# Bias Adjusted Sign Covariance Matrix

Elias Raninen, *Student Member, IEEE*, and Esa Ollila, *Senior Member, IEEE*

***Abstract*—The spatial sign covariance matrix (SSCM), also known as the normalized sample covariance matrix (NSCM), has been widely used in signal processing as a robust alternative to the sample covariance matrix (SCM). It is well-known that the SSCM does not provide consistent estimates of the eigenvalues of the shape matrix (normalized scatter matrix). To alleviate this problem, we propose BASIC (Bias Adjusted SIgn Covariance), which performs an approximate bias correction to the eigenvalues of the SSCM under the assumption that the samples are generated from zero mean unspecified complex elliptically symmetric distributions (the real-valued case is also addressed). We then use the bias correction in order to develop a robust regularized SSCM based estimator, BASIC Shrinkage estimator (BASICS), which is suitable for high dimensional problems, where the dimension can be larger than the sample size. We assess the proposed estimator with several numerical examples as well as in a linear discriminant analysis (LDA) classification problem with real data sets. The simulations show that the proposed estimator compares well to competing robust covariance matrix estimators but has the advantage of being significantly faster to compute.**

***Index Terms*—Covariance matrix, elliptical distribution, spatial sign covariance matrix, robust estimation, shrinkage.**

## I. Introduction

CONSIDER independent and identically distributed (i.i.d.) random samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \subset \mathbb{C}^p$ from a $p$-dimensional elliptically symmetric (ES) distribution $\mathbf{x} \sim F$ with zero mean and positive definite scatter matrix parameter $\mathbf{\Sigma}$. We are interested in estimating $\mathbf{\Sigma}$ up to a scaling constant. Hence, we define the *shape matrix*, $\mathbf{\Lambda}$, and the *scale parameter*, $\eta$, as

$$\mathbf{\Lambda} = \eta^{-1}\mathbf{\Sigma}, \qquad \eta = p^{-1}\mathrm{tr}(\mathbf{\Sigma}).$$

Then $\mathbf{\Lambda}$ satisfies $\mathrm{tr}(\mathbf{\Lambda}) = p$. A conventional estimator for the covariance (scatter) matrix is the (unbiased) sample covariance matrix (SCM), $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}$, which is also the maximum likelihood estimator (MLE) of $\mathbf{\Sigma}$ when sampling from a zero mean multivariate normal (MVN) distribution. However, it is well-known that the SCM is not robust against heavier-tailed distributions or outliers. Hence, robust alternatives such as M-estimators [1] are commonly used. Another commonly used robust estimator is the *spatial sign covariance matrix* (SSCM) [2], [3]

$$\mathbf{S}_{\mathrm{sgn}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}}{\|\mathbf{x}_i\|^2}. \tag{1}$$

Manuscript received October 15, 2021; revised December 3, 2021; accepted December 8, 2021. Date of publication December 13, 2021; date of current version January 28, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Alexander Jung. *(Corresponding author: Elias Raninen.)*

The authors are with the Department of Signal Processing and Acoustics, Aalto University, Aalto FI-00076, Finland (e-mail: elias.raninen@aalto.fi; esa.ollila@aalto.fi).

Digital Object Identifier 10.1109/LSP.2021.3134940

Its corresponding shape estimator given by $\hat{\mathbf{\Lambda}}_{\mathrm{sgn}} = p\mathbf{S}_{\mathrm{sgn}}$ (Since $\mathrm{tr}(\mathbf{S}_{\mathrm{sgn}}) = 1$). We note that $\hat{\mathbf{\Lambda}}_{\mathrm{sgn}}$ is often referred to as the *normalized sample covariance matrix* [4]–[6], and it has become a popular proxy for the unknown clutter covariance matrix in adaptive radar detection problems, especially due to its good robustness properties. The SSCM has the highest possible breakdown point of 1 with fixed location [7] and breakdown point of 1/2 when using the spatial median [8] to estimate the location [9]. The best possible breakdown point of M-estimators is $1/p$ and obtained by Tyler's M-estimator [10]. The SSCM also has the advantage of being faster to compute than complex-valued M-estimators of scatter [11], which are computed using an iterative fixed point (FP) algorithm. However, unlike M-estimators, which are affine equivariant, the SSCM is only unitary equivariant (orthogonally equivariant in the real-valued case) [7], [12]. That is, the transformation $\mathbf{x}_i \mapsto \mathbf{Q}\mathbf{x}_i$, for $i = 1, \ldots, n$, induces $\mathbf{S}_{\mathrm{sgn}} \mapsto \mathbf{Q}\mathbf{S}_{\mathrm{sgn}}\mathbf{Q}^{\mathsf{H}}$ for any unitary matrix $\mathbf{Q}$ (affine equivariance instead means that the same holds for any non-singular matrix).

