Semi-realistic Simulations of Natural Hyperspectral Scenes

Zhipeng Hao, Mark Berman, Yi Guo, Glenn Stone, and Iain Johnstone

***Abstract*—Many papers in the hyperspectral literature use simulations (based on a linear mixture model) to test algorithms, which either estimate the "intrinsic" dimensionality (ID) of the data or endmembers. Usually, these simulations use "real-world" endmembers, proportions distributed according to a uniform or Dirichlet distribution on the endmember simplex, and Gaussian errors which are "spectrally" and "spatially" uncorrelated. When the error standard deviations (SDs) in different bands are assumed to be unequal, they are usually estimated using Roger's method. The simulated and real-world data in these papers are so different that one cannot be confident that the various advocated methods work well with real-world data. We propose a general methodology which produces more realistic simulations, providing us with greater insights into the strengths and weaknesses of various advocated methods. With the aid of the well-known Indian Pines and Cuprite scenes, we compare several specific options within the proposed methodological framework. We also compare the performance of five well-known ID estimators using both real and simulated datasets and demonstrate that Roger's SD estimates are positively biased. A proof that Roger's estimates are always positively biased is given.**

***Index Terms*—Dimensionality, endmembers, hyperspectral, linear mixture model, simulation.**

I. Introduction

SIMULATION is useful for assessing the performance of algorithms in many fields. This is especially true of hyperspectral remote sensing data, where ground-based validation of such algorithms is typically costly and time consuming. The hyperspectral remote sensing literature is dominated by two broad simulation approaches. The first, captured in the Digital Imaging and Remote Sensing Image Generation model, uses sophisticated radiometric and geometric models of real-world scenes to reconstruct closely those scenes [1]. This approach is particularly useful for modeling 3-D objects such as buildings and trees.

The second approach is not based on real-world scenes. It is found in the hyperspectral literature concerned with linear mixture models. These papers are concerned with either 1) estimating the "intrinsic" dimensionality of the data [2]–[5], 2) introducing a new blind unmixing/endmember estimation algorithm [6]–[10] or 3) both [11]. All the simulations are based on the linear mixture model, which we now outline. Let $X_i, i = 1, ..., N$, denote the d-dimensional vector of observations at pixel i (out of N) in a hyperspectral image. Under the linear mixture model, if there are $M(<N)$ spectrally distinct materials in the image, then

$$X_i = \sum_{j=1}^{M} w_{\mathrm{ij}} E_j + \epsilon_i, \quad i = 1, \ldots, N \tag{1}$$

where (i) $E_j, j = 1, \ldots, M$, are the "endmembers" (i.e., pure materials); (ii) w_{ij} are *nonnegative* weights; and (iii) ϵ_i are error terms. The errors are typically a combination of instrumental noise, natural variation in spectra representing the same material, and small nonlinearities in the mixing. M is sometimes called "virtual dimensionality" (VD) [2] and sometimes called "intrinsic dimensionality" (ID) [5], [12]. We shall use the latter term.

Often the weights in (1) are also constrained to be proportions, i.e.,

$$\sum_{j=1}^{M} w_{\mathrm{ij}} = 1, \quad i = 1, \ldots, N. \tag{2}$$

For ease of exposition, we will assume that (2) is satisfied throughout the paper. Small modifications are required when it is not satisfied. Note however that, when it is satisfied, the M endmembers define a simplex in an $(M-1)$-dimensional subspace. For instance, when $M = 3$, the endmembers define a triangle in a 2-D subspace.

The simulations in the ten above-mentioned references vary according to how the three components are chosen. For (i), all use "real-world" endmembers, usually chosen from the USGS or JPL spectral libraries, or other sources [2], [4], [10]. For (ii), the proportions are mostly chosen to be spatially uncorrelated. The proportion distribution is either "random" [2], [5], [10], distributed uniformly [4], [11], or according to a Dirichlet distribution on the simplex defined by the endmembers [3], [7], [9]. In two of these papers [9], [11], the maximum proportions of some endmembers are restricted to being less than 1, modeling the common situation where some materials in a scene are never pure. Papers [6] and [8] incorporate some spatial information. For (iii), all assume Gaussian errors (as shall we in our simulations). Most assume that the errors are "spectrally" and "spatially" uncorrelated. Papers [6]–[11] all assume that the error standard deviations (SDs) in all the bands are equal. Papers [2]–[5] assume that they are different. The first three

Manuscript received July 02, 2015; revised December 28, 2015; accepted May 26, 2016. *(Corresponding author: Mark Berman.)*

Z. Hao, Y. Guo, and G. Stone are with the School of Computing, Engineering and Mathematics, Western Sydney University, Parramatta, NSW 2150, Australia (e-mail: z.hao@westernsydney.edu.au; y.guo@westernsydney.edu.au; g.stone@westernsydney.edu.au).

M. Berman is with the School of Computing, Engineering and Mathematics, Western Sydney University, Parramatta, NSW 2150, Australia, and also with CSIRO Data61, North Ryde, NSW 2113, Australia (e-mail: mark.berman@csiro.au).

I. Johnstone is with the Department of Statistics, Stanford University, CA 94305-4065 USA (e-mail: imj@stanford.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2016.2580178

papers use a (nonspatial) multivariate method suggested by Roger [13] to estimate the SDs, while paper [5] uses a method based on finding homogeneous areas in the scene [14]. For nine of the above-mentioned references, M varies between 3 and 10; one of the simulations carried out by Bioucas-Dias and Nascimento [3] has $M = 15$. For the ten references, N varies between 150 and 10 000.

The above-mentioned papers apply their advocated dimensionality or endmember estimation methods to one or more real-world scenes, in most cases one of the AVIRIS scenes of the Cuprite mining district, but also other scenes [2], [3], [5], [10]. Apart from one of the real-world datasets analyzed by Somers, Zortea, Plaza, and Asner [10] (for which $N = 2700$), the number of pixels in the remaining real-world datasets varies between 21 025 and 122 500, so they are considerably larger than the simulated data sets. The estimated values of M, using one or more of the methods presented in [2]–[5], are, with two exceptions (see [2] and [6]), all greater than 10 and in some cases greater than 20. This reflects the obvious principle that larger (real-world) datasets tend to contain more materials. In this age of "big data," M will tend to become bigger as real-world datasets become bigger! This illustrates obvious differences between the simulated and real-world datasets in the above-mentioned papers. However, the endmembers and the spatial distribution of the proportions also differ. Therefore, in our opinion, the simulated and real-world data in these papers are so different that we cannot really be confident that the various advocated methods work well with real-world data.

In this paper, we propose a general simulation methodology which, while not as sophisticated as a physics-based approach [1], produces simulations which are more like real-world data and provide us with greater insights into the strengths and weaknesses of various ID and endmember estimation methods. We believe that the methodology is potentially useful for simulating "natural" scenes (i.e., those dominated by vegetation, minerals and soils, rather than buildings), where the main interest is in knowing what materials are in a scene, and where and how much they are present. The general methodology is outlined in Section II. Various options within the general framework are also considered. This includes a comparison with methodologies in two recent papers [15], [16], which are similar to ours, but which do not go as far as we propose. In Section III, we describe five well-known ID estimation methods, all of which rely on various eigendecompositions of the data, while in Section IV, we introduce another concept, which we call "effective intrinsic dimensionality" (EID). In some cases, some of the "signal" eigenvalues are smaller than the "noise" eigenvalues. It is unreasonable to expect an eigenvalue-based ID estimation method to identify these. EID represents this concept mathematically. In Section V, we apply the five ID estimation methods to two real-world datasets and to simulated versions of them over a plausible range of values of M, and discuss the results. General conclusions are drawn and future work is discussed in Section VI.

One consequence of our "real-world" approach to simulation is the observation that Roger's method for estimating the band error SDs is positively biased. In Appendix A, we prove that this is always so. In Appendix B, we also develop some theory, which suggests that the average bias depends on the ratio M/d. As this ratio becomes larger, the bias also tends to become larger. The theory is supported by our simulations.

II. General Methodology and Some Options

The general methodology is straightforward. One chooses a real-world scene of interest and applies 1) a method to estimate its dimensionality ($\hat{M}$) and 2) a method to produce $\hat{M}$ endmember estimates ($\hat{E}_j, j = 1, \ldots, \hat{M}$). One can then estimate the weights, $\hat{w}_{\mathrm{ij}}$, usually by least-squares (LS) estimation (with constraints determined by the assumed weight constraints). All these estimates can be plugged into (1) to produce an estimated "true" or "target" image. Finally, the simulated image is produced by adding simulated errors δ_i, following the assumed error distribution, to the "target" image:

$$Y_i = \sum_{j=1}^{\hat{M}} \hat{w}_{\mathrm{ij}} \hat{E}_j + \delta_i, \quad i = 1, \ldots, N. \tag{3}$$

A similar approach to ours is adopted by Gao, Du, Zhang, Yang, and Wu [15]. They apply one of the three VD methods introduced in [2] (but do not state which one) to a 500×500 AVIRIS Cuprite subscene to estimate M and estimate the endmembers using N-FINDR [17] in minimum noise fraction (MNF) space [18]. They then produce the target image in a similar manner to ourselves. Spectrally uncorrelated errors are then added to the target image, using signal-to-noise ratios (SNRs), which are constant across the bands. They compare a number of error ("noise") estimation methods using two metrics based on the difference between the target and estimated errors at each pixel. They do not estimate band error variances themselves.

In [16], N-FINDR [17] is applied to a 350×350 AVIRIS Cuprite subscene, given M, which ranges from 5 to 40 in their simulations. They do not specify whether N-FINDR is run in MNF or some other space. They add spectrally uncorrelated errors, whose variances are both constant and variable. They also investigate the impact of correlated errors. We will not investigate correlated errors in this paper, although we do intend to investigate these in the future. The focus of [16] is to compare the performance of different ID estimation methods using both simulated and real data.

Although our general methodology is somewhat similar to those of [15] and [16], our philosophy is somewhat different. Our objective is to make our simulations as close as possible to real-world scenes by 1) visual inspection of principal component (PC) images (after "whitening" the data), and 2) making the eigenvalues of the real and simulated images as similar as possible. The second aim is particularly important because methods such as VD [2], RSSE [4], and random matrix theory (RMT) [5] are all based on the eigenvalues. If we can make the eigenvalues of real-world and simulated images as close possible, especially for indices equal to and near the true ID, then we are in a stronger position to compare different dimensionality estimation techniques with real data.

With our objective in mind, the approaches of [15] and [16] raise the following questions: 1) Do the errors in real-world datasets have either constant variance or constant SNRs across bands? 2) How important is the space in which the endmembers are found? 3) How important is the endmember estimation algorithm used? We will at least in part address questions 1 and 3 here, with the aid of two well-known scenes: the relatively small 145×145 Indian Pines scene ($N = 21\,025$) and the much larger 512×614 Cuprite scene f970619t01p02_r02_sc03.a.rfl ($N = 314\,368$). For both scenes, some bands have been excluded due to water absorption or noise. For Indian Pines, bands 104–110, 149–165, and 218–220 have been excluded, leaving 193 (out of 220) bands. For Cuprite, bands 1, 2, 104–114, 153–174, and 221–224 have been excluded, leaving 185 (out of 224) bands.

A. *Error Variance Assumptions*

Various methods have been proposed for estimating band error variances. In a forthcoming paper, using simulations similar to those described here, we demonstrate that Roger's method gives much better estimates of error variances than do a variety of spatial methods. This is consistent with observations in [3] and [16] and results in [15]. Roger's method [13] performs a linear regression of each of the d bands on the remaining $d-1$ bands and estimates the variance of each band using the residuals from the fits. Specifically, let $\hat{\epsilon}_{ij}, i = 1, \ldots, N; j = 1, ..., d$, denote the error in the ith pixel when regressing band j on the remaining $d-1$ bands. Roger's estimator of the error variance in band j is

$$\hat{\sigma}^2_{\epsilon,j} = \sum_{i=1}^{N} \hat{\epsilon}^2_{ij}/N, \quad j = 1, \ldots, d. \tag{4}$$

Some properties of this estimator are given in Section II-C. Technical details are given in the Appendixes.

Fig. 1(a) and (b) shows Roger's estimates of the band error SDs for the Indian Pines and Cuprite images. We also show the mean (± 2 SDs) for the SDs of the corresponding simulated images. For Indian Pines, 100 simulations have been used with $\hat{M} = 20$, while for the larger Cuprite scene, only 40 simulations have been used with $\hat{M} = 36$. These values of $\hat{M}$ have been chosen for illustrative purposes only and as initial approximate estimates of the true value of M. Further reasons for choosing these values are given in Section V. The main point here is to show that the error variances are not constant.

What about the assumption of constant SNR? We will use the following common definition for the SNR in band j:

$$SNR_j = 10 \log_{10}\{\text{Var}(s_j)/\text{Var}(\epsilon_j)\} \tag{5}$$

where $\text{Var}(s_j)$ and $\text{Var}(\epsilon_j) \equiv \sigma^2_{\epsilon,j}$ are the variances of the signal and error components, respectively for band j. The usual estimate of $\text{Var}(s_j)$ is just the sample variance in band j. We will use Roger's estimates of $\text{Var}(\epsilon_j)$, given by (4). The estimated SNRs for the Indian Pines and Cuprite datasets are shown in Fig. 2(a) and (b), respectively. In our simulations, we have assumed Gaussian errors, which are both spectrally and spatially uncorrelated, and with SNRs given by those in Fig. 2(a) and (b).

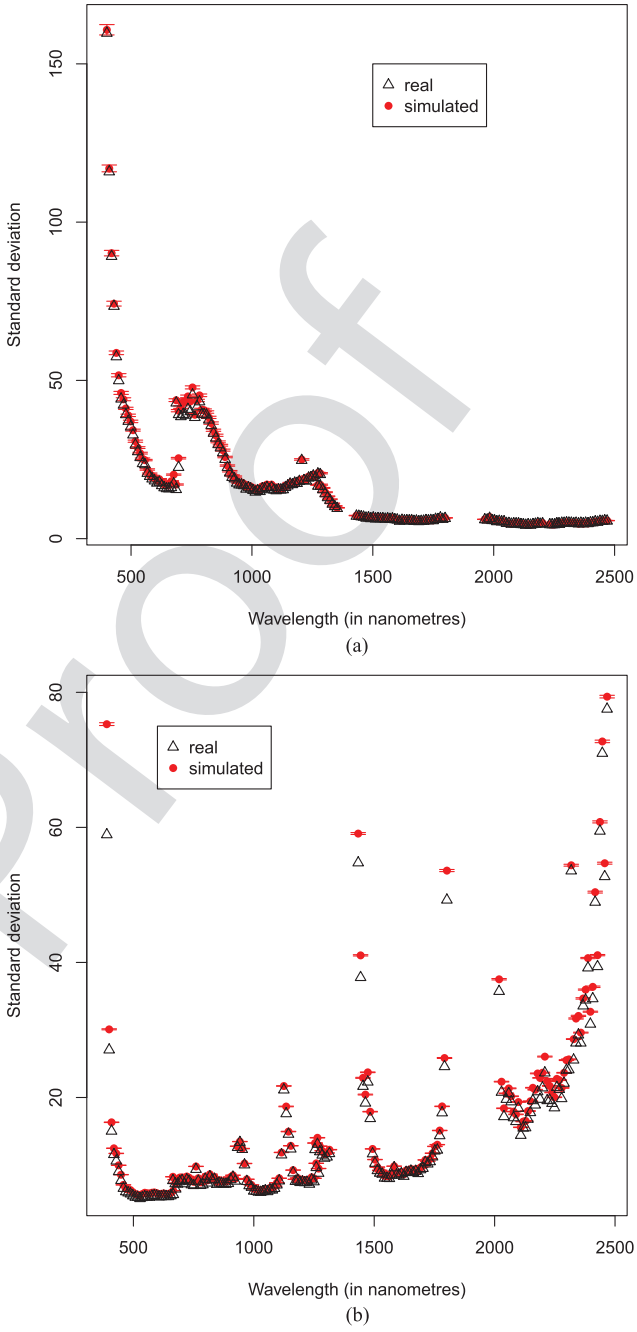


Fig. 1. SDs for real Indian Pines and Cuprite images, and mean (± 2 SDs) for SDs of corresponding simulated images. (a) Indian Pines ($\hat{M} = 20$, 100 simulations). (b) Cuprite ($\hat{M} = 36$, 40 simulations).

We will refer to these as "variable SNR" simulations. For comparison purposes, we have also simulated Gaussian errors with constant SNRs with the values given by the horizontal lines (the average SNR) in Fig. 2(a) and (b).

