# Regularized Tapered Sample Covariance Matrix

Esa Ollila, *Senior member, IEEE* and Arnaud Breloy

***Abstract*—Covariance matrix tapers have a long history in signal processing and related fields. Examples of applications include autoregressive models (promoting a banded structure) or beamforming (widening the spectral null width associated with an interferer). In this paper, the focus is on high-dimensional setting where the dimension $p$ is high, while the data aspect ratio $n/p$ is low. We propose an estimator called TABASCO (TApered or BAnded Shrinkage COvariance matrix) that shrinks the tapered sample covariance matrix towards a scaled identity matrix. We derive optimal and estimated (data adaptive) regularization parameters that are designed to minimize the mean squared error (MSE) between the proposed shrinkage estimator and the true covariance matrix. These parameters are derived under the general assumption that the data is sampled from an unspecified elliptically symmetric distribution with finite 4th order moments (both real- and complex-valued cases are addressed). Simulation studies show that the proposed TABASCO outperforms all competing tapering covariance matrix estimators in diverse setups. An application to space-time adaptive processing (STAP) also illustrates the benefit of the proposed estimator in a practical signal processing setup.**

***Index Terms*—sample covariance matrix, shrinkage, regularization, elliptically symmetric distributions, tapering, banding, sphericity.**

## I. INTRODUCTION

Consider a set of $p$-dimensional (real-valued) vectors $\{\mathbf{x}_i\}_{i=1}^n$ sampled from a distribution of a random vector $\mathbf{x}$ with unknown mean vector $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$ and unknown positive definite symmetric $p \times p$ covariance matrix $\boldsymbol{\Sigma} \equiv \mathrm{cov}(\mathbf{x}) = \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]$. In the high-dimensional case and when the sample size $n$ is of the same order as $p$ ($p = O(n)$) or $p \gg n$, one is required to use regularization (shrinkage) in order to improve the estimation accuracy of the sample covariance matrix (SCM) and to obtain a positive definite matrix estimate. A popular estimate of $\boldsymbol{\Sigma}$ in such a setting is the regularized SCM (RSCM), defined by

$$\mathbf{S}_\beta = \beta\mathbf{S} + (1-\beta)\frac{\mathrm{tr}(\mathbf{S})}{p}\mathbf{I}, \tag{1}$$

where $\beta \in [0,1]$ is the regularization (or shrinkage) parameter, and where

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top, \tag{2}$$

denotes the unbiased sample covariance matrix (SCM), i.e., $\mathbb{E}[\mathbf{S}] = \boldsymbol{\Sigma}$. Above $\mathrm{tr}(\cdot)$ denotes the matrix trace, defined as $\mathrm{tr}(\mathbf{A}) = \sum_i a_{ii}$ for all square matrices $\mathbf{A} = (a_{ij})$. Note also that in (2), $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ denotes the sample mean vector. Automatic data-adaptive computation of optimal (oracle) parameter $\beta$ for which $\mathbf{S}_\beta$ in (1) attains the minimum mean squared error (MMSE) in Frobenius norm has been an active area of research. See for example [1], [2], [3], [4] to name only a few.

In many applications, the estimation accuracy (or another performance criterion) can alternatively be improved by using a so-called *tapered* SCM. Such estimate is defined as $\mathbf{W} \circ \mathbf{S}$, where $\circ$ denotes the Hadamard (or Schur) element-wise product (i.e., $(\mathbf{W} \circ \mathbf{S})_{ij} = w_{ij}s_{ij}$ for $(\mathbf{W})_{ij} = w_{ij}$ and $(\mathbf{S})_{ij} = s_{ij}$), and where $\mathbf{W}$ is a *tapering* matrix (also referred to as covariance matrix taper), i.e., a template that imposes some additional structure to the SCM.

Covariance matrix tapers have been used in many applications in diverse fields. A first main example in statistics is related to covariance matrices with a diagonally dominant structure (e.g., in autoregressive models). This means that the variables have a natural order in the sense that $|i-j|$ large implies that the correlation between the $i$th and the $j$th variables is close to zero. In this settings, popular estimation approaches are to use a banding-type tapering matrices such as thresholding [5], [6]:

$$(\mathbf{W})_{ij} = \begin{cases} 1, & |i-j| < k \\ 0, & |i-j| \geq k \end{cases} \tag{3}$$

for some integer $k \in [\![1, p]\!]$ (called the bandwidth parameter), or softer thresholding variants. Notably, the strong theoretical merits of a linear decay of the form

$$(\mathbf{W})_{ij} = \begin{cases} 1, & |i-j| \leq k/2 \\ 2 - 2\dfrac{|i-j|}{k}, & k/2 < |i-j| < k \\ 0, & |i-j| \geq k \end{cases} \tag{4}$$

were studied in [7]. A second major example concerns the signal processing literature, in which tapering matrices have been developed in order to improve several spectral properties of adaptive beamformers, or to compensate subspace leakage and calibration issues [8]. Most notably, the tapering matrices of the form

$$(\mathbf{W})_{ij} = \mathrm{sinc}((i-j)\Delta/\pi) \tag{5}$$

where $\Delta \in \mathbb{R}^+$, attracted interest as a null broadening technique for fluctuating interference [9], [10], [11], [12], [13].

A first approach to combine regularization with tapering was proposed in [14] with the *shrinkage to tapering* (ST) estimator, defined as the convex combination of the SCM and the tapered SCM:

$$\mathbf{S}_{\mathrm{ST},\beta} = \beta\mathbf{S} + (1-\beta)(\mathbf{W} \circ \mathbf{S}), \tag{6}$$

where $\beta \in [0,1]$ is a shrinkage parameter. The authors then derived the optimal oracle parameter $\beta_o$ minimizing the

E. Ollila is with the Department of Signal Processing and Acoustics, Aalto University, P.O. Box 15400, FI-0007 Aalto, Finland.

Arnaud Breloy is with the LEME (EA4416), University Paris Nanterrre, 92410 Ville-d'Avray, France.

The code used in the presented experiments is available in the Matlab toolbox https://github.com/esollila/Tabasco.

MSE $\mathbb{E}[\|\mathbf{S}_{\mathrm{ST},\beta}-\boldsymbol{\Sigma}\|_{\mathrm{F}}^2]$, and proposed a shrinkage to tapering oracle approximating (STOA) estimator $\hat{\beta}_o$ of $\beta_o$ under the assumption of Gaussian data. Authors in [15] also studied the ST estimator and derived an alternative oracle estimator of the shrinkage parameter both under Gaussian and non-Gaussian data. Data adaptive selection of the bandwidth $k$ in (3) was also addressed with cross validation [14] or oracle estimation [15]. A possible issue with the ST estimate is that it inherently destroys the tapering template structure (e.g., sparsity for banded matrices) since it can be expressed as the modified tapered SCM $\mathbf{S}_{\mathrm{ST},\beta} = (\beta\mathbf{1}\mathbf{1}^\top + (1-\beta)\mathbf{W}) \circ \mathbf{S}$. Hence, shrinkage is applied to the tapering matrix itself rather than to the SCM. In the high dimensional case, it should also be noted that both $\mathbf{W}\circ\mathbf{S}$ and $\mathbf{S}_{\mathrm{ST},\beta}$ are not necessarily positive semidefinite matrices, i.e., they can have negative or null eigenvalues. A possible solution for this problem is to compute their eigenvalue decomposition (EVD) and then replacing the invalid eigenvalues by small positive constants. However, such a post-processing step further deteriorates the template pattern of the covariance matrix estimator, and is computationally restrictive when dealing with high-dimensional data.

In this paper we provide a solution to the aforementioned problems by jointly leveraging shrinkage to identity and tapering: Let $\mathbb{W} = \{\mathbf{W}(k)\}_{k=1}^K$ be a finite set of possible tapering matrices[1] satisfying $\mathbf{W}(k) \in \mathcal{W}^+ \ \forall k \in [\![1, K]\!]$, with

$$\mathcal{W}^+ = \{\mathbf{W} \in \mathbb{R}_{\mathrm{Sym}}^{p\times p} : w_{ii} = 1, w_{ij} \geq 0 \ \forall i,j \in [\![1,p]\!]\} \quad (7)$$

and with $\mathbb{R}_{\mathrm{Sym}}^{p\times p}$ denoting the set of all symmetric $p\times p$ matrices and $[\![1,p]\!] = \{1,\ldots,p\}$. We propose an estimator, referred to as TABASCO (TApered or BAnded Shrinkage COvariance matrix), defined as

$$\hat{\boldsymbol{\Sigma}}_{\beta,k} = \beta(\mathbf{W}(k)\circ\mathbf{S}) + (1-\beta)\frac{\mathrm{tr}(\mathbf{S})}{p}\mathbf{I}, \quad (8)$$

which benefits both from shrinkage (as the classic estimator in (1)) and exploitation of structure via tapering. Note that it also preserves the original scale of the SCM since $\mathrm{tr}(\mathbf{W}\circ\mathbf{S}) = \mathrm{tr}(\mathbf{S}) \ \forall \mathbf{W} \in \mathcal{W}^+$. Obviously, the success of banding and/or tapering depends on one's ability to choose the parameters $\beta$ and $k$ correctly. In this scope, we derive a fully automatic data-adaptive evaluation of the optimal parameters that jointly minimize the mean squared error $\mathbb{E}[\|\hat{\boldsymbol{\Sigma}}_{\beta,k}-\boldsymbol{\Sigma}\|_{\mathrm{F}}^2]$ under the general assumption that the data is sampled from an unspecified elliptically symmetric (ES) distribution with finite 4th order moments. A main interest to consider the general ES model is that it encompasses the standard Gaussian one while still accounting for possibly heavy-tailed distributions. Thus this assumption yields robustness to a large class of possible underlying data distributions. Our empirical experiments evidence that the proposed approach offers a near-to-optimal regularization parameter selection which outperform cross-validation schemes (especially at low sample support).

[1] In this paper, we mostly focus on $k$ implying a notion of bandwidth (or model order), for which $\mathbb{W}$ can be constructed from (3) or (4) with $k \in [\![1,p]\!]$. However, the proposed methodology applies to the general setting where $\mathbb{W}$ corresponds to any finite collection of possibly envisioned templates. Notably, we will also consider an application where $k$ indexes a set of possible $\{\Delta_k\}_{k=1}^K$ used for the template model in (5).

Since both the RSCM in (1) (if $\mathbf{W} = \mathbf{1}\mathbf{1}^\top \in \mathbb{W}$) and the tapered SCM ($\beta = 1$) appear as special cases of (8), TABASCO performs never worse than these two estimators in terms of MSE independently of the underlying structure of the true covariance matrix.

The paper is structured as follows. In Section II expressions for the oracle regularization parameters $\beta$ and $k$ that minimize the MSE are derived in the general case of sampling from an unspecified distribution with finite 4th-order moments. In Section III a practical closed-form expression for the optimal regularization parameters are derived when sampling from an unspecified ES distribution, and an adaptive fully automatic procedure for their computation is proposed. It is shown that the optimal parameters depend on the sphericity of the tapered covariance matrix $\mathbf{W}\circ\boldsymbol{\Sigma}$, and we address the estimation of this quantity in Section IV. Section V extends our results for complex-valued observations. The special cases of known location ($\boldsymbol{\mu} = \mathbf{0}$) is also briefly discussed in Appendix A. Section VI provides simulation studies while in Section VII the estimator is applied to STAP on a real dataset. Finally, Section VIII concludes. The Appendix contains more technical proofs.

## II. ORACLE TABASCO PARAMETERS $\beta$ AND $k$

Recall that the TABASCO estimator $\hat{\boldsymbol{\Sigma}}_{\beta,k}$ is defined by (8) for a set $\mathbb{W} = \{\mathbf{W}(k)\}_{k=1}^K$ of envisioned tapering matrices (*cf.* footnote [1] for examples) and a regularization parameter $\beta \in [0,1]$. In this section, we derive the expression of the oracle parameters $\beta$ and $k$ that minimize the MSE in the general case of sampling from an unspecified $p$-variate distribution with finite 4th-order moments.

### A. Oracle shrinkage parameter $\beta$ for fixed $k$

First, let us introduce some notations and statistical parameters that are elemental in the proposed method. The *scale* is defined as the mean of the eigenvalues, and denoted by

$$\eta = \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{p} \quad (9)$$

while the *sphericity* of $\boldsymbol{\Sigma}$ [16], [17] is defined as

$$\gamma \equiv \gamma(\boldsymbol{\Sigma}) = \frac{p\,\mathrm{tr}(\boldsymbol{\Sigma}^2)}{\mathrm{tr}(\boldsymbol{\Sigma})^2} = \frac{p\|\boldsymbol{\Sigma}\|_{\mathrm{F}}^2}{\mathrm{tr}(\boldsymbol{\Sigma})^2}, \quad (10)$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the *Frobenius matrix norm*, i.e., $\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{A}^\top\mathbf{A})$. The sphericity measures how close $\boldsymbol{\Sigma}$ is to a scaled identity matrix: $\gamma \in [1,p]$, where $\gamma = 1$ if and only if $\boldsymbol{\Sigma} \propto \mathbf{I}$ and $\gamma = p$ if and only if $\boldsymbol{\Sigma}$ has its rank equal to 1. For any $\mathbf{W} \in \mathcal{W}^+$ as in (7), the matrix $\mathbf{W}\circ\boldsymbol{\Sigma}$, is called the tapered covariance matrix and we denote by

$$\gamma_{\mathbf{W}} \equiv \gamma(\mathbf{W}\circ\boldsymbol{\Sigma}) = \frac{p\|\mathbf{W}\circ\boldsymbol{\Sigma}\|_{\mathrm{F}}^2}{\mathrm{tr}(\boldsymbol{\Sigma})^2}, \quad (11)$$

the sphericity parameter of the tapered covariance matrix. Note that in (11) we utilised the fact that for any $p\times p$ matrix $\mathbf{A}$ and any $\mathbf{W} \in \mathcal{W}^+$, it holds that $\mathrm{tr}(\mathbf{W}\circ\mathbf{A}) = \mathrm{tr}(\mathbf{A})$. When $\mathbf{W} = \mathbf{1}\mathbf{1}^\top$, we write $\gamma_{\mathbf{1}\mathbf{1}^\top} \equiv \gamma$ for brevity.

We start by assuming that the index $k$ is fixed. This allows us to simply denote the fixed tapering matrix $\mathbf{W} \equiv \mathbf{W}(k)$ and TABASCO as $\hat{\mathbf{\Sigma}}_\beta \equiv \hat{\mathbf{\Sigma}}_{\beta,k}$. To find the oracle MMSE shrinkage parameter $\beta \in [0,1]$ of $\hat{\mathbf{\Sigma}}_\beta$, the aim is thus to solve

$$\beta_o = \underset{\beta\in[0,1]}{\arg\min}\, \mathbb{E}\big[\|\hat{\mathbf{\Sigma}}_\beta - \mathbf{\Sigma}\|_{\mathrm{F}}^2\big]. \tag{12}$$

Notice that the MSE of the tapered SCM is

$$\begin{aligned}\mathrm{MSE}(\mathbf{W}\circ\mathbf{S}) &= \mathbb{E}\big[\|\mathbf{W}\circ\mathbf{S} - \mathbf{\Sigma}\|_{\mathrm{F}}^2\big] \\ &= \mathbb{E}\big[\|\mathbf{W}\circ\mathbf{S}\|_{\mathrm{F}}^2\big] + \|\mathbf{\Sigma}\|_{\mathrm{F}}^2 - 2\|\mathbf{V}\circ\mathbf{\Sigma}\|_{\mathrm{F}}^2,\end{aligned} \tag{13}$$

where

$$\mathbf{V} = (v_{ij})_{p\times p} \text{ with } v_{ij} = \sqrt{w_{ij}} \text{ for } \mathbf{W}\in\mathcal{W}^+. \tag{14}$$

By normalized MSE (NMSE) we refer to the quantity $\mathrm{NMSE}(\mathbf{W}\circ\mathbf{S}) = \mathrm{MSE}(\mathbf{W}\circ\mathbf{S})/\|\mathbf{\Sigma}\|_{\mathrm{F}}^2$. We are now ready to state the main result of this section.

