Robust Shrinkage Estimation of High-dimensional Covariance Matrices

Yilun Chen, *Student Member, IEEE,* Ami Wiesel, *Member, IEEE,* and Alfred O. Hero III, *Fellow, IEEE*

***Abstract*—We address high dimensional covariance estimation for elliptical distributed samples, which are also known as spherically invariant random vectors (SIRV) or compound-Gaussian processes. Specifically we consider shrinkage methods that are suitable for high dimensional problems with a small number of samples (large p small n). We start from a classical robust covariance estimator [Tyler(1987)], which is distribution-free within the family of elliptical distribution but inapplicable when $n < p$. Using a shrinkage coefficient, we regularize Tyler's fixed point iterations. We prove that, for all n and p, the proposed fixed point iterations converge to a unique limit regardless of the initial condition. Next, we propose a simple, closed-form and data dependent choice for the shrinkage coefficient, which is based on a minimum mean squared error framework. Simulations demonstrate that the proposed method achieves low estimation error and is robust to heavy-tailed samples. Finally, as a real world application we demonstrate the performance of the proposed technique in the context of activity/intrusion detection using a wireless sensor network.**

***Index Terms*—Covariance estimation, large p small n, shrinkage methods, robust estimation, elliptical distribution, activity/intrusion detection, wireless sensor network**

I. Introduction

Estimating a covariance matrix (or a dispersion matrix) is a fundamental problem in statistical signal processing. Many techniques for detection and estimation rely on accurate estimation of the true covariance. In recent years, estimating a high dimensional $p \times p$ covariance matrix under small sample size n has attracted considerable attention. In these "large p small n" problems, the classical sample covariance suffers from a systematically distorted eigen-structure [6], and improved estimators are required.

Much effort has been devoted to high-dimensional covariance estimation, which use Steinian shrinkage [1]–[3] or other types of regularized methods such as [4], [5]. However, most of the high-dimensional estimators assume Gaussian distributed samples. This limits their usage in many important applications involving non-Gaussian and heavy-tailed samples. One exception is the Ledoit-Wolf estimator [2], where the authors shrink the sample covariance towards a scaled identity matrix and proposed a shrinkage coefficient which is asymptotically optimal for any distribution. On the other hand, traditional robust covariance estimators [8]–[10] designed for non-Gaussian samples generally require $n \gg p$ and are not suitable for "large p small n" problems. Therefore, the goal of our work is to develop a covariance estimator for problems that are both high dimensional and non-Gaussian. In this paper, we model the samples using the elliptical distribution [7], which is also referred to as the spherically invariant random vector model (SIRV) [26], [27] or the compound-Gaussian process model [13]. As a flexible and popular alternative, the elliptical family encompasses a large number of important distributions such as Gaussian distribution, the multivariate Cauchy distribution, the multivariate exponential distribution, the multivariate Student-T distribution, the K-distribution and the Weibull distribution. The capability of modelling heavy-tails makes the elliptical distribution appealing in signal processing and related fields. Typical applications include radar detection [13], [17], [20], [22], speech signal processing [23], remote sensing [24], wireless fading channels modelling [27], financial engineering [25] and so forth.

A well-studied covariance estimator for elliptical distributions is the ML estimator based on normalized samples [9], [14], [16]. The estimator is derived as the solution to a fixed point equation by using fixed point iterations. It is distribution-free within the class of elliptical distributions and its performance advantages are well known in the $n \gg p$ regime. However, it is not suitable for the "large p small n" setting. Indeed, when $n < p$, the ML estimator as defined does not even exist. To avoid this problem the authors of [21] propose an iterative regularized ML estimator that employs diagonal loading and uses a heuristic procedure for selecting the regularization parameter. While they did not establish convergence and uniqueness [21] they empirically demonstrated that their algorithm has superior performance in the context of a radar application. Our approach is similar to [21] but is conceived in a Steinian shrinkage framework, where we establish convergence and uniqueness of the resultant iterative estimator. We also propose a general procedure of selecting the shrinkage coefficient for heavy-tailed homogeneous samples. For a fixed shrinkage coefficient, we prove that the regularized fixed iterations converge to a unique solution for all n and p, regardless of the initial condition. Next, following Ledoit-Wolf [2], we provide a simple closed-form expression for the shrinkage coefficient, based on minimizing mean-squared-error. The resultant coefficient is a function of the unknown true covariance and cannot be implemented in practice. Instead, we develop a data-dependent "plug-

Y. Chen and A. O. Hero are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA. Tel: 1-734-763-0564. Fax: 1-734-763-8041. Emails: {yilun, hero}@umich.edu.

A. Wiesel is with the The Rachel and Selim Benin School of Computer Science and Engineering at the Hebrew University of Jerusalem, 91904 Jerusalem, Israel. Tel: 972-2-6584933. Email: ami.wiesel@huji.ac.il.

This work was partially supported by AFOSR, grant number FA9550-06-1-0324. The work of A. Wiesel was supported by a Marie Curie Outgoing International Fellowship within the 7th European Community Framework Programme. Parts of this work were presented at the IEEE Workshop on Sensor Array and Multichannel Signal Processing (SAM), Jerusalem, Israel, Oct. 2010.

in" estimator approximation. Simulation results demonstrate that our estimator achieves superior performance for samples distributed within the elliptical family. Furthermore, for the case that the samples are truly Gaussian, we report very little performance degradation with respect to the shrinkage estimators designed specifically for Gaussian samples [3].

As a real world application we demonstrate the proposed estimator for activity/intrusion detection using an active wireless sensor network. We show that the measured data exhibit strong non-Gaussian behavior and demonstrate significant performance advantages of the proposed robust covariance estimator when used in a covariance-based anomaly detection algorithm.

The paper is organized as follows. Section II provides a brief review of elliptical distributions and of Tyler's covariance estimation method. The regularized covariance estimator is introduced and derived in Section III. We provide simulations and experimental results in Section IV and Section V, respectively. Section VI summarizes our principal conclusions. The proof of theorems and lemmas are provided in the Appendix.

Notations: In the following, we depict vectors in lowercase boldface letters and matrices in uppercase boldface letters. $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and conjugate transpose operator, respectively. $\mathrm{Tr}(\cdot)$ and $\det(\cdot)$ are the trace and the determinant of a matrix, respectively.

II. ML covariance estimation for elliptical distributions

A. Elliptical distribution

Let $\mathbf{x}$ be a $p \times 1$ real random vector generated by the following model

$$\mathbf{x} = \nu \mathbf{u}, \tag{1}$$

where ν is a real, positive random variable and $\mathbf{u}$ is a $p \times 1$ zero-mean, real Gaussian random vector with positive definite covariance $\boldsymbol{\Sigma}$. We assume that ν and $\mathbf{u}$ are statistically independent. The resulting random vector $\mathbf{x}$ is elliptically distributed and its probability density function (pdf) can be expressed by

$$p(\mathbf{x}) = \phi\left(\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}\right), \tag{2}$$

where $\phi(\cdot)$ is the characteristic function (Definition 2, pp. 5, [25]) related to the pdf of ν. The elliptical family encompasses many useful distributions in signal processing and related fields and includes: the Gaussian distribution itself, the K distribution, the Weibull distribution and many others. As stated above, elliptically distributed samples are also referred to as Spherically Invariant Random Vectors (SIRV) or compound Gaussian processes in signal processing.

B. ML estimation

Let $\{\mathbf{x}_i\}_{i=1}^n$ be a set of n independent and identically distributed (i.i.d.) samples drawn according to (1). As the covariance of $\mathbf{x}$ may not exist, our problem is formulated to estimate the covariance (dispersion) matrix $\boldsymbol{\Sigma}$ of $\mathbf{u}$ from $\{\mathbf{x}_i\}_{i=1}^n$. The model (1) is invariant to scaling of the covariance matrix $\boldsymbol{\Sigma}$. Therefore, without loss of generality, we assume that the covariance matrix is trace-normalized in the sense that $\mathrm{Tr}(\boldsymbol{\Sigma}) = p$.

