Random Matrix Derived Shrinkage of Spectral Precision Matrices

A. T. Walden, *Senior Member, IEEE*, and D. Schneider-Luftman

***Abstract*—Much research has been carried out on shrinkage methods for real-valued covariance matrices. In spectral analysis of p-vector-valued time series there is often a need for good shrinkage methods too, most notably when the complex-valued spectral matrix is singular. The equivalent of the Ledoit-Wolf (LW) covariance matrix estimator for spectral matrices can be improved on using a Rao-Blackwell estimator, and using random matrix theory we derive its form. Such estimators can be used to better estimate inverse spectral (precision) matrices too, and a random matrix method has previously been proposed and implemented via extensive simulations. We describe the method, but carry out computations entirely analytically, and suggest a way of selecting an important parameter using a predictive risk approach. We show that both the Rao-Blackwell estimator and the random matrix estimator of the precision matrix can substantially outperform the inverse of the LW estimator in a time series setting. Our new methodology is applied to EEG-derived time series data where it is seen to work well and deliver substantial improvements for precision matrix estimation.**

***Index Terms*—Rao-Blackwell estimators, random matrix theory, shrinkage, spectral matrix.**

I. Introduction

A stationary p-vector-valued time series has, at each frequency f, a $p \times p$ complex-valued spectral matrix $\boldsymbol{S}(f)$, for which an estimator $\hat{\boldsymbol{S}}(f)$, can be derived. If such an estimator is computed by a multitaper scheme involving K tapers (e.g., [32]) then the spectral matrices — complex-valued analogues of covariance matrices — will be singular if $p > K$ (and ill-conditioned if K is only a little larger than p). Unfortunately K cannot be simply increased because of its connection to the implied smoothing bandwidth: if K is made larger, the required resolution may be lost. (Other estimators such as periodograms smoothed over frequencies have analogous properties.) In this paper we look at the estimation of $\boldsymbol{S}(f)$ and more particularly the spectral 'precision' matrix defined as $\boldsymbol{C}(f) = \boldsymbol{S}^{-1}(f)$ when $\hat{\boldsymbol{S}}(f)$ is singular. The precision matrix is used in the computation of partial coherencies in time series graphical modelling (see e.g. [29] and references therein for a neuroscience application). We don't assume a very large p since the moderate p scenario is often encountered in practice and practically is just as important. We shall first give a review of relevant covariance matrix estimation literature, before turning to the contributions of this paper.

 Andrew Walden and Deborah Schneider-Luftman are both at the Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7 2BZ, UK. (e-mail: a.walden@imperial.ac.uk and deborah.schneider-luftman11@imperial.ac.uk)

The estimation of a covariance matrix $\boldsymbol{\Sigma}$ from N samples of p real-valued zero mean random variables has been extensively researched for the case $N > p$. Although the resulting non-singular sample covariance estimator $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ is unbiased its eigenvalues tend to be more spread out than the true eigenvalues. To ameliorate this problem [21] looked at minimax estimation over a certain group, but the estimators depend on the coordinate system. This problem was removed by [10] who considered orthogonally equivariant minimax estimators: an estimator $\mathcal{F}(\hat{\boldsymbol{\Sigma}})$ of $\boldsymbol{\Sigma}$ is said to be orthogonally equivariant if for any orthogonal matrix $\boldsymbol{O}$, we have $\mathcal{F}(\boldsymbol{O}\hat{\boldsymbol{\Sigma}}\boldsymbol{O}^T) = \boldsymbol{O}\mathcal{F}(\hat{\boldsymbol{\Sigma}})\boldsymbol{O}^T$, where ^{T} denotes transposition. In fact such estimators shrink the sample eigenvalues, and so are of the widely researched shrinkage class, see e.g., [11], [18], [36].

For shrinkage estimators which are a combination of the standard covariance matrix and a target matrix proportional to the identity, Ledoit and Wolf (LW) [24], [25] derived the ideal shrinkage parameter, or 'oracle' value, that minimizes a risk measure between $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$. Such LW estimators are (i) suitable for the case $N < p$ when $\hat{\boldsymbol{\Sigma}}$ is singular, (ii) do not assume Gaussianity, and (iii) may be used in large p settings. Modifications to the target matrix were discussed in [35] and [8], the latter shrinking the sample covariance matrix towards its tapered version for high-dimensional matrices; modified estimators for this case were also suggested in [13].

Under the Gaussianity assumption, [9] showed that the LW estimator can be significantly improved upon. They developed the so-called Rao-Blackwell (RB) estimator which is guaranteed at least as good as the LW estimator under any convex loss criterion.

There has also been much interest in accurate estimation of the precision matrix $\boldsymbol{\Sigma}^{-1}$. A weighted combination of $\hat{\boldsymbol{\Sigma}}^{-1}$ and the identity was considered by [11], and improved on by [17]. By looking over the class of orthogonally equivariant estimators for real covariance matrices, Ledoit and Wolf [26] produced nonlinear shrinkage estimators for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$. All these studies assumed that $N > p$. Also the calculations involved in [26] are hugely costly. The singular case has been attracting much attention recently in the context of estimating sparse precision matrices $\boldsymbol{\Sigma}^{-1}$ in high-dimensional situations ($p >> N$), see e.g., [3], [7], [23], [30], [34].

Following some background material on spectral matrix estimation in Section II, the contributions of this paper are as follows.

1) In Section III we study LW oracle estimation for $\boldsymbol{S}(f)$, and give the form of the practical estimator $\hat{\boldsymbol{S}}_{\mathrm{LW}}(f)$. The related Rao-Blackwell estimator for the spectral matrix, $\hat{\boldsymbol{S}}_{\mathrm{RB}}(f)$, is found in Section IV. These oracle

and Rao-Blackwell estimators are surprisingly different in form to the real-valued cases. The Rao-Blackwell estimator is derived making substantial use of random matrix theory and is very simple in form and thus highly usable in practice. The Gaussian assumption is used to derive simple forms for the oracle shrinkage parameter and for the Rao-Blackwell estimator. While in standard real-valued covariance matrix estimation Gaussianity is a problematic assumption and robustness issues arise, in our context this is not dubious because of the Central Limit Theorem effect of the vector Fourier transform used in the time series setting.

2) Section V points out that the inverse of the Rao-Blackwell estimator is in the form of a "Rao-Blackwellized" estimator for $\boldsymbol{C}(f)$. We show that this estimator can substantially outperform the inverse of the LW estimator in a time series setting.
3) In Section VI we examine direct estimation of $\boldsymbol{C}(f)$ from singular estimators $\hat{\boldsymbol{S}}(f)$ using random matrix methods as developed in [28], and formulate a completely analytic (rather than simulation-based) approach to obtain the estimators. A predictive risk approach is given to select a controlling parameter. We show that this estimator can substantially outperform the inverse of the LW estimator in a time series setting.
4) Our new methodology is applied to electroencephalogram (EEG) derived time series data in Section VII, where it is seen to work well and deliver substantial improvements over the inverse LW estimators of $\boldsymbol{C}(f)$.

II. Spectral Matrix Estimation

Here we consider a real p-vector-valued discrete time stochastic process $\{\boldsymbol{X}_t\}$ whose tth element is the column vector $\boldsymbol{X}_t = [X_{1,t}, \ldots, X_{p,t}]^T$, and each component process has zero mean. The sample interval is denoted by Δ_{t}. We assume the p processes are jointly stationary, i.e., for all $l, m = 1, \ldots, p$, $s_{lm,\tau} = \mathrm{cov}\,\{X_{l,t+\tau}, X_{m,t}\}$ is a function of τ only.

The matrix autocovariance sequence $\{\boldsymbol{s}_\tau\}$ is defined by $\boldsymbol{s}_\tau = \mathrm{cov}\{\boldsymbol{X}_{t+\tau}, \boldsymbol{X}_t^T\} = E\{\boldsymbol{X}_{t+\tau}\boldsymbol{X}_t^T\}$, and each component is assumed absolutely summable. The spectral matrix, is then $\boldsymbol{S}(f) = \Delta_{\mathrm{t}} \sum_{\tau=-\infty}^{\infty} \boldsymbol{s}_\tau \, \mathrm{e}^{-\mathrm{i}2\pi f \tau \, \Delta_{\mathrm{t}}}$.

We make use of a set of K orthonormal tapers $\{h_{k,t}\}, k = 0, \ldots, K-1$ and for $t = 0, \ldots, N-1$, form the product $h_{k,t}\boldsymbol{X}_t$ of the tth component of the kth taper with the tth component of the p-vector-valued process, and for $k = 0, \ldots, K-1$ compute the vector Fourier transform

$$\boldsymbol{J}_k(f) \stackrel{\mathrm{def}}{=} \Delta_{\mathrm{t}}^{1/2} \sum_{t=0}^{N-1} h_{k,t}\boldsymbol{X}_t \, \mathrm{e}^{-\mathrm{i}2\pi f t \, \Delta_{\mathrm{t}}}.$$

Let $\boldsymbol{J}(f)$ be the $p \times K$ matrix defined by

$$\boldsymbol{J}(f) = [\boldsymbol{J}_0(f), \ldots, \boldsymbol{J}_{K-1}(f)]. \tag{1}$$

Then the multitaper estimator of the $p \times p$ spectral matrix $\boldsymbol{S}(f)$ is

$$\begin{aligned} \hat{\boldsymbol{S}}(f) &= \frac{1}{K}\sum_{k=0}^{K-1} \hat{\boldsymbol{S}}_k(f) = \frac{1}{K}\sum_{k=0}^{K-1} \boldsymbol{J}_k(f)\boldsymbol{J}_k^H(f) = \\ &= \frac{1}{K}\boldsymbol{J}(f)\boldsymbol{J}^H(f), \end{aligned} \tag{2}$$

where $\hat{\boldsymbol{S}}_k(f) \stackrel{\mathrm{def}}{=} \boldsymbol{J}_k(f)\boldsymbol{J}_k^H(f)$.

Remark 1. *This conveniently mimicks the classical covariance matrix estimator: if $\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_{K-1}$ are K independent p-dimensional Gaussian real-valued random vectors with zero means and covariance matrix $\boldsymbol{\Sigma}$, then the maximum likelihood estimator for $\boldsymbol{\Sigma}$ is $\hat{\boldsymbol{\Sigma}} = \frac{1}{K}\sum_{k=0}^{K-1} \boldsymbol{Y}_k\boldsymbol{Y}_k^T$.*

Letting B denote the bandwidth of the spectral window corresponding to the tapering, then $\boldsymbol{J}_k(f), k = 0, \ldots, K-1$, may be taken to be independently and identically distributed as p-vector-valued complex Gaussian with mean zero and covariance matrix $\boldsymbol{S}(f)$:

$$\boldsymbol{J}_k(f) \stackrel{\mathrm{d}}{=} \mathcal{N}_p^C\{\boldsymbol{0}, \boldsymbol{S}(f)\}, \tag{3}$$

for $B/2 < |f| < f_{\mathcal{N}} - B/2$ for finite N and Gaussian processes, or $0 < |f| < f_{\mathcal{N}}$ asymptotically [5]. Then the estimator of (2) is the maximum-likelihood estimator for $\boldsymbol{S}(f)$, [16]. Further,

$$E\{\hat{\boldsymbol{S}}(f)\} = \frac{1}{K}\sum_{k=0}^{K-1} E\{\boldsymbol{J}_k(f)\boldsymbol{J}_k^H(f)\} = \frac{1}{K}\sum_{k=0}^{K-1} \boldsymbol{S}(f) = \boldsymbol{S}(f), \tag{4}$$

and $E\{\mathrm{tr}\{\hat{\boldsymbol{S}}\}\} = E\{\sum_{j=1}^{p} \hat{S}_{jj}\} = \sum_{j=1}^{p} S_{jj} = \mathrm{tr}\{\boldsymbol{S}\}$, results we shall make use of later. These hold whether $K \geq p$, which corresponds to $\hat{\boldsymbol{S}}(f)$ being non-singular, or $K < p$, when the estimated matrix is singular (both with probability one).