**Theorem 1.** *Let $\{\mathbf{x}_i\}_{i=1}^n$ be an i.i.d. random sample from any $p$-variate distribution with finite 4th order moments. For any fixed $\mathbf{W}\in\mathcal{W}^+$, the oracle parameter $\beta_o$ in* (12) *is*

$$\beta_o = \frac{\|\mathbf{V}\circ\mathbf{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2}{\mathbb{E}\Big[\big\|\mathbf{W}\circ\mathbf{S} - \hat{\eta}\mathbf{I}\big\|_{\mathrm{F}}^2\Big]} \tag{15}$$

$$= \frac{p(\gamma_{\mathbf{V}}-1)\eta^2}{\mathbb{E}\left[\|\,\mathbf{W}\circ\mathbf{S}\,\|_{\mathrm{F}}^2\right] - p^{-1}\mathbb{E}[\mathrm{tr}(\mathbf{S})^2]} \tag{16}$$

$$= \frac{(\gamma_{\mathbf{V}}-1)}{\gamma\cdot\mathrm{NMSE}(\mathbf{W}\circ\mathbf{S}) + 2\gamma_{\mathbf{V}} - \gamma - \mathbb{E}[\hat{\eta}^2]/\eta^2} \tag{17}$$

*where $\gamma_{\mathbf{V}}$ is defined via* (11) *and $\hat{\eta} = \mathrm{tr}(\mathbf{S})/p$. Furthermore, the value of the MSE at the optimum is*

$$\begin{aligned}\mathrm{MSE}(\hat{\mathbf{\Sigma}}_{\beta_o}) &= \frac{\mathbb{E}\big[\,\mathrm{tr}(\mathbf{S})^2\big] - \mathrm{tr}(\mathbf{\Sigma})^2}{p} \\ &+ \|\mathbf{\Sigma}\|_{\mathrm{F}}^2 - \|\mathbf{V}\circ\mathbf{\Sigma}\|_{\mathrm{F}}^2 + (1-\beta_0)\,\|\mathbf{V}\circ\mathbf{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2.\end{aligned} \tag{18}$$

*Proof.* The proof is postponed to Appendix B. □

Notice that Theorem 1 also provides the MMSE shrinkage parameter $\beta_o$ for the RSCM $\mathbf{S}_\beta$ in (1) since $\hat{\mathbf{\Sigma}}_\beta = \mathbf{S}_\beta$ when $\mathbf{W} = \mathbf{1}\mathbf{1}^\top$. For the RSCM the optimal parameter is

$$\beta_o = \frac{(\gamma-1)}{\gamma\cdot\mathrm{NMSE}(\mathbf{S}) + \gamma - \mathbb{E}[\hat{\eta}^2]/\eta^2}, \tag{19}$$

where we used (17) and the facts that $\gamma = \gamma_{\mathbf{V}}$ and $\mathbf{W}\circ\mathbf{S} = \mathbf{S}$ for $\mathbf{W} = \mathbf{1}\mathbf{1}^\top$. The MMSE of the RSCM utilizing the optimal shrinkage parameter in (19) is

$$\mathrm{MSE}(\mathbf{S}_{\beta_o}) = \frac{\mathbb{E}\big[\,\mathrm{tr}(\mathbf{S})^2\big] - \mathrm{tr}(\mathbf{\Sigma})^2}{p} + (1-\beta_0)\,\|\mathbf{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2,$$

where we used (18) and that $\mathbf{V}\circ\mathbf{\Sigma} = \mathbf{\Sigma}$ for $\mathbf{V} = \mathbf{1}\mathbf{1}^\top$.

### B. Oracle index $k$

Notice that $\mathrm{MSE}(\hat{\mathbf{\Sigma}}_{\beta_o})$ in (18) implicitly depends on $k$ through $\mathbf{W}\equiv\mathbf{W}(k)$ and $\mathbf{V}$ defined in (14). We further have the relation

$$\begin{aligned}\mathrm{NMSE}(\hat{\mathbf{\Sigma}}_{\beta_o}) &= C - \frac{\|\mathbf{V}\circ\mathbf{\Sigma}\|_{\mathrm{F}}^2}{\|\mathbf{\Sigma}\|_{\mathrm{F}}^2} + (1-\beta_0)\frac{\|\mathbf{V}\circ\mathbf{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2}{\|\mathbf{\Sigma}\|_{\mathrm{F}}^2} \\ &= C - \frac{\gamma_{\mathbf{V}}}{\gamma} + (1-\beta_0)\frac{\gamma_{\mathbf{V}}-1}{\gamma} \\ &= C - \frac{1}{\gamma} + \frac{\beta_0(1-\gamma_{\mathbf{V}})}{\gamma},\end{aligned} \tag{20}$$

where $C$ is a constant that is not dependent on $k$. Equation (20) then implies that minimizing the MSE with respect to $k$ is equivalent to set

$$k_o = \arg\min_k \beta_0(k)(1-\gamma_{\mathbf{V}}(k)), \tag{21}$$

where $\beta_o(k)$ is given by any of the expressions in (15)-(17) and $\gamma_{\mathbf{V}}(k)$ is defined via (11). Note that we have made explicit the dependence of $\beta_0$ and $\gamma_{\mathbf{V}}$ on $k$ in (21) for clarity of exposition. We use a brute force strategy to determine $k_0$, i.e., compute $\beta_0(k)(1-\gamma_{\mathbf{V}(k)})$ for all indices $k$, and choose $k_0$ as the index that resulted in minimum value of the objective.

## III. DATA ADAPTIVE TABASCO UNDER ES DISTRIBUTIONS

The oracles parameters found in the previous section depend on the true underlying data distribution and covariance matrix through various unknown quantities. A practical implementation of TABASCO thus requires their adaptive evaluation. Rather than resorting to a costly (and potentially inaccurate) cross-validation scheme, we will consider the general case where the data is sampled from an unspecified ES distribution [18], [19] to refine the result of Theorem 1. In this setting, we show that the oracle parameter $\beta$ eventually depend on only few scalar-valued statistics that can be accurately estimated.

### A. ES distributions

Before stating the main results, we first recall some definitions and key results regarding ES distribution [18], [19]. The probability density function of an elliptically distributed random vector, denoted by $\mathbf{x}\sim\mathcal{E}_p(\boldsymbol{\mu},\mathbf{\Sigma},g)$, is given by

$$f(\mathbf{x}) = C_{p,g}|\mathbf{\Sigma}|^{-1/2}g\big((\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\big), \tag{22}$$

where $\mathbf{\Sigma}$ denotes the positive definite symmetric covariance matrix parameter, $\boldsymbol{\mu}$ is the mean vector, $g:\mathbb{R}_{\geq 0}\to\mathbb{R}_{>0}$ is the *density generator*, which is a fixed function that is independent of $\mathbf{x}$, $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$, and $C_{p,g}$ is a normalizing constant ensuring that $f(\mathbf{x})$ integrates to 1. Note that here we define $g$ such that "scatter matrix" parameter $\mathbf{\Sigma}$ coincides with the covariance matrix. This can always be assumed (under assumption of finite 2nd order moments) without any loss of generality [18], [19]. For example, the multivariate normal (MVN) distribution, denoted by $\mathcal{N}_p(\boldsymbol{\mu},\mathbf{\Sigma})$, is obtained when $g(t) = \exp(-t/2)$. The flexibility regarding the density generator $g$ allows for modeling a large class of distributions, including heavy-tailed ones such as the multivariate $t$-distribution (MVT) with $\nu > 2$ degrees of freedom (d.o.f.), denoted by $\mathbf{x}\sim t_\nu(\boldsymbol{\mu},\mathbf{\Sigma})$, where

$\nu > 2$ needs to be assumed for finite 2nd-order moments and $\nu > 4$ for finite 4th-order moments. The density generator in this case is $g(t) = (1 + t/(\nu - 2))^{-\frac{\nu+p}{2}}$.

The elliptical kurtosis [20] parameter $\kappa$ is defined as

$$\kappa = \frac{\mathbb{E}[r^4]}{p(p+2)} - 1 = \frac{1}{3}\mathrm{kurt}(x_i), \tag{23}$$

where the expectation is over the distribution of the random variable $r = \|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})\|$ and $\mathrm{kurt}(x_i)$ denotes the excess kurtosis of any (e.g., $i$th) marginal variable of $\mathbf{x}$, defined as

$$\mathrm{kurt}(x_i) = \frac{\mathbb{E}[(x_i - \mu_i)^4]}{(\mathbb{E}[(x_i - \mu_i)^2])^2} - 3,$$

where $\mu_i = \mathbb{E}[x_i]$. Furthermore, observe that $\mathbb{E}[r^2] = p$. The elliptical kurtosis parameter vanishes (so $\kappa = 0$) when $\mathbf{x}$ has a MVN distribution.

We also recall from [4, Lemma 2] that

$$\mathbb{E}\left[\|\mathbf{S}\|_{\mathrm{F}}^2\right] = (1 + \tau_1 + \tau_2)\,\|\boldsymbol{\Sigma}\|_{\mathrm{F}}^2 + \tau_1\,\mathrm{tr}(\boldsymbol{\Sigma})^2, \tag{24}$$

$$\mathbb{E}\left[\mathrm{tr}(\mathbf{S})^2\right] = 2\tau_1\|\boldsymbol{\Sigma}\|_{\mathrm{F}}^2 + \big(1 + \tau_2\big)\,\mathrm{tr}(\boldsymbol{\Sigma})^2, \tag{25}$$

where the scalars

$$\tau_1 = \frac{1}{n-1} + \frac{\kappa}{n} \qquad \text{and} \quad \tau_2 = \frac{\kappa}{n} \tag{26}$$

are dependent on the elliptical distribution (and hence on the density generator $g$) only via its kurtosis parameter.

### B. Oracle shrinkage parameter $\beta$ under ES distributions

As in the previous derivations, we assume that the index $k$ is fixed and simply denote the fixed tapering matrix $\mathbf{W} \equiv \mathbf{W}(k)$. This section refines the result of Theorem 1 when assuming that the data is sampled from an unspecified ES distribution. We first remark that the optimal $\beta$ in (16) involves the quantity $\mathbb{E}\left[\|\mathbf{W} \circ \mathbf{S}\|_{\mathrm{F}}^2\right]$, which can be specified thanks to derivation of the following lemma.

**Lemma 1.** *Let $\{\mathbf{x}_i\}_{i=1}^n$ be an i.i.d. random sample from $\mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ with finite 4th order moments. Then for any $\mathbf{W} \in \mathcal{W}^+$, it holds that*

$$\mathbb{E}\left[\|\mathbf{W} \circ \mathbf{S}\|_{\mathrm{F}}^2\right] = (1 + \tau_1 + \tau_2)\|\mathbf{W} \circ \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 + \tau_1\,\mathrm{tr}((\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{W})^2)$$

*and*

$$\mathbb{E}\left[\mathrm{tr}((\mathbf{D}_{\mathbf{S}}\mathbf{W})^2)\right] = 2\tau_1\|\mathbf{W} \circ \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 + (1 + \tau_2)\,\mathrm{tr}((\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{W})^2),$$

*where $\mathbf{D}_{\boldsymbol{\Sigma}} = \mathrm{diag}(\boldsymbol{\Sigma})$ and $\mathbf{D}_{\mathbf{S}} = \mathrm{diag}(\mathbf{S})$.* [2]

*Proof.* The proof is postponed to Appendix C. □

Note that if $\mathbf{W} = \mathbf{1}\mathbf{1}^\top$, then $\mathrm{tr}((\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{W})^2) = \mathrm{tr}(\boldsymbol{\Sigma})^2$ and $\mathbf{W} \circ \mathbf{S} = \mathbf{S}$ so the expectations in Lemma 1 coincide with those of [4, Lemma 2] (i.e., (24) and (25)).

Using Lemma 1 we may now derive a simpler closed form expression of the optimal shrinkage parameter $\beta_o$ given in Theorem 1 that depends only on few summary (scalar-valued) statistics. Let us denote

$$\theta_{\mathbf{W}} = \frac{\mathbf{d}_{\boldsymbol{\Sigma}}^\top(\mathbf{W} \circ \mathbf{W})\mathbf{d}_{\boldsymbol{\Sigma}}}{p^2} = \frac{\mathrm{tr}((\mathbf{D}_{\boldsymbol{\Sigma}}\mathbf{W})^2)}{p^2}, \tag{27}$$

where $\mathbf{d}_{\boldsymbol{\Sigma}} = (\sigma_1^2, \ldots, \sigma_p^2)^\top$ contains the variances of the variables, i.e., the diagonal elements of $\boldsymbol{\Sigma}$. Obviously, if $\mathbf{W} = \mathbf{1}\mathbf{1}^\top$, then $\theta_{\mathbf{W}} = \eta^2$. The 2nd equality in (27) follows from [21, Lemma 7.5.2].

**Theorem 2.** *Let $\{\mathbf{x}_i\}_{i=1}^n$ be an i.i.d. random sample from an ES distribution $\mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ with finite 4th order moments. For any $\mathbf{W} \in \mathcal{W}^+$, the oracle parameter $\beta_o$ in (12) is*

$$\beta_o = \frac{t}{t + (n/(n-1))(p\theta_{\mathbf{W}}/\eta^2 + \gamma_{\mathbf{W}} - 2\gamma/p) + \kappa \cdot A}, \tag{28}$$

*where $t = n(\gamma_{\mathbf{V}} - 1)$ and*

$$A = p\theta_{\mathbf{W}}/\eta^2 - 1 + 2\gamma_{\mathbf{W}} - 2\gamma/p. \tag{29}$$

*Proof.* Follows from Theorem 1 after substituting the values of $\mathbb{E}\left[\|\mathbf{W} \circ \mathbf{S}\|_{\mathrm{F}}^2\right]$ given in Lemma 1 and of $\mathbb{E}\left[\mathrm{tr}(\mathbf{S})^2\right]$ given in (25) into the denominator of $\beta_o$ in (16) and simplifying the expression. □

### C. Data adaptive TABASCO implementation

Following from Theorem 2, the proposed data-adaptive implementation of TABASCO consists in applying the oracle procedure of Section II by replacing each of the unknown parameters $\{\eta, \theta_{\mathbf{W}}, \kappa, \gamma, \gamma_{\mathbf{W}}, \gamma_{\mathbf{V}}\}$ in the optimal $\beta_0$ from (28) by carefully chosen estimates:

- For $\eta$ and $\theta_{\mathbf{W}}$, we use the empirical estimates:
$$\hat{\eta} = \mathrm{tr}(\mathbf{S})/p \text{ and } \hat{\theta}_{\mathbf{W}} = \mathrm{tr}((\mathbf{D}_{\mathbf{S}}\mathbf{W})^2)/p^2 \tag{30}$$
- The elliptical kurtosis $\kappa$ can be estimated using an estimator $\hat{\kappa}$ detailed in [4, Sect. IV] as (bias-corrected) average sample excess kurtosis of the marginal variables scaled by $1/3$. Also note that if the data is assumed to follow the MVN distribution, we can set $\kappa = 0$, and the last term $\kappa \cdot A$ can be ignored in the denominator.
- The estimation of the three sphericity statistics: $\gamma$, $\gamma_{\mathbf{W}}$, and $\gamma_{\mathbf{V}}$ is more intricate. It will be addressed in detail in Section IV, which will propose two estimation methods (denoted Ell1 and Ell2) for these quantities. Also notice that $\mathbf{V} = (\sqrt{w_{ij}})_{p\times p}$, so if $\mathbf{W}$ is a selection matrix (i.e., that has only 0-s or 1-s as its off-diagonal elements), as for example in (3), then $\mathbf{W} = \mathbf{V}$ so only $\gamma_{\mathbf{W}}$ needs to be estimated.

