Sparse Matrix Transform for Hyperspectral Image Processing

James Theiler, *Member, IEEE*, Guangzhi Cao, *Member, IEEE*, Leonardo R. Bachega, *Student Member, IEEE*, and Charles A. Bouman, *Fellow, IEEE*

***Abstract*—A variety of problems in remote sensing require that a covariance matrix be accurately estimated, often from a limited number of data samples. We investigate the utility of several variants of a recently introduced covariance estimator—the sparse matrix transform (SMT), a shrinkage-enhanced SMT, and a graph-constrained SMT—in the context of several of these problems. In addition to two more generic measures of quality based on likelihood and the Frobenius norm, we specifically consider weak signal detection, dimension reduction, anomaly detection, and anomalous change detection. The estimators are applied to several hyperspectral data sets, including some randomly rotated data, to elucidate the kinds of problems and the kinds of data for which SMT is well or poorly suited. The SMT is based on the product of K pairwise coordinate (Givens) rotations, and we also introduce and compare two novel approaches for estimating the most effective choice for K.**

***Index Terms*—Anomalous change detection, anomaly detection, change detection, covariance matrix, hyperspectral imagery, matched filter, signal detection, sparse matrix transform (SMT).**

I. Introduction

THE covariance matrix is a key component in a wide array of statistical signal processing tasks applied to remote sensing imagery from multispectral and hyperspectral sensors. If we let $\mathbf{x} \in \mathbb{R}^p$ correspond to the p spectral components at a given pixel, then the distribution of these pixels over the image can be described statistically in terms of an underlying probability distribution. For a Gaussian distribution, the parameters of interest are the mean and the covariance. Let $R \in \mathbb{R}^{p \times p}$ be the "actual" covariance matrix for this distribution, and suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are samples drawn from the distribution. The aim of covariance estimation is to compute a matrix $\widehat{R}$ that is in some sense close to the actual, but unknown, covariance R. What we mean by "in some sense" is that $\widehat{R}$ should be an approximation that is useful for the given task at hand. The maximum-likelihood solution is one such approximation, but particularly when the number of samples, n, is smaller than the number of channels p, this solution tends to over-fit the data. For this reason, a variety of regularization schemes have been investigated [1]–[8]. The sparse matrix transform (SMT) [9]–[11] is a recent addition to this list.

When there are many more pixels than channels, the problem of estimating covariance matrix is not a serious issue, but this is not always the case. Moving-window methods, for instance, seek to better characterize the local statistics of an image and in this case have many fewer pixels with which to estimate those statistics. Cluster-based methods, which segment the image into a large number of spectrally (and in some cases, spatially) distinct regions, have fewer pixels per cluster than are available in the full image. More sophisticated models, such as Gaussian mixture models, also provide fewer pixels per estimated covariance matrix. In addition to reducing the number of pixels available to estimate a covariance matrix of a given size, there are also methods, such as spatio-spectral enhancements, which add many more channels to the image by incorporating local spatial information into each pixel. The choice of window size or cluster number or number of spatio-spectral operators is often influenced by the need to estimate a good covariance matrix. By providing a tool to more accurately estimate a covariance matrix with fewer pixels, these approaches may be further extended.

Many different measures are possible for the quality of an estimate $\widehat{R}$, and the choice of which estimator is best can depend on which measure is used. In [9] and [10], the effectiveness of the covariance estimator was expressed in terms of the Kullback–Leibler distance between Gaussian distributions using R and $\widehat{R}$, while [11] compared estimators based on their utility for weak signal detection.

The purpose of this paper is to evaluate performance on covariance matrices that are observed in real hyperspectral imagery. The evaluation will be in terms that correspond to problems that arise in remote sensing. In addition to the weak signal detection problem that we investigated previously [11], we will consider dimension reduction, anomaly detection, and anomalous change detection. This is in addition to two more generic measures: likelihood and Frobenius distance.

We will begin our exposition in Section II by describing the various covariance estimators that we will compare, including the SMT. In Section III, we will expand on the SMT estimator

Manuscript received April 20, 2010; revised August 10, 2010; accepted December 11, 2010. Date of publication January 06, 2011; date of current version May 18, 2011. The work of J. Theiler and G. Cao was supported by the Laboratory Directed Research and Development (LDRD) program at Los Alamos National Laboratory. The work of L. R. Bachega and C. A. Bouman was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under Contract/Grant 56541-CI. G. Cao, L. R. Bachega, and C. A. Bouman were supported by the National Science Foundation under Contract CCR-0431024. C. A. Bouman was supported by a Xerox Foundation grant. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gustavo Camps-Valls.

J. Theiler is with the Space and Remote Sensing Group, Los Alamos National Laboratory, Los Alamos, NM 87545 USA (e-mail: jt@lanl.gov).

G. Cao is with GE Healthcare Technologies, Waukesha, WI 53188USA.

L. R. Bachega and C. A. Bouman are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA.

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/JSTSP.2010.2103924

by introducing two new approaches for choosing the model order of the SMT. The actual datasets we will use are described in Section IV, and in Section V we compare these estimators on these datasets using a range of generic and remote sensing metrics. These constitute the main results of this paper, but in Section VI, we consider a recently introduced variant, called graph-constrained SMT [12], and apply that to some of these problems. As a final control experiment, we consider in Section VII the problem of estimating randomly rotated covariance matrices. Finally, we summarize our conclusions in Section VIII.

II. Covariance Estimators

The sample covariance is the most natural and most commonly employed choice for estimating covariance from data. In this section, we will review the justification for the sample covariance, and describe several alternatives, all of which use the sample covariance as a starting point.

A. Sample Covariance

Given n data samples (which, in the case of hyperspectral imagery, are pixels) of dimension p (spectral channels), organized into a data matrix $X = [\mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_n] \in \mathbb{R}^{p \times n}$, the sample covariance is given by $S = (1/n)XX^T = \langle \mathbf{x}\mathbf{x}^T \rangle$, where the angle brackets correspond to the average over the n pixels.[1] Here, S is a $p \times p$ matrix: the diagonal components indicate the magnitude of variation of each of the p spectral channels, and the off-diagonal elements measure the extent to which pairs of channels co-vary with each other.

For a p-dimensional Gaussian distribution with zero mean and covariance matrix $R \in \mathbb{R}^{p \times p}$, the likelihood of observing the data X is given by

$$\ell(X|R) = \frac{|R|^{-n/2}}{(2\pi)^{np/2}} \exp\left[-\frac{1}{2}\mathrm{trace}(X^T R^{-1} X)\right]. \quad (1)$$

We note that

$$\mathrm{trace}(X^T R^{-1} X) = \mathrm{trace}(R^{-1} X X^T) = n\,\mathrm{trace}(R^{-1} S) \quad (2)$$

where $S = (1/n)XX^T$ is the sample covariance. This shows that S is a sufficient statistic for characterizing the likelihood of data X, and we can write

$$\ell(S|R) = \frac{|R|^{-n/2}}{(2\pi)^{np/2}} \exp\left[-\frac{n}{2}\mathrm{trace}(R^{-1} S)\right]. \quad (3)$$

When S is full rank, it is a straightforward exercise [10] to show that this likelihood is maximized when $R = S$. That is to say: the sample covariance is the maximum likelihood estimate of true covariance.

Particularly when the number of data samples is small, however, the sample covariance can *over-fit* the data. For instance, when $n < p$, the sample covariance is necessarily singular, whether or not the actual matrix R is singular. For large n, the sample covariance is a useful estimator in its own right; but even for small n, it provides a starting point for other, more sophisticated, estimates.

[1]For convenience of notation, and because the mean is much easier to estimate than the covariance, we assume means have been subtracted from the data, so that $\langle \mathbf{x} \rangle = 0$.

B. Shrinkage

The notion of shrinkage is based on the intuition that a linear combination of an *over-fit* sample covariance S with some simple *under-fit* approximation to R will lead to an intermediate approximation that is "just right." A positive linear combination *shrinks* the over-fit estimate toward the simple approximation. The simplest and most well-known of these uses the identity matrix I as the under-fit approximation. Actually, a better choice is $(1/p)\mathrm{trace}(S)I$ because it has the same "size" (specifically, the same trace) as the sample covariance S. That is,

$$\widehat{R}_{S-I} = (1-\alpha)S + \alpha(1/p)\mathrm{trace}(S)I. \quad (4)$$

This is sometimes called "ridge" regularization or regularized discriminant analysis [1]. An alternative shrinkage target, proposed by Hoffbeck and Landgrebe [2], uses the matrix $D = \mathrm{diag}(S)$ which agrees with the sample covariance on the diagonal entries, but shrinks the off-diagonal entries toward zero:

$$\widehat{R}_{S-D} = (1-\alpha)S + \alpha D. \quad (5)$$

For both of these estimators, we employ the leave-one-out cross-validation scheme suggested by Hoffbeck and Landgrebe [2]. Although not investigated here, we remark that a number of other shrinkage-based estimators have been proposed [3]–[8].

C. Sparse Matrix Transform (SMT)

Consider a decomposition of the actual covariance matrix into the product $R = E\Lambda E^T$, where E is the orthogonal eigenvector matrix and Λ is the diagonal matrix of eigenvalues. We will similarly decompose the problem of estimating R into the two problems of estimating E and estimating Λ.

In particular, jointly maximizing the likelihood in (3) with respect to E and Λ results in the maximum-likelihood (ML) estimates [10]

$$\widehat{E} = \arg\min_{E \in \Omega} \left\{ \left| \mathrm{diag}(E^T S E) \right| \right\} \quad (6)$$

$$\widehat{\Lambda} = \mathrm{diag}(\widehat{E}^T S \widehat{E}) \quad (7)$$

where Ω is the set of allowed orthogonal transforms. Then $\widehat{R} = \widehat{E}\widehat{\Lambda}\widehat{E}^T$ is the ML estimate of the covariance.

If Ω is the set of *all* orthogonal matrices (and S is full rank), then, as previously noted, the ML estimate of the covariance is given by the sample covariance: $\widehat{R} = S$. The key idea with the SMT is to restrict the set Ω.

