise decomposition problem as a binary hypothesis testing problem. Neyman–Pearson detector (NPD) is used to determine how many signal sources are present in the data, subject to a false alarm probability. In particular, the HFC method uses the NPD for binary hypothesis testing, built on the differences in eigenvalues of the sample correlation and sample covariance matrices over the bands. Note that the method runs binary hypothesis test independently on each band image. The HFC method has been improved by incorporating the noise prewhitening step and named as noise-whitened HFC (NWHFC) [4]. This follows the same process as in [15] and formulates the null and alternative hypotheses representing the absence and presence of a signal source, respectively, for a particular spectral band. Subsequently, maximal orthogonal complement algorithm (MOCA) [20] made use of eigenvectors/singular vectors, and followed the similar idea, however, using Bayes detector in contrast to HFC and NWHFC that use the NPD. The assumption made for the MOCA is that both null and alternative hypotheses are equally likely, and the costs for making correct and incorrect decisions are uniform.

The use of eigenanalysis for estimation of VD seems to be underestimated since they do not completely reflect the signal sources, i.e., the eigenvalues and eigenvectors are derived from correlation/covariance matrices [8]. Therefore, maximum orthogonal subspace projection (MOSP) [7] has been introduced that estimates the VD by interpreting spectrally distinct signatures with two important differences when compared to MOCA [20]. One is that MOSP considers the VD problem as an NPD problem with binary hypothesis formulation, where estimated VD varies with false alarm probability and not a single value as in the MOCA. The other is that MOSP uses real targets, i.e., signatures generated by automatic target generation process (ATGP) [21], instead of eigenvectors as in MOCA. The idea paves the way for the later development of higher order statistics based VD (HOS-VD) [22]. This uses HOS generated targets by the NPD, and the VD estimation is varied by using false alarm probability.

The VD estimation techniques are integrated under interband spectral information statistics called target specified VD (TSVD) using spectral target statistics of the kth order developed in [16], [19], and [23]. The TSVD [23] uses the HFC [15] and the endmembers generated by suitable endmember extraction algorithm (EEA) [24]. Extreme value theory has been used for defining probability density functions for the hypothesis formulation. The signal sources generated are tested via binary hypothesis to find the number of endmembers.

Meanwhile, geometry-based approaches named estimation of number of endmembers convex hull (GENE-CH) and geometry

based estimation of number of endmembers affine hull (GENEAH) are developed in [6]. These algorithms are based on the linear mixing model (LMM) with both nonnegativity and sum-to-one constraints on the abundances. Therefore, they consider that all the observed pixel vectors present in the convex hull or affine hull of the endmember signatures. Note that the GENE algorithms iteratively estimate the number of endmembers again by using the NPD-based binary hypothesis testing over the endmembers. In this case, their probability distributions under both hypotheses are found by Chi-square distributions with the degree of freedom as maximum number of endmembers.

There are a few LMM fitting error based techniques developed for the VD estimation, and demonstrated for determining the number of endmembers in the hyperspectral data. It includes signal subspace estimate (SSE) [25], which later improved and called hyperspectral signal identification by minimum error (HySime) [26]. An attempt is made to recover a number of endmembers, their signature values as well as abundances using multitemporal hyperspectral data in an iterative optimization framework [27]. The methods optimize the mean-squared error (MSE) resulting from the LMM as the criterion. However, such MSE-based approaches generally require a reliable estimation of the noise covariance matrix.

III. Methods

In this section, we first formulate the problem of MH testing to estimate VD, then clarify the reasons behind Type-I error in binary hypothesis testing procedure, discuss the motivation to employ MH approach, and finally provide details of the proposed three algorithms for estimating VD using the MH testing algorithms.

A. Problem Formulation

The hyperspectral data are considered as a data cube, $x \times y \times L$, where x and y represent spatial dimensions, and L represents the spectral dimension of the remotely acquired scene. Given N such hyperspectral data vectors, our objective is to estimate VD in the scene by employing Benjamini and Hochberg MH testing procedure; within the eigenanalysis based approaches, the target specified hypothesis testing approach, and geometry-based approaches.

B. Motivation

In binary hypothesis (null/ alternative) testing, Type-I error (false positive) occurs when one rejects the null hypothesis when it is true. If one conducts binary hypothesis test multiple times, as done in [4], [6], [8], and [16] for estimation of VD, and follows the same rejection rule independently for each test; then the resulting probability of committing at least one Type-I error is substantially higher than the nominal level. We substantiate more on this below.

In case of hyperspectral data, suppose there are L (number of bands) independent tests which we wish to test simultaneously. Let q is the nominal level for each p-value, where p-value is probability of observing the value of test statistic which is extreme

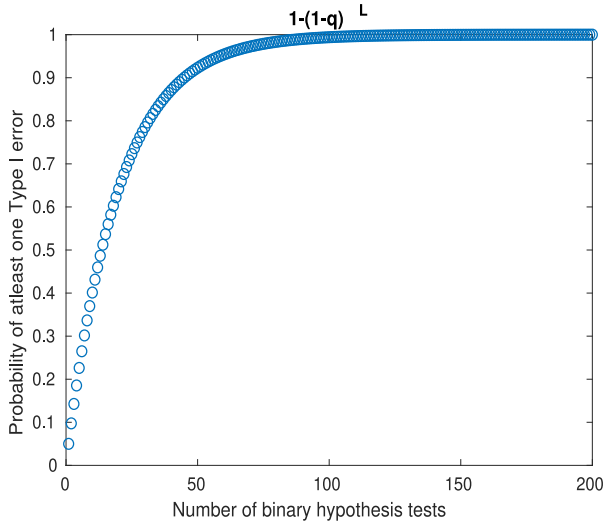


Fig. 1. Illustration of Type-I error: $q = 0.05$; $L = 200$.

TABLE I
Number of Errors Committed When Testing m Null Hypotheses

	Declared non-significant	Declared significant	Total
True null hypotheses	U	V	m_o
Non-true null hypotheses	T	S	m-m_o
	m-B	B	m

or more extreme than observed. The first idea that might come to mind is to test each hypothesis separately, using same level of significance q. If each hypothesis is tested separately, at say q level, then overall Type-I error is scaled to Lq, which is more than q. Note that this may mislead the estimation of VD. Additionally, we know that Probability (at least one Type-I error in L hypotheses) $=$ 1$-$Probability (no Type-I error for all L hypotheses) [28]. Now the probability of making zero Type-I error becomes $(1-q)^{\mathrm{L}}$. Since $0 < q < 1$, it follows that $(1-q)^{\mathrm{L}} < (1-q)$ and so the probability of not making Type-I errors in $L > 1$ tests is much smaller than in the case of a single test. As an example, let us consider the probability of atleast one Type-I error when we perform 200 binary hypothesis tests with $q = 0.05$. See that if the probability of making an error is q, then probability of not making an error is $(1-q)$. Therefore, the probability of not making an error in our example is $(1-0.05)^{200}$. Therefore, the probability of making at least one Type-I error in 200 tests is $1-(1-0.05)^{200} = 0.99996$ referring to Fig. 1. Since the hypotheses are statistically independent with each other even if all the hypotheses are insignificant (worst-case scenario) still the probability of making at least one Type-I error is 0.9999. Since the number of tests for hyperspectral data increases to more than 200, the probability of at least one false positive increases at a rapid rate (as depicted in Fig. 1). In order to address this issue, we propose to use Benjamini and Hochberg method of controlling the false discovery rate (FDR) to estimate VD. The FDR works by estimating some rejection region such that $FDR < q$.

Table I summarizes the number of errors committed when testing m null hypotheses. The specific m hypotheses are assumed to be known in advance. B is an observable random

variable; U, V, S and T are unobservable random variables. We use the equivalent lower case alphabets for their realized values. Consider the problem of testing simultaneously m (null) hypotheses, of which m_o are true. T is the number of rejected hypotheses. The proportion of errors committed by falsely rejecting null hypotheses can be viewed through the random variable $Q = V/(V + S)$, i.e., the proportion of the rejected null hypotheses which are erroneously rejected. Naturally, we define $Q = 0$ when $V + S = 0$, as no error of false rejection can be committed. Q is an unobserved (unknown) random variable, as we do not know v or s, and thus $q = v/(v + s)$, even after experimentation and data analysis. We define the FDR to be the expectation of Q

$$E(Q) = E\{(V/(V + S)\} = P(B > 0)E(V/B|B > 0) \quad (1)$$

where $E(\cdot)$ is the expectation operator.

