Estimating high dimensional covariance matrices: A new look at the Gaussian conjugate framework

Alexis Hannart, Philippe Naveau

Journal of Multivariate Analysis 131 (2014) 149–162

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Estimating high dimensional covariance matrices: A new look at the Gaussian conjugate framework

Alexis Hannart [a,*], Philippe Naveau [b]

[a] CNRS, IFAECI, Argentina

[b] CNRS, LSCE, France

HIGHLIGHTS

- We (re)introduce a class of linear shrinkage estimators of the covariance matrix.
- We follow an empirical Bayesian approach to obtain shrinkage intensity and target.
- The method is generally applicable to any class of target matrices.
- Estimators are found to outperform those of the state-of-the-art Ledoit–Wolf class.
- The implementation is computationally light.

ARTICLE INFO

Article history:
Received 28 May 2013
Available online 11 June 2014

AMS subject classifications:
62H12
62C12
62J07

Keywords:
Covariance matrix estimation
Empirical Bayes
Shrinkage estimation

ABSTRACT

In this paper, we describe and study a class of linear shrinkage estimators of the covariance matrix that is well-suited for high dimensional matrices, has a rather wide domain of applicability, and is rooted into the Gaussian conjugate framework of Chen (1979). We propose here a new look at this framework. The linear shrinkage estimator is thereby obtained as the posterior mean of the covariance, using a Bayesian Gaussian model with conjugate inverse Wishart prior, and deriving the shrinkage intensity and target matrix by marginal likelihood maximization. We introduce some extensions to the seminal approach by deriving a closed-form expression of the marginal likelihood as well as computationally light schemes for its maximization. Further, these developments are implemented in a variety of situations and include a simulation-based performance comparison with a recent, widely used class of linear shrinkage estimators. The Gaussian conjugate estimators are found to outperform these estimators in every tested situation where the latter are available and to be more widely and directly applicable.

1. Context and motivations

Estimating the covariance matrix $\mathbf{\Sigma}$ of a p-dimensional Gaussian model is a common task in statistical analysis. Yet, it is also one which is generally recognized as particularly difficult and challenging (see, e.g., [26]). Recently, the availability of very large datasets from climate science, genomics, finance, marketing applications – among others – has exacerbated this problem with sample sizes n often much smaller than the matrix dimension p (see, e.g., [17,22,27,14]). In situations where $n < p$ the sample covariance matrix $\mathbf{S}$ performs poorly and is not positive definite, i.e. it is non invertible. When p/n has a fixed limit it is known that $\mathbf{S}$ is not consistent [8]. When $n > p$, its positive-definiteness is insured but its eigenvalues

* Corresponding author.
E-mail address: alexis.hannart@cima.fcen.uba.ar (A. Hannart).

http://dx.doi.org/10.1016/j.jmva.2014.06.001
0047-259X/

Table 1
Mapping of the nine illustrative target structures used in the article.

		Variance		
		Unit variance	Common variance	Unequal variances
Correlation	$\rho_{ij} = 0$	A1	A2	A3
	$\rho_{ij} = \rho$	B1	B2	B3
	$\rho_{ij} = \rho^{\lvert i-j\rvert}$	C1	C2	C3

tend to be distorted in such a way that $\mathbf{S}$ is ill-conditioned, implying that inverting it is possible but amplifies the estimation error. Alternative estimators of $\mathbf{\Sigma}$ have been proposed within both frequentist and Bayesian approaches, yielding substantial performance improvements compared to the sample covariance estimator $\mathbf{S}$ for small sample size n. Among these, linear shrinkage estimators are obtained as a weighted average of $\mathbf{S}$ and a covariance matrix $\mathbf{\Delta}$

$$\widehat{\mathbf{\Sigma}} = \alpha\mathbf{\Delta} + (1-\alpha)\mathbf{S}, \tag{1}$$