In particular, we approximate E with a series of K *Givens rotations*, each of which is a simple rotation of angle θ about two axes i and j. Each rotation is given by a matrix of the form $G = I + \Theta(i, j, \theta)$ where

$$\Theta(i,j,\theta)_{rs} = \begin{cases} \cos(\theta) - 1, & \text{if } r = s = i \text{ or } r = s = j \\ \sin(\theta), & \text{if } r = i \text{ and } s = j \\ -\sin(\theta), & \text{if } r = j \text{ and } s = i \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

- Input: Sample covariance matrix S
- Input: Number of rotations K
- Initialize: $S_0 = S$
- Initialize: F so that $F_{ij} = S_{ij}^2 / (S_{ii} S_{jj})$
- Loop over $k = 1 \dots K$
  - Find $G_k = \text{argmin}_G \left| \text{diag}\left(G^T S_{k-1} G\right) \right|$
    - Let $(i,j) = \text{argmax}_{ij} F_{ij}$
    - Let $\theta = \frac{1}{2}\text{atan}\left(-2(S_{k-1})_{ij}, (S_{k-1})_{ii} - (S_{k-1})_{jj}\right)$
    - Let $G_k = I + \Theta(i,j,\theta)$
  - Update: $S_k = G_k^T S_{k-1} G_k$
  - Update: $F_{ij} = (S_k)_{ij}^2 / \left((S_k)_{ii}(S_k)_{jj}\right)$
- Estimate eigenvector matrix: $\widehat{E} = G_1 G_2 \cdots G_K$
- Estimate eigenvalue matrix: $\widehat{\Lambda} = \text{diag}(S_K)$
- Output: estimated covariance matrix $\widehat{R}_{SMT} = \widehat{E}\widehat{\Lambda}\widehat{E}^T$

Fig. 1. Pseudo-code for sparse matrix transform. For simplicity of exposition, the algorithm shown here assumes that the number of rotations K is known beforehand. In practice, we use methods described in Section III to determine K during the iteration.

Let G_k denote the kth Givens rotation in a sequence. Then we can write $E_K = G_1 G_2 \cdots G_K$ as the product of K Givens rotations. The idea is to use E_K to approximate the eigenvector matrix associated with the true covariance matrix R. The algorithm for finding the G_k's, given the sample covariance S, is shown in Fig. 1 and described in fuller detail in [9] and [10]. The aim is to produce an estimate of the eigenvector matrix E that is sparsely parametrized by a limited number K of rotations. Since the optimal E (in the sense of maximum likelihood) minimizes the product of the diagonal elements of the rotated matrix $E^T S E$, each step of the SMT is designed to find the single Givens rotation that does the most to minimize this product. The SMT algorithm nominally requires $O(Kp^2)$ computations, since there are K rotations, and the updates to the matrix F and the "argmax" of the matrix F appear to require $O(p^2)$ effort. However, the updates to F only affect $O(p)$ of the elements, and it is possible to keep track of the maximum at each iteration with only $O(p)$ effort per iteration. Thus, an efficient implementation of SMT can be performed with $O(Kp) + O(p^2)$ effort. In Section VI, we describe an extension of the SMT, based on graphical constraints, that can lead to further reduction in computational resources, ultimately requiring only $O(p \log p)$ effort [12].

Finally, we remark that we can also use SMT as a shrinkage target to produce another covariance estimator

$$\widehat{R}_{S-\text{SMT}} = (1-\alpha)S + \alpha \widehat{R}_{\text{SMT}} \tag{9}$$

which was introduced along with straight SMT by Cao and Bouman [10]. As with the shrinkage estimators in Section II-B, we choose α based on leave-one-out cross-validation [2].

III. Choosing Model Order for the SMT

Cao and Bouman [9], [10] recommended estimating the model order K (i.e., the number of Givens rotations) for the SMT covariance estimator using a cross-validation approach. This is a reasonable and effective approach, but requires roughly a factor of t times the effort, if t-fold cross-validation is used. The authors recommend $t = 3$. In this section, we introduce two alternatives which are simpler to implement, and do not add significantly to the computation time in learning the SMT from data.

A. Heuristic Wishart Criterion

The idea behind the first criterion is to continue applying Givens rotations (i.e., increasing K) until the matrix $E_K^T S E_K$ is "statistically consistent" with a diagonal matrix. Statistical consistency is measured with respect to the Wishart distribution [13], which describes the distribution of sample covariance matrices, computed from samples drawn from a Gaussian distribution with a parent covariance.

Specifically, the Wishart describes the distribution of XX^T, where X is a $p \times n$ matrix, each column of which is drawn from a p-dimensional Gaussian. So the distribution of the sample covariance $S = (1/n)XX^T$ is given by $S \sim (1/n)\mathcal{W}(R,n)$, where R is the covariance of the parent distribution.

If $W \sim \mathcal{W}(I,n)$ is the random variable that corresponds to a normalized Wishart-distributed matrix, then we have $\langle W \rangle = nI$, where $\langle \cdot \rangle$ corresponds to expectation. Further, one can show that the element w_{ij} of the matrix W has variance given by $\text{Var}(w_{ij}) = n$ for $i \neq j$. We also have $\text{Var}(w_{ii}) = 2n$, though we do not use that in the statistic we will propose next.

Suppose after K SMT rotations, providing approximate eigenvector matrix E_K, we write $S_K = E_K^T S E_K$ which is a "nearly" diagonal matrix. Write $\Lambda_K = \text{diag}(S_K)$ as the diagonal elements of S_K, and consider the "correlation" matrix $\widetilde{S}_K = \Lambda_K^{-1/2} S_K \Lambda_K^{-1/2} = \Lambda_K^{-1/2} E_K^T S E_K \Lambda_K^{-1/2}$. If K is adequately large, then we expect $\widetilde{S}_K$ to be adequately close to the identity; roughly, we expect $\widetilde{S}_K$ to be distributed as a normalized Wishart.

Now, if we write S_{Kij} as the ijth component S_K, then the ijth component of $\widetilde{S}_K = \Lambda_K^{-1/2} S_K \Lambda_K^{-1/2}$ will be given by $\widetilde{S}_{Kij} = S_{Kij} / \sqrt{S_{Kii} S_{Kjj}}$. If we have completed enough rotations that $\widetilde{S}_K \sim (1/n)\mathcal{W}(I,n)$, then we expect $\text{Var}(S_{Kij} / \sqrt{S_{Kii} S_{Kjj}}) = \langle \widetilde{S}_{Kij}^2 \rangle = 1/n$.

Note that $F_{Kij} = \widetilde{S}_{Kij}^2$ is a quantity that we track anyway, when we execute the SMT algorithm. We maintain F_{ij} throughout the computation, and use $\arg\max F_{ij}$ to determine the ij pair to next rotate about. If, while we are doing this, we follow the average of F_{ij} (averaged over the non-diagonal elements of the matrix), we can stop rotating when then average is $O(1/n)$. It bears remarking that the average of F_{ij} can be tracked with only $O(p)$ computation per rotation.

Two further comments are in order. One, we have assumed that the correct number of rotations will produce a Wishart distribution, but this would be the case only if the rotations we applied to S were precisely the same rotations that we would have applied to R. In fact, these rotations are adapted to push S (not R) towards a diagonal matrix as fast as possible. By the time the average value of F_{ij} reaches $1/n$, we can expect that some over-fitting has already occurred. As a second comment, we note that we would prefer to slightly *under-fit* rather than over-fit our estimate to R; that way when we shrink against S to produce $\widehat{R}_{\text{S-SMT}}$, we are balancing an over-fit S with an under-fit R_{SMT} covariance. For these two practical considerations, we consider a modified criterion to stop rotating when the average F_{ij} reaches $2/n$.

B. MDL Approach to Estimating Model Order

An alternative, and arguably less heuristic, approach than the Wishart-based method suggested above, employs the concept of minimum description length (MDL).

Using the standard prescription [14] (see also [15]), we consider the description length of a model M_K that contains K continuous parameters (angles θ), and $2K$ discrete parameters (axes i, j), to explain data with pn scalars. The description length for such a model is given by

$$\mathcal{D}_K = -\log \ell(X|M_K) + \frac{1}{2}K\log(pn) + 2K\log p \tag{10}$$

where the first term is the log likelihood associated with the model, the second term corresponds to the bits in the K angle parameters, and the third term is for bits in the K discrete i, j pairs. The log likelihood follows from (3)

$$\log \ell(X|R) = -\frac{n}{2}\left[p\log(2\pi) + \log|R| + \operatorname{trace}(R^{-1}S)\right] \tag{11}$$

where $S = (1/n)XX^T$ is the sample covariance. Let $S_K = E_K^T S E_K$, be the approximately diagonalized sample covariance, and let $\Lambda_K = \operatorname{diag}(S_K)$ be the diagonal elements of S_K. Then, for SMT, we have that $\widehat{R} = E_K \Lambda_K E_K^T$, and so $\operatorname{trace}(\widehat{R}^{-1}S) = \operatorname{trace}(E_K\Lambda_K^{-1}E_K^T S) = \operatorname{trace}(\Lambda_K^{-1}S_K) = p$ regardless of K. Thus, up to an additive constant

$$\begin{aligned}\mathcal{D}_K &= \frac{n}{2}\log|\Lambda_K| + \frac{1}{2}K\log(pn) + 2K\log p \\ &= \frac{n}{2}\sum_{i=1}^{p}\log\lambda_{Ki} + \frac{1}{2}K\log(pn) + 2K\log p\end{aligned} \tag{12}$$

where λ_{Ki} is the ith diagonal element of S_K.

