Maximum likelihood covariance matrix estimation from two possibly mismatched data sets

Olivier Besson

Maximum likelihood covariance matrix estimation from two possibly mismatched data sets

Olivier Besson

ISAE-SUPAERO, 10 Avenue Edouard Belin, Toulouse 31055, France

A B S T R A C T

We consider estimating the covariance matrix from two data sets, one whose covariance matrix $\mathbf{R}_1$ is the sought one and another set of samples whose covariance matrix $\mathbf{R}_2$ slightly differs from the sought one, due e.g. to different measurement configurations. We assume however that the two matrices are rather close, which we formulate by assuming that $\mathbf{R}_1^{1/2}\mathbf{R}_2^{-1}\mathbf{R}_1^{1/2}|\mathbf{R}_1$ follows a Wishart distribution around the identity matrix. It turns out that this assumption results in two data sets with different marginal distributions, hence the problem becomes that of covariance matrix estimation from two data sets which are distribution-mismatched. The maximum likelihood estimator (MLE) is derived and is shown to depend on the values of the number of samples in each set. We show that it involves whitening of one data set by the other one, shrinkage of eigenvalues and colorization, at least when one data set contains more samples than the size p of the observation space. When both data sets have less than p samples but the total number is larger than p, the MLE again entails eigenvalues shrinkage but this time after a projection operation. Simulation results compare the new estimator to state of the art techniques.

Keywords:
Covariance matrix estimation
Maximum likelihood
Mismatch

1. Problem statement

Analysis or processing of multichannel data most often relies on the covariance matrix, which is a fundamental tool e.g., for principal component analysis, spectral analysis, adaptive filtering, detection, direction of arrival estimation among others [1–3]. In practical applications, the $p \times p$ covariance matrix $\mathbf{R}$ needs to be estimated from a finite number n of samples. When the latter are independent and Gaussian distributed, the maximum likelihood estimator of $\mathbf{R}$ is $n^{-1}\mathbf{S}$ where $\mathbf{X}$ is the $p \times n$ data matrix and $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ is the sample covariance matrix (SCM) [1]. However, in low sample support or when deviation from the Gaussian assumption is at hand, the SCM tends to behave poorly. In particular it was observed that the sample covariance matrix is usually less well-conditioned than the true covariance matrix, and therefore considerable effort has been dedicated to regularizing it with a view to improve its performance.

One of the most important approach in this respect is due to Stein [4–6] who, instead of maximizing the likelihood function, advocated to minimize a meaningful loss function within a given class of estimators. Stein hence introduced the concept of admissible estimation and minimax estimators under the so-called Stein's loss. He showed that the SCM-based estimator is not minimax and derived minimax estimators in two important classes, namely estimators of the form $\hat{\mathbf{R}} = \mathbf{G}\mathbf{D}\mathbf{G}^T$ where $\mathbf{D}$ is a diagonal matrix and $\mathbf{G}$ is the Cholesky factor of $\mathbf{S}$, or of the form $\hat{\mathbf{R}} = \mathbf{U}\,\mathrm{diag}\big(\boldsymbol{\varphi}(\boldsymbol{\lambda})\big)\,\mathbf{U}^T$ where $\mathbf{U}\,\mathrm{diag}(\boldsymbol{\lambda})\,\mathbf{U}^T$ is the eigenvalue decomposition of $\mathbf{S}$ and $\boldsymbol{\varphi}(\boldsymbol{\lambda})$ is a non-linear function of $\boldsymbol{\lambda}$. This seminal work of Stein gave rise to a great number of studies, see for instance [7–13] and references therein. A second class of robust estimates is based on linear shrinkage of the SCM to a target matrix (an approach which can be interpreted as an empirical Bayes technique), i.e., estimates of the form $\hat{\mathbf{R}} = \alpha\mathbf{R}_t + \beta\mathbf{S}$ where $\mathbf{R}_t = \mathbf{I}$ is the most widely spread choice, see e.g., [14–20]. Note that these techniques applied with $\mathbf{R}_t = \mathbf{I}$ achieve an affine transformation of the eigenvalues of $\mathbf{S}$, while retaining the eigenvectors, and therefore bear resemblance with Stein's method, although the selection of α, β may not be driven by the same principle. Robustness to a possibly non Gaussian distribution has also been a topic of considerable interest and many papers have focused on robust estimation for elliptically distributed data, see e.g., [21–30] and references therein.

Most of the above cited works deal with estimation of a covariance matrix from a single data set. In this paper, we consider a situation where two data sets $\mathbf{X}_1$ and $\mathbf{X}_2$ are available, with respective covariance matrices $\mathbf{R}_1$ and $\mathbf{R}_2$. This situation typically arises in radar applications when one wishes to detect a target buried in clutter with unknown statistics [31,32]. In order to infer the latter, training samples are generally used, which hopefully share the

E-mail address: olivier.besson@isae-supaero.fr

https://doi.org/10.1016/j.sigpro.2019.107285

same statistics as the clutter in the cell under test (CUT). However, it has been evidenced that clutter is most often heterogeneous [31], with a discrepancy compared to the CUT that may grow with the distance to the CUT [33]. Therefore, one is led to use some clustering that separates training samples, either based on their proximity to the CUT or by means of some statistical criterion, such as the power selected training [34]. The samples so selected are deemed to be representative of the clutter in the CUT while others are less reliable, which corresponds to the situation considered herein. A second example is in the field of synthetic aperture radar in the case where a scene is imaged on two consecutive days, with possible changes in between [35]. Finally, in hyperspectral imagery, the problem of target or anomaly detection leads to a very similar framework. Indeed, the background in a pixel under test has to be estimated from the local pixels around and pixels located further apart [36]. In the present paper, we assume that $\mathbf{R}_2$ is close to $\mathbf{R}_1$, the covariance matrix we wish to estimate. Since $\mathbf{R}_2$ differs from but is close to $\mathbf{R}_1$ we investigate using both $\mathbf{X}_1$ and $\mathbf{X}_2$ to estimate $\mathbf{R}_1$. The reason for using also $\mathbf{X}_2$ is that despite its covariance matrix is not $\mathbf{R}_1$, it is close to. Additionally, one might face situations where the number of samples in $\mathbf{X}_1$ is very small. This paper constitutes a first approach to this specific problem and we focus herein on the most natural approach, namely maximum likelihood estimation. The objective is to figure out the pros and cons of the latter and the conditions under which it is an accurate estimator. The paper is organized as follows. In section 2 we formulate the statistical assumptions: more precisely, we assume that $\mathbf{R}_1^{1/2}\mathbf{R}_2^{-1}\mathbf{R}_1^{1/2}|\mathbf{R}_1$ is a random matrix with a Wishart distribution around the identity matrix, and we derive the joint distribution of $(\mathbf{X}_1, \mathbf{X}_2)$. Section 3 is devoted to the derivation of the maximum likelihood estimator of $\mathbf{R}_1$ from $(\mathbf{X}_1, \mathbf{X}_2)$, taking into account the possible configurations regarding the number of samples in each data set. Numerical simulations illustrate the performance of the MLE and compare it with existing alternatives in section 4. Conclusions and possible extensions of the present work are drawn in section 5.

2. Data model

Let us assume that we have two sets of measurements $\mathbf{X}_1(p\times n_1)$ and $\mathbf{X}_2(p\times n_2)$ which are distributed according to $\mathbf{X}_1 \overset{d}{=} \mathcal{N}(\mathbf{0}, \mathbf{R}_1, \mathbf{I})$ and $\mathbf{X}_2 \overset{d}{=} \mathcal{N}(\mathbf{0}, \mathbf{R}_2, \mathbf{I})$ where $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{\Omega})$ denotes the matrix-variate normal distribution whose density is $(2\pi)^{-pn/2}|\mathbf{\Sigma}|^{-n/2}|\mathbf{\Omega}|^{-p/2}\text{etr}\{-\frac{1}{2}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{\Omega}^{-1}\}$ with $|.|$ the determinant and etr{.} the exponential of the trace of a matrix. Note that we consider real-valued data here whereas in radar applications it is customary to consider complex-valued signals. In Appendix A we show how the results below can be readily extended to the complex case. Our goal in this paper is to estimate $\mathbf{R}_1$, using both $\mathbf{X}_1$ and $\mathbf{X}_2$ even if $\mathbf{R}_1 \neq \mathbf{R}_2$. However we assume that the two matrices are close to each other. In order to define a model that can reflect the proximity between $\mathbf{R}_1$ and $\mathbf{R}_2$, we note that the natural distance between them is given by $d^2(\mathbf{R}_1, \mathbf{R}_2) = \sum_{k=1}^{p} \log^2 \lambda_k(\mathbf{G}_1^T\mathbf{R}_2^{-1}\mathbf{G}_1)$ [37,38] where $\mathbf{G}_1$ is a square-root of $\mathbf{R}_1$, i.e., $\mathbf{R}_1 = \mathbf{G}_1\mathbf{G}_1^T$ and $\lambda_k(\mathbf{G}_1^T\mathbf{R}_2^{-1}\mathbf{G}_1)$ stands for the kth eigenvalue of $\mathbf{G}_1^T\mathbf{R}_2^{-1}\mathbf{G}_1$. This matrix is pivotal in adaptive detection problems also. More precisely, in the case of a covariance mismatch between the training samples and the data under test, it is shown in [39] that the performance of the well-known adaptive matched filter depends essentially on this matrix. Therefore, it becomes natural to encapsulate the difference between $\mathbf{R}_1$ and $\mathbf{R}_2$ through the matrix $\mathbf{W} = \mathbf{G}_1^T\mathbf{R}_2^{-1}\mathbf{G}_1$ and its proximity to the identity matrix. There are of course different ways to translate this constraint in the model. For instance a frequentist approach may be advocated where the joint probability density function of $(\mathbf{X}_1, \mathbf{X}_2)$ would be maximized under the constraint that the distance between $\mathbf{W}$ and $\mathbf{I}$ is smaller than some value. Alternatively, and this is what we elect here, one can resort to an empirical Bayes approach where the random matrix $\mathbf{W}$ follows some prior distribution rather concentrated around $\mathbf{I}$. For mathematical tractability, we choose a conjugate prior for $\mathbf{W}$ and we assume that $\mathbf{W}$ follows a Wishart distribution with ν degrees of freedom and parameter matrix $\mu^{-1}\mathbf{I}$, i.e., $\mathbf{W} \overset{d}{=} \mathcal{W}_p\big(\nu, \mu^{-1}\mathbf{I}\big)$. Of course, this is a rather strong assumption whose validity would be difficult to check, e.g., on real data. However, it is in accordance with the mere knowledge we have about the relation between $\mathbf{R}_1$ and $\mathbf{R}_2$, and it allows for tractable derivations.

Using the fact that $\mathbf{X}_1|\mathbf{R}_1$ and $\mathbf{X}_2|\mathbf{R}_2$ are independent and Gaussian distributed with respective covariance matrices $\mathbf{R}_1$ and $\mathbf{R}_2$, and since $\mathbf{R}_2 = \mathbf{G}_1\mathbf{W}^{-1}\mathbf{G}_1^T$, we thus assume the following stochastic model:

$$p(\mathbf{X}_1, \mathbf{X}_2|\mathbf{R}_1, \mathbf{W}) = (2\pi)^{-p(n_1+n_2)/2}|\mathbf{R}_1|^{-n_1/2}\left|\mathbf{W}^{-1}\mathbf{R}_1\right|^{-n_2/2} \times \text{etr}\left\{-\frac{1}{2}\mathbf{X}_1^T\mathbf{R}_1^{-1}\mathbf{X}_1 - \frac{1}{2}\mathbf{X}_2^T\mathbf{G}_1^{-T}\mathbf{W}\mathbf{G}_1^{-1}\mathbf{X}_2\right\} \quad (1a)$$

$$p(\mathbf{W}) = \frac{\mu^{\nu p/2}}{2^{\nu p/2}\Gamma_p(\nu/2)}|\mathbf{W}|^{(\nu-p-1)/2}\text{etr}\left\{-\frac{1}{2}\mu\mathbf{W}\right\} \quad (1b)$$

Note that $\mathbb{E}\{\mathbf{W}^{-1}\} = (\nu-p-1)^{-1}\mu\mathbf{I}$ so that $\mathbb{E}\{\mathbf{R}_2\} = \mathbb{E}\{\mathbf{G}_1\mathbf{W}^{-1}\mathbf{G}_1^T\} = (\nu-p-1)^{-1}\mu\mathbf{R}_1$: therefore, for $\mathbb{E}\{\mathbf{R}_2\}$ to be equal to $\mathbf{R}_1$, one must select $\mu = \nu-p-1$. Observe also that $\mathbf{W}$ comes closer to $\mathbf{I}$ as ν grows large. Indeed, $\mathbb{E}\{\mathbf{W}\} = \nu(\nu-p-1)^{-1}\mathbf{I}$ and $\mathbb{E}\{(\mathbf{W}-\mathbb{E}\{\mathbf{W}\})^2\} = p\nu(\nu-p-1)^2\mathbf{I}$ which goes to zero as $\nu\to\infty$ [40].