III. Conventional Shrinkage Methodology

The conventional approach to 'covariance matrix' regularization which has been extensively studied involves the forming of a convex combination of the sample covariance matrix and some well-conditioned 'target' matrix. For an estimated $p \times p$ Hermitian spectral matrix $\hat{\boldsymbol{S}}(f)$ this would take the form

$$\boldsymbol{S}^\star(f) = (1 - \rho(f))\hat{\boldsymbol{S}}(f) + \rho(f)\hat{\boldsymbol{T}}(f), \tag{5}$$

where $\rho(f) \in (0, 1)$ is known as the shrinkage parameter and $\hat{\boldsymbol{T}}(f)$ is the target matrix. Provided $\hat{\boldsymbol{S}}(f)$ and $\hat{\boldsymbol{T}}(f)$ are both positive definite, then this convex combination will itself be positive definite. For notational brevity we shall drop the explicit frequency dependence in most of what follows.

Apart from being positive definite, suppose that no *a priori* form is imposed on $\hat{\boldsymbol{T}}$ and our goal is to find an optimal estimator for $\boldsymbol{S}$ of the form of (5) by determining $\rho = \rho_0$ such that

$$\rho_0 = \arg\min E\{||\boldsymbol{S}^\star - \boldsymbol{S}||_{\mathrm{F}}^2\},$$

where, for $\boldsymbol{A} \in \mathbb{C}^{p \times p}$, $||\boldsymbol{A}||_{\mathrm{F}}$ denotes the Frobenius norm $||\boldsymbol{A}||_{\mathrm{F}} = [\mathrm{tr}\{\boldsymbol{A}\boldsymbol{A}^H\}]^{1/2}$, $\mathrm{tr}\{\cdot\}$ denotes trace, and ^{H} denotes complex-conjugate (Hermitian) transpose.

A. Oracle Estimator

Firstly we define

$$\begin{aligned}\alpha^2 &= E\{||\boldsymbol{S}-\hat{\boldsymbol{T}}||_{\rm F}^2\} = E\{{\rm tr}\{[\boldsymbol{S}-\hat{\boldsymbol{T}}][\boldsymbol{S}-\hat{\boldsymbol{T}}]^H\}\}\\ \beta^2 &= E\{||\hat{\boldsymbol{S}}-\boldsymbol{S}||_{\rm F}^2\} = E\{{\rm tr}\{[\hat{\boldsymbol{S}}-\boldsymbol{S}][\hat{\boldsymbol{S}}-\boldsymbol{S}]^H\}\}\\ \delta^2 &= E\{||\hat{\boldsymbol{S}}-\hat{\boldsymbol{T}}||_{\rm F}^2\} = E\{{\rm tr}\{[\hat{\boldsymbol{S}}-\hat{\boldsymbol{T}}][\hat{\boldsymbol{S}}-\hat{\boldsymbol{T}}]^H\}\}\\ \gamma^2 &= E\{{\rm tr}\{[\hat{\boldsymbol{S}}-\boldsymbol{S}][\boldsymbol{S}-\hat{\boldsymbol{T}}]^H\}\}.\end{aligned}$$

Then with $\mathrm{Re}\{\cdot\}$ denoting "real part of,"

$$\begin{aligned}\delta^2 &= E\{||\hat{\boldsymbol{S}}-\hat{\boldsymbol{T}}||_{\rm F}^2\} = E\{||\ [\hat{\boldsymbol{S}}-\boldsymbol{S}]+[\boldsymbol{S}-\hat{\boldsymbol{T}}]\ ||_{\rm F}^2\}\\ &= E\{||\ \boldsymbol{S}-\hat{\boldsymbol{T}}\ ||_{\rm F}^2\} + E\{||\ \hat{\boldsymbol{S}}-\boldsymbol{S}\ ||_{\rm F}^2\}\\ &+ 2{\rm Re}\{E\{{\rm tr}\{[\hat{\boldsymbol{S}}-\boldsymbol{S}][\boldsymbol{S}-\hat{\boldsymbol{T}}]^H\}\}\}\\ &= \alpha^2+\beta^2+2\gamma^2,\end{aligned}$$

since $[\hat{\boldsymbol{S}}-\boldsymbol{S}]$ and $[\boldsymbol{S}-\hat{\boldsymbol{T}}]^H$ are both Hermitian, (each of $\hat{\boldsymbol{S}}, \boldsymbol{S}$ and $\hat{\boldsymbol{T}}$ is Hermitian), and therefore the trace of the product is guaranteed real-valued, so $\mathrm{Re}\{\cdot\}$ is not needed.

The objective function can be written

$$\begin{aligned}E\{||\boldsymbol{S}^\star-\boldsymbol{S}||_{\rm F}^2\} &= E\{||(1-\rho)\hat{\boldsymbol{S}}+\rho\hat{\boldsymbol{T}}-\boldsymbol{S}||_{\rm F}^2\}\\ &= E\{||\ \rho[\hat{\boldsymbol{T}}-\boldsymbol{S}]+(1-\rho)[\hat{\boldsymbol{S}}-\boldsymbol{S}]\ ||_{\rm F}^2\}\\ &= \rho^2\alpha^2+(1-\rho)^2\beta^2-2\rho(1-\rho)\gamma^2.\end{aligned}$$

Differentiating with respect to ρ and setting to zero:

$$\frac{\partial}{\partial\rho}E\{||\boldsymbol{S}^\star-\boldsymbol{S}||_{\rm F}^2\} = 2\rho\alpha^2-2(1-\rho)\beta^2-2(1-2\rho)\gamma^2 = 0$$

so that the solution is [12], [13]

$$\rho_0 = \frac{\beta^2+\gamma^2}{\delta^2} = \frac{\beta^2-\alpha^2+\delta^2}{2\delta^2}. \tag{6}$$

The second derivative is positive so that the objective function is minimized with this ρ_0 value.

The term $\beta^2+\gamma^2$ can be rewritten as

$$\begin{aligned}&E\{{\rm tr}\{[\hat{\boldsymbol{S}}-\boldsymbol{S}][\hat{\boldsymbol{S}}-\boldsymbol{S}]^H\}\} + E\{{\rm tr}\{[\hat{\boldsymbol{S}}-\boldsymbol{S}][\boldsymbol{S}-\hat{\boldsymbol{T}}]^H\}\}\\ =\ & E\{{\rm tr}\{[\hat{\boldsymbol{S}}-\boldsymbol{S}][\hat{\boldsymbol{S}}-\hat{\boldsymbol{T}}]\}\},\end{aligned}$$

where we have used the Hermitian properties of $\hat{\boldsymbol{S}}$ and $\hat{\boldsymbol{T}}$. So ρ_0 in (6) becomes

$$\rho_0 = \frac{E\big\{{\rm tr}\{[\hat{\boldsymbol{S}}-\boldsymbol{S}][\hat{\boldsymbol{S}}-\hat{\boldsymbol{T}}]\}\big\}}{E\big\{{\rm tr}\{[\hat{\boldsymbol{S}}-\hat{\boldsymbol{T}}]^2\}\big\}}. \tag{7}$$

which is of the same form as found in [9, eqn. (6)] for the real-valued case. This form for ρ_0 is distribution invariant. In order to rewrite ρ_0 in (7) in a useful form involving just $\boldsymbol{S}$ and parameters K and p, Gaussianity will be assumed, which is justified as discussed earlier.

B. Stochastic Target

Suppose we define $\mu_0 = {\rm tr}\{\boldsymbol{S}\}/p$ and $\hat{\mu}_0 = {\rm tr}\{\hat{\boldsymbol{S}}\}/p$ and take $\hat{\boldsymbol{T}} = ({\rm tr}\{\hat{\boldsymbol{S}}\}/p)\boldsymbol{I}_p = \hat{\mu}_0\boldsymbol{I}_p$. In this case both $\hat{\boldsymbol{T}}$ and $\hat{\boldsymbol{S}}$ will be subject to estimation error and will in general be correlated. (This was the case developed in [24] for real-valued covariance matrices.)

Theorem 1. *Let $\hat{\boldsymbol{T}} = ({\rm tr}\{\hat{\boldsymbol{S}}\}/p)\boldsymbol{I}_p$. Under the assumption (3), ρ_0 in (7) can be written*

$$\rho_0 = \frac{{\rm tr}^2\{\boldsymbol{S}\} - \frac{1}{p}{\rm tr}\{\boldsymbol{S}^2\}}{[1-\frac{K}{p}]{\rm tr}^2\{\boldsymbol{S}\} + [K-\frac{1}{p}]{\rm tr}\{\boldsymbol{S}^2\}}. \tag{8}$$

Proof: From (7)

$$\rho_0 = \frac{E\big\{{\rm tr}\{[\hat{\boldsymbol{S}}-\boldsymbol{S}][\hat{\boldsymbol{S}}-({\rm tr}\{\hat{\boldsymbol{S}}\}/p)\boldsymbol{I}_p]\}\big\}}{E\big\{{\rm tr}\{[\hat{\boldsymbol{S}}-({\rm tr}\{\hat{\boldsymbol{S}}\}/p)\boldsymbol{I}_p]^2\}\big\}}.$$

The numerator and denominator are then

$$E\big\{{\rm tr}\{\hat{\boldsymbol{S}}^2\} - \frac{1}{p}{\rm tr}^2\{\hat{\boldsymbol{S}}\} - {\rm tr}\{\boldsymbol{S}\hat{\boldsymbol{S}}\} + \frac{1}{p}{\rm tr}\{\boldsymbol{S}\}{\rm tr}\{\hat{\boldsymbol{S}}\}\big\}$$

and $$E\big\{{\rm tr}\{\hat{\boldsymbol{S}}^2\} - \frac{1}{p}{\rm tr}^2\{\hat{\boldsymbol{S}}\}\big\},$$

respectively. Under the assumption (3), $K\hat{\boldsymbol{S}}$ has the complex Wishart distribution with mean $K\boldsymbol{S}$. Then we know (e.g., [27])

$$\begin{aligned}E\big\{{\rm tr}\{\hat{\boldsymbol{S}}^2\}\big\} &= {\rm tr}\{\boldsymbol{S}^2\} + \frac{1}{K}{\rm tr}^2\{\boldsymbol{S}\}\\ E\big\{{\rm tr}^2\{\hat{\boldsymbol{S}}\}\big\} &= {\rm tr}^2\{\boldsymbol{S}\} + \frac{1}{K}{\rm tr}\{\boldsymbol{S}^2\}.\end{aligned}$$

So the numerator and denominator become

$$\frac{1}{K}[{\rm tr}^2\{\boldsymbol{S}\} - \frac{1}{p}{\rm tr}\{\boldsymbol{S}^2\}]$$

and $$\left[1-\frac{1}{pK}\right]{\rm tr}\{\boldsymbol{S}^2\} + \left[\frac{1}{K}-\frac{1}{p}\right]{\rm tr}^2\{\boldsymbol{S}\},$$

respectively, and their ratio gives the required result. ∎

The form (8) is known as an 'oracle' estimator since it involves the unknown quantities ${\rm tr}\{\boldsymbol{S}\}$ and ${\rm tr}\{\boldsymbol{S}^2\}$ and so its value is not known in practical situations.

Remark 2. *The form of the estimator (8) for complex-valued covariance matrix estimators is surprisingly different to that for real-valued covariance matrix estimators: compare (8) with [9, eqn. (7)].*

C. Deterministic Target

If $\hat{\boldsymbol{T}}$ is constant, $\hat{\boldsymbol{T}} = \boldsymbol{T}$ say, then the term $\gamma^2 = {\rm tr}\{E\{[\hat{\boldsymbol{S}}-\boldsymbol{S}]\}[\boldsymbol{S}-\boldsymbol{T}]\} = 0$, and ρ_0 in (6) becomes $\rho_0 = \beta^2/\delta^2$. We now consider the target matrix $\boldsymbol{T} = ({\rm tr}\{\boldsymbol{S}\}/p)\boldsymbol{I}_p = \mu_0\boldsymbol{I}_p$.