In general, $\mathcal{D}_K$ will initially decrease with increasing K (model gets better with more rotations, and λ_{Ki} decreases), but as K continues to grow so does the penalty term. We are seeking the value K, where $\mathcal{D}_K$ is minimized (the so-called minimum description length), and so is no longer improved with increasing rotations. We want the first K for which $\mathcal{D}_{K+1} - \mathcal{D}_K \geq 0$. That is,

$$\begin{aligned}n\sum_{i=1}^{p}\log\lambda_{(K+1)i} + (K+1)\log(pn) + 4(K+1)\log p \\ -n\sum_{i=1}^{p}\log\lambda_{Ki} - K\log(pn) - 4K\log p \geq 0\end{aligned} \tag{13}$$

or

$$n\sum_{i=1}^{p}\log\lambda_{(K+1)i}/\lambda_{Ki} \geq -\log(pn) - 4\log p. \tag{14}$$

If the Kth rotation is over axes i, j, then only those two axes are affected:

$$n\log\frac{\lambda_{(K+1)i}\lambda_{(K+1)j}}{\lambda_{Ki}\lambda_{Kj}} \geq -\log(pn) - 4\log p. \tag{15}$$

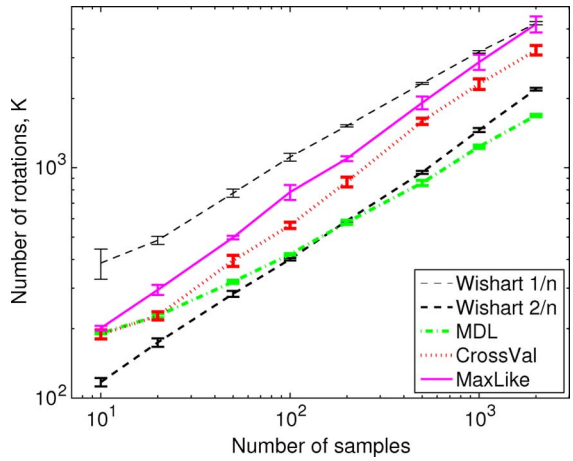


Fig. 2. Number of rotations K chosen by different criteria, as a function of data samples n, for two different hyperspectral covariance matrices. MaxLike is the K that gives the best likelihood function with respect to the real covariance; it is in some sense the "true" K but it is generally unavailable because it requires knowing the true R. CrossVal is three-fold cross-validation. All of the curves are based on three trials, and use the "Cape" dataset.

But we have from [9] that

$$\frac{\lambda_{(K+1)i}\lambda_{(K+1)j}}{\lambda_{Ki}\lambda_{Kj}} = 1 - \frac{S_{Kij}^2}{S_{Kii}S_{Kjj}} = 1 - F_{Kij}. \tag{16}$$

where F_{Kij} is the ij element of the squared correlation matrix that is maintained throughout the computation. As a general trend, F_{Kij} decreases with K though it is not monotonic. Thus, (15) becomes a condition on F_{Kij}

$$n\log(1 - F_{Kij}) \geq -\log(pn) - 4\log p \tag{17}$$

or

$$F_{Kij} \leq 1 - \exp\left[\frac{-(\log n + 5\log p)}{n}\right] \tag{18}$$

and we remark that the ij in this case is the one for which F_{Kij} is maximum. Thus,

$$\max_{ij} F_{Kij} \leq 1 - \exp\left[\frac{-\log n - 5\log p}{n}\right] \tag{19}$$

is the MDL-based stopping condition. When n is large, so that $n \gg \log(pn)$, then this becomes $\max_{ij} F_{Kij} \leq (\log n + 5\log p)/n$.

C. Comparison of Model Order Estimators

In addition to the Heuristic Wishart and MDL-based estimators for K, Fig. 2 includes the cross-validation estimator (CrossVal), and an estimator (MaxLike) that is not generally available in practice because it requires knowledge of the actual covariance R. MaxLike is the choice of K that maximizes the likelihood $\ell(R, \widehat{R}_K)$. The plot shows how the estimated K varies with n for these different model order estimation

TABLE I
DATA SETS AND COVARIANCE MATRICES

Name	Sensor	Location	p	n	Comments
Cape	AVIRIS	Titusville, FL	224	75000	Chip from `f960323t01p02_r04_sc01`
Mall	HYDICE	Washington DC	191	392960	Provided on CD in Landgrebe's book [17]
Blind	HyMap	Cooke City, MT	126	22400	Blind test: http://dirsapps.cis.rit.edu/blindtest/

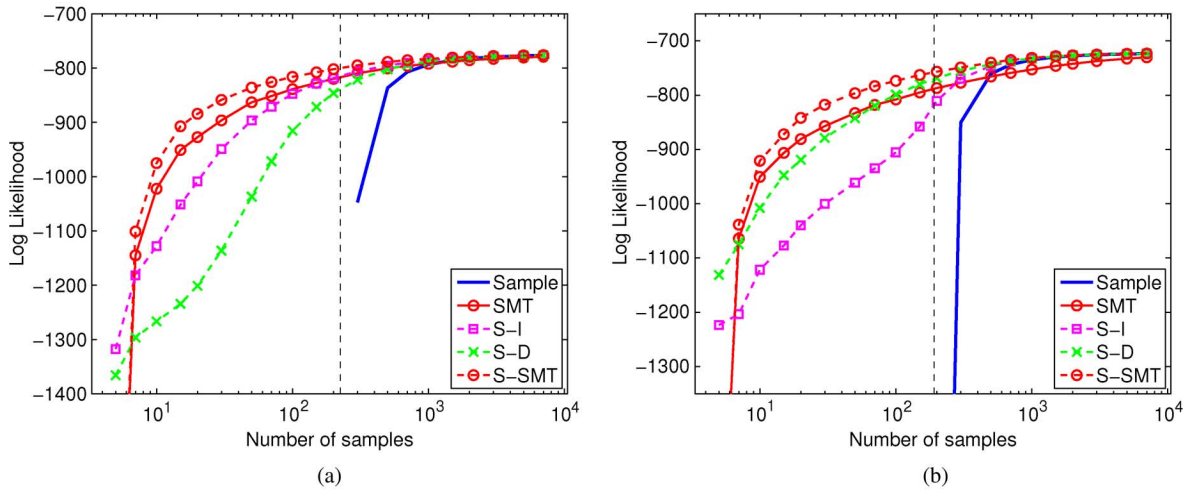


Fig. 3. Log likelihood measure of fitness, as a function of number of samples n, for the (a) "Cape" and (b) "Mall" covariance matrices. Larger values indicate better estimates. The vertical dashed line corresponds to $n = p$.

schemes. For all of these estimators, we observe that as more data is available, larger model orders are called for.

For the experiments in subsequent sections, we use the MDL criterion in (19) to choose the model order K for the sparse matrix transform.

IV. HYPERSPECTRAL DATA SETS AND THEIR COVARIANCE MATRICES

To investigate SMT for hyperspectral signal processing, we have performed some experiments on datasets that are summarized in Table I. The set "Cape" is a 150 $\times$ 500 chip taken over the coast of Florida, near Cape Canaveral, using the 224-channel AVIRIS sensor [16]. We used this AVIRIS data in one of our earlier studies of SMT [11]. The set "Mall" is a 1280 $\times$ 307 pixel 191-channel HYDICE image of the mall in Washington DC [17]. For the "Blind" set, we use one of the four images ("blind radiance") provided as part of the RIT Blind Test dataset [18]. For the change detection experiment, we used both the blind radiance and the "self" radiance images. Both of these images are 800 $\times$ 280 pixels and have 126 spectral channels.

In Section VII, we will consider randomly rotated variants of these datasets; this provides a kind of control experiment which illustrates the importance of the choice of coordinate system for the SMT estimator.

V. MEASURES OF QUALITY FOR COVARIANCE ESTIMATION

In this section, we investigate the performance of the covariance estimators we introduced in Section II. We will begin with some generic measures of quality, the likelihood and two statistics based on Frobenius norm. Then we will consider four measures that are more directly based on problems that arise in hyperspectral imagery: signal detection, dimension reduction, anomaly detection, and anomalous change detection.

A. Likelihood

Following previous comparisons of SMT and other covariance estimators [9], [10], we begin with a likelihood measure. If $\widehat{R}$ is a covariance estimator, then $\ell(S|\widehat{R})$ as defined in (3) is the likelihood of observing a sample covariance S. Our measure of accuracy for $\widehat{R}$ is how likely it would be to observe the actual covariance; in particular we use $\ell(R|\widehat{R})$. Specifically, we use $(1/n)\log\ell(R|\widehat{R})$, or

$$-\frac{1}{2}\left[p\log(2\pi)+\log|\widehat{R}|+\mathrm{trace}(\widehat{R}^{-1}R)\right] \tag{20}$$

as the likelihood measure that is plotted in Fig. 3 for different estimators of different covariance matrices, obtained from the data described in Section IV.

B. Frobenius Norm

The most straightforward way to assess how accurately $\widehat{R}$ approximates R is by element-wise comparison. A natural way to do this is in terms of the Frobenius norm:

$$\|\widehat{R}-R\|_F=\sqrt{\sum_{ij}(\widehat{R}_{ij}-R_{ij})^2} \tag{21}$$

where $\|\cdot\|_F$ indicates the Frobenius norm. It is the square root of the sum of the squares of the elements in the matrix. As Fig. 4(a) shows, the sample covariance provides a good estimate of R in the direct Frobenius sense of (21).

For many remote sensing (and other signal processing) applications, it is not R that one needs to estimate, but its inverse R^{-1}. It is possible to find an approximation $\widehat{R}$ that matches R very closely (in the sense that (21) is small) but for which $\widehat{R}^{-1}$ is a poor approximation to R^{-1}. For instance, for $n < p$, the

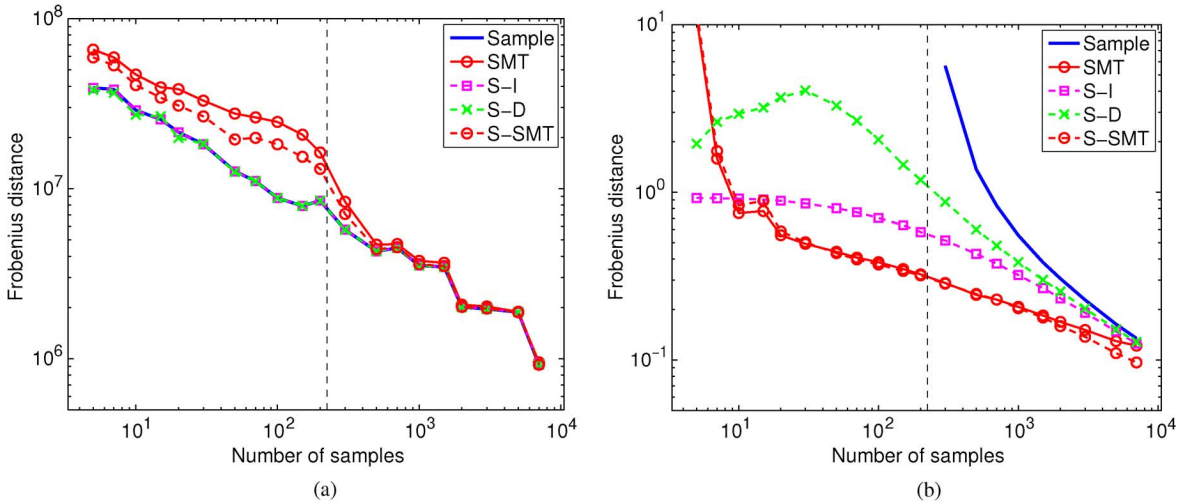


Fig. 4. Best estimator of the inverse is not necessarily the inverse of the best estimator. This figure shows how estimation quality varies with the number n of samples, where quality is defined by the Frobenius norm of the difference, for (a) the covariance matrix directly, and (b) the inverse covariance matrix. For direct covariance estimation, all of the estimators show generally the same performance, and in fact S-I, S-D, and sample covariance all show nearly identical performance. By contrast, there is quite a bit of difference in the performance of these estimators at estimating the inverse. In this case, the SMT algorithms are substantially better over a broad range that includes $n = O(p)$, with S-SMT showing a distinct advantage over straight SMT for larger n. The vertical dashed line corresponds to $n = p$. Results are based on ten trials, using R from the "Cape" dataset. (a) $\|\widehat{R} - R\|_F$. (b) $\|\widehat{R}^{-1} - R^{-1}\|_F$.