The marginal distribution of $(\mathbf{X}_1, \mathbf{X}_2)$ is obtained by integrating (1) with respect to $\mathbf{W}$, which results in

$$\begin{aligned}
p(\mathbf{X}_1, \mathbf{X}_2|\mathbf{R}_1) &= \int_{\mathbf{W}>0} p(\mathbf{X}_1, \mathbf{X}_2|\mathbf{R}_1, \mathbf{W})\,p(\mathbf{W})d\mathbf{W} \\
&= \frac{(2\pi)^{-p(n_1+n_2)/2}\mu^{\nu p/2}}{2^{\nu p/2}\Gamma_p(\nu/2)}|\mathbf{R}_1|^{-(n_1+n_2)/2}\text{etr}\left\{-\frac{1}{2}\mathbf{X}_1^T\mathbf{R}_1^{-1}\mathbf{X}_1\right\} \\
&\times \int_{\mathbf{W}>0}|\mathbf{W}|^{(\nu+n_2-p-1)/2}\text{etr}\left\{-\frac{1}{2}\mathbf{W}\left[\mu\mathbf{I}+\mathbf{G}_1^{-1}\mathbf{X}_2\mathbf{X}_2^T\mathbf{G}_1^{-T}\right]\right\}d\mathbf{W} \\
&= \frac{(2\pi)^{-p(n_1+n_2)/2}\mu^{\nu p/2}}{2^{\nu p/2}\Gamma_p(\nu/2)}2^{(\nu+n_2)p/2}\Gamma_p((\nu+n_2)/2) \\
&\times |\mathbf{R}_1|^{-(n_1+n_2)/2}\left|\mu\mathbf{I}+\mathbf{G}_1^{-1}\mathbf{X}_2\mathbf{X}_2^T\mathbf{G}_1^{-T}\right|^{-(\nu+n_2)/2}\text{etr}\left\{-\frac{1}{2}\mathbf{X}_1^T\mathbf{R}_1^{-1}\mathbf{X}_1\right\} \\
&= (2\pi)^{-pn_1/2}|\mathbf{R}_1|^{-n_1/2}\text{etr}\left\{-\frac{1}{2}\mathbf{X}_1^T\mathbf{R}_1^{-1}\mathbf{X}_1\right\} \\
&\times \frac{\pi^{-pn_2/2}\Gamma_p((\nu+n_2)/2)}{\Gamma_p(\nu/2)}|\mu\mathbf{R}_1|^{-n_2/2}\left|\mathbf{I}+\mathbf{X}_2^T[\mu\mathbf{R}_1]^{-1}\mathbf{X}_2\right|^{-(\nu+n_2)/2}
\end{aligned} \quad (2)$$

In order to obtain the third equality, we made use of the fact that, if $\mathbf{S} \overset{d}{=} \mathcal{W}_p(\nu, \mathbf{\Sigma})$,

$$\int_{\mathbf{S}>0} p(\mathbf{S})d\mathbf{S} = 1 \Rightarrow \int_{\mathbf{S}>0}|\mathbf{S}|^{(\nu-p-1)/2}\text{etr}\left\{-\frac{1}{2}\mathbf{S}\mathbf{\Sigma}^{-1}\right\}d\mathbf{S} = 2^{\nu p/2}\Gamma_p(\nu/2)|\mathbf{\Sigma}|^{\nu/2} \quad (3)$$

Note that $p(\mathbf{X}_1, \mathbf{X}_2|\mathbf{R}_1)$ in (2) can be factored as $p(\mathbf{X}_1, \mathbf{X}_2|\mathbf{R}_1) = f_1(\mathbf{X}_1, \mathbf{R}_1)\times f_2(\mathbf{X}_2, \mathbf{R}_1)$ which shows that $\mathbf{X}_1$ and $\mathbf{X}_2$ are marginally independent and that $p(\mathbf{X}_1, \mathbf{X}_2|\mathbf{R}_1) = p(\mathbf{X}_1|\mathbf{R}_1)p(\mathbf{X}_2|\mathbf{R}_1)$ with $p(\mathbf{X}_1|\mathbf{R}_1) \propto \text{etr}\left\{-\frac{1}{2}\mathbf{X}_1^T\mathbf{R}_1^{-1}\mathbf{X}_1\right\}$ and $p(\mathbf{X}_2|\mathbf{R}_1) \propto \left|\mathbf{I}+\mathbf{X}_2^T[\mu\mathbf{R}_1]^{-1}\mathbf{X}_2\right|^{-(\nu+n_2)/2}$. Due to the model adopted for the random matrix $\mathbf{W} = \mathbf{G}_1^T\mathbf{R}_2^{-1}\mathbf{G}_1$, $\mathbf{X}_2$ follows a matrix variate Student distribution [41]. Therefore, the fact that $\mathbf{R}_2 \neq \mathbf{R}_1$ results here in to two data sets with different distributions: one set

$\mathbf{X}_1$ is Gaussian distributed with covariance matrix $\mathbf{R}_1$ while the uncertainty in $\mathbf{R}_2$ leads to a Student distribution for $\mathbf{X}_2$. This is a rather original situation where one has to carry covariance matrix estimation from two data sets which are mismatched in their distributions. This peculiarity will result in new schemes compared to the conventional case of a single set with given distribution, as detailed now.

3. Maximum likelihood estimation

In this section we address estimation of $\mathbf{R}_1$ from $(\mathbf{X}_1, \mathbf{X}_2)$ and we focus on the most natural estimator, i.e., the maximum likelihood estimator. From (2), the log-likelihood function is, up to an additive and constant term

$$\begin{aligned} f(\mathbf{R}_1) &= -\frac{n_1+n_2}{2}\log|\mathbf{R}_1| - \frac{\nu+n_2}{2}\log\left|\mathbf{I}+\mu^{-1}\mathbf{R}_1^{-1}\mathbf{S}_2\right| - \frac{1}{2}\mathrm{Tr}\left\{\mathbf{R}_1^{-1}\mathbf{S}_1\right\} \\ &= \frac{\nu-n_1}{2}\log|\mathbf{R}_1| - \frac{\nu+n_2}{2}\log\left|\mathbf{R}_1+\mu^{-1}\mathbf{S}_2\right| - \frac{1}{2}\mathrm{Tr}\left\{\mathbf{R}_1^{-1}\mathbf{S}_1\right\} \end{aligned} \tag{4}$$

where $\mathbf{S}_1 = \mathbf{X}_1\mathbf{X}_1^T$ and $\mathbf{S}_2 = \mathbf{X}_2\mathbf{X}_2^T$. Differentiating the previous equation and using the fact that $d|\mathbf{R}| = |\mathbf{R}|\mathrm{Tr}\left\{\mathbf{R}^{-1}d\mathbf{R}\right\}$ and $d\mathbf{R}^{-1} = -\mathbf{R}^{-1}(d\mathbf{R})\mathbf{R}^{-1}$, we obtain the following equation that the ML solution should satisfy

$$(\nu-n_1)\mathbf{R}_1^{-1} - (\nu+n_2)\left(\mathbf{R}_1+\mu^{-1}\mathbf{S}_2\right)^{-1} + \mathbf{R}_1^{-1}\mathbf{S}_1\mathbf{R}_1^{-1} = \mathbf{0} \tag{5}$$

In order to solve (5), we must investigate various configurations for (n_1, n_2) as the solution will depend on them. Before going to the technical details of each case, we give an overview of the results obtained.

3.1. Summary of results

As is illustrated below, the expression of the maximum likelihood estimator depends on the respective values of n_1 and n_2. In the sequel three cases will be distinguished: a first situation where $n_1 < p$ and $n_2 \geq p$, a second one which is the mirror situation, namely $n_1 \geq p$ and $n_2 < p$, and finally a third more challenging case where $n_1 < p$, $n_2 < p$ and $n_1 + n_2 \geq p$.

In the first [respectively second] case, the ML solution is given by (11) [resp. (21)]: it will be shown that the estimation process entails whitening of $\mathbf{X}_1$ [resp. $\mathbf{X}_2$] by the inverse of the square-root of the sample covariance matrix of $\mathbf{X}_2$ [resp. $\mathbf{X}_1$], followed by shrinkage of eigenvalues and finally colorization by the square-root of the sample covariance matrix of $\mathbf{X}_2$ [resp. $\mathbf{X}_1$]. The technique of eigenvalue shrinkage is rather well known but usually applied to the SCM of a single set: herein, due to the presence of two data sets, this technique is applied to one data set after it has been whitened by the second one. Interestingly enough, the ML solution can also be written as (14) [resp. (22)], that is as a weighted sum of the SCM of each data set, where the weighting matrix is diagonal for one set of samples, and non diagonal for the other set.

Finally, when $n_2 < p$, $n_1 < p$ and $n_1 + n_2 \geq p$, the procedure includes a partitioning between the subspace spanned by the columns of $\mathbf{X}_2$ and its orthogonal complement. In the former, shrinkage of eigenvalues is used while, in the latter, projection of the SCM of $\mathbf{X}_1$ is retained.

3.2. Case $n_1 < p$ and $n_2 \geq p$

We consider first the case where $n_1 < p$ and $n_2 \geq p$, i.e., n_1 is not large enough for $\mathbf{S}_1$ to be positive definite and one needs to use $\mathbf{X}_2$ in order to estimate $\mathbf{R}_1$, even though $\mathbf{R}_2 \neq \mathbf{R}_1$. Eq. (5) can be rewritten as

$$\begin{aligned} &(\nu-n_1)\mathbf{R}_1^{-1}\left(\mathbf{R}_1+\mu^{-1}\mathbf{S}_2\right) - (\nu+n_2)\mathbf{I} + \mathbf{R}_1^{-1}\mathbf{S}_1\mathbf{R}_1^{-1}\left(\mathbf{R}_1+\mu^{-1}\mathbf{S}_2\right) = \mathbf{0} \\ &\Rightarrow -(n_1+n_2)\mathbf{I} + (\nu-n_1)\mu^{-1}\mathbf{R}_1^{-1}\mathbf{S}_2 + \mathbf{R}_1^{-1}\mathbf{S}_1 + \mu^{-1}\mathbf{R}_1^{-1}\mathbf{S}_1\mathbf{R}_1^{-1}\mathbf{S}_2 = \mathbf{0} \\ &\Rightarrow -(n_1+n_2)\mathbf{R}_1\mathbf{S}_2^{-1}\mathbf{R}_1 + (\nu-n_1)\mu^{-1}\mathbf{R}_1 + \mathbf{S}_1\mathbf{S}_2^{-1}\mathbf{R}_1 + \mu^{-1}\mathbf{S}_1 = \mathbf{0} \\ &\Rightarrow \mathbf{R}_1\mathbf{S}_2^{-1}\mathbf{R}_1 - \left[\frac{\nu-n_1}{\mu(n_1+n_2)}\mathbf{I} + \frac{1}{n_1+n_2}\mathbf{S}_1\mathbf{S}_2^{-1}\right]\mathbf{R}_1 \\ &- \frac{1}{\mu(n_1+n_2)}\mathbf{S}_1 = \mathbf{0} \end{aligned} \tag{6}$$

