Comparative Analysis of Covariance Matrix Estimation for Anomaly Detection in Hyperspectral Images

Santiago Velasco-Forero, Marcus Chen, Alvina Goh, Sze Kim Pang

Comparative Analysis of Covariance Matrix Estimation for Anomaly Detection in Hyperspectral Images

Santiago Velasco-Forero, Marcus Chen, Alvina Goh and Sze Kim Pang

Abstract

Covariance matrix estimation is fundamental for anomaly detection, especially for the Reed and Xiaoli Yu (RX) detector. Anomaly detection is challenging in hyperspectral images because the data has a high correlation among dimensions, heavy tailed distributions and multiple clusters. This paper comparatively evaluates modern techniques of covariance matrix estimation based on the performance and volume the RX detector. To address the different challenges, experiments were designed to systematically examine the robustness and effectiveness of various estimation techniques. In the experiments, three factors were considered, namely, sample size, outlier size, and modification in the distribution of the sample.

1 Introduction

Hyperspectral (HS) imagery provides rich information both spatially and spectrally. Differing from the conventional RGB camera, HS images measure scientifically the radiance received at fine divided bands across a continuous range of wavelengths. These images enable grain-fine classification of materials otherwise undistinguishable in spectrally reduced sensors. Anomaly detection (AD) using HS images is particularly promising in discovering the subtlest difference among a set of materials. AD is a target detection problem, in which there is no prior knowledge about the spectra of the target of interest [1]. In other words, it aims to detect spectrally anomalous targets. However, the definition of anomalies varies. Practically, HS anomalies are referred to as materials semantically different from the background, such as a target in the homogeneous background [2]. Unfortunately, often the backgrounds are a lot more complex due to presence of multiple materials, which could be spectrally mixed at pixel levels.

Many AD methods have been proposed, and a few literature reviews or tutorials have been thoroughly done [1]–[6]. Recently, the tutorial by [5] gives a good overview of different AD methods in the literature. However, it does not give any experimental comparison. Differing from these reviews, this paper comparatively surveys the existing AD methods via background modeling by covariance matrix estimation techniques. In this manuscript, we analyze the AD in the context of optimal statistical detection, where the covariance matrix of the background is required to be estimated.

The aim of covariance matrix estimation is to compute a matrix $\widehat{\mathbf{\Sigma}}$ that is "close" to the actual, but unknown, covariance $\mathbf{\Sigma}$. We use "close" because that $\widehat{\mathbf{\Sigma}}$ should be an approximation that is useful for he given task at hand. The *sample covariance matrix* (SCM) is the maximum likelihood estimator, but it tends to overfit the data when n does not greatly exceed d. Additionally, in the presence of multiple clusters, this estimation fails to characterize correctly the background. For these reasons, a variety of regularization schemes have been proposed [7], [8], as well as several robust estimation approaches [9]–[17]. In order to comparatively evaluate different methods, a series of experiments have been conducted using synthetic data from distribution with covariance matrix $\mathbf{\Sigma}$ from real HS images. The rest of this manuscript is organized as follows: We study different techniques for covariance matrix estimation in Section 2. Hereafter, in Section 3, we show simulations and real-life HS images to indicate the performance of considered approaches. In Section 4, we discuss several important issues and concluding remarks are given.

2 Anomaly detector in HS: Description and estimation methods

This section briefly describes the RX-detector before reviewing some covariance matrix estimation methods in the literature.

- *Santiago Velasco-Forero is with the CMM-Centre for Mathematical Morphology, PSL Research University, MINES Paristech, France. E-mail: velasco@cmm.ensmp.fr, http://cmm.ensmp.fr/∼velasco. This work was partially carried out during the tenure as a Post-Doctoral Fellowship in the Department of Mathematics at the National University of Singapore.*
- *Marcus Chen is with the Nanyang Technological University, Singapore. E-mail: marcuschen@pmail.ntu.edu.sg.*
- *Alvina Goh and Sze Kim Pang are with the National University of Singapore.*

2.1 The RX-detector

AD may be considered as a binary hypothesis testing problem at every pixel as follows:

$$\begin{aligned} \mathcal{H}_0: &\quad x \sim f_{x|\mathcal{H}_0}(\mathbf{x}), \\ \mathcal{H}_1: &\quad x \sim f_{x|\mathcal{H}_1}(\mathbf{x}), \end{aligned} \tag{1}$$

where $f_{x|\mathcal{H}_i}(\cdot)$ denotes the probability density function (PDF) conditioned on the hypothesis i, i.e., $\mathcal{H}_0$ when the target is absent (*background*), and $\mathcal{H}_1$ when the target is present. Usually, For $\mathcal{H}_0$, the background distribution $f_{x|\mathcal{H}_0}(\mathbf{x})$ is assumed to be a multivariate Gaussian model (MGM) due to theoretical simplicity. The distribution in the presence of target can be assumed to have a multivariate uniform PDF [18]. The well-known *RX anomaly detector* (Reed and Xiaoli Yu [19]) was based on these two assumptions, its test statistics is as follows:

$$\begin{aligned} \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &\underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \tau_0, \\ \Rightarrow \mathtt{AD}_{\mathtt{RX}}(\mathbf{x},\tau_1) = (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau_1, \end{aligned} \tag{2}$$

where $|\boldsymbol{\Sigma}|$ is the determinant of matrix $\boldsymbol{\Sigma}$, and τ_0 and τ_1 are thresholds, above which $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_1$. In other words, the RX-detector is a threshold test on the *Mahalanobis distance* [20]. Thresholding the likelihood ratio provides the hypothesis test that satisfies various optimality criteria including: maximum probability of detection for the given probability of false alarm, minimum expected cost, and minimization of maximal expected cost [21]. However, in most of the cases, $\boldsymbol{\Sigma}$ is unknown and needs to be estimated. It is well-known [22] that given n independent samples, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ from a d-variate Gaussian distribution with known mean $\boldsymbol{\mu} \in \mathbb{R}^d$, the SCM defined by

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T, \tag{3}$$

is the maximum likelihood estimator (MLE) of $\boldsymbol{\Sigma}$.

2.2 The RX-detector in High Dimensional Space

To help better understand the implication of high dimensionality in the RX-detector, we develop an alternative expression for (2) based on the *Singular Value Decomposition* (SVD) of the covariance matrix $\boldsymbol{\Sigma}$, as follows:

$$\mathtt{AD}_{\mathtt{RX}}(\mathbf{x},\tau_1) = (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{U}^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{U}(\mathbf{x}-\boldsymbol{\mu}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau_1,$$

where $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$ with $\boldsymbol{\Lambda}$ a diagonal matrix and $\mathbf{U}$ an orthogonal matrix. The eigenvalues $\{\lambda_i\}_{i=1}^d$ in $\boldsymbol{\Lambda}$ correspond to the variances along the individual eigenvectors and sum up to the total variance of the original data. Let the diagonal matrix $\boldsymbol{\Omega} = \{\omega_{ii}\}_{i=1}^d = \{1/\sqrt{\lambda_i}\}_{i=1}^d$, then $\boldsymbol{\Omega}^2 = \boldsymbol{\Lambda}^{-1}$. Additionally, since $\mathbf{U}$ is a rotation matrix, i.e., $\mathbf{U}^{-1} = \mathbf{U}^T$, we can rewrite the RX-detector as follows:

$$\begin{aligned} \mathtt{AD}_{\mathtt{RX}} =& (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{U}\boldsymbol{\Omega}\boldsymbol{\Omega}\mathbf{U}^T(\mathbf{x}-\boldsymbol{\mu}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau_1 \\ =& ||\boldsymbol{\Omega}\mathbf{U}^T(\mathbf{x}-\boldsymbol{\mu})||_2^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau_1. \end{aligned} \tag{4}$$

As we can see from this decomposition, the RX-detector in (2) is equivalent to the weighted Euclidean norm by the eigenvalues along the principal components. Note that as $\lambda_i \to 0$, the detector $\mathtt{AD}_{\mathtt{RX}}(\mathbf{x},\tau_1) \to \infty$, $\forall \mathbf{x}$, resulting in an unreasonable bias towards preferring $\mathcal{H}_1$ to $\mathcal{H}_0$. This fact is well-known in the literature as bad conditioning, i.e., the condition number [1] of $cond(\boldsymbol{\Sigma}) \to \infty$. Before looking at the possible solutions to the ill-conditioning issue, we would like to have a more detailed analysis of the eigenvalue distribution of covariance matrices in the theory of random matrices [23]–[25]. Denoting the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ by $\lambda_1, \lambda_2, \ldots, \lambda_n$ with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. The *Marchenko-Pastur (M-P) law* states that the distribution of the eigenvalues of empirical covariance matrix, i.e., the empirical spectral density,

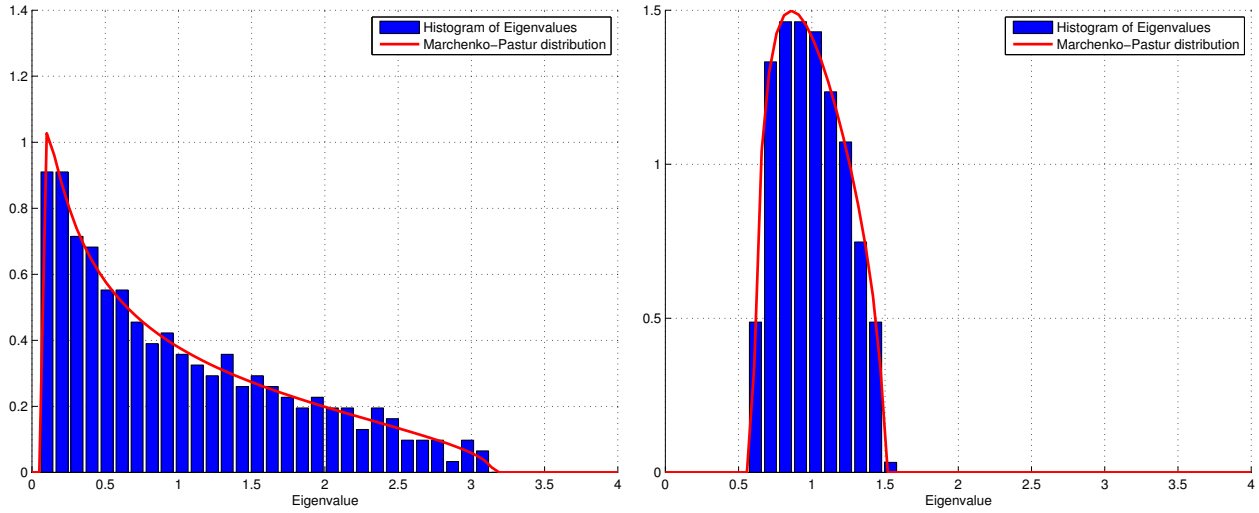
$$f(\lambda) = \frac{1}{n}\delta_{\lambda_i}(\lambda), \tag{5}$$

converges to the deterministic M-P distribution, when $d,\ n \to \infty$ and $\frac{d}{n} \to c$ [23][25]. In the case of $x \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, the M-P law describes the asymptotic behavior of $f(\lambda)$

$$f(\lambda) = \frac{\sqrt{(\lambda-a)(b-\lambda)}}{2\pi c\lambda}, \lambda \geq 0, \tag{6}$$

where $a = (1-\sqrt{c})^2,\ b = (1+\sqrt{c})^2$. A simple analysis of previous equation illustrates that, when n does not greatly exceed d, the SCM will have eigenvalues in the vicinity of zero. This is illustrated in Fig. 1 with two different $\frac{d}{n}$ values. Additionally, one can

1. The condition number of a real matrix $\boldsymbol{\Sigma}$ is the ratio of the largest singular value to the smallest singular value. A well-conditioned matrix means its inverse can be computed with good accuracy.



