# Sparsity-based Cholesky Factorization and Its Application to Hyperspectral Anomaly Detection


**Ahmad W. Bitar & Jean-Philippe Ovarlez & Loong-Fah Cheong**

SONDRA/CentraleSupélec, Plateau du Moulon, 3 rue Joliot-Curie, F-91190 Gif-sur-Yvette, France
ONERA, DEMR/TSI, Chemin de la Hunière, 91120 Palaiseau, France
National University of Singapore (NUS), Singapore, Singapore

**Contact Information:**
SONDRA laboratory
CentraleSupélec
Email: ahmad.bitar@centralesupelec.fr



**Abstract**

Estimating large covariance matrices has been a long-standing important problem in many applications and has attracted increased attention over several decades. This paper deals with two methods based on pre-existing works to impose sparsity on the covariance matrix via its unit lower triangular matrix (aka Cholesky factor) $\mathbf{T}$. The first method serves to estimate the entries of $\mathbf{T}$ using the Ordinary Least Squares (OLS), then imposes sparsity by exploiting some generalized thresholding techniques such as Soft and Smoothly Clipped Absolute Deviation (SCAD). The second method directly estimates a sparse version of $\mathbf{T}$ by penalizing the negative normal log-likelihood with $L_1$ and SCAD penalty functions. The resulting covariance estimators are always guaranteed to be positive definite. Some Monte-Carlo simulations as well as experimental data demonstrate the effectiveness of our estimators for hyperspectral anomaly detection using the Kelly anomaly detector.


## Introduction

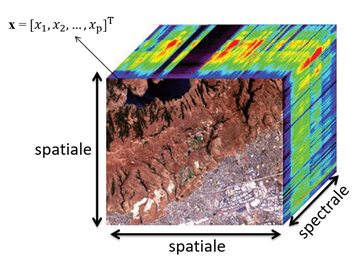



**Fig. 1.** A hyperspectral image (HSI)

A hyperspectral image (HSI) is a three dimensional data cube consisting of a series of images of the same spatial scene in a contiguous and multiple narrow spectral wavelength (color) bands. Each pixel in the HSI is a $p$-dimensional vector, $\mathbf{x} \in \mathbb{R}^p$, where $p$ stands for the total number of spectral bands. With the rich information afforded by the high spectral dimensionality, target detection is not surprisingly one of the most important applications in hyperspectral imagery.

In many situations of practical interest, we do not have sufficient a priori information to specify the statistics of the target class. More precisely, the target's spectra is not provided to the user. This unknown target is referred as "anomaly" having a very different spectra from the surrounding background.

Different Gaussian-based anomaly detectors have been proposed in the literature. The detection performance of these detectors mainly depend on the true unknown covariance matrix (of the background surrounding the test pixel) whose entries have to be carefully estimated specially in large dimensions. Due to the fact that in hyperspectral imagery, the number of covariance matrix parameters to estimate grows with the square of the spectral dimension, it becomes impractical to use traditional covariance estimators where the target detection performance can deteriorate significantly. Many a time, the researchers assume that compounding the large dimensionality problem can be alleviated by leveraging on the assumption that the true unknown covariance matrix is sparse, namely, many entries are zero.

## Covariance estimation via linear regression

Suppose that we observe a sample of $n$ independent and identically distributed $p$-random vectors, $\{\mathbf{x}_i\}_{i\in[1,n]}$, each follows a multivariate Gaussian distribution with zero mean and unknown covariance matrix $\mathbf{\Sigma} = [\sigma_{g,l}]_{p\times p}$. The first traditional estimator we consider in this paper is the Sample Covariance Matrix (SCM), defined as $\hat{\mathbf{\Sigma}}_{SCM} = [\hat{\sigma}_{g,l}]_{p\times p} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$.