Let $\mathbf{S}_2 = \mathbf{L}_2\mathbf{L}_2^T$ and let us define $\tilde{\mathbf{R}}_{12} = \mathbf{L}_2^{-1}\mathbf{R}_1\mathbf{L}_2^{-T}$ and $\tilde{\mathbf{S}}_1 = \mathbf{L}_2^{-1}\mathbf{S}_1\mathbf{L}_2^{-T}$. Then, pre-multiplying the previous equation by $\mathbf{L}_2^{-1}$ and post-multiplying it by $\mathbf{L}_2^{-T}$, we obtain

$$\tilde{\mathbf{R}}_{12}^2 - \left[\frac{\nu-n_1}{\mu(n_1+n_2)}\mathbf{I} + \frac{1}{n_1+n_2}\tilde{\mathbf{S}}_1\right]\tilde{\mathbf{R}}_{12} - \frac{1}{\mu(n_1+n_2)}\tilde{\mathbf{S}}_1 = \mathbf{0} \tag{7}$$

Let $\mathbf{w}$ be an eigenvector of $\tilde{\mathbf{R}}_{12}$ associated with eigenvalue ξ. Then,

$$\begin{aligned} &\xi^2\mathbf{w} - \xi\left[\frac{\nu-n_1}{\mu(n_1+n_2)}\mathbf{I} + \frac{1}{n_1+n_2}\tilde{\mathbf{S}}_1\right]\mathbf{w} - \frac{1}{\mu(n_1+n_2)}\tilde{\mathbf{S}}_1\mathbf{w} = \mathbf{0} \\ &\Rightarrow \left[\frac{1}{\mu(n_1+n_2)} + \frac{\xi}{n_1+n_2}\right]\tilde{\mathbf{S}}_1\mathbf{w} = \xi\left[\xi - \frac{\nu-n_1}{\mu(n_1+n_2)}\right]\mathbf{w} \end{aligned} \tag{8}$$

which implies that $\mathbf{w}$ is also an eigenvector of $\tilde{\mathbf{S}}_1$. Either it is associated with a zero eigenvalue (there are $p-n_1$ of them) and, in this case, $\xi = \frac{\nu-n_1}{\mu(n_1+n_2)}$, or it is associated with a strictly positive eigenvalue λ and ξ satisfies the second-order polynomial equation

$$\xi^2 - \xi\left[\frac{\nu-n_1}{\mu(n_1+n_2)} + \frac{\lambda}{n_1+n_2}\right] - \frac{\lambda}{\mu(n_1+n_2)} = 0 \tag{9}$$

The above polynomial has obviously two real-valued roots, one being negative, the other being positive, and thus the latter is the eigenvalue of $\tilde{\mathbf{R}}_{12}$. To summarize, if we let $\mathbf{L}_2^{-1}\mathbf{X}_1 = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{k=1}^{n_1}\sigma_k\mathbf{u}_k\mathbf{v}_k^T$ be the singular value decomposition of $\mathbf{L}_2^{-1}\mathbf{X}_1$, we have

$$\begin{aligned} \tilde{\mathbf{R}}_{12} &= \sum_{k=1}^{n_1}\xi_k\mathbf{u}_k\mathbf{u}_k^T + \frac{\nu-n_1}{\mu(n_1+n_2)}\sum_{k=n_1+1}^{p}\mathbf{u}_k\mathbf{u}_k^T \\ &= \sum_{k=1}^{n_1}\left[\xi_k - \frac{\nu-n_1}{\mu(n_1+n_2)}\right]\mathbf{u}_k\mathbf{u}_k^T + \frac{\nu-n_1}{\mu(n_1+n_2)}\mathbf{I} \end{aligned} \tag{10}$$

where ξ_k is the positive root of (9) with λ substituted for σ_k^2. The MLE of $\mathbf{R}_1$ is thus

$$\mathbf{R}_1 = \sum_{k=1}^{n_1}\left(\xi_k - \frac{\nu-n_1}{\mu(n_1+n_2)}\right)\mathbf{L}_2\mathbf{u}_k\mathbf{u}_k^T\mathbf{L}_2^T + \frac{\nu-n_1}{\mu(n_1+n_2)}\mathbf{S}_2 \tag{11}$$

It is instructive to study the form of this solution. The original data $\mathbf{X}_1$ is first adaptively whitened by $\mathbf{L}_2^{-1}$ and its sample covariance matrix is computed. The eigenvectors of the latter are retained and the eigenvalues are modified. Then, data is re-colored by $\mathbf{L}_2$. Note that the technique of regularizing eigenvalues while keeping eigenvectors is classical in robust covariance matrix estimation. However, this technique usually applies to one set of samples. Here it applies to one set of samples after it has been "whitened" by the other set. Indeed a whitening-colorization operation is performed pre and post eigenvalues modification. Another important observation is that the transformation $\lambda \to \xi$ preserves the order of the eigenvalues, an important issue in Stein's estimation using eigenvalue decomposition [42–44]. This can be seen by differentiating (9) with respect to λ, which gives

$$\frac{\partial\xi}{\partial\lambda}\left[2\xi - \frac{\nu-n_1}{\mu(n_1+n_2)} - \frac{\lambda}{n_1+n_2}\right] = \frac{\xi}{n_1+n_2} + \frac{1}{\mu(n_1+n_2)} \tag{12}$$

Since the bracketed term on the left-hand side of the previous equation is positive, it follows that $\partial\xi/\partial\lambda > 0$ and therefore the transformation preserves ordering of the eigenvalues. This property will hold true in the other cases developed below.

A comment is also in order regarding the behavior of the MLE when ν grow large, i.e., when $\mathbf{W}$ comes closer to $\mathbf{I}$. Indeed, with $\mu = \nu - p - 1$, one has

$$\begin{aligned}\lim_{\nu\to\infty}\xi_k = \frac{1+\lambda_k}{n_1+n_2} &\Rightarrow \lim_{\nu\to\infty}\tilde{\mathbf{R}}_{12} = \frac{1}{n_1+n_2}\big[\tilde{\mathbf{S}}_1+\mathbf{I}\big]\\ &\Rightarrow \lim_{\nu\to\infty}\mathbf{R}_1 = \frac{1}{n_1+n_2}\mathbf{L}_2\big[\mathbf{L}_2^{-1}\mathbf{S}_1\mathbf{L}_2^{-T}+\mathbf{I}\big]\mathbf{L}_2^T\\ &= \frac{1}{n_1+n_2}[\mathbf{S}_1+\mathbf{S}_2] \end{aligned} \tag{13}$$

which shows that, as $\mathbf{W}$ comes closer to $\mathbf{I}$, i.e., as $\mathbf{R}_2$ comes closer to $\mathbf{R}_1$, the MLE is simply the sample covariance matrix of the whole data, as may be expected.

Finally, another interpretation of the MLE can be obtained by rewriting the MLE in an other form. Noting that the range space of $\mathbf{L}_2^{-1}\mathbf{X}_1$ coincides with the range space of $\mathbf{u}_1,\ldots,\mathbf{u}_{n_1}$, it follows that $\mathbf{u}_k = \mathbf{L}_2^{-1}\mathbf{X}_1\boldsymbol{\eta}_k$ for some vector $\boldsymbol{\eta}_k$. Therefore, (11) can be rewritten as

$$\begin{aligned}\mathbf{R}_1 &= \mathbf{X}_1\left[\sum_{k=1}^{n_1}\Big(\xi_k - \frac{\nu-n_1}{\mu(n_1+n_2)}\Big)\boldsymbol{\eta}_k\boldsymbol{\eta}_k^T\right]\mathbf{X}_1^T + \frac{\nu-n_1}{\mu(n_1+n_2)}\mathbf{S}_2\\ &= \mathbf{X}_1\boldsymbol{\Gamma}_1\mathbf{X}_1^T + \frac{\nu-n_1}{\mu(n_1+n_2)}\mathbf{X}_2\mathbf{X}_2^T \end{aligned} \tag{14}$$

Consequently, the MLE is a weighted version of the sample covariance matrices of each data set. In fact, it can be shown (we omit the details) that if a solution to (5) is sought which is of the form (14), then $\boldsymbol{\Gamma}_1$ is solution to the equation

$$\begin{aligned}\boldsymbol{\Gamma}_1^2 + \left[\frac{\nu-n_1}{\mu(n_1+n_2)}\big(\mathbf{X}_1^T\mathbf{S}_2^{-1}\mathbf{X}_1\big)^{-1} - \frac{1}{n_1+n_2}\mathbf{I}\right]\boldsymbol{\Gamma}_1\\ -\frac{\nu+n_2}{\mu(n_1+n_2)^2}\big(\mathbf{X}_1^T\mathbf{S}_2^{-1}\mathbf{X}_1\big)^{-1} = \mathbf{0}\end{aligned} \tag{15}$$

It ensues that $\boldsymbol{\Gamma}_1$ and $\mathbf{X}_1^T\mathbf{S}_2^{-1}\mathbf{X}_1$ share the same eigenvectors, which are indeed the right singular vectors $\mathbf{v}_k$ of $\mathbf{L}_2^{-1}\mathbf{X}_1$. Moreover, the eigenvalues γ_k of $\boldsymbol{\Gamma}_1$ satisfy

$$\gamma_k^2 + \gamma_k\left[\frac{(\nu-n_1)\sigma_k^{-2}}{\mu(n_1+n_2)} - \frac{1}{n_1+n_2}\right] - \frac{(\nu+n_2)\sigma_k^{-2}}{\mu(n_1+n_2)^2} = 0 \tag{16}$$

To summarize, the MLE of $\mathbf{R}_1$ can either be written as in (11) where the eigenvalues ξ_k are related to the eigenvalues λ_k of $\mathbf{L}_2^{-1}\mathbf{S}_1\mathbf{L}_2^{-T}$ by (9), or as in (14) where $\boldsymbol{\Gamma}_1$ is given by (15).

3.3. *Case* $n_2 < p$ *and* $n_1 \geq p$

We now consider a situation where $n_2 < p$ and $n_1 \geq p$ under which one has a sufficient number of "good" samples $\mathbf{X}_1$ for $\mathbf{S}_1$ to be full-rank. Yet, it might be of interest to use $\mathbf{X}_2$ even though its covariance matrix $\mathbf{R}_2 \neq \mathbf{R}_1$. The derivation of the MLE follows along the same lines as in the previous case, except that now $\mathbf{S}_2$ is rank-deficient and $\mathbf{S}_1$ is full-rank. Starting from the ML Eq. (5), on can write

$$\begin{aligned}&(\nu-n_1)\mathbf{R}_1^{-1}\big(\mathbf{R}_1+\mu^{-1}\mathbf{S}_2\big) - (\nu+n_2)\mathbf{I} + \mathbf{R}_1^{-1}\mathbf{S}_1\mathbf{R}_1^{-1}\big(\mathbf{R}_1+\mu^{-1}\mathbf{S}_2\big) = \mathbf{0}\\ &\Rightarrow -(n_1+n_2)\mathbf{I} + (\nu-n_1)\mu^{-1}\mathbf{R}_1^{-1}\mathbf{S}_2 + \mathbf{R}_1^{-1}\mathbf{S}_1 + \mu^{-1}\mathbf{R}_1^{-1}\mathbf{S}_1\mathbf{R}_1^{-1}\mathbf{S}_2 = \mathbf{0}\\ &\Rightarrow -(n_1+n_2)\mathbf{R}_1\mathbf{S}_1^{-1}\mathbf{R}_1 + (\nu-n_1)\mu^{-1}\mathbf{R}_1\mathbf{S}_1^{-1}\mathbf{S}_2 + \mathbf{R}_1 + \mu^{-1}\mathbf{S}_2 = \mathbf{0}\\ &\Rightarrow \mathbf{R}_1\mathbf{S}_1^{-1}\mathbf{R}_1 - \mathbf{R}_1\left[\frac{(\nu-n_1)}{\mu(n_1+n_2)}\mathbf{S}_1^{-1}\mathbf{S}_2 + \frac{1}{n_1+n_2}\mathbf{I}\right] - \frac{1}{\mu(n_1+n_2)}\mathbf{S}_2 = \mathbf{0}\end{aligned} \tag{17}$$