(a) $n = 300, c = 1$ (b) $n = 6000, c = 0.01$

Fig. 1: Empirical distribution of eigenvalues of SCM $\widehat{\mathbf{\Sigma}}$ and the corresponding Marchenko-Pastur Distribution for two different values of sample size n with $d = 300$.

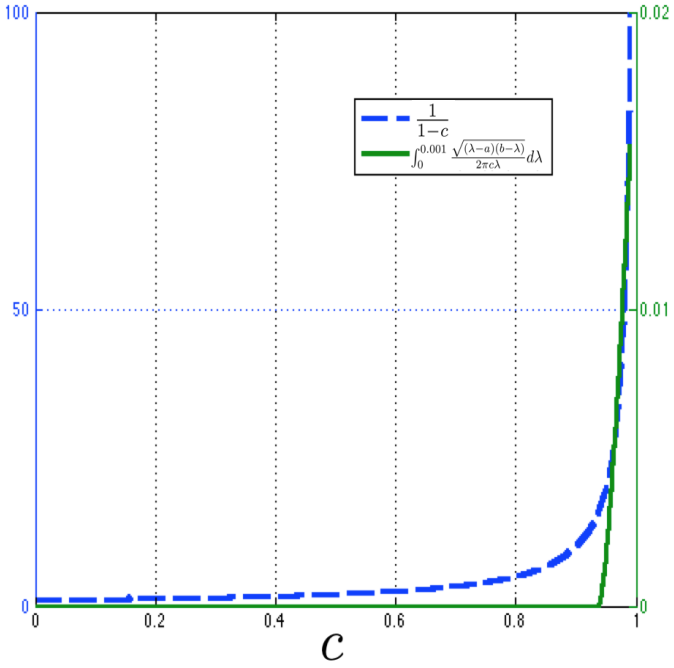


Fig. 2: Approximation of estimation accuracy of $\mathbf{\Sigma}$ by $\widehat{\mathbf{\Sigma}}$ from [27] (in blue) and probability of eigenvalues less than $k = 0.001$ both as function of $c = d/n$ (in green).

compute the integral of (6) between 0 and a small k as function of c to understand the effect of the ratio n/d in (4). This is exactly what Fig. 2 shows for $k = 0.001$. It gives the intuition as soon c is close to one, the probability of having eigenvalues close to zeros increases dramatically and then a malfunction of (4). Similarly, the analysis of the estimation accuracy of $\mathbf{\Sigma}$ elaborated in [26] and [27], with the same distribution assumption, provides more clues about the relationship between c and the performance of the RX-detector. [27] concludes that the precision in the $\mathbf{\Sigma}$ estimation by $\widehat{\mathbf{\Sigma}}$ can be approximated by $\frac{1}{1-c}$ for large d. This simple expression shows that if $c = \frac{d}{n} = 0.1$, there are more 11% overestimation on average (depending on d). Thus, a value less than $c = 0.01$ is needed to achieve 1% estimation error on average. We have also include the results in Fig. 2 normalizing the scale to show are coherent and help us to understand the malfunctioning of the RX-detector when c is going to one.

2.3 Robust Estimation in Non-Gaussian Assumptions

Presence of outliers can distort both mean and covariance estimates in computing Mahalanobis distance. In the following, we describe two types of robust estimators for covariance matrix.

2.3.1 M-estimators

In a Gaussian distribution, the SCM $\widehat{\mathbf{\Sigma}}$ in (3) is the MLE of $\mathbf{\Sigma}$. This can be extended to a larger family of distributions. *Elliptical distributions* is a broad family of probability distributions that generalize the multivariate Gaussian distribution and inherit some of its properties [22], [28]. The d-dimension random vector x has a multivariate elliptical distribution, written as $x \sim E_d(\boldsymbol{\mu}, \mathbf{\Sigma}, \psi)$, if its characteristic function can be expressed as, $\psi_x = \exp(i\mathbf{t}^T\boldsymbol{\mu})\psi\left(\frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\right)$ for some vector $\boldsymbol{\mu}$, positive-definite matrix $\mathbf{\Sigma}$, and for some function ψ, which is called the characteristic generator. From $x \sim E_d(\boldsymbol{\mu}, \mathbf{\Sigma}, \psi)$, it does not generally follow that x has a density $f_x(\mathbf{x})$, but, if it exists, it has the following form:

$$f_x(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}, g_d) = \frac{c_d}{\sqrt{|\mathbf{\Sigma}|}} g_d\left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{7}$$

where c_d is the normalization constant and g_d is some non-negative function with $(\frac{d}{2}-1)$-moment finite. In many applications, including AD, one needs to find a robust estimator for data sets sampled from distributions with heavy tails or outliers. A commonly used robust estimator of covariance is the Maronna's M estimator [29], which is defined as the solution of the equation

$$\widehat{\mathbf{\Sigma}}_M = \frac{1}{n}\sum_{i=1}^{n} u((\mathbf{x}_i-\boldsymbol{\mu})^T\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}))((\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T, \tag{8}$$

where the function $u : (0,\infty) \to [0,\infty)$ determines a whole family of different estimators. In particular, a special case $u(x) = \frac{d}{x}$ is shown to be the most robust estimator of the covariance matrix of an elliptical distribution with form (7), in the sense of minimizing the maximum asymptotic variance. This is the called *Tyler's method* [10] which is given by

$$\widehat{\mathbf{\Sigma}}_{\text{Tyler}} = \frac{d}{n}\sum_{i=1}^{n}\frac{(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T}{(\mathbf{x}_i-\boldsymbol{\mu})^T\widehat{\mathbf{\Sigma}}_{\text{Tyler}}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}. \tag{9}$$

[10] established the conditions for the existence of a solution of the fixed point equation (9). Additionally, [10] shows that the estimator is unique up to a positive scaling factor, i.e., that $\mathbf{\Sigma}$ solves (9) if and only if $c\mathbf{\Sigma}$ solves (9) for some positive scalar $c > 0$. Another interpretation to (9) can be found by considering normalized samples defined as $\{\mathbf{s}_i = \frac{\mathbf{x}_i-\boldsymbol{\mu}}{||\mathbf{x}_i-\boldsymbol{\mu}||}\}_{i=1}^n$. Then, the PDF of $\mathbf{s}$ takes the form [28]:

$$f_{\mathbf{S}}(\mathbf{s}) = \frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}}\det(\mathbf{\Sigma})^{-\frac{1}{2}}(\mathbf{s}^T\mathbf{\Sigma}^{-1}\mathbf{s})^{\frac{-d}{2}},$$

and the MLE of $\mathbf{\Sigma}$ can by obtained by minimizing the negative log-likelihood function:

$$\mathcal{L}(\mathbf{\Sigma}) = \frac{d}{2}\sum_{i=1}^{n}\log(\mathbf{s}_i^T\mathbf{\Sigma}^{-1}\mathbf{s}_i) + \frac{n}{2}\log\det(\mathbf{\Sigma}). \tag{10}$$

If the optimal estimator $\widehat{\mathbf{\Sigma}} > 0$ of (10) exist, it needs to satisfy the equation (9) [28]. When $n > d$, Tyler proposed the following iterative algorithm based on $\{\mathbf{s}_i\}$:

$$\widetilde{\mathbf{\Sigma}}_{k+1} = \frac{d}{n}\sum_{i=1}^{n}\frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\widehat{\mathbf{\Sigma}}_k^{-1}\mathbf{s}_i}, \widehat{\mathbf{\Sigma}}_{k+1} = \frac{\widetilde{\mathbf{\Sigma}}_k}{\text{tr}(\widetilde{\mathbf{\Sigma}}_k)}. \tag{11}$$

It can be shown [10] that the iteration process in (11) converges and does not depend on the initial setting of $\widehat{\mathbf{\Sigma}}_0$. Accordingly, the initial $\widehat{\mathbf{\Sigma}}_0$ is usually set to be the identity matrix of size d. We have denoted the iteration limit $\widehat{\mathbf{\Sigma}}_\infty = \widehat{\mathbf{\Sigma}}_{\text{Tyler}}$. Note that the normalization by the trace in the right side of (11) is not mandatory but it is often used in Tyler based estimation to make easier the comparison and analysis of its spectral properties without any decrement in the detection performance. Recently, a similar M-P law to (6) for the empirical eigenvalues of (11) has been shown in [30], [31].

2.3.2 Multivariate t-distribution Model

Firstly, we evoke a practical advice to perform AD in real-life HS images from [2]. They have indicated that the quality of the AD can be improved by means of considering the correlation matrix $\mathbf{R}$ instead of the covariance matrix $\mathbf{\Sigma}$, also known as the *R-RX-detector* [32]. However, notice that writing the j-th coordinate of the vector $\mathbf{z}$ as $z_{(j)} = \frac{\mathbf{x}_{(j)}-\mu_{(j)}}{\sqrt{\sigma_{(jj)}}}$, we have $\mathbf{z} = (z_1,\ldots,z_d) = \boldsymbol{\sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$, where $\boldsymbol{\sigma} = \text{diag}(\sqrt{\sigma_1,\ldots,\sigma_d})$. Now, $\mathbf{Z} = [\mathbf{z}_1,\ldots,\mathbf{z}_n]$ is zero-mean, and $cov(\mathbf{Z}) = \boldsymbol{\sigma}^{-1/2}\mathbf{\Sigma}\boldsymbol{\sigma}^{-1/2} = \mathbf{R}$, the correlation matrix of $\mathbf{X}$. Thus, the correlation matrix of x is the covariance matrix of $\mathbf{Z}$, i.e., the standardization ensuring that all the variable in $\mathbf{Z}$ are on the same scale. Additionally, note that [32] gives a characterization of the performance of the R-RX-detection. They conclude that the performance of R-RX depends not only on the dimensionality d and the deviation from the anomaly to the background mean but also on the squared magnitude of the background mean. That is an unfavorable point in the case that $\boldsymbol{\mu}$ needs to be estimated. At this point, we are interested in characterizing the MLE solution of the correlation matrix $\mathbf{R}$ by means of *t-distribution*. A d-dimensional random

vector $\mathbf{x}$ is said to have the d-variate $t-$distribution with degrees of freedom v, mean vector $\boldsymbol{\mu}$, and correlation matrix $\mathbf{R}$ (and with $\boldsymbol{\Sigma}$ denoting the corresponding covariance matrix) if its joint PDF is given by:

$$f_x(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},v) = \frac{\Gamma(\frac{v+d}{2})|\mathbf{R}|^{-1/2}}{(\pi v)^{\frac{d}{2}}\Gamma(\frac{v}{2})\left[1+\frac{1}{v}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{\frac{v+d}{2}}}, \tag{12}$$

where the degree of freedom parameter v is also referred to as the shape parameter, because the peakedness of (12) may be diminished or increased by varying v. Note that if $d=1, \boldsymbol{\mu}=0$, and $\mathbf{R}=1$, then (12) is the PDF of the *univariate Student's t distribution* with degrees of freedom v. The limiting form of (12) as $v\to\infty$ is the joint PDF on the d-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Hence, (12) can be viewed as a generalization of the multivariate normal distribution. The particular case of (12) for $\boldsymbol{\mu}=0$ and $\mathbf{R}=\mathbf{I}_d$ is a normal density with zero means and covariance matrix $v\mathbf{I}_d$ in the scale parameter v. However, the MLE does not have closed form and it should be found through *expectation-maximization algorithm* (EM) [33][34]. The EM algorithm takes the form of iterative updates, using the current estimates of $\boldsymbol{\mu}$ and $\mathbf{R}$ to generate the weights. The iterations take the form:

$$\widehat{\boldsymbol{\mu}}_{k+1} = \frac{\sum_{i=1}^n w_k^i \mathbf{x}_i}{\sum_{i=1}^n w_k^i}, \text{ and} \tag{13}$$

$$\widehat{\mathbf{R}}_{k+1} = \frac{1}{n}\sum_{i=1}^n (w_k^i(\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_{k+1})(\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_{(k+1)})^T), \tag{14}$$

where $w_{k+1}^i = \frac{v+d}{v+(\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_k)^T\mathbf{R}_k^{-1}(\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_k)}$. For more details of this algorithm, interested readers may refer to [34], and [35] for faster implementations. In our case, of known zero mean, this approach becomes:

$$\widehat{\mathbf{R}}_{k+1} = \frac{v+d}{n}\sum_{i=1}^n \frac{\mathbf{x}_i\mathbf{x}_i^T}{v+\mathbf{x}_i^T\widehat{\mathbf{R}}_k^{-1}\mathbf{x}_i} \tag{15}$$

For the case of unknown v, [36] showed how to use EM to find the joint MLEs of all parameters $(\boldsymbol{\mu},\mathbf{R},v)$. However, our preliminary work [37] shows that the estimation of v does not give any improvement in AD task. Therefore, we limited ourselves to the case of t-distribution with known value of degrees of freedom v.