In order to address the positivity definiteness constraint problem of $\hat{\mathbf{\Sigma}}_{SCM}$, Pourahmadi [1] has modeled the covariance matrices via linear regressions. This is done by letting $\hat{\mathbf{x}} = [\hat{x}_1, \ldots, \hat{x}_p]^T \in \mathbb{R}^p$, and consider each element $\hat{x}_t$, $t \in [1,\ p]$, as the linear least squares predictor of $x_t$ based on its $t-1$ predecessors $\{x_j\}_{j\in[1,t-1]}$. In particular, for $t \in [1,\ p]$, let

$$\hat{x}_t = \sum_{j=1}^{t-1} C_{t,j}\, x_j, \quad \mathbf{T}\mathbf{\Sigma}\mathbf{T}^T = \mathbf{D}\,. \tag{1}$$

where $\mathbf{T}$ is a unit lower triangular matrix with $-C_{t,j}$ in the $(t,j)$th position for $t \in [2,\ p]$ and $j \in [1,\ t-1]$, and $\mathbf{D}$ is a diagonal matrix with $\theta_t^2 = \mathrm{var}(\epsilon_t)$ as its diagonal entries, where $\epsilon_t = x_t - \hat{x}_t$ is the prediction error for $t \in [1,\ p]$. Note that for $t = 1$, let $\hat{x}_1 = E(x_1) = 0$, and hence, $\mathrm{var}(\epsilon_1) = \theta_1^2 = E\left[(x_1)^2\right]$. Given a sample $\{\mathbf{x}_i\}_{i\in[1,n]}$, with $n > p$, a natural estimate of $\mathbf{T}$ and $\mathbf{D}$, denoted as $\mathbf{T}_{OLS}$ and $\mathbf{D}_{OLS}$ in this paper, is simply done by plugging in the OLS estimates of the regression coefficients and residual variances in (1), respectively. In this paper, we shall denote the second traditional estimator by $\hat{\mathbf{\Sigma}}_{OLS} = \hat{\mathbf{T}}_{OLS}^{-1}\hat{\mathbf{D}}_{OLS}\hat{\mathbf{T}}_{OLS}^{-T}$.

## Main contributions

Before describing the two methods, we want to recall the definition for $\hat{\mathbf{\Sigma}}_{OLS}$. Given a sample $\{\mathbf{x}_i\}_{i\in[1,n]}$, we have:

$$x_{i,t} = \sum_{j=1}^{t-1} C_{t,j}\, x_{i,j} + \epsilon_{i,t} \quad t \in [2,\ p], \quad i \in [1,\ n]. \tag{2}$$

By writing (2) in vector-matrix form for any $t \in [2,\ p]$, one obtains the simple linear regression model:

$$\mathbf{y}_t = \mathbf{A}_{n,t}\boldsymbol{\beta}_t + \mathbf{e}_t\,, \tag{3}$$

where $\mathbf{y}_t = [x_{1,t}, \cdots, x_{n,t}]^T \in \mathbb{R}^n$, $\mathbf{A}_{n,t} = [x_{i,j}]_{n\times(t-1)}$, $\boldsymbol{\beta}_t = [C_{t,1}, \cdots, C_{t,t-1}]^T \in \mathbb{R}^{(t-1)}$, and $\mathbf{e}_t = [\epsilon_{1,t}, \cdots, \epsilon_{n,t}]^T \in \mathbb{R}^n$.
When $n > p$, the OLS estimate of $\boldsymbol{\beta}_t$ and the corresponding residual variance are plugged in $\mathbf{T}$ and D for each $t \in [2,\ p]$, respectively. At the end, one obtains the estimator $\hat{\mathbf{\Sigma}}_{OLS} = \hat{\mathbf{T}}_{OLS}^{-1}\hat{\mathbf{D}}_{OLS}\hat{\mathbf{T}}_{OLS}^{-T}$. Note that $\hat{\mathbf{T}}_{OLS}$ has $-\hat{C}_{t,j}^{OLS}$ in the $(t,j)$th position for $t \in [2,\ p]$ and $j \in [1,\ t-1]$.

## Generalized thresholding based Cholesky Factorization

For any $0 \le \lambda \le 1$, we define a matrix thresholding operator $Th(.)$ and denote by $Th(\hat{\mathbf{T}}_{OLS}) = [Th(-\hat{C}_{t,j}^{OLS})]_{p\times p}$ the matrix resulting from applying a specific thresholding operator $Th(.) \in \{\text{Soft, SCAD}\}$ to each element of the matrix $\mathbf{T}^{OLS}$ for $t \in [2,\ p]$ and $j \in [1,\ t-1]$.
We consider the following minimization problem:

$$Th(\hat{\mathbf{T}}_{OLS}) = \underset{\mathbf{T}}{\mathrm{argmin}} \sum_{t=2}^{p}\sum_{j=1}^{t-1}\left\{\frac{1}{2}(\hat{C}_{t,j}^{OLS} - C_{t,j})^2 + p_\lambda\{|C_{t,j}|\}\right\} \tag{4}$$

where $p_\lambda \in \{p_\lambda^{L_1}, p_{\lambda,a>2}^{SCAD}\}$. We have $p_\lambda^{L_1}(|C_{t,j}|) = \lambda|C_{t,j}|$, and $p_{\lambda,a>2}^{SCAD}(|C_{t,j}|) =$

$$\begin{cases} \lambda|C_{t,j}| & \text{if } |C_{t,j}| \le \lambda \\ -\frac{|C_{t,j}^2| - 2a\lambda|C_{t,j}| + \lambda^2}{2(a-1)} & \text{if } \lambda < |C_{t,j}| \le a\lambda\,. \\ \frac{(a+1)\lambda^2}{2} & \text{if } |C_{t,j}| > a\lambda \end{cases}$$

Solving (4) with $p_\lambda^{L_1}$ and $p_{\lambda,a>2}^{SCAD}$, yields a closed-form Soft and SCAD thresholding rules, respectively [2], [3]. The value $a = 3.7$ was recommended by Fan and Li [3]. Despite the application here is different than in [3], for simplicity, we use the same value throughout the paper.
We shall designate the two thresholded matrices by $\hat{\mathbf{T}}_{Soft}$ and $\hat{\mathbf{T}}_{SCAD}$, that apply Soft and SCAD on $\hat{\mathbf{T}}_{OLS}$, respectively. We denote our first two estimators as:

$$\boxed{\hat{\mathbf{\Sigma}}_{OLS}^{Soft} = \hat{\mathbf{T}}_{Soft}^{-1}\hat{\mathbf{D}}_{OLS}\hat{\mathbf{T}}_{Soft}^{-T}}$$

$$\boxed{\hat{\mathbf{\Sigma}}_{OLS}^{SCAD} = \hat{\mathbf{T}}_{SCAD}^{-1}\hat{\mathbf{D}}_{OLS}\hat{\mathbf{T}}_{SCAD}^{-T}}$$

## A generalization of the estimator in [4]

Note that $\det(\mathbf{T}) = 1$ and $\mathbf{\Sigma} = \mathbf{T}^{-1}\,\mathbf{D}\,\mathbf{T}^{-T}$. It follows that $\det(\mathbf{\Sigma}) = \det(\mathbf{D}) = \prod_{t=1}^{p}\theta_t^2$. Hence, the negative normal log-likelihood of $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$, ignoring an irrelevant constant, satisfies:
$\Lambda = -2\log(L(\mathbf{\Sigma}, \mathbf{x}_1, \cdots, \mathbf{x}_n)) = n\log(\det(\mathbf{D})) + \mathbf{X}^T(\mathbf{T}^T\mathbf{D}^{-1}\mathbf{T})\,\mathbf{X} = n\log(\det(\mathbf{D})) + (\mathbf{T}\,\mathbf{X})^T\,\mathbf{D}^{-1}(\mathbf{T}\,\mathbf{X}) = n\sum_{t=1}^{p}\log\theta_t^2 + \sum_{t=1}^{p}\sum_{i=1}^{n}\epsilon_{i,t}^2/\theta_t^2$. By adding a penalty function $\sum_{t=2}^{p}\sum_{j=1}^{t-1} p_\alpha\{|C_{t,j}|\}$ to $\Lambda$, where $p_\alpha \in \{p_\alpha^{L_1}, p_{\alpha,a>2}^{SCAD}\}$ (see subsection III. A) with $\alpha \in [0, \infty)$, we have:

$$n\log\theta_1^2 + \sum_{i=1}^{n}\frac{\epsilon_{1,1}^2}{\theta_1^2} + \sum_{t=2}^{p}\left(n\log\,\theta_t^2 + \sum_{i=1}^{n}\frac{\epsilon_{i,t}^2}{\theta_t^2} + \sum_{j=1}^{t-1} p_\alpha\{|C_{t,j}|\}\right) \tag{5}$$