Using these plug-in values yields an estimate $\hat{\beta}_o(k)$ for each template in the set $\mathbb{W} = \{\mathbf{W}(k)\}_{k=1}^K$. Similarly, the index $k$ is estimated based on (21) by replacing the unknown $\beta_0(k)$ and $\gamma_{\mathbf{V}}(k)$ by their estimates and solving

$$\hat{k}_o = \arg\min_k \hat{\beta}_0(k)(1 - \hat{\gamma}_{\mathbf{V}}(k)), \tag{31}$$

which allows then to compute a fully data-adaptive TABASCO estimator $\hat{\boldsymbol{\Sigma}}_{\hat{\beta}_0, \hat{k}_0}$ as in (8). The pseudocode of the proposed estimation algorithm is summarized in Algorithm 1.

[2] We denote $\mathrm{diag}(\mathbf{A}) \equiv \mathrm{diag}(a_{11}, \ldots, a_{pp})$ when the operator is applied to any matrix $\mathbf{A} = (a_{ij})_{p\times p}$. Conversely $\mathrm{diag}(\mathbf{a})$ denotes a diagonal matrix with the entries of vector $\mathbf{a}$ on the main diagonal.

**Algorithm 1:** TABASCO

**Input :** Data $\{\mathbf{x}_i\}_{i=1}^n$, templates set $\{\mathbf{W}(k)\}_{k=1}^K$
Compute SCM $\mathbf{S}$ in (2).
Compute $\hat{\eta}$ from (30)
Compute $\hat{\kappa}$ from [4, Sect. IV]
Compute $\hat{\gamma}$ (options in Section IV)
**for** $k \in [\![1, K]\!]$ **do**
  Set $\mathbf{W} = \mathbf{W}(k)$ and $\mathbf{V} = \mathbf{V}(k) = (\sqrt{w_{ij}(k)})_{p\times p}$
  Compute $\hat{\theta}_{\mathbf{W}}$ from (30)
  Compute $\hat{\gamma}_{\mathbf{W}}(k)$ and $\hat{\gamma}_{\mathbf{V}}(k)$ (options in Section IV)
  Compute $\hat{\beta}_o(k)$ from (28) using plug-in estimates
Select optimal $k_0$ as in (31) with $\{\hat{\beta}_o(k), \hat{\gamma}_{\mathbf{V}}(k)\}_{k=1}^K$
Set $\mathbf{W} = \mathbf{W}(\hat{k}_o)$ and $\hat{\beta} = \hat{\beta}_o(k)$
**Output :** $\hat{\mathbf{\Sigma}} = \hat{\beta} \cdot (\mathbf{W} \circ \mathbf{S}) + (1 - \hat{\beta})\hat{\eta}\mathbf{I}$

## IV. ESTIMATORS OF SPHERICITY

In this section, we detail two new alternative estimators of the sphericity of the tapered covariance matrix $\mathbf{W} \circ \mathbf{\Sigma}$ in order to compute the parameters $\{\gamma, \gamma_{\mathbf{W}}, \gamma_{\mathbf{V}}\}$ in TABASCO. These estimators are nontrivial extensions of the (non-tapered) Ell1- and Ell2-sphericity estimators in [4]. Notation "Ell" simply refers to the fact that both estimators assume that the data is drawn from an unspecified elliptically symmetric distribution.

### A. Ell1-estimator of sphericity

The Ell1-estimator is based on the *spatial sign covariance matrix* (SSCM) [22], [23], which has been popular for constructing robust estimates of sphericity [24], [25]. The robust properties of SSCM comes from the fact that it is distribution-free under elliptical models and has the highest possible breakdown point [26], [27]. The Ell1-estimator was theoretically studied in [28] and we propose here its generalization to the sphericity of the tapered covariance matrix $\mathbf{W} \circ \mathbf{\Sigma}$.

First, define the shape matrix (or normalized covariance matrix) as $\mathbf{\Lambda} = p\frac{\mathbf{\Sigma}}{\mathrm{tr}(\mathbf{\Sigma})}$ and note that $\mathrm{tr}(\mathbf{\Lambda}) = p$. The sphericity measures $\gamma$ and $\gamma_{\mathbf{W}}$ for any $\mathbf{W} \in \mathcal{W}^+$ can then be expressed simply in terms of $\mathbf{\Lambda}$ via the formulas:

$$\gamma = \frac{\|\mathbf{\Lambda}\|_{\mathrm{F}}^2}{p} \quad \text{and} \quad \gamma_{\mathbf{W}} = \frac{\|\mathbf{W} \circ \mathbf{\Lambda}\|_{\mathrm{F}}^2}{p}.$$

The (scaled) SSCM is defined by

$$\hat{\mathbf{\Lambda}} = \frac{p}{n}\sum_{i=1}^n \frac{(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top}{\|\mathbf{x}_i - \hat{\boldsymbol{\mu}}\|^2}, \tag{32}$$

where $\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|$ is the *sample spatial median* [29]. The sample mean could also be used, but [30] concludes that the use of spatial median as the location estimator for SSCM is clearly preferable. When $\boldsymbol{\mu}$ is known (and without loss of generality assuming $\boldsymbol{\mu} = \mathbf{0}$), the SSCM is defined as $\hat{\mathbf{\Lambda}} = \frac{p}{n}\sum_{i=1}^n \frac{\mathbf{x}_i\mathbf{x}_i^\top}{\|\mathbf{x}_i\|^2}$.

Recently, it was shown in [28] that the following estimate of sphericity based on the SSCM (when $\boldsymbol{\mu}$ is known),

$$\hat{\gamma} = \frac{n}{n-1}\left(\frac{\|\hat{\mathbf{\Lambda}}\|_{\mathrm{F}}^2}{p} - \frac{p}{n}\right), \tag{33}$$

is asymptotically (as $p \to \infty$) unbiased when sampling from elliptical distributions under the following assumption

(A) The sequence of covariance matrix structures being considered with increasing $p$ satisfies $\gamma = o(p)$ as $p \to \infty$.

In other words, $\mathbb{E}[\hat{\gamma}] \to \gamma$ as $p \to \infty$ when (A) holds. We note that Assumption (A) is sufficiently general and holds for many covariance matrix models as shown in [28, Prop. 3]. The following Theorem presents a modification of the Ell1-estimator [4] for the sphericity of $\mathbf{W} \circ \mathbf{\Sigma}$ with equivalent asymptotic guarantees.

**Theorem 3.** *Let $\{\mathbf{x}_i\}_{i=1}^n$ be an i.i.d. random sample from an ES distribution $\mathcal{E}_p(\boldsymbol{\mu}, \mathbf{\Sigma}, g)$ with known $\boldsymbol{\mu} = \mathbf{0}$. Then, for any $\mathbf{W} \in \mathcal{W}^+$ and under Assumption (A), the following statistic*

$$\hat{\gamma}_{\mathbf{W}} = \frac{n}{n-1}\left(\frac{\|\mathbf{W} \circ \hat{\mathbf{\Lambda}}\|_{\mathrm{F}}^2}{p} - \frac{\mathrm{tr}\left((\mathbf{D}_{\hat{\mathbf{\Lambda}}}\mathbf{W})^2\right)}{np}\right), \tag{34}$$

*where $\mathbf{D}_{\hat{\mathbf{\Lambda}}} = \mathrm{diag}(\hat{\mathbf{\Lambda}})$, is asymptotically, as $p \to \infty$, unbiased estimator of $\gamma_{\mathbf{W}} = \gamma(\mathbf{W} \circ \mathbf{\Sigma})$ in (11), i.e., $\mathbb{E}[\hat{\gamma}_{\mathbf{W}}] \to \gamma_{\mathbf{W}}$ as $p \to \infty$, for any fixed $n$.*

*Proof.* Proof is postponed to the Appendix D. □

Also observe that when $\mathbf{W} = \mathbf{1}\mathbf{1}^\top$, then $\hat{\gamma}_{\mathbf{W}}$ reduces to $\hat{\gamma}$ in (33).

When the mean is unknown and estimated by spatial median $\hat{\boldsymbol{\mu}}$, then the results in [24] derived in the case of $\mathbf{W} = \mathbf{1}\mathbf{1}^\top$ show that when $p = O(n^2)$, the estimator of sphericity $\hat{\gamma}$ needs to be corrected for bias. Thus the reader should bear in mind that $\hat{\gamma}$ (and $\hat{\gamma}_{\mathbf{W}}$ ) may have a non-negligible bias when $\boldsymbol{\mu}$ is estimated by $\hat{\boldsymbol{\mu}}$ and $p$ is orders of magnitude larger than $n$.

### B. Ell2-estimator of sphericity

The Ell2-estimator of sphericity was proposed in [4] and we derive here its adaptation to the sphericity of the tapered covariance matrix $\mathbf{W} \circ \mathbf{\Sigma}$.

The derivation starts by noticing that the obvious plug-in estimate $\|\mathbf{W} \circ \mathbf{S}\|_{\mathrm{F}}^2 / p$ for the parameter

$$\vartheta_{\mathbf{W}} = \frac{\|\mathbf{W} \circ \mathbf{\Sigma}\|_{\mathrm{F}}^2}{p} \tag{35}$$

is biased by resorting to for $\mathbb{E}[\|\mathbf{W} \circ \mathbf{S}\|_{\mathrm{F}}^2]$ in Lemma 1. As a remedy, the following theorem derives a proper unbiased estimator of $\vartheta_{\mathbf{W}}$ which extends [4, Theorem 4] (provided that the elliptical kurtosis parameter $\kappa$ is known).

**Theorem 4.** *Let $\{\mathbf{x}_i\}_{i=1}^n$ be an i.i.d. random sample from a $p$-variate elliptical distribution $\mathcal{E}_p(\boldsymbol{\mu}, \mathbf{\Sigma}, g)$ with finite 4th order moments. Then, an unbiased estimator of $\vartheta_{\mathbf{W}} = \|\mathbf{W} \circ \mathbf{\Sigma}\|_{\mathrm{F}}^2 / p$ for any finite $n$ and $p$ and any $\mathbf{W} \in \mathcal{W}^+$ is*

$$\hat{\vartheta}_{\mathbf{W}} = b_n\left(\frac{\|\mathbf{W} \circ \mathbf{S}\|_{\mathrm{F}}^2}{p} - a_n\frac{\mathrm{tr}\left((\mathbf{D}_{\mathbf{S}}\mathbf{W})^2\right)}{p}\right),$$

*where*

$$a_n = \frac{1}{n+\kappa}\left(\frac{n}{n-1} + \kappa\right) \tag{36}$$

$$b_n = \frac{(\kappa+n)(n-1)^2}{(n-2)(3\kappa(n-1) + n(n+1))}. \tag{37}$$

*Proof.* Note that $a_n$ in (36) can be written as $a_n = \tau_1/(1+\tau_2)$, where definitions of $\tau_1$ and $\tau_2$ are given by (26) while $b_n$ in (37) can be expressed as $b_n = (\tau_1+\tau_2-2\tau_1 a_n)^{-1}$. Then using Lemma 1, we notice that

$$b_n^{-1} p\mathbb{E}[\hat{\vartheta}_{\mathbf{W}}] = (\tau_1 - a_n(1+\tau_2))\,\mathrm{tr}((\mathbf{D}_{\mathbf{\Sigma}}\mathbf{W})^2) + (1+\tau_1+\tau_2-2\tau_1 a_n)\,\|\mathbf{W}\circ\mathbf{\Sigma}\|_{\mathrm{F}}^2 = b_n^{-1}\,\|\mathbf{W}\circ\mathbf{\Sigma}\|_{\mathrm{F}}^2$$

The expressions (36) and (37) are obtained when replacing the values of $\tau_1$ and $\tau_2$ given in (26) into $a_n \equiv a_n(\tau_1,\tau_2)$ and $b_n \equiv b_n(\tau_1,\tau_2)$ and simplifying the obtained expressions. □

Then note that the sphericity of the tapered covariance matrix can also be written as

$$\gamma_{\mathbf{W}} = \vartheta_{\mathbf{W}}/\eta^2,$$

where $\vartheta_{\mathbf{W}}$ and $\eta$ are defined in (35) and (10) respectively. Using this expression, we consider the estimate where $\hat{\vartheta}_{\mathbf{W}}$ is computed from Theorem 4, and $\hat{\eta}^2$ is obtained from (30). This yields the estimator

$$\hat{\gamma}_{\mathbf{W}} = p\hat{b}_n\left(\frac{\|\mathbf{W}\circ\mathbf{S}\|_{\mathrm{F}}^2}{\mathrm{tr}(\mathbf{S})^2} - \hat{a}_n\frac{\mathrm{tr}\left((\mathbf{D}_{\mathbf{S}}\mathbf{W})^2\right)}{\mathrm{tr}(\mathbf{S})^2}\right), \tag{38}$$

where $\hat{a}_n \equiv a_n(\hat{\kappa})$ and $\hat{b}_n \equiv b_n(\hat{\kappa})$ are obtained by replacing the unknown $\kappa$ in (36) and (37) by its estimate $\hat{\kappa}$ [4, Sect. IV]. We refer to (38) as Ell2-estimator of sphericity $\gamma_{\mathbf{W}}$. Also note that, if $n$ is reasonably large, then $\hat{b}_n \approx 1$ and $n/(n+\hat{\kappa}) \approx 1$, its expression can be simplified to

$$\hat{\gamma}_{\mathbf{W}} \approx \frac{p\,\|\mathbf{W}\circ\mathbf{S}\|_{\mathrm{F}}^2}{\mathrm{tr}(\mathbf{S})^2} - (1+\hat{\kappa})\frac{p}{n}\frac{\mathrm{tr}((\mathbf{D}_{\mathbf{S}}\mathbf{W})^2)}{\mathrm{tr}(\mathbf{S})^2}.$$

In the non-tapered case ($\mathbf{W} = \mathbf{1}\mathbf{1}^\top$), the estimator in (38) reduces to the Ell2-estimator of sphericity in [4].

### C. Remarks

Although Ell2-estimator of sphericity does not require knowledge of the underlying elliptically symmetric distribution of the data, it is not a robust estimator. Thus we overall favour Ell1-estimator due to robustness of SSCM, and recommend usage of Ell2-estimator when dealing with data that is not heavy-tailed, i.e., which can be approximated by a Gaussian distribution. The non-robustness of Ell2-estimator is due to its usage of 4th-order moments. Namely, $\mathrm{tr}(\mathbf{S}^2) = \sum_{i=1}^p\sum_{j=1}^p s_{ij}^2$ where $s_{ij} = (\mathbf{S})_{ij} = \frac{1}{n}\sum_{\ell=1}^n x_{\ell i}x_{\ell j} - \bar{x}_i\bar{x}_j$, can be written as a sum of mixed 4th-order sample moments. By central limit theorem, any 4th-order sample moment has a limiting normal distribution if the moments of $\mathbf{x}$ exists up to 8th order. Without this condition, the estimator $\mathrm{tr}(\mathbf{S}^2)/p$ can be highly-variable.

Ell1-estimator is highly robust and performs well for heavier-tailed data. Yet, the assumption $\gamma = o(p)$ needed by Ell1-estimator means that $\gamma \ll p$ when dimension is large. Implicitly, this often means that the eigenvalues of $\mathbf{\Sigma}$ tend to be similar as $p$ grows. In [28, Theorem 2] it was shown that the bias of the SSCM $\hat{\mathbf{\Lambda}}$ is of the order of the sphericity, $\gamma = \mathrm{tr}(\mathbf{\Lambda}^2)/p$, and becomes negligible when $\gamma = o(p)$. Thus for distinctively non-spherical covariance matrices, Ell1-estimator has a visible bias, but the bias vanishes for large $p$ (since the estimator is asymptotically unbiased as shown in Theorem 3 when $\gamma = o(p)$). Thus Ell2-estimator can be favoured when the data is approximately Gaussian, e.g., when the estimated kurtosis is smaller than 1. For example, for MVT distribution with $\nu = 10$, $\mathrm{kurt}(x_i) = 1.0$.