where the so-called shrinkage target $\mathbf{\Delta}$ is assumed to have some degree of similarity with $\mathbf{\Sigma}$. The value of the target matrix $\mathbf{\Delta}$ is usually not assumed to be known; it is commonplace to assume instead that $\mathbf{\Delta}$ has a general structure, i.e. $\mathbf{\Delta}$ is assumed to belong to a given set $\mathcal{F} \subset \mathcal{S}^{+*}$ which reflects a structural constraint (where $\mathcal{S}^{+*}$ denotes the set of symmetric positive definite matrices). We thus refer to the set $\mathcal{F}$ going forward as the *target structure*. The choice of $\mathcal{F}$ is subjective and reflects an a priori belief about $\mathbf{\Sigma}$ that may be more or less precise. For instance, it is commonplace to assume that $\mathbf{\Delta}$ is equal to the identity matrix ($\mathcal{F} = \{\mathbf{I}\}$), is proportional to the identity matrix ($\mathcal{F} = \{\lambda\mathbf{I} \mid \lambda > 0\}$) or is diagonal ($\mathcal{F} = \{\mathbf{\Lambda} \mid \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p),\ \lambda_i > 0\}$). Other more general structures are described in Appendix A and are summarized in Table 1. Beyond these general structures, the choice of $\mathcal{F}$ can also be specific, resulting from considerations that are ad-hoc to a particular context. For instance, an application to portfolio management motivated [16] to choose a structure derived from a stock return model ([24] and Appendix A).

No matter the choice of target structure $\mathcal{F}$, Eq. (1) is thus used to constrain the estimator $\widehat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$. The shrinkage allows this structural constraint to be flexible as $\mathbf{\Sigma}$ does not need to fully match the target structure—i.e. the assumption $\mathbf{\Sigma} \in \mathcal{F}$ is not required. Indeed, the introduction of the weight α – referred to as the shrinkage intensity – enables to adjust the level of structural constraint. If $\mathcal{F}$ is highly relevant, α should be chosen close to one and even equal to one if $\mathbf{\Sigma} \in \mathcal{F}$. Conversely if the relevance of $\mathcal{F}$ is poor, then α should be chosen close to zero. The shrinkage problem is thus to jointly determine an optimal value of $\mathbf{\Delta} \in \mathcal{F}$ together with an optimal value of α in $[0, 1]$.

In a frequentist framework, the shrinkage estimation strategy has been often described as one of building an optimal tradeoff between the bias of the estimator and its variance [16]. Indeed, $\mathbf{S}$ is known to have no bias but has a high variance – especially for small n – whereas on the other hand, $\mathbf{\Delta}$ has a small variance due to the constraint imposed by its underlying structure, but does have a bias if this structure does not perfectly match with that of $\mathbf{\Sigma}$. It is hoped that a weighted average of these two extreme estimators may thus yield a new, improved estimator which would balance bias and variance in an optimal way, thus borrowing strength from both extremes. This intuitive idea is discussed extensively and formalized mathematically in the seminal work of Ledoit and Wolf [16,17]. In line with the intuitive idea of an optimal bias–variance tradeoff, the framework introduced by these authors, hereinafter referred to as the LW framework, consists in minimizing in α and $\mathbf{\Delta}$ over $[0, 1] \times \mathcal{F}$ the mean squared error (mse) $\mathbb{E}(\|\alpha\mathbf{\Delta} + (1-\alpha)\mathbf{S} - \mathbf{\Sigma}\|^2)$ where $\|.\|$ denotes the Frobenius norm defined by $\|\mathbf{A}\|^2 = \mathrm{Tr}(\mathbf{A} \cdot \mathbf{A}')/p$ for any $p \times p$ matrix $\mathbf{A}$, and where $\mathbb{E}(.)$ denotes the expectation w.r.t. the random matrix $\mathbf{S}$. Under this formulation, the shrinkage estimator can be viewed geometrically as the orthogonal projection of $\mathbf{\Sigma}$ on the segment generated by $\mathbf{S}$ and $\mathbf{\Delta}$ (Fig. 1). The minimization yields:

$$\mathbf{\Delta}_o = \underset{\mathbf{\Delta}\in\mathcal{F}}{\mathrm{argmax}} \frac{(\mathbb{E}(\mathrm{Tr}((\mathbf{S}-\mathbf{\Delta})(\mathbf{S}-\mathbf{\Sigma}))))^2}{\mathbb{E}(\mathrm{Tr}((\mathbf{S}-\mathbf{\Delta})^2))} \quad \text{and} \quad \alpha_o = \frac{\mathbb{E}(\mathrm{Tr}((\mathbf{S}-\mathbf{\Delta}_o)(\mathbf{S}-\mathbf{\Sigma})))}{\mathbb{E}(\mathrm{Tr}((\mathbf{S}-\mathbf{\Delta}_o)^2))}. \tag{2}$$