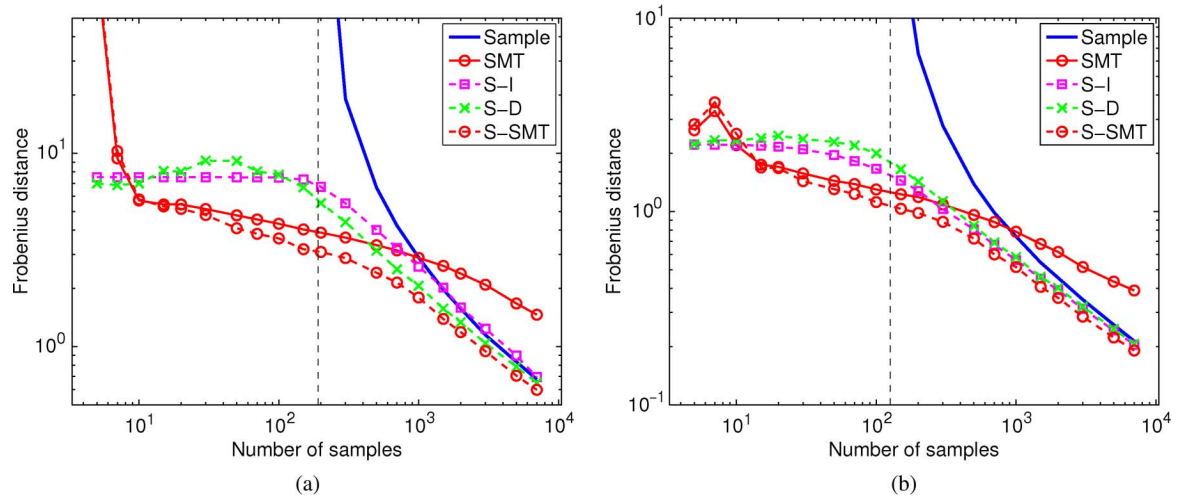


Fig. 5. Same as Fig. 4(b), but for the other two covariance matrices. (a) Mall. (b) Blind.

sample covariance estimator is not even invertible. This suggests a second measure for the utility of an approximation:

$$\|\widehat{R}^{-1} - R^{-1}\|_F \tag{22}$$

In Fig. 4(b), this measure is plotted for the various covariance approximators described in Section II. Further plots in Fig. 5 show the performance for the variety of covariance matrices introduced in Section IV.

C. Matched Filter Signal Detection

Given a weak signal $\mathbf{t}$, one seeks a filter $\mathbf{q} \in \mathbb{R}^p$ which, when applied to an observation $\mathbf{x}$, gives a scalar value $\mathbf{q}^T\mathbf{x}$ which is large when $\mathbf{x}$ contains signal $\mathbf{t}$ and is small otherwise. In particular, the aim is to distinguish two hypotheses:

$$H_0 : \mathbf{x} \sim \mathcal{N}(0, R) \tag{23}$$

$$H_1 : \mathbf{x} \sim \mathcal{N}(a\mathbf{t}, R) \quad \text{for some } a \neq 0 \tag{24}$$

where $\mathcal{N}(\boldsymbol{\mu}, R)$ indicates a normal distribution with mean $\boldsymbol{\mu}$ and covariance R. Following the argument in [11], the signal-to-clutter ratio for a filter $\mathbf{q}$ is given by

$$\mathrm{SCR} = \frac{(\mathbf{q}^T\mathbf{t})^2}{\left\langle (\mathbf{q}^T\mathbf{x})^2 \right\rangle} = \frac{(\mathbf{q}^T\mathbf{t})^2}{\mathbf{q}^T\langle \mathbf{x}\mathbf{x}^T\rangle\mathbf{q}} = \frac{(\mathbf{q}^T\mathbf{t})^2}{\mathbf{q}^T R\mathbf{q}}. \tag{25}$$

The *matched filter* is the vector that optimizes the SCR, and it is given, up to a constant multiplier, by $\mathbf{q} = R^{-1}\mathbf{t}$. Using this $\mathbf{q}$ in (25), we get the optimal SCR:

$$\mathrm{SCR}_o = \frac{(\mathbf{t}^T R^{-1}\mathbf{t})^2}{\mathbf{t}^T R^{-1} R R^{-1}\mathbf{t}} = \mathbf{t}^T R^{-1}\mathbf{t}. \tag{26}$$

If we approximate the matched filter using an approximate covariance matrix $\widehat{R}$, then $\widehat{\mathbf{q}} = \widehat{R}^{-1}\mathbf{t}$ gives

$$\mathrm{SCR} = \frac{(\mathbf{t}^T \widehat{R}^{-1}\mathbf{t})^2}{\mathbf{t}^T \widehat{R}^{-1} R \widehat{R}^{-1}\mathbf{t}} \tag{27}$$

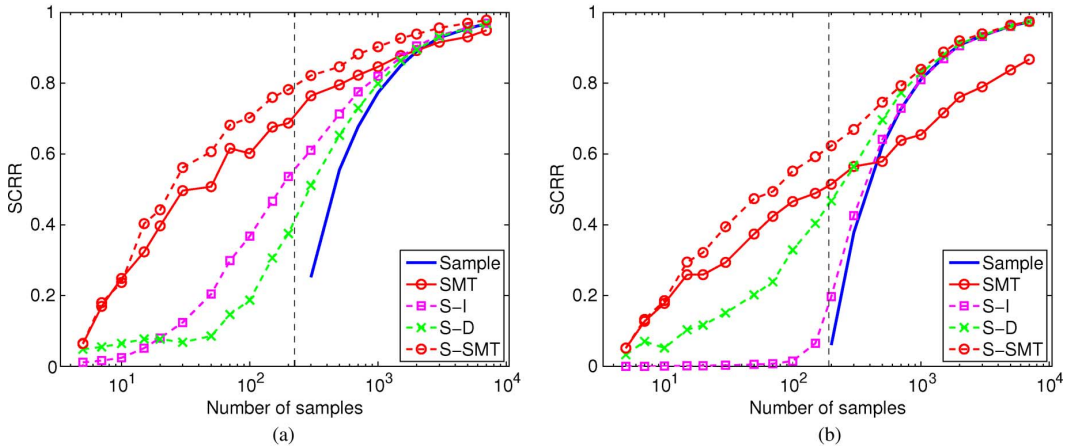


Fig. 6. SCR (signal to clutter ratio) ratio defined in (28) for the (a) "Cape" and (b) "Mall" covariance matrices. Average of ten trials (each trial uses a different signal $\mathbf{t}$ and a different set of n samples from the distribution defined by R). Larger values are better.

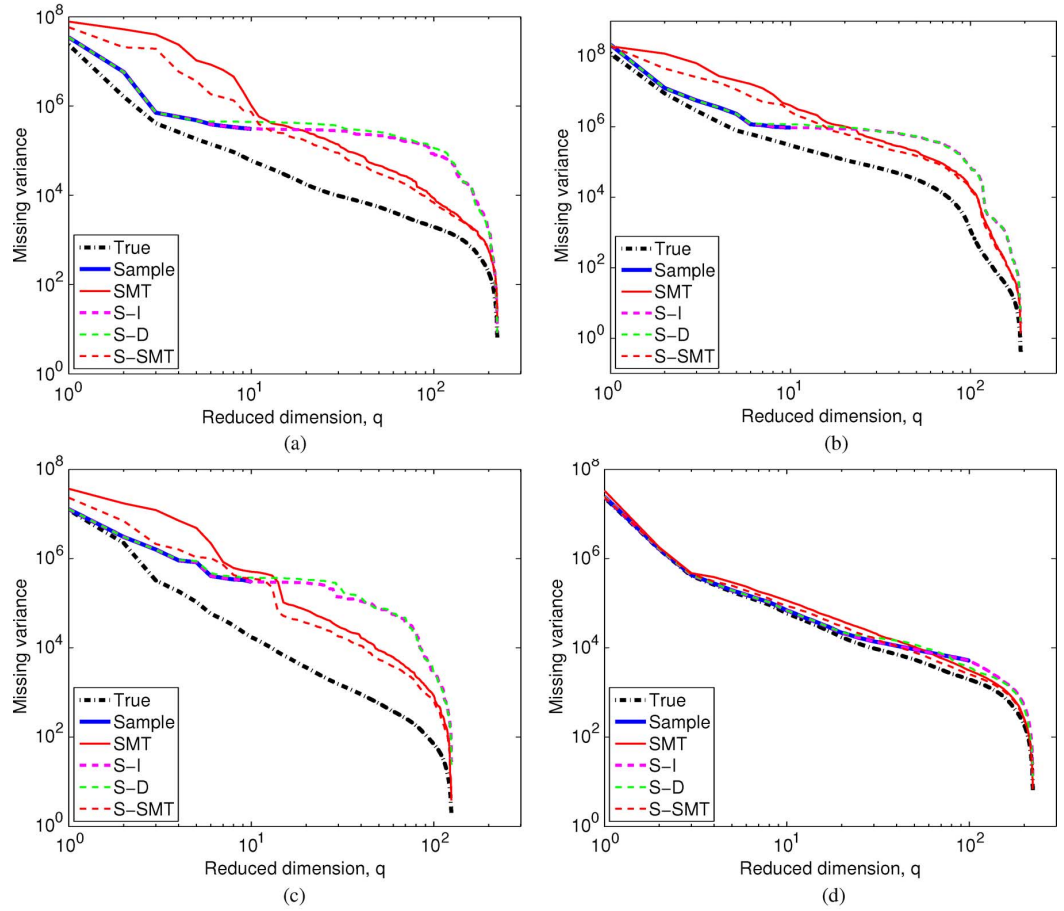


Fig. 7. (a)–(c) Missing variance, defined in (31), when data is reduced to lower dimension, as a function of the lower dimension q; this is for covariance matrices estimated from $n = 10$ samples. Smaller values are better, and the dash-dotted black line is the lower bound obtained when the actual covariance is used. The sample covariance has reduced rank, and does not produce unambiguous estimates beyond $q > n$. (d) For $n = 100$, the trend is still seen, but the effect is much smaller. (a) Cape. (b) Mall. (c) Blind. (d) Cape, $n = 100$.

and the SCRR is the ratio

$$\frac{\mathrm{SCR}}{\mathrm{SCR}_o} = \frac{\left(\mathbf{t}^T \widehat{R}^{-1}\mathbf{t}\right)^2}{(\mathbf{t}^T \widehat{R}^{-1} R \widehat{R}^{-1}\mathbf{t})(\mathbf{t}^T R^{-1}\mathbf{t})}. \tag{28}$$

If $\widehat{R} = R$, then $\mathrm{SCRR} = 1$, but in general $\mathrm{SCRR} \leq 1$.