2.4 Estimators in High Dimensional Space

The SCM $\widehat{\boldsymbol{\Sigma}}$ in (3), offers the advantages of easy computation and being an unbiased estimator, i.e., its expected value is equal to the covariance matrix. However, as illustrated in Section 2.2, in high dimensions the eigenvalues of the SCM are poor estimates for the true eigenvalues. The sample eigenvalues spread over the positive real numbers. That is, the smallest eigenvalues will tend to zero, while the largest tend toward infinity [38], [39]. Accordingly, SCM is unsatisfactory for large covariance matrix estimation problems.

2.4.1 Shrinkage Estimator

To overcome this drawback, it is common to regularize the estimator $\widehat{\boldsymbol{\Sigma}}$ with a highly structured estimator $\mathbf{T}$ via a linear combination $\alpha\widehat{\boldsymbol{\Sigma}}+(1-\alpha)\mathbf{T}$, where $\alpha\in[0,1]$. This technique is called regularization or *shrinkage*, since $\widehat{\boldsymbol{\Sigma}}$ is "shrunk" towards the structured estimator. The shrinkage helps to condition the estimator and avoid the problems of ill-conditioning in (4). The notion of shrinkage is based on the intuition that a linear combination of an *over-fit* sample covariance with some *under-fit* approximation will lead to an intermediate approximation that is "just-right" [13]. A desired property of shrinkage is to maintain eigenvectors of the original estimator while conditioning on the eigenvalues. This is called *rotationally-invariant estimators* [40]. Typically, $\mathbf{T}$ is set to $\rho\mathbf{I}$, where $\mathbf{I}$ is the identity matrix for some $\rho>0$ and ρ is set by $\rho=\sum_{i=1}^d \sigma_{ii}/d$. In this case, the same shrinkage intensity is applied to all sample eigenvalue, regardless of their position. To illustrate the eigenvalues behavior after shrinkage, let us consider the case of linear shrinkage intensity equal to 1/4, 1/2 and 3/4. Fig. 3 illustrates this case. As it was shown in [41], in the case of $\alpha=1/2$, every sample eigenvalue is moved half-way towards the grand mean of all sample eigenvalues. Similarly, for $\alpha=1/4$ eigenvalues are moved a quarter towards the mean of all sample eigenvalues. An alternative is the non-rotationally invariant shrinkage method, proposed by Hoffbeck and Landgrebe [9], uses the diagonal matrix $\mathbf{D}=\text{diag}(\widehat{\boldsymbol{\Sigma}})$ which agrees with the SCM the diagonal entries, but shrinks the off-diagonal entries toward zero:

$$\widehat{\boldsymbol{\Sigma}}_{\text{diag}}^{\alpha} = (1-\alpha)\widehat{\boldsymbol{\Sigma}}+\alpha\text{diag}(\widehat{\boldsymbol{\Sigma}}) \tag{16}$$

However, in the experiments, we use a normalized version of (16), considering the dimension of the data, i.e.,

$$\widehat{\boldsymbol{\Sigma}}_{\text{Stein}}^{\alpha} = (1-\alpha)\widehat{\boldsymbol{\Sigma}}+\alpha\text{Id}(\widehat{\boldsymbol{\Sigma}}) \tag{17}$$

where $\text{Id}(\boldsymbol{\Sigma})=\frac{\text{tr}(\widehat{\boldsymbol{\Sigma}})\mathbf{I}}{d}$. This is sometimes called *ridge* regularization.

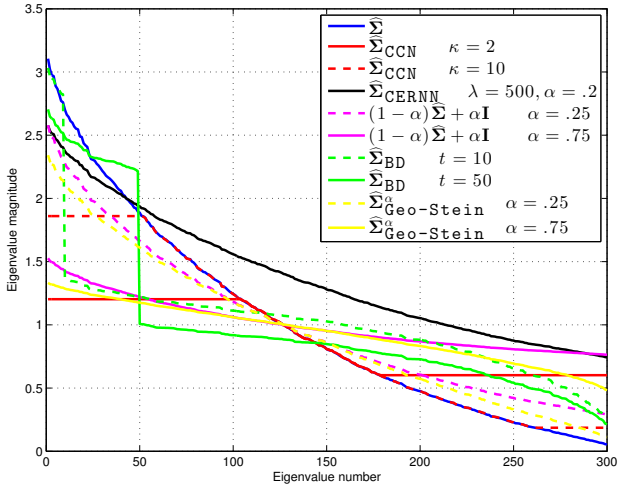


Fig. 3: CCN truncates extreme sample eigenvalues and leaves the moderate ones unchanged and CERNN gives the contrary effect. Linear and geodesic shrinkages moves eigenvalues towards the grand mean of all sample eigenvalues. However, the effect of geodesic shrinkage is more attenuated for extreme eigenvectors than in linear case. The effect of BD-correction depends on the eigenvalues sets defined by t.

2.4.2 Regularized Tyler-estimator

Similarly, shrinkage can be applied to other estimators such as the robust estimator in (11). The idea was proposed in [7], [42], [43]. Wiesel [43] gives the fixed point condition to compute a robust and well-conditioned estimator of $\mathbf{\Sigma}$ by

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_{k+1} &= \frac{d}{n(1+\alpha)}\sum_{i=1}^{n}\frac{(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T}{(\mathbf{x}_i-\boldsymbol{\mu})^T\tilde{\mathbf{\Sigma}}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu})} + \frac{\alpha}{1+\alpha}\frac{d\mathbf{T}}{\mathrm{tr}(\tilde{\mathbf{\Sigma}}_k^{-1}\mathbf{T})}\\ \widehat{\mathbf{\Sigma}}_{k+1} &:= \frac{\tilde{\mathbf{\Sigma}}_{k+1}}{\mathrm{tr}(\tilde{\mathbf{\Sigma}}_{k+1})}.\end{aligned} \tag{18}$$

This estimator is a trade-off between the intrinsic robustness from M-estimators in (11) and the well-conditioning of shrinkage based estimators in section 2.4.1. The existence and uniqueness of this approach has been shown in [17]. Nevertheless, the optimal value of shrinkage parameter α in (18) is still an open question.

2.4.3 Geodesic Interpolation in Riemannian Manifold

The shrinkage methods discussed so far involve the linear interpolation between two matrices, namely, a covariance matrix estimator and a target matrix. It can be extended to other types of interpolations, i.e., other space of representation for $\widehat{\mathbf{\Sigma}}$ and $\mathbf{T}$ different to the Euclidean space. A well-known approach is the Riemannian manifold of covariance matrices, i.e. the space of symmetric matrices with positive eigenvalues [44]. In general, Riemannian manifold are analytical manifolds endowed with a distance measure which allows the measurement of similarity or dissimilarity (closeness or distance) of points. In this representation, the distance, called *geodesic distance*, is the minimum length of the curvature path that connects two points [45], and it can be computed by

$$\mathrm{Dist}_{\text{Geo}}(\mathbf{A},\mathbf{B}) := \sqrt{\mathrm{tr}(\log^2(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}))}. \tag{19}$$

This nonlinear interpolation, here called a *geodesic path* from $\mathbf{A}$ to $\mathbf{B}$ at time t, is defined by $\mathrm{Geo}_t(\mathbf{A},\mathbf{B}) := \mathbf{A}^{1/2}\exp(t\mathbf{M})\mathbf{A}^{1/2}$, where $\mathbf{M} = \log(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2})$ and exp and log are matrix exponential and logarithmic functions respectively. A complete analysis of (19) and the geodesic path via its representation as ellipsoids have been presented in [46]. Additionally, [46] shows that the volume of the geodesic interpolation is smaller than linear interpolation and thus it can increase detection performance in HSI detection problems. Thus, we have included a *Geodesic Stein estimation* with the same intuition behind equation (17) as follows,

$$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}} = \mathrm{Geo}_{\alpha}(\widehat{\mathbf{\Sigma}}, \mathrm{Id}(\widehat{\mathbf{\Sigma}})), \tag{20}$$

where $\alpha \in [0,1]$ determines the trade-off between the original estimation $\widehat{\mathbf{\Sigma}}$ and the well-conditioning $\mathrm{Id}(\widehat{\mathbf{\Sigma}})$.

2.4.4 Constrained MLE

As we have shown in Section 2.2, even when $n > d$, the eigenstructure tends to be systematically distorted unless d/n is extremely small, resulting in ill-conditioned estimators for $\mathbf{\Sigma}$. Recently, several works have proposed regularizing the SCM by explicitly imposing a constraint on the condition number. [16] proposes to solve the following constrained MLE problem:

$$\text{maximize } \mathcal{L}(\mathbf{\Sigma}) \text{ subject to } cond(\mathbf{\Sigma}) \leq \kappa \tag{21}$$

where $\mathcal{L}(\mathbf{\Sigma})$ stands for the log-likelihood function in the Gaussian distributions. This problem is hard to solve in general. However, [16] proves that in the case of rotationally-invariant estimators, (21) reduces to an unconstrained univariate optimization problem. Furthermore, the solution of (21) is a nonlinear function of the sample eigenvalues given by:

$$\widehat{\lambda}_i = \begin{cases} \eta, & \lambda_i(\widehat{\mathbf{\Sigma}}) \leq \eta \\ \lambda_i(\widehat{\mathbf{\Sigma}}), & \eta < \lambda_i(\widehat{\mathbf{\Sigma}}) < \eta\kappa \\ \kappa\eta, & \lambda_i(\widehat{\mathbf{\Sigma}}) \geq \eta\kappa \end{cases} \tag{22}$$

for some η depending on κ and $\lambda(\widehat{\mathbf{\Sigma}})$. We refer this methodology as *Condition Number-Constrained* (CCN) estimation.