Finally, we mention that in practice, we also always use the thresholding

$$\hat{\gamma} = \min(p, \max(1, \hat{\gamma})) \tag{39}$$

for any option in order to guarantee that the final estimator remain in the valid interval, $1 \le \gamma \le p$.

## V. Extension to the complex-valued case

### A. Complex elliptically symmetric (CES) distribution

First we recall some definitions and notations specific to complex-valued case. By $\|\mathbf{x}\|^2 = \mathbf{x}^{\mathsf{H}}\mathbf{x}$ we denote the usual Euclidean norm in complex vector spaces, while $\|\mathbf{B}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(\mathbf{B}^{\mathsf{H}}\mathbf{B})}$ denotes the Frobenius norm of a matrix $\mathbf{B} \in \mathbb{C}^{m\times n}$, where $(\cdot)^{\mathsf{H}} = [(\cdot)^*]^\top$ denotes the conjugate transpose (or Hermitian transpose). For any $x \in \mathbb{C}$, the notation $|\cdot|$ refers to modulus, so $|x|^2 = xx^*$

We now assume that the data $\{\mathbf{x}_i\}_{i=1}^n$ is a random sample from a circular complex elliptically symmetric (CES) distribution, denoted $\mathbf{x} \sim \mathbb{C}\mathcal{E}_p(\boldsymbol{\mu}, \mathbf{\Sigma}, g)$ (*cf.* [19] for a detailed review). Similarly to the real-valued case, the probability density function of a CES distributed random vector $\mathbf{x} \in \mathbb{C}^p$ is given by

$$f(\mathbf{x}) = C_{p,g}|\mathbf{\Sigma}|^{-1} g((\mathbf{x}-\boldsymbol{\mu})^{\mathsf{H}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})),$$

where $\mathbf{\Sigma}$ denotes the positive definite Hermitian covariance matrix, $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$ is the mean vector, $g : \mathbb{R}_{\ge 0} \to \mathbb{R}_{>0}$ is the density generator, and $C_{p,g}$ is a normalizing constant. Again, we also normalize $g$ so that $\mathbf{\Sigma} = \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{H}}]$. The definitions of the scale and sphericity parameters in (10) and (11) remain unchanged. The elliptical kurtosis is however redefined as

$$\kappa = \frac{\mathbb{E}[r^4]}{p(p+1)} - 1 = \frac{1}{2}\mathrm{kurt}(x_i).$$

where the expectation is over $r = \|\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})\|$ and $\mathrm{kurt}(x_i)$ denotes the excess kurtosis of any (e.g., $i$th) marginal variable of $\mathbf{x}$, defined by

$$\mathrm{kurt}(x_i) = \frac{\mathbb{E}[|x_i-\mu_i|^4]}{\sigma_i^4} - 2,$$

where $\mu_i = \mathbb{E}[x_i]$ and $\sigma_i^2 = \mathbb{E}[|x_i-\mu_i|^2]$ denote the mean and variance of $x_i$. The theoretical lower bound of the kurtosis in the complex-valued case is $\kappa^{\mathrm{LB}} = -1/(p+1)$ [19]. Again $\kappa = 0$ if $\mathbf{x}$ has a circular complex multivariate normal distribution ($\mathbf{x} \sim \mathbb{C}\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$).

### B. Oracle $\beta$ and $k$ under CES distributions

The SCM (2) of complex-valued observations is defined by

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathsf{H}} \tag{40}$$

and the complex-valued counterpart TABASCO $\hat{\mathbf{\Sigma}}_\beta$ is still defined as in (8). Extending the previous results to this estimator will require some minor adaptations.

We start by noticing that Theorem 1 still holds for complex-valued observations. Then, the next result provides the complex-valued extension of Lemma 1.

**Lemma 2.** *Let $\{\mathbf{x}_i\}_{i=1}^n$ be an i.i.d. random sample from $\mathbb{C}\mathcal{E}_p(\boldsymbol{\mu}, \mathbf{\Sigma}, g)$ with finite 4th order moments. Then for any $\mathbf{W} \in \mathcal{W}^+$, and for the SCM as in* (40)*, it holds that*

$$\mathbb{E}\left[\|\mathbf{W}\circ\mathbf{S}\|_{\mathrm{F}}^2\right] = (1+\tau_2)\|\mathbf{W}\circ\mathbf{\Sigma}\|_{\mathrm{F}}^2 + \tau_1\,\mathrm{tr}((\mathbf{D}_{\mathbf{\Sigma}}\mathbf{W})^2)$$

*and*

$$\mathbb{E}\left[\mathrm{tr}((\mathbf{D}_{\mathbf{S}}\mathbf{W})^2)\right] = 2\tau_1\|\mathbf{W}\circ\mathbf{\Sigma}\|_{\mathrm{F}}^2 + (1+\tau_2)\,\mathrm{tr}((\mathbf{D}_{\mathbf{\Sigma}}\mathbf{W})^2),$$

*where $\mathbf{D}_{\mathbf{\Sigma}} = \mathrm{diag}(\mathbf{\Sigma})$, $\mathbf{D}_{\mathbf{S}} = \mathrm{diag}(\mathbf{S})$ and $\tau_1$ and $\tau_2$ are defined in* (26)*.*

*Proof.* The proof is postponed to Appendix E. □

This result allows us to derive the complex-valued counterpart of Theorem 2 for the optimal shrinkage parameter $\beta_o$.

**Theorem 5.** *Let $\{\mathbf{x}_i\}_{i=1}^n$ be an i.i.d. random sample from a complex elliptical distribution $\mathbb{C}\mathcal{E}_p(\boldsymbol{\mu}, \mathbf{\Sigma}, g)$ with finite 4th order moments. Then the oracle parameter $\beta_o$ in* (12) *is*

$$\beta_o = \frac{t}{t + (n/(n-1))(p\theta_{\mathbf{W}}/\eta^2 - 2\gamma/p) + \kappa\cdot A},$$

*where $t = n(\gamma_{\mathbf{V}} - 1)$, and*

$$A = p\theta_{\mathbf{W}}/\eta^2 - 1 + \gamma_{\mathbf{W}} - 2\gamma/p.$$

Form this result, the optimal index $k$ can be obtained as in subsection II-B.

### C. Data adaptive adaptations for the complex-valued case

In the complex-valued case, the TABASCO procedure of Algorithm 1 is kept identical. However, the expression of $\beta_o$ is changed in accordance with Theorem 5. The plug in estimates for the side parameters $\{\eta, \theta_{\mathbf{W}}, \kappa, \gamma, \gamma_{\mathbf{W}}, \gamma_{\mathbf{V}}\}$ are then as follows:

- For $\eta$ and $\theta_{\mathbf{W}}$, we keep the empirical estimates as in (30).
- An estimate of elliptical kurtosis $\kappa$ is calculated as in [31] as average sample excess kurtosis of the marginal variables scaled by $1/2$.
- The three sphericity statistics $\gamma$, $\gamma_{\mathbf{W}}$, and $\gamma_{\mathbf{V}}$ can still be estimated with the Ell1 and Ell2 estimators detailed in Section IV. The Ell1 estimator is identical, except that the definition of the SSCM in (32) now involves the Hermitian transpose. The Ell2 estimator is defined as earlier, with changes only in expression of $a_n$ and $b_n$. Indeed, Theorem 4 holds with $a_n$ as in (36) and $b_n$ given by

$$b_n = \frac{n(n-1)^2(\kappa+n)}{2\kappa n(n^2-4n+3) - \kappa^2(n-1)^2 + n^2(n^2-2n-1)}.$$

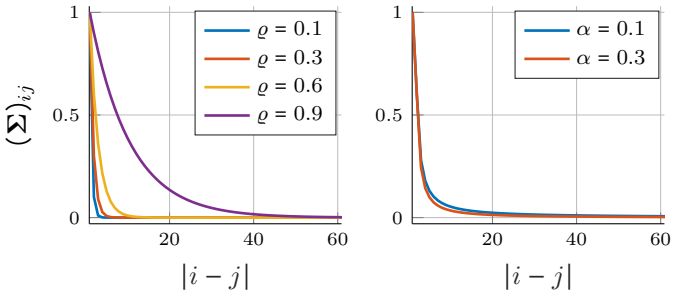


Fig. 1. $(\mathbf{\Sigma})_{ij}$ as a function of $|i-j|$. *Left:* Model 1 in (41) with various correlation parameters $\varrho$. *Right:* Model 2 in (41) with various decay parameters $\alpha$ and $\rho = 0.6$. The dimension is $p = 100$.

## VI. SIMULATION STUDIES

We generate samples from (real-valued) ES distributions with a scatter matrix $\mathbf{\Sigma}$ having a diagonally dominant structure (model 1 and model 2 detailed below). The mean $\boldsymbol{\mu}$ is generated randomly as $\mathcal{N}_p(10\cdot\mathbf{1}, \mathbf{I})$ and kept fixed for all trials, and the number of Monte-Carlo trials is 5000.

The estimators included in the study are: *i*) The Ledoit-Wolf estimator (**LWE**) [2] defined by (1) where $\beta$ is an estimate of an (oracle) MMSE parameter $\beta_o$. *ii*) The shrinkage to tapering oracle approximate (**STOA**) estimator [14] defined by (6) where $\beta$ is an estimate of the oracle parameter computed using an iterative procedure. The bandwidth $k$ is selected using a cross-validation scheme with 60%-to-40% split for training and testing. *iii*) The shrinkage to tapering (ST-)estimators in [15] defined by (6) where both $\beta$ and $k$ are estimates of the oracle MMSE parameters. The estimator **ST-gaus** assumes Gaussian data, while **ST-nong** assumes non-Gaussian data. *iv*) TABASCO (computed via Algorithm 1) using the Ell1-estimator of sphericity.

### A. Model 1

In **Model 1**, $\mathbf{\Sigma}$ possesses an auto-regressive AR(1) structure:

$$(\mathbf{\Sigma})_{ij} = \eta\varrho^{|i-j|}, \tag{41}$$

where $|\varrho| \in [0,1)$. When $\varrho \downarrow 0$, then $\mathbf{\Sigma}$ is close to an identity matrix scaled by $\eta$, and when $\varrho \uparrow 1$, $\mathbf{\Sigma}$ tends to a singular matrix of rank 1. As illustrated in Figure 1, banding matrices allow for a good approximation, so all tapering-type estimators are computed with $\mathbf{W}(k)$ as (3) in this subsection. The optimal bandwidth $\hat{k}_o$ is chosen by consider the set of tapering matrices $\mathbb{W} = \{\mathbf{W}(k) : k \in [\![1, 30]\!] \cup [\![p-30, p]\!]\}$ (this restriction is made to lower the computational cost, but identical results were obtained with $k \in [\![1, p]\!]$).

Figure 2 provides a validation of the theoretical results: it displays the theoretical normalized MSE (NMSE) curves, $L(\beta) = \mathbb{E}[\|\hat{\mathbf{\Sigma}}_\beta - \mathbf{\Sigma}\|_{\mathrm{F}}^2]/\|\mathbf{\Sigma}\|_{\mathrm{F}}^2$ as a function of shrinkage parameter $\beta$ for TABASCO estimators using a fixed bandwidths $k \in [\![1,5]\!]$ and $k = p$ (i.e., $\mathbf{W} = \mathbf{1}\mathbf{1}^\top$). In this setup, the data is generated from MVN distribution $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ with $p = 100$ and $n = 50$ (similar results were obtained for other ES distributions and dimension setups). The black bullet (•) displays the theoretical minimum NMSE in (18) attained for

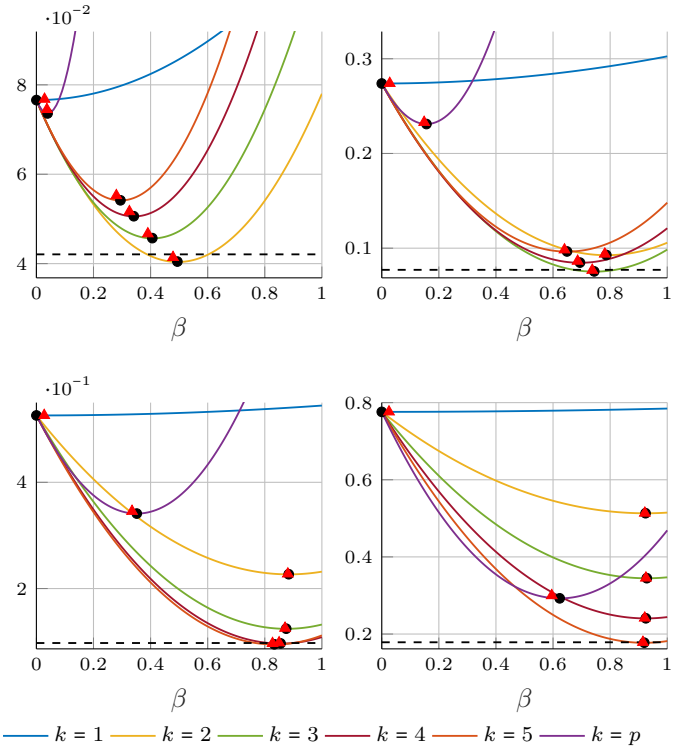


Fig. 2. Solid line display NMSE of TABASCO estimator using fixed $\mathbf{W}(k)$ as in (3) when samples are drawn from a MVN distribution, $\mathbf{\Sigma}$ as in Model 1 in (41) with $\varrho = 0.2$ (top left), $\varrho = 0.4$ (top right), $\varrho = 0.6$ (bottom left), $\varrho = 0.8$ (bottom right); $n = 50$, $p = 100$. The horizontal dashed line correspond to the empirical NMSE obtained by TABASCO of Algorithm 1.

$\beta_o \equiv \beta_o(k)$ for each bandwidth $k$. The empirical average NMSE for TABASCO using estimated $\hat{\beta}_0$ for each fixed $k$ is displayed using red triangle (▲), where the location on $\beta$ axis correspond to empirical average $\hat{\beta}_0$. As can be noted from Figure 2, TABASCO estimates the oracle shrinkage parameter $\beta_o$ very accurately since the black bullets and red triangles are mostly overlapping for each bandwidth. The dashed horizontal line shows the average NMSE obtained by TABASCO when also using an estimated optimal bandwidth $\hat{k}_0$. One can notice that the optimal bandwidth selection using (31) is also accurate. For example, in the case of $\varrho = 0.4$, the optimal bandwidth is $k = 3$ and TABASCO estimator attains an average NMSE that is very close to the theoretical minimum NMSE.

Figure 3 displays the NMSE as a function of bandwidth $k$. As can be noted, the empirical average NMSE of TABASCO using fixed bandwidth displayed using triangle (e.g., ▲) coincides very accurately with the theoretical NMSE. The NMSE of TABASCO with estimated bandwidth is shown using colored star (e.g., ☆) wherein the location on $k$-axis correspond to empirical average $\hat{k}_0$. As can be noted, the optimal bandwidth $k_o$ is also very accurately estimated since $\hat{k}_o \approx k_o$ for each $\varrho$.