B. *Which Endmember Estimation Algorithm Should Be Used?*

Given $\hat{M}$, both [15] and [16] use N-FINDR [17] to estimate the endmembers. N-FINDR finds the simplex of maximum hypervolume with M vertices, constrained to lie among the data points, typically in a subspace of dimensionality $M-1$ (e.g.,

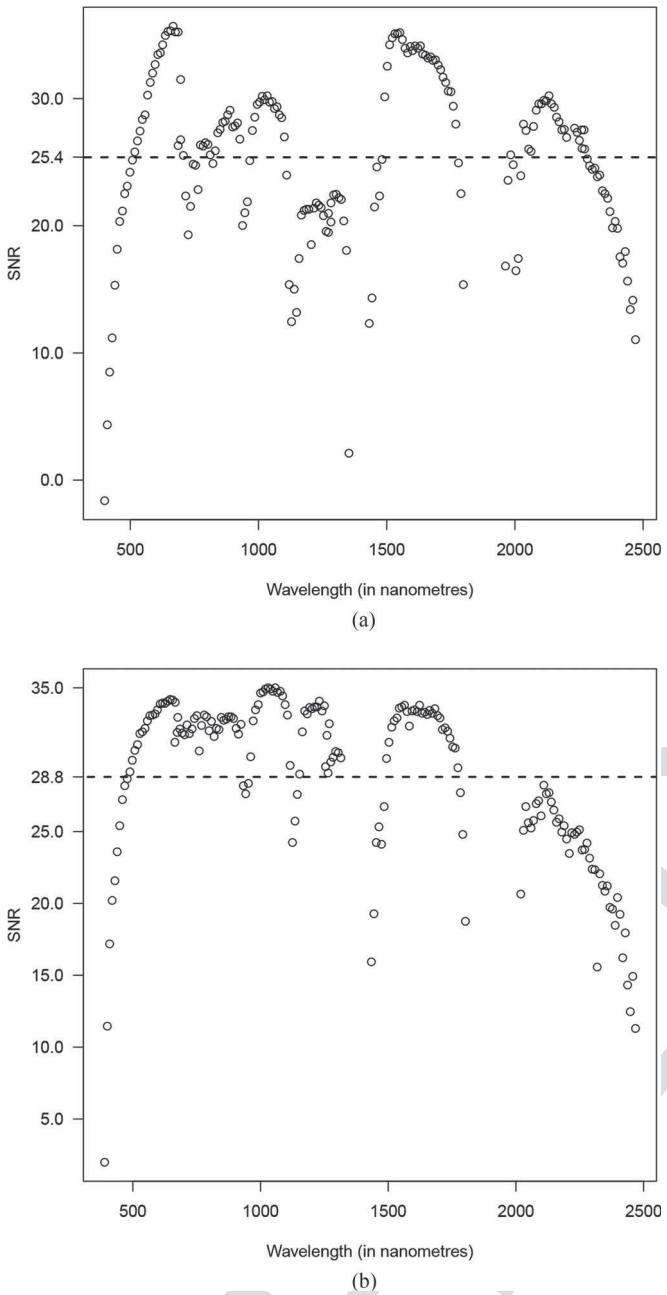


Fig. 2. Estimated SNRs for real Indian Pines and Cuprite images. (a) Indian Pines. (b) Cuprite.

the first $M-1$ PC or MNF bands). Because of this constraint, N-FINDR implicitly assumes that pure versions of all the endmembers exist in the data itself. In our opinion, this assumption is unrealistic. In the last 20 years or so, many endmember estimation algorithms, which seek to overcome this problem have been proposed; [11, Sec. IV.B] provides a good review of these.

Probably, the oldest of these is called the minimum volume transform (MVT) [19]. Since then, there have been several algorithms proposed to calculate the MVT [7], [20], [21]. Unlike N-FINDR, MVT finds the simplex of minimum hypervolume with M vertices, constrained to totally enclose the data projected onto the $(M-1)$-dimensional subspace inside the simplex. This algorithm is unrealistic because it assumes that all the data in the first $M-1$ PC or MNF bands is signal (i.e., with no errors). However, if this endmember estimation method is used, it is particularly simple to produce the target image (3). Note that, in this case, any simplex with M vertices which completely encloses the projected data in the $(M-1)$-dimensional subspace has zero error in that subspace. As a simple example, when $M=3$, this simplex will be any triangle enclosing all the (projected) data. In this case, the target image (the signal component of (3)) can be reconstructed directly from the first $M-1$ PC or MNF bands, with no constraints on the weights.

Several algorithms, which aim to steer a middle course between the "extremes" of N-FINDR and MVT, have been developed. We will use the iterated constrained endmembers (ICE) algorithm [22]. This uses a regularized LS fit of (1) (typically in MNF space), where the regularizing function is proportional to the total variance of the endmembers.

We have simulated both the Indian Pines and Cuprite datasets for a range of values of $\hat{M}$ (discussed in Section V) using both MVT and ICE to estimate the endmembers in (3). The errors δ_i in (3) are uncorrelated with different SDs, assuming either a variable or constant SNR [see Fig. 2(a) and (b)], estimated from the real data using Roger's method [13]. Figs. 3 and 4 show the eigenvalues for one of the simulated Indian Pines ($\hat{M}=20$) and Cuprite ($\hat{M}=36$) datasets using both ICE and MVT for the simulations using variable SNR. For all the ICE simulations shown in this paper, we have used its default regularisation parameter, 0.01. We have first "scaled" each real and simulated image by dividing the data in each band by Roger's estimate of the band SD, so the SNR is variable. For each dataset, we plot the first K (scaled) eigenvalues in plot (a) (where K is a little less than $\hat{M}$) and the remaining eigenvalues in plot (b). In plot (b), we have also drawn a vertical line between the first $\hat{M}-1$ "signal" eigenvalues and the "noise" eigenvalues. In both Figs. 3(b) and 4(b), we see that the real scaled eigenvalues die away smoothly, and that there is no obvious drop between the signal and noise eigenvalues. This is commonly the case with real data. The ICE eigenvalues reflect this behavior. However, the MVT eigenvalues do not; there is a very obvious drop between the "signal" and "noise" eigenvalues. Any reasonable ID estimation method ought to be able to detect this gap and give a perfect estimate of M. The reason for this gap is the implicit (and unrealistic) assumption made by MVT that there is no "noise" in the first $M-1$ (scaled) PCs. Consequently, we will only use the ICE simulations for the rest of the paper.

Although the ICE eigenvalues behave more like the real eigenvalues than the MVT eigenvalues do, there are still some differences. There are at least two reasons for this. First, the values of $\hat{M}$ are a little too small. We will present evidence for this in Section V. Second, Roger's estimates of the band error variances are too large. A careful examination of Fig. 1(a) and (b) suggests that this is so. This is made clearer in Fig. 5(a) and (b), which shows the mean (± 2 SDs) of Roger's estimates of the band error SDs for the simulated datasets, each divided by their target SDs, for the Indian Pines and Cuprite datasets, respectively. We will call these the estimated relative SDs. The error bars are larger for the Indian Pines scene because the image size ($N=21\,025$) is much smaller than that of the Cuprite scene ($N=314\,368$).

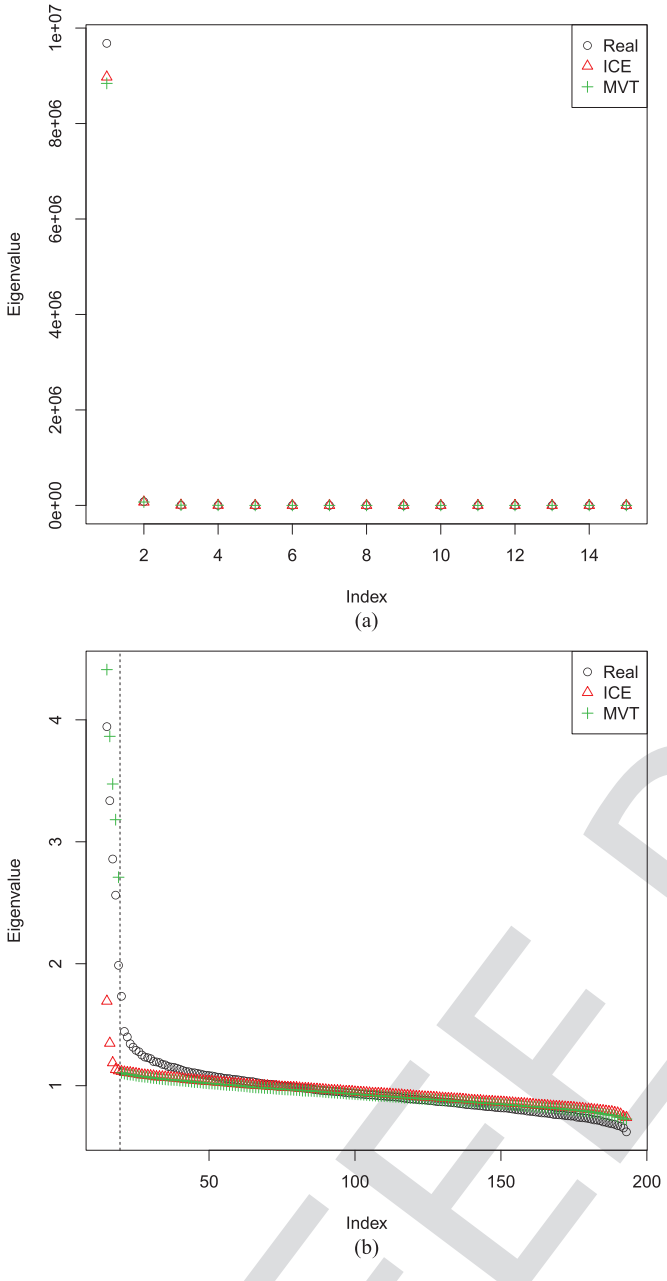


Fig. 3. Indian Pines: Real and simulated ICE and MVT scaled eigenvalues (variable SNR, $\hat{M} = 20$). (a) Eigenvalues 1–15. (b) Eigenvalues 15–193.

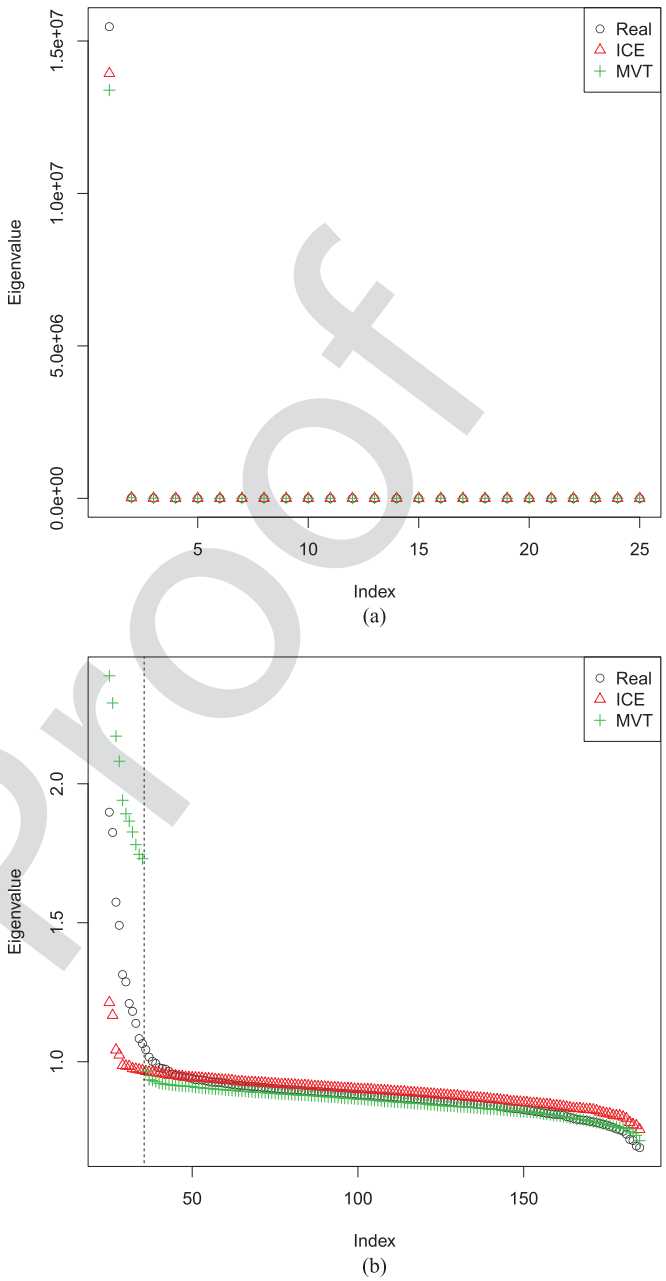


Fig. 4. Cuprite: Real and simulated ICE and MVT scaled eigenvalues (variable SNR, $\hat{M} = 36$). (a) Eigenvalues 1–25. (b) Eigenvalues 25–185.

C. Some Properties of Roger's Estimator

In Appendix A, we prove that, under mild assumptions, Roger's estimate, $\hat{\sigma}^2_{\epsilon,j}$ [defined in (4)] is greater than the true error variance $\sigma^2_{\epsilon,j}$.

Therefore, when scaling data by dividing by Roger's SD estimates, we are dividing by values which are too large, and hence, the real error variances of the scaled data are all too small. Because the sum of the variances of the scaled data equals the sum of the scaled eigenvalues, the larger eigenvalues will also need to be too small to guarantee this. This is what we see in Figs. 3 and 4. The first 85 and 47 real scaled eigenvalues are larger than their simulated counterparts in these two figures, respectively.

Let

$$r_j = \hat{\sigma}^2_{\epsilon,j} / \sigma^2_{\epsilon,j}. \tag{6}$$

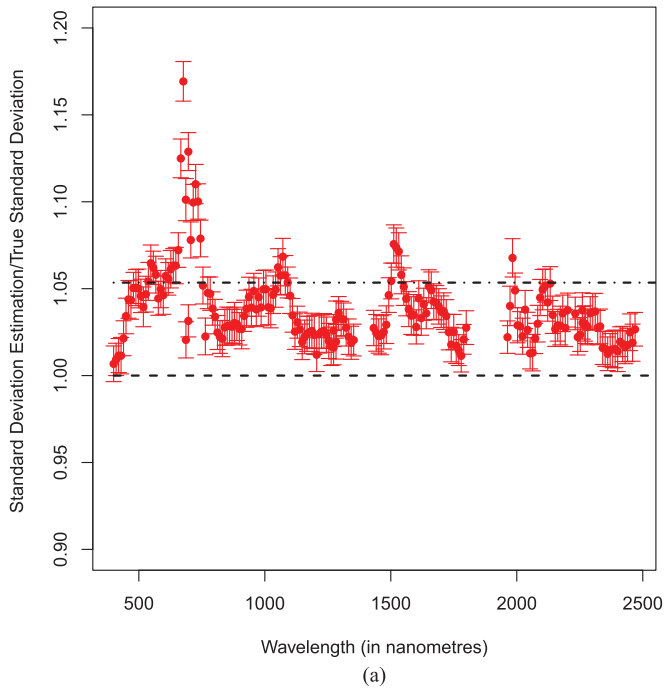
This is the estimated relative variance. The positive bias result above provides the lower bound

$$r_j > 1, \quad j = 1, ..., d. \tag{7}$$

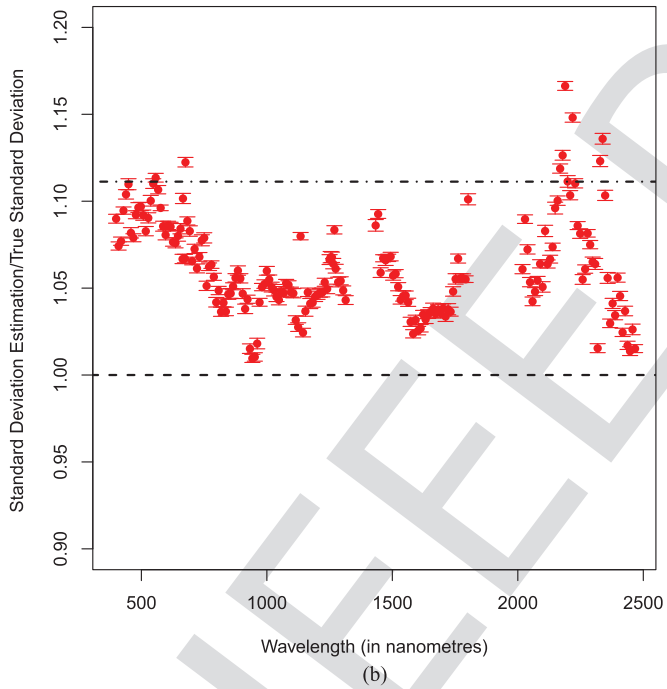
We have also been able to derive a lower bound on the average inverse relative variance:

$$\sum_{j=1}^{d} r_j^{-1} / d > (d - M)/d. \tag{8}$$

The proof of (8) is given in Appendix B.



(a)



(b)

Fig. 5. Mean (± 2 SDs) for relative SDs for simulated Indian Pines and Cuprite images, and upper bounds. (a) Indian Pines ($\hat{M} = 20$, 100 simulations). (b) Cuprite ($\hat{M} = 36$, 40 simulations).

If we are prepared to take a Bayesian approach and assume that the jth row of the eigenvector matrix is uniformly distributed on the d-dimensional sphere, then we can show that

$$E(r_j) < (d-2)/(d-M-2). \tag{9}$$

The proof of (9) is also given in Appendix B.

The upper line in Fig. 5(a) and (b) is $\{(d-2)/(d-M-2)\}^{1/2}$ (since we are plotting SDs rather than variances). Most of the estimated relative SDs in both figures lie below this value. For a single Indian Pines simulation, the left-hand side of (9) is 1.078, while the right-hand side is 1.110. For a single Cuprite simulation, the left-hand side of (9) is 1.132, while the right-hand side is 1.236. So, (9) holds for both simulated datasets.

The value of the upper bound in (9) is that it gives an approximate idea of the likely magnitude of the typical bias in Roger's estimates. In most of the above-mentioned papers, d varies around 200. As mentioned previously, for almost all the simulations in the above papers, $M \leq 10$. When $d = 200$ and $M = 10$, $(d-2)/(d-M-2) = 1.053$, and so the typical bias is less obvious and less important. However, hyperspectral satellites due for launch in the next few years will all have about $d = 200$ bands [23]. The images produced by such satellites will undoubtedly be larger than those produced by the simulations in the above-mentioned papers. Hence, many will have a larger ID, M, and so the typical bias in Roger's estimates will be larger and more significant.

III. Some ID Estimation Methods

Five ID estimation methods will be discussed briefly in this section. The first three, HFC, noise-whitened HFC (NWHFC), and noise subspace projection (NSP), are collectively called VD and were introduced in [2]. All methods assume that the errors are spectrally (and spatially) uncorrelated.

A. HFC

HFC implicitly assumes that all the band error SDs are equal. It tests statistically the equality of the eigenvalues of the sample covariance matrices with and without mean correction. Its ID estimate is the number of eigenvalue pairs determined to be unequal. HFC (as well as NWHFC and NSP) require the user to set a false alarm probability, P_f. In most papers, P_f is set to $10^{-3}, 10^{-4}$, or 10^{-5}.

B. NWHFC

NWHFC does not assume that the band error SDs are equal. It scales the data using Roger's error SD estimates [13] and then applies HFC to the scaled data.

C. NSP

Note that, if we were to scale the data by the "true" error SDs, then the error SDs of the scaled data will all be 1. One would then expect those eigenvalues of the scaled data, which correspond to the noise, to have values which are on average about 1. This is the basis of the third method, called "noise subspace projection" (NSP). The eigenvalues of the mean-corrected sample covariance matrix are tested against 1. When these are not significantly different from 1, one concludes that they correspond to noise. However, NSP has two problems. First, when the data are scaled by the true error SDs, although the "noise" eigenvalues are on average about 1, the first noise eigenvalue is somewhat greater than 1, in a way that can be made explicit mathematically [24], [25]. In this case, the NSP estimate would be too large. However, Roger's estimates are positively biased and consequently, as noted in Section II-C, the leading eigenvalues are too small, so the value of 1 is reached earlier than it

is when scaled by the true error SDs. These two "wrongs" have opposite effects and sometimes make a "right." However, often they do not, as we shall see in Section V.