We now describe how the Type-I error is incurred in the three broad category of approaches to estimate the VD. The HFC [15] and NWHFC [4] are state-of-the-art methods categorized under the eigenvalues based approaches that estimate VD. The methods assume that the hyperspectral signatures are unknown but deterministic signal sources, and the noise in the data are white with zero mean. Under the assumptions, the signatures of interest (giving rise to VD) are those that contribute to the first-order statistics, i.e., sample mean of data. The methods formulate binary hypothesis test using eigenvalues of correlation and covariance matrices of the data. Since the test applies on each band image, it is sensitive to false positives (Type-I) as illustrated in Fig. 1, and that may leads to false identification of VD.

The concept of using real target pixels to estimate the value of VD is first explored in [23] and further investigated in [29] to extend the HFC method to HOS-based HFC [22] methods. Such an approach is useful to unify various VD estimation techniques under the same problem setting and formulation. ATGP [21] generated real targets or signal sources are used, in place of eigenvalues used in HFC [15] and NWHFC [4] or eigenvectors used in MOSP [7], and the method is named as ATGP-HFC [16]. The details of ATGP are found in [21] that finds a set of targets by performing successive orthogonal subspace projections. In such approaches also, Type-I error increases (see Fig. 1) due to independent run of binary hypothesis testing to check for intended signal source in each band.

The GENE-CH [6] and GENE-AH [6] algorithms are devised based on the data geometry that all the observed pixel vectors should lie in the convex hull (CH) and affine hull (AH) of the endmember signatures, respectively. The algorithms exploit the successive estimation property of a pure-pixel-based endmember estimation algorithm until the estimate of the number of endmembers is obtained. In the noisy scenario, the decision of whether the current endmember estimate is in the CH/AH of the previously found endmembers is formulated as a binary hypothesis testing problem, which is solved using the NPD theory. The number of binary hypothesis tests in the estimation of VD for hyperspectral data is large, i.e., typically more than 200. Hence, as depicted in Fig. 1, these data geometry based algorithms are also prone to inflated Type-I error.

We have seen that the three broad state-of-the-art categories of approaches involve larger binary hypothesis tests (> 200) and use the NPD theory. The number of hypotheses tests is very large, and prone to inflated Type-I error. Hence, in the proposed article, our motivation is to control the FDR. The FDR is defined as an expected proportion of Type-I errors. A Type-I error is where we incorrectly reject the null hypothesis; in other words, we get a false positive. We use the Benjamini and Hochberg approach, which controls the FDR by sequentially comparing the observed p-value for each of a family of multiple test statistics, in order from largest to smallest, to a list of computed p-values. The procedure controls the FDR at a certain level, say q, for any configuration of false null hypotheses, assuming independent test statistics.

C. Proposed Approaches

In this section, we propose three algorithms based on the MH testing procedure to control the FDR and improve the estimation of VD in the hyperspectral data. The proposed approach improves the estimation of VD for eigenanalysis, for target specified hypothesis testing, and for geometry-based approaches.

1) MH Testing for Eigenanalysis Based Approaches: In this category of approaches [4], [15], the VD is estimated as the number of alternate hypotheses under the following binary hypothesis formulation:

$$H_{0l} : z_l = \lambda_l{}' - \lambda_l = 0 \text{ vs } H_{1l} : z_l = \lambda_l{}' - \lambda_l > 0 \quad (2)$$

for $l = 1, 2, \ldots, L$. z_l is considered as the observed value for the lth hypothesis, $(\lambda_1{}' \geq \lambda_2{}' \geq, \ldots, \geq \lambda'_L)$ and $(\lambda_1 \geq \lambda_2 \geq , \ldots, \geq \lambda_L)$, respectively, be the eigenvalues of interband sample correlation matrix $\mathbf{R}$

$$\mathbf{R} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbf{r_i} \mathbf{r_i}^{\mathbf{T}} \quad (3)$$

and interband sample covariance matrix $\mathbf{K}$

$$\mathbf{K} = \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathbf{r_i} - \boldsymbol{\mu})(\mathbf{r_i} - \boldsymbol{\mu})^{\mathbf{T}} \quad (4)$$

for given data vectors $\mathbf{r}$ with L bands, and $\boldsymbol{\mu}$ is the mean vector of hyperspectral data cube. In (1), when H_1 is true, i.e., H_0 fails, it implies that there is signal energy contributing to the correlation eigenvalue, in addition to the noise. In HFC [15] and NWHFC [4], the number of times Neyman–Pearson test fails gives an estimation of VD for various false alarm probabilities (P_F). The binary hypothesis test as shown in (2) runs independently on each band and, hence, sensitive to the Type-I error since the test applies L times where $L > 200$ for estimating the VD in the case of hyperspectral data. Recall that FDR is defined as the expected proportion of false rejections. The false rejection in this problem is identifying the number of signal sources, which is not present in the hyperspectral data. Hence, it can lead to false identification of the signal sources.

To address this issue, we propose MH testing approach within the HFC [15] and the NWHFC [4] frameworks for estimating VD. It controls the FDR and restricts the Type-I error at a given

Pseudocode 1: Proposed Multiple Hypothesis Testing in Eigenanalysis-based Approaches.

Input Hyperspectral data $\boldsymbol{r}$,

Step 1 Calculate the eigen values of $\mathbf{R}$ and $\mathbf{K}$ as $(\lambda_1{}' \geq \lambda_2{}' \geq, \ldots, \geq \lambda_L')$ and $(\lambda_1 \geq \lambda_2 \geq, \ldots, \geq \lambda_L)$, respectively, where $\mathbf{R}$ is an interband correlation matrix and $\mathbf{K}$ is a sample covariance matrix of the given data.

Step 2 Consider the test statistic for the lth hypothesis as $z_l = \lambda_l{}' - \lambda_l$, where z_l denotes the observed value.

Step 3 Calculate p-value for the lth spectral band as,

$$p_l = \frac{1}{2}\left[1 - erf\left(\frac{z_l}{\sqrt{2}\sigma_l}\right)\right], \text{ for } l = 1, 2, \ldots, L, \quad (5)$$

where σ_l is the standard deviation in the lth spectral band and $erf(\cdot)$ is an error function.

Step 4 Test L independent null hypotheses, $H_{01}, H_{02}, \ldots, H_{0L}$, with corresponding p-values $p_1, p_2, \ldots, p_{\mathrm{L}}$.

Step 5 Order the p-values $p_1 \leq p_2 \leq, \ldots, \leq p_L$, i.e., arrange the p-values starting from the most likely hypothesis.

Step 6 To control FDR (Type-I error at level q [28]), reject all null hypothesis, i.e., reject $H_{01}, H_{02}, \ldots, H_{0k}$, where k is calculated as

$$k = \max(l : p_l \leq (l/L)q), \text{ for } l = 1, 2, \ldots, L. \quad (6)$$

Output The retained, i.e., not rejected, number of hypotheses is the proposed VD.

level of q, unlike Lq as in [4] and [15]. The pseudocode for the proposed MH testing in eigenanalysis-based approaches is listed in pseudocode 1.

In step 2 of pseudocode 1, under null hypothesis, the test statistic z_l is assumed to be normally distributed with zero mean and $\sqrt{2}\sigma_l$ standard deviation for respective band. L p-values are calculated by integrating the respective the lth distribution from the calculated threshold to infinity in step 3 of pseudocode 1. Note that we assume normal distribution of the test statistic under null hypotheses while calculating p-values. For L hypotheses, we sort and then arrange the p-values starting from the most likely hypothesis in step 5. For a specified level q, suppose i is the largest number for which $p_i \leq (i/L)q$, for $i = 1, 2, \ldots, L$ as in step 6. Then, we reject i null hypotheses whose corresponding p-values are given by $p_1, p_2, \ldots, p_i$. Hence, the statistical procedure employed restricts the expected proportion of falsely rejected null hypotheses to chosen level of significance (q). For more details one may refer to [13], [28].