2.4.5 Covariance Estimate Regularized by Nuclear Norms

Instead of constrain the MLE problem in (21), [47] propose to penalize the MLE as follows,

$$\text{maximize } \mathcal{L}(\mathbf{\Sigma}) + \frac{\lambda}{2}[\alpha||\mathbf{\Sigma}||_* + (1-\alpha)||\mathbf{\Sigma}^{-1}||_*] \tag{23}$$

where the nuclear norm of a matrix $\mathbf{\Sigma}$, is denoted by $||\mathbf{\Sigma}||_*$, is the sum of the eigenvalues of $\mathbf{\Sigma}$, λ is a positive strength constant, and $\alpha \in (0,1)$ is a mixture constant. We refer this approach by the acronym CERNN (Covariance Estimate Regularized by Nuclear Norms).

2.4.6 Ben-David and Davidson correction

Given zero-mean[2] data with normal probability density $\mathbf{x} \sim N(0, \mathbf{\Sigma})$, its sampled covariance matrix $\widehat{\mathbf{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$ follows a central Wishart distribution with n degrees of freedom. The study of covariance estimators in Wishart distribution where the sample size (n) is small in comparison to the dimension (d) is also an active research topic [48]–[50]. Firstly, Efron and Morris proposed a rotationally-invariant estimator of $\mathbf{\Sigma}$ by replacing the sampled eigenvalues with an improved estimation [51]. Their approach is supported by the observation that for any Wishart matrix, the sampled eigenvalues tend to be more spread out than population eigenvalues, in consequence, smaller sampled eigenvalues are underestimated and large sampled eigenvalues are overestimated [49]. Accordingly, they find the best estimator of inverse of the covariance matrix of the form $a\widehat{\mathbf{\Sigma}}^{-1} + b\mathbf{I}/\text{tr}(\widehat{\mathbf{\Sigma}})$ which is achieved by:

$$\widehat{\mathbf{\Sigma}}_{\text{Efron-Morris}} = \left((n-d-1)\widehat{\mathbf{\Sigma}}^{-1} + \frac{d(d+1)-2}{\text{tr}(\widehat{\mathbf{\Sigma}})}\mathbf{I}\right)^{-1}. \tag{24}$$

It is worth mentioning that other estimations have been developed following the idea behind Wishart modeling and assuming a simple model for the eigenvalue structure in the covariance matrix (usually two phases model). Recently, Ben-David and Davidson [49] have introduced a new approach for covariance estimation in HSI, called here *BD-correction*. From the SVD of $\widehat{\mathbf{\Sigma}} = \mathbf{U}\Lambda_{\widehat{\mathbf{\Sigma}}}\mathbf{U}^T$, they proposed a rotationally-invariant estimator by correcting the eigenvalues by means of two diagonal matrices,

$$\widehat{\mathbf{\Sigma}}_{\text{BD}} = \mathbf{U}\Lambda_{\text{BD}}\mathbf{U}^T, \quad \text{with} \quad \Lambda_{\text{BD}} = \Lambda_{\widehat{\mathbf{\Sigma}}}\Lambda_{\text{Mode}}\Lambda_{\text{Energy}}. \tag{25}$$

They firstly estimate the apparent multiplicity p_i of the i-th sample eigenvalue as $p_i = \sum_{j=1}^{d} card[a(j) \leq b(i) \leq b(j)]$, where $a(i) = \Lambda_{\widehat{\mathbf{\Sigma}}}(i)(1-\sqrt{c})^2$ and $b(i) = \Lambda_{\widehat{\mathbf{\Sigma}}}(i)(1+\sqrt{c})^2$. One can interpret the concept of "apparent multiplicity" as the number of distinct eigenvalues that are "close" together and thus represent nearly the same eigenvalue [49]. Secondly, BD-correction affects the i-th sample eigenvalue via its apparent multiplicity p_i as $\Lambda_{Mode}(i) = \frac{(1+p_i/n)}{(1-p_i/n)^2}$ and as

$$\Lambda_{Energy}(i) = \begin{cases} \sum_{i=1}^{t}\Lambda_{\widehat{\mathbf{\Sigma}}}(i) / \sum_{i=1}^{t}(\Lambda_{\widehat{\mathbf{\Sigma}}}(i)\Lambda_{\text{Mode}}(i)) \\ \sum_{i=t+1}^{d}\Lambda_{\widehat{\mathbf{\Sigma}}}(i) / \sum_{i=t+1}^{d}(\Lambda_{\widehat{\mathbf{\Sigma}}}(i)\Lambda_{\text{Mode}}(i)) \end{cases} \tag{26}$$

for a value $t \in [1, \min(n,d)]$ indicating the transition between large and small eigenvalues. Finally, reader can see [49] for an optimal selection of t. A comparison of correction in the eigenvalues by CCN, CERNN, the linear shrinkage in (17), the geodesic Stein in (20) and the BD-correction is illustrated in Fig. 3 for three values of regulation parameter. We can see that CCN truncates extreme sample eigenvalues and leaves the moderate ones unchanged. Compared to the linear estimator, both (21) and (23) pull the larger eigenvalues down more aggressively and pull the smaller eigenvalues up less aggressively.

2.4.7 Sparse Matrix Transform

Recently, [13], [52] introduced the *sparse matrix transform* (SMT). The idea behind is the estimation of the SVD from a series of *Givens rotations*, i.e., $\widehat{\mathbf{\Sigma}}_{\text{SMT}} = \mathbf{V}_k\mathbf{\Lambda}\mathbf{V}_k^T$, where $\mathbf{V}_k = \mathbf{G}_1\mathbf{G}_2\cdots\mathbf{G}_k$ is a product of k *Givens rotation* defined by $\mathbf{G} = \mathbf{I} + \mathbf{\Theta}(i,j,\theta)$ where

$$\mathbf{\Theta}(a,b,\theta) = \begin{cases} \cos(\theta) - 1, & \text{if } r = s = a \text{ or } r = s = b \\ \sin(\theta), & \text{if } r = a \text{ and } s = b \\ -\sin(\theta), & \text{if } r = b \text{ and } s = a \\ 0, & \text{otherwise} \end{cases}$$

2. Or $\boldsymbol{\mu}$ known, in which case, one might subtract $\boldsymbol{\mu}$ from the data.

where each step $i \in \{1, \ldots, k\}$ of the SMT is designed to find the single Givens rotation that minimize $\text{diag}(\mathbf{V}_i^T \widehat{\mathbf{\Sigma}} \mathbf{V}_i)$ the most. The details of this transformation are given in [52], [53]. The number of rotations k is a parameter and it can be estimated from heuristic Wishart estimator as in [13]. However, in the numerical experiments, this method of estimating k tended to over-estimate. As such, SMT is compared with k as function of d in our experiments. Table 1 summarizes the different covariance matrix estimators considered in the experiments.

TABLE 1. Covariance matrix estimators considered in this paper

Name	Notation	Formula
SCM	$\widehat{\mathbf{\Sigma}}$	$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$
Stein Shrinkage [38]	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}}$	$(1-\alpha)\widehat{\mathbf{\Sigma}} + \alpha \text{Id}(\widehat{\mathbf{\Sigma}})$
Tyler [10]	$\widehat{\mathbf{\Sigma}}_{\text{Tyler}}$	$\widehat{\mathbf{\Sigma}}_{j+1} = \frac{d}{n}\sum_{i=1}^{n} \frac{(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T}{(\mathbf{x}_i-\boldsymbol{\mu})^T\widehat{\mathbf{\Sigma}}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}$
Tyler Shrinkage [8]	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$	$\tilde{\mathbf{\Sigma}}_{k+1} = \frac{1}{1+\alpha}\frac{d}{n}\sum_{i=1}^{n}\frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}_i^T\widehat{\mathbf{\Sigma}}_k^{-1}\mathbf{x}_i} + \frac{\alpha}{1+\alpha}\frac{d\mathbf{T}}{\text{tr}(\widehat{\mathbf{\Sigma}}_k^{-1}\mathbf{T})}$
Sparse Matrix Transform (SMT) [13]	$\widehat{\mathbf{\Sigma}}_{\text{SMT}}$	$\mathbf{G}_1\mathbf{G}_2\cdots\mathbf{G}_k\mathbf{\Lambda}(\mathbf{G}_1\mathbf{G}_2\cdots\mathbf{G}_k)^T$
t distribution[36]	$\widehat{\mathbf{\Sigma}}_{\text{t}}$	$\widehat{\mathbf{\Sigma}}_{j+1} = \frac{1}{n}\sum_{i=1}^{n}\frac{(v+d)(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T}{v+(\mathbf{x}_i-\boldsymbol{\mu})^T\widehat{\mathbf{\Sigma}}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}$
Geodesic Stein	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$	$\text{Geo}_{\alpha}(\widehat{\mathbf{\Sigma}}, \text{Id}(\widehat{\mathbf{\Sigma}}))$
Constrained condition number[16]	$\widehat{\mathbf{\Sigma}}_{\text{CCN}}$	(22)
Covariance Estimate Regularized by Nuclear Norms[47]	$\widehat{\mathbf{\Sigma}}_{\text{CERNN}}$	(23)
Efron-Morris Correction [51]	$\widehat{\mathbf{\Sigma}}_{\text{Efron-Morris}}$	(24)
Ben-Davidson Correction [49]	$\widehat{\mathbf{\Sigma}}_{\text{BD}}$	(25)

3 EXPERIMENTAL RESULTS

In this section, we conduct a few experiments to compare the performance of the different methods of covariance matrix estimation. Experiments were carried out using simulation by considering $\mathbf{\Sigma}$ from some well-known HS images. Moreover, they were evaluated for AD. All the covariance matrix estimations are normalized (trace equal to d) to have comparable "size". Note that we are not interested in the joint estimation of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ in this manuscript, then mean vector $\boldsymbol{\mu}$ is assumed known throughout the experiments, i.e. the data matrix is centered by $\boldsymbol{\mu}$. Readers interested in joint estimation of the pair $(\boldsymbol{\mu}, \mathbf{\Sigma})$ might find [54] helpful. The experiments were designed to address the following three issues in the covariance estimation methods:

1) The effect of covariance ill-conditioning due to limited data and high dimension (c close to one).
2) The effect of contamination due to anomalous data being included in the covariance computation.
3) The effect of changes in the distribution such as deviation from Gaussian distributions.

3.1 Performance Evaluation

There are a few performance measures for anomaly detectors. First, we consider the probability of detection (PD) and the false alarm (FA) rate. This view yields a quantitative evaluation in terms of *Receiver Operating Characteristics* (ROC) curves [55]. A detector is good if it has a PD and a low FA rate, i.e., if the curve is closer to the upper left corner. One may reduce the ROC curve to a single value using the *Area under the ROC curve* (AUC). The AUC is estimated by summing trapezoidal areas formed by successive points on the ROC curve. A detector with a greater AUC is said to be "better" than a detector with a smaller AUC. The AUC value depicts the general behavior of the detector and characterizes how near it is to perfect detection (AUC equal to one) or to the worst case (AUC equal to $1/2$) [55].

Besides AUC, another measure is to find the one with better data fitting. That is the intuition behind the approach proposed by [18]. It is a proxy that measures the *volume* inside an anomaly detection surface for a large set of false alarm rates. Since in practical applications, the AD is usually calibrated to a given false alarm rate, one can construct a coverage log-volume versus log-false alarm rate to compare detector performances [56], [57]. Accordingly, for a given threshold radius η, the volume of the ellipsoid contained within $\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x} \leq \eta^2$ is given by

$$\texttt{Volume}(\mathbf{\Sigma}, \eta) = \frac{\pi^{d/2}}{\Gamma(1+d/2)}|\mathbf{\Sigma}|^{1/2}\eta^d. \tag{27}$$

Given an FA rate, a smaller $\texttt{Volume}(\mathbf{\Sigma}, \eta)$ indicates a better fit of real structure of the background and thus is preferred. In this paper, we compute the logarithm of (27) to reduce the effect of numerical issues in the computation.