Obviously, minimizing (5) with respect to $\theta_1^2$ and $\theta_t^2$ gives the solutions $\hat{\theta}_1^2 = \frac{1}{n}\sum_{i=1}^{n}\epsilon_{i,1}^2 = \frac{1}{n}\sum_{i=1}^{n}x_{i,1}^2$ and $\hat{\theta}_t^2 = \frac{1}{n}\sum_{i=1}^{n}\epsilon_{i,t}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{i,t} - \sum_{j=1}^{t-1}C_{t,j}x_{i,j})^2$, respectively.
It remains to estimate the entries of $\mathbf{T}$ by minimizing (5) with respect to $\boldsymbol{\beta}_t$. From equation (2) and (3), the minimization problem to solve for each $t \in [2,\ p]$ is:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_t &= \underset{\boldsymbol{\beta}_t}{\mathrm{argmin}} \sum_{i=1}^{n}\frac{\epsilon_{i,t}^2}{\theta_t^2} + \sum_{j=1}^{t-1} p_\alpha\{|C_{t,j}|\} \\ &= \underset{\boldsymbol{\beta}_t}{\mathrm{argmin}} \frac{1}{\theta_t^2}\sum_{i=1}^{n}\left(x_{i,t} - \sum_{j=1}^{t-1}C_{t,j}x_{i,j}\right)^2 + \sum_{j=1}^{t-1} p_\alpha\{|C_{t,j}|\} \\ &= \underset{\boldsymbol{\beta}_t}{\mathrm{argmin}} \frac{1}{\theta_t^2}||\mathbf{y}_t - \mathbf{A}_{n,t}\boldsymbol{\beta}_t||_F^2 + \sum_{j=1}^{t-1} p_\alpha\{|C_{t,j}|\}\end{aligned} \tag{6}$$

By denoting $l(\boldsymbol{\beta}_t) = \frac{1}{\theta_t^2}||\mathbf{y}_t - \mathbf{A}_{n,t}\boldsymbol{\beta}_t||_F^2$ and $r(\boldsymbol{\beta}_t) = \sum_{j=1}^{t-1} p_\alpha\{|C_{t,j}|\} = \sum_{j=1}^{t-1} r_j(C_{t,j})$, we solve $\boldsymbol{\beta}_t$ iteratively using the General Iterative Shrinkage and Thresholding (GIST) algorithm [5]:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_t^{(k+1)} &= \underset{\boldsymbol{\beta}_t}{\mathrm{argmin}}\; l(\boldsymbol{\beta}_t^{(k)}) + r(\boldsymbol{\beta}_t) + (\nabla l(\boldsymbol{\beta}_t^{(k)}))^T(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t^{(k)}) \\ &\qquad + \frac{w^{(k)}}{2}||\boldsymbol{\beta}_t - \boldsymbol{\beta}_t^{(k)}||^2 \\ &= \underset{\boldsymbol{\beta}_t}{\mathrm{argmin}}\; 0.5||\boldsymbol{\beta}_t - \mathbf{u}_t^{(k)}||^2 + \frac{1}{w^{(k)}}r(\boldsymbol{\beta}_t)\end{aligned} \tag{7}$$

where $\mathbf{u}_t^{(k)} = \boldsymbol{\beta}_t^{(k)} - \nabla l(\boldsymbol{\beta}_t^{(k)})/w^{(k)}$, and $w^{(k)}$ is the step size initialized using the Barzilai-Browein rule [6].
By decomposing (7) into independent (t-1) univariate optimization problems, we have for $j = 1, \cdots, t-1$:

$$C_{t,j}^{(k+1)} = \underset{C_{t,j}}{\mathrm{argmin}}\; 0.5||C_{t,j} - u_{t,j}^{(k)}||^2 + \frac{1}{w^{(k)}}r_j(C_{t,j}) \tag{8}$$

where $\mathbf{u}_t^{(k)} = [u_{t,1}^{(k)}, \cdots, u_{t,t-1}^{(k)}]^T \in \mathbb{R}^{(t-1)}$.
By solving (8) with the $L_1$-norm penalty, $p_\alpha^{L_1}$, we have the following closed form solution:

$$C_{t,j,(L_1)}^{(k+1)} = sign(u_{t,j}^{(k)})\,\max(0,\, |u_{t,j}^{(k)}| - \alpha/w^{(k)}) \tag{9}$$