2) MH Testing for Target Specified Hypothesis Testing Approaches: Binary hypothesis test in target specified hypothesis testing approaches, such as in [16], is formulated as

$$H_{0l} : \boldsymbol{\eta_l} \sim p(\boldsymbol{\eta_l}|H_0) \text{ vs } H_{01} : \boldsymbol{\eta_l} \sim p(\boldsymbol{\eta_l}|H_1) \quad (7)$$

for $l = 1, 2, \ldots, L$. The null hypothesis H_0 represents the maximum residual from the background signal sources, and the alternative hypothesis H_1 represents the maximum residual from the target signal sources. Here, $\boldsymbol{\eta_l}$ is the maximum of complement subspace projections belonging to the two classes, one for target signal class $\mathbf{I_T(l)}$, and the other is background class $\mathbf{I_B(l)}$

$$\boldsymbol{\eta_l} = \max\left\{\max_{i \in \mathbf{I_T(l)}} \|\boldsymbol{P}_{\boldsymbol{S_l}}^{\perp}\boldsymbol{r_i}\|_2^2, \max_{i \in \mathbf{I_B(l)}} \|\boldsymbol{P}_{\boldsymbol{S_l}}^{\perp}\boldsymbol{r_i}\|_2^2\right\} \quad (8)$$

where $\boldsymbol{P}_{\boldsymbol{S_l}}^{\perp}\boldsymbol{r_i} = (\boldsymbol{S_l^T S_l})^{-1}\boldsymbol{S_l^T r_i}$. More specifically for each l, define

$$\boldsymbol{t_l^{SVD}} = \arg\max_{\boldsymbol{r}} \|\boldsymbol{P}_{\boldsymbol{S_l}}^{\perp}\boldsymbol{r_i}\|_2^2 \quad (9)$$

where $\boldsymbol{t_l}^{\mathbf{SVD}}$ is the eigenvector after singular value decomposition (SVD) on data and $\boldsymbol{P}_{\boldsymbol{S_l}}^{\perp}$ maps the observed pixel vector $\boldsymbol{r}$ into the orthogonal complement of $\boldsymbol{S_l}$ denoted by $\boldsymbol{S_l^{\perp}}$. $\boldsymbol{S_l}$ is the space linearly spanned by signal vectors found by the SVD. The orthogonal complement subspace projections of data sample vectors $\boldsymbol{P}_{\boldsymbol{S_l}}^{\perp}\boldsymbol{r_i}$ under H_0 are the noise sample vectors. Hence, it is reasonable for MOCA [20] to assume that the vector $\boldsymbol{P}_{\boldsymbol{S_l}}^{\perp}\boldsymbol{r_i}$ under H_0 behaves as independent and identically distributed (IID) Gaussian random variables. Moreover, $\boldsymbol{\eta_l}$ is the maximum residuals from orthogonal projection obtained in $\boldsymbol{S_l^{\perp}}$ under H_0. Using extreme value theory [30], $\boldsymbol{\eta_l}$ under H_0 is modeled as a Gumbel distribution, i.e., $F_{v_l}(\boldsymbol{\eta_l})$ is the cumulative distribution function (cdf) of v_l as in [7]. Now, if $\boldsymbol{t_l}^{\mathbf{SVD}}$ in (9) is replaced with $\boldsymbol{t_l}^{\mathrm{ATGP}}$ for each band, where $\boldsymbol{t_l}^{\mathrm{ATGP}}$ is a spectral signature generated by ATGP, then problem turns to be target specified approach and a new hypothesis testing problem is formulated as follows :

$$H_{0l} : \boldsymbol{\eta_l} \sim p_0(\boldsymbol{\eta_l}) \text{ versus } H_{1l} : \boldsymbol{\eta_l} \sim p_1(\boldsymbol{\eta_l}) \quad (10)$$

where $\boldsymbol{\eta_l} = \|\boldsymbol{t_l}^{\mathrm{ATGP}}\|_2^2$ for $l = 1, 2, \ldots, L$. The signal sources under the hypotheses are now real target signal sources rather than eigenvectors in (9) or eigenvalues in (2). Hence, ATGP-specified VD is now defined as

$$\arg\min_{1 \leq l \leq L} F_{v_l}(\boldsymbol{\eta_l}) \leq 1 - P_F = VD^{\mathrm{ATGP}}(P_F). \quad (11)$$

The signal source $\boldsymbol{\eta_l} = \|\boldsymbol{t_l}^{\mathbf{ATGP}}\|_2^2$ is a random variable representing the signal energy of real target signal source $\boldsymbol{t_l^{ATGP}}$. Further if one chooses signal strength rather than signal energy then the problem can be reformulated as the signal-strength-based hypothesis testing problem with $\boldsymbol{\eta_l} = \|\boldsymbol{t_l}^{\mathbf{ATGP}}\|_2^2$ replaced by $\sqrt{\boldsymbol{\eta_l}} = \|\boldsymbol{t_l}^{\mathbf{ATGP}}\|$ given by

$$H_{0l} : \sqrt{\boldsymbol{\eta_l}} \sim p_0(\sqrt{\boldsymbol{\eta_l}}) \text{ versus } H_{1l} : \sqrt{\boldsymbol{\eta_l}} \sim p_1(\sqrt{\boldsymbol{\eta_l}}) \quad (12)$$

for $l = 1, 2, \ldots, L$. ATGP-specified VD is now defined as

$$\arg\min_{1 \leq l \leq L} F_{v_l}(\sqrt{\boldsymbol{\eta_l}}) \leq 1 - P_F = VD^{\sqrt{\mathrm{ATGP}}}(P_F). \quad (13)$$

In both the cases of ATGP-specified VD, i.e., (10) and (12), Type-I error is inflated due to independent run of binary hypothesis testing on each l and give rise to total of Lq Type-I error. We propose to estimate the VD using MH testing to control the FDR and bring it down to q. The pseudocode for the proposed MH testing in such target specified approaches is listed in pseudocode 2.

For L hypotheses, we sort the p-values calculated in step 4 from each of the hypotheses as $p_1 \leq p_2 \leq, \ldots, \leq p_L$. For a

Pseudocode 2: Proposed MH Testing Algorithm in Target Specified Approaches.

Input Hyperspectral data $\mathbf{r}$.
Step 2 Compute ATGP generated signal sources ${t_l}^{\mathrm{ATGP}}$ for $l = 1, 2, \ldots, L$.
Step 3 Consider the test statistic for the lth hypothesis as $\eta_l = \|{t_l}^{\mathrm{ATGP}}\|^2$, where η_l is the observed value.
Step 4 p-value for the lth observed value is calculated from the cumulative distribution function (cdf) of Gumbel distribution [7] as follows,

$$p_l = 1 - F_{v_l}(\eta_l), \text{ for } l = 1, 2, \ldots, L. \tag{14}$$

Step 5 Test L independent null hypothesis, $H_{01}, H_{02}, \ldots, H_{0L}$, with corresponding p-values $p_1, p_2, \ldots, p_{\mathrm{L}}$.
Step 6 Order the p-values $p_1 \leq p_2 \leq, \ldots, \leq p_L$, i.e., arrange the p-values starting from the most likely hypothesis.
Step 7 To control FDR at level q, reject all null hypothesis, i.e., reject $H_{01}, H_{02}, \ldots, H_{0\mathrm{k}}$, where k is calculated as

$$\mathrm{k} = \max(l : p_l \leq (l/L)q), \text{ for } l = 1, 2, \ldots, L. \tag{15}$$

Output The retained i.e., not rejected number of hypotheses is the proposed VD.

specified level q, i is the largest number for which $p_i \leq (i/L)q$. Then, we reject i null hypotheses whose corresponding p-values are given by $p_1, p_2, \ldots, p_i$ as in step 7. For independent test statistics and for configuration of false null hypothesis, pseudocode 2 controls the FDR at chosen level of significance q. Thus, Type-I error is effectively avoided and in turn the accuracy of VD estimation is improved.