3.2 Simulations on Elliptical Distribution

We start on experiments in the case of multivariate t distribution in (12) with v degrees of freedom. It can be interpreted as generalization of the Gaussian distribution (or conversely, the Gaussian as special case of the t-distribution when the degree of freedom tends to infinity). As $\mathbf{\Sigma}$, we have used the covariance matrix of one homogeneous zone in Pavia University HSI (pixels in rows 555 to 590 and columns 195 to 240) [58] in 90 bands (from 11 to 100). $\mathbf{\Sigma}$ is normalized to have trace equal to one. Its condition number is large,

2.7616×10^5. Anomalies in this example are generated by the same distribution but using an identity matrix of trace equal to one as parameter of the distribution. We perform estimations varying three components:

1) The degrees of freedom of the distribution from where the multivariate sample is generated.
2) The size of the sample to calculate covariance matrix estimators in Table 1.
3) The number of anomalies included in the sample to compute covariance matrix estimators in Table 1.

With that in mind, we have generated 4000 random vectors (half of them anomalies) and we have set the parameters in each estimator by minimizing the volume calculated in the threshold corresponding to a false alarm rate of 0.001. Different volumes by varying parameters in the estimation can be compared in (b,d,f,h,j,l) of Fig. 4, 5 and 6 in all the explored cases. In the experiments, the number of rotations in $\widehat{\mathbf{\Sigma}}_{\text{SMT}}$ is fixed to i times the dimension d, for $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$, the regularization parameter α is i, for $\widehat{\mathbf{\Sigma}}_{\text{CCN}}$ the regularization parameter is 2^{i+1}, in $\widehat{\mathbf{\Sigma}}_{\text{CERNN}}$ and $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}}$ the value $\alpha = i/20$, and for $\widehat{\mathbf{\Sigma}}_{\text{BD}}$ the value t is equal to $i + 1$. Different values of i from 1 to 20 are shown in x-axis. We highlight that the estimators yield detectors with AUC close to one. Additionally, to compare the general performance from the "best estimation" in each approach, we have plot the coverage log-volume versus log-false alarm rate in (a,c,e,g,i,k) of Fig. 4, 5 and 6. The interpretation of these figures can be done in three directions:

- *From left to right*, we provide the evolution of the performance by varying the degrees of freedom v. Note, that the limiting form of (12) as $v \to \infty$ is the joint pdf of the d-variate normal distribution with covariance matrix $\mathbf{\Sigma}$. Hence, we use a large value of degrees of freedom, $v = 1000$, to generate the Gaussian case. In $v = 1$, is the case of multivariate Cauchy distributions. Note that Cauchy distributions look similar to Gaussian distributions. However, they have much heavier tails. Thus, it is a good indicator of how sensitive the estimators are to heavy-tail departures from normality.
- *From up to down*, we illustrate the effect of the relative value $c = d/n$ in the performance of the estimation. We have used, in the first row, five times the number of sample than the dimension, i.e. $c = 0.2$, and in the second row, only 100 samples which correspond to a difficult scenario where $c = 0.9$.
- *From Fig. 4 to Fig. 6*, we show the consequence of including anomalies in the sample where the estimation is performed. Three cases are considered: Fig. 4 is a free noise case, Fig. 5 includes a low level of contamination (1%), and Fig. 6 shows a level of noisy samples equal to 10%.

At this stage, we can have some conclusion about the performance of studied estimators:

- In the more "classical" scenario, i.e., Gaussian distribution, no contamination and much more samples than dimensions ($c = 0.2$ in Fig. 4), the approaches $\widehat{\mathbf{\Sigma}}_{\text{BD}}$ and $\widehat{\mathbf{\Sigma}}_{\text{Efron-Morris}}$ based on correction of eigenvalues (section 2.4.6) performed slightly better than the other alternatives. However, as soon as the sample size was reduced, the data was contaminated or the distribution of data was "less" Gaussian, their performances seemed to be drastically affected.
- In Gaussian cases with contaminated samples and $c = 0.2$, the robust approaches, $\widehat{\mathbf{\Sigma}}_{\text{t}}$ and $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$ performed better than other approaches. However, $\widehat{\mathbf{\Sigma}}_{\text{t}}$ was unquestionably affected by the nocuous decreasing of the sample size in the case of $c = 0.9$, producing detector with huge volumes.
- In the scenario of Gaussian data and $c = 0.9$, $\widehat{\mathbf{\Sigma}}_{\text{SMT}}$ did the best job followed by the shrinkage approaches, i.e., $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}}$, $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$ and $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$. Another important point to note is that $\widehat{\mathbf{\Sigma}}_{\text{SMT}}$ was more affected by the contamination in the data than shrinkage-based methods.
- In the case of Cauchy distributions, $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$ was in general less affected by heavy-tails than other approaches. Additionally, geodesic interpolation ($\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$) clearly outperformed linear interpolations ($\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}}$) in these heavy-tails scenarios. $\widehat{\mathbf{\Sigma}}_{\text{CERNN}}$ and $\widehat{\mathbf{\Sigma}}_{\text{CCN}}$ were robust in this difficult case of heavy tails with contaminated data.

Finally, to summarize, the best three performances according to the coverage log-volume versus log-false alarm curve in each scenario are included in Table 2. Now, we move forward along the difficulty of the studied problems by including simulations on a more complex scenario.

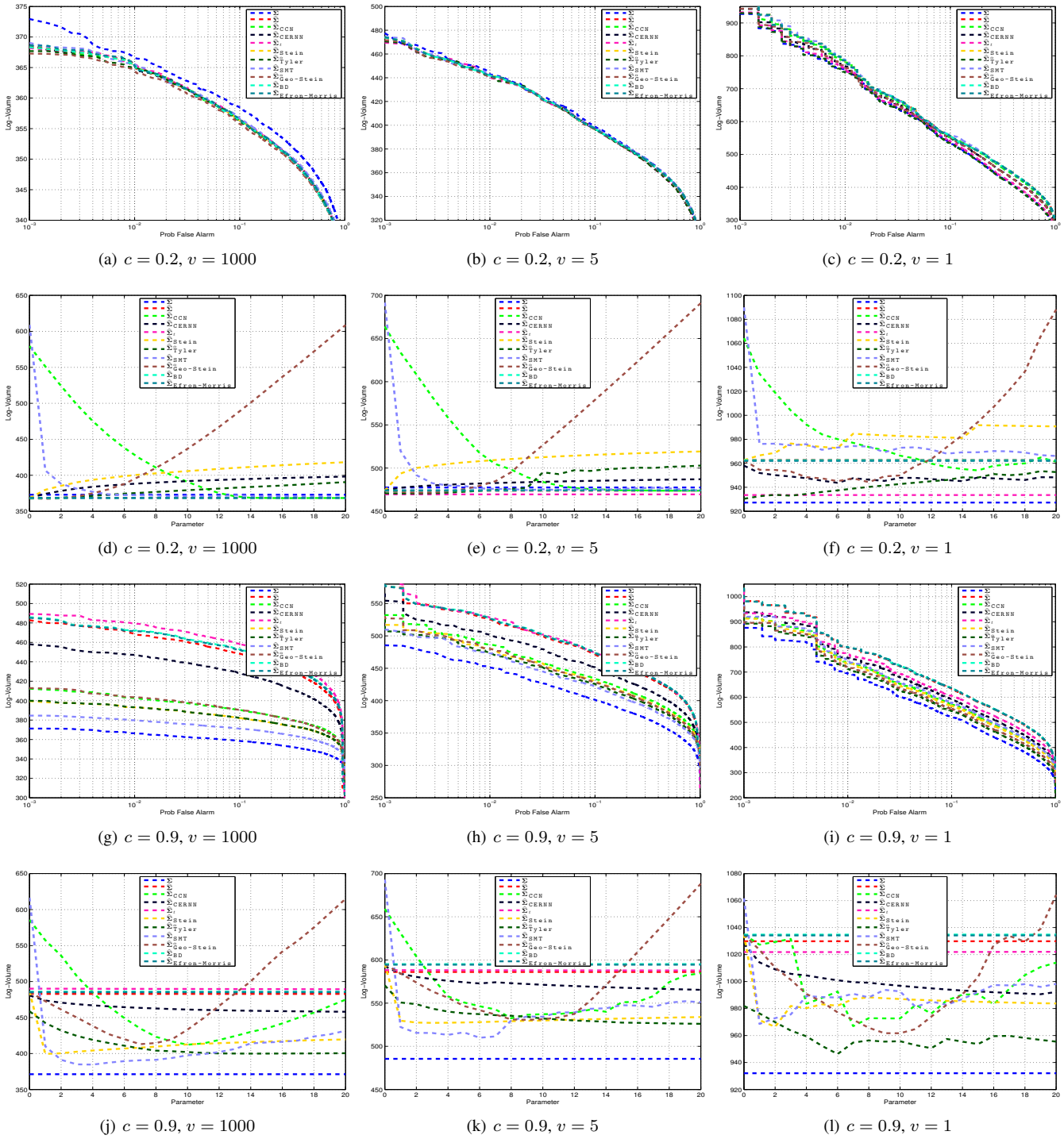
3.3 Simulations on Dirichlet Distributions

In HS images, spectral diversity can be considered in a set of proportions, called abundances [59]. In this type of data, called *compositional data* [60], the Dirichlet family of distributions is usually the first candidate employed for modeling the data. The rationale behind this choice is the following [61]:

1) The Dirichlet density automatically enforces the non-negativity and sum-to-one constraint, which is natural in the linear mixture model.
2) Mixtures of density allow one to model complex distributions in which the mass probability is scattered over a number of clusters inside the simplex [61].