In Fig. 6, SCRR is plotted against number of samples n for the three covariance matrices under consideration. Although the magnitude of $\mathbf{t}$ does not affect SCRR, the direction does, and the results shown are based on an average of ten trials, each one using a different randomly chosen $\mathbf{t}$. The distribution of directions for $\mathbf{t}$ is isotropic, since each component is drawn from an independent Gaussian distribution.

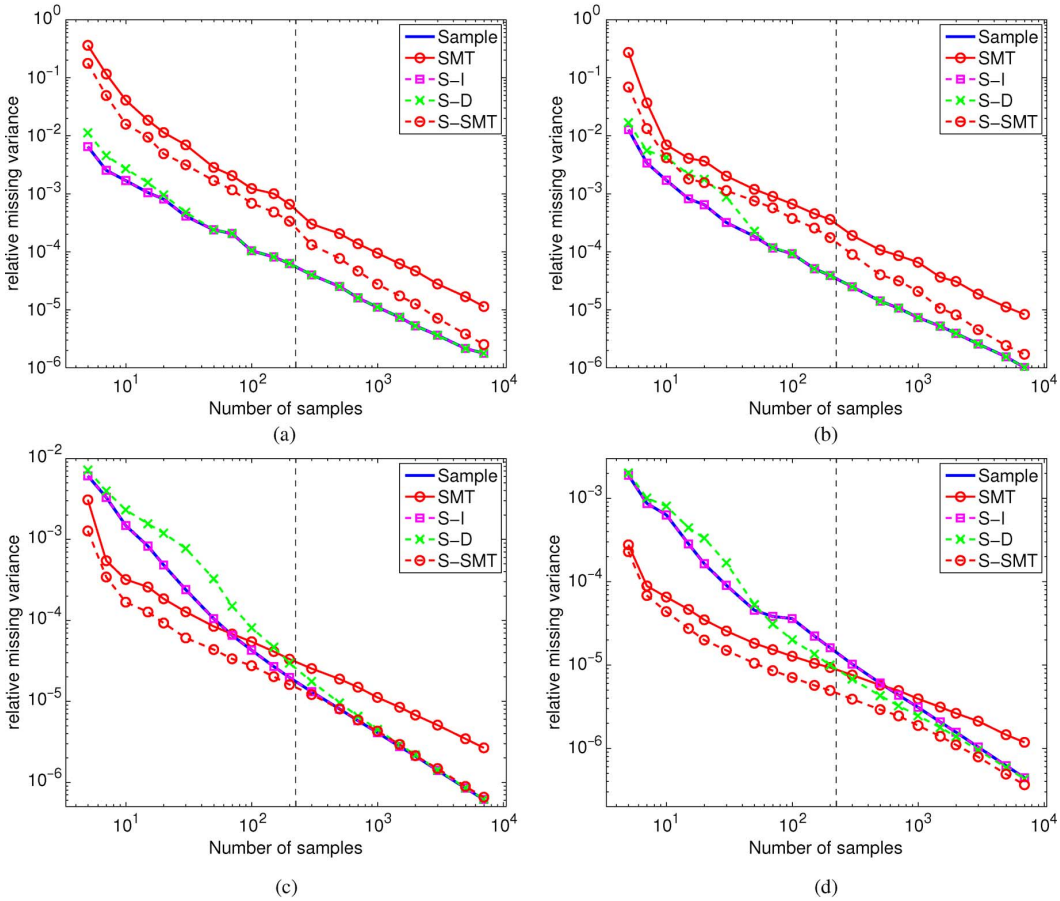


Fig. 8. Relative missing variance, as defined in (32), as a function of n for various q. For small q, the sample covariance is the best choice, but for small n and larger q, the SMT variants outperform the sample covariance. Results are based on the "Cape" covariance. (a) $q = 5$. (b) $q = 10$. (c) $q = 50$. (d) $q = 100$.

D. Dimension Reduction

In the first q principal components, a large fraction of the signal variance is typically captured. We can write $R = E\Lambda E^T$, where E is the matrix of eigenvectors, and Λ is a diagonal matrix of non-negative eigenvalues, which we will order from largest to smallest: $\lambda_1 \geq \cdots \geq \lambda_p$. Let E_q correspond to the first q columns of E. Then the q largest principal components, given by $\mathbf{x}_q = E_q^T\mathbf{x}$ have a variance

$$\begin{aligned}\left\langle \mathbf{x}_q^T\mathbf{x}_q\right\rangle &= \operatorname{trace}\left(\left\langle \mathbf{x}_q\mathbf{x}_q^T\right\rangle\right) = \operatorname{trace}\left(\left\langle E_q^T\mathbf{x}\mathbf{x}^TE_q\right\rangle\right)\\ &= \operatorname{trace}\left(E_q^TRE_q\right) = \operatorname{trace}\left(E_q^TE\Lambda E^TE_q\right)\\ &= \sum_{i=1}^{q}\lambda_i. \end{aligned} \tag{29}$$

But if R is approximated by $\widehat{R} = \widehat{E}\widehat{\Lambda}\widehat{E}^T$, then the first q principal components will be given by $\widehat{E}_q$ instead of E_q, and the variance will be smaller. Specifically,

$$\begin{aligned}\left\langle \widehat{\mathbf{x}}_q^T\widehat{\mathbf{x}}_q\right\rangle &= \operatorname{trace}\left(\left\langle \widehat{\mathbf{x}}_q\widehat{\mathbf{x}}_q^T\right\rangle\right) = \operatorname{trace}\left(\left\langle \widehat{E}_q^T\mathbf{x}\mathbf{x}^T\widehat{E}_q\right\rangle\right)\\ &= \operatorname{trace}\left(\widehat{E}_q^TR\widehat{E}_q\right) = \operatorname{trace}\left(\widehat{E}_q^TE\Lambda E^T\widehat{E}_q\right)\\ &= \sum_{i=1}^{p}\alpha_i\lambda_i \end{aligned} \tag{30}$$

where $\alpha_i = \sum_j \left(\widehat{E}_q^TE\right)_{ji}^2$. Note that $0 \leq \alpha_i \leq 1$ and $\sum_{i=1}^p \alpha_i = q$. Since the λ_i's are sorted in descending order, it follows that $\sum_{i=1}^p \alpha_i\lambda_i \leq \sum_{i=1}^q \lambda_i$. Equality holds when $\widehat{E}_q$ spans the same subspace as E_q; that is: $\widehat{E}_q = E_qU$ for some orthogonal $q \times q$ matrix U.

Using q instead of p principal components, one will fail to capture the variance in the last $p - q$ components: $\sum_{i=1}^p \lambda_i - \sum_{i=1}^q \lambda_i = \sum_{i=q+1}^p \lambda_i$. This is the best case, and is achieved when accurate principal components are used. If approximate principal components are used, then the missing variance is given by $\sum_{i=1}^p \lambda_i - \sum_{i=1}^q \alpha_i\lambda_i$. Thus, the cost of using the approximation, in terms of missing variance, is given by the difference

$$\left[\sum_{i=1}^{p}\lambda_i - \sum_{i=1}^{q}\alpha_i\lambda_i\right] - \left[\sum_{i=1}^{p}\lambda_i - \sum_{i=1}^{q}\lambda_i\right] = \sum_{i=1}^{q}(1-\alpha_i)\lambda_i \tag{31}$$

To make this quantity dimensionless, we divide by the missing variance that would be obtained using accurate principal components. This gives the *relative missing variance* that we use in our plots:

$$\frac{\sum_{i=1}^{q}(1-\alpha_i)\lambda_i}{\sum_{i=q+1}^{p}\lambda_i}. \tag{32}$$

We remark that for the SMT case (but not the S-SMT, unfortunately), the encoding of the $\widehat{E}_q$ matrix as a product