The commonly used sample covariance, defined as

$$\widehat{\mathbf{S}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T, \tag{3}$$

is known to be a poor estimator of $\boldsymbol{\Sigma}$, especially when the samples are high-dimensional (large p) and/or heavy-tailed. Tyler's method [9] addresses this problem by working with the normalized samples:

$$\mathbf{s}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2} = \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|_2}, \tag{4}$$

for which the term ν in (1) drops out. The pdf of $\mathbf{s}_i$ is given by [25]

$$p(\mathbf{s}_i; \boldsymbol{\Sigma}) = \frac{\Gamma(p/2)}{2\pi^{p/2}} \cdot \sqrt{\det(\boldsymbol{\Sigma}^{-1})} \cdot \left(\mathbf{s}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{s}_i\right)^{-p/2}. \tag{5}$$

Taking the derivative and equating to zero, the maximum likelihood estimator based on $\{\mathbf{s}_i\}_{i=1}^n$ is the solution to

$$\boldsymbol{\Sigma} = \frac{p}{n} \cdot \sum_{i=1}^n \frac{\mathbf{s}_i \mathbf{s}_i^T}{\mathbf{s}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{s}_i}. \tag{6}$$

When $n > p$, the ML estimator can be found using the following fixed point iterations:

$$\widehat{\boldsymbol{\Sigma}}_{j+1} = \frac{p}{n} \cdot \sum_{i=1}^n \frac{\mathbf{s}_i \mathbf{s}_i^T}{\mathbf{s}_i^T \widehat{\boldsymbol{\Sigma}}_j^{-1} \mathbf{s}_i}, \tag{7}$$

where the initial $\widehat{\boldsymbol{\Sigma}}_0$ is usually set to the identity matrix. Assuming that $n > p$ and that any p samples out of $\{\mathbf{s}_i\}_{i=1}^n$ are linearly independent with probability one, it can be shown that the iteration process in (7) converges and that the limiting value is unique up to constant scale, which does not depend on the initial value of $\widehat{\boldsymbol{\Sigma}}_0$. In practice, a final normalization step is needed, which ensures that the iteration limit $\widehat{\boldsymbol{\Sigma}}_\infty$ satisfies $\mathrm{Tr}(\widehat{\boldsymbol{\Sigma}}_\infty) = p$.

The ML estimate corresponds to the Huber-type M-estimator and has many good properties when $n \gg p$, such as asymptotic normality and strong consistency. Furthermore, it has been pointed out [9] that the ML estimate (7) is the "most robust" covariance estimator in the class of elliptical distributions in the sense of minimizing the maximum asymptotic variance. We note that (7) can be also motivated from other approaches as proposed in [14], [16].

III. Robust shrinkage covariance estimation

Here we extend Tyler's method to the high dimensional setting using shrinkage regularization. It is easy to see that there is no solution to (6) when $n < p$ (its left-hand-side is full rank whereas its right-hand-side of is rank deficient). This motivates us to develop a regularized covariance estimator for elliptical samples. Following [2], [3], we propose to regularize the fixed point iterations as

$$\widetilde{\boldsymbol{\Sigma}}_{j+1} = (1-\rho)\frac{p}{n} \sum_{i=1}^n \frac{\mathbf{s}_i \mathbf{s}_i^T}{\mathbf{s}_i^T \widehat{\boldsymbol{\Sigma}}_j^{-1} \mathbf{s}_i} + \rho \mathbf{I}, \tag{8}$$

$$\widehat{\boldsymbol{\Sigma}}_{j+1} = \frac{\widetilde{\boldsymbol{\Sigma}}_{j+1}}{\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_{j+1})/p}, \tag{9}$$

where ρ is the so-called shrinkage coefficient, which is a constant between 0 and 1. When $\rho = 0$ and $n > p$ the proposed shrinkage estimator reduces to Tyler's unbiased method in (6) and when $\rho = 1$ the estimator reduces to the trivial uncorrelated case yielding a scaled identity matrix. The term $\rho\mathbf{I}$ ensures that $\widehat{\mathbf{\Sigma}}_{j+1}$ is always well-conditioned and thus allows continuation of the iterative process without the need for restarts. Therefore, the proposed iteration can be applied to high dimensional estimation problems. We emphasize that the normalization (9) is important and necessary for convergence. We establish provable convergence and uniqueness of the limit in the following theorem.

Theorem 1. *Let $0 < \rho < 1$ be a shrinkage coefficient. Then, the fixed point iterations in (8) and (9) converge to a unique limit for any positive definite initial matrix $\widehat{\mathbf{\Sigma}}_0$.*

The proof of Theorem 1 follows directly from the concave Perron-Frobenius theory [28] and is provided in the Appendix. We note that the regularization presented in (8) and (9) is similar to diagonal loading [21]. However, unlike the diagonal loading approach of [21], the proposed shrinkage approach provides a systematic way to choose the regularization parameter ρ, discussed in the next section.

A. Choosing the shrinkage coefficient

We now turn to the problem of choosing a good, data-dependent, shrinkage coefficient ρ, as as an alternative to cross-validation schemes which incur intensive computational costs. As in Ledoit-Wolf [2], we begin by assuming we "know" the true covariance $\mathbf{\Sigma}$. Then we define the following clairvoyant "estimator":

$$\widetilde{\mathbf{\Sigma}}(\rho) = (1-\rho)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\mathbf{\Sigma}^{-1}\mathbf{s}_i} + \rho\mathbf{I}, \tag{10}$$

where the coefficient ρ is chosen to minimize the minimum mean-squared error:

$$\rho_O = \arg\min_{\rho} E\left\{\left\|\widetilde{\mathbf{\Sigma}}(\rho) - \mathbf{\Sigma}\right\|_F^2\right\}. \tag{11}$$

The following theorem shows that there is a closed-form solution to the problem (11), which we refer to as the "oracle" coefficient.

Theorem 2. *For i.i.d. elliptical distributed samples the solution to (11) is*

$$\rho_O = \frac{p^2 + (1-2/p)\mathrm{Tr}(\mathbf{\Sigma}^2)}{(p^2 - np - 2n) + (n+1+2(n-1)/p)\mathrm{Tr}(\mathbf{\Sigma}^2)}, \tag{12}$$

under the condition $\mathrm{Tr}(\mathbf{\Sigma}) = p$.

The proof of Theorem 2 requires the calculation of the fourth moments of an isotropically distributed random vector [30]–[32] and is provided in the Appendix.

The oracle coefficient cannot be implemented since ρ_O is a function of the unknown true covariance $\mathbf{\Sigma}$. Therefore, we propose a plug-in estimate for ρ_O:

$$\hat{\rho} = \frac{p^2 + (1-2/p)\mathrm{Tr}(\widehat{\mathbf{M}}^2)}{(p^2 - np - 2n) + (n+1+2(n-1)/p)\mathrm{Tr}(\widehat{\mathbf{M}}^2)}, \tag{13}$$

where $\widehat{\mathbf{M}}$ can be any consistent estimator of $\mathbf{\Sigma}$, *e.g.,* the trace-normalized Ledoit-Wolf estimator. Another appealing candidate for plug-in is the (trace-normalized) normalized sample covariance $\widehat{\mathbf{R}}$ [12] defined by:

$$\widehat{\mathbf{R}} = \frac{p}{n}\sum_{i=1}^{n}\mathbf{s}_i\mathbf{s}_i^T. \tag{14}$$

We note that the only requirement on the covariance estimator $\widehat{\mathbf{M}}$ is that it provide a good approximation to $\mathrm{Tr}(\mathbf{\Sigma}^2)$. It does not have to be well-conditioned nor does it have to be an accurate estimator of the true covariance matrix $\mathbf{\Sigma}$.

By using the plug-in estimate $\hat{\rho}$ in place of ρ, the robust shrinkage estimator is computed via the fixed point iteration in (8) and (9).

We note that our proposed minimum MSE based approach to estimate the shrinkage coefficient is completely general and makes a minimum of assumptions on the provenance of the data. In specific applications such as adaptive beamforming, a specifically tailored coefficient estimator may be advantageous [21].

B. Extension to the complex case

Here we consider the scenario where the random vector $\mathbf{x}$ in (1) is complex elliptical distributed. In this case, ν is still a real, positive random variable but $\mathbf{u}$ is a complex Gaussian random vector with covariance matrix $\mathbf{\Sigma}$. Note that $\mathbf{\Sigma}$ is Hermitian and is assumed to be trace-normalized, *i.e.,* $\mathrm{Tr}(\mathbf{\Sigma}) = p$. The complex version of our fixed point iterations is

$$\begin{aligned} \widetilde{\mathbf{\Sigma}}_{j+1} &= (1-\rho)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{s}_i\mathbf{s}_i^H}{\mathbf{s}_i^H\widehat{\mathbf{\Sigma}}_j^{-1}\mathbf{s}_i} + \rho\mathbf{I}, \\ \widehat{\mathbf{\Sigma}}_{j+1} &= \frac{\widetilde{\mathbf{\Sigma}}_{j+1}}{\mathrm{Tr}(\widetilde{\mathbf{\Sigma}}_{j+1})/p}, \end{aligned} \tag{15}$$

where $\mathbf{s}_i$ is defined by (4). As in the real case, the shrinkage coefficient ρ is chosen to minimize (11), where the complex clairvoyant estimator $\widetilde{\mathbf{\Sigma}}(\rho)$ is re-defined as

$$\widetilde{\mathbf{\Sigma}}(\rho) = (1-\rho)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{s}_i\mathbf{s}_i^H}{\mathbf{s}_i^H\mathbf{\Sigma}^{-1}\mathbf{s}_i} + \rho\mathbf{I}. \tag{16}$$

The following theorem extends Theorems 1 and 2 to the complex case.