Assuming ES distributions, define the expectations:

$$\mathbf{\Sigma}_{\mathrm{sgn}} = \mathbb{E}[\mathbf{S}_{\mathrm{sgn}}] \text{ and } \mathbf{\Lambda}_{\mathrm{sgn}} = \mathbb{E}[\hat{\mathbf{\Lambda}}_{\mathrm{sgn}}].$$

It is well-known that the eigenvalues of $\mathbf{\Lambda}_{\mathrm{sgn}}$ and $\mathbf{\Lambda}$ are not equal (except when $\mathbf{\Sigma} \propto \mathbf{I}$). However, their ordering is preserved, and their corresponding eigenvectors are the same [2], [3], [7]. In [13, Theorem 2] it was shown that for a sequence of covariance matrices of increasing dimension, the bias of the SSCM is of the order of the *sphericity*, $\gamma = \mathrm{tr}(\mathbf{\Lambda}^2)/p$, and becomes negligible as the dimension increases provided $\gamma/p \to 0$ as $p \to \infty$. Since $\gamma$ measures how close the shape matrix is to the identity matrix, the implication is that the bias of the SSCM is smaller for approximately spherical covariance matrices ($\gamma$ small) and larger for spiked covariance matrices, where there are only a few large eigenvalues ($\gamma$ large). Letting $\boldsymbol{\lambda} = (\lambda_i)$ denote the eigenvalues of $\mathbf{\Lambda}$ and $\boldsymbol{\delta} = (\delta_i)$ the eigenvalues of $\mathbf{\Sigma}_{\mathrm{sgn}}$, there exists a one-to-one continuously differentiable function $h_i$ such that $\delta_i = h_i(\boldsymbol{\lambda})$ [7]. In [14] a formula for $h_i$ was derived for complex ES distributions in the case that all eigenvalues are distinct. In [15] a one-dimensional integral formula for $h_i$ was derived for real-valued ES distributions, which does not assume distinct eigenvalues.

The main contributions of this letter are as follows. We derive the expectation of a more general weighted SCM (2) (of which the SSCM is a special case) under the assumption that the data is generated from a complex ES distribution, and without any restrictions on the multiplicities of the eigenvalues. The result is easily generalized to real ES distributions. We then use the result to propose BASIC (Bias Adjusted SIgn Covariance): a bias correction that shifts the eigenvalues $(p\delta_i)$ of $\mathbf{\Lambda}_{\mathrm{sgn}}$ closer to the eigenvalues $(\lambda_i)$ of $\mathbf{\Lambda}$. Furthermore, for high-dimensional problems, we develop a SSCM based shrinkage estimator, called BASIC Shrinkage estimator (BASICS).

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

We note that in [7] it was shown that the SSCM is asymptotically inadmissible, which means that the asymptotic covariance matrix of the consistency corrected SSCM is uniformly larger than that of Tyler's [16] M-estimator. However, the computational simplicity of the bias corrected SSCM makes it an appealing alternative to M-estimators that are commonly computed using an iterative FP algorithm.

The paper is organized as follows. In Section II, we derive the expectation of the weighted SCM (Theorem 1). Section III discusses the bias correction scheme and develops the regularized SSCM based estimator. Section IV provides simulation studies and an application in *linear discriminant analysis* (LDA) classification. Section V concludes. All proofs are kept in the appendix.

*Notations:* For a matrix $\mathbf{A}$, $\mathbf{A} \succ \mathbf{0}$ means that $\mathbf{A}$ is positive definite. The notations $(\cdot)^*$, $(\cdot)^\top$, and $(\cdot)^\mathsf{H}$ denote the complex conjugate, transpose, and Hermitian (conjugate transpose). The notation $\lambda(\mathbf{A})$ denotes the eigenvalues of $\mathbf{A}$. The identity matrix is denoted by $\mathbf{I}$ and $|\mathbf{A}|$ is the determinant of $\mathbf{A}$. The notation $\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}[\cdot]$ denotes the expectation taken over a distribution with mean $\boldsymbol{\mu}$ and scatter parameter $\boldsymbol{\Sigma}$. The uniform distribution on the complex unit sphere is denoted by $\mathcal{U}(\mathbb{C}\mathcal{S}^p)$, where $\mathbb{C}\mathcal{S}^p = \{\mathbf{u} \in \mathbb{C}^p : \|\mathbf{u}\| = 1\}$.

## II. Expectation of a Weighted SCM

A (non-degenerate) centered (zero mean) random vector $\mathbf{x} \in \mathbb{C}^p$ is said to have a circular complex elliptically symmetric (CES) distribution if it admits the *stochastic representation* [17], [18]: $\mathbf{x} = r\boldsymbol{\Sigma}^{1/2}\mathbf{u}$, where $r > 0$ is a random variable called the *modular variate* that is independent of $\mathbf{u} \sim \mathcal{U}(\mathbb{C}\mathcal{S}^p)$, and $\boldsymbol{\Sigma}^{1/2}$ is the unique positive definite square-root of $\boldsymbol{\Sigma} \succ \mathbf{0}$. The distribution of $r$ determines the particular ES distribution. If the cumulative distribution function of $r$ is absolutely continuous, the density of $\mathbf{x}$ is up to a constant of the form $|\boldsymbol{\Sigma}|^{-1}g(\mathbf{x}^\mathsf{H}\boldsymbol{\Sigma}^{-1}\mathbf{x})$, where $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ is called the *density generator*. We then write $\mathbf{x} \sim \mathbb{C}\mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$.

Define a weighted SCM by

$$\hat{\boldsymbol{\Sigma}}_\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{x}_i\mathbf{x}_i^\mathsf{H}}{\mathbf{x}_i^\mathsf{H}\mathbf{Z}\mathbf{x}_i} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{v}_i\mathbf{v}_i^\mathsf{H}}{\mathbf{v}_i^\mathsf{H}\mathbf{Z}\mathbf{v}_i}, \tag{2}$$

where $\mathbf{v}_i = \mathbf{x}_i/\|\mathbf{x}_i\|$, and $\mathbf{Z} = \mathbf{Z}^\mathsf{H} \succ \mathbf{0}$ is a fixed (non-random) $p \times p$ matrix, yielding the SSCM for $\mathbf{Z} = \mathbf{I}$. Since $\hat{\boldsymbol{\Sigma}}_\mathbf{Z}$ does not depend on the modular variate $r$, it is distribution-free within the class of ES distributions. The expectation of $\hat{\boldsymbol{\Sigma}}_\mathbf{Z}$ is given next.