D. HySime

HySime [3, Algorithm 2] uses Roger's method to estimate the signal and noise components at each pixel, and from these to estimate signal and noise covariance matrices. They then decompose the expected mean square error (MSE) of the true signal into estimated signal and noise components via an eigendecomposition of the estimated signal covariance matrix. Their ID estimate is the dimensionality of the subspace of eigenvectors which minimizes the MSE.

E. Random Matrix Theory

RMT is used by [26] to estimate the ID, under the assumption that the band error variances are equal. This approach is adapted by [5, Algorithm 1] to deal with the case where the band error variances are unequal. They estimate these using a method based on finding homogeneous areas in the scene [14]. However, they do not scale the data by the estimated band error SDs, but subtract the estimated error covariance matrix (i.e., which is assumed to be diagonal) from the data covariance matrix, to obtain an estimate of the signal covariance matrix. For the purposes of comparison with the other ID estimation methods described above, we will use [5, Algorithm 1], but with Roger's estimates of the band error variances instead of those based on the method described in [14].

IV. Effective Intrinsic Dimensionality

There are a number of definitions of ID in the hyperspectral literature. Cawse-Nicholson, Damelin, Robin, and Sears [5, Introduction] give a useful review of these. In particular, Chang and Du [2] define VD as "the minimum number of spectrally distinct signal sources that characterize the hyperspectral data from the perspective view of target detection and classification." This is a very informal definition. A more formal definition is given by Cawse-Nicholson, Damelin, Robin, and Sears [5, Definition 1], who state that ID is the dimensionality of the signal subspace. In [12], ID is simply defined as the number of endmembers, which is also the definition that we use in this paper.

Assuming that the linear mixture model (which of course is an approximation to reality) holds, we propose an alternative concept, which we call Effective Intrinsic Dimensionality (EID). It reflects the idea that, in the presence of errors, some signals in the data may be so weak as to be undetectable, and so is a means of bridging the gap between the definitions in [2] and [5]. When the band error variances are all equal, to σ^2 say, then [26, eq. (11)] and [5, eq. (8)] state that there is a threshold value (called the "asymptotic limit of detection" by the former), below which signal eigenvalues cannot be successfully identified, at least asymptotically. This value is $\sigma^2\sqrt{d/N}$. The proof of this result is contained in [27, Th. 1.1].

TABLE I
Various ID Estimates for the Real Indian Pines and Cuprite Images

Scene	HFC	NWHFC	NSP	HySime	RMT
Indian Pines	29, 28, 27	18, 18, 17	62, 60, 59	11	19
Cuprite	36, 29, 22	26, 24, 22	37, 37, 37	16	29

This concept is easily generalised to the case where the errors have unequal variances and/or are spectrally correlated. Before we do this, we need to make some definitions and make two mild assumptions. Let

$$S_i = \sum_{j=1}^{M} w_{\mathrm{ij}} E_j, \quad i = 1, \ldots, N, \tag{10}$$

denote the "signal" component of (1).

Let Σ_S and Σ_ϵ denote the expected values of the signal and error covariance matrices, respectively.

Assumption 1: Σ_ϵ is positive definite. Of course, if Σ_ϵ is diagonal, all of whose diagonal entries are "strictly" positive (which we have assumed in this paper), then it will be positive definite.

Assumption 2: Σ_S is positive semidefinite.

Equation (10) implies that the rank of Σ_S is $M(<d)$. However, this is not needed for what follows.

The two assumptions imply that there exists a $d * d$ matrix A satisfying

$$A^T \Sigma_S A = \Lambda, \quad A^T \Sigma_\epsilon A = I \tag{11}$$

where Λ is a diagonal matrix with nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ [28, Th. 12.2.13]. Of course, if the rank of Σ_S is M, then $\lambda_j = 0, j = M+1, \ldots, d$. When Σ_ϵ is diagonal, the simultaneous diagonalization (11) is achieved by simply dividing the data by the band error SDs and then applying a standard eigendecomposition.

The transformation (11) has converted the problem of unequal error variances to one with error variances all equal to 1. So the "asymptotic limit of detection" becomes

$$\lambda_{\mathrm{crit}} = \sqrt{d/N}. \tag{12}$$

We define the EID as the number of λ_j's greater than λ_{crit}. Of course, EID $\leq$ ID, because $\lambda_{\mathrm{crit}} > \lambda_{M+1} = 0$. We shall see that, in our simulations, for smaller values of $\hat{M}$, EID = ID, while for larger values of $\hat{M}$, EID $<$ ID, reflecting the fact that, as $\hat{M}$ becomes larger, the additional endmembers represent signals in the real scene, which are 1) minor variations of other more ubiquitous endmembers, 2) rare, or 3) nonexistent altogether.

V. Simulations

In order to decide on a plausible range of values of $\hat{M}$, we first apply the five ID estimation methods described in Section III to the real datasets. For HFC, NWHFC, and NSP, we give the results for $P_f = 10^{-3}, 10^{-4}$, and 10^{-5} in that order. The results are shown in Table I.

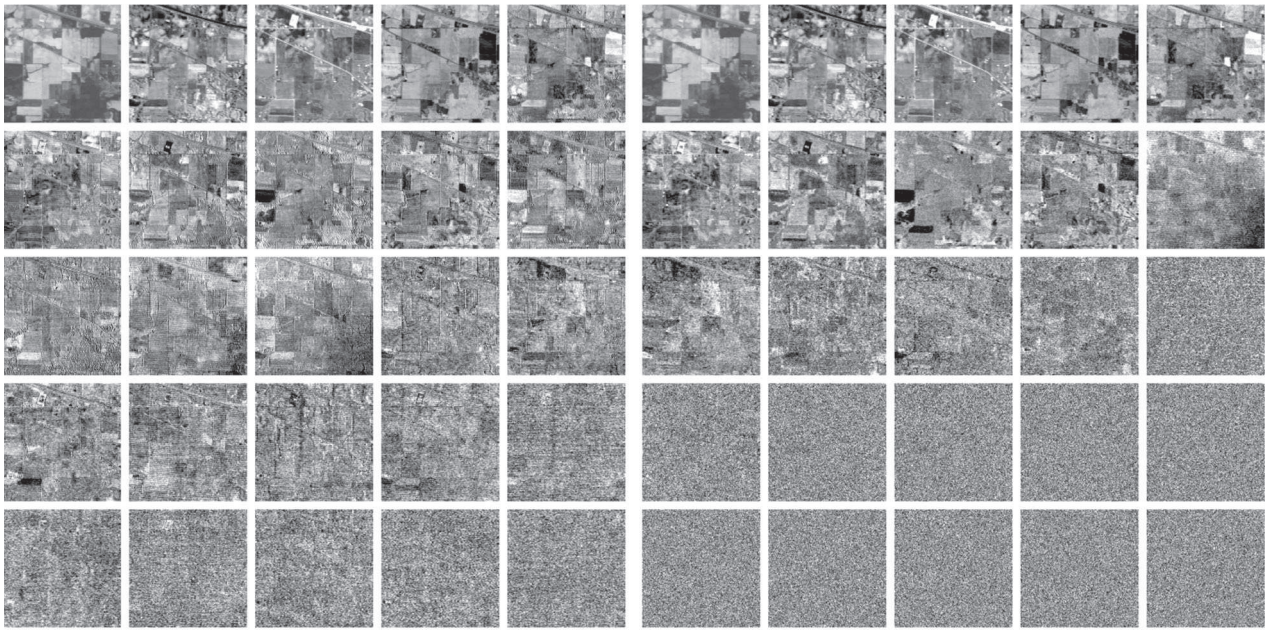


Fig. 6. Indian Pines: Real and simulated scaled PCs ($\hat{M} = 20$). (a) Real scaled PCs. (b) Simulated scaled PCs ($\hat{M} = 20$).

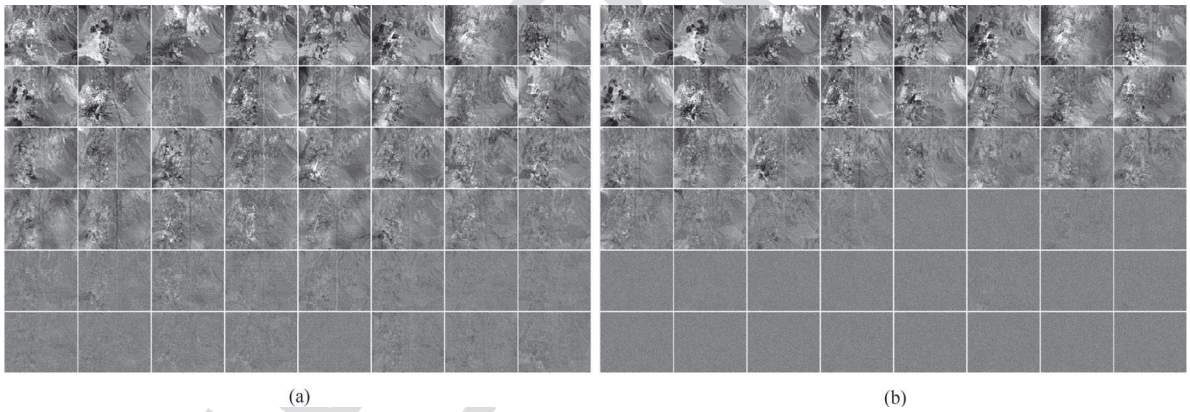


Fig. 7. Cuprite: Real and simulated scaled PCs ($\hat{M} = 36$). (a) Real scaled PCs. (b) Simulated scaled PCs ($\hat{M} = 36$).

Our initial assumption is that, for each image, the true ID is somewhere in the range of estimates in Table I. However, the ranges are very large: 11–62 for Indian Pines and 16–37 for Cuprite. HySime gives the lowest estimates and NSP gives the highest estimates for both datasets. This is a great concern because the papers which introduce these two methods, [2] and [3], are probably the two most highly cited papers on ID estimation in the hyperspectral literature. The NSP estimates for Indian Pines (62, 60, and 59) are particularly hard to believe; see the discussion in Section III-C for a possible explanation of this extreme behavior. Low HySime estimates were also obtained by Robin, Cawse-Nicholson, Mahmood, and Sears [16], when applied to several real datasets (different from those analyzed here), but not in their simulated datasets, which used N-FINDR [17] to generate their endmembers. Another issue which this table highlights is the difficulty in deciding on the appropriate value of P_f for the three VD methods. Chang and Du [2] use $P_f = 10^{-3}$, while Robin, Cawse-Nicholson, Mahmood, and Sears [16, Sec. II.C] state: "We have confirmed the stability of this value in preliminary experiments, as using 10^{-3} or 10^{-4} made no difference in the results." The HFC estimates for the Cuprite scene are inconsistent with their observations, suggesting potential problems with HFC at least.

In order to assist us with narrowing the range of plausible values of $\hat{M}$, Figs. 6(a) and 7(a) show the first 25 and 48 "scaled" PCs for the Indian Pines and Cuprite scenes, respectively (i.e., after dividing each band by Roger's error SD estimate). Each PC image has been linearly stretched over the range of its mean +/− 2.5 SDs, so that any signal is apparent.

Real signal is apparent in perhaps the first 20 Indian Pines PCs, which is why we have chosen this value of $\hat{M}$ for illustrative purposes. However, there is some form of horizontal

instrumental noise in later PCs (which our model will treat as "signal" because it is spatially correlated), as well as some vertical and "speckly" features in some of the earlier PCs. It is difficult to assess whether these features are in the scene or are instrumental effects. Fig. 6(b) shows the simulated scaled PCs when $\hat{M} = 20$. Signal is apparent in the first 14 PCs (which suggests that $\hat{M} = 20$ may be too small). Of these, the first nine are similar to the corresponding real images. The horizontal, vertical, and many of the speckle features are absent from the simulated PCs, which highlights the need to incorporate such "artefacts" in mixture models. However, their absence from the simulated images makes it easier to compare the various ID methods under the model assumptions of this paper. Based on our interpretation of Fig. 6(a), we have decided to omit the HySime and NSP estimates of the real Indian Pines dataset and simulate it with $\hat{M}$ varying between 17 and 29.

Strong signal is apparent in perhaps the first 31 or 32 Cuprite PCs. However, there are weaker "spatially coherent" signals apparent in most of the remaining PCs, up to PC 48. Perhaps these weaker signals are due to local variants of some endmembers and/or small nonlinearities in the mixing. However, from the point of view of a linear mixture model, they are real signals. There are no apparent nonrandom instrumental effects. Based on our interpretation of Fig. 7(a), we have decided to omit the much smaller HySime estimate of the real Cuprite dataset and simulate it with $\hat{M}$ varying between 22 and 37. Fig. 7(b) shows the simulated scaled PCs when $\hat{M} = 36$. This allows for the strong signals plus a few weak ones. Signal is apparent in the first 27 or 28 PCs, although there is a faint signal in PC 31. Again, this suggests that $\hat{M} = 36$ may be too small. The first 14 PCs of the real and simulated images are similar.

Fig. 8(a) and (b) shows the mean (± 2 SDs) for the five ID estimates versus $\hat{M}$ for the simulated Indian Pines and Cuprite images. Following [2] and [16], we use $P_f = 10^{-3}$ for HFC, NWHFC, and NSP. The first thing to note is that, for smaller values of $\hat{M}$, EID = ID. However, as $\hat{M}$ becomes larger, EID $<$ ID, probably due to the reasons given at the end of Section IV.

For the Indian Pines simulations, all five estimators are almost independent of $\hat{M}$. This probably reflects the fact that the true ID is at the lower end of the range of values of $\hat{M}$ chosen, and the inability of any of the estimators to detect weak or rare signals in the simulated (and possibly real) data. The mean estimates of the simulations of each of four of the five ID estimation methods are in the same order as their corresponding estimates for the real data (NSP $>$ HFC $>$ NWHFC $>$ HySime). HFC and NWHFC are on average about the same as for the real data, while HySime is a little smaller. NSP drops from 62 to the high 40s (which is still far too high), while RMT drops from 19 to a range of 10–12. The drop in the HySime, NSP, and RMT estimates may, at least in part, be due to the positive bias in Roger's band estimators. Because HFC and NWHFC compare the eigenvalues of covariances with and without mean correction, they are perhaps somewhat less sensitive to this bias, whose effects may be about the same on both sets of eigenvalues.

The most striking feature for the Cuprite simulations is that the mean values of the HFC estimates are much too high for $\hat{M} < 32$, but very good for $\hat{M} \geq 32$ (although its SDs are very

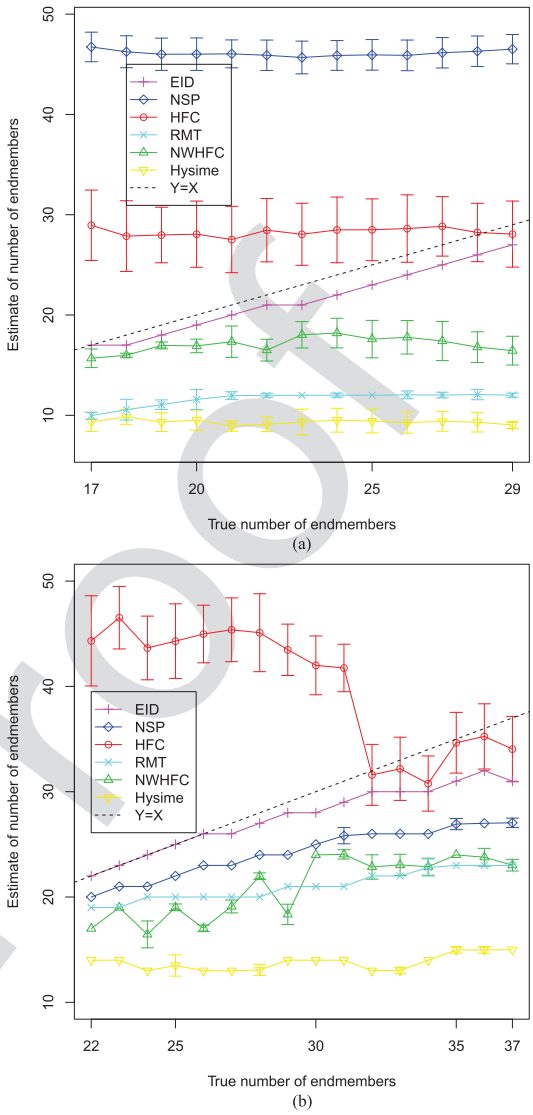


Fig. 8. Mean (± 2 SDs) for five ID estimates versus $\hat{M}$ for simulated Indian Pines and Cuprite images with variable SNR ($P_f = 10^{-3}$). (a) Indian Pines. (b) Cuprite.

large). We have no explanation for this as yet. All the remaining estimates lie below EID, but gradually increase with $\hat{M}$, probably reflecting the fact that the true ID is at the upper end of the range of values of $\hat{M}$ chosen. If one restricts attention to values of $\hat{M}$ above 32, the mean estimates of simulated HFC estimates comes closest to the corresponding estimate for the real data (36), while the means of the remaining simulated estimators tend to be a little less than their counterparts for real data. This is consistent with the behavior seen with the simulated Indian Pines data.

Overall, HFC is clearly the most variable of the estimators. Because NWHFC is a whitened version of HFC, its variability is greatly reduced and is comparable to that of HySime. Both are a little more variable than RMT. NSP's variability appears to be a function of its mean.

The results in Fig. 8(a) and (b) are for variable SNR. We have also carried out simulations using constant SNR, whose values

are shown in Fig. 2(a) and (b). We do not plot the results here. Generally speaking, however, the estimators are a little larger than for their variable-SNR counterparts. For the Cuprite data, the HFC and NWHFC estimates are even more variable than are their variable-SNR counterparts.