For the SCAD penalty function, $p_{\alpha,a>2}^{SCAD}$, we can observe that it contains three parts for three different conditions. In this case, by recasting problem (8) into three minimization sub-problems for each condition, and after solving them, one can obtain the following three sub-solutions $h_{t,j}^1$, $h_{t,j}^2$, and $h_{t,j}^3$, where:

$$\begin{aligned}h_{t,j}^1 &= sign(u_{t,j}^{(k)})\,\min(\alpha, \max(0, |u_{t,j}^{(k)}| - \alpha/w^{(k)})), \\ h_{t,j}^2 &= sign(u_{t,j}^{(k)})\,\min(a\alpha, \max(\alpha, \tfrac{w^{(k)}|u_{t,j}^{(k)}|(a-1) - a\alpha}{w^{(k)}(a-2)})), \\ h_{t,j}^3 &= sign(u_{t,j}^{(k)})\,\max(a\alpha, |u_{t,j}^{(k)}|).\end{aligned}$$

| Models | $\hat{\mathbf{\Sigma}}$ | $\hat{\mathbf{\Sigma}}_{SCM}$ | $\hat{\mathbf{\Sigma}}_{OLS}$ | $\hat{\mathbf{\Sigma}}_{OLS}^{Soft}$ | $\hat{\mathbf{\Sigma}}_{OLS}^{SCAD}$ | $\hat{\mathbf{\Sigma}}_{L_1}$ | $\hat{\mathbf{\Sigma}}_{SCM}^{Soft}$ | $\hat{\mathbf{\Sigma}}_{SMT}$ | $B_k(\hat{\mathbf{\Sigma}}_{SM})$ | $\hat{\mathbf{\Sigma}}_{SCM}^{Soft}$ | $\hat{\mathbf{\Sigma}}_{SCM}^{SCAD}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model 1** | 0.9541 | 0.7976 | 0.8331 | 0.9480 | 0.9480 | **0.9509** | **0.9509** | 0.9503 | **0.9509** | **0.9509** | **0.9509** |
| **Model 2** | 0.9540 | 0.7977 | 0.8361 | 0.9124 | 0.9124 | 0.9264 | 0.9264 | 0.9184 | **0.9478** | 0.9274 | 0.9270 |
| **Model 3** | 0.9541 | 0.7978 | 0.8259 | 0.8169 | 0.8257 | 0.8236 | **0.8261** | 0.7798 | 0.5321 | 0.5969 | 0.5781 |
| **MUSE** | Not known | 0.6277 | 0.6575 | 0.9620 | **0.9643** | 0.8844 | 0.8844 | 0.7879 | 0.9277 | 0.7180 | 0.7180 |

**Table 1.** List of AUC values.

Hence, we have the following closed form solution:

$$\begin{aligned}C_{t,j,(SCAD)}^{(k+1)} = \underset{q_{t,j}}{\mathrm{argmin}}\; &0.5||q_{t,j} - u_{t,j}^{(k)}||^2 + \frac{1}{w^{(k)}}r_j(q_{t,j}) \\ &s.t.\quad q_{t,j} \in \{h_{t,j}^1,\ h_{t,j}^2,\ h_{t,j}^3\}\end{aligned} \tag{10}$$

At the end, we denote our last two estimators as:

$$\boxed{\hat{\mathbf{\Sigma}}_{L_1} = \hat{\mathbf{T}}_{L_1}^{-1}\hat{\mathbf{D}}\,\hat{\mathbf{T}}_{L_1}^{-T}}$$

$$\boxed{\hat{\mathbf{\Sigma}}_{SCAD} = \hat{\mathbf{T}}_{SCAD}^{-1}\hat{\mathbf{D}}\,\hat{\mathbf{T}}_{SCAD}^{-T}}$$

where $\hat{\mathbf{T}}_{L_1}$ and $\hat{\mathbf{T}}_{SCAD}$ have respectively $-\hat{C}_{t,j,(L_1)}$ and $-\hat{C}_{t,j,(SCAD)}$ in the $(t,j)$th position for $t \in [2,\ p]$ and $j \in [1,\ t-1]$, whereas $\hat{\mathbf{D}}$ has the entries $(\hat{\theta}_1^2, \hat{\theta}_t^2)$ on its diagonal.