Figure 4 compares the performance of TABASCO with the state of the art in various setups. The upper panel displays the NMSE curves as a function of the sample size $n$ for four choices of correlation parameter $\varrho$ when the data follows a MVN distribution. The lower panel displays the same results when the data follows a MVT distribution with $\nu = 5$, which is heavy-tailed with marginal kurtosis $\mathrm{kurt}(x_i) = 6$ and elliptical kurtosis $\kappa = \mathrm{kurt}(x_i)/3 = 2$. In the Gaussian

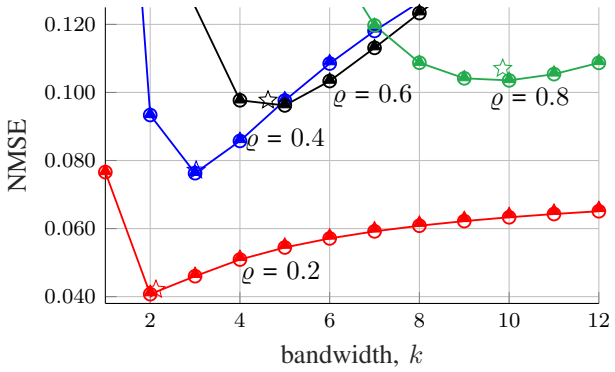


Fig. 3. Oracle NMSE of TABASCO estimator with fixed bandwidth $k$ (⊖) when $\mathbf{\Sigma}$ has an AR(1) structure and $\varrho \in \{0.2, 0.4, 0.6, 0.8\}$. The average NMSE of estimated TABASCO using fixed $k$ and estimated $k$ over 5000 MC trials is superimposed to the curves using symbols ▲ and ☆, respectively. Samples are drawn from a MVN distribution, $n = 50$ and $p = 100$. Banding matrices $\mathbf{W}(k)$ as in (3) are used.

case, all banding-type estimators outperform LWE thanks to the exploitation of the diagonally dominant structure of the covariance matrix. In the heavy-tailed case, this is no longer true for STOA and ST-gaus, while ST-nong and TABASCO remain robust. In all scenarios, TABASCO offers the lowest NMSE, and especially improves the performance when $n \ll p$.

Figure 5 displays the obtained (average) estimated shrinkage parameter $\hat{\beta}_o$ of TABASCO and LWE as a function of $n$. The average shrinkage parameter of TABASCO is generally much larger than that of LWE. This means that it assign overall more weight on the banded SCM $\mathbf{W} \circ \mathbf{S}$ compared to LWE, which uses $\mathbf{W}(p) = \mathbf{1}\mathbf{1}^\top$. This behavior is expected since banding the SCM should naturally improve the MSE when the true covariance matrix has a diagonally dominant structure.

Figure 6 presents a comparison similar to Figure 4 when the variables are permuted at random for each Monte Carlo trial, thus destroying the diagonally dominant structure of the AR(1) covariance matrix[3]. The hypothesis is that any banding estimator with optimal bandwidth selection should be able to select the bandwidth $k = p$ accordingly. Note that LWE is invariant to variable permutations, and hence its results stays the same for both of these scenarios. In this setup, TABASCO performs better that LWE for $n \ll p$ and equally well as LWE for $n$ large enough. This result implies that bandwidth selection of TABASCO is consistent: it chooses $k = p$ since the true covariance matrix does not have a diagonally dominant structure. The improvement brought at low sample support can be explained by the fact that an ES distribution is assumed by TABASCO, which allows for a better estimation of the oracle parameter (LWE only assumes finite 4th order moments). This example confirms that TABASCO always benefits from banding and bandwidth selection: it offers significantly improved NMSE compared to RSCM when banding structure is present in the covariance matrix, while it does not perform worse when such structure does not exist, thanks to its robust and efficient bandwidth selection.

[3]Prominent algorithms for recovering hidden ordering-structure in the variables are the Best Permutation Analysis (BPA) [32] or Isoband [33]. The perspective of their joint use with TABASCO is left for further studies.

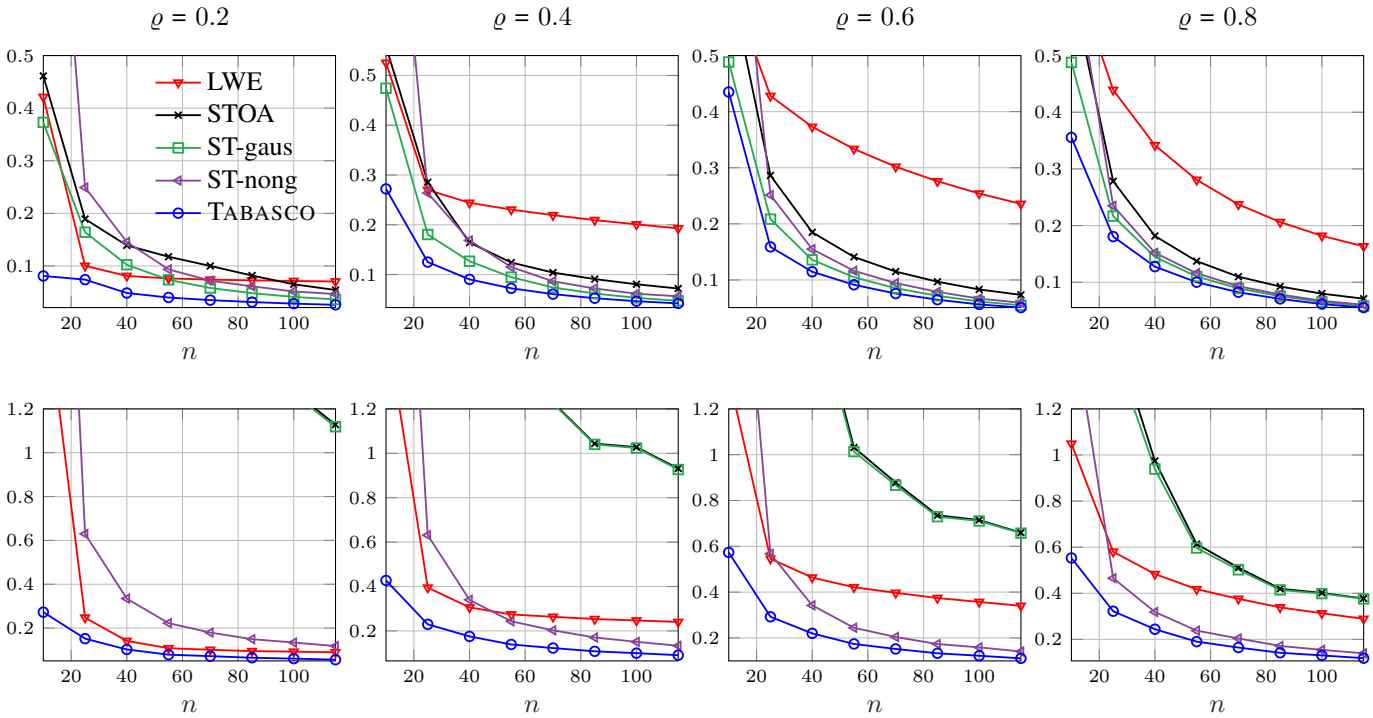


Fig. 4. Average NMSE curves when samples are from a MVN distribution (upper panel) and $t$-distribution with $\nu = 5$ d.o.f. (lower panel), $\mathbf{\Sigma}$ has an AR(1) structure with $\varrho \in \{0.2, 0.4, 0.6, 0.8\}$ from left to right. Dimension is $p = 100$ and banding matrices are used in STOA, ST-gaus, ST-nong and TABASCO.

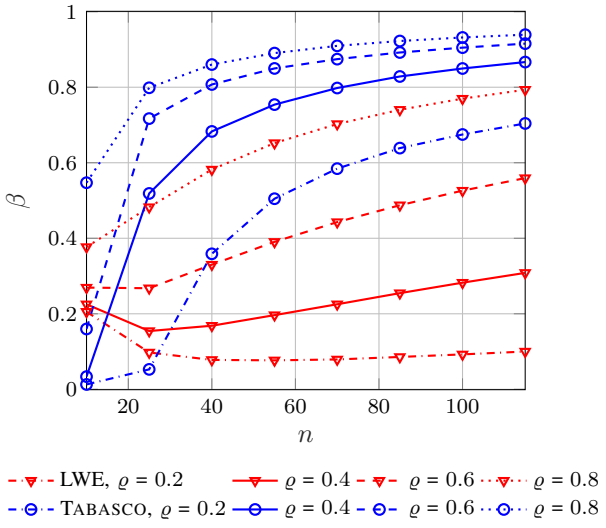


Fig. 5. Average estimated shrinkage parameter $\beta$ for LWE and TABASCO when samples are from a MVN distribution, $\mathbf{\Sigma}$ has an AR(1) structure ($\varrho \in \{0.2, 0.4, 0.6, 0.8\}$) and $p = 100$. Banding matrices are used in TABASCO.

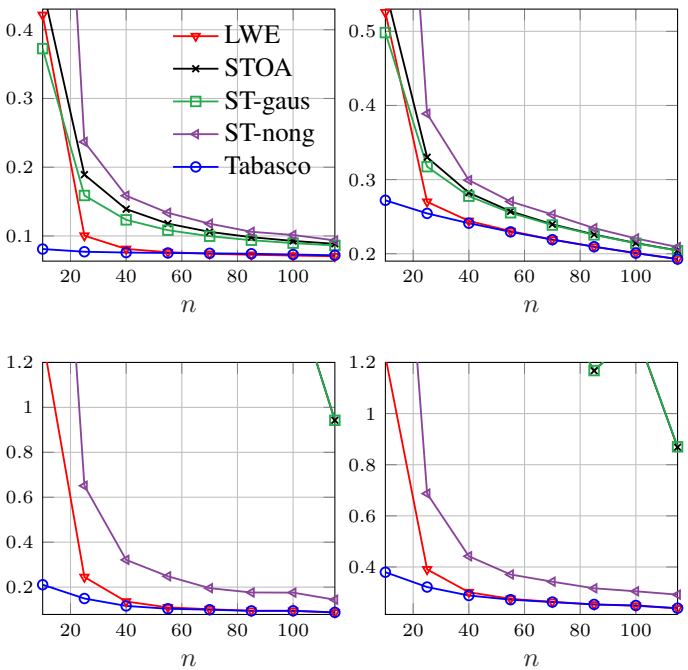


Fig. 6. Average NMSE curves when samples are from a MVN distribution (top row) and MVT distribution (bottom row) with $\nu = 5$ d.o.f., $\mathbf{\Sigma}$ has a permuted AR(1) structure with $\varrho = 0.2$ (left panel) and $\varrho = 0.4$ (right panel), and dimension is $p = 100$.

### *B. Model 2*

In **Model 2** [7], $\mathbf{\Sigma}$ is defined by

$$(\mathbf{\Sigma})_{ij} = \begin{cases} 1 & , i = j \\ \rho |i-j|^{-(\alpha+1)} & , i \neq j, \end{cases} \tag{42}$$

where $\alpha$ is a decay parameter and $\rho$ is a correlation parameter. As in the study of [7], we set $\rho = 0.6$, and Figure 1 illustrates the effect of decay parameter $\alpha$ in the case of $p = 100$.

Figure 7 presents a comparison similar to Figure 4 where we also included the minimax risk tapering (**MnMx-Taper**) estimator $\mathbf{W}(k^*) \circ \mathbf{S}$, where $k^* = \lfloor n^{1/(2(\alpha+1))} \rfloor$ is the optimal

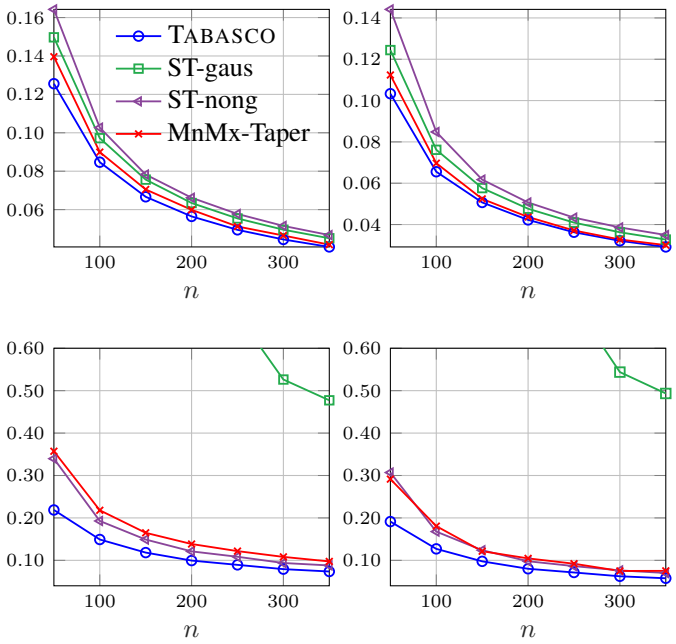


Fig. 7. Average NMSE curves when samples are from a MVN distribution (upper panel) and MVT distribution with $\nu = 5$ d.o.f. (lower panel), $\mathbf{\Sigma}$ follows model 2 with $\alpha = 0.1$ (left panel) and $\alpha = 0.3$ (right panel), $p = 250$.

(oracle) bandwidth [7, Section 6]. The dimension is $p = 250$. It should be noted that MnMx-Taper has advantage over the other estimators since it uses the true decay parameter $\alpha$, which is unknown in practice. TABASCO also uses tapering matrices $\mathbf{W}(k)$ as in (4), but ST-gaus and ST-nong are restricted to tapering matrices whose off-diagonal elements are 0-s or 1-s. Hence, these are still computed with banding matrices $\mathbf{W}(k)$ as in (3). In either case, the optimal bandwidth $\hat{k}_o$ is chosen by consider the set of tapering matrices $\mathbb{W} = \{\mathbf{W}(k) : k \in [\![1, 30]\!] \cup [\![p-30, p]\!]\}$. As can be noted, TABASCO again outperforms other estimators for all values of $n$ and $\alpha$ and for both sampling distributions. In the MVN case (top panel), TABASCO outperforms MnMx-Taper with a clear margin when $n$ is very small. This can be attributed to its ability to optimally shrink the tapered SCM towards a scaled identity matrix when $n/p < 1$. However for $n \geq p$, TABASCO and MnMx-Taper estimator have similar performance, especially when $\alpha = 0.3$.

In the MVT case (lower panel of Figure 7), the performance differences are more clear. TABASCO outperforms MnMx-taper by a large margin. ST-gaus estimator completely fails due to the impulsive nature of the underlying sampling distributions. The results also illustrate that the performance of tapered SCM estimator is dependent on the underlying sampling distribution more heavily than TABASCO. This is illustrated further in Figure 8 where we compare the true theoretical NMSE curves of tapered SCM $\mathbf{W} \circ \mathbf{S}$ and TABASCO estimator $\hat{\mathbf{\Sigma}}_{\beta_0}$ as a function of bandwidth $k$ in the case where $n = 100$ and when sampling from a MVN distribution (left panel) and MVT distribution (right panel) with $\nu = 5$ d.o.f. following model 2 with $\alpha = 0.1$. Figure 8 shows two important points. First, the performance differences between the tapered SCM and TABASCO are larger when the distribution is heavier tailed,

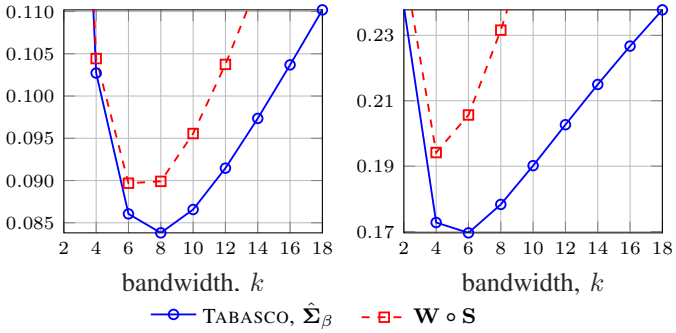


Fig. 8. The true (theoretical) NMSE curves as a function of bandwidth $k$ for the tapered SCM $\mathbf{W} \circ \mathbf{S}$ and TABASCO $\hat{\mathbf{\Sigma}}_\beta$ when sampling from a MVN distribution (left panel) and MVT distribution (right panel) with $\nu = 5$ d.o.f., $\mathbf{\Sigma}$ follows model 2 with $\alpha = 0.1$, $n = 100$ and $p = 250$.

which was already evident in Figure 7. Second, TABASCO with optimal bandwidth selection is able to estimate the optimal bandwidth rather accurately since the average (empirical) NMSE value seen in Figure 7 at $n = 100$ is close to the minimum true (theoretical) NMSE value.