*Theorem 1:* Let $\mathbf{x} \sim \mathbb{C}\mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$ with $\boldsymbol{\Sigma} = \eta\boldsymbol{\Lambda}$. For any fixed $\mathbf{Z} = \mathbf{Z}^\mathsf{H} \succ \mathbf{0}$,

$$\mathbb{E}\left[\frac{\mathbf{x}\mathbf{x}^\mathsf{H}}{\mathbf{x}^\mathsf{H}\mathbf{Z}\mathbf{x}}\right] = \int_0^\infty \frac{\boldsymbol{\Lambda}(\mathbf{I} + t\mathbf{Z}\boldsymbol{\Lambda})^{-1}}{|\mathbf{I} + t\mathbf{Z}\boldsymbol{\Lambda}|}\mathrm{d}t \tag{3}$$

$$= \int_0^1 \frac{\boldsymbol{\Lambda}((1-t)\mathbf{I} + t\mathbf{Z}\boldsymbol{\Lambda})^{-1}}{|(1-t)\mathbf{I} + t\mathbf{Z}\boldsymbol{\Lambda}|}(1-t)^{p-1}\mathrm{d}t. \tag{4}$$

The expectation of the SSCM (1) is obtained when $\mathbf{Z} = \mathbf{I}$ while for $\mathbf{Z} = \boldsymbol{\Lambda}^{-1}$ the expectation is equal to $\boldsymbol{\Lambda}/p$. Previously, the expected value of the SSCM (for $p > 2$) was derived in [15] for real ES distributions in the form

$$\delta_i = \frac{\lambda_i}{2}\int_0^\infty \frac{1}{(1+t\lambda_i)\prod_{j=1}^p(1+t\lambda_j)^{1/2}}\mathrm{d}t, \tag{5}$$

where $\delta_i$ denotes the $i$th eigenvalue of $\boldsymbol{\Sigma}_{\text{sgn}}$ and $\lambda_i$ the $i$th eigenvalue of $\boldsymbol{\Lambda}$. The corresponding expression in the complex-valued case can easily be derived from (3) to be

$$\delta_i = \lambda_i \int_0^\infty \frac{1}{(1+t\lambda_i)\prod_{j=1}^p(1+t\lambda_j)}\mathrm{d}t \tag{6}$$

$$= \lambda_i \int_0^1 \frac{(1-t)^{p-1}}{(1-t+t\lambda_i)\prod_{j=1}^p(1-t+t\lambda_j)}\mathrm{d}t \tag{7}$$

$$= (\lambda_i/p)F_D^{(p)}(1,\ldots,1,p+1;1-\lambda_1,\ldots,1-\lambda_p),$$

where $F_D^{(p)}$ denotes Lauricella's hypergeometric series defined via analytic continuation if $|1-\lambda_i| \geq 1$. The integrals above can be computed efficiently using numerical integration such as with the adaptive quadratic method [19].

## III. Estimating the Shape Matrix

For real-valued ES distributions, in [20], a heuristic FP algorithm involving numerical integration was proposed for estimating the eigenvalues of the shape matrix using the eigenvalues of the SSCM. We propose an alternative approach, where we approximate the integral (7) and then invert it. This enables simple computation of the bias correction.

First, we show that the integrand in (7) decreases close to zero well before $t$ reaches the upper limit 1. We can find an upper limit $u$ of the integral corresponding to $\delta_{\max}$ so that the error is at most $\epsilon_{\text{UB}}$. From (7), we have that

$$\int_u^1 \frac{\lambda_{\max}(1-t)^{p-1}}{(1-t+t\lambda_{\max})\prod_{j=1}^p(1-t+t\lambda_j)}\mathrm{d}t$$
$$\leq \int_u^1 \frac{\lambda_{\max}(1-t)^{p-1}}{(1-t+t\lambda_{\max})^2(1-t+t\lambda_{\min})^{p-1}}\mathrm{d}t$$
$$= \frac{\lambda_{\min}(1-u)^p}{p\lambda_{\max}(1-(1-\lambda_{\min})u)^p}$$
$$\cdot\, {}_2F_1\left(2,p;p+1;\frac{(\lambda_{\max}-\lambda_{\min})(1-u)}{\lambda_{\max}(1-(1-\lambda_{\min})u)}\right) = \epsilon_{\text{UB}},$$

where ${}_2F_1$ is the Gaussian hypergeometric function. With $u = 0.15$ an error level $\epsilon_{\text{UB}} \leq 0.05 \cdot \frac{\lambda_{\max}}{p}$ is attained both for DoA 1 and DoA 2 setups of Section IV corresponding to a 4–6 % relative error ($\epsilon_{\text{UB}}/\delta_{\max}$). Using $u = 0.3$, the relative error decreases to 0.2–0.5 %.