VI. Conclusion and Future Work

The ten papers cited near the beginning of this paper (a number of which are highly cited) use simulations to introduce or test either 1) methods for estimating the ID of hyperspectral data, or 2) methods for estimating endmembers in such data. In our opinion, these simulations are so different from real-world data that one cannot be confident that the various advocated methods work well with real-world data. We have introduced a general methodology which aims to make the simulations more realistic. We have also explored various options within the general framework, including 1) comparing the use of variable and constant SNRs for the errors, and 2) comparing the MVT and ICE algorithms as inputs to the simulation method. We believe that by attempting to produce more realistic simulations, we have obtained greater insights into the strengths and weaknesses of various advocated methods. In particular, we have observed that Roger's error variance estimates are positively biased. This has led us to prove (in Appendix A) that this is always so. We have also demonstrated that the average relative bias of Roger's error variance estimates is an increasing function of M/d and developed some theory in Appendix B to support this.

Although there are differences between the ID estimates for the real images and their simulated counterparts, they are mostly consistent. In addition, our approach strongly suggests that HFC and NSP are highly variable (and hence unreliable) ID estimators. On the other hand, NWHFC, HySime, and RMT appear to underestimate ID (or even EID). This may in part be due to the bias in Roger's variance estimates. So a first goal is to see if this bias can be corrected.

Toward this end, Mahmood, Robin, and Sears [29] have recently developed a first-order correction to Roger's error SD estimate. We have applied this correction to three real and simulated hyperspectral images (including the two presented here). For the three simulated images, although the bias of the corrected estimates is certainly reduced, it still appears to be a little positive. In addition, for the real image not presented here (which has worse artefacts than those in the Indian Pines scene), some of the bands have very small error variances. Unfortunately, the first-order correction produces negative estimates of some of these. So, although promising, the new method needs some further development, which we intend to investigate.

Depending on the results of these investigations, modifications of some of the above ID estimation methods (or possibly entirely new ones) may need to be developed to make them more accurate than Fig. 8(a) and (b) suggest that they are.

Recently, Robin, Cawse-Nicholson, Mahmood, and Sears [16] and Cawse-Nicholson, Robin, and Sears [30] have investigated the estimation of ID in the presence of correlated noise, which is known to have a significant presence in some instruments. They used various multivariate and spatial methods to estimate the ID. It is our intention to investigate these methods, again with the aid of semirealistic simulations of natural hyperspectral scenes, and if necessary either to improve them further or to develop new methods for correlated errors.

Another issue that has not been addressed in this paper is sparsity. Once the endmembers have been estimated, one needs to estimate the proportions. In our simulations, we have fitted the spectrum in each pixel using LS fitting in MNF space with the constraints that the weights are nonnegative and sum to 1. However, even though this constraint forces some of the weights to be zero, it generally leaves too many endmembers with small (positive) weights. In recent years, "sparse unmixing" techniques have been introduced [31]–[33]. These force many of these small weights to be zero. It will be interesting to see what impact "sparsity" has on the simulated data. In particular, for large $\hat{M}$, the *maximum* proportion of some of the endmembers in a simulated scene will be very small. Forcing sparsity may push the maximum proportion of some of these endmembers to zero, thus reducing the EID.

Appendix A
Proof of Positive Bias of Roger's Band SD Estimators

Before we prove that Roger's band SD estimators are always positively biased, we need to carefully state our assumptions, which are fairly standard. Because we are assuming that the error covariance matrix is diagonal, we need to modify Assumption 1 appropriately.

Assumption 1a: Σ_ϵ is diagonal, all of whose diagonal entries, $\sigma^2_{\epsilon,j}, j = 1, \ldots, d$, are "strictly" positive.

Equation (10) implies that the rank of Σ_S is $M(<d)$. We, therefore, modify Assumption 2 to the following.

Assumption 2a: Σ_S is positive semidefinite with (unknown) rank $M(<d)$.

In addition, we assume the following.

Assumption 3: S_i and ϵ_i are uncorrelated.

Let Σ_X denote the expected value of the data. It follows from (1), (10), and Assumption 3 that

$$\Sigma_X = \Sigma_S + \Sigma_\epsilon. \tag{13}$$

Let $\bar{X} = \sum_{i=1}^N X_i/N$, let $\hat{\Sigma}_X = \sum_{i=1}^N X_i X_i^T/N$ denote the sample covariance matrix of the observed data, and let v_j denote the jth diagonal entry of $\hat{\Sigma}_X^{-1}$. Quoting other references, Roger [13, Sec. 2.3.1] pointed out that

$$v_j^{-1} = \hat{\sigma}^2_{\epsilon,j}, \quad j = 1, \ldots, d. \tag{14}$$

Either this formula or (4) can be used to provide Roger's estimator of the band error variances (and hence their SDs).

The flaw in Roger's approach is easily seen from the regression interpretation, on which (4) is based. When regressing one band on the remaining $d-1$ bands, the $d-1$ explanatory variables themselves have errors. This is the cause of the bias, and as it happens, it is always positive.

In the proof that follows, we make no distinction between the theoretical covariance matrices in (13) and their sample versions. The differences are negligible if N is large enough.

We do not know how large N needs to be, but it will certainly be the case for most real-world datasets.

Proof: It follows from (13) and (11) that

$$\Sigma_X^{-1} = A(\Lambda + I)^{-1} A^T \tag{15}$$

and

$$\Sigma_\epsilon^{-1} = AA^T. \tag{16}$$

Let $a_j^T \equiv (a_{j1}, \ldots, a_{\mathrm{jd}})$ denote the jth row of A. It then follows from (16) and Assumption 1 a that

$$a_j^T a_j = \sum_{k=1}^{d} a_{\mathrm{jk}}^2 = \sigma_{\epsilon,j}^{-2}. \tag{17}$$

Let e_j denote the d-vector, which has zeroes everywhere, except for the jth position, which equals 1. It then follows from the definitions of v_j, a_j, (15), and (17) that

$$\begin{aligned} v_j &= e_j^T \Sigma_X^{-1} e_j \\ &= e_j^T A(\Lambda + I)^{-1} A^T e_j \\ &= a_j^T (\Lambda + I)^{-1} a_j \\ &= \sum_{k=1}^{d} a_{\mathrm{jk}}^2/(\lambda_k + 1) \qquad (18) \\ &< \sum_{k=1}^{d} a_{\mathrm{jk}}^2 \qquad (19) \\ &= \sigma_{\epsilon,j}^{-2} \qquad (20) \end{aligned}$$

by (17). Invert both sides of this inequality and use (14) to obtain

$$\hat{\sigma}_{\epsilon,j}^2 > \sigma_{\epsilon,j}^2, \quad j = 1, \ldots, d. \tag{21}$$

Note that the strict inequality (19) holds provided that the largest eigenvalue is positive.

Appendix B
Derivation of Inequalities (7) and (8)

At this point, it will be convenient to let

$$\Sigma_S^* = \Sigma_\epsilon^{-1/2} \Sigma_S \Sigma_\epsilon^{-1/2}, \Sigma_X^* = \Sigma_\epsilon^{-1/2} \Sigma_X \Sigma_\epsilon^{-1/2}. \tag{22}$$

These are just the covariance matrices of the signal and data respectively after each band has been divided (i.e., "scaled") by the true error SD in each band. Then, (13) becomes

$$\Sigma_X^* = \Sigma_S^* + I \tag{23}$$

and (11) can be converted into a "standard" eigendecomposition, in which the band error variances all equal 1:

$$B^T \Sigma_S^* B = \Lambda, \quad B^T B = BB^T = I \tag{24}$$

where $B = \Sigma_\epsilon^{1/2} A$. Note that, because B is orthogonal, (24) includes one more equality than (11) does.

The analog of (15) is

$$(\Sigma_X^*)^{-1} = B(\Lambda + I)^{-1} B^T. \tag{25}$$

Take the trace of both sides of (25) and use its cyclic property and the last equality in (24) to obtain

$$\begin{aligned} tr((\Sigma_X^*)^{-1}) &= tr((\Lambda + I)^{-1}) \\ &= \sum_{k=1}^{d} (\lambda_k + 1)^{-1} \\ &> d - M \qquad (26) \end{aligned}$$

because Assumption 2a implies that the last $d - M$ λ_k's are all zero. However, it follows from the second equality in (6), (14), and (22) that the left-hand side of (26) is just $\sum_{j=1}^{d} r_j^{-1}$. Substitute this into (26) and divide both sides by d to obtain (8).

To obtain inequality (9), let $(b_{j1}, \ldots, b_{\mathrm{jd}})$ denote the jth row of B. Using an analogous argument leading to (6), (14), and (17), we obtain

$$\begin{aligned} r_j^{-1} &= \sum_{k=1}^{d} b_{\mathrm{jk}}^2/(\lambda_k + 1) \\ &> \sum_{k=M+1}^{d} b_{\mathrm{jk}}^2 \end{aligned}$$

again because the last $d - M$ λ_k's are all zero. Invert this inequality to obtain

$$r_j < \left(\sum_{k=M+1}^{d} b_{\mathrm{jk}}^2 \right)^{-1}. \tag{27}$$

In order to make further progress, we need to make a distributional assumption about the b_{jk}'s. Note first that, analogous to (17), we have from the second equality in (24) that

$$\sum_{k=1}^{d} b_{\mathrm{jk}}^2 = 1. \tag{28}$$

We make the following (Bayesian) assumption:

Assumption 4: For each j, conditional on (28), $b_j^T \equiv (b_{j1}, \ldots, b_{\mathrm{jd}})$ is uniformly distributed on the d-dimensional sphere of radius 1.

Under this assumption, $\sum_{k=M+1}^{d} b_{\mathrm{jk}}^2$ has a Beta distribution with parameters $\frac{1}{2}(d - M)$ and $\frac{1}{2}M$ [34, p. 49]. It is then straightforward to show that $E((\sum_{k=M+1}^{d} b_{\mathrm{jk}}^2)^{-1}) = (d - 2)/(d - M - 2)$. Substitute this into (27) to obtain (9).

References

[1] E. Ientilucci and S. Brown, "Advances in wide-area hyperspectral image simulation," *Proc. SPIE*, vol. 5075, pp. 110–121, 2003.

[2] 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.

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

[4] N. Acito, M. Diani, and G. Corsini, "A new algorithm for robust estimation of the signal subspace in hyperspectral images in the presence of rare signal components," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3844–3856, Nov. 2009.

[5] K. Cawse-Nicholson, S. Damelin, A. Robin, and M. Sears, "Determining the intrinsic dimension of a hyperspectral image using random matrix

theory," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1301–1310, Apr. 2013.
[6] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 3, pp. 765–777, Mar. 2007.
[7] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4418–4432, Nov. 2009.
[8] M. Zortea and A. Plaza, "Spatial preprocessing for endmember extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2679–2693, Aug. 2009.
[9] A. Huck, M. Guillaume, and J. Blanc-Talon, "Minimum dispersion constrained nonnegative matrix factorization to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 6, pp. 2590–2602, Jun. 2010.
[10] B. Somers, M. Zortea, A. Plaza, and G. Asner, "Automated extraction of image-based endmember bundles for improved spectral unmixing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 396–408, Apr. 2012.
[11] J. Bioucas-Dias *et al.*, "Hyperspectral unmixing overview: Geometrical, statistical and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, Apr. 2012.
[12] J. Bioucas-Dias and J. Nascimento, "Estimation of signal subspace on hyperspectral data," *Proc. SPIE*, vol. 5982, pp. 191–198, 2005.
[13] R. Roger, "Principal Components transform with simple automatic noise adjustment," *Int. J. Remote Sens.*, vol. 17, pp. 2719–2727, 1996.
[14] P. Meer, J.-M. Jolion, and A. Rosenfeld, "A fast parallel algorithm for blind estimation of noise variance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 2, pp. 216–223, Feb. 1990.
[15] L. Gao, Q. Du, B. Zhang, W. Yang, and Y. Wu, "A comparative study on linear regression-based noise estimation for hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 488–498, Apr. 2013.
[16] A. Robin, K. Cawse-Nicholson, A. Mahmood, and M. Sears, "Estimation of the intrinsic dimension of hyperspectral images: Comparison of current methods," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2854–2861, Jun. 2015.
[17] M. E. Winter, "N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data," *Proc. SPIE* vol. 3753, pp. 266–275, 1999.
[18] A. Green, M. Berman, P. Switzer, and M. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65–74, Jan. 1988.
[19] M. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 3, pp. 542–552, May 1994.
[20] D. Fuhrmann, "A simplex shrink-wrap algorithm," *Proc. SPIE*, vol. 3718, pp. 501–511, 1999.
[21] J. Li and J. Bioucas-Dias, "Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Boston, MA, USA, 2008, vol. 3, pp. 250–253.
[22] M. Berman, H. Kiiveri, R. Lagerstrom, A. Ernst, R. Dunne, and J. Huntington, "ICE: A statistical approach to identifying endmembers," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 10, pp. 2085–2095, Oct. 2004.
[23] K. Staenz, A. Mueller, and U. Heiden, "Overview of terrestrial imaging spectroscopy missions," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Melbourne, Australia, Aug. 2013, pp. 3502–3505.
[24] S. Geman, "A limit theorem for the norm of random matrices," *Ann. Probab.*, vol. 8, pp. 252–261, 1980.
[25] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Ann. Statist.*, vol. 29, pp. 295–327, 2001.
[26] S. Kritchman and B. Nadler, "Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 3930–3941, Oct. 2009.
[27] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *J. Multivariate Anal.*, vol. 97, no. 6, pp. 1382–1408, 2006.
[28] F. Graybill, *Matrices with Applications in Statistics*, 2nd ed. Belmont, CA, USA: Wadsworth, 1983.
[29] A. Mahmood, A. Robin, and M. Sears, "Modified residual method for estimation of noise statistics in hyperspectral images," presented at the 7th Workshop Hyperspectral Image Signal Process.: Evol. Remote Sens., Tokyo, Japan, 2015.
[30] 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.
[31] M.-D. Iordache, J. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2014–2039, Jun. 2011.
[32] Y. Guo and M. Berman, "A comparison between subset selection and L1 regularisation with an application in spectroscopy," *Chemometrics Intell. Lab. Syst.*, vol. 118, pp. 127–138, 2012.
[33] J. B. Greer, "Sparse demixing of hyperspectral images," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 219–228, Jan. 2012.
[34] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, vol. 197. Hoboken, NJ, USA: Wiley, 2009.

Zhipeng Hao received the B.Sc. degree in mathematics from Nankai University, Tianjin, China in 2004, and the master's and Ph.D. degrees in statistics from Macquarie University, Sydney, Australia, in 2009 and 2015, respectively.

He is currently a Postdoctoral Research Fellow with the Centre for Research in Mathematics, Western Sydney University, Parramatta, Australia. His main research interests include hyperspectral data analysis, flow cytometry data analysis, and survival data analysis.

Mark Berman received the B.Sc. (Hons.) (University Medal) degree in mathematical statistics and the Master of Statistics from the University of New South Wales, Sydney, Australia, in 1974 and 1976, respectively, and Ph.D. and D.I.C. degrees in mathematical statistics from the Imperial College of Science and Technology, London, U.K., in 1978.

He was a Visiting Lecturer with the Department of Statistics, University of California, Berkeley, during 1978–1979. Between 1979 and 2014, he was with the CSIRO Division of Mathematics, Informatics and Statistics (CMIS), Sydney, where he reached the position of Chief Research Scientist. He led CMIS' Image Analysis Group from 1989 to 2000. In 1988, he took leave from CSIRO to establish the Image Processing and Data Analysis Group with the Melbourne Research Laboratories of Broken Hill Proprietary Ltd. He has given Ph.D. courses in spectroscopy and hyperspectral imaging with the Technical University of Denmark (2007) and Stanford University (2008 and 2014). He is currently an Adjunct Fellow with the Centre for Research in Mathematics, Western Sydney University, Sydney. He is the author of several patents. His research interests include image analysis (especially hyperspectral), spectroscopy, and spatial data analysis.

Dr. Berman was an Associate Editor of *Computational Statistics and Data Analysis* (2001–2006) and an Associate Editor of *Environmetrics* (2010–2015). He has received Best Paper Awards from the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 1990 and the *Journal of Chemometrics* in 2011.

Yi Guo received the B.Eng. (Hons.) in instrumentation from the North China University of Technology, Beijing, China, in 1998, the M.Eng. degree in automatic control from Central South University, Changsha, China, in 2002, and the Ph.D. degree in computer science, focusing on dimensionality reduction for structured data with no vectorial representation, from the University of New England, Armidale, Australia, in 2008.

From 2008 until 2016, he was with CSIRO, working as a computational statistician on various projects in spectroscopy, remote sensing, and materials science. He recently joined the Centre for Research in Mathematics, Western Sydney University, Sydney, Australia. His recent research interests include machine learning, computational statistics, and big data.

Glenn Stone received the B.A. degree in mathematics from the University of Oxford, Oxford, U.K., the M.Sc. degree in computer science from the University of Manchester, Manchester, U.K., and the Ph.D. degree in statistics from the University of Bath, Bath, U.K.

After a period of teaching statistics with the University of Bath, he moved to Australia and joined the CSIRO in 1993. There, he worked on the development and application of statistical methods for a wide range of areas, including spatial smoothing, data mining, insurance risk, and biostatistics. From 1999 to 2003, he was a Principal Research Analyst with Insurance Australia Group, returning to CSIRO in 2003 to work in bioinformatics and biostatistics and, more recently, in remote sensing. He joined Western Sydney University, Sydney, Australia, in 2011, where he is currently a Professor of Data Science. His research interests include computationally intensive statistical methods with applications in flow cytometry, remote sensing, ecogenomics, and wavelet methods for non-Gaussian data.

Iain Johnstone received the B.Sc. (Hons) degree in pure mathematics and statistics and the M.Sc. degree in statistics from the Australian National University, Canberra, Australia, in 1977 and 1978, respectively, and the Ph.D. degree in statistics from Cornell University, Ithaca, NY, USA, in 1981.

Since 1981, he has been an Assistant, Associate, and then Full Professor with the Department of Statistics, Stanford University, Stanford, CA, USA. Since 1989, he has also held a 50% time appointment in biostatistics with the Stanford University School of Medicine. His research in theoretical statistics has used ideas from harmonic analysis, such as wavelets, to understand noise-reduction methods in signal and image processing. More recently, he has applied random matrix theory to the study of high-dimensional multivariate statistical methods, such as principal components and canonical correlation analysis. In biostatistics, he has collaborated with investigators in cardiology and prostate cancer.