3) MH Testing for Geometry-Based Approaches: Under LMM, the geometry-based approaches such as [6] assume that the data are corrupted with IID zero-mean Gaussian noise. Estimation of the number of endmembers (e) is carried out from the given N hyperspectral data vectors. The binary decision rule for geometric approaches is to decide H_0 if $\psi > P_F$ and decide H_1 if $\psi < P_F$, where

$$H_{0j}(\tilde{\boldsymbol{r}}[l_j] \in \mathrm{conv}(\tilde{\boldsymbol{r}}[l_1], \ldots, \tilde{\boldsymbol{r}}[l_{j-1}]) : o_j \sim f_{\chi^2}(r, e_{\max} - 1) \tag{16}$$

versus

$$H_{1j}(\tilde{\boldsymbol{r}}[l_j] \notin \mathrm{conv}(\tilde{\boldsymbol{r}}[l_1], \ldots, \tilde{\boldsymbol{r}}[l_{j-1}]) : o_j \sim f_{N\chi^2}(r, e_{\max} - 1, \boldsymbol{\mu_j}) \tag{17}$$

where $\tilde{\boldsymbol{r}}$ is the dimensionally reduced vector of data vector $\boldsymbol{r}$, $f_{\chi^2}(\cdot)$ is the probability density function of central Chi-square distribution, $f_{N\chi^2}(\cdot)$ is the noncentral Chi-square probability density function, ψ is the cumulative distribution function of central Chi-square distribution with mean vector $\boldsymbol{\mu_j}$, and $\mathrm{conv}(\cdot)$ is the convex hull of a set of vectors. Given the hyperspectral data $\boldsymbol{r}$, maximum number of endmembers ($e_{\max}$) and estimate of noise covariance matrix $\widehat{\boldsymbol{D}}$, it obtains an affine set fitting parameter $(\boldsymbol{C}, \boldsymbol{d}) \in \mathbb{R}^{L \times (e_{\max}-1)} \times \mathbb{R}^L$ as in [6]. Here $e - 1$ is the affine dimension of $\mathrm{aff}(\boldsymbol{m_1}, \ldots, \boldsymbol{m_e})$, where $\mathrm{aff}(\cdot)$ is affine hull of a set of endmembers $(\boldsymbol{m_1}, \ldots, \boldsymbol{m_e})$. The first pixel index l_1 is computed by the successive EEA and obtain the jth pixel index l_j using the successive EEA and compute dimensionally reduced vector for it. Keep accumulating dimensionally reduced vectors for pixel indices from 1 to $(j-1)$ and form matrix $\widehat{\boldsymbol{A}}_{\boldsymbol{j-1}}$. The constrained least squares $\boldsymbol{\theta}^*$ is then solved for GENE-CH and GENE-AH as in [6] and calculate $\boldsymbol{\epsilon_j} = \tilde{\mathbf{r}}[\mathbf{l_j}] - \widehat{\boldsymbol{A}}_{\boldsymbol{j-1}}\boldsymbol{\theta}^*$. Now, we can compute $o_j = \boldsymbol{\epsilon_j^T}(\xi^* \sum)^{-1}$, where $\xi^* = 1 + \boldsymbol{\theta}^{*T}\boldsymbol{\theta}^*$ and $\sum = \widehat{\boldsymbol{C}}^{\boldsymbol{T}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{C}}$. Calculate cdf ψ using the probability density function of the central Chi-square distribution $f_{\chi^2}(\boldsymbol{r}, e_{\max-1})$. Then by Neyman–Pearson lemma, the optimal threshold κ for the hypothesis testing problem satisfies

$$\psi(\kappa) = P_F \tag{18}$$

where P_F is the preassigned acceptable false alarm rate. The binary testing procedure for the total number of endmembers with varying P_F is sensitive to the Type-I error since the test applies k times for estimating the VD. We propose MH testing approach to control FDR as available in pseudocode 3.

As shown in pseudocode 3, step 7, p-values are calculated using the probability density function of central Chi-square distribution. Then we reject K null hypotheses whose corresponding p-values are given by $p_1, p_2, \ldots, p_K$ as shown in step 10 of the pseudocode 3.

It is clear that use of proposed three algorithms effectively avoids Type-I error. Referring to Table I, the expectation of falsely rejected null hypotheses is controlled to q with the hypothesis correction procedure used. It is more realistic to see the data at hand as part of the evidence that was and will be accumulated about these hypotheses. This is a major reason why presenting the result of the test as a p-value is more helpful to research than presenting only the "significant–nonsignificant" dichotomy.

IV. Results

In this section, we discuss the results obtained for the estimation of number of endmembers using proposed algorithms on different hyperspectral datasets. The signatures of interest in hyperspectral image analysis are materials substances, which generally cannot be identified a *a priori* or by visual inspection such as endmembers, anomalies, or man-made objects. In this article, we demonstrate efficacies of our algorithms to estimate VD for identifying the number of endmembers in hyperspectral unmixing application. Note that since our problem in this article is to improve VD, we show detail results and analysis for estimating number of endmembers in the synthetic as well as real hyperspectral datasets. We refrain performing endmember extraction and abundance estimation similar to many state-of-the-art approaches for VD estimation [6], [8], [25], [26], [31].

We first conduct experiments on four synthesized hyperspectral data cubes using two sets of five spectral signatures of the United States Geological Survey (USGS) digital spectral library [32]. To assess the noise sensitivity of the proposed algorithms, we have added different levels of white Gaussian noises in

Pseudocode 3: Proposed Multiple Hypothesis Testing Algorithm in Geometry-Based Approaches.

Input Hyperspectral data $\boldsymbol{r}$.

Step 1 Initialize with maximum number of endmembers $e \leq e_{\max} \leq L$, and estimate of noise co-variance matrix $\widehat{\boldsymbol{D}}$.

Step 2 Consider the test statistic for the jth hypothesis as ψ, where ψ is the cdf of central Chi-square distribution defined by $f_{\chi^2}(r, e_{max-1})$.

Step 3 Obtain an affine set fitting parameter $(\boldsymbol{C}, \boldsymbol{d}) \in \mathbb{R}^{L\times(e_{\max}-1)} \times \mathbb{R}^{L}$ as in [6], where $(e-1)$ is the affine dimension of $aff(\boldsymbol{m_1}, \ldots, \boldsymbol{m_e})$.

Step 4 Obtain the first pixel index l_1 by the successive EEA and compute $\tilde{\boldsymbol{r}}[l_1]=\widehat{\boldsymbol{C}}^T(\boldsymbol{r}[l_1]-\widehat{\boldsymbol{d}}) \in \mathbb{R}^{e_{\max}-1}$ and set $k=2$. Obtain the kth pixel index $\boldsymbol{l_k}$ using the successive EEA and compute $\tilde{\boldsymbol{r}}[l_k]=\widehat{\boldsymbol{C}}^T(\boldsymbol{r}[l_k]-\widehat{\boldsymbol{d}}) \in \mathbb{R}^{e_{\max}-1}$ and form $\widehat{\boldsymbol{A}}_{\boldsymbol{j-1}} = [\tilde{\boldsymbol{r}}[l_1], \ldots, \tilde{\boldsymbol{r}}[l_{j-1}]] \in \mathbb{R}^{(e_{\max}-1)\times(j-1)}$ for $j = 1, 2, \ldots, k$.

Step 5 The constrained least squares for geometrical approaches is solved as in [6] and named as θ^*.

Step 6 Calculate $\boldsymbol{\epsilon_j} = \tilde{\boldsymbol{r}}[l_j] - \widehat{\boldsymbol{A}}_{\boldsymbol{j-1}}\boldsymbol{\theta^*}$. Compute $o_j = \boldsymbol{\epsilon_j^T}(\xi^* \sum)^{-1}$, where $\xi^* = 1 + \boldsymbol{\theta^{*T}}\boldsymbol{\theta^*}$ and $\sum = \widehat{\boldsymbol{C}}^{\boldsymbol{T}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{C}}$.

Step 7 p-value for j^{th} index is calculated as,

$$p_j = 1 - \psi_j, \text{ for } j = 1, 2, \ldots, k, \tag{19}$$

where ψ_j is calculated as in [6].