## Hyperspectral anomaly detection

Suppose the following signal model:

$$\begin{cases} H_0 : \mathbf{x} = \mathbf{n}, & \mathbf{x}_i = \mathbf{n}_i, \quad i = 1, \cdots, n \\ H_1 : \mathbf{x} = \gamma\,\mathbf{d} + \mathbf{n}, & \mathbf{x}_i = \mathbf{n}_i, \quad i = 1, \cdots, n \end{cases} \tag{11}$$

where $\mathbf{n}_1, \cdots, \mathbf{n}_n$ are $n$ i.i.d $p$-vectors, each follows a multivariate Normal distribution $\mathcal{N}(0, \mathbf{\Sigma})$. $\mathbf{d}$ is an unknown steering vector and which denotes the presence of an anomalous signal with unknown amplitude $\gamma > 0$. The Kelly anomaly detector [7] is described as follows:

$$D_{KellyAD\hat{\mathbf{\Sigma}}}(\mathbf{x}) = \mathbf{x}^T\hat{\mathbf{\Sigma}}_{SCM}^{-1}\mathbf{x} \underset{H_0}{\overset{H_1}{\gtrless}} \delta\,, \tag{12}$$

where $\delta$ is a prescribed threshold value. In the following two subsections, the detection performances of the estimators, when are plugged in $D_{KellyAD,\hat{\mathbf{\Sigma}}}$ are evaluated by the Receiver Operating Characteristics (ROC) curves and their corresponding Area Under Curves (AUC) values.
**Note that the tuning parameter $\lambda$ and $\alpha$ are chosen automatically using a 5-fold crossvalidated loglikelihood procedure (see Subsection 4.2 in [4] for details).**

## Monte-Carlo simulations

The experiments are conducted on three covariance models:

- Model 1: $\mathbf{\Sigma} = \mathbf{I}$, the identity matrix,
- Model 2: the Autoregressive model order 1, AR(1), $\mathbf{\Sigma} = [\sigma_{gl}]_{p\times p}$, where $\sigma_{gl} = c^{|g-l|}$, for $c = 0.3$,
- Model 3: $\Sigma = [\sigma_{gl}]_{p\times p}$, where $\sigma_{gl} = (1 - ((|g-l|)/r))_+$, for $r = p/2$: the triangular matrix.

The computations have been made through $10^5$ Monte-Carlo trials and the ROC curves are drawn for a signal to noise ratio equal to 15dB. We choose $n = 80$ for covariance estimation under Gaussian assumption, and set $p = 60$. The artificial anomaly we consider is a vector containing normally distributed pseudorandom numbers (to have fair results, the same vector is used for the three models). The ROC curves for Model 1, 2 and 3 are shown in Fig. 2, and their corresponding AUC values are presented in Table 1.
The estimators used in comparison are: $\hat{\mathbf{\Sigma}}_{SCM}$, $\hat{\mathbf{\Sigma}}_{OLS}$, $\hat{\mathbf{\Sigma}}_{SMT}$ [8], $B_k(\hat{\mathbf{\Sigma}}_{SCM})$ [9], $\hat{\mathbf{\Sigma}}_{SCM}^{Soft}$ [2], et $\hat{\mathbf{\Sigma}}_{SCM}^{SCAD}$ [2].

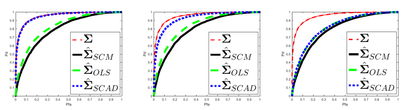
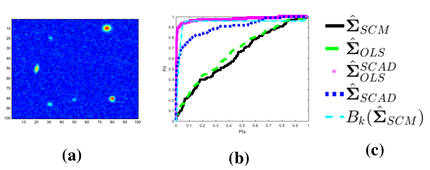



Fig. 2. ROC curves for the three Models. Columns from left to right: Model 1, Model 2, Model 3.