Let $\mathbf{S}_1 = \mathbf{L}_1\mathbf{L}_1^T$ and let us define $\tilde{\mathbf{R}}_{11} = \mathbf{L}_1^{-1}\mathbf{R}_1\mathbf{L}_1^{-T}$ and $\tilde{\mathbf{S}}_2 = \mathbf{L}_1^{-1}\mathbf{S}_2\mathbf{L}_1^{-T}$. Then, taking the transpose of the previous equation, pre-multiplying by $\mathbf{L}_1^{-1}$ and post-multiplying by $\mathbf{L}_1^{-T}$, we obtain

$$\tilde{\mathbf{R}}_{11}^2 - \left[\frac{(\nu-n_1)}{\mu(n_1+n_2)}\tilde{\mathbf{S}}_2 + \frac{1}{n_1+n_2}\mathbf{I}\right]\tilde{\mathbf{R}}_{11} - \frac{1}{\mu(n_1+n_2)}\tilde{\mathbf{S}}_2 = \mathbf{0} \tag{18}$$

As before, it can be seen that $\tilde{\mathbf{R}}_{11}$ and $\tilde{\mathbf{S}}_2$ share the same eigenvectors. The $p-n_2$ eigenvectors of $\tilde{\mathbf{S}}_2$ associated with zero eigenvalue will correspond to a constant eigenvalue for $\tilde{\mathbf{R}}_{11}$ equal to $(n_1+n_2)^{-1}$. A strictly positive eigenvalue ζ of $\tilde{\mathbf{R}}_{11}$ is related to its counterpart λ of $\tilde{\mathbf{S}}_2$ by

$$\zeta^2 - \zeta\left[\frac{\lambda(\nu-n_1)}{\mu(n_1+n_2)} + \frac{1}{n_1+n_2}\right] - \frac{\lambda}{\mu(n_1+n_2)} = 0 \tag{19}$$

Now, if we let $\mathbf{L}_1^{-1}\mathbf{X}_2 = \mathbf{Y}\boldsymbol{\Theta}\mathbf{Z}^T = \sum_{k=1}^{n_2}\theta_k\mathbf{y}_k\mathbf{z}_k^T$ be the singular value decomposition of $\mathbf{L}_1^{-1}\mathbf{X}_2$, we have

$$\begin{aligned}\tilde{\mathbf{R}}_{11} &= \sum_{k=1}^{n_2}\zeta_k\mathbf{y}_k\mathbf{y}_k^T + \frac{1}{n_1+n_2}\sum_{k=n_2+1}^{p}\mathbf{y}_k\mathbf{y}_k^T\\ &= \sum_{k=1}^{n_2}\Big(\zeta_k - \frac{1}{n_1+n_2}\Big)\mathbf{y}_k\mathbf{y}_k^T + \frac{1}{n_1+n_2}\mathbf{I}\end{aligned} \tag{20}$$

where ζ_k is the positive root of (19) with λ substituted for θ_k^2. The MLE of $\mathbf{R}_1$ becomes

$$\mathbf{R}_1 = \sum_{k=1}^{n_2}\Big(\zeta_k - \frac{1}{n_1+n_2}\Big)\mathbf{L}_1\mathbf{y}_k\mathbf{y}_k^T\mathbf{L}_1^T + \frac{1}{n_1+n_2}\mathbf{S}_1 \tag{21}$$

Again, since the range space of $\mathbf{L}_1^{-1}\mathbf{X}_2$ is spanned by $\mathbf{y}_1,\ldots,\mathbf{y}_{n_2}$, one has $\mathbf{y}_k = \mathbf{L}_1^{-1}\mathbf{X}_2\boldsymbol{\chi}_k$ and hence

$$\begin{aligned}\mathbf{R}_1 &= \mathbf{X}_2\left[\sum_{k=1}^{n_2}\Big(\zeta_k - \frac{1}{n_1+n_2}\Big)\boldsymbol{\chi}_k\boldsymbol{\chi}_k^T\right]\mathbf{X}_2^T + \frac{1}{n_1+n_2}\mathbf{X}_1\mathbf{X}_1^T\\ &= \mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2^T + \frac{1}{n_1+n_2}\mathbf{X}_1\mathbf{X}_1^T\end{aligned} \tag{22}$$

Note that (22) differs from (14) in that the weighting matrix applied between $\mathbf{X}_1$ and $\mathbf{X}_1^T$ is now diagonal while that applied between $\mathbf{X}_2$ and $\mathbf{X}_2^T$ is no longer diagonal. Furthermore, if one looks for a solution of the form (22) then $\boldsymbol{\Gamma}_2$ is the solution to

$$\begin{aligned}\boldsymbol{\Gamma}_2^2 + \left[\frac{(\mathbf{X}_2^T\mathbf{S}_1^{-1}\mathbf{X}_2)^{-1}}{n_1+n_2} - \frac{(\nu-n_1)}{\mu(n_1+n_2)}\mathbf{I}\right]\boldsymbol{\Gamma}_2\\ -\frac{\nu+n_2}{\mu(n_1+n_2)^2}(\mathbf{X}_2^T\mathbf{S}_1^{-1}\mathbf{X}_2)^{-1} = \mathbf{0}\end{aligned} \tag{23}$$

$\boldsymbol{\Gamma}_2$ and $\mathbf{X}_2^T\mathbf{S}_1^{-1}\mathbf{X}_2$ share the same eigenvectors (actually $\mathbf{z}_k$) and the eigenvalue γ_k of $\boldsymbol{\Gamma}_2$ is obtained as the positive solution to

$$\gamma_k^2 + \gamma_k\left[\frac{\theta_k^{-2}}{n_1+n_2} - \frac{(\nu-n_1)}{\mu(n_1+n_2)}\right] - \frac{\theta_k^{-2}(\nu+n_2)}{\mu(n_1+n_2)^2} = 0 \tag{24}$$

Remark 1. When $n_1 \geq p$ and $n_2 \geq p$, the previous techniques can still be used, with slight variations. In this case, $\tilde{\mathbf{S}}_1$ and $\tilde{\mathbf{S}}_2$ are now full-rank, and therefore the MLE of $\mathbf{R}_1$ is given by (11) but with the first sum extended to p eigenvectors ($\tilde{\mathbf{S}}_1$ has now p non-zero eigenvalues), and the second term vanishes. The ML solution is also given by (21) with the first term extended to p eigenvectors and the second term vanishing.

3.4. *Case* $n_1 < p$, $n_2 < p$ *and* $n_1 + n_2 \geq p$

We now consider the more challenging case where neither of the two data sets contains enough samples for their respective sample covariance matrices to be full rank, and thus it becomes mandatory to combine both sets. This situation is a bit trickier and requires some carefulness. Going back to (5), the MLE of $\mathbf{R}_1$ should satisfy

$$\begin{aligned}&(\nu-n_1)\mathbf{R}_1^{-1} - (\nu+n_2)\big(\mathbf{R}_1+\mu^{-1}\mathbf{S}_2\big)^{-1} + \mathbf{R}_1^{-1}\mathbf{S}_1\mathbf{R}_1^{-1} = \mathbf{0}\\ &\Rightarrow (\nu-n_1)\mathbf{R}_1^{-1} + \mathbf{R}_1^{-1}\mathbf{S}_1\mathbf{R}_1^{-1}\end{aligned}$$

$$-(\nu+n_2)\left[\mathbf{R}_1^{-1}-\mu^{-1}\mathbf{R}_1^{-1}\mathbf{X}_2\left(\mathbf{I}+\mu^{-1}\mathbf{X}_2^T\mathbf{R}_1^{-1}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^T\mathbf{R}_1^{-1}\right]=\mathbf{0}$$
$$\Rightarrow (n_1+n_2)\mathbf{R}_1=(\nu+n_2)\mu^{-1}\mathbf{X}_2\left(\mathbf{I}+\mu^{-1}\mathbf{X}_2^T\mathbf{R}_1^{-1}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^T+\mathbf{X}_1\mathbf{X}_1^T \tag{25}$$

Before pursuing, it is worthy looking at the previous equation to get some insight. We observe that the projection of $\mathbf{R}_1$ onto the subspace orthogonal to $\mathbf{X}_2$ will be equal to the projection of $\mathbf{S}_1$ on this same subspace. This suggests to use a decomposition that splits data in $\mathcal{R}(\mathbf{X}_2)$ and its orthogonal complement. To do so, let us consider the SVD of $\mathbf{X}_2$ as $\mathbf{X}_2=\mathbf{CDE}^T=\begin{bmatrix}\mathbf{C}_a & \mathbf{C}_b\end{bmatrix}\begin{bmatrix}\mathbf{D}_a\\ \mathbf{0}\end{bmatrix}\mathbf{E}^T=\mathbf{C}_a\mathbf{D}_a\mathbf{E}^T$, where $\mathbf{C}$ is $p\times p$, $\mathbf{C}_a$ is $p\times n_2$ and $\mathbf{D}_a$ is the $n_2\times n_2$ diagonal matrix of singular values. Let us also operate a change of coordinates and define

$$\mathbf{\Sigma}=\mathbf{C}^T\mathbf{R}_1\mathbf{C}=\begin{pmatrix}\mathbf{C}_a^T\mathbf{R}_1\mathbf{C}_a & \mathbf{C}_a^T\mathbf{R}_1\mathbf{C}_b\\ \mathbf{C}_b^T\mathbf{R}_1\mathbf{C}_a & \mathbf{C}_b^T\mathbf{R}_1\mathbf{C}_b\end{pmatrix}=\begin{pmatrix}\mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab}\\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb}\end{pmatrix} \tag{26}$$

With these definitions, it is straightforward to show that $\mathbf{X}_2^T\mathbf{R}_1^{-1}\mathbf{X}_2=\mathbf{E}\mathbf{D}_a\mathbf{\Sigma}_{a.b}^{-1}\mathbf{D}_a\mathbf{E}^T$ where $\mathbf{\Sigma}_{a.b}=\mathbf{\Sigma}_{aa}-\mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba}$ and thus

$$\begin{aligned}\mathbf{X}_2\left(\mathbf{I}+\mu^{-1}\mathbf{X}_2^T\mathbf{R}_1^{-1}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^T&=\mathbf{C}_a\mathbf{D}_a\left[\mathbf{I}+\mu^{-1}\mathbf{D}_a\mathbf{\Sigma}_{a.b}^{-1}\mathbf{D}_a\right]^{-1}\mathbf{D}_a\mathbf{C}_a^T\\&=\mathbf{C}_a\left[\mathbf{D}_a^{-2}+\mu^{-1}\mathbf{\Sigma}_{a.b}^{-1}\right]^{-1}\mathbf{C}_a^T\end{aligned} \tag{27}$$

Therefore, pre-multiplying (25) by $\mathbf{C}^T$ and post-multiplying it by $\mathbf{C}$, we obtain

$$\begin{aligned}(n_1+n_2)\begin{pmatrix}\mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab}\\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb}\end{pmatrix}&=(\nu+n_2)\mu^{-1}\begin{pmatrix}\left[\mathbf{D}_a^{-2}+\mu^{-1}\mathbf{\Sigma}_{a.b}^{-1}\right]^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\\&+\begin{pmatrix}\mathbf{C}_a^T\mathbf{S}_1\mathbf{C}_a & \mathbf{C}_a^T\mathbf{S}_1\mathbf{C}_b\\ \mathbf{C}_b^T\mathbf{S}_1\mathbf{C}_a & \mathbf{C}_b^T\mathbf{S}_1\mathbf{C}_b\end{pmatrix}\end{aligned} \tag{28}$$

which immediately implies that

$$\begin{aligned}(n_1+n_2)\mathbf{\Sigma}_{ba}&=\mathbf{C}_b^T\mathbf{S}_1\mathbf{C}_a\\(n_1+n_2)\mathbf{\Sigma}_{bb}&=\mathbf{C}_b^T\mathbf{S}_1\mathbf{C}_b\end{aligned} \tag{29}$$