Theorem 2. *Let $\boldsymbol{T} = ({\rm tr}\{\boldsymbol{S}\}/p)\boldsymbol{I}_p$. Under the assumption (3), $\rho_0 = \beta^2/\delta^2$ can be written*

$$\rho_0 = \frac{{\rm tr}^2\{\boldsymbol{S}\}}{[1-\frac{K}{p}]{\rm tr}^2\{\boldsymbol{S}\} + K{\rm tr}\{\boldsymbol{S}^2\}}. \tag{9}$$

Proof: This proceeds along the same lines as for Theorem 1. ∎

This case was extensively studied in [25] who made many interesting observations. When $\boldsymbol{T} = \mu_0\boldsymbol{I}_p$, then using (5) the eigenvalues of $\hat{\boldsymbol{S}}$ are shrunk according to $\hat{\lambda}_i \to (1-\rho_0)\hat{\lambda}_i + \rho_0\mu_0$, thus reducing the condition number. μ_0 is the "grand mean" of both true and sample eigenvalues [25] and thus the sample eigenvalues will be shrunk towards their grand mean. In practice we will know neither μ_0 nor $\rho_0 = \beta^2/\delta^2$ since they both involve the unknown $\boldsymbol{S}$. These quantities can be estimated via "plug-in" values. Following the derivation of consistent

estimators in [2] we first take $\hat{\mu}_0$ for μ and next note that δ^2 could be estimated by omitting the expected value:

$$\begin{aligned}\hat{\delta^2} &= ||\hat{\boldsymbol{S}} - \hat{\mu}_0 \boldsymbol{I}_p||_{\mathrm{F}}^2 = \mathrm{tr}\{[\hat{\boldsymbol{S}} - \hat{\mu}_0 \boldsymbol{I}_p][\hat{\boldsymbol{S}} - \hat{\mu}_0 \boldsymbol{I}_p]^H\} \\ &= \mathrm{tr}\{\hat{\boldsymbol{S}}^2\} - \frac{\mathrm{tr}^2\{\hat{\boldsymbol{S}}\}}{p} = \sum_{i=1}^{p}\sum_{j=1}^{p} |\hat{S}_{ij} - \hat{\mu}_0 \delta_{i,j}|^2,\end{aligned}$$

where $\delta_{i,j}$ is the usual Kronecker delta, equal to unity when $i = j$, and zero otherwise. The estimation of $\beta^2 = E\{||\hat{\boldsymbol{S}} - \boldsymbol{S}||_{\mathrm{F}}^2\}$ is less simple. Using (4), β^2 can be written

$$\beta^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} E\{|\hat{S}_{ij} - E\{\hat{S}_{ij}\}|^2\} = \sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{var}\{\hat{S}_{ij}\}, \quad (10)$$

so it can be estimated using a form of sample variance: $\hat{\beta}^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} \widehat{\mathrm{var}}\{\hat{S}_{ij}\}$. Given (3), for the multitaper spectral matrix estimator we know $\mathrm{var}\{\hat{S}_{ij}\} = \mathrm{var}\left\{(1/K)\sum_{k=0}^{K-1} \hat{S}_{k,ij}\right\} = (1/K)\,\mathrm{var}\{\hat{S}_{k,ij}\}$, where $\hat{S}_{k,ij} = (\hat{\boldsymbol{S}}_k)_{ij}$. An estimator for $\mathrm{var}\{\hat{S}_{k,ij}\}$ is $\widehat{\mathrm{var}}\{\hat{S}_{k,ij}\} = (1/K)\sum_{k=0}^{K-1} |\hat{S}_{k,ij} - \hat{S}_{ij}|^2$, so we get $\widehat{\mathrm{var}}\{\hat{S}_{ij}\} = (1/K^2)\sum_{k=0}^{K-1} |\hat{S}_{k,ij} - \hat{S}_{ij}|^2$, which gives an estimator of β^2 in (10) of the form

$$\hat{\beta}^2 = \frac{1}{K^2}\sum_{k=0}^{K-1} ||\hat{\boldsymbol{S}}_k - \hat{\boldsymbol{S}}||_{\mathrm{F}}^2, \quad (11)$$

so the estimator of ρ_0 becomes

$$\hat{\rho}_0 = \frac{\hat{\beta}^2}{\hat{\delta}^2} = \frac{\sum_{k=0}^{K-1} ||\hat{\boldsymbol{S}}_k - \hat{\boldsymbol{S}}||_{\mathrm{F}}^2}{K^2\left[\mathrm{tr}\{\hat{\boldsymbol{S}}^2\} - (\mathrm{tr}^2\{\hat{\boldsymbol{S}}\}/p)\right]} \stackrel{\mathrm{def}}{=} \hat{\rho}_{\mathrm{LW}}, \quad (12)$$

where we have defined this estimator to be $\hat{\rho}_{\mathrm{LW}}$ because it is of the same form as derived in [25, pp. 379–380] for real-valued covariance matrices.

Finally then the proposed shrinkage estimator of the spectrum is, from (5), given by

$$\hat{\boldsymbol{S}}_{\mathrm{LW}} = [1 - \hat{\rho}_{\mathrm{LW}}]\,\hat{\boldsymbol{S}} + \hat{\rho}_{\mathrm{LW}}\hat{\mu}_0 \boldsymbol{I}_p, \quad (13)$$

exactly mimicking [25, p. 380]. As a result the empirical shrinkage of the eigenvalues is given by $\hat{\lambda}_i \to (1 - \hat{\rho}_{\mathrm{LW}})\hat{\lambda}_i + \hat{\rho}_{\mathrm{LW}}\hat{\mu}_0$. This approach can be used if $\hat{\boldsymbol{S}}$ is singular or ill-conditioned. Notice that if $K < p$, so that $\hat{\boldsymbol{S}}$ is singular, the resulting zero eigenvalues will be modified to $\hat{\rho}_{\mathrm{LW}}\hat{\mu}_0$.

Note that since $\delta^2 = \alpha^2 + \beta^2$ if we define $\bar{\beta}^2 = \min\{\hat{\beta}^2, \hat{\delta}^2\}$ then $\bar{\beta}^2/\hat{\delta}^2 = \min\{\hat{\rho}_{\mathrm{LW}}, 1\}$ provides an estimate for the shrinkage parameter which is constrained by its theoretical upper bound of unity. This would be used in practical applications.

Remark 3. *The form of $\hat{\beta}^2$ given in (11) for the multitaper approach is very appealing as the averaging is all carried out at the frequency of interest, and is done over tapers. In the approach of [2, p. 921] the "local variance" averaging must be done over different frequencies.*

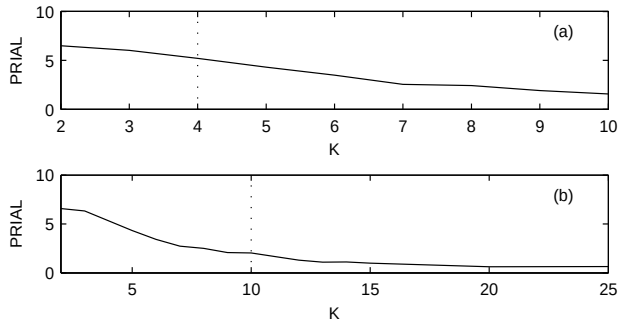

Fig. 1. Simulated PRIAL values for (a) $\boldsymbol{S}_A$ for which $p = 4$ and (b) $\boldsymbol{S}_B$ for which $p = 10$. In each case the dotted line indicates p.

IV. Rao-Blackwell Estimation

It is possible to produce another estimator from $\hat{\boldsymbol{S}}_{\mathrm{LW}}$ which is at least as good under any convex loss criterion. The transformed estimator to be derived is known as the Rao-Blackwell estimator and was developed for real-valued covariance matrices in the context of (13) by [9]. The idea is that if $T(\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1})$ is a sufficient statistic for $\boldsymbol{S}$, and if $\mathcal{S}(\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1})$ is an estimator for $\boldsymbol{S}$, then the conditional expectation $\mathcal{S}'(\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1}) \stackrel{\mathrm{def}}{=} E\{\mathcal{S}(\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1})|T\}$ is never worse than $\mathcal{S}(\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1})$ under any convex loss criterion. To see this, start with the risk $R(\boldsymbol{S}, \mathcal{S})$ of the original estimator [4, p. 483]

$$\begin{aligned}R(\boldsymbol{S}, \mathcal{S}) &= E_{\boldsymbol{S}}\{L(\boldsymbol{S}, \mathcal{S}(\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1}))\} \quad (14)\\ &= E_{\boldsymbol{S}}\{E\{L(\boldsymbol{S}, \mathcal{S}(\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1}))|T\}\} \\ &\geq E_{\boldsymbol{S}}\{L(\boldsymbol{S}, E\{\mathcal{S}(\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1})|T\})\} \\ &= E_{\boldsymbol{S}}\{L(\boldsymbol{S}, \mathcal{S}'(\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1}))\} \quad (15)\\ &= R(\boldsymbol{S}, \mathcal{S}').\end{aligned}$$

(Here the second line uses the rule of iterated expectation and the third line follows from Jensen's inequality and the assumed convexity of the loss function.)

In the context of spectral matrix estimation we note that under the independent complex Gaussian assumption for the $\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1}$, (3), that $\hat{\boldsymbol{S}}$ is a sufficient statistic for estimating $\boldsymbol{S}$, [16, Theorem 4.2]; this is true for $K \geq p$ and $K < p$. Then, the Rao-Blackwell estimator takes the form $\hat{\boldsymbol{S}}_{\mathrm{RB}} = E\{\hat{\boldsymbol{S}}_{\mathrm{LW}}|\hat{\boldsymbol{S}}\}$ and

$$\begin{aligned}R(\boldsymbol{S}, \hat{\boldsymbol{S}}_{\mathrm{LW}}) &= E_{\boldsymbol{S}}\{||\hat{\boldsymbol{S}}_{\mathrm{LW}} - \boldsymbol{S}||_{\mathrm{F}}^2\} \\ &= E_{\boldsymbol{S}}\{E\{||\hat{\boldsymbol{S}}_{\mathrm{LW}} - \boldsymbol{S}||_{\mathrm{F}}^2|\hat{\boldsymbol{S}}\}\} \\ &\geq E_{\boldsymbol{S}}\{||E\{\hat{\boldsymbol{S}}_{\mathrm{LW}}|\hat{\boldsymbol{S}}\} - \boldsymbol{S}||_{\mathrm{F}}^2\} \\ &= E_{\boldsymbol{S}}\{||\hat{\boldsymbol{S}}_{\mathrm{RB}} - \boldsymbol{S}||_{\mathrm{F}}^2\} = R(\boldsymbol{S}, \hat{\boldsymbol{S}}_{\mathrm{RB}}).\end{aligned}$$

So,

$$\begin{aligned}\hat{\boldsymbol{S}}_{\mathrm{RB}} &= E\{\hat{\boldsymbol{S}}_{\mathrm{LW}}|\hat{\boldsymbol{S}}\} = E\{[1 - \hat{\rho}_{\mathrm{LW}}]\,\hat{\boldsymbol{S}} + \hat{\rho}_{\mathrm{LW}}\hat{\mu}_0 \boldsymbol{I}_p|\hat{\boldsymbol{S}}\} \\ &= [1 - E\{\hat{\rho}_{\mathrm{LW}}|\hat{\boldsymbol{S}}\}]\hat{\boldsymbol{S}} + E\{\hat{\rho}_{\mathrm{LW}}\hat{\mu}_0|\hat{\boldsymbol{S}}\}\boldsymbol{I}_p \\ &\stackrel{\mathrm{def}}{=} [1 - \hat{\rho}_{\mathrm{RB}}]\hat{\boldsymbol{S}} + \hat{\rho}_{\mathrm{RB}}\hat{\mu}_0 \boldsymbol{I}_p,\end{aligned}$$

where the Rao-Blackwell shrinkage parameter $\hat{\rho}_{\rm RB}$ is

$$\hat{\rho}_{\rm RB} \stackrel{\rm def}{=} E\{\hat{\rho}_{\rm LW}|\hat{\boldsymbol{S}}\} = \frac{E\left\{\sum_{k=0}^{K-1} ||\hat{\boldsymbol{S}}_k - \hat{\boldsymbol{S}}||_{\rm F}^2 \; |\hat{\boldsymbol{S}}\right\}}{K^2\left[{\rm tr}\{\hat{\boldsymbol{S}}^2\} - ({\rm tr}^2\{\hat{\boldsymbol{S}}\}/p)\right]}. \quad (16)$$

The form of the shrinkage parameter was derived in [9] for real-valued covariance matrices. For our complex-valued case the form is substantially different.