Dr. Johnstone is a member of the U.S. National Academy of Sciences and the American Academy of Arts and Sciences and a former president of the Institute of Mathematical Statistics.

Semi-realistic Simulations of Natural Hyperspectral Scenes

Zhipeng Hao, Mark Berman, Yi Guo, Glenn Stone, and Iain Johnstone

***Abstract*—Many papers in the hyperspectral literature use simulations (based on a linear mixture model) to test algorithms, which either estimate the "intrinsic" dimensionality (ID) of the data or endmembers. Usually, these simulations use "real-world" endmembers, proportions distributed according to a uniform or Dirichlet distribution on the endmember simplex, and Gaussian errors which are "spectrally" and "spatially" uncorrelated. When the error standard deviations (SDs) in different bands are assumed to be unequal, they are usually estimated using Roger's method. The simulated and real-world data in these papers are so different that one cannot be confident that the various advocated methods work well with real-world data. We propose a general methodology which produces more realistic simulations, providing us with greater insights into the strengths and weaknesses of various advocated methods. With the aid of the well-known Indian Pines and Cuprite scenes, we compare several specific options within the proposed methodological framework. We also compare the performance of five well-known ID estimators using both real and simulated datasets and demonstrate that Roger's SD estimates are positively biased. A proof that Roger's estimates are always positively biased is given.**

***Index Terms*—Dimensionality, endmembers, hyperspectral, linear mixture model, simulation.**

I. Introduction

SIMULATION is useful for assessing the performance of algorithms in many fields. This is especially true of hyperspectral remote sensing data, where ground-based validation of such algorithms is typically costly and time consuming. The hyperspectral remote sensing literature is dominated by two broad simulation approaches. The first, captured in the Digital Imaging and Remote Sensing Image Generation model, uses sophisticated radiometric and geometric models of real-world scenes to reconstruct closely those scenes [1]. This approach is particularly useful for modeling 3-D objects such as buildings and trees.

The second approach is not based on real-world scenes. It is found in the hyperspectral literature concerned with linear mixture models. These papers are concerned with either 1) estimating the "intrinsic" dimensionality of the data [2]–[5], 2) introducing a new blind unmixing/endmember estimation algorithm [6]–[10] or 3) both [11]. All the simulations are based on the linear mixture model, which we now outline. Let $X_i, i = 1, ..., N$, denote the d-dimensional vector of observations at pixel i (out of N) in a hyperspectral image. Under the linear mixture model, if there are $M(<N)$ spectrally distinct materials in the image, then

$$X_i = \sum_{j=1}^{M} w_{\mathrm{ij}} E_j + \epsilon_i, \quad i = 1, \ldots, N \tag{1}$$

where (i) $E_j, j = 1, \ldots, M$, are the "endmembers" (i.e., pure materials); (ii) w_{ij} are *nonnegative* weights; and (iii) ϵ_i are error terms. The errors are typically a combination of instrumental noise, natural variation in spectra representing the same material, and small nonlinearities in the mixing. M is sometimes called "virtual dimensionality" (VD) [2] and sometimes called "intrinsic dimensionality" (ID) [5], [12]. We shall use the latter term.

Often the weights in (1) are also constrained to be proportions, i.e.,

$$\sum_{j=1}^{M} w_{\mathrm{ij}} = 1, \quad i = 1, \ldots, N. \tag{2}$$

For ease of exposition, we will assume that (2) is satisfied throughout the paper. Small modifications are required when it is not satisfied. Note however that, when it is satisfied, the M endmembers define a simplex in an $(M-1)$-dimensional subspace. For instance, when $M = 3$, the endmembers define a triangle in a 2-D subspace.

The simulations in the ten above-mentioned references vary according to how the three components are chosen. For (i), all use "real-world" endmembers, usually chosen from the USGS or JPL spectral libraries, or other sources [2], [4], [10]. For (ii), the proportions are mostly chosen to be spatially uncorrelated. The proportion distribution is either "random" [2], [5], [10], distributed uniformly [4], [11], or according to a Dirichlet distribution on the simplex defined by the endmembers [3], [7], [9]. In two of these papers [9], [11], the maximum proportions of some endmembers are restricted to being less than 1, modeling the common situation where some materials in a scene are never pure. Papers [6] and [8] incorporate some spatial information. For (iii), all assume Gaussian errors (as shall we in our simulations). Most assume that the errors are "spectrally" and "spatially" uncorrelated. Papers [6]–[11] all assume that the error standard deviations (SDs) in all the bands are equal. Papers [2]–[5] assume that they are different. The first three

Manuscript received July 02, 2015; revised December 28, 2015; accepted May 26, 2016. *(Corresponding author: Mark Berman.)*

Z. Hao, Y. Guo, and G. Stone are with the School of Computing, Engineering and Mathematics, Western Sydney University, Parramatta, NSW 2150, Australia (e-mail: z.hao@westernsydney.edu.au; y.guo@westernsydney.edu.au; g.stone@westernsydney.edu.au).

M. Berman is with the School of Computing, Engineering and Mathematics, Western Sydney University, Parramatta, NSW 2150, Australia, and also with CSIRO Data61, North Ryde, NSW 2113, Australia (e-mail: mark.berman@csiro.au).

I. Johnstone is with the Department of Statistics, Stanford University, CA 94305-4065 USA (e-mail: imj@stanford.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2016.2580178

papers use a (nonspatial) multivariate method suggested by Roger [13] to estimate the SDs, while paper [5] uses a method based on finding homogeneous areas in the scene [14]. For nine of the above-mentioned references, M varies between 3 and 10; one of the simulations carried out by Bioucas-Dias and Nascimento [3] has $M = 15$. For the ten references, N varies between 150 and 10 000.

The above-mentioned papers apply their advocated dimensionality or endmember estimation methods to one or more real-world scenes, in most cases one of the AVIRIS scenes of the Cuprite mining district, but also other scenes [2], [3], [5], [10]. Apart from one of the real-world datasets analyzed by Somers, Zortea, Plaza, and Asner [10] (for which $N = 2700$), the number of pixels in the remaining real-world datasets varies between 21 025 and 122 500, so they are considerably larger than the simulated data sets. The estimated values of M, using one or more of the methods presented in [2]–[5], are, with two exceptions (see [2] and [6]), all greater than 10 and in some cases greater than 20. This reflects the obvious principle that larger (real-world) datasets tend to contain more materials. In this age of "big data," M will tend to become bigger as real-world datasets become bigger! This illustrates obvious differences between the simulated and real-world datasets in the above-mentioned papers. However, the endmembers and the spatial distribution of the proportions also differ. Therefore, in our opinion, the simulated and real-world data in these papers are so different that we cannot really be confident that the various advocated methods work well with real-world data.

In this paper, we propose a general simulation methodology which, while not as sophisticated as a physics-based approach [1], produces simulations which are more like real-world data and provide us with greater insights into the strengths and weaknesses of various ID and endmember estimation methods. We believe that the methodology is potentially useful for simulating "natural" scenes (i.e., those dominated by vegetation, minerals and soils, rather than buildings), where the main interest is in knowing what materials are in a scene, and where and how much they are present. The general methodology is outlined in Section II. Various options within the general framework are also considered. This includes a comparison with methodologies in two recent papers [15], [16], which are similar to ours, but which do not go as far as we propose. In Section III, we describe five well-known ID estimation methods, all of which rely on various eigendecompositions of the data, while in Section IV, we introduce another concept, which we call "effective intrinsic dimensionality" (EID). In some cases, some of the "signal" eigenvalues are smaller than the "noise" eigenvalues. It is unreasonable to expect an eigenvalue-based ID estimation method to identify these. EID represents this concept mathematically. In Section V, we apply the five ID estimation methods to two real-world datasets and to simulated versions of them over a plausible range of values of M, and discuss the results. General conclusions are drawn and future work is discussed in Section VI.

One consequence of our "real-world" approach to simulation is the observation that Roger's method for estimating the band error SDs is positively biased. In Appendix A, we prove that this is always so. In Appendix B, we also develop some theory, which suggests that the average bias depends on the ratio M/d. As this ratio becomes larger, the bias also tends to become larger. The theory is supported by our simulations.

II. General Methodology and Some Options

The general methodology is straightforward. One chooses a real-world scene of interest and applies 1) a method to estimate its dimensionality ($\hat{M}$) and 2) a method to produce $\hat{M}$ endmember estimates ($\hat{E}_j, j = 1, \ldots, \hat{M}$). One can then estimate the weights, $\hat{w}_{\mathrm{ij}}$, usually by least-squares (LS) estimation (with constraints determined by the assumed weight constraints). All these estimates can be plugged into (1) to produce an estimated "true" or "target" image. Finally, the simulated image is produced by adding simulated errors δ_i, following the assumed error distribution, to the "target" image:

$$Y_i = \sum_{j=1}^{\hat{M}} \hat{w}_{\mathrm{ij}} \hat{E}_j + \delta_i, \quad i = 1, \ldots, N. \tag{3}$$

A similar approach to ours is adopted by Gao, Du, Zhang, Yang, and Wu [15]. They apply one of the three VD methods introduced in [2] (but do not state which one) to a 500×500 AVIRIS Cuprite subscene to estimate M and estimate the endmembers using N-FINDR [17] in minimum noise fraction (MNF) space [18]. They then produce the target image in a similar manner to ourselves. Spectrally uncorrelated errors are then added to the target image, using signal-to-noise ratios (SNRs), which are constant across the bands. They compare a number of error ("noise") estimation methods using two metrics based on the difference between the target and estimated errors at each pixel. They do not estimate band error variances themselves.

In [16], N-FINDR [17] is applied to a 350×350 AVIRIS Cuprite subscene, given M, which ranges from 5 to 40 in their simulations. They do not specify whether N-FINDR is run in MNF or some other space. They add spectrally uncorrelated errors, whose variances are both constant and variable. They also investigate the impact of correlated errors. We will not investigate correlated errors in this paper, although we do intend to investigate these in the future. The focus of [16] is to compare the performance of different ID estimation methods using both simulated and real data.

Although our general methodology is somewhat similar to those of [15] and [16], our philosophy is somewhat different. Our objective is to make our simulations as close as possible to real-world scenes by 1) visual inspection of principal component (PC) images (after "whitening" the data), and 2) making the eigenvalues of the real and simulated images as similar as possible. The second aim is particularly important because methods such as VD [2], RSSE [4], and random matrix theory (RMT) [5] are all based on the eigenvalues. If we can make the eigenvalues of real-world and simulated images as close possible, especially for indices equal to and near the true ID, then we are in a stronger position to compare different dimensionality estimation techniques with real data.

With our objective in mind, the approaches of [15] and [16] raise the following questions: 1) Do the errors in real-world datasets have either constant variance or constant SNRs across bands? 2) How important is the space in which the endmembers are found? 3) How important is the endmember estimation algorithm used? We will at least in part address questions 1 and 3 here, with the aid of two well-known scenes: the relatively small 145×145 Indian Pines scene ($N = 21\,025$) and the much larger 512×614 Cuprite scene f970619t01p02_r02_sc03.a.rfl ($N = 314\,368$). For both scenes, some bands have been excluded due to water absorption or noise. For Indian Pines, bands 104–110, 149–165, and 218–220 have been excluded, leaving 193 (out of 220) bands. For Cuprite, bands 1, 2, 104–114, 153–174, and 221–224 have been excluded, leaving 185 (out of 224) bands.

A. Error Variance Assumptions

Various methods have been proposed for estimating band error variances. In a forthcoming paper, using simulations similar to those described here, we demonstrate that Roger's method gives much better estimates of error variances than do a variety of spatial methods. This is consistent with observations in [3] and [16] and results in [15]. Roger's method [13] performs a linear regression of each of the d bands on the remaining $d-1$ bands and estimates the variance of each band using the residuals from the fits. Specifically, let $\hat{\epsilon}_{ij}, i = 1, \ldots, N; j = 1, ..., d$, denote the error in the ith pixel when regressing band j on the remaining $d-1$ bands. Roger's estimator of the error variance in band j is

$$\hat{\sigma}^2_{\epsilon,j} = \sum_{i=1}^{N} \hat{\epsilon}^2_{ij}/N, \quad j = 1, \ldots, d. \tag{4}$$

Some properties of this estimator are given in Section II-C. Technical details are given in the Appendixes.

Fig. 1(a) and (b) shows Roger's estimates of the band error SDs for the Indian Pines and Cuprite images. We also show the mean (± 2 SDs) for the SDs of the corresponding simulated images. For Indian Pines, 100 simulations have been used with $\hat{M} = 20$, while for the larger Cuprite scene, only 40 simulations have been used with $\hat{M} = 36$. These values of $\hat{M}$ have been chosen for illustrative purposes only and as initial approximate estimates of the true value of M. Further reasons for choosing these values are given in Section V. The main point here is to show that the error variances are not constant.

What about the assumption of constant SNR? We will use the following common definition for the SNR in band j:

$$SNR_j = 10 \log_{10}\{\text{Var}(s_j)/\text{Var}(\epsilon_j)\} \tag{5}$$

where $\text{Var}(s_j)$ and $\text{Var}(\epsilon_j) \equiv \sigma^2_{\epsilon,j}$ are the variances of the signal and error components, respectively for band j. The usual estimate of $\text{Var}(s_j)$ is just the sample variance in band j. We will use Roger's estimates of $\text{Var}(\epsilon_j)$, given by (4). The estimated SNRs for the Indian Pines and Cuprite datasets are shown in Fig. 2(a) and (b), respectively. In our simulations, we have assumed Gaussian errors, which are both spectrally and spatially uncorrelated, and with SNRs given by those in Fig. 2(a) and (b).

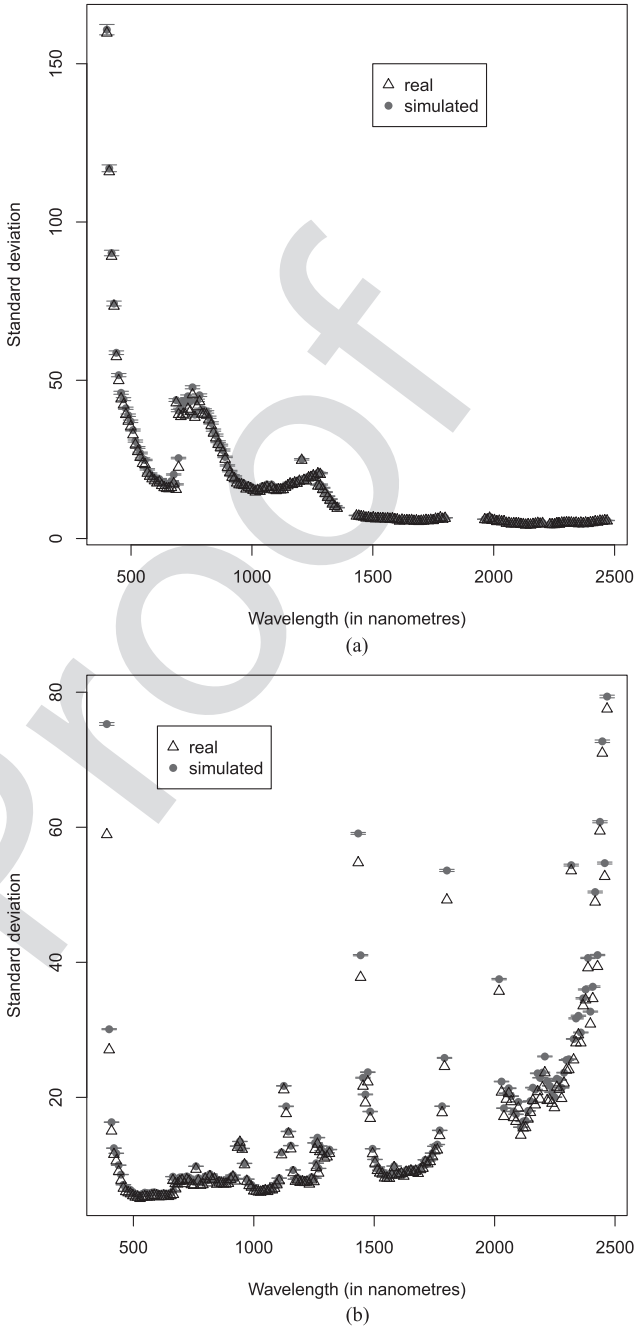
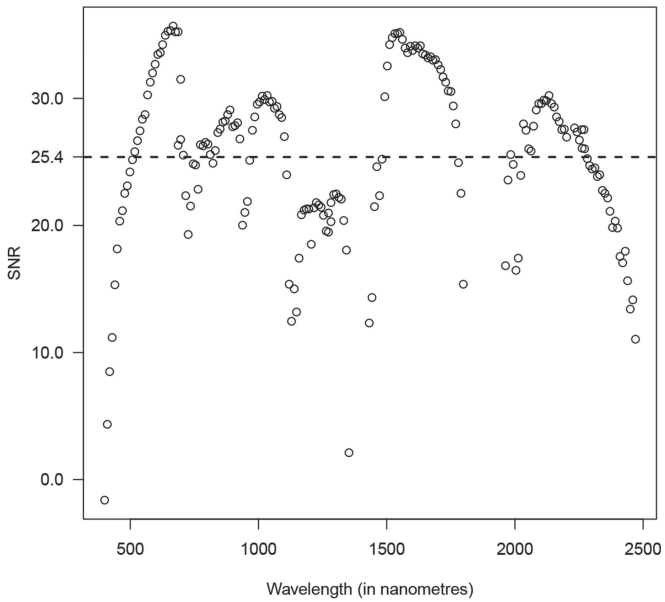


Fig. 1. SDs for real Indian Pines and Cuprite images, and mean (± 2 SDs) for SDs of corresponding simulated images. (a) Indian Pines ($\hat{M} = 20$, 100 simulations). (b) Cuprite ($\hat{M} = 36$, 40 simulations).