The integral (7) simplifies a great deal when the determinant in the denominator is replaced by the approximation

$$\prod_{i=1}^p(1-t+t\lambda_i) = e^{\log\left(\prod_{i=1}^p(1-t+t\lambda_i)\right)}$$
$$= e^{\sum_{i=1}^p \log(1+(\lambda_i-1)t)} \leq e^{\sum_{i=1}^p(\lambda_i-1)t} = 1, \tag{8}$$

where we used $\log(1+x) \approx x$ with $x = (\lambda_i - 1)t$. The approximation error is small for eigenvalues close to the mean eigenvalue, $\mathrm{tr}(\boldsymbol{\Lambda})/p = 1$, and larger for eigenvalues that deviate more from the mean. The approximation also has the interpretation of replacing the geometric mean with the arithmetic mean using the inequality $\prod_{i=1}^p(1+(\lambda_i-1)t)^{1/p} \leq$

$\frac{1}{p}\sum_{i=1}^{p}(1+(\lambda_i-1)t)=1$. Hence, using (8) in (7), we obtain

$$\delta_i \geq \int_0^1 \frac{\lambda_i(1-t)^{p-1}}{1-t+t\lambda_i}\mathrm{d}t = (\lambda_i/p)\cdot {}_2F_1(1,1;p+1;1-\lambda_i) \equiv \tilde{\delta}(\lambda_i), \quad (9)$$

where analytic continuation is used if $|1-\lambda_i|\geq 1$. The approximation[1] only depends on the dimension $p$ and the eigenvalue $\lambda_i$. Therefore, it is possible to form a precalculated table with a one-to-one mapping between $\lambda_i$ and $\tilde{\delta}_i(\lambda_i)$.

The approximate bias correction for the eigenvalues of the SSCM is computed as follows. Given the data $(\mathbf{x}_i)\sim\mathbb{C}\mathcal{E}_p(\mathbf{0},\eta\mathbf{\Lambda},g)$, compute the eigendecomposition $\mathbf{S}_{\text{sgn}}=\mathbf{U}\text{diag}((\hat{\delta}_i))\mathbf{U}^{\mathsf{H}}$ and then compute $\tilde{\lambda}_i=\tilde{\delta}^{-1}(\hat{\delta}_i)$ (the inverse of (9)) using a precomputed table and interpolation between points. As a final step, normalize the estimated eigenvalues via $\hat{\lambda}_i=p\tilde{\lambda}_i/\sum_{j=1}^{p}\tilde{\lambda}_j$. We then obtain the BASIC estimator $\hat{\mathbf{\Lambda}}_{\text{BASIC}}=\mathbf{U}\text{diag}((\hat{\lambda}_i))\mathbf{U}^{\mathsf{H}}$.

When $n<p$, the SSCM is rank deficient. In this case, we propose to use regularization (shrinkage). The regularized SSCM (RSSCM), $\hat{\mathbf{\Lambda}}_{\text{RSSCM}}$, is defined by

$$\text{minimize}_{\alpha\in[0,1]}\ \mathbb{E}[\|\hat{\mathbf{\Lambda}}_{\text{RSSCM}}(\alpha)-\mathbf{\Lambda}_{\text{sgn}}\|_{\text{F}}^2], \quad (10)$$

$$\text{subject to } \hat{\mathbf{\Lambda}}_{\text{RSSCM}}(\alpha)=\alpha\hat{\mathbf{\Lambda}}_{\text{sgn}}+(1-\alpha)\mathbf{I}. \quad (11)$$

By applying bias correction to $\hat{\mathbf{\Lambda}}_{\text{RSSCM}}(\alpha)$, we then obtain the BASICS estimator $\hat{\mathbf{\Lambda}}_{\text{BASICS}}(\alpha)$. Thus, the idea is to first obtain a good estimate of $\mathbf{\Lambda}_{\text{sgn}}$ with $\hat{\mathbf{\Lambda}}_{\text{RSSCM}}$ and then apply BASIC bias correction to its eigenvalues to obtain an estimate of $\mathbf{\Lambda}$.

*Proposition 1:* Let $\mathbf{x}_1,\ldots,\mathbf{x}_n\sim\mathbb{C}\mathcal{E}_p(\mathbf{0},\eta\mathbf{\Lambda},g)$. The $\alpha$, which minimizes (10) is given by

$$\alpha^\star=\frac{\frac{n}{n-1}\left(a-\frac{p}{n}\right)-1}{a-1},\ \text{where } a=p^{-1}\,\mathbb{E}[\text{tr}(\hat{\mathbf{\Lambda}}_{\text{sgn}}^2)].$$

For an estimate $\hat{\alpha}^\star$ of $\alpha^\star$, we only need to estimate $a$, which can be obtained by its sample counterpart $\text{tr}(\hat{\mathbf{\Lambda}}_{\text{sgn}}^2)/p$. Lastly, we project the estimate $\hat{\alpha}^\star$ to the interval [0,1].

## IV. Numerical Experiments

We consider a setting where a uniform linear array (ULA) of $p$ sensors with a half wavelength spacing receives $K$ zero mean independent complex source signals of power $\sigma_{s,i}^2$. The array covariance matrix is then $\mathbf{\Sigma}=\mathbf{A}(\boldsymbol{\theta})\mathbf{\Sigma}_s\mathbf{A}(\boldsymbol{\theta})^{\mathsf{H}}+\sigma_n^2\mathbf{I}$, where $\mathbf{A}(\boldsymbol{\theta})=(\mathbf{a}(\theta_1)\cdots\mathbf{a}(\theta_K))$ is the steering matrix with steering vectors $\mathbf{a}(\theta_k)=(1,e^{-\jmath\pi\sin(\theta_k)},\ldots,e^{-\jmath\pi(p-1)\sin(\theta_k)})^\top$, $\mathbf{\Sigma}_s=\text{diag}(\sigma_{s,1}^2,\ldots,\sigma_{s,K}^2)$ contains the source powers, and $\sigma_n^2$ denotes the noise power. We consider two setups: **DoA 1** has $p=20$ sensors, $K=1$ sources with power $\sigma_{s,1}^2=2$ and direction $\theta_1=15^o$, and noise power $\sigma_n^2=1$ while **DoA 2** has $p=40$ sensors, $K=3$ sources with powers $\mathbf{\Sigma}_s=\text{diag}\ (2,1,0.5)$ and directions $\boldsymbol{\theta}=(-5^o,0^o,5^o)^\top$, and noise power $\sigma_n^2=1$. Fig. 1 displays the first five eigenvalues $(\lambda_i)$ of $\mathbf{\Lambda}$, $(p\delta_i)$ of $\mathbf{\Lambda}_{\text{sgn}}$ (computed using (7)), and $(\tilde{\delta}^{-1}(\delta_i))$ of $\mathbf{\Lambda}_{\text{BASIC}}$ (computed from the inverse function of (9)). Using BASIC, the bias of $\lambda(\mathbf{\Lambda}_{\text{sgn}})$ is alleviated in both models.