Theorem 3. *Under the assumption (3), $\hat{\rho}_{\rm RB}$ in (16) takes the simple form*

$$\hat{\rho}_{\rm RB} = \frac{tr^2\{\hat{\boldsymbol{S}}\} - (tr\{\hat{\boldsymbol{S}}^2\}/K)}{(K+1)\left[tr\{\hat{\boldsymbol{S}}^2\} - (tr^2\{\hat{\boldsymbol{S}}\}/p)\right]}. \quad (17)$$

Proof: This uses invariance properties of the random matrix $\boldsymbol{J}$ and the random unitary matrices arising from its singular value decomposition. Details are given in AppendixB: put the results of Lemma 6 and Lemma 7 into the numerator of (16), then (17) readily follows. ∎

From (14) and (15) we have that $E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{S}}_{\rm LW} - \boldsymbol{S}\|_{\rm F}^2\} \geq E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{S}}_{\rm RB} - \boldsymbol{S}\|_{\rm F}^2\}$. It is common to look at such a difference via the percentage relative improvement in average loss (PRIAL) defined as

$${\rm PRIAL} \stackrel{\rm def}{=} 100\frac{E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{S}}_{\rm LW} - \boldsymbol{S}\|_{\rm F}^2\} - E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{S}}_{\rm RB} - \boldsymbol{S}\|_{\rm F}^2\}}{E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{S}}_{\rm LW} - \boldsymbol{S}\|_{\rm F}^2\}}.$$

To illustrate this quantity two different Hermitian matrices, $\boldsymbol{S}_A$ and $\boldsymbol{S}_B$ were utilized. $\boldsymbol{S}_A$ is the 4×4 'random' choice

$$\boldsymbol{S}_A = \begin{bmatrix} 10 & 7+{\rm i} & 8 & 4 \\ 7-{\rm i} & 12 & 6+2{\rm i} & 5-{\rm i} \\ 8 & 6-2{\rm i} & 15 & 9-3{\rm i} \\ 4 & 5+{\rm i} & 9+3{\rm i} & 10 \end{bmatrix}$$

and the second $\boldsymbol{S}_B$ is set equal to a 10×10 estimated spectral matrix from an EEG dataset. From each of these $\boldsymbol{S}$ matrices, a set of $m = 5000$ matrix estimates $\hat{\boldsymbol{S}}_1, \ldots, \hat{\boldsymbol{S}}_m$ were simulated satisfying (2) and (3). For each replication, estimates were constructed of the form $\hat{\boldsymbol{S}}_{\rm LW}$ and $\hat{\boldsymbol{S}}_{\rm RB}$, and the Frobenius norm between the estimate and the true matrix ($\boldsymbol{S}_A$ or $\boldsymbol{S}_B$) was found. The results were averaged over the 5000 replications to give estimates of $E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{S}}_{\rm LW} - \boldsymbol{S}\|_{\rm F}^2\}$ and $E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{S}}_{\rm RB} - \boldsymbol{S}\|_{\rm F}^2\}$. This was done for $K < p$ (singular case) and $K \geq p$ (non-singular). The results are shown in Fig. 1. Behaviour seems quite smooth as K crosses from the singular to non-singular cases. The Rao-Blackwell estimator offers a useful improvement over the Ledoit-Wolf estimator. In these examples the PRIAL decreases almost monotonically with increasing degrees of freedom, K, but this behaviour need not hold for other choices for $\boldsymbol{S}$.

Note that, analogously to the Ledoit-Wolf estimate of the shrinkage parameter, $\min\{\hat{\rho}_{\rm RB}, 1\}$ provides an estimate for the shrinkage parameter which is constrained by its theoretical upper bound of unity, and would be used in practice.

Remark 4. *In [9] an oracle approximating shrinkage (OAS) estimator was given. The analogous estimator in the complex case for (8) was found to be unpredictable. For example, for $\boldsymbol{S}_A$ while for $K = 2$ the PRIAL (comparing to the Ledoit-Wolf estimator) was increased from 6.5% (Rao-Blackwell) to 15% (OAS), for $K = 4$ it decreased from 5.2% (Rao-Blackwell) to 1.0% (OAS). The behaviour of the Rao-Blackwell estimator seems better suited for practical use. It should also be pointed out that the oracle in (8) is optimal for the stochastic target, while $\hat{\rho}_{\rm LW}$ and $\hat{\rho}_{\rm RB}$ were developed for the deterministic target optimization.*

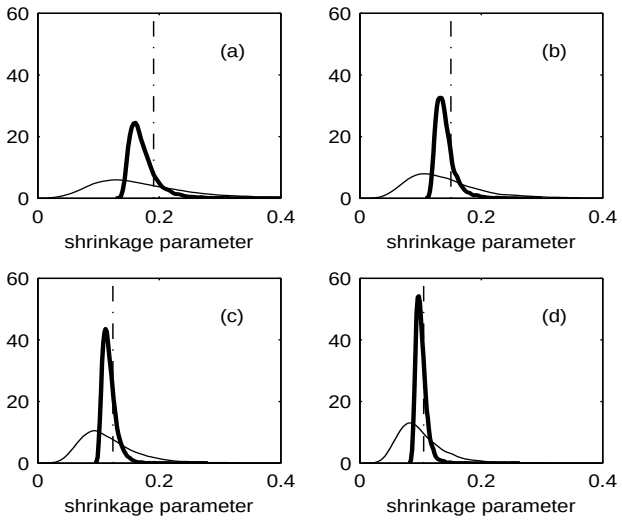

Fig. 2. Simulated distributions for $\hat{\rho}_{\rm LW}$ (thin line) and $\hat{\rho}_{\rm RB}$ (thick line) for the 10×10 matrix $\boldsymbol{S}_B$ for (a) $K = 6$, (b) $K = 8$, (c) $K = 10$ and (d) $K = 12$. The vertical dash-dot line shows the oracle solution ρ_0 of (9).

Fig. 2 compares the empirical distributions of $\hat{\rho}_{\rm LW}$ and $\hat{\rho}_{\rm RB}$ for the matrix $\boldsymbol{S}_B$ ($p = 10$) for (a) $K = 6$, (b) $K = 8$, (c) $K = 10$ and (d) $K = 12$. As expected as K increases, $\hat{\rho}_{\rm LW}$ and $\hat{\rho}_{\rm RB}$ reduce in variance and converge toward the oracle solution. The distribution of $\hat{\rho}_{\rm RB}$ is always preferable to that of $\hat{\rho}_{\rm LW}$.

In the rest of the paper we turn our attention to estimation of inverse spectral matrices.

V. Rao-Blackwell Estimation for Inverse Spectral Matrices

We denote the inverse of the spectral matrix, i.e., the precision matrix, by $\boldsymbol{C} \stackrel{\rm def}{=} \boldsymbol{S}^{-1}$. We shall firstly show that $\hat{\boldsymbol{S}}_{\rm RB}^{-1}$ is actually a "Rao-Blackwellized" estimator for $\boldsymbol{C}$.

Lemma 1. *The inverse, $\hat{\boldsymbol{S}}_{\rm RB}^{-1}$, of the Rao-Blackwell estimator, $\hat{\boldsymbol{S}}_{\rm RB}$, is in the form of a "Rao-Blackwellized" estimator for $\boldsymbol{C}$.*

Proof: Firstly we note that $\hat{\boldsymbol{S}}$ is a sufficient statistic for $\boldsymbol{C}$. To see this we note that the probability density function for $\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1}$ can be written

$$p(\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1}; \boldsymbol{C}) = \pi^{-pK} \det{}^K\{\boldsymbol{C}\} \exp[-K{\rm tr}\{\boldsymbol{C}\hat{\boldsymbol{S}}\}].$$

The part that depends on $\boldsymbol{C}$ only depends on the sample through $\hat{\boldsymbol{S}}$, so this is a sufficient statistic for $\boldsymbol{C}$ by the factorization theorem [19]. Now $\hat{\boldsymbol{S}}_{\rm RB}(\hat{\boldsymbol{S}}) = E\{\hat{\boldsymbol{S}}_{\rm LW}|\hat{\boldsymbol{S}}\}$ is an estimator for $\boldsymbol{S}$, so $\hat{\boldsymbol{S}}_{\rm RB}^{-1}(\hat{\boldsymbol{S}})$ is an estimator for $\boldsymbol{C}$. Recall

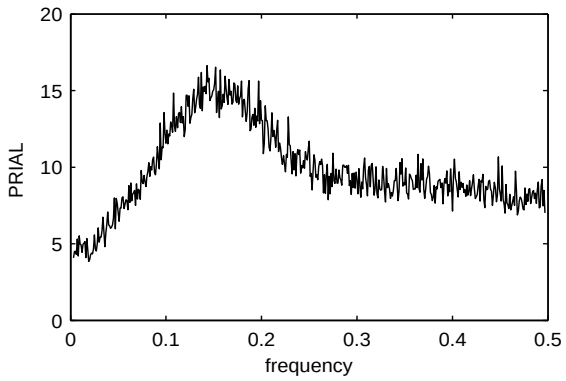

Fig. 3. Estimated PRIAL (%) (improvement of $\hat{\boldsymbol{C}}_{\rm RB}$ over $\hat{\boldsymbol{C}}_{\rm LW}$) for a $\mathrm{VAR}_5(1)$ time series example.

the general result that for a function $h(\cdot)$,

$$E\{h(\hat{\boldsymbol{S}})|\hat{\boldsymbol{S}}\} = h(\hat{\boldsymbol{S}}),$$

so

$$E\{\hat{\boldsymbol{S}}_{\rm RB}^{-1}(\hat{\boldsymbol{S}})|\hat{\boldsymbol{S}}\} = \hat{\boldsymbol{S}}_{\rm RB}^{-1}(\hat{\boldsymbol{S}}) \stackrel{\rm def}{=} \hat{\boldsymbol{C}}_{\rm RB}(\hat{\boldsymbol{S}}),$$

which completes the proof. ∎

Clearly we can use $\hat{\boldsymbol{C}}_{\rm RB}(\hat{\boldsymbol{S}})$ to estimate $\boldsymbol{C}$ when $\hat{\boldsymbol{S}}$ is singular, $K < p$, or non-singular, $K \geq p$.

In order to illustrate the Rao-Blackwellized estimator for $\boldsymbol{C}$ a stable and stationary vector autoregressive process of order 1 and dimension $p = 5$ ($\mathrm{VAR}_5(1)$) was utilized. The process was simulated 5000 times with $N = 1000$ and $K = 4$. Fig. 3 shows the resulting (estimated) PRIAL

$$\mathrm{PRIAL} \stackrel{\rm def}{=} 100\frac{E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{C}}_{\rm LW} - \boldsymbol{C}\|_{\rm F}^2\} - E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{C}}_{\rm RB} - \boldsymbol{C}\|_{\rm F}^2\}}{E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{C}}_{\rm LW} - \boldsymbol{C}\|_{\rm F}^2\}}, \tag{18}$$

where $\hat{\boldsymbol{C}}_{\rm LW} = \hat{\boldsymbol{S}}_{\rm LW}^{-1}$. The PRIAL reaches as much as 15% for some frequencies showing that the Rao-Blackwell approach can be a worthwhile improvement over the Ledoit-Wolf estimator even for dimension $p = 5$.

VI. Random Matrix Approach to Inverse Spectral Matrices

Marzetta *et al.* [28] examined how to manipulate a singular ($K < p$) covariance matrix constructed from circularly-symmetric complex vectors to obtain a non-singular version. In the context of spectral matrices, we can explain their idea as follows.

Firstly an ensemble of $L \times p$ *random* matrices $\boldsymbol{\Phi} \in \mathbb{C}^{L\times p}$, with $L \leq K < p$, is introduced, which have orthonormal rows, so that $\boldsymbol{\Phi}\boldsymbol{\Phi}^H = \boldsymbol{I}_L$. Such matrices are often called 'semi-unitary' and were chosen to be bi-unitarily invariant (see AppendixA). Such matrices are called "isotropically random" with the Haar distribution in [28].