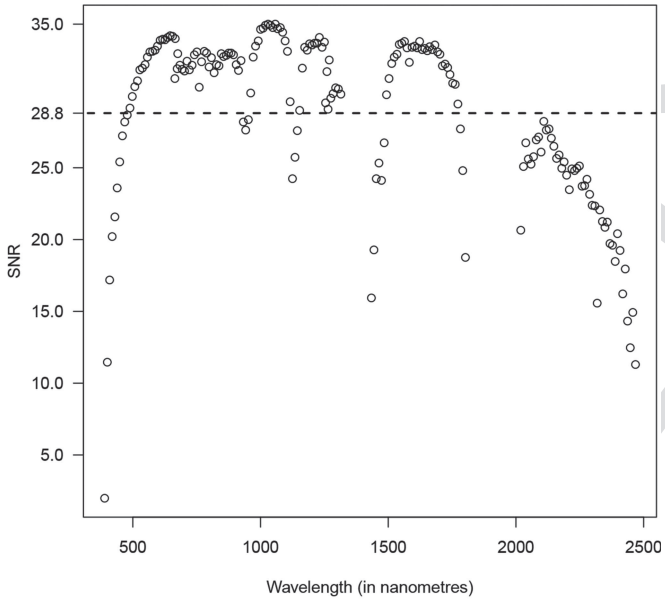
We will refer to these as "variable SNR" simulations. For comparison purposes, we have also simulated Gaussian errors with constant SNRs with the values given by the horizontal lines (the average SNR) in Fig. 2(a) and (b).

B. Which Endmember Estimation Algorithm Should Be Used?

Given $\hat{M}$, both [15] and [16] use N-FINDR [17] to estimate the endmembers. N-FINDR finds the simplex of maximum hypervolume with M vertices, constrained to lie among the data points, typically in a subspace of dimensionality $M-1$ (e.g.,



(a)



(b)

Fig. 2. Estimated SNRs for real Indian Pines and Cuprite images. (a) Indian Pines. (b) Cuprite.

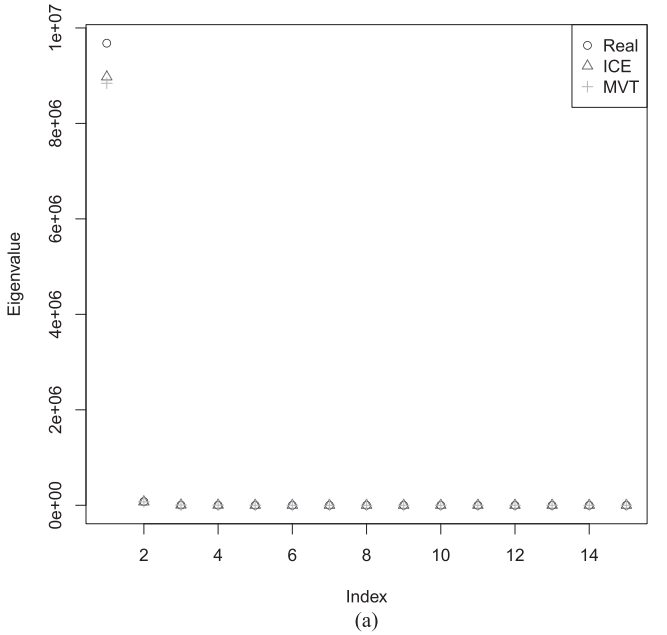
the first $M-1$ PC or MNF bands). Because of this constraint, N-FINDR implicitly assumes that pure versions of all the endmembers exist in the data itself. In our opinion, this assumption is unrealistic. In the last 20 years or so, many endmember estimation algorithms, which seek to overcome this problem have been proposed; [11, Sec. IV.B] provides a good review of these.

Probably, the oldest of these is called the minimum volume transform (MVT) [19]. Since then, there have been several algorithms proposed to calculate the MVT [7], [20], [21]. Unlike N-FINDR, MVT finds the simplex of minimum hypervolume with M vertices, constrained to totally enclose the data projected onto the $(M-1)$-dimensional subspace inside the simplex. This algorithm is unrealistic because it assumes that all the data in the first $M-1$ PC or MNF bands is signal (i.e., with no errors). However, if this endmember estimation method is used, it is particularly simple to produce the target image (3). Note that, in this case, any simplex with M vertices which completely encloses the projected data in the $(M-1)$-dimensional subspace has zero error in that subspace. As a simple example, when $M=3$, this simplex will be any triangle enclosing all the (projected) data. In this case, the target image (the signal component of (3)) can be reconstructed directly from the first $M-1$ PC or MNF bands, with no constraints on the weights.

Several algorithms, which aim to steer a middle course between the "extremes" of N-FINDR and MVT, have been developed. We will use the iterated constrained endmembers (ICE) algorithm [22]. This uses a regularized LS fit of (1) (typically in MNF space), where the regularizing function is proportional to the total variance of the endmembers.

We have simulated both the Indian Pines and Cuprite datasets for a range of values of $\hat{M}$ (discussed in Section V) using both MVT and ICE to estimate the endmembers in (3). The errors δ_i in (3) are uncorrelated with different SDs, assuming either a variable or constant SNR [see Fig. 2(a) and (b)], estimated from the real data using Roger's method [13]. Figs. 3 and 4 show the eigenvalues for one of the simulated Indian Pines ($\hat{M}=20$) and Cuprite ($\hat{M}=36$) datasets using both ICE and MVT for the simulations using variable SNR. For all the ICE simulations shown in this paper, we have used its default regularisation parameter, 0.01. We have first "scaled" each real and simulated image by dividing the data in each band by Roger's estimate of the band SD, so the SNR is variable. For each dataset, we plot the first K (scaled) eigenvalues in plot (a) (where K is a little less than $\hat{M}$) and the remaining eigenvalues in plot (b). In plot (b), we have also drawn a vertical line between the first $\hat{M}-1$ "signal" eigenvalues and the "noise" eigenvalues. In both Figs. 3(b) and 4(b), we see that the real scaled eigenvalues die away smoothly, and that there is no obvious drop between the signal and noise eigenvalues. This is commonly the case with real data. The ICE eigenvalues reflect this behavior. However, the MVT eigenvalues do not; there is a very obvious drop between the "signal" and "noise" eigenvalues. Any reasonable ID estimation method ought to be able to detect this gap and give a perfect estimate of M. The reason for this gap is the implicit (and unrealistic) assumption made by MVT that there is no "noise" in the first $M-1$ (scaled) PCs. Consequently, we will only use the ICE simulations for the rest of the paper.

Although the ICE eigenvalues behave more like the real eigenvalues than the MVT eigenvalues do, there are still some differences. There are at least two reasons for this. First, the values of $\hat{M}$ are a little too small. We will present evidence for this in Section V. Second, Roger's estimates of the band error variances are too large. A careful examination of Fig. 1(a) and (b) suggests that this is so. This is made clearer in Fig. 5(a) and (b), which shows the mean (± 2 SDs) of Roger's estimates of the band error SDs for the simulated datasets, each divided by their target SDs, for the Indian Pines and Cuprite datasets, respectively. We will call these the estimated relative SDs. The error bars are larger for the Indian Pines scene because the image size ($N=21\,025$) is much smaller than that of the Cuprite scene ($N=314\,368$).



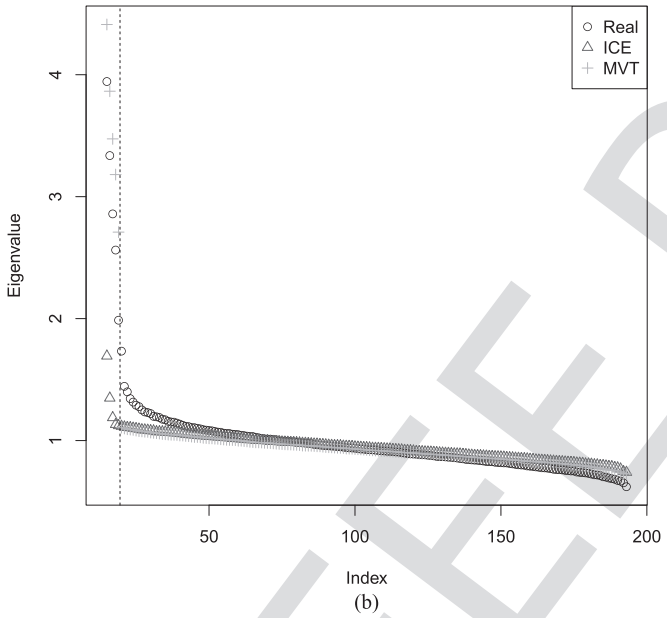
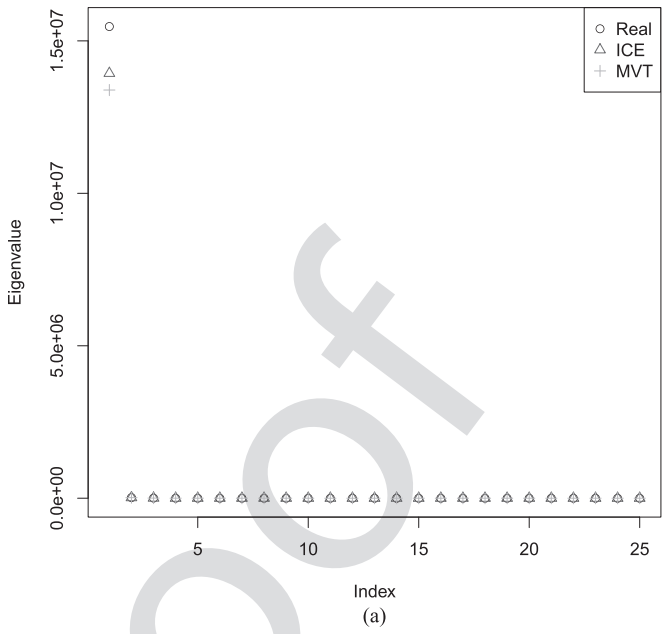


Fig. 3. Indian Pines: Real and simulated ICE and MVT scaled eigenvalues (variable SNR, $\hat{M} = 20$). (a) Eigenvalues 1–15. (b) Eigenvalues 15–193.



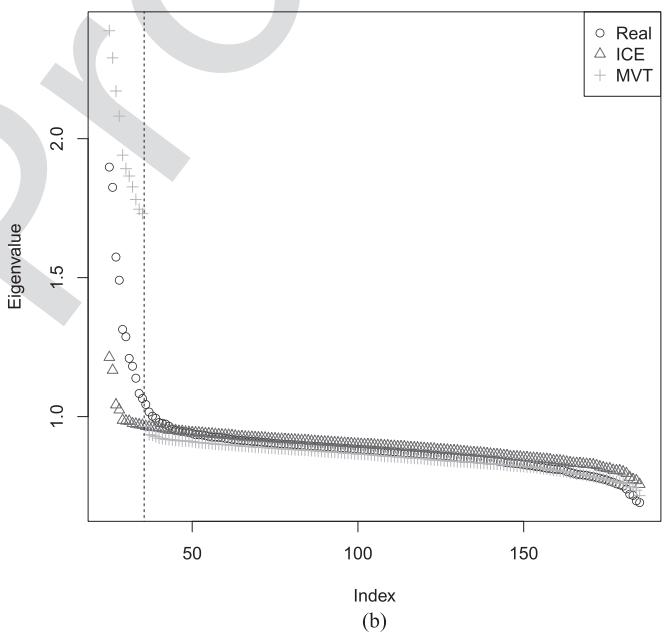


Fig. 4. Cuprite: Real and simulated ICE and MVT scaled eigenvalues (variable SNR, $\hat{M} = 36$). (a) Eigenvalues 1–25. (b) Eigenvalues 25–185.

C. Some Properties of Roger's Estimator

In Appendix A, we prove that, under mild assumptions, Roger's estimate, $\hat{\sigma}^2_{\epsilon,j}$ [defined in (4)] is greater than the true error variance $\sigma^2_{\epsilon,j}$.

Therefore, when scaling data by dividing by Roger's SD estimates, we are dividing by values which are too large, and hence, the real error variances of the scaled data are all too small. Because the sum of the variances of the scaled data equals the sum of the scaled eigenvalues, the larger eigenvalues will also need to be too small to guarantee this. This is what we see in Figs. 3 and 4. The first 85 and 47 real scaled eigenvalues are larger than their simulated counterparts in these two figures, respectively.

Let

$$r_j = \hat{\sigma}^2_{\epsilon,j} / \sigma^2_{\epsilon,j}. \tag{6}$$

This is the estimated relative variance. The positive bias result above provides the lower bound

$$r_j > 1, \quad j = 1, ..., d. \tag{7}$$

We have also been able to derive a lower bound on the average inverse relative variance:

$$\sum_{j=1}^{d} r_j^{-1} / d > (d - M)/d. \tag{8}$$

The proof of (8) is given in Appendix B.

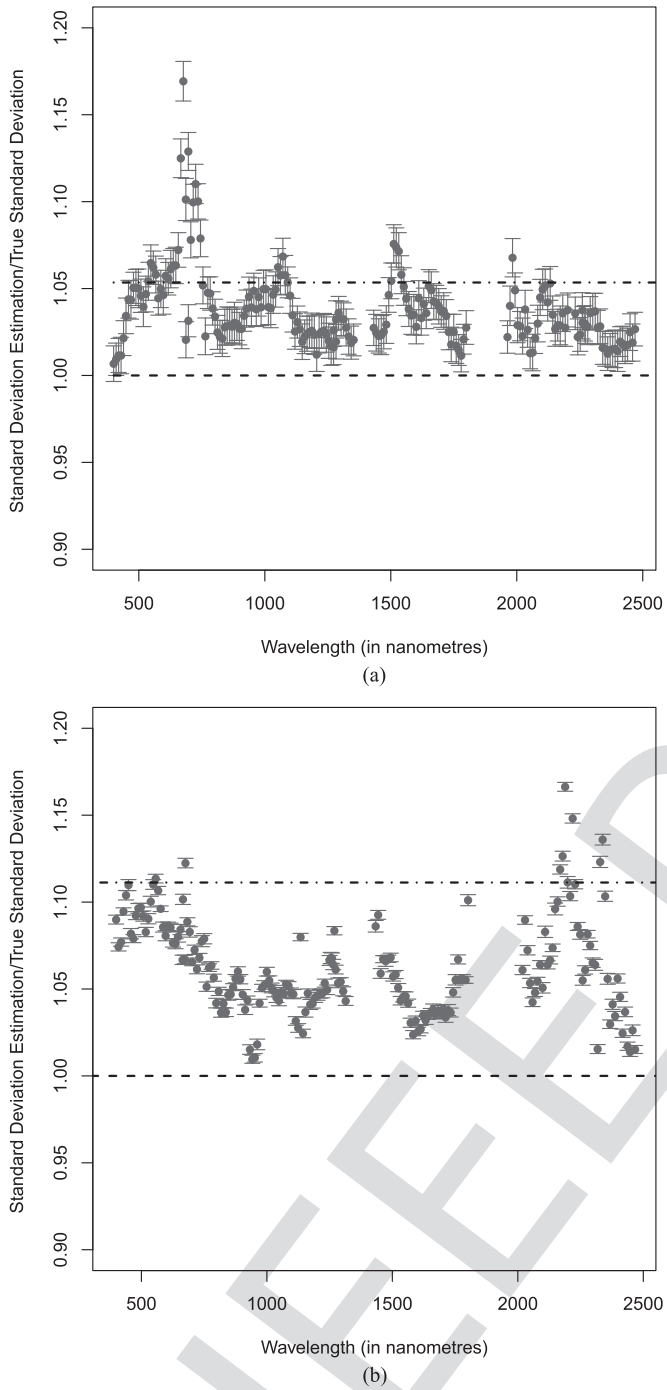


Fig. 5. Mean (± 2 SDs) for relative SDs for simulated Indian Pines and Cuprite images, and upper bounds. (a) Indian Pines ($\hat{M} = 20$, 100 simulations). (b) Cuprite ($\hat{M} = 36$, 40 simulations).

If we are prepared to take a Bayesian approach and assume that the jth row of the eigenvector matrix is uniformly distributed on the d-dimensional sphere, then we can show that

$$E(r_j) < (d-2)/(d-M-2). \tag{9}$$

The proof of (9) is also given in Appendix B.

The upper line in Fig. 5(a) and (b) is $\{(d-2)/(d-M-2)\}^{1/2}$ (since we are plotting SDs rather than variances). Most of the estimated relative SDs in both figures lie below this value. For a single Indian Pines simulation, the left-hand side of (9) is 1.078, while the right-hand side is 1.110. For a single Cuprite simulation, the left-hand side of (9) is 1.132, while the right-hand side is 1.236. So, (9) holds for both simulated datasets.

The value of the upper bound in (9) is that it gives an approximate idea of the likely magnitude of the typical bias in Roger's estimates. In most of the above-mentioned papers, d varies around 200. As mentioned previously, for almost all the simulations in the above papers, $M \leq 10$. When $d = 200$ and $M = 10$, $(d-2)/(d-M-2) = 1.053$, and so the typical bias is less obvious and less important. However, hyperspectral satellites due for launch in the next few years will all have about $d = 200$ bands [23]. The images produced by such satellites will undoubtedly be larger than those produced by the simulations in the above-mentioned papers. Hence, many will have a larger ID, M, and so the typical bias in Roger's estimates will be larger and more significant.

III. Some ID Estimation Methods

Five ID estimation methods will be discussed briefly in this section. The first three, HFC, noise-whitened HFC (NWHFC), and noise subspace projection (NSP), are collectively called VD and were introduced in [2]. All methods assume that the errors are spectrally (and spatially) uncorrelated.

A. HFC

HFC implicitly assumes that all the band error SDs are equal. It tests statistically the equality of the eigenvalues of the sample covariance matrices with and without mean correction. Its ID estimate is the number of eigenvalue pairs determined to be unequal. HFC (as well as NWHFC and NSP) require the user to set a false alarm probability, P_f. In most papers, P_f is set to $10^{-3}, 10^{-4}$, or 10^{-5}.

B. NWHFC

NWHFC does not assume that the band error SDs are equal. It scales the data using Roger's error SD estimates [13] and then applies HFC to the scaled data.

C. NSP

Note that, if we were to scale the data by the "true" error SDs, then the error SDs of the scaled data will all be 1. One would then expect those eigenvalues of the scaled data, which correspond to the noise, to have values which are on average about 1. This is the basis of the third method, called "noise subspace projection" (NSP). The eigenvalues of the mean-corrected sample covariance matrix are tested against 1. When these are not significantly different from 1, one concludes that they correspond to noise. However, NSP has two problems. First, when the data are scaled by the true error SDs, although the "noise" eigenvalues are on average about 1, the first noise eigenvalue is somewhat greater than 1, in a way that can be made explicit mathematically [24], [25]. In this case, the NSP estimate would be too large. However, Roger's estimates are positively biased and consequently, as noted in Section II-C, the leading eigenvalues are too small, so the value of 1 is reached earlier than it

is when scaled by the true error SDs. These two "wrongs" have opposite effects and sometimes make a "right." However, often they do not, as we shall see in Section V.