This corroborates the comments we made after Eq. (25) since one has

$$\begin{aligned}(n_1+n_2)\mathbf{C}_b\mathbf{\Sigma}_{bb}\mathbf{C}_b^T=(n_1+n_2)\mathbf{C}_b\mathbf{C}_b^T\mathbf{R}_1\mathbf{C}_b\mathbf{C}_b^T&=\mathbf{P}_{\mathbf{X}_2}^{\perp}\mathbf{R}_1\mathbf{P}_{\mathbf{X}_2}^{\perp}\\&=\mathbf{C}_b\mathbf{C}_b^T\mathbf{S}_1\mathbf{C}_b\mathbf{C}_b^T=\mathbf{P}_{\mathbf{X}_2}^{\perp}\mathbf{S}_1\mathbf{P}_{\mathbf{X}_2}^{\perp}\end{aligned} \tag{30}$$

It now remains to find $\mathbf{\Sigma}_{aa}$ or equivalently $\mathbf{\Sigma}_{a.b}$. Towards this end, note that

$$(n_1+n_2)\mathbf{\Sigma}_{aa}=(\nu+n_2)\mu^{-1}\left[\mathbf{D}_a^{-2}+\mu^{-1}\mathbf{\Sigma}_{a.b}^{-1}\right]^{-1}+\mathbf{C}_a^T\mathbf{S}_1\mathbf{C}_a \tag{31}$$

However,

$$\begin{aligned}(n_1+n_2)\mathbf{\Sigma}_{aa}&=(n_1+n_2)\left[\mathbf{\Sigma}_{a.b}+\mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba}\right]\\&=(n_1+n_2)\mathbf{\Sigma}_{a.b}+\left(\mathbf{C}_a^T\mathbf{S}_1\mathbf{C}_b\right)\left(\mathbf{C}_b^T\mathbf{S}_1\mathbf{C}_b\right)^{-1}\left(\mathbf{C}_b^T\mathbf{S}_1\mathbf{C}_a\right)\end{aligned} \tag{32}$$

which leads to

$$(n_1+n_2)\mathbf{\Sigma}_{a.b}=(\nu+n_2)\mu^{-1}\left[\mathbf{D}_a^{-2}+\mu^{-1}\mathbf{\Sigma}_{a.b}^{-1}\right]^{-1}+\left[\mathbf{C}^T\mathbf{S}_1\mathbf{C}\right]_{a.b} \tag{33}$$

For the sake of notational convenience, let us denote $\mathbf{F}=\left[\mathbf{C}^T\mathbf{S}_1\mathbf{C}\right]_{a.b}$. Post-multiplying the previous equation by $\left[\mathbf{D}_a^{-2}+\mu^{-1}\mathbf{\Sigma}_{a.b}^{-1}\right]$ results in

$$\begin{aligned}&(n_1+n_2)\mathbf{\Sigma}_{a.b}\mathbf{D}_a^{-2}-\left[(\nu-n_1)\mu^{-1}\mathbf{I}+\mathbf{F}\mathbf{D}_a^{-2}\right]-\mu^{-1}\mathbf{F}\mathbf{\Sigma}_{a.b}^{-1}=\mathbf{0}\\&\Rightarrow\mathbf{\Sigma}_{a.b}\mathbf{D}_a^{-2}\mathbf{\Sigma}_{a.b}-\left[\frac{\nu-n_1}{\mu(n_1+n_2)}\mathbf{I}+\frac{1}{n_1+n_2}\mathbf{F}\mathbf{D}_a^{-2}\right]\mathbf{\Sigma}_{a.b}-\frac{1}{\mu(n_1+n_2)}\mathbf{F}=\mathbf{0}\\&\Rightarrow\tilde{\mathbf{\Sigma}}_{a.b}^2-\left[\frac{\nu-n_1}{\mu(n_1+n_2)}\mathbf{I}+\frac{1}{n_1+n_2}\tilde{\mathbf{F}}\right]\tilde{\mathbf{\Sigma}}_{a.b}-\frac{1}{\mu(n_1+n_2)}\tilde{\mathbf{F}}=\mathbf{0}\end{aligned} \tag{34}$$

where $\tilde{\mathbf{\Sigma}}_{a.b}=\mathbf{D}_a^{-1}\mathbf{\Sigma}_{a.b}\mathbf{D}_a^{-1}$ and $\tilde{\mathbf{F}}=\mathbf{D}_a^{-1}\mathbf{F}\mathbf{D}_a^{-1}$. Similarly to what was done before, $\tilde{\mathbf{\Sigma}}_{a.b}$ and $\tilde{\mathbf{F}}$ share the same eigenvectors. When the eigenvalue λ of $\tilde{\mathbf{F}}$ is zero (there are actually $p-n_1$ of them [45]) the corresponding eigenvalue ϕ of $\tilde{\mathbf{\Sigma}}_{a.b}$ is $\frac{\nu-n_1}{\mu(n_1+n_2)}$. For each of the $r=n_1+n_2-p$ non-zero λ, the corresponding ϕ is the unique positive root of

$$\phi^2-\left[\frac{\nu-n_1}{\mu(n_1+n_2)}+\frac{\lambda}{n_1+n_2}\right]\phi-\frac{\lambda}{\mu(n_1+n_2)}=0 \tag{35}$$

Therefore, if $\tilde{\mathbf{u}}_k$ are the eigenvectors of $\tilde{\mathbf{F}}$, $\tilde{\mathbf{\Sigma}}_{a.b}$ is given by

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_{a.b}&=\sum_{k=1}^{r}\phi_k\tilde{\mathbf{u}}_k\tilde{\mathbf{u}}_k^T+\frac{\nu-n_1}{\mu(n_1+n_2)}\sum_{k=r+1}^{p}\tilde{\mathbf{u}}_k\tilde{\mathbf{u}}_k^T\\&=\sum_{k=1}^{r}\left[\phi_k-\frac{\nu-n_1}{\mu(n_1+n_2)}\right]\tilde{\mathbf{u}}_k\tilde{\mathbf{u}}_k^T+\frac{\nu-n_1}{\mu(n_1+n_2)}\mathbf{I}\end{aligned} \tag{36}$$

Once $\tilde{\mathbf{\Sigma}}_{a.b}$ is computed, $\mathbf{\Sigma}_{a.b}=\mathbf{D}_a\tilde{\mathbf{\Sigma}}_{a.b}\mathbf{D}_a$ and $\mathbf{\Sigma}_{aa}$ can be obtained from (32). Finally, the MLE of $\mathbf{R}_1$ is given by $\mathbf{C}\mathbf{\Sigma}\mathbf{C}^T$.

We now present an alternative way to compute the solution. From (25), it appears that $\mathbf{R}_1$ can be written as $(n_1+n_2)\mathbf{R}_1=\mathbf{X}_1\mathbf{X}_1^T+\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{X}_2^T$ where $\mathbf{\Gamma}_2=(\nu+n_2)\mu^{-1}\left(\mathbf{I}+\mu^{-1}\mathbf{X}_2^T\mathbf{R}_1^{-1}\mathbf{X}_2\right)^{-1}$. Let $\mathbf{X}=\begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}$ and let $\mathbf{X}^T=\mathbf{Q}\mathbf{R}$ be the QR decomposition of $\mathbf{X}^T$ with $\mathbf{Q}$ a $(n_1+n_2)\times p$ semi-unitary matrix, i.e., $\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_p$. Let us partition $\mathbf{Q}$ as $\mathbf{Q}=\begin{bmatrix}\mathbf{Q}_1\\ \mathbf{Q}_2\end{bmatrix}$ so that $\mathbf{X}_1^T=\mathbf{Q}_1\mathbf{R}$ and $\mathbf{X}_2^T=\mathbf{Q}_2\mathbf{R}$. Then, one has

$$\begin{aligned}(n_1+n_2)\mathbf{R}_1&=\mathbf{X}_1\mathbf{X}_1^T+\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{X}_2^T\\&=\mathbf{R}^T\left[\mathbf{Q}_1^T\mathbf{Q}_1+\mathbf{Q}_2^T\mathbf{\Gamma}_2\mathbf{Q}_2\right]\mathbf{R}\end{aligned}$$

and therefore

$$\begin{aligned}(n_1+n_2)^{-1}\mathbf{X}_2^T\mathbf{R}_1^{-1}\mathbf{X}_2&=\mathbf{Q}_2\left[\mathbf{Q}_1^T\mathbf{Q}_1+\mathbf{Q}_2^T\mathbf{\Gamma}_2\mathbf{Q}_2\right]^{-1}\mathbf{Q}_2^T\\&=\mathbf{Q}_2\left[\mathbf{I}+\mathbf{Q}_2^T(\mathbf{\Gamma}_2-\mathbf{I})\mathbf{Q}_2\right]^{-1}\mathbf{Q}_2^T\\&=\mathbf{Q}_2\left[\mathbf{I}-\mathbf{Q}_2^T\left[(\mathbf{\Gamma}_2-\mathbf{I})^{-1}+\mathbf{Q}_2\mathbf{Q}_2^T\right]^{-1}\mathbf{Q}_2\right]\mathbf{Q}_2^T\\&=\mathbf{Q}_2\mathbf{Q}_2^T-\mathbf{Q}_2\mathbf{Q}_2^T\left[(\mathbf{\Gamma}_2-\mathbf{I})^{-1}+\mathbf{Q}_2\mathbf{Q}_2^T\right]^{-1}\mathbf{Q}_2\mathbf{Q}_2^T\\&=\left[\left(\mathbf{Q}_2\mathbf{Q}_2^T\right)^{-1}+\mathbf{\Gamma}_2-\mathbf{I}\right]^{-1}\end{aligned} \tag{37}$$

Consequently, if we define $\mathbf{B}_2=\left(\mathbf{Q}_2\mathbf{Q}_2^T\right)^{-1}-\mathbf{I}$

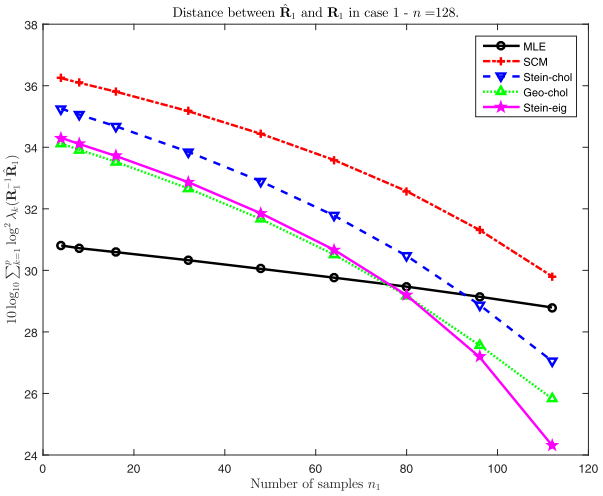
$$\begin{aligned}\mathbf{\Gamma}_2^{-1}&=(\nu+n_2)^{-1}\mu\left[\mathbf{I}+\mu^{-1}\mathbf{X}_2^T\mathbf{R}_1^{-1}\mathbf{X}_2\right]\\&=(\nu+n_2)^{-1}\mu\mathbf{I}+(\nu+n_2)^{-1}\mathbf{X}_2^T\mathbf{R}_1^{-1}\mathbf{X}_2\\&=(\nu+n_2)^{-1}\mu\mathbf{I}+(\nu+n_2)^{-1}(n_1+n_2)(\mathbf{\Gamma}_2+\mathbf{B}_2)^{-1}\end{aligned} \tag{38}$$

Pre-multiplying the previous equation by $(\mathbf{\Gamma}_2+\mathbf{B}_2)$ and post-multiplying by $\mathbf{\Gamma}_2$, we obtain the following second-order polynomial equation:

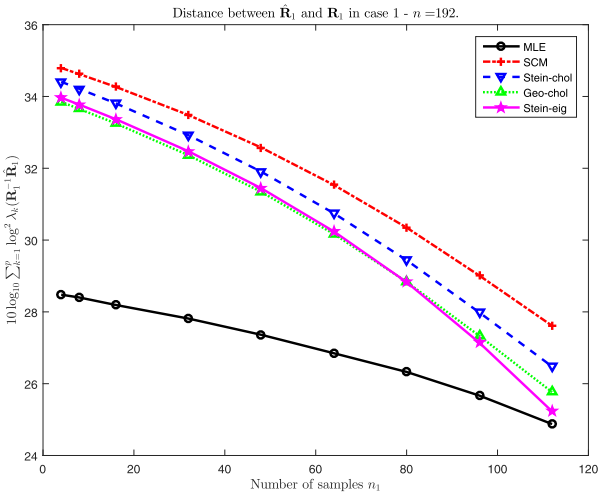
$$\mathbf{\Gamma}_2^2+\left[\mathbf{B}_2-\frac{\nu-n_1}{\mu}\mathbf{I}\right]\mathbf{\Gamma}_2-\frac{\nu+n_2}{\mu}\mathbf{B}_2=\mathbf{0} \tag{39}$$

It follows that $\mathbf{\Gamma}_2$ and $\mathbf{B}_2$ share the same eigenvectors. If λ is a non-zero eigenvalue of $\mathbf{B}_2$ (there are n_1+n_2-p of them), then the corresponding eigenvalue γ of $\mathbf{\Gamma}_2$ is the unique positive root to the following polynomial equation

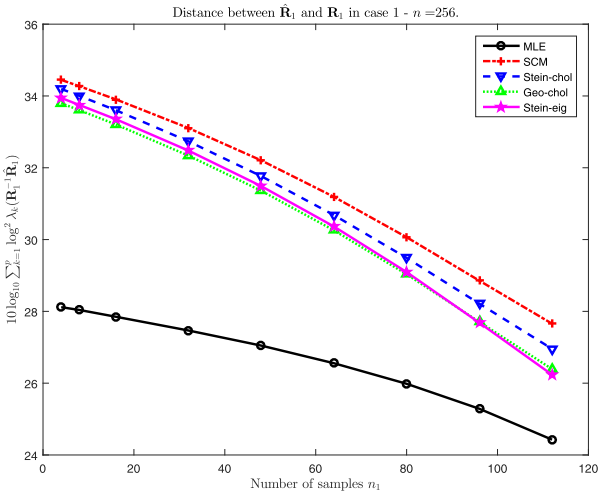
$$\gamma^2+\left[\lambda-\frac{\nu-n_1}{\mu}\right]\gamma-\frac{\nu+n_2}{\mu}\lambda=0 \tag{40}$$



(a) $n_1 + n_2 = p$

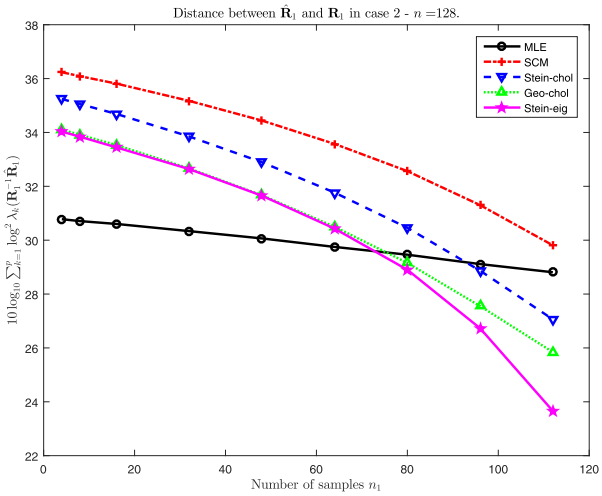


(b) $n_1 + n_2 = 3p/2$

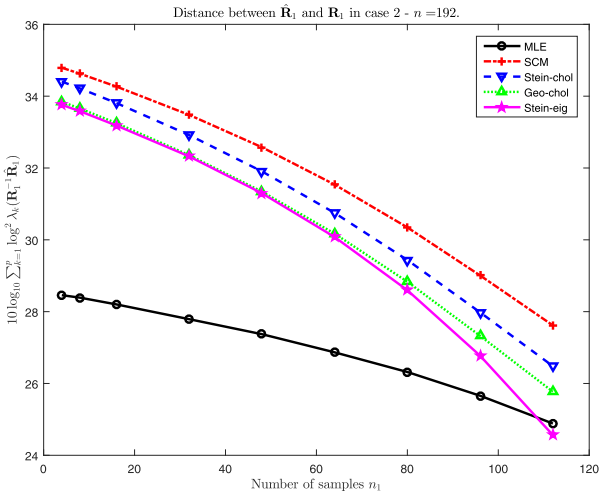


(c) $n_1 + n_2 = 2p$

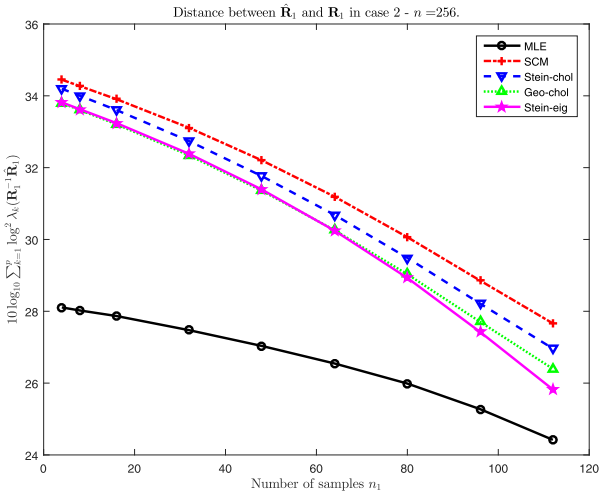
Fig. 1. Average distance between $\hat{\mathbf{R}}_1$ and $\mathbf{R}_1$ in case 1.



(a) $n_1 + n_2 = p$



(b) $n_1 + n_2 = 3p/2$



(c) $n_1 + n_2 = 2p$

Fig. 2. Average distance between $\hat{\mathbf{R}}_1$ and $\mathbf{R}_1$ in case 2.

If $\lambda = 0$ then $\gamma = (\nu - n_2)\mu^{-1}$. Finally, the solution $\mathbf{\Gamma}_2$ is given by

$$\begin{aligned}\mathbf{\Gamma}_2 &= \sum_{k=1}^{r} \gamma_k \mathbf{b}_k \mathbf{b}_k^T + \frac{\nu - n_1}{\mu} \sum_{k=r+1}^{n_2} \gamma_k \mathbf{b}_k \mathbf{b}_k^T \\ &= \sum_{k=1}^{r} \left[\gamma_k - \frac{\nu - n_1}{\mu}\right] \mathbf{b}_k \mathbf{b}_k^T + \frac{\nu - n_1}{\mu} \mathbf{I} \end{aligned} \tag{41}$$

where $\mathbf{b}_k$ are the eigenvectors of $\mathbf{B}_2$. Note that

$$\begin{aligned}\mathbf{B}_2\mathbf{b} = \lambda\mathbf{b} &\Rightarrow (\mathbf{Q}_2\mathbf{Q}_2^T)^{-1}\mathbf{b} - \mathbf{b} = \lambda\mathbf{b} \\ &\Rightarrow (\mathbf{Q}_2\mathbf{Q}_2^T)^{-1}\mathbf{b} = (1+\lambda)\mathbf{b} \\ &\Rightarrow (\mathbf{Q}_2\mathbf{Q}_2^T)\mathbf{b} = (1+\lambda)^{-1}\mathbf{b}\end{aligned}$$

and hence $\mathbf{b}$ is an eigenvector of $\mathbf{Q}_2\mathbf{Q}_2^T$ associated with eigenvalue $(1+\lambda)^{-1}$, or equivalently a right singular vector of $\mathbf{Q}_2^T$. Observe also that, since $\mathbf{X}_2^T = \mathbf{Q}_2\mathbf{R}$ and $\mathbf{X}\mathbf{X}^T = \mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R} = \mathbf{R}^T\mathbf{R}$, one has

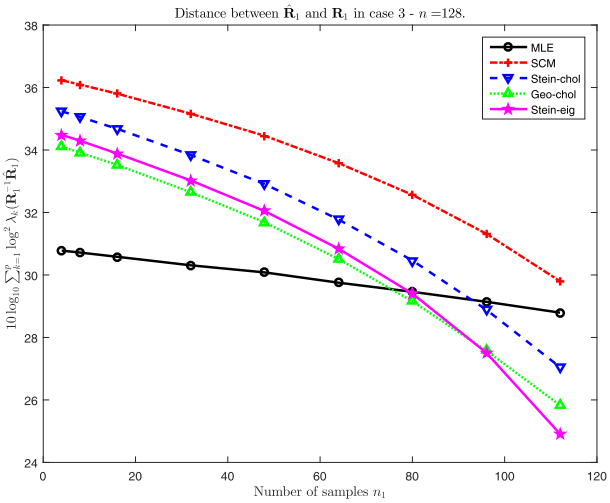
$$\begin{aligned}\mathbf{Q}_2\mathbf{Q}_2^T &= \mathbf{X}_2^T\mathbf{R}^{-1}\mathbf{R}^{-T}\mathbf{X}_2 = \mathbf{X}_2^T(\mathbf{R}^T\mathbf{R})^{-1}\mathbf{X}_2 \\ &= \mathbf{X}_2^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}_2 = \mathbf{X}_2^T(\mathbf{X}_1\mathbf{X}_1^T + \mathbf{X}_2\mathbf{X}_2^T)^{-1}\mathbf{X}_2\end{aligned}$$

Hence, if we let $\mathbf{S} = \mathbf{X}_1\mathbf{X}_1^T + \mathbf{X}_2\mathbf{X}_2^T = \mathbf{L}\mathbf{L}^T$, then $\mathbf{Q}_2^T$ and $\mathbf{L}^{-1}\mathbf{X}_2$ share the same right singular vectors.

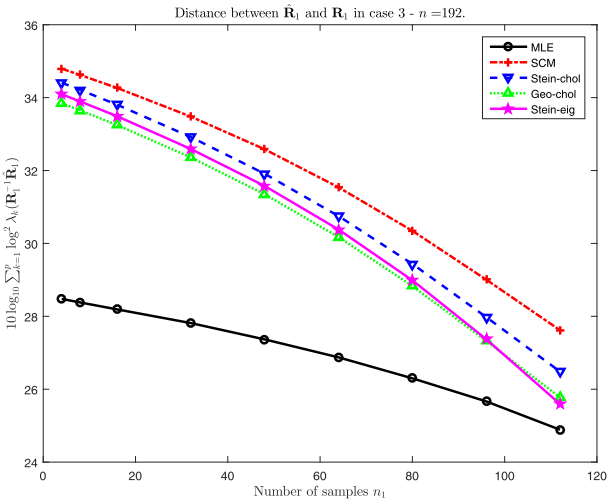
4. Numerical simulations

In this section, we evaluate numerically the performance of the MLE presented above through Monte-Carlo simulations. We consider a scenario where the size of the observation space is $p = 128$. Three cases will be considered for the covariance matrix $\mathbf{R}_1$, which correspond to different kind of processes. In the first case the (k, ℓ) element is $\mathbf{R}_1(k,\ell) = P\rho^{|k-\ell|} + \delta(k,\ell)$ with $\rho = 0.7$. The second case assumes that $\mathbf{R}_1(k,\ell) = Pe^{-0.5(2\pi\sigma_f|k-\ell|)^2} + \delta(k,\ell)$ with $\sigma_f = 0.02$. In the third case, $\mathbf{R}_1(k,\ell) = r_{\mathrm{AR}}(|k-\ell|) + \delta(k,\ell)$ where $r_{\mathrm{AR}}(|k-\ell|)$ corresponds to the correlation of an autoregressive process whose poles are located at $0.95e^{\pm i2\pi 0.05}$, $0.9e^{\pm i2\pi 0.15}$, $0.9e^{\pm i2\pi 0.18}$. Finally $P = 100$ and $r_{\mathrm{AR}}(0) = 100$. The corresponding processes are rather lowpass in case 1 and 2, while case 3 concerns processes with sharp peaks in their spectrum. In each simulation $\mathbf{X}_1$ is generated from a Gaussian distribution with covariance matrix $\mathbf{R}_1$. Then $\mathbf{W}$ is generated from a Wishart distribution with $\nu = p+2$ degrees of freedom and parameter matrix $(\nu - p - 1)^{-1}\mathbf{I}$ and $\mathbf{R}_2$ is computed as $\mathbf{R}_2 = \mathbf{G}_1\mathbf{W}^{-1}\mathbf{G}_1^T$. Then $\mathbf{X}_2$ is generated from a Gaussian distribution with covariance matrix $\mathbf{R}_2$.