The $L \times L$ matrix $\boldsymbol{\Phi}\hat{\boldsymbol{S}}\boldsymbol{\Phi}^H$ is invertible (with probability one). [28] advocate inverting this matrix and projecting out the result to a $p \times p$ matrix again using the random semi-unitary matrix $\boldsymbol{\Phi}$. Then taking the conditional expectation over the semi-unitary ensemble, gives

$$\hat{\boldsymbol{C}}_L^\star(\hat{\boldsymbol{S}}) \stackrel{\rm def}{=} (p/L)\, E_{\boldsymbol{\Phi}}\{\boldsymbol{\Phi}^H[\boldsymbol{\Phi}\hat{\boldsymbol{S}}\boldsymbol{\Phi}^H]^{-1}\boldsymbol{\Phi} \,\big|\, \hat{\boldsymbol{S}}\},$$

as an estimator for $\boldsymbol{C}$. Although not given explicitly in [28] a rescaling by (p/L) has been included as in [38] so that the estimate of the inverse of the identity matrix is the identity. The term L such that $L < K < p$ is a parameter to be chosen; its determination is discussed later.

Since here $K < p$, the Hermitian matrix $\hat{\boldsymbol{S}}$ has rank $r = \min\{p, K\} = K$ with probability 1. Its spectral decomposition is $\hat{\boldsymbol{S}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^H$, where

$$\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_K, \underbrace{0, \ldots, 0}_{p-K \text{ times}}\}$$

is the diagonal matrix of estimated eigenvalues, (ordered largest to smallest), and $\boldsymbol{U}$ is the unitary matrix having corresponding eigenvectors for its columns. From [28] it follows that

$$\hat{\boldsymbol{C}}_L^\star(\hat{\boldsymbol{S}}) = (p/L)\boldsymbol{U}\,\hat{\boldsymbol{C}}_L^\star(\boldsymbol{\Lambda})\,\boldsymbol{U}^H, \tag{19}$$

so the required estimator can be constructed from $\hat{\boldsymbol{C}}_L^\star(\boldsymbol{\Lambda})$. Further, [28] show that

$$\hat{\boldsymbol{C}}_L^\star(\boldsymbol{\Lambda}) = \mathrm{diag}\{\lambda_1^\star, \ldots, \lambda_K^\star, \lambda^\star, \ldots, \lambda^\star\}, \tag{20}$$

where $\lambda_i^\star$, $i = 1, \ldots, K$ are modified versions of λ_i, $i = 1, \ldots, K$, and the $p - K$ zero eigenvalues of $\hat{\boldsymbol{S}}$ have been replaced by $p - K$ copies of a single value, $\lambda^\star$.

A. Computations via simulations

The computation of $\lambda_i^\star$, $i = 1, \ldots, K$ and $\lambda^\star$ can be carried out purely via simulation, as done by [28] (personal correspondence with Gabriel Tucci). However, for a given $\hat{\boldsymbol{S}}$, in order to get good agreement between the estimator of $\boldsymbol{S}$ derived by averaging many copies of $\boldsymbol{\Phi}^H[\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^H]^{-1}\boldsymbol{\Phi}$ for different $\boldsymbol{\Phi}$, (followed by premultiplication by $\boldsymbol{U}$ and post-multiplication by $\boldsymbol{U}^H$), and the analytic estimator to be described below, the number of copies needing to be averaged is typically very large. For example the order of 10^6 $\boldsymbol{\Phi}$'s were required for the $p = 10$ channel EEG example to achieve agreement to two significant figures. The corresponding compute-time cost turned out to be around 5000 times as heavy, about 500s for the simulation approach versus 0.1s for the analytic scheme at any frequency. Even with modern computational power this sort of simulation burden is not suitable in a spectral matrix context where $\boldsymbol{C}$ must be estimated at possibly thousands of frequencies.

B. Computations using analytic methods

We now examine how to compute (20) using analytic methods. Define $\boldsymbol{D}_K = \mathrm{diag}\{\lambda_1, \ldots, \lambda_K\}$. Then [28, Theorem 1], for a continuous function $g(\cdot)$,

$$\int_{\Omega_0} \frac{1}{K}\mathrm{tr}\{g(\boldsymbol{\Phi}_0^H\boldsymbol{D}_K\boldsymbol{\Phi}_0)\}\mathrm{d}\boldsymbol{\Phi}_0 = \sum_{k=0}^{L-1}\frac{(K-(k+1))!\det\{\boldsymbol{G}_k\}}{(L-(k+1))!\det\{\boldsymbol{V}_K\}} \tag{21}$$

Here $\Omega_0 \stackrel{\rm def}{=} \{\boldsymbol{\Phi}_0 \in \mathbb{C}^{K\times L} : \boldsymbol{\Phi}_0^H\boldsymbol{\Phi}_0 = \boldsymbol{I}_L\}$, these matrices with orthonormal columns again being bi-unitarily invariant (Haar distributed) — see Lemma 4 of AppendixA. $\boldsymbol{V}_K$ is

the Vandermonde matrix associated with $\boldsymbol{D}_K$ given in the 'flipped' form

$$\boldsymbol{V}_K = \begin{bmatrix} \lambda_1^{K-1} & \lambda_2^{K-1} & \cdots & \lambda_K^{K-1} \\ \lambda_1^{K-2} & \lambda_2^{K-2} & \cdots & \lambda_K^{K-2} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1 & \lambda_2 & \cdots & \lambda_K \\ 1 & 1 & \cdots & 1 \end{bmatrix},$$

and $\boldsymbol{G}_k$ is the matrix defined by replacing row $(k+1)$ of the Vandermonde matrix $\boldsymbol{V}_K$, namely $[\lambda_1^{K-(k+1)}, \ldots \lambda_K^{K-(k+1)}]$, by the row

$$\left[I^{(K-L)}\{x^{L-(k+1)} g(x)\}\Big|_{x=\lambda_1}, \ldots, I^{(K-L)}\{x^{L-(k+1)} g(x)\}\Big|_{x=\lambda_K} \right], \quad (22)$$

where $I^{(q)}\{f(x)\}$ denotes q integrations of $f(x)$.

We consider first the computation of $\lambda_i^\star$, for which [28, p. 6265]

$$\lambda_i^\star = \frac{\partial}{\partial \lambda_i} \int_{\Omega_0} \frac{1}{K} \mathrm{tr}\{\log(\boldsymbol{\Phi}_0^H \boldsymbol{D}_K \boldsymbol{\Phi}_0)\} \mathrm{d}\boldsymbol{\Phi}_0. \quad (23)$$

The integral component is given by (21) with $g(\cdot) \equiv \log(\cdot)$. So to compute $\boldsymbol{G}_k$ via (22) we need to know terms like $I^{(q)}\{x^n \log x\}$ for $q \geq 1, n \geq 0$. This is found to be,

$$I^{(q)}\{x^n \log x\} = \frac{x^{n+q} n!}{(n+q)!} \left[\log x - \sum_{j=1}^{q} \frac{1}{n+j} \right].$$

To calculate $\lambda_i^\star$ in (23) we can now use (21),

$$\lambda_i^\star = \sum_{k=0}^{L-1} \frac{(K-(k+1))!}{(L-(k+1))!} \frac{\partial}{\partial \lambda_i} \left[\frac{\det\{\boldsymbol{G}_k\}}{\det\{\boldsymbol{V}_K\}} \right].$$

The partial derivative on the right is given by

$$\frac{\det\{\boldsymbol{V}_K\} \frac{\partial}{\partial \lambda_i} \det\{\boldsymbol{G}_k\} - \det\{\boldsymbol{G}_k\} \frac{\partial}{\partial \lambda_i} \det\{\boldsymbol{V}_K\}}{\det^2\{\boldsymbol{V}_K\}}.$$

To find the derivative of the determinant of a $K \times K$ matrix $\boldsymbol{M}$ ($\boldsymbol{G}_k$ or $\boldsymbol{V}_K$) we first differentiate all entries of the matrix $\boldsymbol{M}$ by λ_i; denote the (l,m)th resulting entry by $A_{l,m}$. Now let $\boldsymbol{B}$ be the cofactor matrix corresponding to $\boldsymbol{M}$. For $1 \leq l, m \leq K$ define $D_{l,m} = A_{l,m} B_{l,m}$, the element-by-element multiplication of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Then the derivative of the determinant is given by [15, eqn. 6]

$$\frac{\partial}{\partial \lambda_i} \det\{\boldsymbol{M}\} = \sum_{l,m=1}^{K} D_{l,m}.$$

For the matrix $\boldsymbol{V}_K$,

$$A_{l,m} = \begin{cases} (K-l)\lambda_i^{K-(l+1)}, & \text{if } m = i; \\ 0, & \text{otherwise.} \end{cases}$$

For $\boldsymbol{G}_k$, entry $A_{l,m}$ is given by

$$\begin{cases} (K-l)\lambda_i^{K-(l+1)}, & \text{if } m = i, l \neq k+1; \\ \frac{\partial}{\partial \lambda_i} I^{(K-L)}\{x^{L-(k+1)} \log(x)\}\Big|_{x=\lambda_i}, & \text{if } m = i, l = k+1; \\ 0, & \text{otherwise,} \end{cases}$$

where of course we can simplify the second term to

$$I^{(K-L-1)}\{x^{L-(k+1)} \log(x)\}\Big|_{x=\lambda_i}.$$

The cofactor matrices for $\boldsymbol{G}_k$ or $\boldsymbol{V}_K$ can be readily found using standard matrix software. Hence we are able to compute $\lambda_i^\star$, $i = 1, \ldots, K$.

The computation of $\lambda^\star$ is straightforward. We know [28, p. 6264] that for $L < K$, $\lambda^\star = \det\{\boldsymbol{G}\}/\det\{\boldsymbol{V}_K\}$ with $\boldsymbol{G}$ being the matrix defined by replacing the Lth row of the Vandermonde matrix $\boldsymbol{V}_K$, namely $[\lambda_1^{K-L}, \ldots, \lambda_K^{K-L}]$, by the row $\left[\lambda_1^{K-(L+1)} \log \lambda_1, \ldots, \lambda_K^{K-(L+1)} \log \lambda_K\right]$. We are thus able to compute all the components of (20) and therefore $\hat{\boldsymbol{C}}_L^\star(\hat{\boldsymbol{S}})$ in (19).

C. Choice of L

In practice we must choose a suitable value of L to use. Use of the analytic results means we require $L < K$ and we are interested in the singular case $K < p$. To select L we proceed by seeking $L = \hat{L}$ that minimizes the predictive risk defined as

$$\mathrm{PR}(\ell) = E\left\{ E_{\tilde{\boldsymbol{J}}}\{\|\hat{\boldsymbol{C}}_\ell^\star \tilde{\boldsymbol{J}} \tilde{\boldsymbol{J}}^H - \boldsymbol{I}_p\|_{\mathrm{F}}^2 | \boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1}\} \right\}$$

where $\hat{\boldsymbol{C}}_\ell^\star$ is the estimated inverse spectral matrix found from $\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1}$ when $L = \ell$, and $\tilde{\boldsymbol{J}}$ is independent of the $\boldsymbol{J}_k$'s and from the same distribution. Here we have used quadratic loss which does not involve any further matrix inversions. We approximate the predictive risk using leave-one-out cross-validation. Specifically, the estimate of the predictive risk is

$$\widehat{\mathrm{PR}}(\ell) = \frac{1}{K} \sum_{j=1}^{K} \|\hat{\boldsymbol{C}}_\ell^{\star[j]} \boldsymbol{J}_j \boldsymbol{J}_j^H - \boldsymbol{I}_p\|_{\mathrm{F}}^2,$$

where $\hat{\boldsymbol{C}}_\ell^{\star[j]}$ denotes the estimated inverse spectral matrix found from $\boldsymbol{J}_0, \ldots, \boldsymbol{J}_{K-1}$ excluding $\boldsymbol{J}_j$. Then we take

$$\hat{L} = \arg\min_\ell \widehat{\mathrm{PR}}(\ell). \quad (24)$$

Note that using this scheme it is only possible to consider values of $\ell < K-1$ since we know that ordinarily L must be less than K but additionally here $\hat{\boldsymbol{C}}_\ell^{[j]}$ is derived from $K-1$ of the $\boldsymbol{J}_j$'s.