[1]The corresponding approximation of (9) in the real-valued case is obtained by $\delta_i\geq\int_0^1\frac{\lambda_i(1-t)^{p/2-1}}{2(1-t+t\lambda_i)}\mathrm{d}t=(\lambda_i/p)\cdot{}_2F_1(1,1;p/2+1;1-\lambda_i)$.

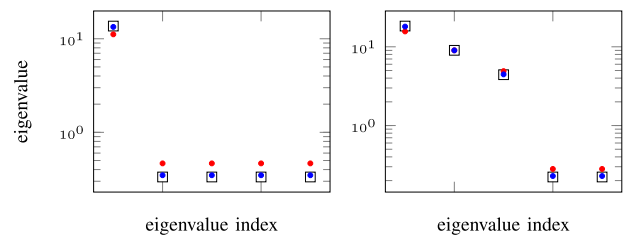


Fig. 1. First five eigenvalues of DoA 1 (*left panel*) and DoA 2 (*right panel*). □ $\lambda(\mathbf{\Lambda})=(\lambda_i)$, • $\lambda(\mathbf{\Lambda}_{\text{sgn}})=(p\delta_i)$, and • $\lambda(\mathbf{\Lambda}_{\text{BASIC}})=(\tilde{\delta}^{-1}(\delta_i))$.

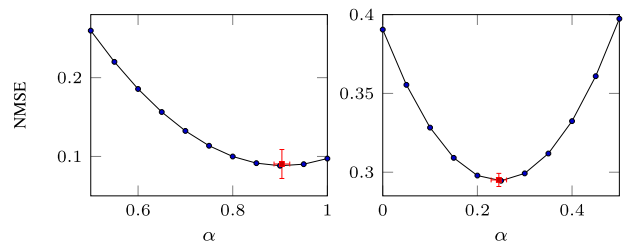


Fig. 2. Theoretical NMSE of $\hat{\mathbf{\Lambda}}_{\text{RSSCM}}(\alpha)$ as a function of $\alpha\in[0,1]$ in black and the average estimated $\hat{\alpha}^\star$ and the corresponding NMSE of $\hat{\mathbf{\Lambda}}_{\text{RSSCM}}(\hat{\alpha}^\star)$ shown in red (error bars denote the standard deviation) for the CS model (*left panel*) and the AR(1) model (*right panel*); $n=50$, $p=100$, and $\varrho=0.5$.

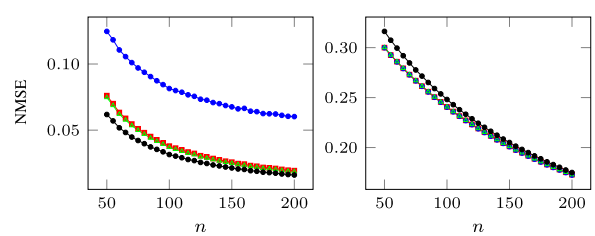


Fig. 3. NMSE of BASICS (red), RSSCM (blue), RFP (green), and CWH (black) for CS (*left panel*) and AR(1) (*right panel*); $p=100$ and $\varrho=0.5$.

The next two simulations assess the regularized estimators using two different models: the compound symmetry (**CS**) model, where $(\mathbf{\Sigma})_{ii}=1$ and $(\mathbf{\Sigma})_{ij}=\varrho$, $i\neq j$, and the first-order autoregressive model (**AR(1)**), where $(\mathbf{\Sigma})_{ij}=\varrho^{|i-j|}$. In both models $\varrho=0.5$ and the dimension is $p=100$.

Fig. 2 depicts the theoretical NMSE in (10) for $\alpha\in[0,1]$ as well as the mean and standard deviation of $\hat{\alpha}^\star$ and the NMSE of $\hat{\mathbf{\Lambda}}_{\text{RSSCM}}(\hat{\alpha}^\star)$ averaged over $21\cdot 2000$ independent Monte Carlo (MC) trials. For each MC, $n=50$ samples were drawn from the complex MVN distribution $\mathbb{C}\mathcal{N}_p(\mathbf{0},\mathbf{\Sigma})$. The estimate $\hat{\alpha}^\star$ coincides well with the minimum NMSE.

Next, we compare BASICS, RSSCM, RFP (RSSCM with bias correction using the FP algorithm [20]) and the regularized Tyler's M-estimator (CWH) [21] (implemented as in [22]) in the estimation of the shape matrix. Fig. 3 displays the NMSE, Ave $\|\hat{\mathbf{\Lambda}}-\mathbf{\Lambda}\|_{\text{F}}^2/\|\mathbf{\Lambda}\|_{\text{F}}^2$, as a function of $n$ averaged over 2000 independent MC trials for each $n$, when sampling from the multivariate $t$-distribution $\mathbb{C}t_{\nu,p}(\mathbf{0},\mathbf{\Sigma})$ with $\nu=2$ degrees of freedom. In the CS case, the bias correction significantly improved the performance, whereas in the AR(1) case, the performance was unchanged.