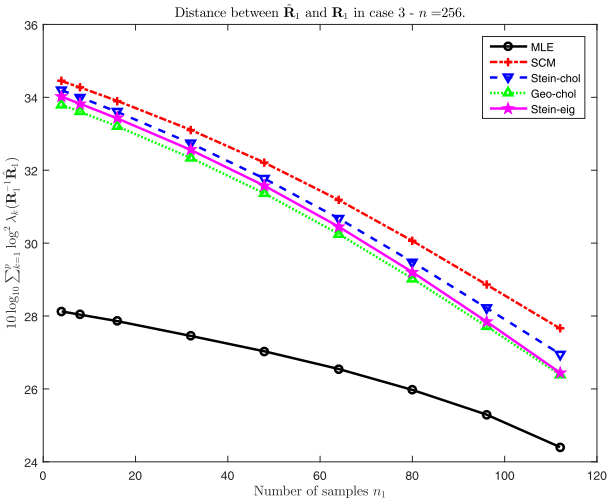
The MLE is compared with four competitors. The first is the sample covariance matrix based on all samples, i.e., $(n_1+n_2)^{-1}\mathbf{S}$ where $\mathbf{S} = \mathbf{X}_1\mathbf{X}_1^T + \mathbf{X}_2\mathbf{X}_2^T$. The second is of the form $\mathbf{G}_{\mathrm{SCM}}\mathbf{D}\mathbf{G}_{\mathrm{SCM}}^T$, where $\mathbf{G}_{\mathrm{SCM}}$ is the Cholesky factor of $\mathbf{S}$, and $\mathbf{D}$ is a diagonal matrix which is chosen to minimize Stein's loss and is given by $\mathbf{D}_{k,k} = 1/(n_1+n_2+p-2k+1)$. The third is of the same form but is meant at minimizing the natural distance between $\mathbf{R}_1$ and its estimate: as shown in [13], it amounts to choosing $\mathbf{D}_{k,k} = \exp\left\{-\mathbb{E}\left\{\log\chi^2_{n_1+n_2-i+1}\right\}\right\}$. Finally, we consider the class of orthogonally invariant estimators of the form $\mathbf{U}_{\mathrm{SCM}}\mathrm{diag}\big(\boldsymbol{\varphi}(\boldsymbol{\lambda})\big)\mathbf{U}_{\mathrm{SCM}}^T$ where $\mathbf{S} = \mathbf{U}_{\mathrm{SCM}}\mathrm{diag}\big(\boldsymbol{\lambda}\big)\mathbf{U}_{\mathrm{SCM}}^T$ is the eigenvalue decomposition of $\mathbf{S}$ and $\boldsymbol{\varphi}(\boldsymbol{\lambda}) = \begin{bmatrix}\varphi_1(\boldsymbol{\lambda}) & \dots & \varphi_p(\boldsymbol{\lambda})\end{bmatrix}$. Stein showed that the choice $\varphi_k = \lambda_k/(n_1+n_2-p+1+2\lambda_k\sum_{j\neq k}(\lambda_k-\lambda_j)^{-1})$ is the best with respect to Stein's loss. However this choice has two drawbacks: it can result in some $\varphi_k < 0$ and it does not preserve the order of the eigenvalues λ_k, which is a problem [42]. In order to overcome these problems, Stein proposed an isotonizing scheme that guarantees $\varphi_k > 0$ and preserves order, see [46] for details of this scheme. We consider this improved estimator as the fourth alternative. The figure of merit for all estimators will be the natural



(a) $n_1 + n_2 = p$



(b) $n_1 + n_2 = 3p/2$



(c) $n_1 + n_2 = 2p$

Fig. 3. Average distance between $\hat{\mathbf{R}}_1$ and $\mathbf{R}_1$ in case 3.

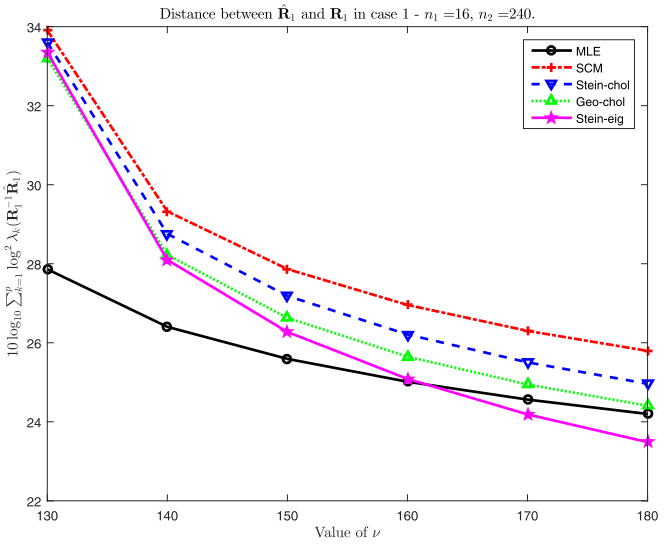


Fig. 4. Average distance between $\hat{\mathbf{R}}_1$ and $\mathbf{R}_1$ in case 1 versus ν. $n_1 + n_2 = 2p$.

distance between the true and the estimated covariance matrices $d^2(\mathbf{R}_1, \hat{\mathbf{R}}_1) = \sum_{k=1}^{p} \log^2 \lambda_k(\mathbf{R}_1^{-1}\hat{\mathbf{R}}_1)$.

The simulation results are shown in Figs. 1-4 where we consider different values for the total number of samples $n = n_1 + n_2$, namely $n = p$, $n = 3p/2$ and $n = 2p$. The main conclusions regarding these simulations are the following:

- the MLE is shown to outperform its competitors when n_1 is small and n is large enough, typically it has the best performance for $n = 3p/2$ and $n = 2p$. One can observe that the improvement achieved by the MLE is more important when $n = 2p$ and n_1 is small, i.e., when one has very few samples drawn from $\mathbf{R}_1$ and a large majority of samples drawn from $\mathbf{R}_2$.
- in contrast, when $n = p$ the other methods can perform better than the MLE, especially when n_1 is above a threshold, i.e., when the number of "good" samples is large enough.
- among the Stein-like methods, that based on eigenvalue decomposition (with isotonizing) is the best, but the method based on Cholesky factorization and minimization of the geodesic distance comes very close.

In a final simulation, we evaluate the influence of ν: recall that, as ν increases, $\mathbf{W}$ is closer to $\mathbf{I}$ and thus $\mathbf{R}_2$ is closer to $\mathbf{R}_1$, which means that $\mathbf{X}_2$ should be nearly as informative as $\mathbf{X}_1$. In Fig. 4 we display the average distance as a function of ν in case 1 with $n_1 + n_2 = 2p$. It is observed that, as ν increases, the performance of all estimators improve. The proposed MLE is no longer the most accurate above a threshold, where it is dominated by the Stein's estimator based on the eigenvalues of the whole sample covariance matrix. However, the proposed MLE still performs better than all other estimators.

5. Conclusions

In this paper, we considered the problem of estimating a covariance matrix $\mathbf{R}_1$ from two data sets, one set $\mathbf{X}_1$ whose covariance matrix is actually $\mathbf{R}_1$ and another set $\mathbf{X}_2$ whose covariance matrix $\mathbf{R}_2$ is different but close to $\mathbf{R}_1$. Since the distance between $\mathbf{R}_1$ and $\mathbf{R}_2$ depends on the eigenvalues of $\mathbf{W} = \mathbf{G}_1^T\mathbf{R}_2^{-1}\mathbf{G}_1$, we embedded the latter in a statistical model and assumed that it followed a Wishart distribution around the identity matrix. We showed that the problem is that of estimating $\mathbf{R}_1$ from two data sets with different distributions. The maximum likelihood estimator was derived and its expression was shown to depend on the number of samples in $\mathbf{X}_1$ and $\mathbf{X}_2$. The MLE was shown to perform quite well, as compared to state of the art algorithms, at least when the number of samples in $\mathbf{X}_1$ is small and the total number of samples n is large enough. However, as in a classical framework with a single data set, there is room from improvement of the MLE, especially in low sample support. Therefore, future work should be devoted to improving the MLE in this situation. For instance, one could study how the MLE could be regularized or could investigate whether a Stein-like approach is possible for this two data sets framework. Alternatively, a frequentist approach where joint estimation of $\mathbf{R}_1$ and $\mathbf{W}$ is performed under some constraints constitutes a worthy path of investigation.

Declaration of Competing Interest

None.

Appendix A. Extension to complex-valued data

In this appendix, we briefly show that the derivations concerning the maximum likelihood estimator can be extended in a straightforward manner to the complex case. Let us assume here that $\mathbf{X}_1|\mathbf{R}_1 \overset{d}{=} \mathcal{CN}(\mathbf{0}, \mathbf{R}_1, \mathbf{I})$ and $\mathbf{X}_2|\mathbf{R}_2 \overset{d}{=} \mathcal{CN}(\mathbf{0}, \mathbf{R}_2, \mathbf{I})$ are complex-valued data and distributed according to a circularly symmetric complex-valued matrix-variate normal distribution. Let $\mathbf{R}_1 = \mathbf{G}_1\mathbf{G}_1^H$ -where ^{H} stands for the Hermitian transpose- and $\mathbf{R}_2 = \mathbf{G}_1\mathbf{W}^{-1}\mathbf{G}_1^H$ where $\mathbf{W} \overset{d}{=} \mathcal{CW}_p(\nu, \mu^{-1}\mathbf{I})$ follows a complex Wishart distribution. The statistical (complex-valued) model is thus

$$p(\mathbf{X}_1, \mathbf{X}_2|\mathbf{R}_1, \mathbf{W}) = \pi^{-p(n_1+n_2)}|\mathbf{R}_1|^{-n_1}\left|\mathbf{W}^{-1}\mathbf{R}_1\right|^{-n_2} \times \text{etr}\left\{-\mathbf{X}_1^H\mathbf{R}_1^{-1}\mathbf{X}_1 - \mathbf{X}_2^H\mathbf{G}_1^{-H}\mathbf{W}\mathbf{G}_1^{-1}\mathbf{X}_2\right\} \quad \text{(A.1a)}$$

$$p(\mathbf{W}) = \frac{\mu^{\nu p}}{\tilde{\Gamma}_p(\nu)}|\mathbf{W}|^{\nu-p}\text{etr}\{-\mu\mathbf{W}\} \quad \text{(A.1b)}$$

Note that, in the complex case, $\mathbb{E}\left\{\mathbf{W}^{-1}\right\} = (\nu - p)^{-1}\mu\mathbf{I}$ [40] so that $\mathbb{E}\{\mathbf{R}_2\} = \mathbb{E}\left\{\mathbf{G}_1\mathbf{W}^{-1}\mathbf{G}_1^H\right\} = (\nu - p)^{-1}\mu\mathbf{R}_1$. Therefore, for $\mathbb{E}\{\mathbf{R}_2\}$ to be equal to $\mathbf{R}_1$, one must have $\mu = \nu - p$ in the complex case, instead of $\mu = \nu - p - 1$ in the real case.