A d-dimensional vector $\mathbf{p} = (p_1, p_2, \ldots, p_d)$ is said to have the *Dirichlet distribution* with parameter vector $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_d)$, $\rho_i > 0$, if it has the joint density

$$f(\mathbf{p}) = B(\rho) \prod_{i=1}^{d} p_i^{\rho_i - 1}, \tag{28}$$



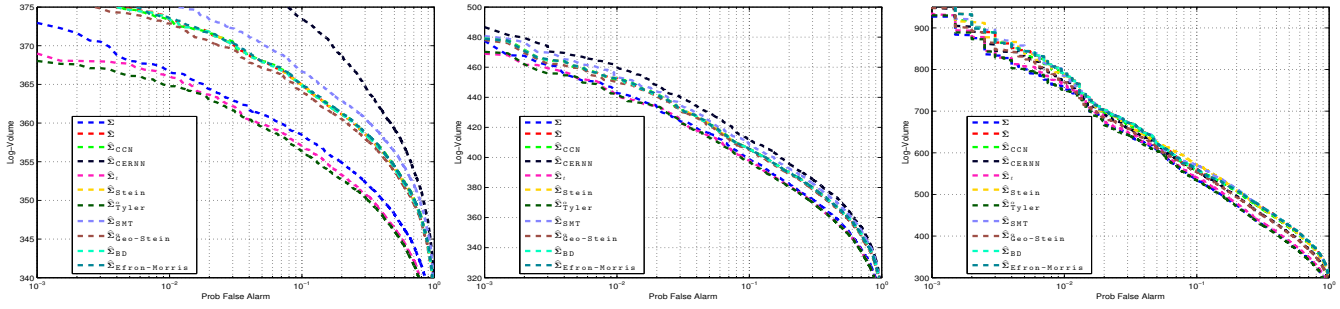
(a) $c = 0.2$, $v = 1000$ (b) $c = 0.2$, $v = 5$ (c) $c = 0.2$, $v = 1$

(d) $c = 0.2$, $v = 1000$ (e) $c = 0.2$, $v = 5$ (f) $c = 0.2$, $v = 1$

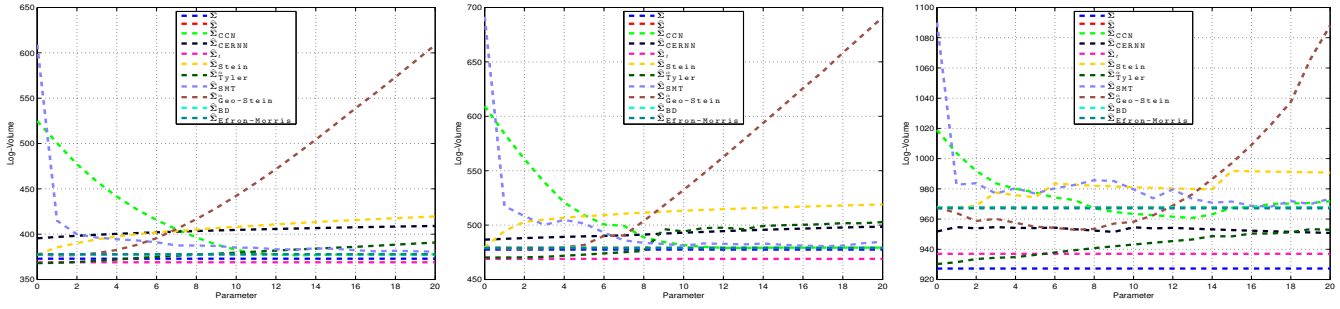
(g) $c = 0.9$, $v = 1000$ (h) $c = 0.9$, $v = 5$ (i) $c = 0.9$, $v = 1$

(j) $c = 0.9$, $v = 1000$ (k) $c = 0.9$, $v = 5$ (l) $c = 0.9$, $v = 1$

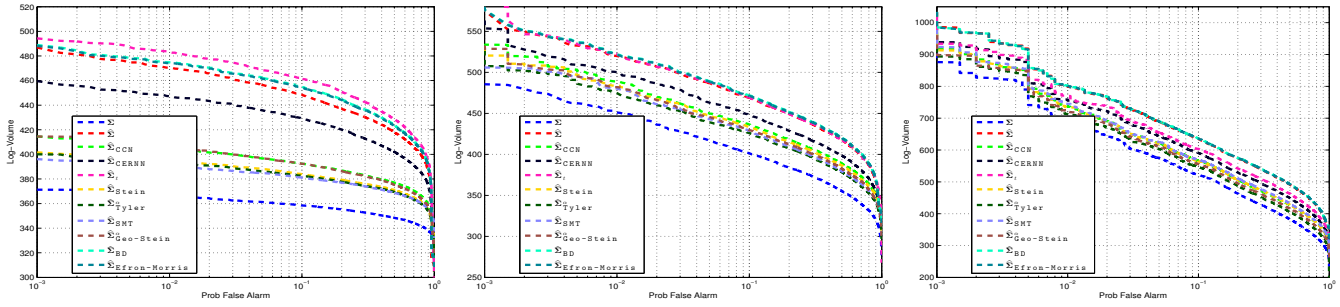
Fig. 4: **Non-contamination case:** Covariance matrices are estimated considering only background vectors in $d = 90$. From left to right, we can analyze the effect of distribution shape in the performance of the estimations, i.e., from Multivariate Gaussian (large v) to Multivariate Cauchy distribution (v=1). First row: $n = 450, c = 0.2$. Second row: $n = 100, c = 0.9$. In (d,e,f,j,k,l) volumes are calculated in the threshold corresponding to a false alarm rate of 0.001.



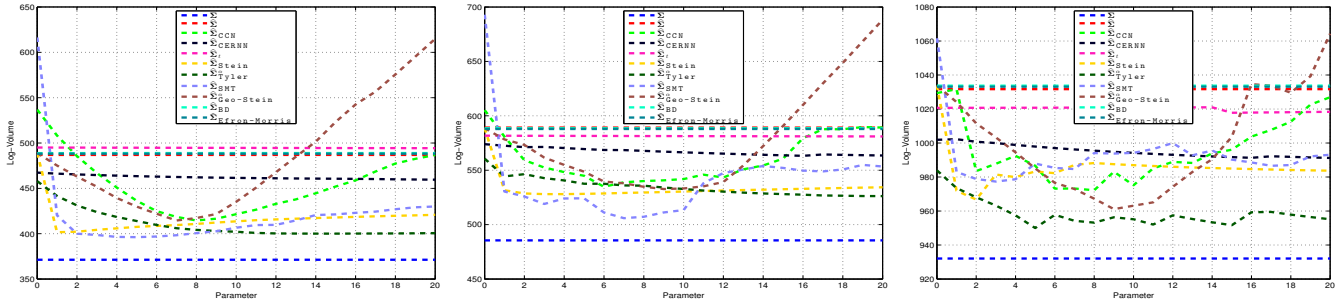
(a) $c = 0.2$, $v = 1000$ (b) $c = 0.2$, $v = 5$ (c) $c = 0.2$, $v = 1$



(d) $c = 0.2$, $v = 1000$ (e) $c = 0.2$, $v = 5$ (f) $c = 0.2$, $v = 1$

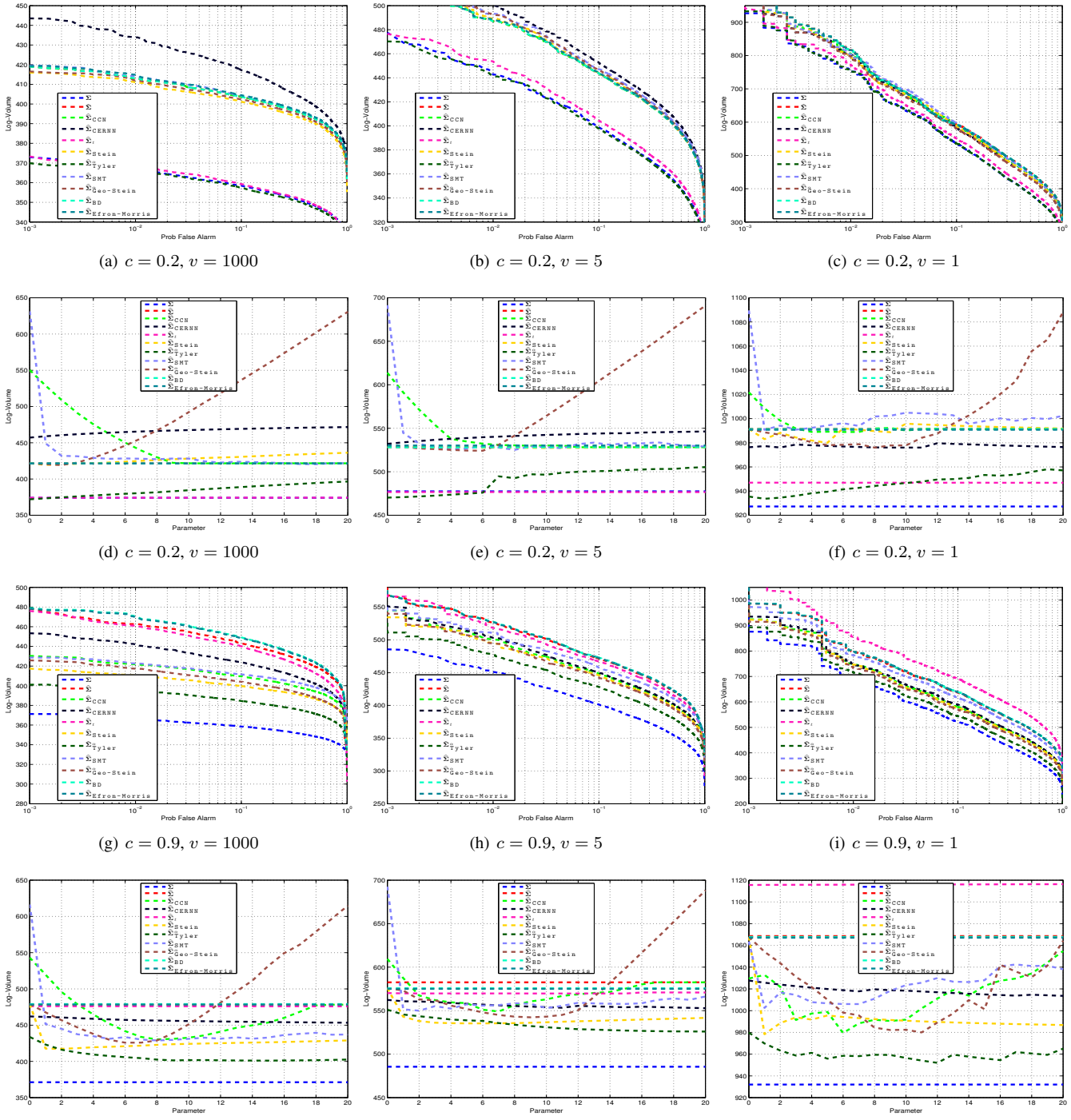


(g) $c = 0.9$, $v = 1000$ (h) $c = 0.9$, $v = 5$ (i) $c = 0.9$, $v = 1$



(j) $c = 0.9$, $v = 1000$ (k) $c = 0.9$, $v = 5$ (l) $c = 0.9$, $v = 1$

Fig. 5: **Contamination case** 1%**:** Covariance matrices are estimated considering background vectors in $d = 90$ and 1% of anomalies. From left to right, we can analyze the effect of v in the performance of the estimations. First row: $n = 450, c = 0.2$. Second row: $n = 100, c = 0.9$. In (d,e,f,j,k,l) volumes are calculated in the threshold corresponding to a false alarm rate of 0.001.