Theorem 3. *For any $0 < \rho < 1$ the complex valued iterations in (15) converge to a unique limit for any positive definitive Hermitian matrix $\widehat{\mathbf{\Sigma}}_0$. For $\widetilde{\mathbf{\Sigma}}(\rho)$ defined in (16), the solution to (11) is*

$$\rho_O = \frac{p^2 - 1/p\mathrm{Tr}(\mathbf{\Sigma}\mathbf{\Sigma}^H)}{(p^2 - pn - n) + (n + (n-1)/p)\mathrm{Tr}(\mathbf{\Sigma}\mathbf{\Sigma}^H)} \tag{17}$$

under the condition $\mathrm{Tr}(\mathbf{\Sigma}) = p$.

The proof of Theorem 3 is similar to that of Theorem 1 and Theorem 2 and is provided in the Appendix. In practice, ρ_O can be approximated by plugging any consistent estimator of $\mathrm{Tr}(\mathbf{\Sigma}\mathbf{\Sigma}^H)$ in (17).

IV. NUMERICAL SIMULATION

In this section we use simulations to demonstrate the superior performance of the proposed shrinkage approach. First we show that our estimator outperforms other estimators for the case of heavy-tailed samples generated by a multivariate Student-T distribution, where ν in (1) is a function of a Chi-square random variable:

$$\nu = \sqrt{\frac{d}{\chi_d^2}}; \tag{18}$$

The degree-of-freedom d of this multivariate Student-T statistic is set to 3. The dimensionality p is chosen to be 100 and we let $\mathbf{\Sigma}$ be the covariance matrix of an AR(1) process,

$$\mathbf{\Sigma}(i,j) = r^{|i-j|}, \tag{19}$$

where $\mathbf{\Sigma}(i,j)$ denotes the entry of $\mathbf{\Sigma}$ in row i and column j. The parameter r is set to 0.7 in this simulation. The sample size n varies from 5 to 225 with step size 10. All the simulations are repeated for 100 trials and the average empirical performance is reported.

We use (13) with $\widehat{\mathbf{M}} = \widehat{\mathbf{R}}$ and employ iterations defined by (8) and (9) with $\rho = \hat{\rho}$. For comparison, we also plot the results of the trace-normalized oracle in (12), the trace-normalized Ledoit-Wolf estimator [2], and the non-regularized solution in (7) (when $n > p$). As the Ledoit-Wolf estimator operates on the sample covariance which is sensitive to outliers, we also compare to a trace-normalized version of a clairvoyant Ledoit-Wolf estimator implemented according to the procedure in [2] with known ν. More specifically, the samples $\mathbf{x}_i$ are firstly normalized by the known realizations ν_i, yielding truly Gaussian samples; then the sample covariance of the normalized $\mathbf{x}_i$'s is computed, which is used to estimate the Ledoit-Wolf shrinkage parameters and estimate the covariance via equation (14) in [2]. The MSE is plotted in Fig. 1 for the case that $r = 0.7$. It can be observed that the proposed method performs significantly better than the Ledoit-Wolf estimator in Fig. 1, and that the performance is very close to the ideal oracle estimator using the optimal shrinkage parameter (12). Even the clairvoyant Ledoit-Wolf implemented with known ν_i does not outperform the proposed estimator in the small sample size regime. These results demonstrate the robustness of the proposed approach. Although the Ledoit-Wolf estimator performs better when $r = 0$, the case where $\mathbf{\Sigma} = \mathbf{I}$, the proposed approach still significantly outperforms it, especially for small sample size n (results not shown).

As a graphical illustration, in Fig. 2 we provide covariance visualizations for a realization of the estimated covariances using the Ledoit-Wolf method and the proposed approach. The covariance matrix estimates are rendered as a heatmap in Fig. 2(a). The sample size in this example is set to 50, which is smaller than the dimension 100. Compared to the true covariance, it is clear that the proposed covariance estimator preserves the structure of the true covariance, while in the Ledoit-Wolf covariance procudure produces "block pattern" artifacts caused by heavy-tails of the multivariate Student-T.

When $n > p$, we also observe a substantial improvement of the proposed method over the ML covariance estimate, which provides further evidence of the power of Steinian shrinkage for reducing MSE.

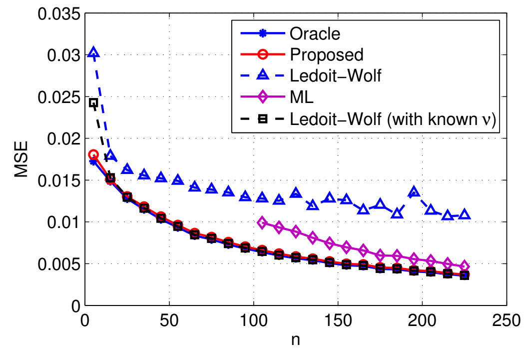
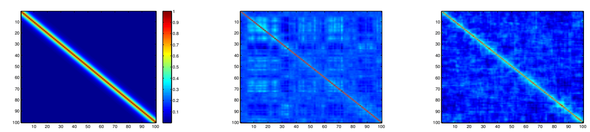


Fig. 1. Multivariate Student-T samples: Comparison of different trace-normalized covariance estimators when $p = 100$, where r is set to 0.7.



(a) True covariance (b) Ledoit-Wolf (c) Proposed

Fig. 2. Multivariate Student-T samples: Heatmap visualizations of the covariance matrix estimates using the Ledoit-Wolf and the proposed approaches. $p = 100, n = 50$. Note that $n < p$ in this case.

In order to assess the tradeoff between accuracy and robustness we investigate the case when the samples are truly Gaussian distributed. We use the same simulation parameters as in the previous example, the only difference being that the samples are now generated from a Gaussian distribution. The performance comparison is shown in Fig. 3, where four different (trace-normalized) methods are included: the oracle estimator derived from Gaussian assumptions (Gaussian oracle) [3], the iterative approximation of the Gaussian oracle (Gaussian OAS) proposed in [3], the Ledoit-Wolf estimator and the proposed method. It can be seen that for truly Gaussian samples the proposed method performs very closely to the Gaussian OAS, which is specifically designed for Gaussian distributions. Indeed, for small sample size ($n < 20$), the proposed method performs even better than the Ledoit-Wolf estimator. This indicates that, although the proposed robust method is developed for the entire elliptical family, it actually sacrifices very little performance for the case that the distribution is Gaussian.

V. APPLICATION TO ANOMALY DETECTION IN WIRELESS SENSOR NETWORKS

In this section we demonstrate the proposed robust covariance estimator in a real application: activity detection using a wireless sensor network.

The experiment was set up on an Mica2 sensor network platform, as shown in Fig. 4, which consists of 14 sensor nodes randomly deployed inside and outside a laboratory at the University of Michigan. Wireless sensors communicated

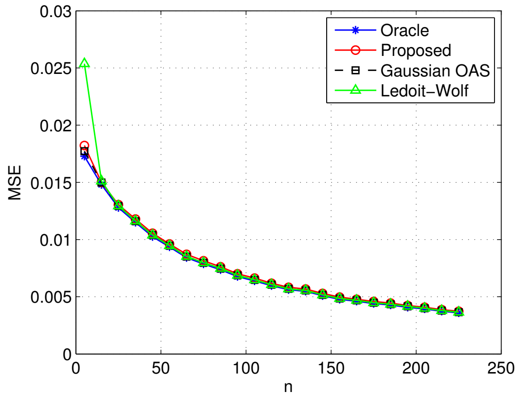


Fig. 3. Gaussian samples: Comparison of trace-normalized different covariance estimators when $p = 100$.

with each other asynchronously by broadcasting an RF signal every 0.5 seconds. The received signal strength (RSS), defined as the voltage measured by a receiver's received signal strength indicator circuit (RSSI), was recorded for each pair of transmitting and receiving nodes. There were $14 \times 13 = 182$ pairs of RSSI measurements over a 30 minute period, and samples were acquired every 0.5 sec. During the experiment period, persons walked into and out of the lab at random times, causing anomaly patterns in the RSSI measurements. Finally, for ground truth, a web camera was employed to record the actual activity.