D. HySime

HySime [3, Algorithm 2] uses Roger's method to estimate the signal and noise components at each pixel, and from these to estimate signal and noise covariance matrices. They then decompose the expected mean square error (MSE) of the true signal into estimated signal and noise components via an eigendecomposition of the estimated signal covariance matrix. Their ID estimate is the dimensionality of the subspace of eigenvectors which minimizes the MSE.

E. Random Matrix Theory

RMT is used by [26] to estimate the ID, under the assumption that the band error variances are equal. This approach is adapted by [5, Algorithm 1] to deal with the case where the band error variances are unequal. They estimate these using a method based on finding homogeneous areas in the scene [14]. However, they do not scale the data by the estimated band error SDs, but subtract the estimated error covariance matrix (i.e., which is assumed to be diagonal) from the data covariance matrix, to obtain an estimate of the signal covariance matrix. For the purposes of comparison with the other ID estimation methods described above, we will use [5, Algorithm 1], but with Roger's estimates of the band error variances instead of those based on the method described in [14].

IV. Effective Intrinsic Dimensionality

There are a number of definitions of ID in the hyperspectral literature. Cawse-Nicholson, Damelin, Robin, and Sears [5, Introduction] give a useful review of these. In particular, Chang and Du [2] define VD as "the minimum number of spectrally distinct signal sources that characterize the hyperspectral data from the perspective view of target detection and classification." This is a very informal definition. A more formal definition is given by Cawse-Nicholson, Damelin, Robin, and Sears [5, Definition 1], who state that ID is the dimensionality of the signal subspace. In [12], ID is simply defined as the number of endmembers, which is also the definition that we use in this paper.

Assuming that the linear mixture model (which of course is an approximation to reality) holds, we propose an alternative concept, which we call Effective Intrinsic Dimensionality (EID). It reflects the idea that, in the presence of errors, some signals in the data may be so weak as to be undetectable, and so is a means of bridging the gap between the definitions in [2] and [5]. When the band error variances are all equal, to σ^2 say, then [26, eq. (11)] and [5, eq. (8)] state that there is a threshold value (called the "asymptotic limit of detection" by the former), below which signal eigenvalues cannot be successfully identified, at least asymptotically. This value is $\sigma^2\sqrt{d/N}$. The proof of this result is contained in [27, Th. 1.1].

TABLE I
Various ID Estimates for the Real Indian Pines and Cuprite Images

Scene	HFC	NWHFC	NSP	HySime	RMT
Indian Pines	29, 28, 27	18, 18, 17	62, 60, 59	11	19
Cuprite	36, 29, 22	26, 24, 22	37, 37, 37	16	29

This concept is easily generalised to the case where the errors have unequal variances and/or are spectrally correlated. Before we do this, we need to make some definitions and make two mild assumptions. Let

$$S_i = \sum_{j=1}^{M} w_{\mathrm{ij}} E_j, \quad i = 1, \ldots, N, \tag{10}$$

denote the "signal" component of (1).

Let Σ_S and Σ_ϵ denote the expected values of the signal and error covariance matrices, respectively.

Assumption 1: Σ_ϵ is positive definite. Of course, if Σ_ϵ is diagonal, all of whose diagonal entries are "strictly" positive (which we have assumed in this paper), then it will be positive definite.

Assumption 2: Σ_S is positive semidefinite.

Equation (10) implies that the rank of Σ_S is $M(<d)$. However, this is not needed for what follows.

The two assumptions imply that there exists a $d * d$ matrix A satisfying

$$A^T \Sigma_S A = \Lambda, \quad A^T \Sigma_\epsilon A = I \tag{11}$$

where Λ is a diagonal matrix with nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ [28, Th. 12.2.13]. Of course, if the rank of Σ_S is M, then $\lambda_j = 0, j = M+1, \ldots, d$. When Σ_ϵ is diagonal, the simultaneous diagonalization (11) is achieved by simply dividing the data by the band error SDs and then applying a standard eigendecomposition.

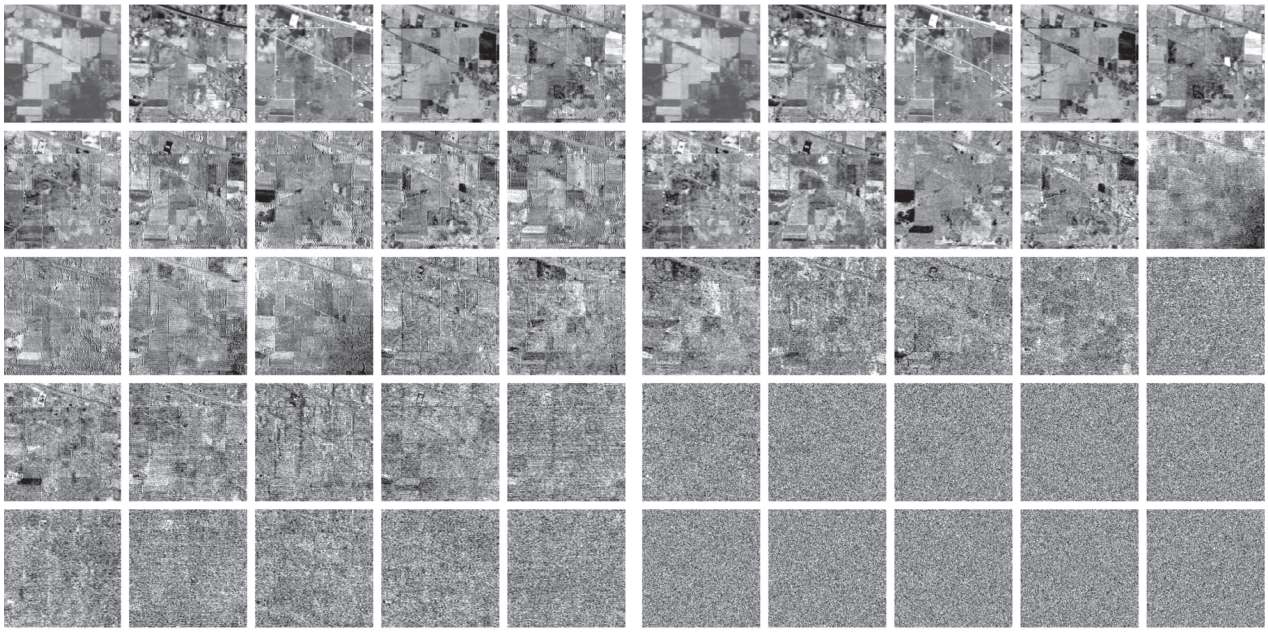
The transformation (11) has converted the problem of unequal error variances to one with error variances all equal to 1. So the "asymptotic limit of detection" becomes

$$\lambda_{\mathrm{crit}} = \sqrt{d/N}. \tag{12}$$

We define the EID as the number of λ_j's greater than λ_{crit}. Of course, EID $\leq$ ID, because $\lambda_{\mathrm{crit}} > \lambda_{M+1} = 0$. We shall see that, in our simulations, for smaller values of $\hat{M}$, EID = ID, while for larger values of $\hat{M}$, EID $<$ ID, reflecting the fact that, as $\hat{M}$ becomes larger, the additional endmembers represent signals in the real scene, which are 1) minor variations of other more ubiquitous endmembers, 2) rare, or 3) nonexistent altogether.

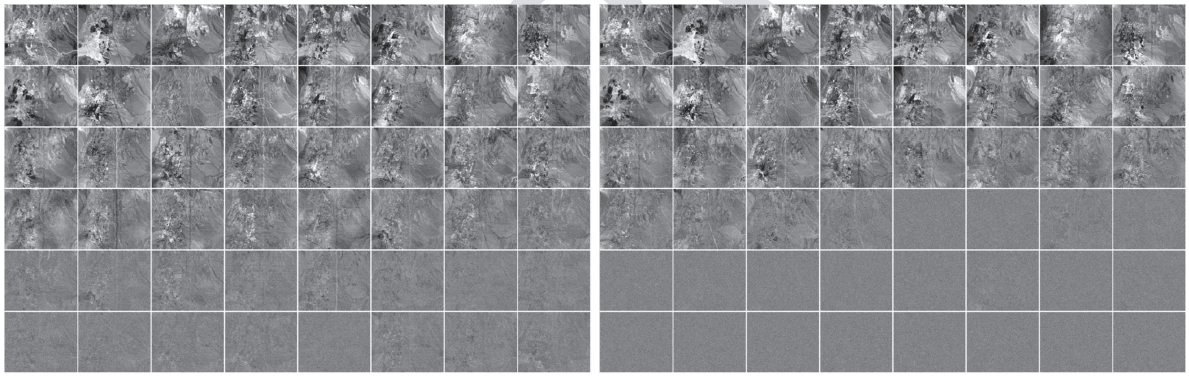
V. Simulations

In order to decide on a plausible range of values of $\hat{M}$, we first apply the five ID estimation methods described in Section III to the real datasets. For HFC, NWHFC, and NSP, we give the results for $P_f = 10^{-3}, 10^{-4}$, and 10^{-5} in that order. The results are shown in Table I.



(a) (b)

Fig. 6. Indian Pines: Real and simulated scaled PCs ($\hat{M} = 20$). (a) Real scaled PCs. (b) Simulated scaled PCs ($\hat{M} = 20$).



(a) (b)

Fig. 7. Cuprite: Real and simulated scaled PCs ($\hat{M} = 36$). (a) Real scaled PCs. (b) Simulated scaled PCs ($\hat{M} = 36$).

Our initial assumption is that, for each image, the true ID is somewhere in the range of estimates in Table I. However, the ranges are very large: 11–62 for Indian Pines and 16–37 for Cuprite. HySime gives the lowest estimates and NSP gives the highest estimates for both datasets. This is a great concern because the papers which introduce these two methods, [2] and [3], are probably the two most highly cited papers on ID estimation in the hyperspectral literature. The NSP estimates for Indian Pines (62, 60, and 59) are particularly hard to believe; see the discussion in Section III-C for a possible explanation of this extreme behavior. Low HySime estimates were also obtained by Robin, Cawse-Nicholson, Mahmood, and Sears [16], when applied to several real datasets (different from those analyzed here), but not in their simulated datasets, which used N-FINDR [17] to generate their endmembers. Another issue which this table highlights is the difficulty in deciding on the appropriate value of P_f for the three VD methods. Chang and Du [2] use $P_f = 10^{-3}$, while Robin, Cawse-Nicholson, Mahmood, and Sears [16, Sec. II.C] state: "We have confirmed the stability of this value in preliminary experiments, as using 10^{-3} or 10^{-4} made no difference in the results." The HFC estimates for the Cuprite scene are inconsistent with their observations, suggesting potential problems with HFC at least.

In order to assist us with narrowing the range of plausible values of $\hat{M}$, Figs. 6(a) and 7(a) show the first 25 and 48 "scaled" PCs for the Indian Pines and Cuprite scenes, respectively (i.e., after dividing each band by Roger's error SD estimate). Each PC image has been linearly stretched over the range of its mean +/− 2.5 SDs, so that any signal is apparent.

Real signal is apparent in perhaps the first 20 Indian Pines PCs, which is why we have chosen this value of $\hat{M}$ for illustrative purposes. However, there is some form of horizontal

instrumental noise in later PCs (which our model will treat as "signal" because it is spatially correlated), as well as some vertical and "speckly" features in some of the earlier PCs. It is difficult to assess whether these features are in the scene or are instrumental effects. Fig. 6(b) shows the simulated scaled PCs when $\hat{M} = 20$. Signal is apparent in the first 14 PCs (which suggests that $\hat{M} = 20$ may be too small). Of these, the first nine are similar to the corresponding real images. The horizontal, vertical, and many of the speckle features are absent from the simulated PCs, which highlights the need to incorporate such "artefacts" in mixture models. However, their absence from the simulated images makes it easier to compare the various ID methods under the model assumptions of this paper. Based on our interpretation of Fig. 6(a), we have decided to omit the HySime and NSP estimates of the real Indian Pines dataset and simulate it with $\hat{M}$ varying between 17 and 29.

Strong signal is apparent in perhaps the first 31 or 32 Cuprite PCs. However, there are weaker "spatially coherent" signals apparent in most of the remaining PCs, up to PC 48. Perhaps these weaker signals are due to local variants of some endmembers and/or small nonlinearities in the mixing. However, from the point of view of a linear mixture model, they are real signals. There are no apparent nonrandom instrumental effects. Based on our interpretation of Fig. 7(a), we have decided to omit the much smaller HySime estimate of the real Cuprite dataset and simulate it with $\hat{M}$ varying between 22 and 37. Fig. 7(b) shows the simulated scaled PCs when $\hat{M} = 36$. This allows for the strong signals plus a few weak ones. Signal is apparent in the first 27 or 28 PCs, although there is a faint signal in PC 31. Again, this suggests that $\hat{M} = 36$ may be too small. The first 14 PCs of the real and simulated images are similar.

Fig. 8(a) and (b) shows the mean (± 2 SDs) for the five ID estimates versus $\hat{M}$ for the simulated Indian Pines and Cuprite images. Following [2] and [16], we use $P_f = 10^{-3}$ for HFC, NWHFC, and NSP. The first thing to note is that, for smaller values of $\hat{M}$, EID = ID. However, as $\hat{M}$ becomes larger, EID $<$ ID, probably due to the reasons given at the end of Section IV.

For the Indian Pines simulations, all five estimators are almost independent of $\hat{M}$. This probably reflects the fact that the true ID is at the lower end of the range of values of $\hat{M}$ chosen, and the inability of any of the estimators to detect weak or rare signals in the simulated (and possibly real) data. The mean estimates of the simulations of each of four of the five ID estimation methods are in the same order as their corresponding estimates for the real data (NSP $>$ HFC $>$ NWHFC $>$ HySime). HFC and NWHFC are on average about the same as for the real data, while HySime is a little smaller. NSP drops from 62 to the high 40s (which is still far too high), while RMT drops from 19 to a range of 10–12. The drop in the HySime, NSP, and RMT estimates may, at least in part, be due to the positive bias in Roger's band estimators. Because HFC and NWHFC compare the eigenvalues of covariances with and without mean correction, they are perhaps somewhat less sensitive to this bias, whose effects may be about the same on both sets of eigenvalues.

The most striking feature for the Cuprite simulations is that the mean values of the HFC estimates are much too high for $\hat{M} < 32$, but very good for $\hat{M} \geq 32$ (although its SDs are very

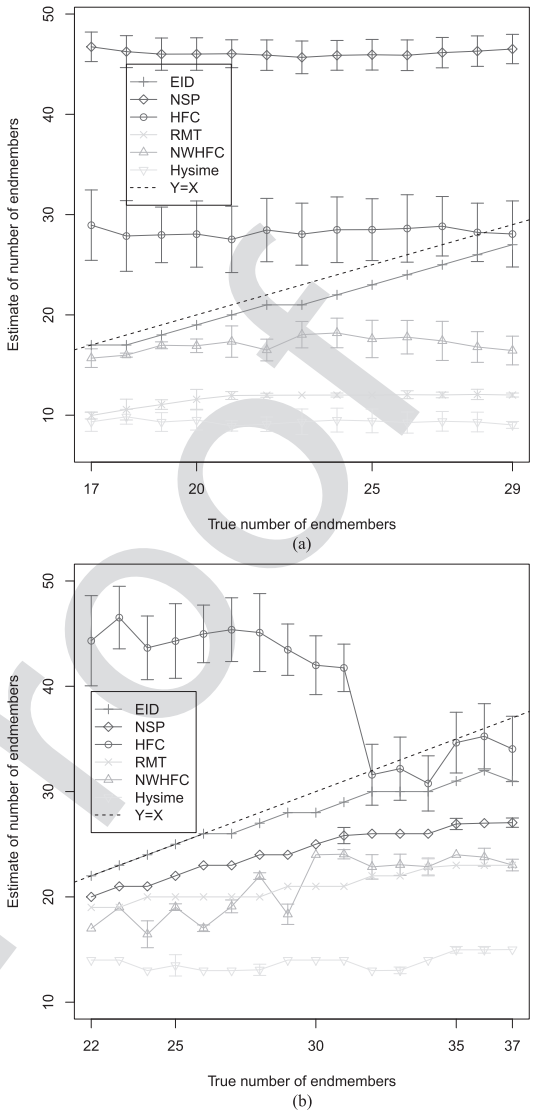


Fig. 8. Mean (± 2 SDs) for five ID estimates versus $\hat{M}$ for simulated Indian Pines and Cuprite images with variable SNR ($P_f = 10^{-3}$). (a) Indian Pines. (b) Cuprite.

large). We have no explanation for this as yet. All the remaining estimates lie below EID, but gradually increase with $\hat{M}$, probably reflecting the fact that the true ID is at the upper end of the range of values of $\hat{M}$ chosen. If one restricts attention to values of $\hat{M}$ above 32, the mean estimates of simulated HFC estimates comes closest to the corresponding estimate for the real data (36), while the means of the remaining simulated estimators tend to be a little less than their counterparts for real data. This is consistent with the behavior seen with the simulated Indian Pines data.

Overall, HFC is clearly the most variable of the estimators. Because NWHFC is a whitened version of HFC, its variability is greatly reduced and is comparable to that of HySime. Both are a little more variable than RMT. NSP's variability appears to be a function of its mean.

The results in Fig. 8(a) and (b) are for variable SNR. We have also carried out simulations using constant SNR, whose values

are shown in Fig. 2(a) and (b). We do not plot the results here. Generally speaking, however, the estimators are a little larger than for their variable-SNR counterparts. For the Cuprite data, the HFC and NWHFC estimates are even more variable than are their variable-SNR counterparts.