Step 8 Test k independent null hypothesis, $H_{01}, H_{02}, \ldots, H_{0k}$, with corresponding p-values $p_1, p_2, \ldots, p_k$

Step 9 Starting from the most likely hypothesis, arrange the p-values in order as $p_1 \leq p_2 \leq, \ldots, \leq p_k$.

Step 10 To control FDR at level q [28], reject all null hypothesis, i.e., reject $H_{01}, H_{02}, \ldots, H_{0K}$, where K is calculated as

$$K = \max(j : p_j \leq (j/k)q), \text{ for } j = 1, 2, \ldots, k. \tag{20}$$

Output The retained, i.e., not rejected, number of hypotheses is the proposed VD.

the datasets and compare the results with state-of-the-art approaches. Moreover, we conduct parameters sensitivity analyses, compute time complexities, and compare average execution times of the algorithms. Next, we show the experiments on two well-known real hyperspectral data. First is collected by AVIRIS over the cuprite mining site area, USA [33]–[35]. The second data are captured by the HYDICE sensor at the location of Copperas Cove, near Fort Hood, Texas, USA, in October 1995 [33]–[35].

A. Experiments on Four Synthetic Hyperspectral Datasets

The LMM is used to construct the hyperspectral data. A L-dimensional hyperspectral data vector $\boldsymbol{r}$ is considered as linear combination of the endmembers, i.e., matrix $\boldsymbol{M}$, as follows:

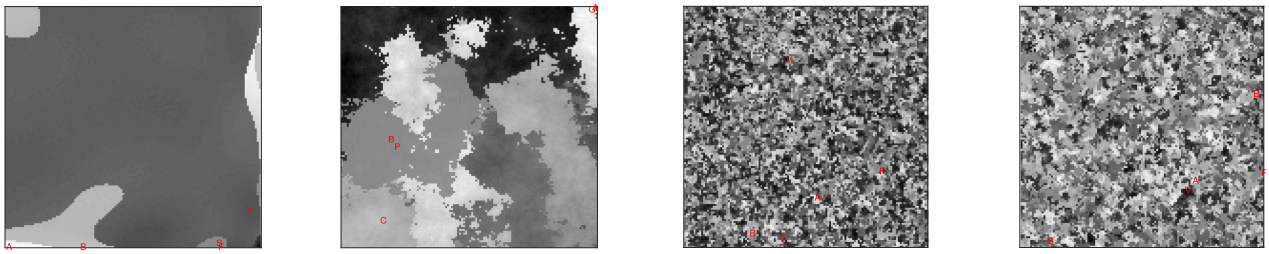
$$\boldsymbol{r} = \boldsymbol{M}\boldsymbol{\alpha} + \mathbf{n} \tag{21}$$

where for e number of endmembers, size of $\boldsymbol{M}$ is $L \times e$, $\boldsymbol{\alpha}$ is abundance vector of size $e \times 1$, and $\boldsymbol{n}$ represents IID Gaussian noise vector of size $L \times 1$.

1) Data Generation: We first conduct experiments on four sets of synthesized hyperspectral data cubes using two different sets of five spectral signatures from the USGS. Therefore, the actual number of endmembers for all four synthetic datasets is five (5) in each. The selected endmembers for constructing first data are Asphalt (gds317), Brick (gds350), Fiberglass (gds374), Sheetmetal (gds352), and Vinylplastic (gds372). They form the $\boldsymbol{M}$ of size 431×5. The Gaussian fields with different parameters have been considered for generating different abundance maps. In particular, nonnegative and sum-to-one abundance vectors of size 5×1 are randomly generated for every location using the Gaussian field. The abundance maps are generated using different Gaussian fields, and hence they contain both smooth regions and edges with significantly high heterogenous regions making estimation challenging. Finally, LMM (21) is used to generate datasets. Using an abundance vector $\boldsymbol{\alpha}$ of size 5×1 and endmember matrix M of size 431×5, a pixel vector $\boldsymbol{r}$ is constructed as $\boldsymbol{r} = \boldsymbol{M}\boldsymbol{\alpha}$, generating a 431×1 reflectance data vector. Considering five different abundance maps each of size 128×128, a 431-band ground truth data cube is generated, i.e., $128 \times 128 \times 431$, and we call it synthetic data 1. Now, synthetic data 2 is constructed (LMM) using another five reflectance spectra Ammonium Chloride, Cyanide Potassium, Green Slime, Brucite, and Kerogen Bic while using another random Gaussian field for abundance patterns. Similarly using the same LMM process, another two data cubes of size $128 \times 128 \times 431$ are constructed and we call them synthetic data 3 and synthetic data 4. The mean images of all synthetic data sets along with marked endmembers are shown in Fig. 2(a)–(d).

2) Parameters and Sensitivity Analysis: All the algorithms are implemented in MATLAB (R2019a) platform and run on a laptop computer with Intel(R) Core(TM) i3-7100 CPU at 2.40 GHz with 4 GB of RAM. IID Gaussian noises have been added to all the synthetic datasets at different proportions of signal to noise ratio (SNR). The detailed sensitivity analysis of parameter P_F for all the state-of-the-art algorithms and significance level q for the proposed methods are carried out. Fig. 3 shows the sample sensitivity analysis of parameter P_F for all the four synthetic datasets at 60 dB SNR using state-of-the-algorithms. Fig. 4 shows the sensitivity analysis of parameter q for all the proposed algorithms at 60 dB SNR for every synthetic dataset. The reliable P_F values are obtained for all state-of-the-art methods, $e_{\max}$ is set to 50, and the true noise covariance matrix is supplied for each simulated dataset. Note that for the proposed approaches, q is set to 0.05 in all the experiments. In all our experiments, we have used optimum values of the various parameters while implementing different algorithms as mentioned in the respective papers.

3) Results Discussion: Table II provides the values of estimated number of endmembers by proposed MH testing based



(a) (b) (c) (d)

Fig. 2. Mean images of synthetic datasets: (a) mean image of synthetic data 1. Endmembers and their respective locations are marked: A-Asphalt, B-Brick, F-Fiberglass, S-Sheetmetal, V-Vinylplastic, and (b) mean image of synthetic data 2. Endmembers marked are A-Ammonium Chloride, C-Cyanide Potassium, G-Green Slime, B-Brucite, P-Kerogin Bic. (c) mean image of synthetic data 3, and (d) mean image of synthetic data 4. Endmembers and their respective locations for synthetic data 3-4 are marked: A-Asphalt, B-Brick, F-Fiberglass, S-Sheetmetal, V-Vinylplastic.

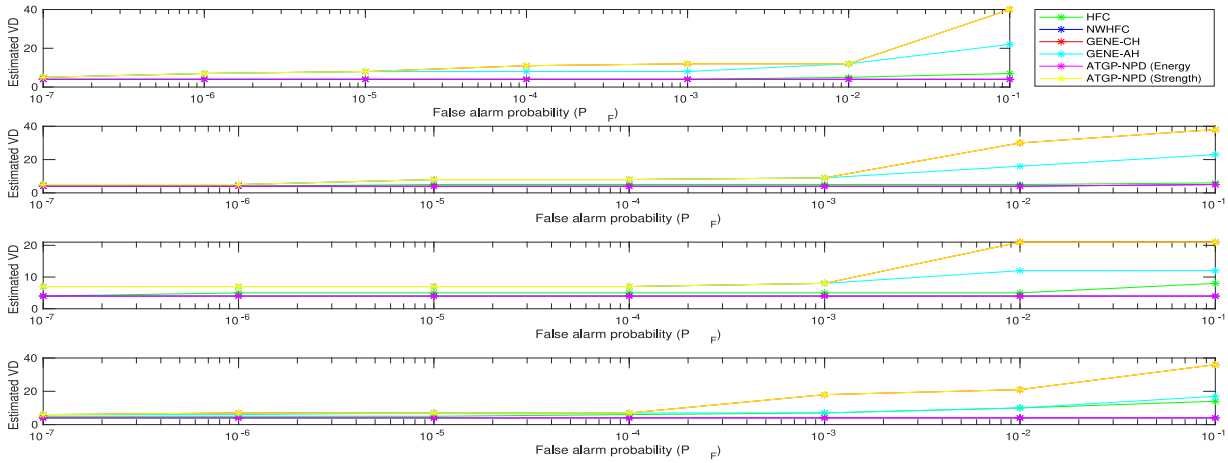


Fig. 3. Parameter sensitivity analysis for false alarm probability (P_F) for the four synthetic datasets 1–4 (top to bottom) at SNR of 60 dB.