Fig. 4. Experimental platform: wireless Mica2 sensor nodes.

Fig. 5 shows all the received signals and the ground truth indicator extracted from the video. The objective of this experiment was intrusion (anomaly) detection. We emphasize that, with the exception of the results shown in Fig. 10, the ground truth indicator is only used for performance evaluation and the detection algorithms presented here were completely *unsupervised*.

To remove temperature drifts [36] of receivers we detrended the data as follows. Let $x_i[k]$ be the k-th sample of the i-th RSS signal and denote

$$\mathbf{x}[k] = \left(x_1[k], x_2[k], \dots, x_{182}[k]\right)^T. \tag{20}$$

The local mean value of $\mathbf{x}[k]$ is defined by

$$\bar{\mathbf{x}}[k] = \frac{1}{2m+1} \sum_{i=k-m}^{k+m} \mathbf{x}[k], \tag{21}$$

where the integer m determines local window size and is set to 50 in this study. We detrend the data by subtracting this local mean

$$\mathbf{y}[k] = \mathbf{x}[k] - \bar{\mathbf{x}}[k], \tag{22}$$

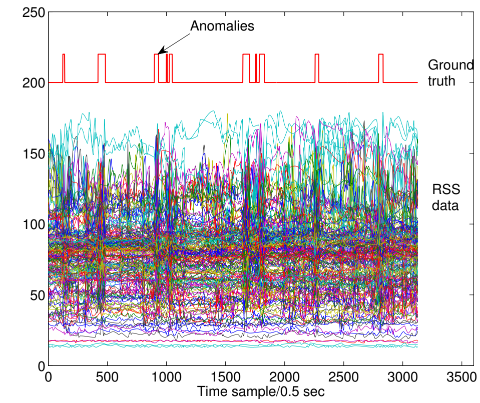


Fig. 5. At bottom 182 RSS sequences sampled from each pair of transmitting and receiving nodes in intrusion detection experiment. Ground truth indicators at top are extracted from video captured from a web camera that recorded the scene.

yielding a detrended sample $\mathbf{y}[k]$ used in our anomaly detection.

We established that the detrended measurements were heavy-tailed non-Gaussian by performing several statistical tests. First the Lilliefors test [37] of Gaussianity was performed on the detrended RSS measurements. This resulted in rejection of the Gaussian hypothesis at a level of significance of 10^{-6}. As visual evidence, we show the quantile-quantile plot (QQ plot) for one of the detrended RSS sequences in Fig. 6 which illustrates that the samples are non-Gaussian. In Fig. 7, we plot the histograms and scatter plots of two of the detrended RSS sequences, which shows the heavy-tail nature of the sample distribution. This strongly suggests that the RSS samples can be better described by a heavy-tailed elliptical distribution than by a Gaussian distribution. As additional evidence, we fitted a Student-T distribution to the first detrended RSS sequence, and used maximum likelihood to estimate the degree-of-freedom as $d = 2$ with a 95% confidence interval (CI) $[1.8460, 2.2879]$.

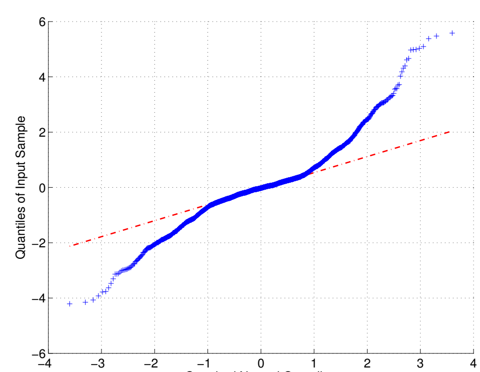


Fig. 6. QQ plot of data versus the standard Gaussian distribution.

Consider the following discriminant based on the detrended data:

$$t_k = \mathbf{s}^T[k] \mathbf{\Sigma}^{-1} \mathbf{s}[k], \tag{23}$$

where, similarly to (4), $\mathbf{s}[k] = \mathbf{y}[k] / \|\mathbf{y}[k]\|_2$ and $\mathbf{\Sigma}$ is given by

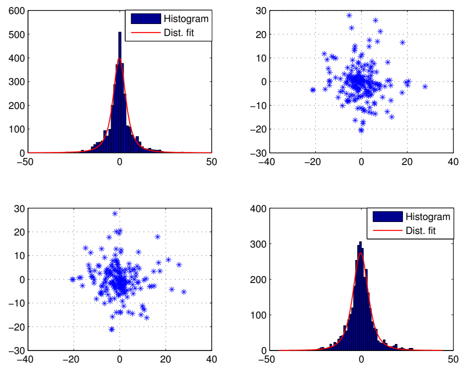


Fig. 7. Histograms and scatter plots of the first two de-trended RSS sequences, which are fit by a multivariate Student-T distribution with degree-of-freedom $d = 2$.

the solution to (6). A time sample is declared to be anomalous if the test statistic t_k exceeds a specified threshold. The statistic (23) is equivalent to a robustified version of the Mahalanobis distance anomaly detector [33]. Note that direct application of shrinkage to the sample covariance of $\mathbf{y}[k]$ would be problematic since a multivariate Student-T vector with 2 degrees of freedom has no second order moments. The test statistic (23) can be interpreted as a shrinkage robustified Mahalanobis distance test applied to the better behaved variable $\mathbf{s}[k]$ that has finite moments of all orders. Specifically, we constructed the 182×182 sample covariance by randomly subsampling 200 time slices from the RSS data shown in Fig. 5. Note, that these 200 samples correspond to a training set that is contaminated by anomalies at the same anomaly rate (approximately 10%) as the entire sample set. The detection performance was evaluated using the receiver operating characteristic (ROC) curve, where the averaged curves from 200 independent Monte-Carlo trials are shown in Fig. 8. For comparison, we also implemented the activity detector (23) with other covariance estimates including: the sample covariance, the Ledoit-Wolf estimator and Tyler's ML estimator.

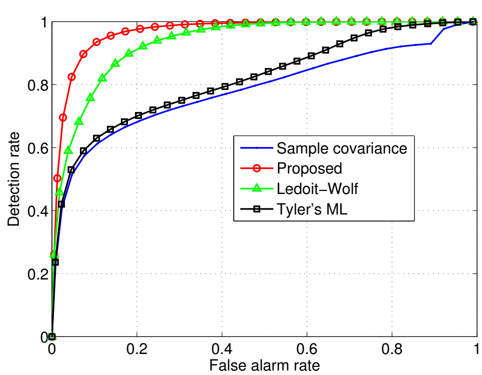


Fig. 8. Performance comparison for different covariance estimators, $p = 182, n = 200$.

From the mean ROCs we can see that the detection performances are rank ordered as follows: Proposed > Ledoit-Wolf > Tyler's ML > Sample covariance. The sample covariance performs poorly in this setting due to the small sample size ($n = 200, p = 182$) and its sensitivity to the heavy-tailed distribution shown in Fig. 6 and 7. The Tyler ML method and the Ledoit-Wolf estimator improve upon the sample covariance since they compensate for heavy tails and for small sample size, respectively. Our proposed method compensates for both effects simultaneously and achieves the best detection performance.

We also plot the 90% confidence envelopes, determined by cross-validation, on the ROCs in Fig. 9. The width of the confidence interval reflects the sensitivity of the anomaly detector to variations in the training set. Indeed, the upper and lower endpoints of the confidence interval are the optimistic and the pessimistic predictions of detection performance. The proposed method achieves the smallest width among the four computed 90% confidence envelopes.

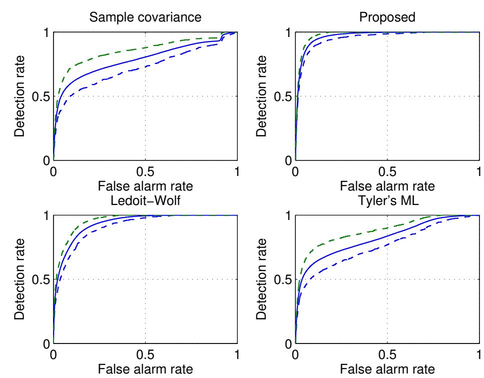


Fig. 9. Performance comparison for different covariance estimators, including the mean value and 90% confidence intervals. (a): Sample covariance. (b): Proposed. (c): Ledoit-Wolf. (d): Tyler's ML. The 200 training samples are randomly selected from the entire data set.