D. Example

In order to illustrate the random matrix estimator $\hat{\boldsymbol{C}}_L^\star(\hat{\boldsymbol{S}})$ for $\boldsymbol{C}$ in a time series context, a stable and stationary vector autoregressive process of order 1 and dimension $p = 10$ ($\mathrm{VAR}_{10}(1)$) was utilized with $N = 1000$ and $K = 8$. At each frequency (24) was used to choose L. Fig. 4 shows the resulting (estimated) PRIAL

$$\mathrm{PRIAL} \stackrel{\mathrm{def}}{=} 100 \frac{E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{C}}_{\mathrm{LW}} - \boldsymbol{C}\|_{\mathrm{F}}^2\} - E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{C}}_L^\star - \boldsymbol{C}\|_{\mathrm{F}}^2\}}{E_{\boldsymbol{S}}\{\|\hat{\boldsymbol{C}}_{\mathrm{LW}} - \boldsymbol{C}\|_{\mathrm{F}}^2\}}. \quad (25)$$

This estimated PRIAL was found from 100 replications and because of the need to produce the replications computations were carried out only at every 10th Fourier frequency. The PRIAL reaches nearly 20% for some frequencies again showing a worthwhile improvement over the Ledoit-Wolf estimator.

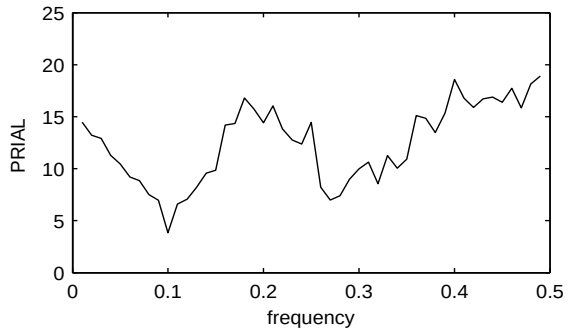

Fig. 4. Estimated PRIAL (%) (improvement of $\hat{\boldsymbol{C}}_L^\star$ over $\hat{\boldsymbol{C}}_{\mathrm{LW}}$) for a $\mathrm{VAR}_{10}(1)$ time series example.

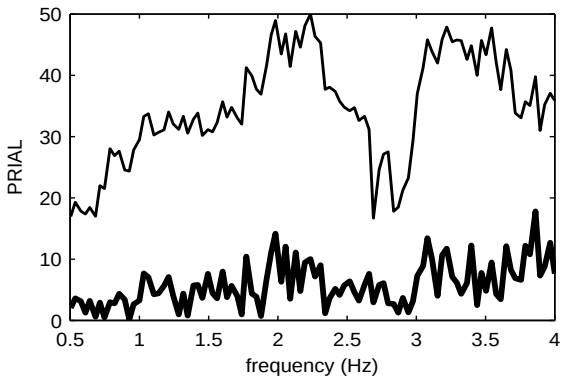

Fig. 5. Estimated PRIAL (%) for EEG data. Improvement of $\hat{\boldsymbol{C}}_{\mathrm{RB}}$ over $\hat{\boldsymbol{C}}_{\mathrm{LW}}$ is shown by the thick line. Improvement of $\hat{\boldsymbol{C}}_L^\star$ over $\hat{\boldsymbol{C}}_{\mathrm{LW}}$ is shown by the thin line.

VII. Application to EEG data

We now compute $\hat{\boldsymbol{C}}_{\mathrm{RB}}$ and $\hat{\boldsymbol{C}}_L^\star$ for electroencephalogram (EEG) data, (resting conditions with eyes closed), for a patient diagnosed with positive syndrome schizophrenia. Interest was in the delta frequency range, $0.5 < f \leq 4$Hz, see [29]. EEG was recorded on the scalp at 10 sites, so $\{\boldsymbol{X}_t\}$ is a $p = 10$ vector-valued process, using a bandpass filter of 0.5–45Hz and sample interval of $\Delta_{\mathrm{t}} = 0.01$s. To remove the dominant and contaminating 10Hz alpha rhythm, which would otherwise cause severe spectral leakage, the data was low-pass filtered and resampled to a sample interval of $\Delta_{\mathrm{t}} = 0.05$s. After this downsampling $N = 612$.

Using this real data the spectral matrix $\boldsymbol{S}(f)$ was estimated as $\boldsymbol{S}_0(f)$, say, for $|f| \leq f_{\mathcal{N}}$, using $K = 40$ tapers. Using the vector-valued circulant embedding approach, [6], 100 independent Gaussian p-vector-valued time series ($p = 10$) were computed, each having $\boldsymbol{S}_0(f)$, $|f| \leq f_{\mathcal{N}}$, as its true spectral matrix. For each of these time series the singular matrix $\hat{\boldsymbol{S}}(f)$ was computed using multitaper estimation with $K = 8$ tapers for 100 frequencies equally spaced between 0.5 and 4Hz, and from these estimates $\hat{\boldsymbol{C}}_{\mathrm{RB}}$ and $\hat{\boldsymbol{C}}_L^\star$ were computed, (with (24) choosing L for $\hat{\boldsymbol{C}}_L^\star$). The estimated PRIAL — with $\boldsymbol{C} = \boldsymbol{S}_0^{-1}$ — was then found over the 100 replications. In this way the simulation experiment mimicks the spectral properties of the EEG data while providing calibrated results, which are shown in Fig. 5. We see that both schemes improve on the LW method, but that $\hat{\boldsymbol{C}}_L^\star$ does particularly well, with PRIAL reaching 50%.

VIII. Concluding Discussion

We have described two analytical estimators (Rao-Blackwell and random matrix) for the spectral precision matrix. Interestingly, $\hat{\boldsymbol{C}}_{\mathrm{RB}}$ is the inverse of a shrinkage estimator where the shrinkage parameter is obtained as a conditional expectation, conditional on $\hat{\boldsymbol{S}}$, while the random matrix estimator $\hat{\boldsymbol{C}}_L^\star$ is also a conditional expectation, again conditioned on $\hat{\boldsymbol{S}}$. We have shown that both hold promise for being useful in practice, offering possibly substantial improvements over the inverse of the LW estimator of $\boldsymbol{C}$. Further investigation of their properties seems worthwhile.

Appendix

To simplify notation we drop explicit frequency dependence.

A. Bi-unitary invariance

Definition 1. *A complex-valued $n \times m$ random matrix $\boldsymbol{Z}$ is right(left)-unitarily invariant if its distribution is invariant under the transformation $\boldsymbol{Z} \to \boldsymbol{Z}\boldsymbol{\Theta}$ ($\boldsymbol{Z} \to \boldsymbol{\Upsilon}\boldsymbol{Z}$) where $\boldsymbol{\Theta} \in \mathcal{U}(m), \boldsymbol{\Upsilon} \in \mathcal{U}(n)$, where $\mathcal{U}(n)$ is the compact group of all $n \times n$ complex unitary matrices, i.e., $\mathcal{U}(n) = \{\boldsymbol{U}_{n\times n} : \boldsymbol{U}^H\boldsymbol{U} = \boldsymbol{I}_n\}$. If both are true we say $\boldsymbol{Z}$ is bi-unitarily invariant.*

Lemma 2. *The matrix $\boldsymbol{J}$ defined in (1) with $\boldsymbol{J}_k$ given by (3) is right-unitarily invariant. (If $\boldsymbol{S} = \boldsymbol{I}_p$ it is bi-unitarily invariant.)*

Proof: This follows from [22, p. 487]. ∎

Lemma 3. *When considered as a metric space $\mathcal{U}(n)$ is measurable. There is a unique left-unitarily invariant probability measure μ for $\mathcal{U}(n)$ such that $\mu(\boldsymbol{\Theta}\boldsymbol{A}) = \mu(\boldsymbol{A})$ for any measurable $\boldsymbol{A} \subset \mathcal{U}(n)$ and any $\boldsymbol{\Theta} \in \mathcal{U}(n)$. Moreover, since $\mathcal{U}(n)$ is compact, the same measure μ is also right-unitarily invariant. The Haar measure is this unique probability measure μ on $\mathcal{U}(n)$ that is bi-unitarily invariant. See [37, p. 108].*

Remark 5. *Let $\boldsymbol{\Upsilon} \in \mathcal{U}(n)$. If $\boldsymbol{\Upsilon}$ has Haar measure then for all $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \mathcal{U}(n)$, $p(\boldsymbol{\Theta}_1\boldsymbol{\Upsilon}\boldsymbol{\Theta}_2) = p(\boldsymbol{\Upsilon})$, where $p(\boldsymbol{\Upsilon})$ denotes the joint probability density function of the components of the unitary matrix.*

Lemma 4. *Let $\boldsymbol{\Upsilon} \in \mathcal{U}(n)$ equipped with Haar measure. We now consider two specific truncations of the $n \times n$ unitary matrices. Suppose we partition $\boldsymbol{\Upsilon}$ in two ways:*

$$\boldsymbol{\Upsilon} = \left[\begin{array}{c} \boldsymbol{\Phi} \\ \hline \boldsymbol{P}_{(n-m)\times n} \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{\Phi}_0 & \boldsymbol{Q}_{n\times(n-m)} \end{array}\right],$$

where $\boldsymbol{\Phi}$ is $m \times n, m < n$ and $\boldsymbol{\Phi}_0$ is $n \times m, m < n$. Then $\boldsymbol{\Upsilon} \to \boldsymbol{\Phi}$ maps the unitary group onto the Stiefel manifold of $m \times n$ matrices with orthonormal rows, $\boldsymbol{\Phi}\boldsymbol{\Phi}^H = \boldsymbol{I}_m$. The image of the Haar measure under this map is bi-unitarily invariant. Likewise, $\boldsymbol{\Upsilon} \to \boldsymbol{\Phi}_0$ maps the unitary group onto the Stiefel manifold of $n \times m$ matrices with orthonormal columns, $\boldsymbol{\Phi}_0^H\boldsymbol{\Phi}_0 = \boldsymbol{I}_m$. The image of the Haar measure under this map is again bi-unitarily invariant. See [14].

B. Results required for proof of Theorem 3

Theorem 4. *We know that the singular value decomposition (SVD) for the $p \times K$ random matrix $\boldsymbol{J}$ defined by (1) and (3) is [1, p. 182] $\boldsymbol{J} = \boldsymbol{U}\boldsymbol{\Psi}\boldsymbol{V}^H$, where $\boldsymbol{U} \in \mathcal{U}(p), \boldsymbol{V} \in \mathcal{U}(K)$ and $\boldsymbol{\Psi}$ is the $p \times K$ matrix*

$$\boldsymbol{\Psi} = \left[\begin{array}{c|c} \boldsymbol{\Omega} & \boldsymbol{0}_{r\times(K-r)} \\ \hline \boldsymbol{0}_{(p-r)\times r} & \boldsymbol{0}_{(p-r)\times(K-r)} \end{array}\right],$$

$\boldsymbol{\Omega}$ is the diagonal matrix $\boldsymbol{\Omega} = \mathrm{diag}\{\omega_1, \ldots, \omega_r\}$, $\omega_i = \lambda_i^{1/2}$, the square root of the ith ordered eigenvalue $\lambda_i(\boldsymbol{J}\boldsymbol{J}^H) = \lambda_i(\boldsymbol{J}^H\boldsymbol{J})$. Here $r = \mathrm{rank}\{\boldsymbol{J}\} = \mathrm{rank}\{\boldsymbol{J}\boldsymbol{J}^H\} = \mathrm{rank}\{\boldsymbol{J}^H\boldsymbol{J}\}$. Further $r = \min\{p, K\}$ with probability 1. Then,

1) *$\{\boldsymbol{U}, \boldsymbol{\Omega}\}$ and $\boldsymbol{V}$ are statistically independent.*
2) *$\boldsymbol{V}$ is a bi-unitarily invariant unitary matrix.*

Proof: 1. We firstly show that $\{\boldsymbol{U}, \boldsymbol{\Omega}\}$ and $\boldsymbol{V}$ are statistically independent.