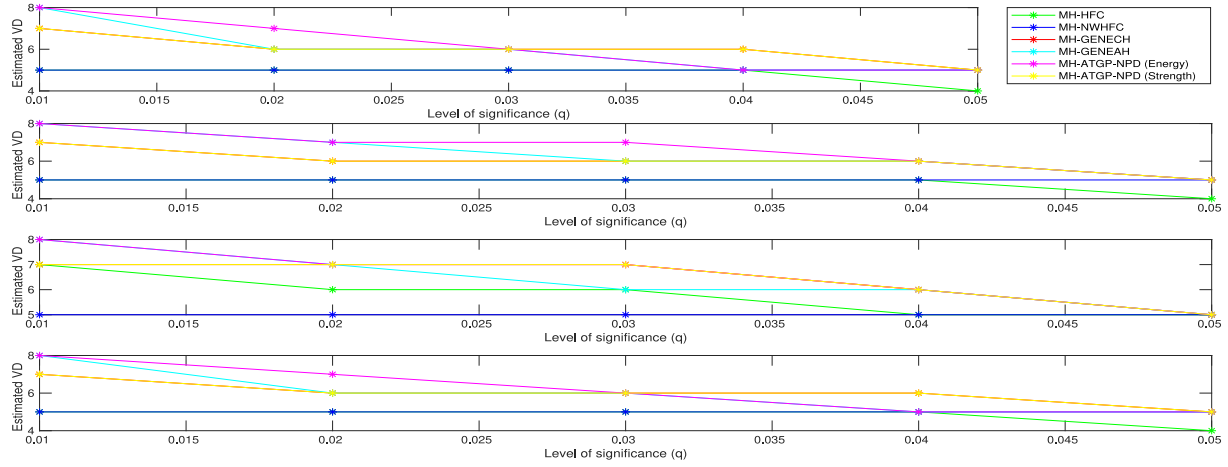


Fig. 4. Parameter sensitivity analysis for significance level (q) for the four synthetic datasets 1–4 (top to bottom) at SNR of 60 dB.

approaches and compares the results with state-of-the-art methods. HFC method estimates six endmembers for $P_F = 0.001$. HFC overestimates number of endmembers as noise increases in all the synthetic datasets. Proposed MH-HFC is giving consistent results for all the synthetic datasets for all the noise levels except at SNR of 20 dB. NWHFC [4] remove the second-order statistical correlation such that the noise variance in the corresponding correlation eigenvalue and covariance eigenvalue will be the same. On the other hand, the number of endmembers is more accurately estimated using proposed MH-NWHFC (see Table II) since the noise variances have now been decorrelated and this do not affect the subsequent eigenvalue comparison. It can be observed from Table II that for different SNRs, the estimation accuracy of the GENE algorithms is considerably

TABLE II
ESTIMATED NUMBER OF ENDMEMBERS BY VARIOUS ALGORITHMS FOR FOUR SYNTHETIC HYPERSPECTRAL DATA AT DIFFERENT NOISE LEVELS

Algorithms			Estimated # endmembers															
			SNR in synthetic Data 1				SNR in synthetic Data 2				SNR in synthetic Data 3				SNR in synthetic Data 4			
	P_F	q	80 dB	60 dB	40 dB	20 dB	80 dB	60 dB	40dB	20 dB	80 dB	60 dB	40dB	20 dB	80 dB	60 dB	40dB	20 dB
HFC [4]	0.001	N/A	8	7	8	6	6	10	12	14	6	7	8	9	6	6	7	9
NWHFC [4]	0.001	N/A	5	5	5	5	5	6	7	8	5	5	6	7	5	5	5	6
GENE-CH [6]	0.0001	N/A	6	6	7	9	6	7	8	9	6	6	7	8	6	6	8	9
GENE-AH [6]	0.0001	N/A	6	7	9	9	6	9	10	11	8	9	10	11	7	8	9	10
ATGP-NPD (Energy) [16]	0.0001	N/A	5	6	7	8	6	7	8	9	6	6	7	8	6	7	8	9
ATGP-NPD (strength) [8]	0.00001	N/A	5	5	6	7	5	6	7	8	6	7	7	8	6	7	8	9
MH-HFC [14]	N/A	0.05	5	5	5	5	5	5	5	5	5	5	6	6	5	5	5	6
MH-NWHFC [14]	N/A	0.05	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Proposed MH-GENE-CH	N/A	0.05	5	6	6	7	5	6	6	7	5	5	6	7	5	5	5	6
Proposed MH-GENE-AH	N/A	0.05	5	5	5	5	5	5	5	5	5	6	6	7	6	6	7	8
Proposed MH-ATGP-NPD (Energy)	N/A	0.05	5	5	5	5	5	5	5	5	5	6	6	5	6	6	7	7
Proposed MH-ATGP-NPD (strength)	N/A	0.05	5	5	5	5	5	5	5	5	5	6	6	7	5	5	5	6

TABLE III
TIME COMPLEXITY AND EXECUTION TIMES OF VARIOUS ALGORITHMS

Algorithms	Time complexity	Average execution time (seconds)
HFC [4]	$O(n^3)$	0.5395
MH-HFC [14]	$O(n^3)$	**0.4379**
NWHFC [4]	$O(n^3)$	0.6412
MH-NWHFC [14]	$O(n^3)$	**0.5931**
GENE-CH [6]	$O(n^4)$	30.1407
Proposed MH-GENE-CH	$O(n^4)$	**23.0464**
GENE-AH [6]	$O(n^4)$	11.9095
Proposed MH-GENE-AH	$O(n^4)$	**8.2432**
ATGP-NPD (Energy) [16]	$O(n^5)$	218.2109
Proposed MH-ATGP-NPD (Energy)	$O(n^5)$	**218.1049**
ATGP-NPD (strength) [8]	$O(n^5)$	217.6153
Proposed MH-ATGP-NPD (strength)	$O(n^5)$	**217.4120**

Algorithms are implemented in MATLAB (R2019a) and run on a laptop with Intel(R) Core(TM) i3-7100 CPU at 2.40 GHz with 4 GB RAM.

robust to the $e_{\max}$ values, and the closer the $e_{\max}$ value is to the true e, the better will be the estimation accuracy. ATGP-NPD (Energy) gives an overestimation because of real targets generated by ATGP. Table II shows that the proposed approaches consistently yield five (5) endmembers at all the noise levels except for MH-GENE-CH. It is evident since the MH-GENE-CH algorithm is suitable for data with pure pixels. It usually occurs only for the hyperspectral images taken with a reasonably high spatial resolution. Further, MH-GENE-CH gives an overestimation of endmembers due to the fact that in noise-free case, any dimensionally reduced pixel vectors lie in the convex hull of the dimensionally reduced endmember signatures. The proposed MH-GENE-AH is a better choice when the pure pixel assumption is violated.

4) Time Complexity and Average Execution Times: The time complexity and execution time (seconds) over all the scenarios under consideration of each algorithm are now discussed in Table III. Time complexities of all the algorithms are first calculated and listed in Table III. It can be observed that the time complexities of the proposed algorithms are remaining the same as that of state-of-the art algorithms since we do not require newer computing resources in our proposed algorithms. We then compute execution times by all algorithms and shown in the same Table III. One can see from the table that ATGP-based target specified VD approaches require much more time than rest of algorithms. Execution times of the proposed algorithms found lesser when compared to existing algorithms (see Table III). Note that the proposed MH testing procedure only requires calculation of p-values for null hypotheses when compared to binary hypothesis testing (other approaches) on total number of bands based on Neyman–Pearson detection criterion.