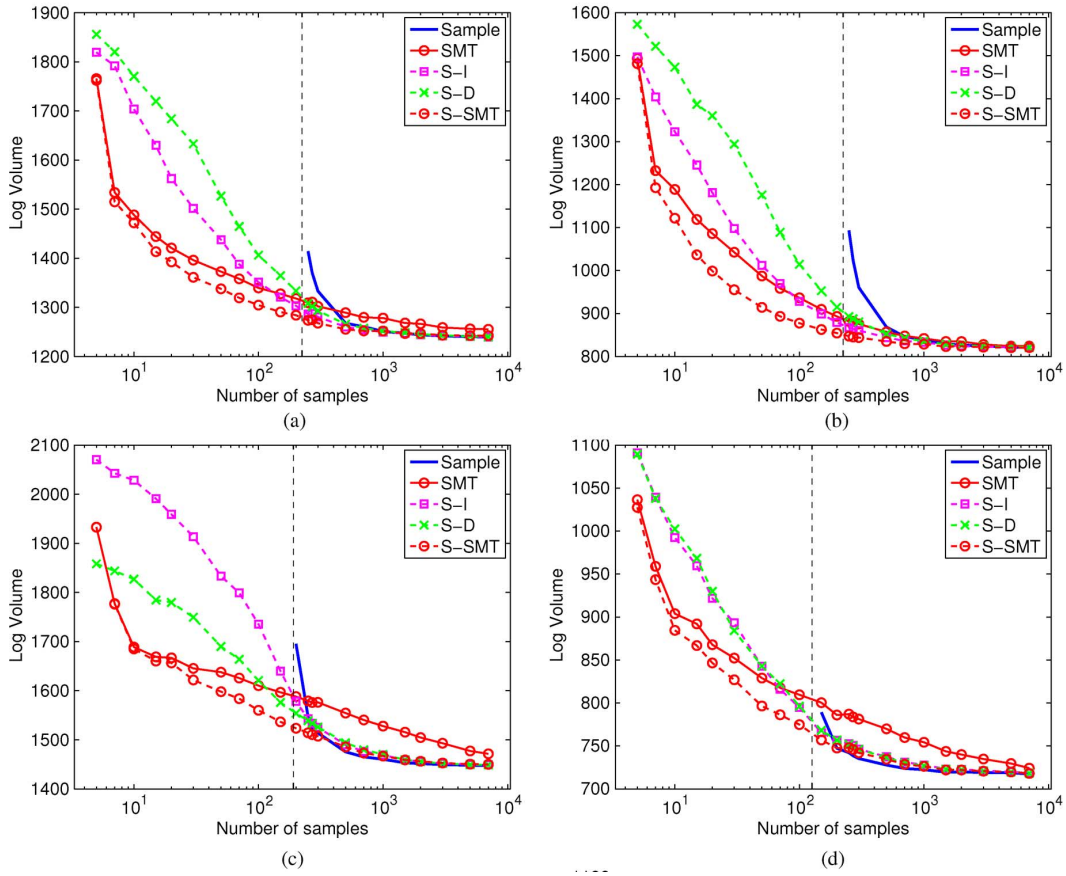


Fig. 9. Volume of ellipsoids that cover 99.9% of the data. Smaller values are "tighter" (and therefore better) fits. (a) Cape. (b) Cape, Gaussian. (c) Mall. (d) Blind.

of K Givens rotations provides a computational advantage in the application of dimension reduction. Instead of directly multiplying $\widehat{\mathbf{x}} = E_q^T \mathbf{x}$ which requires $O(pq)$ computations, use $\widehat{\mathbf{x}} = P_q G_K^T G_{K-1}^T \cdots G_1^T \mathbf{x}$, where G_k is the kth Givens rotation, and P_q is the projection matrix that picks off the q channels with highest variance. This is only $O(K)$ computations [19] and typically $K = O(p)$, as has been shown previously [9], [10], [12].

In Fig. 7 and Fig. 8, we see that SMT does well at minimizing the missing variance when n is small and q is large.

E. Anomaly Detection

The aim of anomaly detection is to distinguish "typical" data, which are presumed to be sampled from a parent distribution, from anomalies (or outliers) that are not. The RX anomaly detector [20] treats this parent distribution as Gaussian with covariance matrix R, and measures anomalousness in terms of the squared Mahalanobis distance from the origin. That is, when

$$\mathbf{x}^T R^{-1} \mathbf{x} > \eta^2 \tag{33}$$

for some threshold radius η, then $\mathbf{x}$ is considered anomalous.

A proxy for missed detection rate is the volume of the ellipsoid contained within $\mathbf{x}^T R^{-1} \mathbf{x} \leq \eta^2$; it is given by

$$V(R, \eta) = \frac{\pi^{p/2}}{\Gamma(1 + p/2)} |R|^{1/2} \eta^p. \tag{34}$$

The false alarm rate is the fraction of the measure for which $\mathbf{x}^T \widehat{R}^{-1} \mathbf{x} > \eta^2$. Thus, one can for instance vary η and plot out a "coverage" plot of volume versus false alarm rate, as proposed in [21]. Alternatively, as done here, one can choose η so that a given false alarm rate is achieved, and then use volume as a measure of quality. This is done in Fig. 9, and it is seen that SMT provides covariance approximations that provide smaller volumes for a given coverage (here, chosen as 99.9% of the data). Smaller volumes correspond to fewer missed detections, and the fixed coverage corresponds to a constant false alarm rate (here, of 0.1%).

We remark that the anomaly detection problem is closely related to the likelihood metric described in Section V-A. The coverage is the fraction of data enclosed by this ellipsoid for which $\mathbf{x}^T \widehat{R}^{-1} \mathbf{x} < \eta^2$, where $\mathbf{x}$ is drawn from the Gaussian with covariance given by R. We want this fraction to be large, so we want to choose $\widehat{R}$ so that $\mathbf{x}^T \widehat{R}^{-1} \mathbf{x}$ is small, but we can write $\langle \mathbf{x}^T \widehat{R}^{-1} \mathbf{x} \rangle = \mathrm{trace}(\widehat{R}^{-1} R)$. This encourages us to make $\widehat{R}$ large (i.e., to have a large ellipsoid) because that gives a lower false alarm rate. But we also want the ellipsoid to be small, since the missed detection rate will scale with the volume of the ellipsoid. This volume in turn scales with the determinant $|\widehat{R}|$. To trade off these conflicting interests, one can try to minimize a formula such as

$$\log |\widehat{R}| + \mathrm{trace}(\widehat{R}^{-1} R) \tag{35}$$

which has the same form as the negative log likelihood function in (20).

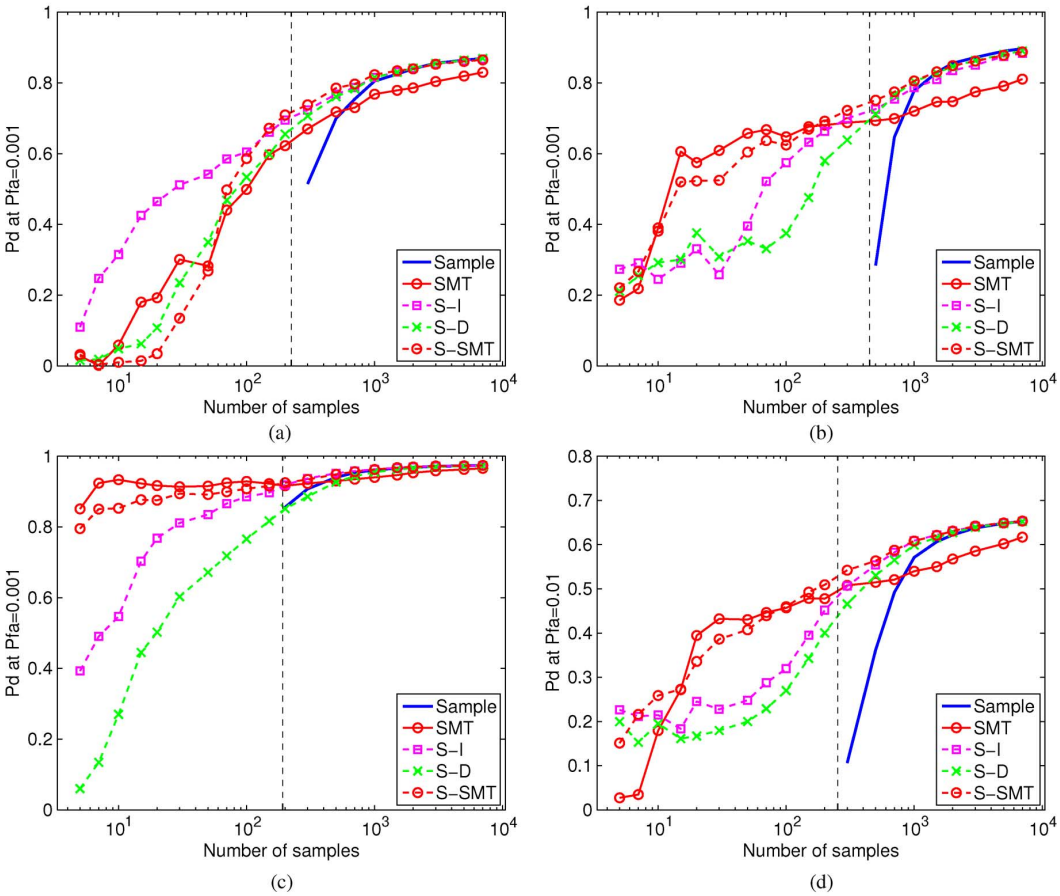


Fig. 10. Performance is measured by the detection rate achieved at a specified false alarm rate. Larger values are better. For the "Cape" data set, the pervasive differences are simulated: (a) channels are split, so the first 112 channels are considered the first image, and the last 112 channels are taken to be the second image. (b) The image is smoothed with a 3×3 kernel and translated by one pixel to simulate misregistration. (c) For the "Mall" data, the channels are similarly split, with 95 bands going to the first image, and 96 to the second image. (d) For the "Blind" dataset, two separate images were used, the "blind radiance" image and the "self radiance" image. (a) Cape-split. (b) Cape-misreg. (c) Mall-split. (d) Blind.

F. Anomalous Change Detection

In the anomalous change detection problem [22]–[25], the aim is to find the small rare changes, without being confounded by the pervasive differences between the two images that might, for instance, be due to camera calibration, atmospheric conditions, or seasonal variation.

Many of the algorithms for anomalous change detection can be recast in a formulation that requires the estimation of a large covariance matrix [24]. For this study, we employed the hyperbolic anomalous change detection (HACD) algorithm, first introduced in [25]. Following the simulation framework described in [24], we simulated both the pervasive differences and the anomalous changes. For one of the cases we used the actual pervasive differences observed in two images taken approximately an hour apart [18]. The pervasive differences are simulated in Fig. 10(a)–(c), but actual pervasive differences are used in Fig. 10(d).

Whereas SMT provided substantial gains for the straight anomaly detection problem, the results for anomalous change detection are more ambiguous. For large sample size n, we found that SMT did not provide competitive performance (though S-SMT generally did). For the small n case, there were some examples (Cape-misreg, Mall-split, and Blind) where SMT and S-SMT both outperformed the competitors, and one case (Cape-split) where S-I was superior.

VI. SMT With Graph-Based Constraints

This section describes results from a recently introduced variant of SMT, called graph-constrained SMT [12], that has, as its main advantage, a more computationally efficient implementation when the dimension p is particularly large. The method also exploits *a priori* information about the structure of a covariance matrix, and has the potential for more accurate estimation than standard SMT provides.