For both Model 1 and 2, our estimators significantly improve the detection performances comparing to those of the traditional estimators ($\hat{\mathbf{\Sigma}}_{SCM}$, $\hat{\mathbf{\Sigma}}_{OLS}$), and have competitive detection results with state-of-the-art. An important finding is that even for a non sparse covariance model (that is, Model 3) , our estimators do not show a harm on the detection when compared to those of $\hat{\mathbf{\Sigma}}_{SCM}$, $\hat{\mathbf{\Sigma}}_{OLS}$. Despite $\hat{\mathbf{\Sigma}}_{OLS}^{Soft}$, $\hat{\mathbf{\Sigma}}_{OLS}^{SCAD}$ and $\hat{\mathbf{\Sigma}}_{L_1}$ have slightly lower AUC values than for $\hat{\mathbf{\Sigma}}_{OLS}$, this is still a negligible degradation on the detection. Thus, considering that $\hat{\mathbf{\Sigma}}_{OLS}^{Soft}$, $\hat{\mathbf{\Sigma}}_{OLS}^{SCAD}$ and $\hat{\mathbf{\Sigma}}_{L_1}$ have no worse detection results than to that of $\hat{\mathbf{\Sigma}}_{OLS}$ is still acceptable.

## Application on experimental data

Our estimators are now evaluated for galaxy detection on the Multi Unit Spectroscopic Explorer (MUSE) data cube (see [10]). It is a $100 \times 100$ image and consists of 3600 bands in wavelengths ranging from 465-930 nm. We used one band of each 60, so that 60 bands in total. Figure 3**(a)** exhibits the mean power in dB over the 60 bands. The covariance matrix is estimated using a sliding window of size $9 \times 9$, having $n = 80$ secondary data (after excluding only the test pixel). The mean has been removed from the given HSI. Figure 3**(b)** exhibits the ROC curves of our estimators when compared to some others, and their AUC values are shown in Table 1.
The estimators $\hat{\mathbf{\Sigma}}_{OLS}^{Soft}$, $\hat{\mathbf{\Sigma}}_{OLS}^{SCAD}$ achieve higher detection results than for all the others, whereas both $\hat{\mathbf{\Sigma}}_{L_1}$ and $\hat{\mathbf{\Sigma}}_{SCAD}$ achieve only a lower AUC values than for $B_k(\hat{\mathbf{\Sigma}}_{SCM})$.




**(a)** **(b)** **(c)**

**Fig. 3. (a)** MUSE HSI (average). **(b)** ROC curves for MUSE. **(c)** Legend.

## References


[1] M. Pourahmadi, "Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation," *Biometrika*, vol. 86, no. 3, pp. 677–690, 1999. [Online]. Available: http://biomet.oxfordjournals.org/content/86/3/677.abstract

[2] A. J. Rothman, E. Levina, and J. Zhu, "Generalized thresholding of large covariance matrices," *Journal of the American Statistical Association*, vol. 104, no. 485, pp. 177–186, 2009. [Online]. Available: http://www.jstor.org/stable/40591909

[3] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001. [Online]. Available: http://dx.doi.org/10.1198/016214501753382273

[4] J. Z. Huang, N. Liu, M. Pourahmadi, and L. Liu, "Covariance matrix selection and estimation via penalised normal likelihood," *Biometrika*, vol. 93, no. 1, pp. 85–98, 2006. [Online]. Available: http://biomet.oxfordjournals.org/content/93/1/85.abstract

[5] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye, "Gist: General iterative shrinkage and thresholding for non-convex sparse learning," 2013.

[6] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA Journal of Numerical Analysis*, vol. 8, no. 1, pp. 141–148, 1988.

[7] E. J. Kelly, "An adaptive detection algorithm," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 23, no. 1, pp. 115–127, November 1986.

[8] G. Cao and C. Bouman, "Covariance estimation for high dimensional data vectors using the sparse matrix transform," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 225–232. [Online]. Available: http://papers.nips.cc/paper/3409-covariance-estimation-for-high-dimensional-data-vectors-using-the-sparse-matrix-transform.pdf

[9] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008.

[10] official website of the MUSE Project, "http://muse.univ-lyon1.fr/."

[11] W. B. Wu and M. Pourahmadi, "Nonparametric estimation of large covariance matrices of longitudinal data," *Biometrika*, vol. 90, no. 4, pp. 831–844, 2003. [Online]. Available: http://biomet.oxfordjournals.org/content/90/4/831.abstract

[12] C. Z. P. Gong, J. Ye, "Multi-stage multi-task feature learning," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1988–1996.

[13] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye, "A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems." in *ICML (2)*, 2013, pp. 37–45.