## VII. APPLICATION TO SPACE-TIME ADAPTIVE PROCESSING

Space time adaptive processing (STAP) is a technique used in airborne phased array radar to detect moving target embedded in an interference background such as jamming or strong clutter [34]. The radar receiver consists in an array of $Q$ antenna elements processing $P$ pulses in a coherent processing interval. Within the tested sample $\mathbf{x}_0 \in \mathbb{C}^p$ with $p = P \cdot Q$, the received signal is composed of $i$) possible unknown targets responses; $ii$) unknown interferences (ground clutter) plus thermal noise. A detection problem for a given steering vector $\mathbf{p}$ is classically formalized as a binary hypothesis test: under $H_0$, $\mathbf{x}_0$ only contains the interference plus noise, or under $H_1$, $\mathbf{x}_0$ additionally contains a scaled observation of $\mathbf{p}$, i.e.:

$$\begin{cases} H_0: & \mathbf{x}_0 = \mathbf{n}_0 & ; \mathbf{x}_i = \mathbf{n}_i, \forall\, i \in [\![1, n]\!] \\ H_1: & \mathbf{x}_0 = \alpha \mathbf{p} + \mathbf{n}_0 & ; \mathbf{x}_i = \mathbf{n}_i, \forall\, i \in [\![1, n]\!] \end{cases}$$

where $\mathbf{x}_i \in \mathbb{C}^p$, $i = 1, \ldots, n$ is a secondary data set, assumed to contain i.i.d. and target-free realizations of the interference plus noise. Usually, this disturbance $\mathbf{n}_i$ is modeled as centered complex Gaussian (or elliptically) distributed with covariance matrix $\mathbf{\Sigma}$. In this context, efficient adaptive detection statistics can be built from the expression of the adaptive coherence estimator (ACE) detector [35]:

$$\hat{\Lambda}(\hat{\mathbf{\Sigma}}) = \frac{|\mathbf{p}^{\mathsf{H}} \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_0|^2}{|\mathbf{p}^{\mathsf{H}} \hat{\mathbf{\Sigma}}^{-1} \mathbf{p}|\, |\mathbf{x}_0^{\mathsf{H}} \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_0|} \underset{H_0}{\overset{H_1}{\gtrless}} \delta_{\hat{\mathbf{\Sigma}}}, \tag{43}$$

where $\hat{\mathbf{\Sigma}}$ is a *plug-in* estimate of $\mathbf{\Sigma}$ computed from $\{\mathbf{x}_i\}_{i=1}^n$. More specifically in STAP, the target $\mathbf{p}$ follows the steering vector model of [34], which is function of the target angle of arrival (AoA) $\theta$ and velocity $v$. The statistic (43) can thus be computed for a dictionary of steering vectors covering a 2D-grid on $\theta$ and $v$, yielding an adaptive detection map.

Using the SCM as estimate in (43) yields a generalized likelihood ratio test (GLRT) [36]. However, plug-in detectors can benefit from refined estimation processes in order to improve

robustness, or to deal with limited sample support issues. For example shrinkage to identity (also referred to as diagonal loading or robust beamforming [37]) is a common procedure to improve several properties of the detector's output. In the context of interference cancellation, tapering templates have been considered as a spectrum notch-widening technique [11], or to deal with modulation effects [8].

This section presents an experimental validation of TABASCO to illustrate the interest of both approach on real data. The STAP data is provided by the French agency DGA/MI: the clutter is real but the targets were synthetically added in the dataset conception. The number of sensors is $Q = 4$ and the number of coherent pulses is $P = 64$, the size of the data is then $p = QP = 256$. The center frequency and the bandwidth are respectively equal to $f_0 = 10$GHz and the bandwidth $B = 5$MHz. The radar celerity is $V = 100$m/s. The inter-element spacing is $d = 0.3$m and the pulse repetition frequency is $f_r = 1$kHz. The clutter to noise ratio is evaluated around 20dB. We consider two different scenarios: one where the tested cell is under $H_1$ with a single target at ($\theta = 0$, $v = 4$m/s), and one where the tested cell is under $H_1$ with 10 targets at various speed/angle. The signal to clutter ratio of each target was estimated to be around $-5$dB. In both scenarios, $n = 397$ (all available) target-free secondary data are used to estimate the interference covariance matrix.

The tapering matrix is constructed as proposed in [11][4], i.e.,

$$\begin{aligned} \mathbf{W}(k) &= \mathbf{T}_f \otimes \mathbf{T}_\theta \\ \left[\mathbf{T}_f\right]_{ij} &= (1 + \operatorname{sinc}((i-j)k/\pi))/2 \in \mathbb{R}^{P\times P} \\ \left[\mathbf{T}_\theta\right]_{ij} &= (1 + \operatorname{sinc}((i-j)k/\pi))/2 \in \mathbb{R}^{Q\times Q} \end{aligned} \tag{44}$$

Note that index $k$ is here a "null-spectrum width" parameter in $\mathbb{R}^+$ and not a bandwidth parameter in $[\![1, p]\!]$ as in (3) or (4). We also point out that the banding-type tapering matrices (even involving a Kronecker-product structure) were tested but appeared not well suited to the data, nor beneficial to the detection process. Thus they will not be discussed in the following.

Figure 9 presents the detection map of $\hat{\Lambda}(\hat{\mathbf{\Sigma}})$ constructed with: $i$) the SCM; $ii$) the tapered SCM $\mathbf{W}(k) \circ \mathbf{S}$ using bandwidth $k = 0.05$ (selected manually to obtain the best visual results); $iii$) TABASCO with the proposed adaptive selection of $\beta$ for $k = 0$ (equivalent to RSCM, yielding $\beta = 0.9324$); $iv$) TABASCO with the proposed adaptive estimation of $\beta$ and $k$ allowing $k \in [10^{-3}, 10^{-1}]$ (TABASCO, yielding $k = 0.0143$ and $\beta = 0.9929$). ST-type estimators presented results visually identical to the tapered-SCM so they are omitted. This result can be explained because the oracle shrinkage coefficient from ST tends, in practice, to push the estimate towards the tapered matrix only: for example, the oracle coefficient $(1-\beta)$ in [14] is always greater than 0.9 and is most likely close to 1.

[4]The tapering in [11] actually uses $\left[\mathbf{T}_f\right]_{ij} = \operatorname{sinc}((i-j)k/\pi)$ and $[\mathbf{T}_\theta]_{ij} = \operatorname{sinc}((i-j)k/\pi)$, which performs a sliding window average on the estimated signal spectrum. The one considered here performs a linear combination of the original spectrum with such average. This modification was made so that the tapering matrix always conforms to the theoretical requirements $w_{ii} = 1$ and $w_{ij} \geq 0$, but did not significantly impacted the output of the tested detectors.

First we can notice that the SCM provides an unreliable detection maps in both scenarios, which is due to insufficient sample support in this configuration ($n < 2p$). As observed in [11] on another dataset, the covariance matrix tapering can widen the clutter notch, i.e., properly cancel the response of the detector on the anti-diagonal of the detection map. This permits to clearly distinguish several targets when compared to the detection map of the SCM. However, this improvement is at the cost of canceling the response of slower targets in the second scenario. This was to be expected since slow targets are hard to distinguish from the ground response in practice. Within the considered framework, hard cancellation of the ground clutter generally implies to also cancel these targets, which implies a trade-off when selecting the null-spectrum width of the tapering matrix. The shrinkage to identity of RSCM also greatly improves the detection process, as it allows us to detect all the 10 targets in the second scenario. However, it still presents some high false alarms on the clutter ridge. Finally, TABASCO appears as an interesting trade-off by combining the two effects, and illustrates that the proposed NMSE-driven method still allows for a reasonable regularization parameters (both $\beta$ and $k$) selection in this detection application. More precisely, TABASCO yields the best detection map in the first scenario (no null-spectrum pattern and no false alarms around the clutter ridge). In the second scenario, slower targets are still canceled but TABASCO allows for recovering some targets that were initially canceled by a "tapering only" approach.

## VIII. CONCLUSIONS AND PERSPECTIVES

We proposed TABASCO: a new covariance matrix estimator that jointly benefits from shrinkage to a scaled identity matrix and tapering of the SCM. By assuming the samples to be generated from an unspecified ES distribution, we also derived an efficient and robust estimation method for the oracle regularization parameters that minimize the MSE. Simulations studies illustrated that TABASCO outperforms existing regularized and tapered estimators in numerous setups. Interestingly, if $\mathbf{W} = \mathbf{1}\mathbf{1}^\top$ belongs to the set of tapering matrices $\mathbb{W}$ considered, the estimator can avoid applying tapering if this option does not provide reduction to the MSE. Thus TABASCO performs similarly to the regularized SCM proposed in [4] in this case, while significantly outperforming it when the tapering templates are valid. We also proposed two new novel estimators that measure the sphericity of the tapered covariance matrix.

## APPENDIX

### A. Known location $\boldsymbol{\mu}$

In some applications, the mean vector $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$ is known and assumed to be $\boldsymbol{\mu} = \mathbf{0}$ without loss of generality. This is the case in STAP application for example. In this case, the covariance matrix $\mathbf{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ is estimated by the SCM, defined by

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top. \tag{45}$$

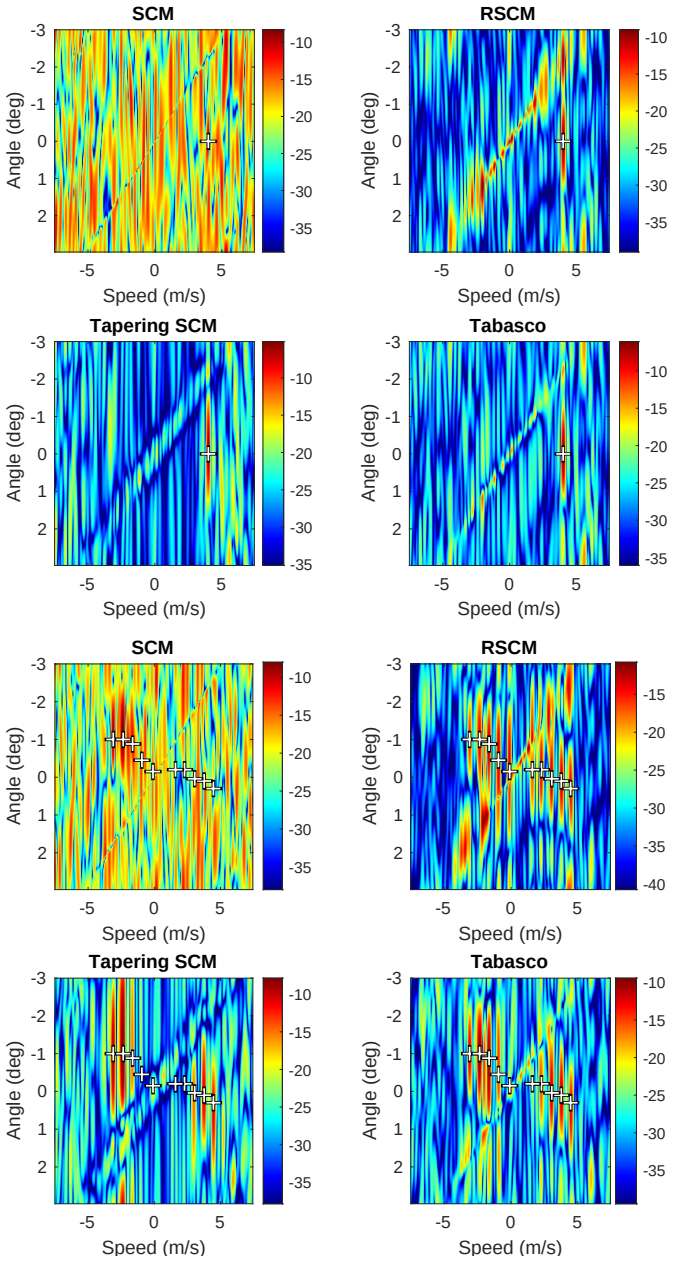


Fig. 9. Output of various STAP detectors for the first (top) and second (bottom) scenarios

which is also unbiased estimator of $\mathbf{\Sigma}$, i.e., $\mathbb{E}[\mathbf{S}] = \mathbf{\Sigma}$. The known location case implies only small changes in our estimation procedure since Theorem 1 holds for both known and unknown location cases.

When the location is known, the expectation $\mathbb{E}\big[\|\mathbf{S}\|_{\mathrm{F}}^2\big]$ and $\mathbb{E}\Big[\mathrm{tr}(\mathbf{S})^2\Big]$ are of the form (24) and (25) with $\tau_1$ and $\tau_2$ given by

$$\tau_1 = \frac{1+\kappa}{n} \qquad \text{and} \qquad \tau_2 = \frac{\kappa}{n}. \tag{46}$$

This result follows as a special case of [31, Lemma 1] for a Gaussian weight function. Similarly Lemma 1 holds when using $\tau_1$ and $\tau_2$ in (46). The change to the optimal $\beta_0$ parameter is also minimal: one may ignore the term $(n/(n-1))$ that appears as the multiplier of the 2nd last term $p\theta_{\mathbf{W}}/\eta^2 + \gamma_{\mathbf{W}} - 2\gamma/p$ in the denominator of $\beta_0$ in Theorem 2. Theorem 4 also holds with

$$a_n = \frac{1+\kappa}{n+\kappa} \qquad \text{and} \qquad b_n = \frac{n(n+\kappa)}{(n-1)(n+2+3\kappa)}.$$

### B. Proof of Theorem 1

Write $L(\beta) = \mathrm{MSE}(\hat{\mathbf{\Sigma}}_\beta) = \mathbb{E}[\|\hat{\mathbf{\Sigma}}_\beta - \mathbf{\Sigma}\|_{\mathrm{F}}^2]$. Then note that

$$\begin{aligned} L(\beta) &= \mathbb{E}\Big[\Big\|\beta(\mathbf{W}\circ\mathbf{S}) + (1-\beta)p^{-1}\,\mathrm{tr}(\mathbf{S})\mathbf{I} - \mathbf{\Sigma}\Big\|_{\mathrm{F}}^2\Big] \\ &= \mathbb{E}\Big[\Big\|\beta(\mathbf{W}\circ\mathbf{S}-\mathbf{\Sigma}) + (1-\beta)\big(p^{-1}\,\mathrm{tr}(\mathbf{S})\mathbf{I} - \mathbf{\Sigma}\big)\Big\|_{\mathrm{F}}^2\Big] \\ &= \beta^2 a_1 + (1-\beta)^2 a_2 + 2\beta(1-\beta)a_3 \end{aligned} \tag{47}$$

where the constants $a_i$-s are defined by

$$\begin{aligned} a_1 &= \mathrm{MSE}(\mathbf{W}\circ\mathbf{S}), \\ a_2 &= \mathbb{E}\Big[\Big\|p^{-1}\,\mathrm{tr}(\mathbf{S})\mathbf{I} - \mathbf{\Sigma}\Big\|_{\mathrm{F}}^2\Big], \\ a_3 &= \mathbb{E}\big[\,\mathrm{tr}((\mathbf{W}\circ\mathbf{S}-\mathbf{\Sigma})(p^{-1}\,\mathrm{tr}(\mathbf{S})\mathbf{I}-\mathbf{\Sigma}))\big]. \end{aligned}$$