Finally, for completeness we provide performance comparison of covariance-based *supervised* activity detection algorithms in Fig. 10. The training period is selected to be $[251, 450]$ based on ground truth where no anomalies appear. It can be observed that, by excluding the outliers caused by anomalies, the performance of the Ledoit-Wolf based intrusion detection algorithm is close to that of the proposed method. We conclude that the activity detection performance of the proposed covariance estimator is more robust than the other three estimators with respect to outlier contamination in the training samples.

VI. Conclusion

In this paper, we proposed a shrinkage covariance estimator which is robust over the class of elliptically distributed samples. The proposed estimator is obtained by fixed point iterations, and we established theoretical guarantees for existence, convergence and uniqueness. The optimal shrinkage coefficient was derived using a minimum mean-squared-error framework and has a closed-form expression in terms of the unknown true covariance. This expression can be well approximated by a simple plug-in estimator. Simulations suggest that the iterative approach converges to a limit which is robust to heavy-tailed multivariate Student-T samples. Furthermore, we show that

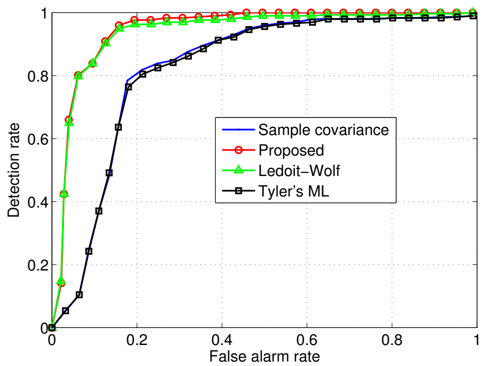


Fig. 10. Performance comparison for different covariance estimators, $p = 182, n = 200$. The covariance matrix is estimated in a *supervised* manner.

for the Gaussian case, the proposed estimator performs very closely to previous estimators designed expressly for Gaussian samples.

As a real world application we demonstrated the performance of the proposed estimator in intrusion detection using a wireless sensor network. Implementation of a standard covariance-based detection algorithm using our robust covariance estimator achieved superior performances as compared to conventional covariance estimators.

The basis of the proposed method is the ML estimator originally proposed by Tyler in [9]. However, the approach presented in this paper can be extended to other M-estimators.

One of the main contributions of our work is the proof of uniqueness and convergence of the estimator. This proof extends the results of [9], [18] to the regularized case. Recently, an alternative proof to the non-regularized case using convexity on manifolds was presented in [35]. This latter proof highlights the geometrical structure of the problem and gives additional insight. We are currently investigating its extension to the regularized case.

VII. Acknowledgement

The authors would like to thank Neal Patwari for designing the intrusion detection experiment and collecting the data. We also thank Harry Schmidt and his group at Raytheon, Tucson AZ, for lending us the Mica2 motes for this experiment. We would like to thank Prof. Drton for fruitful discussions and a preprint of [38].

VIII. Appendix

A. Proof of Theorem 1

In this appendix we prove Theorem 1. The original convergence proof for the non-regularized case in [9], [18] is based on careful exploitation of the specific form of (6). In the contrast, our proof for the regularized case is based on a direct connection from concave Perron-Frobenius theory [28], [29]. We begin by summarizing the required concave Perron-Frobenius result in the following lemma.

Lemma 1 ([28]). *Let $(E, \|\cdot\|)$ be a Banach space with $K \subset E$ being a closed, convex cone on which $\|\cdot\|$ is increasing, i.e., for which $x \leq y$ implies $\|x\| \leq \|y\|$, where the operator $\leq$ on the convex cone K means that if $x \leq y$ then $y - x \in K$. Define $U = \{x | x \in K, ||x|| = 1\}$. Let $T : K \to K$ be a concave operator such that*

$$T(\mu x + (1-\mu)y) \geq \mu T(x) + (1-\mu)T(y), \quad \text{for all } \mu \in [0,1], \quad \text{all } x, y \in K. \tag{24}$$

If for some $e \in K - \{0\}$ and constants $a > 0$, $b > 0$ there is

$$ae \leq T(x) \leq be, \quad \text{for all } x \in U, \tag{25}$$

then there exists a unique $x^* \in U$ to which the iteration of the normalized operator $\tilde{T}(x) = T(x)/\|T(x)\|, x \in K - \{0\}$ converges:

$$\lim_{k \to \infty} \tilde{T}^k(x) = x^*, \quad \text{for all } x \in K - \{0\}. \tag{26}$$

Lemma 1 can be obtained by combining results from Lemma 2 and Theorem in Section 4 of [28]. Here we show that the proof of Theorem 1 is a direct result of applying Lemma 1 with proper definitions of E, K, U and T:

- E: the set of all symmetric matrices;
- K: the set of all positive semi-definite matrices on E;
- $\|\mathbf{\Sigma}\|$: the normalized nuclear norm of $\mathbf{\Sigma}$, *i.e.,*

$$\|\mathbf{\Sigma}\| = \frac{1}{p}\sum_{j=1}^{p} |\lambda_j|, \tag{27}$$

  where λ_j is the j-th eigenvalue of $\mathbf{\Sigma}$ and $|\cdot|$ is the absolute value operator. Note that for any $\mathbf{\Sigma} \in K$, the nuclear norm $\|\cdot\|$ is identical to $\mathrm{Tr}(\cdot)/p$ and is increasing;
- U: the set $U = \{\mathbf{\Sigma} | \mathbf{\Sigma} \in K, \|\mathbf{\Sigma}\| = 1\}$;
- T: the mapping from K to K defined by

$$T(\mathbf{\Sigma}) = (1-\rho)\frac{p}{n}\sum_{i=1}^{n} w(\mathbf{s}_i, \mathbf{\Sigma})\mathbf{s}_i\mathbf{s}_i^T + \rho\mathbf{I}, \tag{28}$$

  where the weight function $w(\mathbf{s}_i, \mathbf{\Sigma})$ is defined as

$$w(\mathbf{s}_i, \mathbf{\Sigma}) = \inf_{\mathbf{z}^T\mathbf{s}_i \neq 0} \frac{\mathbf{z}^T\mathbf{\Sigma}\mathbf{z}}{(\mathbf{s}_i^T\mathbf{z})^2}, \tag{29}$$

  for any $\mathbf{\Sigma} \in K$.

Proof: With the above definitions we show that Theorem 1 is a direct result of Lemma 1. We begin by showing that the mapping operator T is concave. Indeed, it is sufficient to show that $w(\mathbf{s}_i, \mathbf{\Sigma})$ in concave in $\mathbf{\Sigma}$, which is true because it is the infinimum of affine functions of $\mathbf{\Sigma}$.

Next, we show that T satisfies condition (25) with $e = \mathbf{I}$. It is easy to see that

$$\rho\mathbf{I} \leq T(\mathbf{\Sigma}), \tag{30}$$

for any $\mathbf{\Sigma} \in U$. Then we show that

$$w(\mathbf{s}_i, \mathbf{\Sigma}) \leq p, \tag{31}$$

for any $\mathbf{\Sigma} \in U$. Indeed,

$$w(\mathbf{s}_i, \mathbf{\Sigma}) = \inf_{\mathbf{z}^T\mathbf{s}_i \neq 0} \frac{\mathbf{z}^T\mathbf{\Sigma}\mathbf{z}}{(\mathbf{s}_i^T\mathbf{z})^2} \leq \frac{\mathbf{s}_i^T\mathbf{\Sigma}\mathbf{s}_i}{(\mathbf{s}_i^T\mathbf{s}_i)^2} \leq \frac{\lambda_{\max}}{\mathbf{s}_i^T\mathbf{s}_i} = \lambda_{\max}, \tag{32}$$

where $\lambda_{\max}$ is the maximum eigenvalue of $\mathbf{\Sigma}$. The last equality in the right-hand-side of (32) comes from the fact that $\mathbf{s}_i$ is of unit norm by definition (4). (31) is thus obtained by noticing that $\mathbf{\Sigma} \in U$ and $\lambda_{\max} \le p$. Substituting (31) into (28) we have

$$T(\mathbf{\Sigma}) \le (1-\rho)p^2\widehat{\mathbf{R}} + \rho\mathbf{I} \le \left((1-\rho)p^2\alpha_{\max} + \rho\right)\mathbf{I}, \quad (33)$$

where

$$\widehat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{s}_i\mathbf{s}_i^T,$$

and $\alpha_{\max}$ is the maximum eigenvalue of $\widehat{\mathbf{R}}$. Again, as $\mathbf{s}_i$ is of unit norm, $\alpha_{\max} \le \mathrm{Tr}(\widehat{\mathbf{R}}) = 1$ and

$$T(\mathbf{\Sigma}) \le \left((1-\rho)p^2 + \rho\right)\mathbf{I}. \quad (34)$$

Therefore, we have shown that T satisfies condition (25), where $e = \mathbf{I}$, $a = \rho$ and $b = (1-\rho)p^2 + \rho$. In addition, (25) establishes that the mapping T from U always yields a positive definite matrix. Therefore, the convergent limit of the fixed-point iteration is positive definite.