Let $\boldsymbol{U} = [\boldsymbol{U}_0, \boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_p] = [\boldsymbol{U}_0 \,|\, \boldsymbol{U}_1]$ and let $\boldsymbol{V} = [\boldsymbol{V}_0, \boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_K] = [\boldsymbol{V}_0 \,|\, \boldsymbol{V}_1]$. The full SVD $\boldsymbol{J} = \boldsymbol{U}\boldsymbol{\Psi}\boldsymbol{V}^H$ can be written in the form

$$\boldsymbol{J} = [\boldsymbol{U}_0 \,|\, \boldsymbol{U}_1]\boldsymbol{\Psi}\left[\begin{array}{c} \boldsymbol{V}_0^H \\ \hline \boldsymbol{V}_1^H \end{array}\right].$$

Now consider two cases

- $K \le p$. In this case, $r = K$ and

$$\boldsymbol{J} = [\boldsymbol{U}_0 \,|\, \boldsymbol{U}_1]\left[\begin{array}{c} \boldsymbol{\Omega} \\ \hline \boldsymbol{0}_{(p-K)\times K} \end{array}\right]\boldsymbol{V}^H. \tag{26}$$

- $K > p$ In this case, $r = p$ and

$$\boldsymbol{J} = \boldsymbol{U}\left[\begin{array}{c|c} \boldsymbol{\Omega} & \boldsymbol{0}_{p\times(K-p)} \end{array}\right]\left[\begin{array}{c} \boldsymbol{V}_0^H \\ \hline \boldsymbol{V}_1^H \end{array}\right]. \tag{27}$$

Write $\boldsymbol{J} = \boldsymbol{A} + \mathrm{i}\boldsymbol{B}$. The probability density is given by [22, eqn. 78]

$$\pi^{-pK}|\boldsymbol{S}|^{-K}\exp^{-\mathrm{tr}\{\boldsymbol{S}^{-1}\boldsymbol{J}\boldsymbol{J}^H\}}\prod_{i=1}^{p}\prod_{j=1}^{K}\mathrm{d}A_{ij}\mathrm{d}B_{ij}. \tag{28}$$

$\mathrm{d}A_{ij}$ is the i,j-th element of $\mathrm{d}\boldsymbol{A}$ and $\prod_{i=1}^{p}\prod_{j=1}^{K}\mathrm{d}A_{ij}\mathrm{d}B_{ij}$ is the volume element. Since we are interested in transforming $\boldsymbol{J}$ it is convenient to use another notation for the volume element, viz $(\mathrm{d}\boldsymbol{J})$, so that (28) becomes

$$\pi^{-pK}|\boldsymbol{S}|^{-K}\exp^{-\mathrm{tr}\{\boldsymbol{S}^{-1}\boldsymbol{J}\boldsymbol{J}^H\}}(\mathrm{d}\boldsymbol{J}) \tag{29}$$

which relates the volume element to the exterior product notation:

$$(\mathrm{d}\boldsymbol{J}) \stackrel{\mathrm{def}}{=} (\mathrm{d}\boldsymbol{A})(\mathrm{d}\boldsymbol{B}).$$

where $(\mathrm{d}\boldsymbol{A}) = \wedge_{j=1}^{K}\wedge_{i=1}^{p}\mathrm{d}A_{ij}$; see [31, Chapter 2]. Now we return to the case of $K \le p$ and consider the 'thin' SVD corresponding to (26). It takes the form

$$\boldsymbol{J} = \boldsymbol{U}_0\boldsymbol{\Omega}\boldsymbol{V}^H. \tag{30}$$

The transformation $\boldsymbol{J} \to \boldsymbol{U}_0\boldsymbol{\Omega}\boldsymbol{V}^H$ was studied in [33] who found the volume element $(\mathrm{d}\boldsymbol{J})$ to be proportional to

$$[\det\{\boldsymbol{\Omega}\}]^{2p-2K+1}\prod_{k<l}^{K}(\omega_k^2 - \omega_l^2)^2(\boldsymbol{\Omega})(\boldsymbol{U}_0\mathrm{d}\boldsymbol{U}_0)(\boldsymbol{V}\mathrm{d}\boldsymbol{V}^H). \tag{31}$$

In (29), $\pi^{-pK}|\boldsymbol{S}|^{-K}\exp^{-\mathrm{tr}\{\boldsymbol{S}^{-1}\boldsymbol{J}\boldsymbol{J}^H\}}$ becomes

$$\pi^{-pK}|\boldsymbol{S}|^{-K}\exp^{-\mathrm{tr}\{\boldsymbol{S}^{-1}\boldsymbol{U}_0\boldsymbol{\Omega}^2\boldsymbol{U}_0^H\}}. \tag{32}$$

The product of (32) and the volume element (31) shows that the probability density can be factored into functions of $\{\boldsymbol{U}_0, \boldsymbol{\Omega}\}$ and $\boldsymbol{V}$. Now $\boldsymbol{U} = [\boldsymbol{U}_0 \,|\, \boldsymbol{U}_1]$ and in order for $\boldsymbol{U}$ to be unitary, $\boldsymbol{U}_1$ depends totally on $\boldsymbol{U}_0$. Hence $\boldsymbol{V}$ is independent of $\boldsymbol{U}$ and $\boldsymbol{\Omega}$.

For the case $K > p$ consider the 'thin' SVD corresponding to (27), i.e., $\boldsymbol{J} = \boldsymbol{U}\boldsymbol{\Omega}\boldsymbol{V}_0^H$. Then the probability density can be factored into functions of $\{\boldsymbol{U}, \boldsymbol{\Omega}\}$ and $\boldsymbol{V}_0$. Now $\boldsymbol{V} = [\boldsymbol{V}_0 \,|\, \boldsymbol{V}_1]$ and in order for $\boldsymbol{V}$ to be unitary, $\boldsymbol{V}_1$ depends totally on $\boldsymbol{V}_0$. Hence $\boldsymbol{V}$ is again independent of $\boldsymbol{U}$ and $\boldsymbol{\Omega}$. ■

2. We now show that the unitary matrix $\boldsymbol{V}$ is bi-unitarily invariant.

Proof: Note that $\boldsymbol{J}^H\boldsymbol{J} = \boldsymbol{V}\boldsymbol{\Psi}^2\boldsymbol{V}^H = \boldsymbol{V}\boldsymbol{\Lambda}_K\boldsymbol{V}^H$, with

$$\boldsymbol{\Lambda}_K = \left[\begin{array}{ccc|c} \lambda_1 & & & \\ & \ddots & & \boldsymbol{0}_{r\times(K-r)} \\ & & \lambda_r & \\ \hline \multicolumn{3}{c|}{\boldsymbol{0}_{(K-r)\times r}} & \boldsymbol{0}_{(K-r)\times(K-r)} \end{array}\right].$$

Since $\boldsymbol{J}$ is right-unitarily invariant (Lemma 2) we know that $\boldsymbol{J}$ and $\boldsymbol{J}\boldsymbol{\Theta}^H$ have the same distribution for $\boldsymbol{\Theta}^H \in \mathcal{U}(K)$. Hence, with $\stackrel{\mathrm{d}}{=}$ denoting "equal in distribution,"

$$\boldsymbol{J}^H\boldsymbol{J} \stackrel{\mathrm{d}}{=} (\boldsymbol{J}\boldsymbol{\Theta}^H)^H(\boldsymbol{J}\boldsymbol{\Theta}^H) = \boldsymbol{\Theta}\boldsymbol{J}^H\boldsymbol{J}\boldsymbol{\Theta}^H = (\boldsymbol{\Theta}\boldsymbol{V})\boldsymbol{\Lambda}_K(\boldsymbol{\Theta}\boldsymbol{V})^H$$

and so $\boldsymbol{V}\boldsymbol{\Lambda}_K\boldsymbol{V}^H \stackrel{\mathrm{d}}{=} (\boldsymbol{\Theta}\boldsymbol{V})\boldsymbol{\Lambda}_K(\boldsymbol{\Theta}\boldsymbol{V})^H$. The random components of $\boldsymbol{\Lambda}_K$ are functions of the random components of $\boldsymbol{\Omega}$, and $\boldsymbol{V}$ is independent of $\boldsymbol{U}$ and $\boldsymbol{\Omega}$, so $\boldsymbol{V}$ and $\boldsymbol{\Lambda}_K$ are independent. Then, $\boldsymbol{V} \stackrel{\mathrm{d}}{=} \boldsymbol{\Theta}\boldsymbol{V}$. Since the distribution of $\boldsymbol{V}$ is left-unitarily invariant and $\boldsymbol{V} \in \mathcal{U}(K)$, we know from Lemma 3 of AppendixA that it is also right-unitarily invariant, and hence is a bi-unitarily invariant unitary matrix. This completes the proof. ■

Lemma 5. *With the $K \times K$ matrix $\boldsymbol{V}$ defined as in Theorem 4, let $v_{jk} = (\boldsymbol{V})_{jk}$. Then for $1 \le j, k, l \le K, j \ne l$,*

$$E\{|v_{kj}|^4\} = 2/[K(K+1)] \tag{33}$$
$$E\{|v_{kj}|^2 \cdot |v_{kl}|^2\} = 1/[K(K+1)]. \tag{34}$$

Proof: The bi-unitarily invariant nature of the unitary matrix $\boldsymbol{V}$ is sufficient [20, p. 812] for the stated moment results of [20, Proposition 1.2] to hold, in particular (33) and (34). ■

Lemma 6. *We can write*

$$E\left\{\sum_{k=0}^{K-1}||\hat{\boldsymbol{S}}_k - \hat{\boldsymbol{S}}||_{\mathrm{F}}^2 \;|\hat{\boldsymbol{S}}\right\} = \sum_{k=0}^{K-1}E\{||\boldsymbol{J}_k||_2^4|\hat{\boldsymbol{S}}\} - Ktr\{\hat{\boldsymbol{S}}^2\}.$$

Proof: Expanding the expectation on the left we get

$$\sum_{k=0}^{K-1}E\{\mathrm{tr}\{\boldsymbol{J}_k\boldsymbol{J}_k^H\boldsymbol{J}_k\boldsymbol{J}_k^H\}|\hat{\boldsymbol{S}}\} - \sum_{k=0}^{K-1}E\{\mathrm{tr}\{\hat{\boldsymbol{S}}\boldsymbol{J}_k\boldsymbol{J}_k^H\}|\hat{\boldsymbol{S}}\} - \sum_{k=0}^{K-1}E\{\mathrm{tr}\{\boldsymbol{J}_k\boldsymbol{J}_k^H\hat{\boldsymbol{S}}\}|\hat{\boldsymbol{S}}\} + \sum_{k=0}^{K-1}E\{\mathrm{tr}\{\hat{\boldsymbol{S}}^2\}|\hat{\boldsymbol{S}}\}$$

Now,

$$\text{tr}\{\boldsymbol{J}_k\boldsymbol{J}_k^H\boldsymbol{J}_k\boldsymbol{J}_k^H\} = \text{tr}\{\boldsymbol{J}_k^H\boldsymbol{J}_k\boldsymbol{J}_k^H\boldsymbol{J}_k\} = (\boldsymbol{J}_k^H\boldsymbol{J}_k)^2 = ||\boldsymbol{J}_k||_2^4,$$

so the first term is simply $\sum_{k=0}^{K-1} E\{||\boldsymbol{J}_k||_2^4|\hat{\boldsymbol{S}}\}$. For the second term in the expansion we get

$$-E\{\text{tr}\{\hat{\boldsymbol{S}}\sum_k \boldsymbol{J}_k\boldsymbol{J}_k^H\}|\hat{\boldsymbol{S}}\} = -E\{\text{tr}\{K\hat{\boldsymbol{S}}^2\}|\hat{\boldsymbol{S}}\} = -K\text{tr}\{\hat{\boldsymbol{S}}^2\}.$$

Terms three and four follow likewise to give the result. ■

Lemma 7.

$$E\{||\boldsymbol{J}_k||_2^4|\hat{\boldsymbol{S}}\} = \frac{K}{K+1}\left[tr\{\hat{\boldsymbol{S}}^2\} + tr^2\{\hat{\boldsymbol{S}}\}\right].$$