Then define

$$\begin{aligned} \tilde{a}_3 &= p^{-1}\mathbb{E}\left[\mathrm{tr}\left(\mathbf{W}\circ\mathbf{S}-\mathbf{\Sigma}\right)\mathrm{tr}(\mathbf{S})\right] \\ &= p^{-1}\mathbb{E}\big[\,\mathrm{tr}(\mathbf{S})^2\big] - \eta^2 p \end{aligned} \tag{48}$$

where we used that $\mathbb{E}[\mathbf{S}] = \mathbf{\Sigma}$, $\mathrm{tr}(\mathbf{W}\circ\mathbf{S}) = \mathrm{tr}(\mathbf{S})$ and $\mathbb{E}[\mathrm{tr}(\mathbf{S})] = \mathrm{tr}(\mathbb{E}[\mathbf{S}]) = \eta p$ with $\eta = \mathrm{tr}(\mathbf{\Sigma})/p$ denoting the scale. We may write $a_3$ in the form

$$\begin{aligned} a_3 &= \tilde{a}_3 + \mathbb{E}\big[\,\mathrm{tr}((\mathbf{W}\circ\mathbf{S}-\mathbf{\Sigma})\mathbf{\Sigma})\big] \\ &= \tilde{a}_3 + \|\mathbf{\Sigma}\|_{\mathrm{F}}^2 - \|\mathbf{V}\circ\mathbf{\Sigma}\|_{\mathrm{F}}^2 \end{aligned} \tag{49}$$

by using $\mathbb{E}[\mathbf{S}] = \mathbf{\Sigma}$ and $\mathbb{E}[\mathrm{tr}((\mathbf{W}\circ\mathbf{S})\mathbf{\Sigma})] = \mathrm{tr}((\mathbf{W}\circ\mathbf{\Sigma})\mathbf{\Sigma}) = \mathrm{tr}((\mathbf{V}\circ\mathbf{\Sigma})^2)$ while $a_2$ can be expressed as

$$a_2 = \tilde{a}_3 + \|\mathbf{\Sigma}\|_{\mathrm{F}}^2 - p\eta^2. \tag{50}$$

Note that $L(\beta)$ is a convex quadratic function in $\beta$ with a unique minimum given by

$$\beta_o = \frac{a_2 - a_3}{(a_1 - a_3) + (a_2 - a_3)}. \tag{51}$$

Using (49) and (50), the numerator is

$$\begin{aligned} a_2 - a_3 &= \|\mathbf{V}\circ\mathbf{\Sigma}\|_{\mathrm{F}}^2 - p\eta^2 \\ &= \|\mathbf{V}\circ\mathbf{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2 = p(\gamma_{\mathbf{V}} - 1)\eta^2, \end{aligned} \tag{52}$$

where we used that $\mathrm{tr}(\mathbf{V}\circ\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{\Sigma})$. Using the expression for $a_1 = \mathrm{MSE}(\mathbf{W}\circ\mathbf{S})$ given in (13) together with equations (49) and (48), the first term in the denominator of $\beta_o$ is

$$a_1 - a_3 = \mathbb{E}\big[\|\mathbf{W}\circ\mathbf{S}\|_{\mathrm{F}}^2\big] - \|\mathbf{V}\circ\mathbf{\Sigma}\|_{\mathrm{F}}^2 - p^{-1}\mathbb{E}\big[\,\mathrm{tr}(\mathbf{S})^2\big] + \eta^2 p.$$

Summing this with term $a_2 - a_3$ from (52) shows that the denominator of $\beta_o$ is $\mathbb{E}[\|\mathbf{W}\circ\mathbf{S}\|_{\mathrm{F}}^2] - p^{-1}\mathbb{E}[\mathrm{tr}(\mathbf{S})^2]$. These results thus yield expression given in (15) and (16) for $\beta_o$. The final expression (17) for $\beta_o$ can be deduced from (16) by using (13) and then simplifying the expression.

The expression for MSE of $\hat{\mathbf{\Sigma}}_{\beta_o}$ follows by substituting $\beta_o$ into expression for $L(\beta)$ in (47) and using the relation,

$(1-\beta_o)(a_2-a_3)=\beta_o(a_1-a_3)$, which follows from (51). This gives

$$L(\beta_o)=a_2-\beta_o(a_2-a_3)=a_3+(1-\beta_o)(a_2-a_3).$$

This gives the stated MSE expression since $a_2-a_3=\|\mathbf{V}\circ\mathbf{\Sigma}-\eta\mathbf{I}\|_{\mathrm{F}}^2$ and using (49) together with (48) for $a_3$.

### C. Proof of Lemma 1

Before proceeding with the proof we introduce some definitions and results that are used in the sequel. First, we let $\mathbf{K}_p$ denote the $p^2\times p^2$ commutation matrix defined as a block matrix whose $ij$th block is equal to a $p\times p$ matrix that has a 1 at element $ji$ and zeros elsewhere, i.e., $\mathbf{K}_p=\sum_{i,j}\mathbf{e}_i\mathbf{e}_j^{\top}\otimes\mathbf{e}_j\mathbf{e}_i^{\top}$. It also has the following important properties [38]: $\mathbf{K}_p\mathrm{vec}(\mathbf{A})=\mathrm{vec}(\mathbf{A}^{\top})$ and $\mathbf{K}_p(\mathbf{A}\otimes\mathbf{B})\mathbf{K}_p=(\mathbf{B}\otimes\mathbf{A})$ for any $p\times p$ matrices $\mathbf{A}$ and $\mathbf{B}$, where $\mathrm{vec}(\mathbf{A})$ vectorizes matrix $\mathbf{A}$ by stacking the columns of the matrix on top of each other. We then have the following identities.

**Lemma 3.** *The following holds:*

*a)* $\|\mathbf{A}\circ\mathbf{B}\|_{\mathrm{F}}^2=\mathrm{tr}\big(\mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{A})^{\top}\circ\mathrm{vec}(\mathbf{B})\mathrm{vec}(\mathbf{B})^{\top}\big)$ *for all* $\mathbf{A},\mathbf{B}\in\mathbb{R}^{m\times n}$.

*b)* $\|\mathbf{A}\circ\mathbf{B}\|_{\mathrm{F}}^2=\mathrm{tr}\big(\mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{A})^{\top}\circ\mathbf{K}_p(\mathbf{B}\otimes\mathbf{B})\big)\ \forall\mathbf{A}\in\mathbb{R}^{m\times m}$ *and* $\forall\mathbf{B}\in\mathbb{R}_{\mathrm{Sym}}^{m\times m}$.

*c)* $\mathbf{d}_{\mathbf{B}}^{\top}(\mathbf{A}\circ\mathbf{A})\mathbf{d}_{\mathbf{B}}=\mathrm{tr}\big(\mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{A})^{\top}\circ(\mathbf{B}\otimes\mathbf{B})\big)$ $\forall\mathbf{A},\mathbf{B}\in\mathbb{R}^{m\times m}$.

*d)* $\mathrm{tr}\big((\mathbf{D}_{\mathbf{B}}\mathbf{A})^2\big)=\mathbf{d}_{\mathbf{B}}^{\top}(\mathbf{A}\circ\mathbf{A})\mathbf{d}_{\mathbf{B}}$ *for all* $\mathbf{A}\in\mathbb{R}_{\mathrm{Sym}}^{m\times m}$ *and* $\mathbf{B}\in\mathbb{R}^{m\times m}$.

*Proof.* Let $\mathbf{A}=(a_{ij})$ and $\mathbf{B}=(b_{ij})$. *a)* First note that

$$\begin{aligned}\|\mathbf{A}\circ\mathbf{B}\|_{\mathrm{F}}^2&=\mathrm{tr}\big(\mathrm{vec}(\mathbf{A}\circ\mathbf{B})\mathrm{vec}(\mathbf{A}\circ\mathbf{B})^{\top}\big)\\&=\mathrm{tr}\big(\mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{A})^{\top}\circ\mathrm{vec}(\mathbf{B})\mathrm{vec}(\mathbf{B})^{\top}\big).\end{aligned}$$

*b)* It is a simple matter to verify that for all $\mathbf{B}\in\mathbb{R}_{\mathrm{Sym}}^{m\times m}$ it holds that $\mathrm{diag}(\mathrm{vec}(\mathbf{B})\mathrm{vec}(\mathbf{B})^{\top})=\mathrm{diag}(\mathbf{K}_p(\mathbf{B}\otimes\mathbf{B}))$. Thus

$$\begin{aligned}&\mathrm{tr}\big(\mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{A})^{\top}\circ\mathbf{K}_p(\mathbf{B}\otimes\mathbf{B})\big)\\&\quad=\mathrm{tr}\big(\mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{A})^{\top}\circ\mathrm{vec}(\mathbf{B})\mathrm{vec}(\mathbf{B})^{\top}\big)\end{aligned}$$

which gives the stated result due to a)-part. *c)* It is a simple task to verify that the trace of the Hadamard product of $\mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{A})^{\top}$ with $\mathbf{B}\otimes\mathbf{B}$ equals $\sum_{i,j}b_{ii}a_{ij}^2b_{jj}$ which is equivalent with $\mathbf{d}_{\mathbf{B}}^{\top}(\mathbf{A}\circ\mathbf{A})\mathbf{d}_{\mathbf{B}}$. *d)* Follows from [21, Lemma 7.5.2]. □

Write $\mathbf{w}=\mathrm{vec}(\mathbf{W})$. Using Lemma 3a) we first notice that

$$\mathbb{E}\Big[\|\mathbf{W}\circ\mathbf{S}\|_{\mathrm{F}}^2\Big]=\mathrm{tr}\Big(\mathbf{w}\mathbf{w}^{\top}\circ\mathbb{E}[\mathrm{vec}(\mathbf{S})\mathrm{vec}(\mathbf{S})^{\top}]\Big).\quad(53)$$

We then recall that the (variance-)covariance matrix of $\mathbf{S}$ when sampling from an elliptical population $\mathcal{E}_p(\boldsymbol{\mu},\mathbf{\Sigma},g)$ is given by [4, Theorem 2]:

$$\begin{aligned}\mathrm{cov}(\mathrm{vec}(\mathbf{S}))&=\mathbb{E}[\mathrm{vec}(\mathbf{S})\mathrm{vec}(\mathbf{S})^{\top}]-\mathrm{vec}(\mathbf{\Sigma})\mathrm{vec}(\mathbf{\Sigma})^{\top}\quad(54)\\&=\tau_1(\mathbf{I}+\mathbf{K}_p)(\mathbf{\Sigma}\otimes\mathbf{\Sigma})+\tau_2\mathrm{vec}(\mathbf{\Sigma})\mathrm{vec}(\mathbf{\Sigma})^{\top},\quad(55)\end{aligned}$$

where $\tau_1$ and $\tau_2$ are constants defined in (26). Equations (54) and (55) then imply that

$$\begin{aligned}\mathbb{E}[\mathrm{vec}(\mathbf{S})\mathrm{vec}(\mathbf{S})^{\top}]&=\mathrm{cov}(\mathrm{vec}(\mathbf{S}))+\mathrm{vec}(\mathbf{\Sigma})\mathrm{vec}(\mathbf{\Sigma})^{\top}\\&=\tau_1(\mathbf{I}+\mathbf{K}_p)(\mathbf{\Sigma}\otimes\mathbf{\Sigma})+(1+\tau_2)\mathrm{vec}(\mathbf{\Sigma})\mathrm{vec}(\mathbf{\Sigma})^{\top}.\quad(56)\end{aligned}$$

Inserting (56) into (53) yields

$$\mathbb{E}\Big[\|\mathbf{W}\circ\mathbf{S}\|_{\mathrm{F}}^2\Big]=(1+\tau_1+\tau_2)\|\mathbf{W}\circ\mathbf{\Sigma}\|_{\mathrm{F}}^2+\tau_1\,\mathrm{tr}\big((\mathbf{D}_{\mathbf{\Sigma}}\mathbf{W})^2\big).$$

simply by invoking identities in Lemma 3. This proves the first identity.

Next we note that

$$\mathbb{E}\big[\,\mathrm{tr}((\mathbf{D}_{\mathbf{S}}\mathbf{W})^2)\big]=\sum_{i,j=1}^{p}\mathbb{E}[s_{ii}s_{jj}]w_{ij}^2.\quad(57)$$

Equation (56) implies that

$$\mathbb{E}[s_{ii}s_{jj}]=2\tau_1\sigma_{ij}^2+(1+\tau_2)\sigma_i^2\sigma_j^2.\quad(58)$$

Thus inserting (58) into (57) yields

$$\begin{aligned}\mathbb{E}\big[\,\mathrm{tr}((\mathbf{D}_{\mathbf{S}}\mathbf{W})^2)\big]&=2\tau_1\sum_{i,j=1}^{p}w_{ij}^2\sigma_{ij}^2+(1+\tau_2)\sum_{i,j=1}^{p}w_{ij}^2\sigma_i^2\sigma_j^2\\&=2\tau_1\|\mathbf{W}\circ\mathbf{\Sigma}\|_{\mathrm{F}}^2+(1+\tau_2)\,\mathrm{tr}((\mathbf{D}_{\mathbf{\Sigma}}\mathbf{W})^2)\end{aligned}$$

which proves the latter claim.

### D. Proof of Theorem 3

Let us express the SSCM as

$$\hat{\mathbf{\Lambda}}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i\mathbf{v}_i^{\top},\quad\text{where }\mathbf{v}_i=\sqrt{p}\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}.$$

Hence

$$\begin{aligned}\frac{\|\mathbf{W}\circ\hat{\mathbf{\Lambda}}\|_{\mathrm{F}}^2}{p}&=\frac{1}{pn^2}\mathrm{tr}\Big(\big(\mathbf{W}\circ\mathbf{v}_1\mathbf{v}_1^{\top}+\cdots+\mathbf{W}\circ\mathbf{v}_n\mathbf{v}_n^{\top}\big)^2\Big)\\&=\sum_{i=1}^{n}\frac{\|\mathbf{W}\circ\mathbf{v}_i\mathbf{v}_i^{\top}\|_{\mathrm{F}}^2}{pn^2}+\sum_{i\neq j}^{n}\frac{\mathrm{tr}\big((\mathbf{W}\circ\mathbf{v}_i\mathbf{v}_i^{\top})(\mathbf{W}\circ\mathbf{v}_j\mathbf{v}_j^{\top})\big)}{pn^2}.\end{aligned}$$

Then since $\mathbf{v}_i$-s are i.i.d., and $\mathbb{E}[\hat{\mathbf{\Lambda}}]=\mathbb{E}[\mathbf{v}_i\mathbf{v}_i^{\top}]$ for all $i$, the expectation of the 2nd term is

$$\sum_{i\neq j}\frac{\mathbb{E}\big[\,\mathrm{tr}\big((\mathbf{W}\circ\mathbf{v}_i\mathbf{v}_i^{\top})(\mathbf{W}\circ\mathbf{v}_j\mathbf{v}_j^{\top})\big)\big]}{pn^2}=\frac{n-1}{n}\frac{\|\mathbf{W}\circ\mathbf{\Lambda}_{\mathrm{sgn}}\|_{\mathrm{F}}^2}{p},$$

where $\mathbf{\Lambda}_{\mathrm{sgn}}=\mathbb{E}[\hat{\mathbf{\Lambda}}]$. The expectation of the 1st terms is

$$\begin{aligned}\sum_{i}\frac{\mathbb{E}\big[\|\mathbf{W}\circ\mathbf{v}_i\mathbf{v}_i^{\top}\|_{\mathrm{F}}^2\big]}{pn^2}&=\frac{\mathbb{E}\big[\|\mathbf{W}\circ\mathbf{v}\mathbf{v}^{\top}\|_{\mathrm{F}}^2\big]}{pn}\\&=\frac{\mathbb{E}\Big[\mathbf{d}^{\top}(\mathbf{W}\circ\mathbf{W})\mathbf{d}\Big]}{pn}\end{aligned}$$

where $\mathbf{d}=(v_1^2,\ldots,v_p^2)^{\top}$ contains the diagonal elements of $\mathbf{v}\mathbf{v}^{\top}$, where $\mathbf{v}=_d\mathbf{v}_i$ and $=_d$ reads "has the same distribution as". Furthermore, write $\mathbf{D}=\mathrm{diag}(\mathbf{v}\mathbf{v}^{\top})$. Thus we have that

$$\begin{aligned}&\frac{n}{n-1}\cdot\frac{\mathbb{E}\big[\|\mathbf{W}\circ\hat{\mathbf{\Lambda}}\|_{\mathrm{F}}^2\big]}{p}\\&\quad=\frac{\|\mathbf{W}\circ\mathbf{\Lambda}_{\mathrm{sgn}}\|_{\mathrm{F}}^2}{p}+\frac{\mathbb{E}\big[\mathbf{d}^{\top}(\mathbf{W}\circ\mathbf{W})\mathbf{d}\big]}{p(n-1)}.\quad(59)\end{aligned}$$