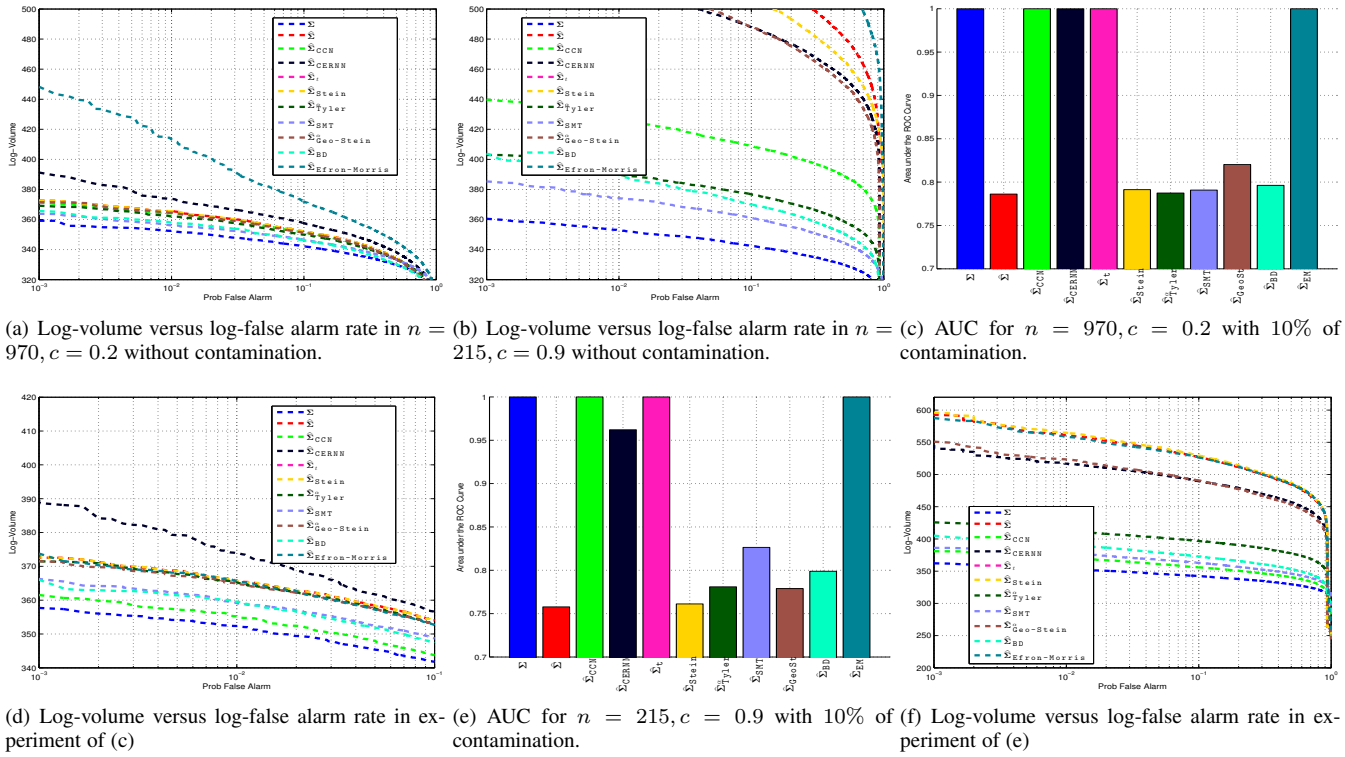
(a) $c = 0.2$, $v = 1000$ (b) $c = 0.2$, $v = 5$ (c) $c = 0.2$, $v = 1$

(d) $c = 0.2$, $v = 1000$ (e) $c = 0.2$, $v = 5$ (f) $c = 0.2$, $v = 1$

(g) $c = 0.9$, $v = 1000$ (h) $c = 0.9$, $v = 5$ (i) $c = 0.9$, $v = 1$

(j) $c = 0.9$, $v = 1000$ (k) $c = 0.9$, $v = 5$ (l) $c = 0.9$, $v = 1$

Fig. 6: **Contamination case** 10%**:** Covariance matrices are estimated considering background vectors in $d = 90$ and 10% of anomalies. From left to right, we can analyze the effect of v in the performance of the estimations. First row: $n = 450, c = 0.2$. Second row: $n = 100, c = 0.9$. In (d,e,f,j,k,l) volumes are calculated in the threshold corresponding to a false alarm rate of 0.001.



(a) Log-volume versus log-false alarm rate in $n = 970, c = 0.2$ without contamination.

(b) Log-volume versus log-false alarm rate in $n = 215, c = 0.9$ without contamination.

(c) AUC for $n = 970, c = 0.2$ with 10% of contamination.

(d) Log-volume versus log-false alarm rate in experiment of (c)

(e) AUC for $n = 215, c = 0.9$ with 10% of contamination.

(f) Log-volume versus log-false alarm rate in experiment of (e)

Fig. 7: **Dirichlet case:** Covariance matrices are estimated considering background vectors in $d = 194$. Parameters in each covariance matrix estimator are set to minimize the volume in the threshold corresponding to a false alarm rate of 0.001.

where $B(\rho) = \frac{\Gamma(\sum_{i=1}^{d}\rho_i)}{\prod_{i=1}^{d}\Gamma(\rho_i)}$, $p_i \geq 0, \sum_{i=1}^{d} p_i = 1$. We write $\mathbf{p} \sim \mathcal{D}(\rho)$. A complex experiment is carried out by selecting fifteen endmembers from a real HS image (World Trade Center) by means of *Vertex Component Analysis* (VCA)[62]. After that, we use (28) to generate an abundance matrix and then spectral information by using a *linear mixture model* [63] with a random Gaussian noise. Our motivation is to have a realistic low-rank covariance matrix, which appears often in many HS images [21]. In this case, the population covariance matrix $\mathbf{\Sigma}$ is not a parameter in the simulation. We generate two millions of vectors by the same abundance matrix and we consider its covariance matrix as $\mathbf{\Sigma}$. Its condition number is equal to 2.7202×10^5. We have generated 4000 random vectors from three Dirichlet distributions $\mathcal{D}_1([9,\ldots,9]), \mathcal{D}_2([3,9,1,\ldots,1])$ and $\mathcal{D}_3([1,1,3,9,9,1,\ldots,1,1])$. In this experiment, the covariance matrix is estimated only with vector from D_1 in two sample sizes ($n = 215$ and $n = 970$). Results of the detection in the 4000 vectors from the three classes (considering 500 vectors from each D_2 and D_3 as anomalies) are illustrated in Fig. 7. We can see that some techniques fail to correctly detect the anomalies. Among the estimators with AUC close to one, the best performance according to volume is clearly given by the $\widehat{\mathbf{\Sigma}}_{\text{SMT}}$. After that, $\widehat{\mathbf{\Sigma}}_{\text{BD}}$ and $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$ performed better than other approaches. From this point, we would like to analyze the behavior of the estimator in the presence of contaminated samples. Accordingly, we substitute 10% of the sample with vectors from $\mathcal{D}_2$. Thus, the AUC and the volume vs false alarm rate for the studied estimators are illustrated in Fig. 7 (c) and (e) with 10% and two sample sizes (215 and 970). We can see, that the idea of constrain the estimation of covariance matrix by the condition number provides detectors ($\widehat{\mathbf{\Sigma}}_{\text{CCN}}$), which outperformed all the other methods in the particular task of AD for Dirichlet distributions even if we reduce the sample size from 970 to 215. Finally, to summarize, the best performances according to the coverage log-volume versus log-false alarm curve in each scenario have been included in Table 2 to make easier the comparison with the result of previous sections.

4 Conclusions and Future Work

This article presents a comparison of many covariance matrix estimators for anomaly detection in HS image processing. We have shown that due to high dimensionality in HS data, classical estimation techniques could fail in AD problem. We evaluated the performance of covariance matrix estimators in the AD problem with three considerations: ratio between sample size and dimension, contamination in the sample, and modification in the distributions of the sample (Gaussian, Cauchy and linear mixing model from Dirichlet distribution). In the Gaussian case with no contamination and much more samples than dimensions, $\widehat{\mathbf{\Sigma}}_{\text{BD}}$ outperformed the other alternatives. However, its performance decreased when the samples contained some contaminated data, or there were insufficient the samples. In Gaussian scenarios with a small contamination rate, $\widehat{\mathbf{\Sigma}}_{\text{t}}$ could obtain satisfactory performance, but its behavior declined with a decrease the sample size in a fixed dimension. Additionally, Geodesic interpolations ($\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$) performed better than linear interpolations ($\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}}$) in most of the cases, especially in heavy-tails distributions. Overall, $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$ and $\widehat{\mathbf{\Sigma}}_{\text{SMT}}$ showed the best performance in most of

TABLE 2. Top-3 performances in different analyzed scenarios

Distribution	Contamination	$c = \frac{d}{n}$	Top three performances		
Gaussian	No	0.2	$\widehat{\mathbf{\Sigma}}_{\text{BD}}$	$\widehat{\mathbf{\Sigma}}_{\text{Efron-Morris}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$
Gaussian	1%	0.2	$\widehat{\mathbf{\Sigma}}_{\text{t}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$	—
Gaussian	10%	0.2	$\widehat{\mathbf{\Sigma}}_{\text{t}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$	—
Gaussian	No	0.9	$\widehat{\mathbf{\Sigma}}_{\text{SMT}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$
Gaussian	1%	0.9	$\widehat{\mathbf{\Sigma}}_{\text{SMT}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$
Gaussian	10%	0.9	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$
Cauchy	No	0.2	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$	$\widehat{\mathbf{\Sigma}}_{\text{t}}$
Cauchy	1%	0.2	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$	$\widehat{\mathbf{\Sigma}}_{\text{CERNN}}$
Cauchy	10%	0.2	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$	$\widehat{\mathbf{\Sigma}}_{\text{t}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$
Cauchy	No	0.9	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$	$\widehat{\mathbf{\Sigma}}_{\text{CERNN}}$
Cauchy	1%	0.9	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$	$\widehat{\mathbf{\Sigma}}_{\text{CERNN}}$
Cauchy	10%	0.9	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$	$\widehat{\mathbf{\Sigma}}_{\text{CCN}}$
Dirichlet	No	0.2	$\widehat{\mathbf{\Sigma}}_{\text{SMT}}$	$\widehat{\mathbf{\Sigma}}_{\text{BD}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$
Dirichlet	10%	0.2	$\widehat{\mathbf{\Sigma}}_{\text{CCN}}$	—	—
Dirichlet	No	0.9	$\widehat{\mathbf{\Sigma}}_{\text{SMT}}$	$\widehat{\mathbf{\Sigma}}_{\text{BD}}$	$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$
Dirichlet	10%	0.9	$\widehat{\mathbf{\Sigma}}_{\text{CCN}}$	—	—

the explored cases. However, note that $\widehat{\mathbf{\Sigma}}_{\text{SMT}}$ was more affected by the contamination than shrinkage-based methods. In contrast, $\widehat{\mathbf{\Sigma}}_{\text{SMT}}$ could adapt better to the data samples generated from linear mixture models. Finally, the recent approach by constraining the condition number ($\widehat{\mathbf{\Sigma}}_{\text{CCN}}$) performed exceptionally well in the difficult case of heavy tails distributions with contaminated data, in addition to all the explored cases in Dirichlet simulations. Future work includes the addition of other techniques of AD based on nonparametric estimation, random subspaces and machine learning techniques. Additionally, we are planning to explore automatic selection of the parameter α for regularized estimators by considering ideas from [41], [64] and [65]. Finally, similar analysis can be done in other important aspects of HS analysis, for instance, target detection, band selection, and so on.

REFERENCES

[1] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 58–69, 2002.

[2] C-I. Chang and S-S. Chiang, "Anomaly detection and classification for hyperspectral imagery," *IEEE Transactions on Geosc. and Rem. Sens.*, vol. 40, no. 6, pp. 1314–1325, 2002.

[3] D. Borghys, V. Kasen, I.and Achard, and C. Perneel, "Hyperspectral anomaly detection: comparative evaluation in scenes with diverse complexity," *Journal of Electrical and Computer Engineering*, vol. 2012, pp. 5, 2012.

[4] S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images," *Aerospace and Electronic Systems Magazine, IEEE*, vol. 25, no. 7, pp. 5–28, 2010.

[5] S. Matteoli, M. Diani, and J. Theiler, "An overview of background modeling for detection of targets and anomalies in hyperspectral remotely sensed imagery," *IEEE Journal of Sel. Topics in Applied Earth Observations and Rem. Sens.*, vol. 7, no. 6, pp. 2317–2336, 2014.

[6] A.P. Schaum, "Hyperspectral anomaly detection beyond RX," in *Defense and Security Symposium*. International Society for Optics and Photonics, 2007, pp. 656502–656502.