Lastly, we apply the methods in *linear discriminant analysis* (LDA) where $\mathbf{x}$ is classified to one of $K$ classes by the rule

$$\hat{k}=\arg\min_{k\in\{1,\ldots,K\}}-2\mathbf{x}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k+\boldsymbol{\mu}_k^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k, \quad (12)$$

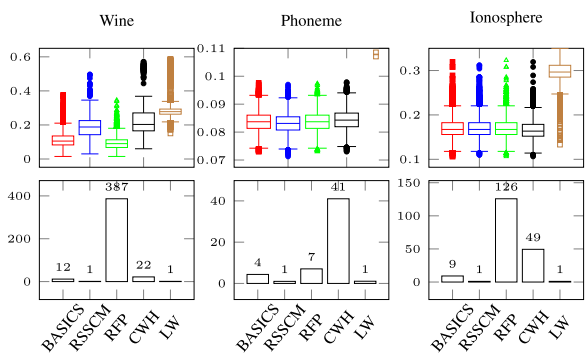


Fig. 4. LDA misclassification rates (*top panel*) and the median computation times relative to the median computation time of RSSCM (*bottom panel*).

where it is assumed that class distributions are $\mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ [23]. LDA using regularized estimators of covariance is an active area of research, where recent improvements especially focus on utilization of random matrix theory [24]–[26]. Here, we illustrate the benefits of the proposed BASICS for LDA in a heavy-contamination setting. We consider the following data sets: Wine [27] ($n_1 = 59, n_2 = 71, n_3 = 48, p = 13$), Phoneme [23] ($n_1 = 695$, $n_2 = 1022$, $n_3 = 757$, $n_4 = 1163$, $n_5 = 872$, $p = 256$), and Ionosphere[2] [27] ($n_1 = 225$, $n_2 = 126$, $p = 32$). We used 25% of randomly selected samples from each class as the training set for the estimation of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}$. The training data set was contaminated by replacing 10 % of the samples from each class by a random sample from a Cauchy distribution, $\mathcal{C}_p((\hat{\mu}_i), \mathrm{diag}((10\hat{\sigma}_i^2)))$, where $\hat{\mu}_i$ is the sample mean and $\hat{\sigma}_i^2$ is the sample variance of the $i$th variable of the whole data set, respectively. The shape matrix $\boldsymbol{\Lambda}$ was estimated from the pooled centered training data by first centering each data class $k$ with respect to its spatial median. The misclassification rate was computed using the uncontaminated remaining samples (test set).

Fig. 4 depicts the misclassification rates and median computation times for BASICS, RSSCM, RFP of [20], CWH of [21] and (regularized SCM based non-robust) LW of [28] averaged over 10000 independent MC trials.[3] The reported computation times are relative to the median computation time of the RSSCM and exclude the computation time for centering the data.[4] As can be noted, for the Wine data set, BASICS and RFP had the best performance while for the other data sets, the differences between the performances of the robust methods were minor. RFP and CWH were, however, computationally significantly more demanding and their computation times varied a lot between the data sets. RSSCM, BASICS, and LW were computationally most efficient.

## V. Conclusion

We proposed a BASIC estimator of the shape matrix that approximately corrects for the bias of the eigenvalues of the SSCM. We also proposed the BASIC Shrinkage estimator (BASICS), which applies bias correction for the regularized SSCM and is useful for $n < p$ settings. The usefulness of the proposed estimators were demonstrated in simulations as well as in LDA classification with contaminated data. BASICS compared well to the regularized Tyler's M-estimator, but had a significantly lower computation time. Codes for the proposed methods are available at github.com/EliasRaninen.

[2]First two variables were removed as they were constant for the other class.

[3]Stopping criterions were $\|\mathbf{I} - \hat{\boldsymbol{\Lambda}}_{k-1}^{-1}\hat{\boldsymbol{\Lambda}}_k\|_\infty < 10^{-5}$ for CWH, and $\|\hat{\boldsymbol{\lambda}}_{k-1} - \hat{\boldsymbol{\lambda}}_k\|/\|\hat{\boldsymbol{\lambda}}_k\| < 10^{-5}$ for RFP, where $k$ denotes the current iterate.

[4]The CPU model used for computation was Intel Core i5-8365 U.

## Appendix

*Proof of Theorem 1*: Since $\mathbf{x}\mathbf{x}^\mathsf{H}/(\mathbf{x}^\mathsf{H}\mathbf{Z}\mathbf{x})$ is distribution-free within the class of ES distributions, we can without loss of generality assume that $\mathbf{x} \sim \mathbb{C}\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Lambda})$. Define

$$\mathbf{F}(t) = \frac{\mathbf{x}\mathbf{x}^\mathsf{H}}{\mathbf{x}^\mathsf{H}\mathbf{Z}\mathbf{x}} e^{-t\mathbf{x}^\mathsf{H}\mathbf{Z}\mathbf{x}}, \quad t \geq 0. \tag{13}$$

The goal is then to solve for $\mathbb{E}[\mathbf{F}(0)]$. Since $|\frac{\partial}{\partial t}(\mathbf{F}(t))_{ij}| = |-x_i x_j^* e^{-t\mathbf{x}^\mathsf{H}\mathbf{Z}\mathbf{x}}| \leq |x_i x_j^*|$ for $t \geq 0$, and $\mathbb{E}[|x_i x_j^*|] < \infty$, we apply the Dominated Convergence Theorem (see [29, pp. 26], [30, Sec. 2.4]) to obtain