Graph-constrained SMT employs graph-based constraints in order both to further limit size of the set Ω from (6), and to reduce the computational effort required to identify the Givens rotations $G_1, \ldots, G_K$. During the computation of each of the K Givens rotations in the SMT design, the greedy search for the most correlated pair of coordinates ij is replaced by a constrained search over the pairs of nodes in the graphical structure. As each rotation is designed, the graphical structure evolves in a way that the neighborhood relations are re-examined and pruned, with each coordinate keeping correlation information of no more than m neighbors.

The motivation for using the graphical-SMT in the context of hyperspectral image processing is based on the observation that

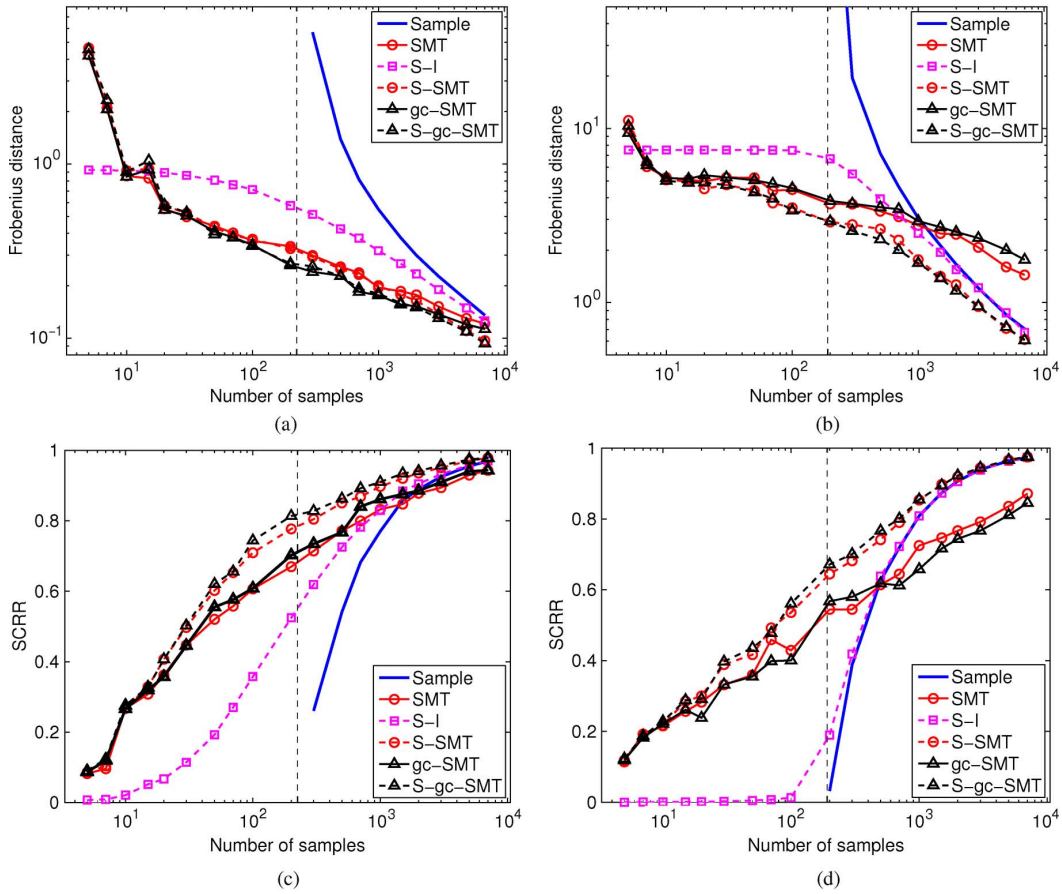


Fig. 11. Graph-constrained SMT using $m = 16$ neighbors. These runs are for a single trial. Panels (a), (b) correspond to Fig. 4(b) and Fig. 5(a). Panels (c), (d) correspond to Fig. 6(a), (b). We observe that the cheaper graph-constrained SMT provides comparable performance to standard SMT on these problems.

two neighboring bands i and j, such that $|i - j| \leq m/2$ for a small threshold m, tend to be among the most correlated ones. Therefore, we derive a form of graphical constraint among the pairs of bands and use the graphical-SMT algorithm to estimate the covariance matrix of the hyperspectral data at a much lower computational cost than with the standard SMT algorithm. This suits our purpose particularly well when p is very large and m can be set to be relatively small, resulting in $O(p \log p)$ computational complexity as opposed to $O(p^2)$ of the standard SMT algorithm without significant differences in the quality of the resulting estimators [12].

Fig. 11 compares the covariance estimators produced with both the SMT and the graphical-SMT algorithms as well as with other methods. The results suggest that both variants of the SMT algorithm produce similar results over a wide range of sizes of the training set.

Finally, the number of degrees of freedom is smaller in the graphical-SMT than in the standard SMT since only the m most correlated neighbors of each coordinate are considered during the SMT computation. This fact suggests a minor modification in the MDL criterion used to choose K. Instead of (19), we use

$$\max_{ij} F_{Kij} \leq 1 - \exp\left[\frac{-\log n - 3\log p - 2\log m}{n}\right] \quad (36)$$

which accounts for the neighborhood of size m imposed by the graphical constraint.

VII. Randomly Rotated Covariance Matrices

In this section, we perform a control experiment to help explain and delimit the range of situations for which SMT is most suitable.

One reason SMT outperforms other covariance estimators is that it takes advantage of the tendency for real data, and for hyperspectral data in particular, to have its eigenvectors aligned with the natural axes of the system. This argument suggests that SMT would lose its power if this alignment were randomly rotated. In particular, given a covariance matrix R, and a random orthogonal matrix Q, one can generate a new covariance $R' = Q^T R Q$ which will have the same eigenvalues as R, but the eigenvectors will essentially be random. In Cao *et al.* [11], it was argued that structure in the eigenvectors is what enabled SMT to outperform its competitors.

With that in mind, we performed some experiments with rotated covariance matrices, shown in Fig. 12. As expected, neither SMT nor S-SMT did as well as simple shrinkage against the scaled identity (S-I, in the plots). Even S-D was outperformed by the rotationally invariant S-I.

VIII. Conclusion

We have identified a series of signal processing problems in remote sensing that require, at some stage of the computation, an estimation of covariance matrices from limited data. For each of these problems, and in each case for different hyperspectral

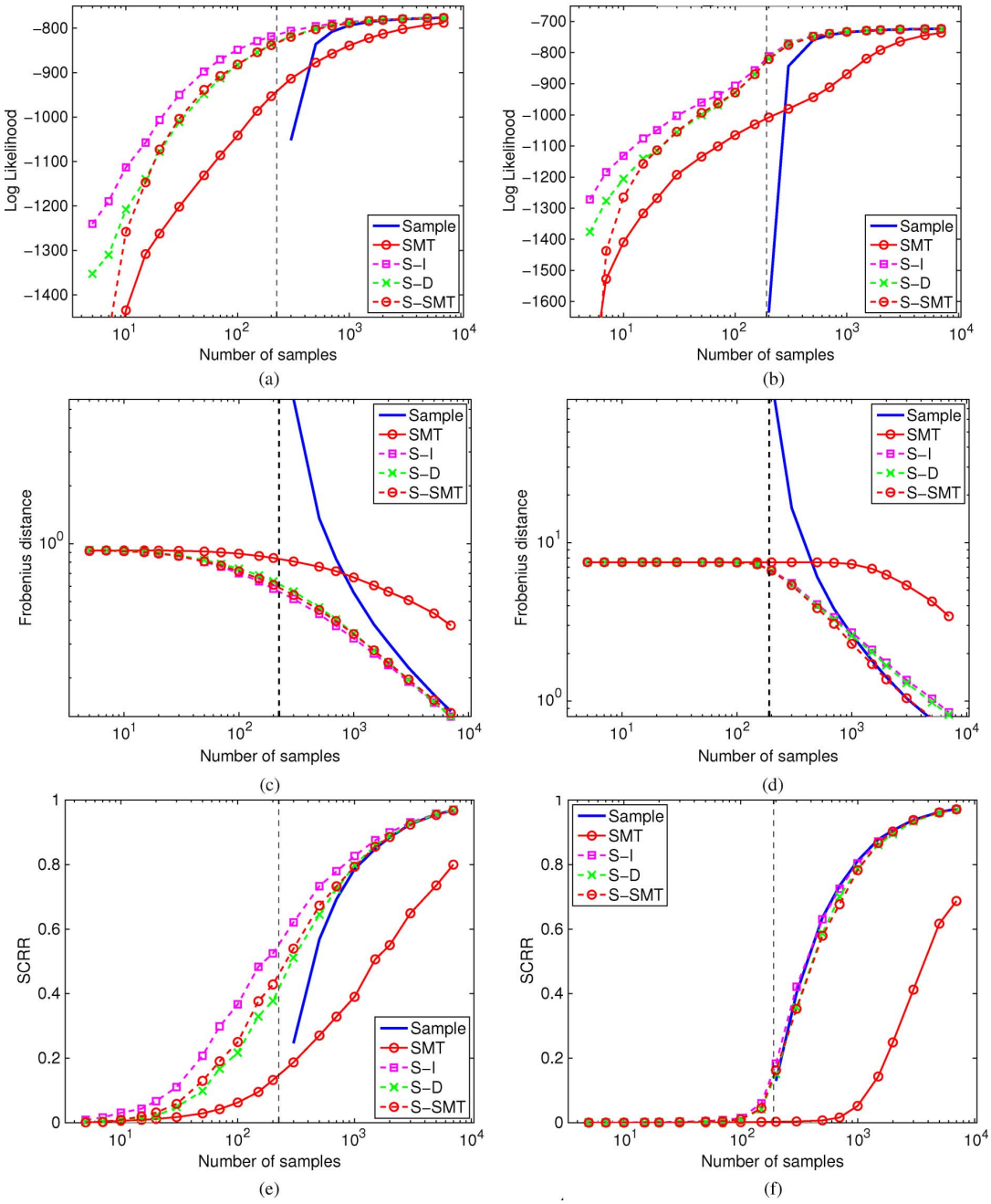


Fig. 12. Performance against randomly rotated covariance matrices for the likelihood measure of goodness. Average of 10 trials. (a), (b) Log likelihood metric in (20): larger values are better. (c), (d) Frobenius distance between inverses, defined in (22): smaller values are better. (e), (f) SCR ratio in (28): larger values are better. (a) Cape. (b) Mall. (c) Cape. (d) Mall. (e) Cape. (f) Mall.

data sets, we have applied a suite of covariance approximation schemes, and compared their performance.