[7] Y. Chen, A. Wiesel, and A. O. Hero, "Shrinkage estimation of high dimensional covariance matrices," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 2937–2940.

[8] Y. Chen, A. Wiesel, and A. O. Hero, "Robust shrinkage estimation of high-dimensional covariance matrices," *Signal Processing, IEEE Transactions on*, vol. 59, no. 9, pp. 4097–4107, 2011.

[9] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 763–767, 1996.

[10] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *The Annals of Statistics*, vol. 15, no. 1, pp. 234–251, 1987.

[11] S. Matteoli, M. Diani, and G. Corsini, "Improved estimation of local background covariance matrix for anomaly detection in hyperspectral images," *Optical Engineering*, vol. 49, no. 4, pp. 1–16, 2010.

[12] J. Frontera-Pons, M. Mahot, J.P. Ovarlez, and F. Pascal, "Robust Detection using M- estimators for Hyperspectral Imaging," in *IEEE Workshop on Hyperspectral Image and Signal Processing*, 2012.

[13] J. Theiler, G. Cao, L.R. Bachega, and C.A. Bouman, "Sparse matrix transform for hyperspectral image processing," *IEEE Journal of Sel. Topics in Signal Processing*, vol. 5, no. 3, pp. 424–437, 2011.

[14] Y. I. Abramovich and O. Besson, "Regularized Covariance Matrix Estimation in Complex Elliptically Symmetric Distributions Using the Expected Likelihood Approach- Part 1: The Over-Sampled Case," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5807–5818, 2013.

[15] O. Besson and Y. I. Abramovich, "Regularized Covariance Matrix Estimation in Complex Elliptically Symmetric Distributions Using the Expected Likelihood Approach– Part 2: The Under-Sampled Case," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5819–5829, 2013.

[16] J-H. Won, J. Lim, S-J. Kim, and B. Rajaratnam, "Condition-number-regularized covariance estimation," *Journal of the Royal Statistical Society: Series B*, vol. 75, no. 3, pp. 427–450, 2013.

[17] Y. Sun, P. Babu, and D. Palomar, "Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms," *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5143–5156, 2014.

[18] J. Theiler, "By definition undefined: Adventures in anomaly (and anomalous change) detection," in *IEEE Workshop on Hyperspectral Image and Signal Processing*, 2014.

[19] 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.

[20] P. C. Mahalanobis, "On the generalized distance in statistics," *Proc. of the National Institute of Sciences (Calcutta)*, vol. 2, pp. 49–55, 1936.

[21] D. Manolakis, D. Marden, and G. A. Shaw, "Hyperspectral image processing for automatic target detection applications," *Lincoln Laboratory Journal*, vol. 14, no. 1, pp. 79–116, 2003.

[22] T. W. Anderson, *An introduction to multivariate statistical analysis*, Wiley, 1958.

[23] V. A. Marvcenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Sbornik: Mathematics*, vol. 1, no. 4, pp. 457–483, 1967.

[24] A. Edelman and N. R. Rao, "Random matrix theory," *Acta Numerica*, vol. 14, no. 1, pp. 233–297, 2005.

[25] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, Number 118. Cambridge University Press, 2010.

[26] I. S. Reed, J. D. Mallett, and L. E. Brennan, "Rapid convergence rate in adaptive arrays," *Aerospace and Electronic Systems, IEEE Transactions on*, , no. 6, pp. 853–863, 1974.

[27] C. E. Davidson and A. Ben-David, "Performance loss of multivariate detection algorithms due to covariance estimation," in *SPIE Europe Remote Sensing*, 2009, pp. 74770–74770.

[28] G. Frahm, *Generalized elliptical distributions: theory and applications*, Ph.D. thesis, Universität zu Köln, 2004.

[29] R. Maronna, "Robust M-estimators of multivariate location and scatter," *The Annals of Statistics*, pp. 51–67, 1976.

[30] R. Couillet, F. Pascal, and J. W. Silverstein, "The random matrix regime of maronna's m-estimator with elliptically distributed samples," *Journal of Multivariate Analysis*, vol. 139, no. 0, pp. 56 – 78, 2015.

[31] T. Zhang, X. Cheng, and A. Singer, "Marchenko-pastur law for Tyler's and Maronna's M-estimators," *arXiv:1401.3424*, 2014.

[32] C. E. Davidson and A. Ben-David, "On the use of covariance and correlation matrices in hyperspectral detection," in *Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, 2011, pp. 1–6.

[33] T. K. Moon, "The expectation-maximization algorithm," *Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[34] C. Liu and D. B. Rubin, "ML estimation of the t distribution using EM and its extensions, ECM and ECME," *Statistica Sinica*, vol. 5, no. 1, pp. 19–39, 1995.

[35] S. Nadarajah and S. Kotz, "Estimation methods for the multivariate t distribution," *Acta Applicandae Mathematicae*, vol. 102, no. 1, pp. 99–118, 2008.

[36] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989.

[37] S. Velasco-Forero, M. Chen, A. Goh, and S. K. Pang, "A comparative analysis of covariance matrix estimation in anomaly detection," in *IEEE Workshop on Hyperspectral Image and Signal Processing*, 2014.

[38] O. Ledoit and M. Wolf, "Honey, i shrunk the sample covariance matrix," *UPF Economics and Business Working Paper*, , no. 691, 2003.

[39] B. Efron and C. Morris, "Data analysis using Steinś estimator and its generalizations," *Journal of the American Statistical Association*, vol. 70, no. 350, pp. 311–319, 1975.

[40] C. Stein, "Estimation of a covariance matrix," *Rietz Lecture*, 1975.

[41] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of multivariate analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[42] Y. I. Abramovich and N. K. Spencer, "Diagonally loaded normalised sample matrix inversion (lnsmi) for outlier-resistant adaptive filtering," in *ICASSP 2007*. IEEE, 2007, vol. 3, pp. 1111–1105.

[43] A. Wiesel, "Unified framework to regularized covariance estimation in scaled gaussian models," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 29–38, 2012.

[44] X. Pennec, "Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements," *Journal of Mathematical Imaging and Vision*, vol. 25, no. 1, pp. 127–154, 2006.

[45] M. Chen, S.K. Pang, T.J. Cham, and A. Goh, "Visual tracking with generative template model based on riemannian manifold of covariances," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*. IEEE, 2011, pp. 1–8.

[46] Avishai Ben-David and Justin Marks, "Geodesic paths for time-dependent covariance matrices in a Riemannian manifold," *IEEE Transactions on Geosc. and Rem. Sens. Letters*, vol. 11, pp. 1499–1503, 2014.

[47] E. C. Chi and K. Lange, "Stable estimation of a covariance matrix guided by nuclear norm penalties," *Computational statistics & data analysis*, vol. 80, pp. 117–128, 2014.

[48] R. Nadakuditi and J. W. Silverstein, "Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples," *IEEE Journal of Sel. Topics in Signal Processing*, vol. 4, no. 3, pp. 468–480, 2010.

[49] A. Ben-David and C. E. Davidson, "Eigenvalue estimation of hyperspectral wishart covariance matrices from limited number of samples," *IEEE Transactions on Geosc. and Rem. Sens.*, vol. 50, no. 11, pp. 4384–4396, 2012.

[50] R. Menon, P. Gerstoft, and W.S. Hodgkiss, "Asymptotic eigenvalue density of noise covariance matrices," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3415–3424, 2012.

[51] B. Efron and C. Morris, "Multivariate empirical bayes and estimation of covariance matrices," *The Annals of Statistics*, pp. 22–32, 1976.

[52] G. Cao, L. R. Bachega, and C. A. Bouman, "The Sparse Matrix Transform for Covariance Estimation and Analysis of High Dimensional Signals," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 625–640, mar 2011.

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

[54] J. Frontera-Pons, M. Mahot, J. P. Ovarlez, F. Pascal, S.K. Pang, and J. Chanussot, "A class of robust estimates for detection in hyperspectral images using elliptical distributions background," in *International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2012, pp. 4166–4169.

[55] J. P. Egan, *Signal Detection and ROC Analysis*, Academic Press, 1975.

[56] J. Theiler and D. Hush, "Statistics for characterizing data on the periphery," in *International Geoscience and Remote Sensing Symposium*. IEEE, 2010, pp. 4764–4767.

[57] J. Theiler, "Ellipsoid-simplex hybrid for hyperspectral anomaly detection," in *IEEE Workshop on Hyperspectral Image and Signal Processing*, 2011.

[58] S. Velasco-Forero and J. Angulo, "Classification of hyperspectral images by tensor modeling and additive morphological decomposition," *Pattern Recognition*, vol. 46, no. 2, pp. 566–577, 2013.

[59] N. Keshava and J. F. Mustard, "Spectral unmixing," *Signal Processing Magazine*, vol. 19, no. 1, pp. 44–57, 2002.

[60] J. Aitchison, "The statistical analysis of compositional data," *Journal of the Royal Statistical Society. Series B*, pp. 139–177, 1982.

[61] J. M. P. Nascimento and J. M. Bioucas-Dias, "Hyperspectral unmixing based on mixtures of Dirichlet components," *IEEE Transactions on Geosc. and Rem. Sens.*, vol. 50, no. 3, pp. 863–878, 2012.

[62] J. M. P. Nascimento and J. M. Bioucas-Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Transactions on Geosc. and Rem. Sens.*, vol. 43, no. 4, pp. 898–910, 2005.

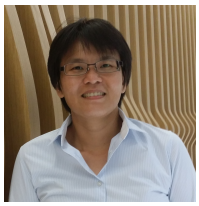
[63] A. Plaza, P. Martínez, R. Pérez, and J. Plaza, "A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data," *IEEE Transactions on Geosc. and Rem. Sens.*, vol. 42, no. 3, pp. 650–663, 2004.

[64] J. Theiler, "The incredible shrinking covariance estimator," in *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 2012, pp. 83910P–83910P.

[65] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, pp. 199–227, 2008.

Santiago Velasco-Forero received a B.Sc. in Statistics from the National University of Colombia, M.Sc. in Mathematics in University of Puerto Rico, and Ph.D. in image processing at École des Mines de Paris, France. During the period 2013-2014, he pursued research on multivariate image analysis and processing with the ITWM - Fraunhofer Institute in Kaiserlautern, Germany and the Department of Mathematics of the National University of Singapore. His research interests included image processing, multivariate statistics, computer vision, and mathematical morphology. Currently, he is tenure track researcher at the École des Mines de Paris in France.

Marcus Chen received the Bachelor of Science degree in Electrical and Computer Engineering from Carnegie Mellon University, Pittsburgh, PA, in 2007, and the Master of Science degree in Electrical Engineering from Stanford University, Stanford, CA, in 2008. He is currently working toward the PhD degree at Nanyang Technological University, Singapore. His research interests include pattern recognition, optimization, large scale image search, and remote sensing.



Alvina Goh completed her Ph.D. in Biomedical Engineering and M.S.E. in Applied Mathematics and Statistics from the Johns Hopkins University. She received her M.S. and B.S.E. in Electrical Engineering from California Institute of Technology and University of Michigan at Ann Arbor, respectively. Currently, she is a researcher at DSO National Laboratories in Singapore and an adjunct assistant professor at the National University of Singapore. Her research interests include speech processing, machine learning, computer vision, and medical imaging.

Sze Kim Pang completed his Ph.D. in Bayesian Statistical Signal Processing in the University of Cambridge and M.Eng. in the Imperial College London. Currently, he is a Principal Member of technical staff at DSO National Laboratories in Singapore.