Finally, we note that, for any $\mathbf{\Sigma} \succ \mathbf{0}$, we have

$$\|\mathbf{\Sigma}\| = \frac{\mathrm{Tr}(\mathbf{\Sigma})}{p}, \quad (35)$$

and

$$w(\mathbf{s}_i, \mathbf{\Sigma}) = \inf_{\mathbf{z}^T\mathbf{s}_i \neq 0} \frac{\mathbf{z}^T\mathbf{\Sigma}\mathbf{z}}{(\mathbf{s}_i^T\mathbf{z})^2} = \frac{1}{\mathbf{s}_i^T\mathbf{\Sigma}^{-1}\mathbf{s}_i}. \quad (36)$$

The limit (26) is then identical to the limit of proposed iterations (8) and (9) for any $\mathbf{\Sigma} \succ \mathbf{0}$. Therefore, Theorem 1 has been proved.

B. Proof of Theorem 2

Proof: To ease the notation we define $\widetilde{\mathbf{C}}$ as

$$\widetilde{\mathbf{C}} = \frac{p}{n}\sum_{i=1}^{n} \frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\mathbf{\Sigma}^{-1}\mathbf{s}_i}. \quad (37)$$

The shrinkage estimator in (10) is then

$$\widetilde{\mathbf{\Sigma}}(\rho) = (1-\rho)\widetilde{\mathbf{C}} + \rho\mathbf{I}. \quad (38)$$

By substituting (38) into (10) and taking derivatives of ρ, we obtain that

$$\begin{aligned}\rho_O &= \frac{E\left\{\mathrm{Tr}\left((\mathbf{I}-\widetilde{\mathbf{C}})(\mathbf{\Sigma}-\widetilde{\mathbf{C}})\right)\right\}}{E\left\{\left\|\mathbf{I}-\widetilde{\mathbf{C}}\right\|_F^2\right\}} \\ &= \frac{m_2 - m_{11} - m_{12} + \mathrm{Tr}(\mathbf{\Sigma})}{m_2 - 2m_{11} + p},\end{aligned} \quad (39)$$

where

$$m_2 = E\left\{\mathrm{Tr}(\widetilde{\mathbf{C}}^2)\right\}, \quad (40)$$

$$m_{11} = E\left\{\mathrm{Tr}(\widetilde{\mathbf{C}})\right\}, \quad (41)$$

and

$$m_{12} = E\left\{\mathrm{Tr}(\widetilde{\mathbf{C}}\mathbf{\Sigma})\right\}. \quad (42)$$

Next, we calculate the moments. We begin by eigen-decomposing $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^T, \quad (43)$$

and denote

$$\mathbf{\Lambda} = \mathbf{U}\mathbf{D}^{1/2}. \quad (44)$$

Then, we define

$$\mathbf{z}_i = \frac{\mathbf{\Lambda}^{-1}\mathbf{s}_i}{\|\mathbf{\Lambda}^{-1}\mathbf{s}_i\|_2} = \frac{\mathbf{\Lambda}^{-1}\mathbf{u}_i}{\|\mathbf{\Lambda}^{-1}\mathbf{u}_i\|_2}. \quad (45)$$

Noting that $\mathbf{u}_i$ is a Gaussian distributed random vector with covariance $\mathbf{\Sigma}$, it is easy to see that $\|\mathbf{z}_i\|_2 = 1$ and $\mathbf{z}_i$ and $\mathbf{z}_j$ are independent with each other for $i \neq j$. Furthermore, $\mathbf{z}_i$ is isotropically distributed [30]–[32] and satisfies [3]

$$E\left\{\mathbf{z}_i\mathbf{z}_i^T\right\} = \frac{1}{p}\mathbf{I}, \quad (46)$$

$$\begin{aligned}E\left\{\left(\mathbf{z}_i^T\mathbf{D}\mathbf{z}_i\right)^2\right\} &= \frac{1}{p(p+2)}\left(2\mathrm{Tr}(\mathbf{D}^2) + \mathrm{Tr}^2(\mathbf{D})\right) \\ &= \frac{1}{p(p+2)}\left(2\mathrm{Tr}(\mathbf{\Sigma}^2) + \mathrm{Tr}^2(\mathbf{\Sigma})\right),\end{aligned} \quad (47)$$

and

$$E\left\{\left(\mathbf{z}_i^T\mathbf{D}\mathbf{z}_j\right)^2\right\} = \frac{1}{p^2}\mathrm{Tr}(\mathbf{D}^2) = \frac{1}{p^2}\mathrm{Tr}(\mathbf{\Sigma}^2), \ i \neq j. \quad (48)$$

Expressing $\widetilde{\mathbf{C}}$ in terms of $\mathbf{z}_i$, there is

$$\widetilde{\mathbf{C}} = \frac{p}{n}\mathbf{\Lambda}\sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i^T\mathbf{\Lambda}^T. \quad (49)$$

Then,

$$E\left\{\widetilde{\mathbf{C}}\right\} = \frac{p}{n}\mathbf{\Lambda}\sum_{i=1}^{n} E\left\{\mathbf{z}_i\mathbf{z}_i^T\right\}\mathbf{\Lambda}^T = \mathbf{\Sigma}, \quad (50)$$

and accordingly we have

$$m_{11} = E\left\{\mathrm{Tr}(\widetilde{\mathbf{C}})\right\} = \mathrm{Tr}(\mathbf{\Sigma}), \quad (51)$$

and

$$m_{12} = E\left\{\mathrm{Tr}(\widetilde{\mathbf{C}}\mathbf{\Sigma})\right\} = \mathrm{Tr}(\mathbf{\Sigma}^2). \quad (52)$$

For m_2 there is

$$\begin{aligned}m_2 &= \frac{p^2}{n^2}E\left\{\mathrm{Tr}\left(\mathbf{\Lambda}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i^T\mathbf{\Lambda}^T\mathbf{\Lambda}\sum_{j=1}^{n}\mathbf{z}_j\mathbf{z}_j^T\mathbf{\Lambda}^T\right)\right\} \\ &= \frac{p^2}{n^2}E\left\{\mathrm{Tr}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{z}_i\mathbf{z}_i^T\mathbf{\Lambda}^T\mathbf{\Lambda}\mathbf{z}_j\mathbf{z}_j^T\mathbf{\Lambda}^T\mathbf{\Lambda}\right)\right\} \\ &= \frac{p^2}{n^2}E\left\{\mathrm{Tr}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{z}_i\mathbf{z}_i^T\mathbf{D}\mathbf{z}_j\mathbf{z}_j^T\mathbf{D}\right)\right\} \\ &= \frac{p^2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left\{\left(\mathbf{z}_i^T\mathbf{D}\mathbf{z}_j\right)^2\right\}.\end{aligned} \quad (53)$$

Now substitute (47) and (48) to (53):

$$\begin{aligned}m_2 &= \frac{p^2}{n^2}\left(\frac{n}{p(p+2)}\left(2\mathrm{Tr}(\mathbf{\Sigma}^2) + \mathrm{Tr}^2(\mathbf{\Sigma})\right) + \frac{n(n-1)}{p^2}\mathrm{Tr}(\mathbf{\Sigma}^2)\right) \\ &= \frac{1}{n(1+2/p)}\left(2\mathrm{Tr}(\mathbf{\Sigma}^2) + \mathrm{Tr}^2(\mathbf{\Sigma})\right) + (1-\frac{1}{n})\mathrm{Tr}(\mathbf{\Sigma}^2) \\ &= \left(1 - \frac{1}{n} + \frac{2}{n(1+2/p)}\right)\mathrm{Tr}(\mathbf{\Sigma}^2) + \frac{\mathrm{Tr}^2(\mathbf{\Sigma})}{n(1+2/p)}.\end{aligned} \quad (54)$$

Recalling $\mathrm{Tr}(\mathbf{\Sigma}) = p$, (12) is finally obtained by substituting (51), (52) and (54) into (39).