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\mathbf{F}(t)\right] &= \mathbb{E}\left[\frac{\partial}{\partial t}\mathbf{F}(t)\right] \\
&= -\int \mathbf{x}\mathbf{x}^\mathsf{H} e^{-t\mathbf{x}^\mathsf{H}\mathbf{Z}\mathbf{x}} \cdot |\pi\boldsymbol{\Lambda}|^{-1} e^{-\mathbf{x}^\mathsf{H}\boldsymbol{\Lambda}^{-1}\mathbf{x}} \mathrm{d}\mathbf{x} \\
&= -|\pi\boldsymbol{\Lambda}|^{-1} \int \mathbf{x}\mathbf{x}^\mathsf{H} e^{-\mathbf{x}^\mathsf{H}(\boldsymbol{\Lambda}^{-1}+t\mathbf{Z})\mathbf{x}} \mathrm{d}\mathbf{x} \\
&= -|\pi\boldsymbol{\Lambda}|^{-1}|\pi\mathbf{T}|\,\mathbb{E}_{\mathbf{0},\mathbf{T}}[\mathbf{x}\mathbf{x}^\mathsf{H}] = -|\boldsymbol{\Lambda}|^{-1}|\mathbf{T}|\mathbf{T},
\end{aligned}$$

where we let $\mathbf{T}^{-1} = \boldsymbol{\Lambda}^{-1} + t\mathbf{Z}$. Since $\mathbf{T} = (\boldsymbol{\Lambda}^{-1} + t\mathbf{Z})^{-1} = \boldsymbol{\Lambda}(\mathbf{I} + t\mathbf{Z}\boldsymbol{\Lambda})^{-1}$, we can further write

$$\frac{|\mathbf{T}|}{|\boldsymbol{\Lambda}|}\mathbf{T} = \frac{|\boldsymbol{\Lambda}(\mathbf{I} + t\mathbf{Z}\boldsymbol{\Lambda})^{-1}|}{|\boldsymbol{\Lambda}|}\boldsymbol{\Lambda}(\mathbf{I} + t\mathbf{Z}\boldsymbol{\Lambda})^{-1} = \frac{\boldsymbol{\Lambda}(\mathbf{I} + t\mathbf{Z}\boldsymbol{\Lambda})^{-1}}{|\mathbf{I} + t\mathbf{Z}\boldsymbol{\Lambda}|}.$$

Then solving for $\mathbb{E}[\mathbf{F}(0)]$ by integration yields

$$\begin{aligned}
\mathbb{E}[\mathbf{F}(0)] &= -\left(\mathbb{E}[\mathbf{F}(\infty)] - \mathbb{E}[\mathbf{F}(0)]\right) = -\int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\mathbf{F}(t)\right]\mathrm{d}t \\
&= -\int_0^\infty \mathbb{E}\left[\frac{\partial}{\partial t}\mathbf{F}(t)\right]\mathrm{d}t = \int_0^\infty \frac{\boldsymbol{\Lambda}(\mathbf{I} + t\mathbf{Z}\boldsymbol{\Lambda})^{-1}}{|\mathbf{I} + t\mathbf{Z}\boldsymbol{\Lambda}|}\mathrm{d}t,
\end{aligned}$$

since $\mathbf{F}(\infty) = \mathbf{0}$. The indefinite integral can be rewritten as an integral over [0,1] via the substitution $t = x/(1-x)$, $\mathrm{d}t = \mathrm{d}x/(1-x)^2$, so that we obtain

$$\mathbb{E}[\mathbf{F}(0)] = \int_0^1 \frac{\boldsymbol{\Lambda}((1-x)\mathbf{I} + x\mathbf{Z}\boldsymbol{\Lambda})^{-1}}{|(1-x)\mathbf{I} + x\mathbf{Z}\boldsymbol{\Lambda}|}(1-x)^{p-1}\mathrm{d}x.$$

This concludes the proof for the complex-valued case. The real-valued case follows similarly by choosing $\mathbf{F}(t) = \mathbf{x}\mathbf{x}^\top/(\mathbf{x}^\top\mathbf{Z}\mathbf{x})e^{-t\mathbf{x}^\top\mathbf{Z}\mathbf{x}/2}$ in (13) and $\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Lambda})$. ■

*Proof of Proposition 1*: Solving the zero derivative equation of the convex function $\mathbb{E}[\|\hat{\boldsymbol{\Lambda}}(\alpha) - \boldsymbol{\Lambda}_{\mathrm{sgn}}\|_\mathrm{F}^2] = \alpha^2\,\mathbb{E}[\|\hat{\boldsymbol{\Lambda}}_{\mathrm{sgn}} - \mathbf{I}\|_\mathrm{F}^2] + \|\boldsymbol{\Lambda}_{\mathrm{sgn}} - \mathbf{I}\|_\mathrm{F}^2 - 2\alpha\,\mathbb{E}[\mathrm{tr}((\hat{\boldsymbol{\Lambda}}_{\mathrm{sgn}} - \mathbf{I})(\boldsymbol{\Lambda}_{\mathrm{sgn}} - \mathbf{I}))]$ results in $\alpha^\star = \frac{\mathrm{tr}(\boldsymbol{\Lambda}_{\mathrm{sgn}}^2)/p - 1}{a-1}$. Since $\mathrm{tr}(\boldsymbol{\Lambda}_{\mathrm{sgn}}^2) = \mathrm{tr}(\mathbb{E}[\hat{\boldsymbol{\Lambda}}_{\mathrm{sgn}}]^2)$ and from [13, eq. (38)], we have $\mathrm{tr}(\mathbb{E}[\hat{\boldsymbol{\Lambda}}_{\mathrm{sgn}}]^2)/p = n/(n-1)(a - p/n)$, the proof is complete. ■

## REFERENCES


[1] R. A. Maronna, "Robust M-estimators of multivariate location and scatter," *Ann. Statist.*, vol. 4, no. 1, Jan. 1976.