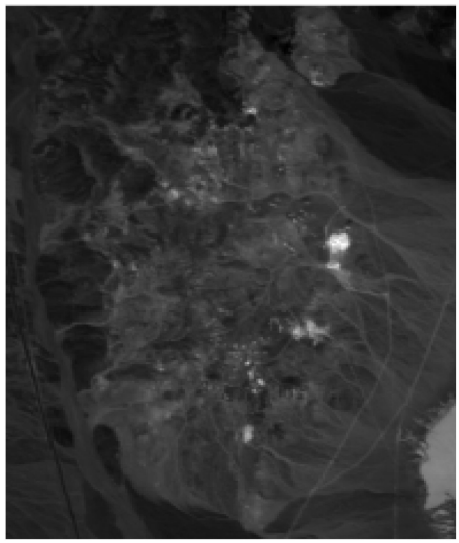


Fig. 5. Mean image of portion of AVIRIS Cuprite data.

B. Experiment on Real AVIRIS Cuprite Data

In this section, we evaluate the proposed approach on the well-known hyperspectral data collected by AVIRIS at the Cuprite mining site, USA [35]. The data are collected in 224 contiguous wavelength bands ranging from 370 to 2480 nm with a spectral resolution of 10 nm. After the removal of noisy and water absorption bands, 188 bands are retained, and a region of 250×191 is considered [35]. The mean image of the portion of AVIRIS Cuprite data are shown in Fig. 5. The estimated number of endmembers by different algorithms are listed in Table IV. Though the ground truth is not acquired for the Cuprite data, nevertheless, researchers and practitioners have considered upto 30 endmembers in the scene. It can be seen from Table IV that the proposed approaches are showing the improvement in estimation of number of endmembers. The proposed approaches MH-HFC, MH-NWHFC, and MH-GENE-AH have shown better performance. As with the synthetic datasets, the proposed MH-GENE-CH gives an overestimation of endmembers in real dataset too. One may note that the estimate of number of endmembers is better for NWHFC [4] and the proposed MH-NWHFC with

TABLE IV
ESTIMATED NUMBER OF ENDMEMBERS BY VARIOUS ALGORITHMS FOR AVIRIS CUPRITE DATA

Algorithms	P_F	q	Estimated # endmembers
HFC [4]	0.01	N/A	23
NWHFC [4]	0.01	N/A	18
GENE-CH [6]	0.00001	N/A	34
GENE-AH [6]	0.00001	N/A	31
ATGP-NPD (Energy) [16]	0.0001	N/A	37
ATGP-NPD (strength) [8]	0.0001	N/A	36
MH-HFC [14]	N/A	0.05	21
MH-NWHFC [14]	N/A	0.05	18
Proposed MH-GENE-CH	N/A	0.05	34
Proposed MH-GENE-AH	N/A	0.05	24
Proposed MH-ATGP-NPD (Energy)	N/A	0.05	25
Proposed MH-ATGP-NPD (strength)	N/A	0.05	25

Fig. 6. Mean image of HYDICE Urban data.

TABLE V
ESTIMATED NUMBER OF ENDMEMBERS BY VARIOUS ALGORITHMS FOR HYDICE URBAN DATA

Algorithms	P_F	q	Estimated # endmembers
HFC [4]	0.0001	N/A	4
NWHFC [4]	0.0001	N/A	4
GENE-CH [6]	0.0001	N/A	5
GENE-AH [6]	0.0001	N/A	4
ATGP-NPD (Energy) [16]	0.0001	N/A	7
ATGP-NPD (strength) [8]	0.0001	N/A	6
MH-HFC [14]	N/A	0.05	4
MH-NWHFC [14]	N/A	0.05	4
Proposed MH-GENE-CH	N/A	0.05	5
Proposed MH-GENE-AH	N/A	0.05	4
Proposed MH-ATGP-NPD (Energy)	N/A	0.05	4
Proposed MH-ATGP-NPD (strength)	N/A	0.05	4

the fact that the noise variances have been decorrelated and not affected the eigenvalue comparison.

C. Experiment on Real HYDICE Urban Data

In this section, the proposed approaches are evaluated on another real data collected by HYDICE. Referring to mean image of HYDICE Urban data in Fig. 6, there are 307×307 pixels, each of which corresponds to a $2 \times 2\,\mathrm{m}^2$ area. These data have 210 bands ranging from 400 to 2500 nm with a spectral resolution of 10 nm. Some bands are removed due to dense water vapor and atmospheric effects resulting in 162 band dataset. Four land cover types are generally estimated in this dataset viz. Asphalt, Grass, Tree, and Roof. The estimated number of endmembers obtained by different algorithms are shown in Table V. As shown in table, the proposed approaches are showing better and comparable results with respect to state-of-the-art methods.

V. CONCLUSION

In this article, we proposed a novel statistical approach to control the inflated Type-I error rate and provided statistical correction to the state-of-the-art approaches while estimating virtual dimensionality of the hyperspectral data. We have employed Benjamini and Hochberg MH testing procedure within the HFC, NWHFC, GENE-CH, GENE-AH, as well as in ATGP-NPD methods. Note that the proposed approach controls the false discovery rate at the specified level by incorporating consistency in the hypothesis testing. We found that the proposed strategy improves the performance of the state-of-the-art methods under the three broad approaches, i.e., eigenanalysis, target specified, and geometry-based approaches, for estimating VD. The experiments validate the efficacies of the proposed algorithms on estimating the number of endmembers while performing spectral unmixing of hyperspectral data. Further, the proposed algorithms have lower execution time while having the same time complexities as those of state-of-the-art approaches. Hence, our algorithms are statistically more robust and provide better estimate of VD.

The hyperspectral data are considered as a data cube of size $x \times y \times L$, where x and y represent spatial dimensions, and L represents the spectral dimension of the remotely acquired scene. Now, in most cases L is finite and large (say > 200) whereas x or y or both may be large. Hence, the number of hypotheses L is always finite but not extremely very large (say maximum 500). For a hyperspectral data with large x and y (which may be classified in the bigdata), the discussed inference procedure may involve new set of challenges including computational power and time. Similar challenges would be faced while analyzing multitemporal hyperspectral datasets as well, i.e., another case of bigdata where x, y, and L would be very large. However, this would more affect to the binary hypothesis testing (other approaches) when compared to proposed MH testing algorithms (see Table III). Nevertheless, one way to deal with the issue is to apply the usual inference procedure on a representative sample of the full dataset drawn using coreset sampling which is a type of importance sampling procedure. For more details on coreset, interested researchers and practitioners may refer to [36]–[39].

In this article, we control FDR for estimating VD of the hyperspectral data, where FDR is defined as in (1). However, [40], [41] and others gave arguments against including $P(B > 0)$, where B is the number of hypotheses declared significant while defining FDR. In such a scenario, one may alternatively use positive false discovery rate instead of FDR as in [41] to carry out the inference procedure.

We understand that the current work is a first step toward addressing the important issue of MH correction in the state-of-the-art methods for estimating VD. We believe this article will lead to future research on estimating VD while retaining statistical correctness.

ACKNOWLEDGMENT

The authors would like to thank the Editor, the Associate editor, and four anonymous reviewers for their insightful comments. They also acknowledge the USGS for the spectral library of minerals.

REFERENCES

[1] A. F. H. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, "Imaging spectrometry for earth remote sensing," *Sci.*, vol. 228, no. 4704, pp. 1147–1153, 1985.

[2] P. Ghamisi *et al.*, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.

[3] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2012.

[4] C.-I. Chang and Q. Du, "Estimation of number of spectrally distinct signal sources in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 608–619, Mar. 2004.

[5] C.-I. Chang, *Hyper. Imag.: Tech. for Spe. Det. and Class.*, vol. 1. Berlin, Germany: Springer-Verlag, 2003.

[6] A. Ambikapathi, T. Chan, C. Chi, and K. Keizer, "Hyperspectral data geometry-based estimation of number of endmembers using p-norm-based pure pixel identification algorithm," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2753–2769, May 2013.

[7] C.-I. Chang, W. Xiong, H.-M. Chen, and J.-W. Chai, "Maximum orthogonal subspace projection approach to estimating the number of spectral signal sources in hyperspectral imagery," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 504–520, Jun. 2011.