Next note that $\mathbf{D}_{\hat{\mathbf{\Lambda}}} = \operatorname{diag}(\hat{\mathbf{\Lambda}})$ can be written as

$$\mathbf{D}_{\hat{\mathbf{\Lambda}}} = \frac{1}{n}(\mathbf{D}_1 + \ldots + \mathbf{D}_n),$$

where $\mathbf{D}_i = \operatorname{diag}(\mathbf{v}_i\mathbf{v}_i^\top)$. Furthermore, let $\mathbf{d}_i = (v_{i1}^2, \ldots, v_{ip}^2)^\top$ denote a random vector containing the diagonal elements of $\mathbf{v}_i\mathbf{v}_i^\top$. Then we get

$$\begin{aligned}
&\operatorname{tr}\Big((\mathbf{D}_{\hat{\Lambda}}\mathbf{W})^2\Big) = \frac{1}{n^2}\operatorname{tr}\Big(\Big(\mathbf{D}_1\mathbf{W} + \ldots + \mathbf{D}_n\mathbf{W}\Big)^2\Big) \\
&= \frac{1}{n^2}\sum_{i=1}^n \operatorname{tr}\Big(\Big(\mathbf{D}_i\mathbf{W}\Big)^2\Big) + \frac{1}{n^2}\sum_{i\neq j}\operatorname{tr}\Big((\mathbf{D}_i\mathbf{W})\mathbf{D}_j\mathbf{W}\Big) \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbf{d}_i(\mathbf{W}\circ\mathbf{W})\mathbf{d}_i + \frac{1}{n^2}\sum_{i\neq j}\operatorname{tr}\Big((\mathbf{D}_i\mathbf{W})\mathbf{D}_j\mathbf{W}\Big).
\end{aligned}$$

Thus

$$\begin{aligned}
&\frac{1}{p(n-1)}\mathbb{E}\Big[\operatorname{tr}\Big((\mathbf{D}_{\hat{\Lambda}}\mathbf{W})^2\Big)\Big] \\
&= \frac{\mathbb{E}[\mathbf{d}(\mathbf{W}\circ\mathbf{W})\mathbf{d}]}{pn(n-1)} + \frac{1}{pn}\operatorname{tr}\Big((\mathbb{E}[\mathbf{D}]\mathbf{W})^2\Big) \\
&= \frac{\mathbb{E}[\mathbf{d}(\mathbf{W}\circ\mathbf{W})\mathbf{d}]}{pn(n-1)} + \frac{1}{pn}\mathbb{E}[\mathbf{d}]^\top(\mathbf{W}\circ\mathbf{W})\mathbb{E}[\mathbf{d}]. \qquad (60)
\end{aligned}$$

Using (59) and (60) we then obtain that

$$\mathbb{E}[\hat{\gamma}_{\mathbf{W}}] = \frac{\|\mathbf{W}\circ\mathbf{\Lambda}_{\text{sgn}}\|_{\text{F}}^2}{p} + \frac{1}{n}\varepsilon, \qquad (61)$$

where

$$\begin{aligned}
\varepsilon &= \frac{1}{p}\Big(\mathbb{E}[\mathbf{d}(\mathbf{W}\circ\mathbf{W})\mathbf{d}] - \mathbb{E}[\mathbf{d}]^\top(\mathbf{W}\circ\mathbf{W})\mathbb{E}[\mathbf{d}]\Big) \\
&= \frac{1}{p}\Big(\sum_{i=1}^p \operatorname{var}(v_i^2) + \sum_{i\neq j}^p w_{ij}\operatorname{cov}(v_iv_j)\Big) \to 0 \quad \text{as } p\to\infty. \qquad (62)
\end{aligned}$$

Next note that $\mathbf{\Lambda}_{\text{sgn}} = \mathbb{E}[\hat{\mathbf{\Lambda}}] = \mathbf{\Lambda} + o(\|\mathbf{\Lambda}\|_{\text{F}})$ when (A) holds by [28, Theorem 2]. This fact together with (61) and (62) imply that

$$\mathbb{E}[\hat{\gamma}_{\mathbf{W}}] \to \frac{\|\mathbf{W}\circ\mathbf{\Lambda}\|_{\text{F}}^2}{p} = \gamma_{\mathbf{W}}$$

as $p\to\infty$ under assumption (A). Thus we have proven the claim.

### *E. Proof of Lemma 2: complex case*

In our proof we will use the following identities.

**Lemma 4.** *The following holds:*

*a)* $\|\mathbf{A}\circ\mathbf{B}\|_{\text{F}}^2 = \operatorname{tr}\Big(\operatorname{vec}(\mathbf{A})\operatorname{vec}(\mathbf{A})^{\mathsf{H}}\circ\operatorname{vec}(\mathbf{B})\operatorname{vec}(\mathbf{B})^{\mathsf{H}}\Big)$ *for all* $\mathbf{A},\mathbf{B}\in\mathbb{C}^{m\times n}$.

*b)* $\mathbf{d}_{\mathbf{B}}^\top(\mathbf{A}\circ\mathbf{A})\mathbf{d}_{\mathbf{B}} = \operatorname{tr}\Big(\operatorname{vec}(\mathbf{A})\operatorname{vec}(\mathbf{A})^\top\circ(\mathbf{B}^*\otimes\mathbf{B})\Big)$ $\forall\,\mathbf{A}\in\mathbb{C}^{m\times m}$ *and* $\mathbf{B}\in\mathbb{C}_{\text{Sym}}^{m\times m}$.

*c)* $\operatorname{tr}\big((\mathbf{D}_{\mathbf{B}}\mathbf{A})^2\big) = \mathbf{d}_{\mathbf{B}}^\top(\mathbf{A}\circ\mathbf{A})\mathbf{d}_{\mathbf{B}}$ *for all* $\mathbf{A}\in\mathbb{R}_{\text{Sym}}^{m\times m}$ *and* $\mathbf{B}\in\mathbb{C}^{m\times m}$.

*Proof. a,b)* proofs of the identities are as proofs of Lemma 3a),b). *c)* follows directly from [21, Lemma 7.5.2]. □

Write $\mathbf{w} = \operatorname{vec}(\mathbf{W})$. Using Lemma 4a) we first notice that

$$\mathbb{E}\Big[\|\mathbf{W}\circ\mathbf{S}\|_{\text{F}}^2\Big] = \operatorname{tr}\Big(\mathbf{w}\mathbf{w}^\top\circ\mathbb{E}[\operatorname{vec}(\mathbf{S})\operatorname{vec}(\mathbf{S})^{\mathsf{H}}]\Big). \qquad (63)$$

We then recall that the (variance-)covariance matrix of $\mathbf{S}$ when sampling from a complex elliptically symmetric distribution $\mathbb{C}\mathcal{E}_p(\boldsymbol{\mu},\mathbf{\Sigma},g)$ is [39, Theorem 3]:

$$\begin{aligned}
\operatorname{cov}(\operatorname{vec}(\mathbf{S})) &= \mathbb{E}[\operatorname{vec}(\mathbf{S})\operatorname{vec}(\mathbf{S})^{\mathsf{H}}] - \operatorname{vec}(\mathbf{\Sigma})\operatorname{vec}(\mathbf{\Sigma})^{\mathsf{H}} \qquad (64)\\
&= \tau_1(\mathbf{\Sigma}^*\otimes\mathbf{\Sigma}) + \tau_2\operatorname{vec}(\mathbf{\Sigma})\operatorname{vec}(\mathbf{\Sigma})^{\mathsf{H}}, \qquad (65)
\end{aligned}$$

where $\tau_1$ and $\tau_2$ are constants defined in (26). Equations (64) and (65) then imply that

$$\begin{aligned}
\mathbb{E}[\operatorname{vec}(\mathbf{S})\operatorname{vec}(\mathbf{S})^{\mathsf{H}}] &= \operatorname{cov}(\operatorname{vec}(\mathbf{S})) + \operatorname{vec}(\mathbf{\Sigma})\operatorname{vec}(\mathbf{\Sigma})^{\mathsf{H}} \\
&= \tau_1(\mathbf{\Sigma}^*\otimes\mathbf{\Sigma}) + (1+\tau_2)\operatorname{vec}(\mathbf{\Sigma})\operatorname{vec}(\mathbf{\Sigma})^{\mathsf{H}}. \qquad (66)
\end{aligned}$$

Inserting (66) into (63) yields

$$\mathbb{E}\Big[\|\mathbf{W}\circ\mathbf{S}\|_{\text{F}}^2\Big] = (1+\tau_2)\|\mathbf{W}\circ\mathbf{\Sigma}\|_{\text{F}}^2 + \tau_1\operatorname{tr}\big((\mathbf{D}_{\mathbf{\Sigma}}\mathbf{W})^2\big).$$

simply by invoking identities in Lemma 4. This proves the first identity. The proof of latter part $\mathbb{E}\big[\operatorname{tr}((\mathbf{D}_{\mathbf{S}}\mathbf{W})^2)\big]$ is as earlier in the real-valued case in Appendix C.

## REFERENCES


[1] L. Du, J. Li, and P. Stoica, "Fully automatic computation of diagonal loading levels for robust adaptive beamforming," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 1, pp. 449–458, 2010.

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

[3] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for MMSE covariance estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5016–5029, 2010.

[4] E. Ollila and E. Raninen, "Optimal shrinkage covariance matrix estimation under random sampling from elliptical distributions," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2707–2719, 2019.

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

[6] ——, "Covariance regularization by thresholding," *Ann. Stat.*, vol. 36, no. 6, pp. 2577–2604, 2008.

[7] T. T. Cai, C.-H. Zhang, H. H. Zhou *et al.*, "Optimal rates of convergence for covariance matrix estimation," *Ann. Stat.*, vol. 38, no. 4, pp. 2118–2144, 2010.

[8] J. Guerci and J. Bergin, "Principal components, covariance matrix tapers, and the subspace leakage problem," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 38, no. 1, pp. 152–162, 2002.

[9] R. Mailloux, "Covariance matrix augmentation to produce adaptive array pattern troughs," *Electronics Letters*, vol. 31, no. 10, pp. 771–772, 1995.

[10] M. Zatman, "Production of adaptive array troughs by dispersion synthesis," *Electronics Letters*, vol. 31, no. 25, pp. 2141–2142, 1995.

[11] J. R. Guerci, "Theory and application of covariance matrix tapers for robust adaptive beamforming," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 47, no. 4, pp. 977–985, 1999.

[12] H. Song, W. Kuperman, W. Hodgkiss, P. Gerstoft, and J. S. Kim, "Null broadening with snapshot-deficient covariance matrices in passive sonar," *IEEE journal of Oceanic Engineering*, vol. 28, no. 2, pp. 250–261, 2003.

[13] L. Rugini, P. Banelli, and S. Cacopardi, "Regularized MMSE multiuser detection using covariance matrix tapering," in *IEEE International Conference on Communications, 2003. ICC'03.*, vol. 4. IEEE, 2003, pp. 2460–2464.

[14] X. Chen, Z. J. Wang, and M. J. McKeown, "Shrinkage-to-tapering estimation of large covariance matrices," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5640–5656, 2012.

[15] J. Li, J. Zhou, and B. Zhang, "Estimation of large covariance matrices by shrinking to structured target in normal and non-normal distributions," *IEEE Access*, vol. 6, pp. 2158–2169, 2018.

[16] O. Ledoit and M. Wolf, "Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size," *Ann. Stat.*, vol. 30, no. 4, pp. 1081–1102, 2002.

[17] M. S. Srivastava, "Some tests concerning the covariance matrix in high dimensional data," *Journal of the Japan Statistical Society*, vol. 35, no. 2, pp. 251–272, 2005.
[18] K.-T. Fang, S. Kotz, and K.-W. Ng, *Symmetric Multivariate and Related Distributions*. London: Chapman and hall, 1990.
[19] 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, 2012.
[20] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. New York: Wiley, 1982, 704 pages.
[21] R. A. Horn and C. A. Johnson, *Matrix Analysis*, 2nd ed. Cambridge: Cambridge University Press, 2012.
[22] J. I. Marden, "Some robust estimates of principal components," *Stat. Probab. Lett.*, vol. 43, no. 4, pp. 349–359, 1999.
[23] S. Visuri, V. Koivunen, and H. Oja, "Sign and rank covariance matrices," *J. Statist. Plann. Inference*, vol. 91, pp. 557–575, 2000.
[24] C. Zou, L. Peng, L. Feng, and Z. Wang, "Multivariate sign-based high-dimensional tests for sphericity," *Biometrika*, vol. 101, no. 1, 2014.
[25] T. Zhang and A. Wiesel, "Automatic diagonal loading for Tyler's robust covariance estimator," in *IEEE Statistical Signal Processing Workshop (SSP'16)*, 2016, pp. 1–5.
[26] 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, 2014.
[27] C. Croux, C. Dehon, and A. Yadine, "The k-step spatial sign covariance matrix," *Advances in data analysis and classification*, vol. 4, no. 2, pp. 137–150, 2010.
[28] E. Raninen, D. E. Tyler, and E. Ollila, "Linear pooling of sample covariance matrices," *IEEE Trans. Signal Process.*, vol. 70, pp. 659–672, 2021.
[29] B. Brown, "Statistical Uses of the Spatial Median," *J. Royal Stat. Soc., Ser. B*, vol. 45, no. 1, pp. 25–30, 1983.
[30] A. Dürre, D. Vogel, and D. E. Tyler, "The spatial sign covariance matrix with unknown location," *J. Mult. Anal.*, vol. 130, pp. 107–117, 2014.
[31] 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.
[32] B. Rajaratnam and J. Salzman, "Best permutation analysis," *J. Mult. Anal.*, vol. 121, pp. 193–223, 2013.
[33] A. Wagaman and E. Levina, "Discovering sparse covariance structures with the isomap," *J. Comput. Graph Stat.*, vol. 18, no. 3, pp. 551–572, 2009.
[34] J. Ward, "Space time adaptive processing for airborne radar," MIT, Lexington, Mass., USA, Tech. Rep., December 1994.
[35] S. Kraut, L. L. Scharf, and R. W. Butler, "The adaptive coherence estimator: a uniformly most-powerful-invariant adaptive detection statistic," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 427–438, Feb 2005.
[36] S. Kraut and L. L. Scharf, "The cfar adaptive subspace detector is a scale-invariant glrt," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2538–2541, Sep 1999.
[37] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1702–1715, 2003.
[38] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 2nd ed. Chichester: Wiley, 1999, 422 pages.
[39] E. Raninen, E. Ollila, and D. E. Tyler, "On the variability of the sample covariance matrix under complex elliptical distributions," *IEEE Signal Processing Letters*, vol. 28, pp. 2092–2096, 2021.