C. Proof of Theorem 3

The proof of convergence and uniqueness is a simple extension of the proof in Appendix A by re-defining E as the set of all Hermitian matrices and the mapping function $T(\mathbf{\Sigma})$ as

$$T(\mathbf{\Sigma}) = (1-\rho)\frac{p}{n}\sum_{i=1}^{n} w(\mathbf{s}_i, \mathbf{\Sigma})\mathbf{s}_i\mathbf{s}_i^H + \rho\mathbf{I}, \tag{55}$$

where

$$w(\mathbf{s}_i, \mathbf{\Sigma}) = \inf_{\mathbf{z}^H\mathbf{s}_i \neq 0} \frac{\mathbf{z}^H\mathbf{\Sigma}\mathbf{z}}{|\mathbf{s}_i^H\mathbf{z}|^2}. \tag{56}$$

For the oracle coefficient, define

$$\widetilde{\mathbf{C}} = \frac{n}{p}\sum_{i=1}^{n}\frac{\mathbf{s}_i\mathbf{s}_i^H}{\mathbf{s}_i^H\mathbf{\Sigma}^{-1}\mathbf{s}_i}. \tag{57}$$

Then $\widetilde{\mathbf{\Sigma}}(\rho) = (1-\rho)\widetilde{\mathbf{C}} + \rho\mathbf{I}$. It can be shown that

$$\begin{aligned}\rho_O &= \frac{E\left\{\mathrm{Re}\left(\mathrm{Tr}\left((\mathbf{I}-\widetilde{\mathbf{C}})(\mathbf{\Sigma}-\widetilde{\mathbf{C}})^H\right)\right)\right\}}{E\left\{\|\mathbf{I}-\widetilde{\mathbf{C}}\|_F^2\right\}} \\ &= \frac{m_2 - m_{11} - m_{12} + \mathrm{Tr}(\mathbf{\Sigma}^H)}{m_2 - 2m_{11} + p},\end{aligned} \tag{58}$$

where m_2, m_{11} and m_{12} are re-defined as

$$m_2 = E\left\{\mathrm{Tr}\left(\widetilde{\mathbf{C}}\widetilde{\mathbf{C}}^H\right)\right\}, \tag{59}$$

$$m_{11} = E\left\{\mathrm{Re}\left(\mathrm{Tr}\left(\widetilde{\mathbf{C}}^H\right)\right)\right\} \tag{60}$$

and

$$m_{12} = E\left\{\mathrm{Re}\left(\mathrm{Tr}\left(\widetilde{\mathbf{C}}\mathbf{\Sigma}^H\right)\right)\right\}, \tag{61}$$

respectively.

Next, we eigen-decompose the Hermitian positive definite matrix $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^H, \tag{62}$$

where $\mathbf{D}$ is a real diagonal matrix and $\mathbf{U}$ is a complex unitary matrix. Define

$$\mathbf{\Lambda} = \mathbf{U}\mathbf{D}^{1/2} \tag{63}$$

and

$$\mathbf{z}_i = \frac{\mathbf{\Lambda}^{-1}\mathbf{s}_i}{\|\mathbf{\Lambda}^{-1}\mathbf{s}_i\|_2}. \tag{64}$$

$\{\mathbf{z}_i\}_{i=1}^n$ are then complex valued isotropically distributed random vectors and are independent to each other [30]–[32]. Using results from [34], it can be shown that

$$E\left\{\mathbf{z}_i\mathbf{z}_i^H\right\} = \frac{1}{p}\mathbf{I}, \tag{65}$$

$$\begin{aligned}E\left\{\left|\mathbf{z}_i^H\mathbf{D}\mathbf{z}_i\right|^2\right\} &= \frac{1}{p(p+1)}\left(\mathrm{Tr}(\mathbf{D}^2) + \mathrm{Tr}^2(\mathbf{D})\right) \\ &= \frac{1}{p(p+1)}\left(\mathrm{Tr}(\mathbf{\Sigma}\mathbf{\Sigma}^H) + \mathrm{Tr}^2(\mathbf{\Sigma})\right)\end{aligned} \tag{66}$$

and

$$E\left\{\left|\mathbf{z}_i^H\mathbf{D}\mathbf{z}_j\right|^2\right\} = \frac{1}{p^2}\mathrm{Tr}(\mathbf{D}^2) = \frac{1}{p^2}\mathrm{Tr}(\mathbf{\Sigma}\mathbf{\Sigma}^H),\ i \neq j. \tag{67}$$

Eq. (65) $\sim$ (67) are complex versions of (46) $\sim$ (48). Expressing $\widetilde{\mathbf{C}}$ in terms of $\mathbf{z}_i$, there is

$$\widetilde{\mathbf{C}} = \frac{p}{n}\mathbf{\Lambda}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i^H\mathbf{\Lambda}^H, \tag{68}$$

and accordingly

$$E\left\{\widetilde{\mathbf{C}}\right\} = \mathbf{\Sigma}. \tag{69}$$

As $\mathrm{Re}(\cdot)$, $\mathrm{Tr}(\cdot)$ and $E\{\cdot\}$ are exchangeable to each other, we have

$$m_{11} = \mathrm{Re}\left(\mathrm{Tr}(\mathbf{\Sigma}^H)\right) = \mathrm{Tr}(\mathbf{\Sigma}) \tag{70}$$

and

$$m_{12} = \mathrm{Tr}(\mathbf{\Sigma}\mathbf{\Sigma}^H). \tag{71}$$

For m_2, using a similar derivation as in Appendix B, it can be shown that

$$\begin{aligned}m_2 &= \frac{p^2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left\{\left|\mathbf{z}_i^H\mathbf{D}\mathbf{z}_j\right|^2\right\} \\ &= \left(1 - \frac{1}{n(p+1)}\right)\mathrm{Tr}(\mathbf{\Sigma}\mathbf{\Sigma}^H) + \frac{p}{n(p+1)}\mathrm{Tr}^2(\mathbf{\Sigma}).\end{aligned} \tag{72}$$

As $\mathrm{Tr}(\mathbf{\Sigma}^H) = \mathrm{Tr}(\mathbf{\Sigma}) = p$, (17) can be finally obtained by substituting (70), (71) and (72) to (58).

REFERENCES

[1] C. Stein, "Estimation of a covariance matrix," In *Rietz Lecture, 39th Annual Meeting*, IMS, Atlanta, GA, 1975.

[2] O. Ledoit, M. Wolf, "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices", *Journal of Multivariate Analysis*, vol. 88, iss. 2, pp. 365 - 411, Feb. 2004.

[3] Y. Chen, A. Wiesel, Y. C. Eldar, A. O. Hero III, "Shrinkage Algorithms for MMSE Covariance Estimation ," *IEEE Trans. on Sig. Process.*, vol. 58, iss. 10, pp. 5016 - 5029, 2010.

[4] J. Friedman, T. Hastie, R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432 - 441, 2008.

[5] P. Bickel, E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, vol. 36, pp. 199-227, 2008.

[6] I. M. Jonestone, "On the distribution of the largest eigenvalue in principal components analysis," *The Annals of Statistics*, vol. 29, pp. 295 - 327, 2001.

[7] D. Kelker, "Distribution Theory of Spherical Distributions and a Location-Scale Parameter Generalization." *Sankhya* vol. A32, no. 4, pp. 419 - 430, 1970.

[8] P.J. Huber, "Robust statistics," Wiley, 1981.

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

[10] P. Rousseeuw, "Multivariate estimation with high breakdown point," *Mathematical Statistics and Applications*, vol. B, pp. 283 - 297, Reidel, 1985.

[11] E. Conte and M. Longo, "Characterization of radar clutter as a spherically invariant random process, in *Proc. Inst. Elect. Eng.*, vol. 134, pp. 191 - 197, Apr. 1987.

[12] E. Conte, M. Lops, and G. Ricci, "Adaptive radar detection in compound- Gaussian clutter, *Proc. Eusipco94*, Edinburgh, U.K., pp. 526 - 529, Sep. 1994.

[13] E. Conte, M. Lops, and G. Ricci, Asymptotically optimum radar detection in compound-Gaussian clutter, IEEE Trans. Aerosp. Electron. Syst., vol. 31, pp. 617 - 625, Apr. 1995.

[14] E. Conte, A. De Maio and G. Ricci, "Recursive Estimation of the Covariance Matrix of a Compound-Gaussian Process and its Application to Adaptive CFAR Detection, *IEEE Trans. on Sig. Process.*, vol. 50, no. 8, pp. 1908 - 1915, Aug. 2002.