VI. Conclusion and Future Work

The ten papers cited near the beginning of this paper (a number of which are highly cited) use simulations to introduce or test either 1) methods for estimating the ID of hyperspectral data, or 2) methods for estimating endmembers in such data. In our opinion, these simulations are so different from real-world data that one cannot be confident that the various advocated methods work well with real-world data. We have introduced a general methodology which aims to make the simulations more realistic. We have also explored various options within the general framework, including 1) comparing the use of variable and constant SNRs for the errors, and 2) comparing the MVT and ICE algorithms as inputs to the simulation method. We believe that by attempting to produce more realistic simulations, we have obtained greater insights into the strengths and weaknesses of various advocated methods. In particular, we have observed that Roger's error variance estimates are positively biased. This has led us to prove (in Appendix A) that this is always so. We have also demonstrated that the average relative bias of Roger's error variance estimates is an increasing function of M/d and developed some theory in Appendix B to support this.

Although there are differences between the ID estimates for the real images and their simulated counterparts, they are mostly consistent. In addition, our approach strongly suggests that HFC and NSP are highly variable (and hence unreliable) ID estimators. On the other hand, NWHFC, HySime, and RMT appear to underestimate ID (or even EID). This may in part be due to the bias in Roger's variance estimates. So a first goal is to see if this bias can be corrected.

Toward this end, Mahmood, Robin, and Sears [29] have recently developed a first-order correction to Roger's error SD estimate. We have applied this correction to three real and simulated hyperspectral images (including the two presented here). For the three simulated images, although the bias of the corrected estimates is certainly reduced, it still appears to be a little positive. In addition, for the real image not presented here (which has worse artefacts than those in the Indian Pines scene), some of the bands have very small error variances. Unfortunately, the first-order correction produces negative estimates of some of these. So, although promising, the new method needs some further development, which we intend to investigate.

Depending on the results of these investigations, modifications of some of the above ID estimation methods (or possibly entirely new ones) may need to be developed to make them more accurate than Fig. 8(a) and (b) suggest that they are.

Recently, Robin, Cawse-Nicholson, Mahmood, and Sears [16] and Cawse-Nicholson, Robin, and Sears [30] have investigated the estimation of ID in the presence of correlated noise, which is known to have a significant presence in some instruments. They used various multivariate and spatial methods to estimate the ID. It is our intention to investigate these methods, again with the aid of semirealistic simulations of natural hyperspectral scenes, and if necessary either to improve them further or to develop new methods for correlated errors.

Another issue that has not been addressed in this paper is sparsity. Once the endmembers have been estimated, one needs to estimate the proportions. In our simulations, we have fitted the spectrum in each pixel using LS fitting in MNF space with the constraints that the weights are nonnegative and sum to 1. However, even though this constraint forces some of the weights to be zero, it generally leaves too many endmembers with small (positive) weights. In recent years, "sparse unmixing" techniques have been introduced [31]–[33]. These force many of these small weights to be zero. It will be interesting to see what impact "sparsity" has on the simulated data. In particular, for large $\hat{M}$, the *maximum* proportion of some of the endmembers in a simulated scene will be very small. Forcing sparsity may push the maximum proportion of some of these endmembers to zero, thus reducing the EID.

Appendix A
Proof of Positive Bias of Roger's Band SD Estimators

Before we prove that Roger's band SD estimators are always positively biased, we need to carefully state our assumptions, which are fairly standard. Because we are assuming that the error covariance matrix is diagonal, we need to modify Assumption 1 appropriately.

Assumption 1a: Σ_ϵ is diagonal, all of whose diagonal entries, $\sigma^2_{\epsilon,j}, j = 1, \ldots, d$, are "strictly" positive.

Equation (10) implies that the rank of Σ_S is $M(<d)$. We, therefore, modify Assumption 2 to the following.

Assumption 2a: Σ_S is positive semidefinite with (unknown) rank $M(<d)$.

In addition, we assume the following.

Assumption 3: S_i and ϵ_i are uncorrelated.

Let Σ_X denote the expected value of the data. It follows from (1), (10), and Assumption 3 that

$$\Sigma_X = \Sigma_S + \Sigma_\epsilon. \tag{13}$$

Let $\bar{X} = \sum_{i=1}^N X_i/N$, let $\hat{\Sigma}_X = \sum_{i=1}^N X_i X_i^T/N$ denote the sample covariance matrix of the observed data, and let v_j denote the jth diagonal entry of $\hat{\Sigma}_X^{-1}$. Quoting other references, Roger [13, Sec. 2.3.1] pointed out that

$$v_j^{-1} = \hat{\sigma}^2_{\epsilon,j}, \quad j = 1, \ldots, d. \tag{14}$$

Either this formula or (4) can be used to provide Roger's estimator of the band error variances (and hence their SDs).

The flaw in Roger's approach is easily seen from the regression interpretation, on which (4) is based. When regressing one band on the remaining $d-1$ bands, the $d-1$ explanatory variables themselves have errors. This is the cause of the bias, and as it happens, it is always positive.

In the proof that follows, we make no distinction between the theoretical covariance matrices in (13) and their sample versions. The differences are negligible if N is large enough.

We do not know how large N needs to be, but it will certainly be the case for most real-world datasets.

Proof: It follows from (13) and (11) that

$$\Sigma_X^{-1} = A(\Lambda + I)^{-1} A^T \tag{15}$$

and

$$\Sigma_\epsilon^{-1} = AA^T. \tag{16}$$

Let $a_j^T \equiv (a_{j1}, \ldots, a_{\mathrm{jd}})$ denote the jth row of A. It then follows from (16) and Assumption 1 a that

$$a_j^T a_j = \sum_{k=1}^{d} a_{\mathrm{jk}}^2 = \sigma_{\epsilon,j}^{-2}. \tag{17}$$

Let e_j denote the d-vector, which has zeroes everywhere, except for the jth position, which equals 1. It then follows from the definitions of v_j, a_j, (15), and (17) that

$$\begin{aligned} v_j &= e_j^T \Sigma_X^{-1} e_j \\ &= e_j^T A(\Lambda + I)^{-1} A^T e_j \\ &= a_j^T (\Lambda + I)^{-1} a_j \\ &= \sum_{k=1}^{d} a_{\mathrm{jk}}^2/(\lambda_k + 1) \qquad (18) \\ &< \sum_{k=1}^{d} a_{\mathrm{jk}}^2 \qquad (19) \\ &= \sigma_{\epsilon,j}^{-2} \qquad (20) \end{aligned}$$

by (17). Invert both sides of this inequality and use (14) to obtain

$$\hat{\sigma}_{\epsilon,j}^2 > \sigma_{\epsilon,j}^2, \quad j = 1, \ldots, d. \tag{21}$$

Note that the strict inequality (19) holds provided that the largest eigenvalue is positive.

Appendix B
Derivation of Inequalities (7) and (8)

At this point, it will be convenient to let

$$\Sigma_S^* = \Sigma_\epsilon^{-1/2} \Sigma_S \Sigma_\epsilon^{-1/2}, \Sigma_X^* = \Sigma_\epsilon^{-1/2} \Sigma_X \Sigma_\epsilon^{-1/2}. \tag{22}$$

These are just the covariance matrices of the signal and data respectively after each band has been divided (i.e., "scaled") by the true error SD in each band. Then, (13) becomes

$$\Sigma_X^* = \Sigma_S^* + I \tag{23}$$

and (11) can be converted into a "standard" eigendecomposition, in which the band error variances all equal 1:

$$B^T \Sigma_S^* B = \Lambda, \quad B^T B = BB^T = I \tag{24}$$

where $B = \Sigma_\epsilon^{1/2} A$. Note that, because B is orthogonal, (24) includes one more equality than (11) does.

The analog of (15) is

$$(\Sigma_X^*)^{-1} = B(\Lambda + I)^{-1} B^T. \tag{25}$$

Take the trace of both sides of (25) and use its cyclic property and the last equality in (24) to obtain

$$\begin{aligned} tr((\Sigma_X^*)^{-1}) &= tr((\Lambda + I)^{-1}) \\ &= \sum_{k=1}^{d} (\lambda_k + 1)^{-1} \\ &> d - M \qquad (26) \end{aligned}$$

because Assumption 2a implies that the last $d - M$ λ_k's are all zero. However, it follows from the second equality in (6), (14), and (22) that the left-hand side of (26) is just $\sum_{j=1}^{d} r_j^{-1}$. Substitute this into (26) and divide both sides by d to obtain (8).

To obtain inequality (9), let $(b_{j1}, \ldots, b_{\mathrm{jd}})$ denote the jth row of B. Using an analogous argument leading to (6), (14), and (17), we obtain

$$\begin{aligned} r_j^{-1} &= \sum_{k=1}^{d} b_{\mathrm{jk}}^2/(\lambda_k + 1) \\ &> \sum_{k=M+1}^{d} b_{\mathrm{jk}}^2 \end{aligned}$$

again because the last $d - M$ λ_k's are all zero. Invert this inequality to obtain

$$r_j < \left(\sum_{k=M+1}^{d} b_{\mathrm{jk}}^2 \right)^{-1}. \tag{27}$$

In order to make further progress, we need to make a distributional assumption about the b_{jk}'s. Note first that, analogous to (17), we have from the second equality in (24) that

$$\sum_{k=1}^{d} b_{\mathrm{jk}}^2 = 1. \tag{28}$$

We make the following (Bayesian) assumption:

Assumption 4: For each j, conditional on (28), $b_j^T \equiv (b_{j1}, \ldots, b_{\mathrm{jd}})$ is uniformly distributed on the d-dimensional sphere of radius 1.

Under this assumption, $\sum_{k=M+1}^{d} b_{\mathrm{jk}}^2$ has a Beta distribution with parameters $\frac{1}{2}(d - M)$ and $\frac{1}{2}M$ [34, p. 49]. It is then straightforward to show that $E((\sum_{k=M+1}^{d} b_{\mathrm{jk}}^2)^{-1}) = (d - 2)/(d - M - 2)$. Substitute this into (27) to obtain (9).

References

[1] E. Ientilucci and S. Brown, "Advances in wide-area hyperspectral image simulation," *Proc. SPIE*, vol. 5075, pp. 110–121, 2003.

[2] 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.

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

[4] N. Acito, M. Diani, and G. Corsini, "A new algorithm for robust estimation of the signal subspace in hyperspectral images in the presence of rare signal components," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3844–3856, Nov. 2009.

[5] K. Cawse-Nicholson, S. Damelin, A. Robin, and M. Sears, "Determining the intrinsic dimension of a hyperspectral image using random matrix

theory," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1301–1310, Apr. 2013.
[6] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 3, pp. 765–777, Mar. 2007.
[7] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4418–4432, Nov. 2009.
[8] M. Zortea and A. Plaza, "Spatial preprocessing for endmember extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2679–2693, Aug. 2009.
[9] A. Huck, M. Guillaume, and J. Blanc-Talon, "Minimum dispersion constrained nonnegative matrix factorization to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 6, pp. 2590–2602, Jun. 2010.
[10] B. Somers, M. Zortea, A. Plaza, and G. Asner, "Automated extraction of image-based endmember bundles for improved spectral unmixing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 396–408, Apr. 2012.
[11] J. Bioucas-Dias *et al.*, "Hyperspectral unmixing overview: Geometrical, statistical and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, Apr. 2012.
[12] J. Bioucas-Dias and J. Nascimento, "Estimation of signal subspace on hyperspectral data," *Proc. SPIE*, vol. 5982, pp. 191–198, 2005.
[13] R. Roger, "Principal Components transform with simple automatic noise adjustment," *Int. J. Remote Sens.*, vol. 17, pp. 2719–2727, 1996.
[14] P. Meer, J.-M. Jolion, and A. Rosenfeld, "A fast parallel algorithm for blind estimation of noise variance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 2, pp. 216–223, Feb. 1990.
[15] L. Gao, Q. Du, B. Zhang, W. Yang, and Y. Wu, "A comparative study on linear regression-based noise estimation for hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 488–498, Apr. 2013.
[16] A. Robin, K. Cawse-Nicholson, A. Mahmood, and M. Sears, "Estimation of the intrinsic dimension of hyperspectral images: Comparison of current methods," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2854–2861, Jun. 2015.
[17] M. E. Winter, "N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data," *Proc. SPIE* vol. 3753, pp. 266–275, 1999.
[18] A. Green, M. Berman, P. Switzer, and M. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65–74, Jan. 1988.
[19] M. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 3, pp. 542–552, May 1994.
[20] D. Fuhrmann, "A simplex shrink-wrap algorithm," *Proc. SPIE*, vol. 3718, pp. 501–511, 1999.
[21] J. Li and J. Bioucas-Dias, "Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Boston, MA, USA, 2008, vol. 3, pp. 250–253.
[22] M. Berman, H. Kiiveri, R. Lagerstrom, A. Ernst, R. Dunne, and J. Huntington, "ICE: A statistical approach to identifying endmembers," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 10, pp. 2085–2095, Oct. 2004.
[23] K. Staenz, A. Mueller, and U. Heiden, "Overview of terrestrial imaging spectroscopy missions," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Melbourne, Australia, Aug. 2013, pp. 3502–3505.
[24] S. Geman, "A limit theorem for the norm of random matrices," *Ann. Probab.*, vol. 8, pp. 252–261, 1980.
[25] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Ann. Statist.*, vol. 29, pp. 295–327, 2001.
[26] S. Kritchman and B. Nadler, "Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 3930–3941, Oct. 2009.
[27] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *J. Multivariate Anal.*, vol. 97, no. 6, pp. 1382–1408, 2006.
[28] F. Graybill, *Matrices with Applications in Statistics*, 2nd ed. Belmont, CA, USA: Wadsworth, 1983.
[29] A. Mahmood, A. Robin, and M. Sears, "Modified residual method for estimation of noise statistics in hyperspectral images," presented at the 7th Workshop Hyperspectral Image Signal Process.: Evol. Remote Sens., Tokyo, Japan, 2015.
[30] 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.
[31] M.-D. Iordache, J. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2014–2039, Jun. 2011.
[32] Y. Guo and M. Berman, "A comparison between subset selection and L1 regularisation with an application in spectroscopy," *Chemometrics Intell. Lab. Syst.*, vol. 118, pp. 127–138, 2012.
[33] J. B. Greer, "Sparse demixing of hyperspectral images," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 219–228, Jan. 2012.
[34] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, vol. 197. Hoboken, NJ, USA: Wiley, 2009.

Zhipeng Hao received the B.Sc. degree in mathematics from Nankai University, Tianjin, China in 2004, and the master's and Ph.D. degrees in statistics from Macquarie University, Sydney, Australia, in 2009 and 2015, respectively.

He is currently a Postdoctoral Research Fellow with the Centre for Research in Mathematics, Western Sydney University, Parramatta, Australia. His main research interests include hyperspectral data analysis, flow cytometry data analysis, and survival data analysis.

Mark Berman received the B.Sc. (Hons.) (University Medal) degree in mathematical statistics and the Master of Statistics from the University of New South Wales, Sydney, Australia, in 1974 and 1976, respectively, and Ph.D. and D.I.C. degrees in mathematical statistics from the Imperial College of Science and Technology, London, U.K., in 1978.

He was a Visiting Lecturer with the Department of Statistics, University of California, Berkeley, during 1978–1979. Between 1979 and 2014, he was with the CSIRO Division of Mathematics, Informatics and Statistics (CMIS), Sydney, where he reached the position of Chief Research Scientist. He led CMIS' Image Analysis Group from 1989 to 2000. In 1988, he took leave from CSIRO to establish the Image Processing and Data Analysis Group with the Melbourne Research Laboratories of Broken Hill Proprietary Ltd. He has given Ph.D. courses in spectroscopy and hyperspectral imaging with the Technical University of Denmark (2007) and Stanford University (2008 and 2014). He is currently an Adjunct Fellow with the Centre for Research in Mathematics, Western Sydney University, Sydney. He is the author of several patents. His research interests include image analysis (especially hyperspectral), spectroscopy, and spatial data analysis.

Dr. Berman was an Associate Editor of *Computational Statistics and Data Analysis* (2001–2006) and an Associate Editor of *Environmetrics* (2010–2015). He has received Best Paper Awards from the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 1990 and the *Journal of Chemometrics* in 2011.

Yi Guo received the B.Eng. (Hons.) in instrumentation from the North China University of Technology, Beijing, China, in 1998, the M.Eng. degree in automatic control from Central South University, Changsha, China, in 2002, and the Ph.D. degree in computer science, focusing on dimensionality reduction for structured data with no vectorial representation, from the University of New England, Armidale, Australia, in 2008.

From 2008 until 2016, he was with CSIRO, working as a computational statistician on various projects in spectroscopy, remote sensing, and materials science. He recently joined the Centre for Research in Mathematics, Western Sydney University, Sydney, Australia. His recent research interests include machine learning, computational statistics, and big data.

Glenn Stone received the B.A. degree in mathematics from the University of Oxford, Oxford, U.K., the M.Sc. degree in computer science from the University of Manchester, Manchester, U.K., and the Ph.D. degree in statistics from the University of Bath, Bath, U.K.

After a period of teaching statistics with the University of Bath, he moved to Australia and joined the CSIRO in 1993. There, he worked on the development and application of statistical methods for a wide range of areas, including spatial smoothing, data mining, insurance risk, and biostatistics. From 1999 to 2003, he was a Principal Research Analyst with Insurance Australia Group, returning to CSIRO in 2003 to work in bioinformatics and biostatistics and, more recently, in remote sensing. He joined Western Sydney University, Sydney, Australia, in 2011, where he is currently a Professor of Data Science. His research interests include computationally intensive statistical methods with applications in flow cytometry, remote sensing, ecogenomics, and wavelet methods for non-Gaussian data.

Iain Johnstone received the B.Sc. (Hons) degree in pure mathematics and statistics and the M.Sc. degree in statistics from the Australian National University, Canberra, Australia, in 1977 and 1978, respectively, and the Ph.D. degree in statistics from Cornell University, Ithaca, NY, USA, in 1981.

Since 1981, he has been an Assistant, Associate, and then Full Professor with the Department of Statistics, Stanford University, Stanford, CA, USA. Since 1989, he has also held a 50% time appointment in biostatistics with the Stanford University School of Medicine. His research in theoretical statistics has used ideas from harmonic analysis, such as wavelets, to understand noise-reduction methods in signal and image processing. More recently, he has applied random matrix theory to the study of high-dimensional multivariate statistical methods, such as principal components and canonical correlation analysis. In biostatistics, he has collaborated with investigators in cardiology and prostate cancer.

Dr. Johnstone is a member of the U.S. National Academy of Sciences and the American Academy of Arts and Sciences and a former president of the Institute of Mathematical Statistics.