Proof: We adopt the approach of [9, Lemma 3], although details and the result are different. Now

$$K\hat{\boldsymbol{S}} = \boldsymbol{J}\boldsymbol{J}^H = \boldsymbol{U}\boldsymbol{\Psi}\boldsymbol{\Psi}^H\boldsymbol{U}^H = \boldsymbol{U}\boldsymbol{\Lambda}_p\boldsymbol{U}^H, \qquad (35)$$

where, with $\lambda_i \in \mathbb{R}$,

$$\boldsymbol{\Lambda}_p = \left[\begin{array}{ccc|c} \lambda_1 & & & \\ & \ddots & & \boldsymbol{0}_{r\times(p-r)} \\ & & \lambda_r & \\ \hline \multicolumn{3}{c|}{\boldsymbol{0}_{(p-r)\times r}} & \boldsymbol{0}_{(p-r)\times(p-r)} \end{array}\right].$$

Let $\boldsymbol{V}^H = [\boldsymbol{\nu}_0, \ldots, \boldsymbol{\nu}_{K-1}]$ so that $\boldsymbol{J}_k = \boldsymbol{U}\boldsymbol{\Psi}\boldsymbol{\nu}_k$ and

$$\boldsymbol{J}_k^H\boldsymbol{J}_k = \boldsymbol{\nu}_k^H\boldsymbol{\Psi}^H\boldsymbol{\Psi}\boldsymbol{\nu}_k = \boldsymbol{\nu}_k^H\boldsymbol{\Lambda}_K\boldsymbol{\nu}_k.$$

Consequently,

$$\begin{aligned} E\{||\boldsymbol{J}_k||_2^4|\hat{\boldsymbol{S}}\} &= E\{(\boldsymbol{\nu}_k^H\boldsymbol{\Lambda}_K\boldsymbol{\nu}_k)^2|\hat{\boldsymbol{S}}\} \\ &= E\{E\{(\boldsymbol{\nu}_k^H\boldsymbol{\Lambda}_K\boldsymbol{\nu}_k)^2|\hat{\boldsymbol{S}},\boldsymbol{\Lambda}_K\}|\hat{\boldsymbol{S}}\}. \quad (36) \end{aligned}$$

- $\hat{\boldsymbol{S}}$ depends on $\boldsymbol{U}$ and $\boldsymbol{\Lambda}_p$ and the random components of $\boldsymbol{\Lambda}_p$ are functions of the random components of $\boldsymbol{\Omega}$.
- The random components of $\boldsymbol{\Lambda}_K$ are functions of the random components of $\boldsymbol{\Omega}$.
- $\boldsymbol{\nu}_k$ is a function of $\boldsymbol{V}$.

Now $\boldsymbol{V}$ is independent of $\boldsymbol{U}$ and $\boldsymbol{\Omega}$ by Theorem 4. Therefore, for the inner conditional expectation of (36) we know that $E\{(\boldsymbol{\nu}_k^H\boldsymbol{\Lambda}_K\boldsymbol{\nu}_k)^2|\hat{\boldsymbol{S}},\boldsymbol{\Lambda}_K\}$ is given by

$$\begin{aligned} &\sum_{j=1}^r \lambda_j^2 E\{|\nu_{jk}|^4\} + \sum_{j\neq l}^r \lambda_j\lambda_l E\{|\nu_{jk}|^2|\nu_{lk}|^2\} \\ &\quad = \sum_{j=1}^r \lambda_j^2 E\{|v_{kj}|^4\} + \sum_{j\neq l}^r \lambda_j\lambda_l E\{|v_{kj}|^2|v_{kl}|^2\} \end{aligned}$$

where $v_{kl} = (\boldsymbol{V})_{kl}$. Then using (33) and (34), we see that

$$\begin{aligned} E\{(\boldsymbol{\nu}_k^H\boldsymbol{\Lambda}_K\boldsymbol{\nu}_k)^2|\hat{\boldsymbol{S}},\boldsymbol{\Lambda}_K\} &= \frac{1}{K(K+1)}\left[2\sum_{j=1}^r\lambda_j^2 + \sum_{j\neq l}^r \lambda_j\lambda_l\right] \\ &= \frac{1}{K(K+1)}\left[\sum_{j=1}^r\lambda_j^2 + \sum_{j,l}^r \lambda_j\lambda_l\right] \\ &= \frac{1}{K(K+1)}\left[\text{tr}\{\boldsymbol{\Lambda}_p^2\} + \text{tr}^2\{\boldsymbol{\Lambda}_p\}\right] \\ &= \frac{K}{K+1}\left[\text{tr}\{\hat{\boldsymbol{S}}^2\} + \text{tr}^2\{\hat{\boldsymbol{S}}\}\right], \end{aligned}$$

since from (35) we have that

$$\text{tr}\{\boldsymbol{\Lambda}_p^2\} = K^2\text{tr}\{\hat{\boldsymbol{S}}^2\} \quad \text{and} \quad \text{tr}^2\{\boldsymbol{\Lambda}_p\} = K^2\text{tr}^2\{\hat{\boldsymbol{S}}\}.$$

Taking the outer expectation conditional on $\hat{\boldsymbol{S}}$ changes nothing, which completes the proof. ■

ACKNOWLEDGMENT

The work of Deborah Schneider-Luftman was supported by EPSRC (UK).

REFERENCES

[1] D. S. Bernstein, *Matrix Mathematics*. Princeton, NJ: Princeton University Press, 2005.

[2] H. Böhm, and R. von Sachs, "Shrinkage sstimation in the frequency domain of multivariate time series," *Journal of Multivariate Analysis*, vol. 100, pp. 919–35, 2009.

[3] T. Cai, W. Liu, and X. Luo, ' 'A constrained l1 minimization approach to sparse precision matrix estimation", *J. Amer. Statist. Assoc.*, vol. 106, pp. 594–607, 2011.

[4] G. Casella and R. L. Berger, *Statistical Inference*. Belmont, CA: Duxbury, 1990.

[5] S. Chandna and A. T. Walden, "Statistical properties of the estimator of the rotary coefficient," *IEEE Trans. Signal Process.*, vol. 59, pp. 1298–1303, 2011.

[6] S. Chandna and A. T. Walden, "Simulation methodology for inference on physical parameters of complex vector-valued signals," *IEEE Trans. Signal Process.*, vol. 60, pp. 5260–5269, 2013.

[7] X. Chen, Y-H. Kim, and Z. Jane Wang, "Efficient minimax estimation of a class of high-dimensional sparse precision matrices," *IEEE Trans. Signal Process.* vol. 60, pp. 2899–2912, 2012.

[8] X. Chen, Z. Jane Wang, and M. J. McKeown, "Shrinkage-to-tapering estimation of large covariance matrices," *IEEE Trans. Signal Process.*, vol. 60, pp. 5640–5656, 2012.

[9] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for MMSE covariance estimation," *IEEE Trans. Signal Process.* vol. 58, pp. 5016–5029, 2010.

[10] D. K. Dey and C. Srinivasan, "Estimation of covariance matrix under Stein's loss," *The Annals of Statistics*, vol. 13, pp. 1581–1591, 1985.

[11] B. Efron and C. Morris, "Multivariate empirical Bayes and estimation of covariance matrices," *The Annals of Statistics*, vol. 4, pp. 22–32, 1976.

[12] M. Fiecas and H. Ombao, "The generalized shrinkage estimator for the analysis of functional connectivity of brain signals," *The Annals of Applied Statistics*, vol. 5, pp. 1102–25, 2011.

[13] T. J. Fisher and X. Sun, "Improved Stein-type shrinkage estimators for the high-dimensional normal covariance matrix," *Computational Statistics and Data Analysis*, vol. 55, pp. 1909–18, 2011.

[14] Y. V. Fyodorov and B. A. Khoruzhenko, "A few remarks on colour-flavour transformations, truncations of random unitary matrices, Berezin reproducing kernels and Selberg-type integrals," *J. Phys. A: Math. Theor.*, vol. 40, pp. 669–699, 2007.

[15] M. A. Golberg, "The derivative of a determinant," *The American Mathematical Monthly*, vol. 79, pp. 1124–1126, 1972.

[16] N. R. Goodman, "Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction)," *Ann. Math. Statist.*, vol. 34, pp. 152–77, 1963.

[17] L. R. Haff, "Estimation of the inverse covariance matrix: random mixtures of the inverse Wishart matrix and the identity," *The Annals of Statistics*, vol. 7, pp. 1264–1276, 1979.

[18] L. R. Haff, "Empirical Bayes estimation of the multivariate normal covariance matrix," *The Annals of Statistics*, vol. 8, pp. 586–597, 1980.

[19] P. R. Halmos and L. J. Savage, "Applications of the Radon-Nikodym Theorem to the theory of sufficient statistics," *Annals of Mathematical Statistics*, vol. 20, pp. 225-41, 1949.

[20] F. Hiai and D. Petz, "Asymptotic freeness almost everywhere for random matrices," *Acta. Sci. Math. (Szeged)*, vol. 66, pp. 809–34, 2000.

[21] W. James and C. Stein, C. "Estimation with quadratic loss," in *Proceedings of the Fourth Berke- ley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Berkeley, CA: University of California Press, pp. 361–379, 1961.

[22] A. T. James, "Distributions of matrix variates and latent roots derived from normal samples," *Ann. Math. Statist.*, vol. 35, pp. 475–501, 1964.

[23] C. Lam and J. Fan, “Sparsistency and rates of convergence in large covariance matrix estimation”, *Ann. Statist.*, vol. 37, pp. 4254–4278, 2009.
[24] O. Ledoit and M. Wolf, “Improved estimation of the covariance matrix of stock returns with an application to portfolio selection,” *Journal of Empirical Finance*, vol. 10, pp. 603–21, 2003.
[25] O. Ledoit and M. Wolf, “A well-conditioned estimator for large-dimensional covariance matrices,” *Journal of Multivariate Analysis* vol. 88, pp. 365–411, 2004.
[26] O. Ledoit and M. Wolf, “Nonlinear shrinkage estimation of large-dimensional covariance matrices,” *The Annals of Statistics*, vol. 40, pp. 1024–1060, 2012.
[27] D. Maiwald and D. Kraus, “Calculation of moments of complex Wishart and complex inverse Wishart distributed matrices,” *IEE Proceedings Radar, Sonar Navigation*, vol. 147, pp. 162–168, 2000.
[28] T. L. Marzetta, G. H. Tucci and S. H. Simon, “A random matrix-theoretic approach to handling singular covariance matrices,” *IEEE Trans. Information Theory* vol. 57, pp. 6256–6271, 2011.
[29] T. Medkour, A. T. Walden, A. P. Burgess & V. B. Strelets, “Brain connectivity in positive and negative syndrome schizophrenia,” *Neuroscience*, vol. 169, pp. 1779–88.
[30] N. Meinshausen and P. Bühlmann, “High-dimensional graphs and” variable selection with the Lasso”, *Ann. Statist.*, vol. 34, pp. 1436–1462, 2006.
[31] R. J. Muirhead, *Aspects of Multivariate Statistical Theory.* Hoboken NJ: John Wiley, 1982.
[32] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications*. Cambridge, UK: Cambridge University Press, 1993.
[33] T. Ratnarajah and R. Vaillancourt, “Complex singular Wishart matrices and applications,” *Computers and Mathematics with Applications*, vol. 50, pp. 399–411, 2005.
[34] A. Rothman, P. Bickel, E. Levina, and J. Zhu, ”Sparse permutation invariant covariance estimation”, *Electron. J. Statist.*, vol. 2, pp. 494–515, 2008.
[35] J. Schäfer and K. Strimmer, “A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics,” *Statist. Appl. Genet. Molec. Biol.*, vol 4, no. 1, 2005.
[36] C. Stein, “Estimation of a covariance matrix.” Rietz lecture, 39th Annual Meeting IMS. Atlanta, GA, 1975.
[37] C. Tracy and H. Widom, “Introduction to random matrices,” in *Geometric and Quantum Aspects of Integrable Systems*, edited by G. F. Helminck (Lecture Notes in Physics, Volume 424), Berlin: Springer, pp. 103-130, 1993.
[38] G. H. Tucci and K. Wang, “An innovative approach for analysing rank deficient covariance matrices,” In *Proc. IEEE Symposium on Information Theory*, Boston, pp. 2596–2600, 2012.