[15] F. Gini, "Sub-optimum Coherent Radar Detection in a Mixture of Kdistributed and Gaussian Clutter, *IEE Proc. Radar, Sonar and Navigation*, vol. 144, no. 1, pp. 39 - 48, Feb. 1997.

[16] F. Gini and M. V. Greco, "Covariance matrix estimation for CFAR detection in correlated heavy tailed clutter, *Signal Process. (Special Section on Signal Processing with Heavy Tailed Distributions)*, vol. 82, no. 12, pp. 1847 - 1859, Dec. 2002.

[17] J. Wang, A. Dogandzic, A. Nehorai, "Maximum Likelihood Estimation of Compound-Gaussian Clutter and Target Parameters," *IEEE Trans. on Sig. Process.*, vol. 54, no. 10, pp. 3884 - 3898, 2006.

[18] F. Pascal, Y. Chitour, J.-P. Ovarlez, P. Forster, and P. Larzabal, "Covariance Structure Maximum-Likelihood Estimates in Compound Gaussian Noise: Existence and Algorithm Analysis," *IEEE Trans. on Sig. Process.*, vol. 56, no. 1, pp. 34 - 48, Jan. 2008.

[19] Y. Chitour, F. Pascal, "Exact Maximum Likelihood Estimates for SIRV Covariance Matrix: Existence and Algorithm Analysis," *IEEE Trans. on Sig. Process.*, vol. 56, no. 10, pp. 4563 - 4573, Oct. 2008.

[20] M. Rangaswamy, "Statistical analysis of the nonhomogeneity detector for non-Gaussian interference backgrounds," *IEEE Trans. on Sig. Process.*, vol. 53, no. 6, pp. 2101 - 2111, June 2005.

[21] Y. I. Abramovich and N. K. Spencer, "Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering," *IEEE Intl Conf. on Acoust., Speech, and Signal Processing*, vol. 3, pp. 1105 - 1108, 2007.

[22] J.B. Billingsley, "Ground Clutter Measurements for Surface-Sited Radar, *Technical Report 780*, MIT, Feb. 1993.

[23] H. Brehm, W. Stammler, "Description and generation of spherically invariant speech-model signals," *Signal Processing*, vol. 12, iss. 2, pp. 119 - 141, Mar. 1987.

[24] G. Vasile, *et al.*, "Normalized Coherency Matrix Estimation Under the SIRV Model. Alpine Glacier Polsar Data Analysis," *IEEE International Geoscience and Remote Sensing Symposium*, vol. 1, pp. 74 - 77, 2008.

[25] G. Frahm, "Generalized Elliptical Distributions: Theory and Applications," *Dissertation*, University of Cologne, Department of Economic and Social Statistics, 2004.

[26] K. Yao, "A representation theorem and its applications to spherically invariant random processes, *IEEE Trans. Inf. Theory*, vol. IT-19, no. 5, pp. 600 - 608, Sep. 1973.

[27] K. Yao, M.K. Simon and E. Biglieri, "A unified theory on wireless communication fading statistics based on SIRP, *Signal Processing Advances in Wireless Communications, 2004 IEEE 5th Workshop on*, pp. 135 - 139, 2004.

[28] U. Krause, "Relative stability for ascending and positively homogeneous operators on Banach spaces," *J. Math. Anal. Appl.* pp. 182 - 202, 1994.

[29] U. Krause, "Concave Perron-Frobenius theory and applications," *Nonlinear Analysis,* vol. 47, iss. 3, pp. 1457 - 1466, 2001.

[30] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multipleantenna communication link in Rayleigh flat fading," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 139 - 157, 1999.

[31] B. Hassibi and T. L. Marzetta,"Multiple-antennas and isotropically random unitary inputs: the received signal density in closed form," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1473 - 1484, Jun. 2002.

[32] Y. C. Eldar and S. Shamai, "A covariance shaping framework for linear multiuser detection," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2426 - 2446, 2005.

[33] V. Chandola, A. Banerjee, V. Kumar, "Anomaly detection: a survey," *ACM Computing Surveys*, vol. 41, no. 15, pp. 1 - 58, Jul. 2009.

[34] A. Tulino, S. Verdu, "Random matrix theory and wireless communciations," Now Publishers Inc, 2004.

[35] C. Auderset, C. Mazza, E. A. Ruh, "Angular Gaussian and Cauchy estimation," *Journal of Multivariate Analysis*, vol. 93, iss. 1, pp. 180 - 197, 2005.

[36] D. A. Johns, K. Martin, "Analog integrated circuit design," Wiley, 1997.

[37] H. W. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, vol. 62, pp. 399 - 402, 1967.

[38] M. Finegold, M. Drton, "Robust Graphical Modeling with Classical and Alternative T-Distributions," *preprint available in arXiv:1009.3669*, http://arxiv.org/pdf/1009.3669.

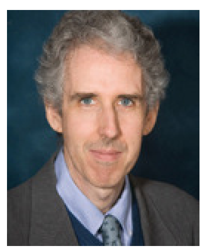
Yilun Chen (S'05) received the B.Eng. and M.Sc. degrees in electrical engineering from Tsinghua University, Beijing, China, in 2005 and 2007, respectively. He is currently a Ph.D candidate with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor.

He performed research in statistical algorithms of radar imaging at High Speed Signal Processing and Network Research Institute, Department of Electrical Engineering, Tsinghua University, between 2005 and 2007. His current research interests include theories and applications of statistical and adaptive signal processing, machine learning and convex optimization.

Ami Wiesel (S'02–M'09) received the B.Sc. and M.Sc. degrees in electrical engineering from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 2000 and 2002, respectively, and the Ph.D. degree in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 2007. From 2007 to 2009, he was a Postdoctoral Fellow with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor.

Currently, he is a faculty member at the The Rachel and Selim Benin School of Computer Science and Engineering at the Hebrew University of Jerusalem, Israel.

Dr. Wiesel was a recipient of the Young Author Best Paper Award for a 2006 paper in the IEEE Transactions in Signal Processing and a Student Paper Award for the 2005 Workshop on Signal Processing Advances in Wireless Communications (SPAWC) paper. He was awarded the Weinstein Study Prize in 2002, the Intel Award in 2005, the Viterbi Fellowship in 2005 and 2007, and the Marie Curie Fellowship in 2008.



Alfred O. Hero (S'79–M'84–SM'96–F'98) received the B.S. (*summa cum laude*) from Boston University (1980) and the Ph.D from Princeton University (1984), both in Electrical Engineering. Since 1984 he has been with the University of Michigan, Ann Arbor, where he is the R. Jamison and Betty Professor of Engineering. His primary appointment is in the Department of Electrical Engineering and Computer Science and he also has appointments, by courtesy, in the Department of Biomedical Engineering and the Department of Statistics. In 2008 he was awarded the the Digiteo Chaire d'Excellence, sponsored by Digiteo Research Park in Paris, located at the Ecole Superieure d'Electricite, Gif-sur-Yvette, France. He has held other visiting positions at LIDS Massachussets Institute of Technology (2006), Boston University (2006), I3S University of Nice, Sophia-Antipolis, France (2001), Ecole Normale Supérieure de Lyon (1999), Ecole Nationale Supérieure des Télécommunications, Paris (1999), Lucent Bell Laboratories (1999), Scientific Research Labs of the Ford Motor Company, Dearborn, Michigan (1993), Ecole Nationale Superieure des Techniques Avancees (ENSTA), Ecole Superieure d'Electricite, Paris (1990), and M.I.T. Lincoln Laboratory (1987 - 1989).

Alfred Hero is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE). He has been plenary and keynote speaker at major workshops and conferences. He has received several best paper awards including: a IEEE Signal Processing Society Best Paper Award (1998), the Best Original Paper Award from the Journal of Flow Cytometry (2008), and the Best Magazine Paper Award from the IEEE Signal Processing Society (2010). He received a IEEE Signal Processing Society Meritorious Service Award (1998), a IEEE Third Millenium Medal (2000) and a IEEE Signal Processing Society Distinguished Lecturership (2002). He was President of the IEEE Signal Processing Society (2006-2007). He sits on the Board of Directors of IEEE (2009-2011) where he is Director Division IX (Signals and Applications).

Alfred Hero's recent research interests have been in detection, classification, pattern analysis, and adaptive sampling for spatio-temporal data. Of particular interest are applications to network security, multi-modal sensing and tracking, biomedical imaging, and genomic signal processing.