[2] N. Locantore *et al.*, "Robust principal component analysis for functional data," *Test*, vol. 8, no. 1, pp. 1–73, Jun. 1999.

[3] S. Visuri, V. Koivunen, and H. Oja, "Sign and rank covariance matrices," *J. Stat. Plan. Inference*, vol. 91, no. 2, pp. 557–575, Dec. 2000.

[4] E. Conte, M. Lops, and G. Ricci, "Adaptive radar detection in compound-Gaussian clutter," in *Proc. Eur. Signal Process. Conf.*, Edinburgh, Scotland, 1994, pp. 526–529.

[5] E. Conte, M. Lops, and G. Ricci, "Adaptive matched filter detection in spherically invariant noise," *IEEE Signal Process. Lett.*, vol. 3, no. 8, pp. 248–250, Aug. 1996.

[6] E. Conte, M. Lops, and G. Ricci, "Adaptive detection schemes in compound-Gaussian clutter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 34, no. 4, pp. 1058–1069, Oct. 1998.

[7] A. F. Magyar and D. E. Tyler, "The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions," *Biometrika*, vol. 101, no. 3, pp. 673–688, Sep. 2014.

[8] B. M. Brown, "Statistical uses of the spatial median," *J. Roy. Stat. Soc.: Ser. B (Methodological)*, vol. 45, no. 1, pp. 25–30, Sep. 1983.

[9] C. Croux, C. Dehon, and A. Yadine, "The k-step spatial sign covariance matrix," *Adv. Data Anal. Classification*, vol. 4, no. 2, pp. 137–150, Sep. 2010.

[10] L. Dümbgen and D. E. Tyler, "On the breakdown properties of some multivariate M-functionals," *Scand. J. Statist.*, vol. 32, no. 2, pp. 247–264, 2005.

[11] E. Ollila and V. Koivunen, "Robust antenna array processing using M-estimators of pseudo-covariance," in *Proc. 14th IEEE Int. Symp. Pers., Indoor Mobile Radio Commun.*, 2003, vol. 3, pp. 2659–2663.

[12] C. Croux, E. Ollila, and H. Oja, "Sign and rank covariance matrices: Statistical properties and application to principal components analysis," in *Statistical Data Analysis Based on the L1-Norm and Related Methods*, Y. Dodge, Ed. Basel, Switzerland: Birkhäuser, 2002, pp. 257–269.

[13] E. Raninen, D. E. Tyler, and E. Ollila, "Linear pooling of sample covariance matrices," *IEEE Trans. Signal Process.*, 2021, accepted for publication.

[14] S. Bausson, F. Pascal, P. Forster, J.-P. Ovarlez, and P. Larzabal, "First- and second-order moments of the normalized sample covariance matrix of spherically invariant random vectors," *IEEE Signal Process. Lett.*, vol. 14, no. 6, pp. 425–428, Jun. 2007.

[15] A. Dürre, D. E. Tyler, and D. Vogel, "On the eigenvalues of the spatial sign covariance matrix in more than two dimensions," *Statist. Probability Lett.*, vol. 111, pp. 80–85, Apr. 2016.

[16] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *Ann. Statist.*, vol. 15, no. 1, pp. 234–251, 1987.

[17] K.-T. Fang, S. Kotz, and K. W. Ng, *Symmetric Multivariate and Related Distributions*. London, U.K.: Chapman & Hall, 1990.

[18] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor, "Complex elliptically symmetric distributions: Survey, new results and applications," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5597–5625, Nov. 2012.

[19] L. F. Shampine, "Vectorized adaptive quadrature in MATLAB," *J. Comput. Appl. Math.*, vol. 211, no. 2, pp. 131–140, Feb. 2008.

[20] A. Dürre, R. Fried, and D. Vogel, "The spatial sign covariance matrix and its application for robust correlation estimation," *Austrian J. Statist.*, vol. 46, no. 3-4, pp. 13–22, Apr. 2017.

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

[22] E. Ollila, D. P. Palomar, and F. Pascal, "Shrinking the eigenvalues of M-estimators of covariance matrix," *IEEE Trans. Signal Process.*, vol. 69, pp. 256–269, 2021.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[24] X. Yang, K. Elkhalil, A. Kammoun, T. Y. Al-Naffouri, and M.-S. Alouini, "Regularized discriminant analysis: A large dimensional study," in *Proc. IEEE Int. Symp. Inf. Theory*, 2018, pp. 536–540.

[25] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "A large dimensional study of regularized discriminant analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 2464–2479, 2020.

[26] N. Auguin, D. Morales-Jimenez, and M. R. McKay, "Large-dimensional characterization of robust linear discriminant analysis," *IEEE Trans. Signal Process.*, vol. 69, pp. 2625–2638, 2021.

[27] D. Dua and C. Graff, "UCI machine learning repository," Irvine, CA: University of California, School of Information and Computer Science. 2019. [Online]. Available: https://archive.ics.uci.edu/ml

[28] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivariate Anal.*, vol. 88, no. 2, pp. 365–411, 2004.

[29] W. Rudin, *Real and Complex Analysis (Mathematics Series)*, 3rd ed. New York, NY, USA: McGraw-Hill, 1987.

[30] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Boston, MA, USA: Cengage Learning, 2001.