[8] C.-I. Chang, "A review of virtual dimensionality for hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1285–1305, Apr. 2018.

[9] N. Yokoya, J. Chanussot, and A. Iwasaki, "Nonlinear unmixing of hyperspectral data using semi-nonnegative matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1430–1437, Feb. 2014.

[10] C. Li, T. Sun, K. F. Kelly, and Y. Zhang, "A compressive sensing and unmixing scheme for hyperspectral data processing," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1200–1210, Mar. 2012.

[11] R. Huang, X. Li, and L. Zhao, "Spectral–spatial robust nonnegative matrix factorization for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8235–8254, Oct. 2019.

[12] W. Zhang, X. Lu, and X. Li, "Similarity constrained convex nonnegative matrix factorization for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4810–4822, Jul. 2019.

[13] S. R. Austin, I. Dialsingh, and N. Altman, "Multiple hypothesis testing: A review," *J. Indian Soc. Agricultural Statist.*, vol. 68, no. 2, pp. 303–14, 2014.

[14] S. S. Vijayashekhar, J. S. Bhatt, and B. Chattopadhyay, "A statistical approach to improve virtual dimensionality of hyperspectral data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 2272–2275.

[15] J. C. Harsanyi, W. Farrand, and C.-I. Chang, "Detection of subpixel spectral signatures in hyperspectral image sequences," in *Proc. Annu. Meeting, Amer. Soc. Photogrammetry Remote Sens.*, 1994, pp. 236–247.

[16] D. Paylor and C.-I. Chang, "A theory of least-squares target-specified virtual dimensionality in hyperspectral imagery," in *Proc. SPIE*, vol. 9124, 2014, Art. no. 912402.

[17] E. R. Malinowski, "Determination of the number of factors and the experimental error in a data matrix," *Analytical Chem.*, vol. 49, no. 4, pp. 612–617, 1977.

[18] K. Cawse-Nicholson, A. Robin, and M. Sears, "The effect of correlation on determining the intrinsic dimension of a hyperspectral image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 482–487, Apr. 2013.

[19] D. Paylor and C. I. Chang, "Second-order statistics-specified virtual dimensionality," *Proc. SPIE*, vol. 8743, 2013, Art. no. 87430X.

[20] O. Kuybeda, D. Malah, and M. Barzohar, "Rank estimation and redundancy reduction of high-dimensional noisy signals with preservation of rare vectors," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5579–5592, Dec. 2007.

[21] H. Ren and C.-I. Chang, "Automatic spectral target recognition in hyperspectral imagery," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1232–1249, Oct. 2003.

[22] C.-I. Chang, W. Xiong, and C.-H. Wen, "A theory of high-order statistics-based virtual dimensionality for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 188–208, Jan. 2014.

[23] C.-I. Chang, "A unified theory for target-specified virtual dimensionality of hyperspectral imagery," *Proc. SPIE* , vol. 8539, 2012, Art. no. 85390J.

[24] P. J. Martínez, R. M. Pérez, A. Plaza, P. L. Aguilar, M. C. Cantero, and J. Plaza, "Endmember extraction algorithms from hyperspectral images," *Ann. Geophys.*, vol. 49, no. 1, pp. 93–101, 2006.

[25] J. M. Bioucas-Dias and J. M. P. Nascimento, "Estimation of signal subspace on hyperspectral data," *Proc. SPIE*, vol. 5982, 2005, Art. no. 59820L.

[26] J. M. Bioucas-Dias and J. M. P. Nascimento, "Hyperspectral subspace identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 8, pp. 2435–2445, Aug. 2008.

[27] J. S. Bhatt, M. V. Joshi, and S. S. Vijayashekhar, "A multitemporal linear spectral unmixing: An iterative approach accounting for abundance variations," in *Proc. 9th Workshop Hyperspectral Image Signal Process., Evol. Remote Sens.*, 2018, pp. 1–5.

[28] Y. Benjamini *et al.*, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc.*, pp. 289–300, 1995.

[29] C.-I. Chang, *Real-Time Recursive Hyperspectral Sample and Band Processing: Algorithm Architecture and Implementation*. Berlin, Germany: Springer-Verlag, 2017.

[30] M. R. Leadbetter, G. Lindgren, and H. Rootzén, *Extremes and Related Properties of Random Sequences and Processes*. Berlin, Germany: Springer-Verlag, 2012.

[31] C.-I. Chang, X. Jiao, C.-C. Wu, E. Y. Du, and H.-M. Chen, "Component analysis-based unsupervised linear spectral mixture analysis for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol 49, no. 11, pp. 4123–4137, Nov. 2011.

[32] R. F. Kokaly *et al.*, "USGS spectral library version 7 data: US geological survey data release," 2017.

[33] F. Zhu, Y. Wang, B. Fan, G. Meng, and C. Pan, "Effective spectral unmixing via robust representation and learning-based sparsity," *CoRR*, abs/1409.0685, 2014.

[34] F. Zhu, Y. Wang, B. Fan, G. Meng, S. Xiang, and C. Pan, "Spectral unmixing via data-guided sparsity," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5412–5427, Dec. 2014.

[35] F. Zhu, Y. Wang, S. Xiang, B. Fan, and C. Pan, "Structured sparse method for hyperspectral unmixing," *ISPRS J. Photogrammetry Remote Sens.*, vol. 88, pp. 101–118, 2014.

[36] Q. Liu *et al.*, "The coreset variational bayes (CVB) algorithm for mixture analysis," *Brazilian J. Probability Statist.*, vol. 33, no. 2, pp. 267–279, 2019.

[37] O. Bachem, M. Lucic, and A. Krause, "Scalable k-means clustering via lightweight coresets," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 1119–1127.

[38] O. Bachem, M. Lucic, and S. Lattanzi, "One-shot coresets: The case of k-clustering," 2017, *arXiv:1711.09649*.

[39] D. Feldman, M. Schmidt, and C. Sohler, "Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering," in *Proc. 24th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2013, pp. 1434–1453.

[40] J. P. Shaffer, "Multiple hypothesis testing," *Annu. Rev. Psychol.*, vol. 46, no. 1, pp. 561–584, 1995.

[41] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proc. Nat. Acad. Sci.*, vol. 100, no. 16, pp. 9440–9445, 2003.

Vijayashekhar S S (Student Member, IEEE) received the B.E. degree in electronics and communications engineering from Visvesvaraya Technological University, Karnataka, India, in 2009, and the M.Tech. degree in digital electronics and advanced communication systems from Manipal university, Karnataka, India, in 2014. He is currently a full-time Ph.D. Research Scholar with the Indian Institute of Information Technology Vadodara, Gandhinagar, India.

He was an Assistant Professor for over five years in Karnataka, India. His research interests include signal processing, hyperspectral unmixing, and machine learning.

Jignesh S. Bhatt received the B.E. degree in electronics and communications engineering from Hemchandracharya North Gujarat University, Patan, India, in 2000, the M.Tech. degree in communication systems from the Sardar Vallabhbhai National Institute of Technology, Surat, India, in 2008, and the Ph.D. degree in information and communication technology from the Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India, in 2014.

Since 2014, he has been an Assistant Professor with the Indian Institute of Information Technology Vadodara, Gandhinagar, India. He has authored a book entitled *Regularization in Hyperspectral Unimixing* (Bellingham, WA, USA: SPIE Press, 2016). He has developed image registration software for meteorological payloads with the Department of Space, Indian Space Research Organization, Bengaluru, India. His research interests include remote sensing, signal processing, computer vision, deep learning, and inverse ill-posed problems, especially in hyperspectral imagery and medical data.

Bhargab Chattopadhyay received the Ph.D. degree in statistics from the Department of Statistics, University of Connecticut, Storrs, CT, USA, in 2012.

He is currently an Assistant Professor with the Department of Decision Sciences, Indian Institute of Management Visakhapatnam, Visakhapatnam, Andhra Pradesh. Post Ph.D., he worked as Assistant Professor with the University of Texas at Dallas and the Indian Institute of Information Technology Vadodara. His main research interests include sequential analysis, statistical inference and change point detection with application in econometrics, psychology, actuarial science, and others.