The marginal distribution of $(\mathbf{X}_1, \mathbf{X}_2)$ is now

$$\begin{aligned}
p(\mathbf{X}_1, \mathbf{X}_2|\mathbf{R}_1) &= \int_{\mathbf{W}>0} p(\mathbf{X}_1, \mathbf{X}_2|\mathbf{R}_1, \mathbf{W})p(\mathbf{W})d\mathbf{W} \\
&= \frac{\pi^{-p(n_1+n_2)}\mu^{\nu p}}{\tilde{\Gamma}_p(\nu)}|\mathbf{R}_1|^{-(n_1+n_2)}\text{etr}\left\{-\mathbf{X}_1^H\mathbf{R}_1^{-1}\mathbf{X}_1\right\} \\
&\times \int_{\mathbf{W}>0}|\mathbf{W}|^{\nu+n_2-p}\text{etr}\left\{-\mathbf{W}\left[\mu\mathbf{I} + \mathbf{G}_1^{-1}\mathbf{X}_2\mathbf{X}_2^H\mathbf{G}_1^{-H}\right]\right\}d\mathbf{W} \\
&= \frac{\pi^{-p(n_1+n_2)}\mu^{\nu p}\tilde{\Gamma}_p(\nu)}{\tilde{\Gamma}_p(\nu)}|\mathbf{R}_1|^{-(n_1+n_2)} \\
&\text{etr}\left\{-\mathbf{X}_1^H\mathbf{R}_1^{-1}\mathbf{X}_1\right\}\left|\mu\mathbf{I} + \mathbf{G}_1^{-1}\mathbf{X}_2\mathbf{X}_2^H\mathbf{G}_1^{-H}\right|^{-(\nu+n_2)} \\
&= \pi^{-pn_1}|\mathbf{R}_1|^{-n_1}\text{etr}\left\{-\mathbf{X}_1^H\mathbf{R}_1^{-1}\mathbf{X}_1\right\} \\
&\times \frac{\pi^{-pn_2}\tilde{\Gamma}_p(\nu)}{\tilde{\Gamma}_p(\nu)}|\mu\mathbf{R}_1|^{-n_2}\left|\mathbf{I} + \mathbf{X}_2^H[\mu\mathbf{R}_1]^{-1}\mathbf{X}_2\right|^{-(\nu+n_2)}
\end{aligned} \quad \text{(A.2)}$$

and we recover the fact that $\mathbf{X}_1|\mathbf{R}_1$ is Gaussian distributed and that $\mathbf{X}_2|\mathbf{R}_1$ is Student distributed. From (A.2), the log-likelihood function is, up to an additive and constant term

$$\begin{aligned}
\tilde{f}(\mathbf{R}_1) &= -(n_1 + n_2)\log|\mathbf{R}_1| - (\nu + n_2)\log\left|\mathbf{I} + \mu^{-1}\mathbf{R}_1^{-1}\mathbf{S}_2\right| - \text{Tr}\left\{\mathbf{R}_1^{-1}\mathbf{S}_1\right\} \\
&= (\nu - n_1)\log|\mathbf{R}_1| - (\nu + n_2)\log\left|\mathbf{R}_1 + \mu^{-1}\mathbf{S}_2\right| - \text{Tr}\left\{\mathbf{R}_1^{-1}\mathbf{S}_1\right\}
\end{aligned} \quad \text{(A.3)}$$

where $\mathbf{S}_1 = \mathbf{X}_1\mathbf{X}_1^H$ and $\mathbf{S}_2 = \mathbf{X}_2\mathbf{X}_2^H$. Differentiating the previous equation, it follows that the maximum likelihood estimator of $\mathbf{R}_1$ should satisfy

$$(\nu - n_1)\mathbf{R}_1^{-1} - (\nu + n_2)\left(\mathbf{R}_1 + \mu^{-1}\mathbf{S}_2\right)^{-1} + \mathbf{R}_1^{-1}\mathbf{S}_1\mathbf{R}_1^{-1} = \mathbf{0} \qquad \text{(A.4)}$$

which is exactly (5), the equation in the real case. From there, all previous derivations follow simply by replacing the transpose by the Hermtian transpose.

References

[1] R.J. Muirhead, Aspects of Multivariate Statistical Theory, John Wiley & Sons, Hoboken, NJ, 1982.
[2] L.L. Scharf, Statistical Signal Processing: Detection, Estimation and Time Series Analysis, Addison Wesley, Reading, MA, 1991.
[3] M.S. Srivastava, Methods of Multivariate Statistics, John Wiley & Sons., New York, 2002.
[4] C. Stein, Inadmissibility of the usual estimator for the mean of a multivariate distribution, in: Proceedings 3rd Berkeley Symposium on Mathematical Statistics and Probability, 1956, pp. 197–206.
[5] C. Stein, Lectures on the theory of estimation of many parameters, J. Math. Sci. 34 (1986) 1373–1403.
[6] W. James, C. Stein, Estimation with Quadratic Loss, Springer Series in Statistics (Perspectives in Statistics), Springer, pp. 443–460.
[7] D.K. Dey, C. Srinivasan, Estimation of a covariance matrix under stein's loss, Ann. Stat. 13 (4) (1985) 1581–1591.
[8] D.K. Dey, C. Srinivasan, Trimmed minimax estimator of a covariance matrix, Ann. Inst. Stat. Math. 38 (1986) 101–108.
[9] F. Perron, Minimax estimators of a covariance matrix, J. Multivar. Anal. 43 (1) (1992) 16–28.
[10] T. Ma, L. Jia, Y. Su, A new estimator of covariance matrix, J. Stat. Plan. Inference 142 (2) (2012) 529–536.
[11] H. Tsukuma, Estimation of a high-dimensional covariance matrix with the stein loss, J. Multivar. Anal. 148 (2016) 1–17.
[12] H. Tsukuma, Minimax estimation of a normal covariance matrix with the partial Iwasawa decomposition, J. Multivar. Anal. 145 (2016) 190–207.
[13] M.-T. Tsai, On the maximum likelihood estimation of a covariance matrix, Math. Method. Stat. 27 (2018) 71–82.
[14] L.R. Haff, Empirical Bayes estimation of the multivariate normal covariance matrix, Ann. Stat. 8 (3) (1980) 586–597.
[15] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivar. Anal. 88 (2) (2004) 365–411.
[16] P. Stoica, J. Li, X. Zhu, J.R. Guerci, On using a priori knowledge in space–time adaptive processing, IEEE Trans.Signal Process. 56 (6) (2008) 2598–2602.
[17] Y. Chen, A. Wiesel, Y.C. Eldar, A.O. Hero, Shrinkage algorithms for MMSE covariance estimation, IEEE Trans.Signal Process. 58 (10) (2010) 5016–5029.
[18] T. Fisher, X. Sun, Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix, Comput. Stat. Data Anal. 55 (5) (2011) 1909–1918.
[19] A. Coluccia, Regularized covariance matrix estimation via empirical Bayes, IEEE Signal Process. Lett. 22 (11) (2015) 2127–2131.
[20] Y. Ikeda, T. Kubokawa, M.S. Srivastava, Comparison of linear shrinkage estimators of a large covariance matrix in normal and non-normal distributions, Comput. Stat. Data Anal. 95 (2016) 95–108.
[21] T. Kubokawa, M.S. Srivastava, Robust improvement in estimation of a covariance matrix in an elliptically contoured distribution, Ann. Stat. 27 (2) (1999) 600–609.
[22] F. Pascal, P. Forster, J.-P. Ovarlez, P. Larzabal, Performance analysis of covariance matrix estimates in impulsive noise, IEEE Trans. Signal Process. 56 (61) (2008) 2206–2217.
[23] Y. Chen, A. Wiesel, A.O. Hero, Robust shrinkage estimation of high-dimensional covariance matrices, IEEE Trans. Signal Process. 59 (9) (2011) 4097–4107.
[24] E. Ollila, D. Tyler, V. Koivunen, H. Poor, Complex elliptically symmetric distributions: survey, new results and applications, IEEE Trans. Signal Process. 60 (11) (2012) 5597–5625.
[25] A. Wiesel, Unified framework to regularized covariance estimation in scaled Gaussian models, IEEE Trans.Signal Process. 60 (1) (2012) 29–38.
[26] M. Mahot, F. Pascal, P. Forster, J.-P. Ovarlez, Asymptotic properties of robust complex covariance matrix estimates, IEEE Trans. Signal Process. 61 (13) (2013) 3348–3356.
[27] Y.I. Abramovich, O. Besson, Regularized covariance matrix estimation in complex elliptically symmetric distributions using the expected likelihood approach - part 1: the oversampled case, IEEE Trans. Signal Process. 61 (23) (2013) 5807–5818.
[28] O. Besson, Y.I. Abramovich, Regularized covariance matrix estimation in complex elliptically symmetric distributions using the expected likelihood approach - part 2: the under-sampled case, IEEE Trans.Signal Process. 61 (23) (2013) 5819–5829.
[29] F. Pascal, Y. Chitour, Y. Quek, Generalized robust shrinkage estimator and its application to STAP detection problem, IEEE Trans. Signal Process. 62 (21) (2014) 5640–5651.
[30] E. Ollila, E. Raninen, Optimal high-dimensional shrinkage covariance estimation for elliptical distributions, IEEE Trans. Signal Process. 67 (10) (2019) 2707–2719.
[31] W.L. Melvin, Space-time adaptive radar performance in heterogeneous clutter, IEEE Trans. Aerospace Electron.Syst. 36 (2) (2000) 621–633.
[32] , Principles of Modern Radar: Advanced Principles, W.L. Melvin, J.A. Scheer (Eds.), 2, Institution Engineering Technology, 2012.
[33] R. Nitzberg, An effect of range-heterogeneous clutter on adaptive doppler filters, IEEE Trans.Aerospace Electron.Syst. 26 (3) (1990) 475–480.
[34] D.J. Rabideau, A.O. Steinhardt, Improved adaptive clutter cancellation through data-adaptive training, IEEE Trans.Aerospace Electron.Syst. 35 (3) (1999) 879–891.
[35] L.M. Novak, Change detection for multi-polarization multi-pass SAR, in: Proceedings SPIE 5808, Algorithms for Synthetic Aperture Radar Imagery XII, 2005, pp. 234–246.
[36] N.M. Nasrabadi, Hyperspectral target detection : an overview of current and future challenges, IEEE Signal Process. Mag. 31 (1) (2014) 34–44.
[37] R. Bhatia, Positive Definite Matrices, Princeton University Press, 2007.
[38] S.T. Smith, Covariance, subspace and intrinsic Cramér-Rao bounds, IEEE Trans. Signal Process. 53 (5) (2005) 1610–1630.
[39] R.S. Raghavan, False alarm analysis of the AMF algorithm for mismatched training, IEEE Trans. Signal Process. 67 (1) (2019) 83–96.
[40] J.A. Tague, C.I. Caldwell, Expectations of useful complex Wishart forms, Multidimensional Systems and Signal Processing 5 (1994) 263–279.
[41] A.K. Gupta, D.K. Nagar, Matrix Variate Distributions, Chapman & Hall/CRC, Boca Raton, FL, 2000.
[42] Y. Sheena, A. Takemura, Inadmissibility of non-order preserving orthogonally invariant estimators of the covariance matrix in the case of Stein's loss, J. Multivar. Anal. 41 (1992) 117–131.
[43] B. Rajaratnam, D. Vincenzi, A theoretical study of Stein's covariance estimator, Biometrika 103 (2016) 653–666.
[44] B. Naul, B. Rajaratnam, D. Vincenzi, The role of isotonizing algorithm in Stein's covariance matrix estimator, Comput. Stat. 31 (4) (2016) 1453–1476.
[45] L. Guttman, General theory and methods for matric factoring, Psychometrica 9 (1) (1944) 1–16.
[46] S. Lin, M. Perlman, A Monte Carlo comparison of four estimators of a covariance matrix, in: P.R. Krishnaiah (Ed.), Multivariate Analysis VI, North Holland, Amsterdam, 1985, pp. 411–429.