For the likelihood, the Frobenius distance between inverses, the matched filter detection of small signals, and the anomaly detection experiments, the sparse matrix transform with shrinkage (S-SMT) consistently outperformed the other estimators. This was most noticeable when the number of samples n was smaller than the number of channels p. For $n \gg p$, all of the estimators gave similar performance.

The dimension reduction experiments were somewhat enlightening. When the reduced dimension q was very small, we found that the sample covariance actually did a better job than the other estimators. However, for large q and small n, the SMT and S-SMT estimators did the best job. This is consistent with the intuition that sample covariance does a better job at estimating the performance of the large eigenvalues, but for tasks where the small eigenvalues are important, SMT appears advantageous.

The results with anomalous change detection were mixed. In some cases the detection rates were higher for SMT and in other cases, they were lower. This inconclusiveness was observed even with the same data set ("Cape") using different simulated pervasive differences. For the one case ("Blind") where the pervasive differences were not simulated, the S-SMT

outperformed the other estimators for $n < p$. We found these results initially surprising, given the definitive advantage SMT provided for straight anomaly detection. But where straight anomaly detection seeks anomalies that have significant components in the eigenspace of the low eigenvalues (where SMT seems to work better than competitors), the anomalous changes have more of (but not all of) their energy in large eigenvalue components.

We have introduced two new model order estimation schemes for automatically determining the number K of Givens rotations to be applied in an SMT estimator. Both of these schemes can be computed with minimal overhead since they are both based on values of the squared correlation matrix F which is already computed as an intermediate step in the SMT algorithm.

Numerical experiments suggest that the heuristic Wishart estimator tends to provide an over-estimate of the number of needed rotations, whereas the MDL approach tends to provide an underestimate. When used with a shrinkage, one generally prefers a smaller K, and this favors the MDL approach.

To some extent, the occasional failure of S-SMT to outperform S-D or the sample covariance can be attributed to a failure in the choice of model order K. (This failure was sometimes observed when the number of samples was very small, $n < 10$.) That is because SMT with $K = 0$ is equivalent to the diagonal matrix D, and SMT with $K = O(p^2)$ approaches the sample covariance.

We described a smaller number of experiments with the recently introduced graph-constrained SMT [12], and found that results essentially identical to SMT were obtained using this more restricted and more computationally efficient variant. We attribute this to the tendency of hyperspectral data to be more highly correlated for adjacent wavelengths.

Finally, by comparing the performance of these estimators on randomly rotated variants of the real hyperspectral covariance matrices, and finding that the advantages of SMT were lost, we confirmed the interpretation that SMT is exploiting an approximate alignment that is empirically observed to occur the eigenvectors and the natural coordinates of the data.

REFERENCES

[1] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, pp. 165–175, 1989.

[2] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp. 763–767, Jul. 1996.

[3] C. Lee and D. A. Landgrebe, "Analyzing high-dimensional multispectral data," *IEEE Trans. Geoscience and Remote Sensing*, vol. 31, no. 4, pp. 792–800, Jul. 1993.

[4] P. V. Villeneuve, H. A. Fry, J. Theiler, B. W. Smith, and A. D. Stocker, "Improved matched-filter detection techniques," in *Proc. SPIE*, 1999, vol. 3753, pp. 278–285.

[5] J. Schafer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statist. Applicat. Genetics Molec. Biol.*, vol. 4, no. 1, 2005, article 32.

[6] M. J. Daniels and R. E. Kass, "Shrinkage estimators for covariance matrices," *Biometrics*, vol. 57, no. 4, pp. 1173–1184, 2001.

[7] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Ann. Statist.*, vol. 36, no. 1, pp. 199–227, 2008.

[8] N. M. Nasrabadi, "Regularization for spectral matched filter and RX anomaly detector," in *Proc. SPIE*, 2008, vol. 6966, p. 696604.

[9] G. Cao and C. A. Bouman, "Covariance estimation for high dimensional data vectors using the sparse matrix transform," School of Elect. and Comput. Eng., Purdue Univ., West Lafayette, IN, Tech. Rep. TR-ECE-08-05, April 2008.

[10] G. Cao and C. A. Bouman, "Covariance estimation for high dimensional data vectors using the sparse matrix transform," in *Advances in Neural Information Processing Systems 21*. Cambridge, MA: MIT Press, 2009, pp. 225–232.

[11] G. Cao, C. A. Bouman, and J. Theiler, "Weak signal detection in hyperspectral imagery using sparse matrix transform (SMT) covariance estimation," in *Proc. WHISPERS (Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing)*, 2009, pp. 1–4.

[12] L. R. Bachega, G. Cao, and C. A. Bouman, "Fast signal analysis and decomposition on graphs using the sparse matrix transform," in *Proc. ICASSP*, 2010, pp. 5426–5429.

[13] J. Wishart, "The generalized product moment distribution in samples from a normal multivariate population," *Biometrika*, vol. 20A, pp. 32–52, 1928.

[14] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[15] G. E. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[16] G. Vane, R. O. Green, T. G. Chrien, H. T. Enmark, E. G. Hansen, and W. M. Porter, "The Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)," *Remote Sens. Environ.*, vol. 44, pp. 127–143, 1993.

[17] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. New York: Wiley, 2003.

[18] D. Snyder, J. Kerekes, I. Fairweather, R. Crabtree, J. Shive, and S. Hager, "Development of a web-based application to evaluate target finding algorithms," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2008, vol. 2, pp. 915–918.

[19] J. Theiler, G. Cao, and C. A. Bouman, "Sparse matrix transform for fast projection to reduced dimension," in *Proc. IGARSS*, 2010, pp. 4362–4365.

[20] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 10, pp. 1760–1770, Oct. 1990.

[21] J. Theiler and D. Hush, "Statistics for characterizing data on the periphery," in *Proc. IGARSS*, 2010, pp. 4764–4767.

[22] A. Schaum and A. Stocker, "Long-interval chronochrome target detection," in *Proc. 1997 Int. Symp. Spectral Sensing Res.*, 1998.

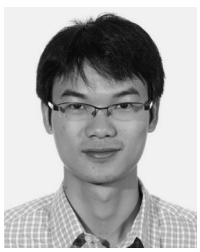
[23] A. Schaum and A. Stocker, "Hyperspectral change detection and supervised matched filtering based on covariance equalization," in *Proc. SPIE*, 2004, vol. 5425, pp. 77–90.

[24] J. Theiler, "Quantitative comparison of quadratic covariance-based anomalous change detectors," *Appl. Opt.*, vol. 47, pp. F12–F26, 2008.

[25] J. Theiler and S. Perkins, "Proposed framework for anomalous change detection," in *Proc. ICML Workshop Mach. Learn. Algorithms Surveill. Event Detect.*, 2006, pp. 7–14.

James Theiler (M'04) received the Ph.D. degree in physics from the California Institute of Technology, Pasadena, in 1987.

After the Ph.D. degree, he held appointments at UCSD, MIT Lincoln Laboratory, Los Alamos National Laboratory, and the Santa Fe Institute. He joined the Space and Remote Sensing Sciences Group, Los Alamos National Laboratory, Los Alamos, NM, in 1994, and was named a Laboratory Fellow in 2005. His professional interests include image processing, remote sensing, and machine learning. Also, covariance matrices.



Guangzhi Cao (S'07–M'10) received the B.S. and M.S. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2002 and 2004, respectively, and the Ph.D. from Purdue University, West Lafayette, IN, in 2009. He is a Scientist of advanced algorithms in the CT System Group, GE Healthecare Technologies, Waukesha, WI. He had an internship with the GE Global Research Center, Shanghai, China, in 2004 and the Los Alamos National Laboratory, Los Alamos, NM, in 2009. His current research interests include statistical signal and image processing, computed tomography, inverse problems, machine learning, and computer vision.

Leonardo R. Bachega (S'10) received the B.S. degree in physics and the M.S. degree in electrical and computer engineering from the University of Sao Paulo, Sao Paulo, Brazil, in 2000 and 2004, respectively, and the M.S.E.C.E. degree from Purdue University, West Lafayette, IN, in 2010. He is currently pursuing the Ph.D. degree in electrical and computer engineering at Purdue University.

From 2002 to 2004, he was with the Blue Gene/L Software Team at the IBM T. J. Watson Research Center, Yorktown Heights, NY. His research focuses on developing pattern recognition and machine learning algorithms for high-dimensional data using the sparse matrix transform. His main interests include statistical signal processing, machine learning, image processing, and high-performance computing.

Charles A. Bouman (S'86–M'89–SM'97–F'01) received the B.S.E.E. degree from the University of Pennsylvania, Philadelphia, in 1981, the M.S. degree from the University of California at Berkeley in 1982, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, in 1989.

From 1982 to 1985, he was a Full Staff Member at MIT Lincoln Laboratory. Lexington, MA, He joined the faculty of Purdue University, West Lafayette, IN, in 1989 where he is currently the Michael J. and Katherine R. Birck Professor of Electrical and Computer Engineering. He also holds a courtesy appointment in the School of Biomedical Engineering and is co-director of Purdue's Magnetic Resonance Imaging Facility located in Purdue's Research Park. His research focuses on the use of statistical image models, multiscale techniques, and fast algorithms in applications including tomographic reconstruction, medical imaging, and document rendering and acquisition.

Prof. Bouman is a Fellow of the American Institute for Medical and Biological Engineering (AIMBE), a Fellow of the society for Imaging Science and Technology (IS&T), and a Fellow of the SPIE professional society. He is also a recipient of IS&T's Raymond C. Bowman Award for outstanding contributions to digital imaging education and research, has been a Purdue University Faculty Scholar, and received the College of Engineering Engagement/Service Award, and Team Award. He was previously the Editor-in-Chief for the IEEE TRANSACTIONS ON IMAGE PROCESSING, and is currently a member of the Board of Governors and a Distinguished Lecturer for the IEEE Signal Processing Society. He has been an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He has also been Co-Chair of the 2006 SPIE/IS&T Symposium on Electronic Imaging, Co-Chair of the SPIE/IS&T conferences on Visual Communications and Image Processing 2000 (VCIP), a Vice President of Publications and a member of the Board of Directors for the IS&T Society, and he is the founder and Co-Chair of the SPIE/IS&T conference on Computational Imaging.