Of course, Eq. (2) cannot be applied straightforwardly because the expectations in $\mathbf{S}$ therein must be evaluated to approximate the so-called *oracle estimators* α_o and $\mathbf{\Delta}_o$. The latter quantities depend on $\mathbf{\Sigma}$ and are thus not known in practice (hence the term "oracle") and must be replaced by empirical estimates to obtain the final estimators α_{lw} and $\mathbf{\Delta}_{lw}$. In favorable situations where explicit calculations can be conducted, this approach yields estimators of α and $\mathbf{\Delta}$ that have a closed form and may also have some suitable asymptotic properties. Whether or not such explicit calculations are possible depends on the choice of the target structure $\mathcal{F}$. This approach was successfully applied for the first time to our knowledge in [16] to the aforementioned stock return target structure. In [17], the same authors then adapted this approach to the generally applicable case $\mathcal{F} = \{\lambda\mathbf{I} \mid \lambda > 0\}$ to obtain:

$$\mathbf{\Delta}_{lw} = \frac{\mathrm{Tr}(\mathbf{S})}{p}\mathbf{I} \quad \text{and} \quad \alpha_{lw} = \min\left\{\frac{\sum_{i=1}^{n}\|\mathbf{S} - x_i x_i'\|^2}{n^2\left(\mathrm{Tr}(\mathbf{S}^2) - \mathrm{Tr}^2(\mathbf{S})/p\right)}, 1\right\}. \tag{3}$$

Then, [23] developed further adaptation and extension in the LW framework to cover four additional target structures (A1, A3, B1, B2). More recently, [3] have shown that the estimators of Eq. (3) can be improved substantially, especially for

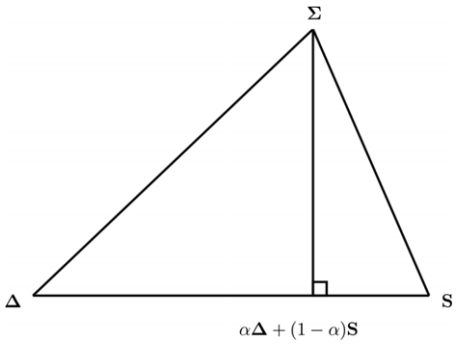

Fig. 1. After [16]. Geometric interpretation of the oracle linear shrinkage estimator as the orthogonal projection of $\mathbf{\Sigma}$ onto segment [$\mathbf{S}$, $\mathbf{\Delta}$]. Orthogonality among symmetric matrices is here defined by the inner product associated with the Frobenius norm.

small values of n, by application of the Rao–Blackwell theorem in combination with an iterative scheme insuring improved convergence towards the oracle. Most recent developments in the LW framework by [15,18] proposed a nonlinear shrinkage estimator also yielding substantial improvement but only in the case $n > p$.

On the other hand, shrinkage can be interpreted straightforwardly in the Bayesian framework as the effect of introducing an informative a priori on $\mathbf{\Sigma}$. One of the first studies illustrating this idea is [9] who showed that if the prior used on $\mathbf{\Sigma}$ is the standard conjugate, i.e. an inverse Wishart distribution having mean parameter $\mathbf{\Delta}$ and shape parameter $\nu > p - 1$, then the posterior covariance mean is obtained by

$$\mathbb{E}(\mathbf{\Sigma} \mid \mathbf{S}, \mathbf{\Delta}, \alpha) = \alpha\mathbf{\Delta} + (1 - \alpha)\mathbf{S}, \tag{4}$$

with $\alpha = (\nu - p - 1)/(n + \nu - p - 1)$. Calculations underpinning Eq. (4) are omitted at this stage but will be recalled in Section 2. Under this setting, the estimator of Eq. (1) is therefore interpretable as the a posteriori mean of $\mathbf{\Sigma}$ under an a priori mean equal to the shrinkage target $\mathbf{\Delta}$ and an a priori variance driven by the shrinkage intensity α. Thus, shrinkage occurs here as a consequence of introducing prior information into the estimation. Under this perspective, estimating the covariance is sometimes approached as a problem of defining appropriate a priori distribution and loss function. Studies such as [7,19,28,5,6,22] illustrate the latter.

A fully Bayesian approach in which the prior is set based on expert knowledge does not solve the problem of deriving optimal values of α and $\mathbf{\Delta}$ based on the data. Yet, the latter can be tackled with an approach referred to as *empirical* Bayesian which is described for instance by [10]. An empirical Bayesian approach was used for the first time in the context of covariance estimation by [2] and is hereinafter referred to as the GC approach ('Gaussian Conjugate' and/or 'Gaussian Chen'). [2] followed the conjugate formulation of [9] and obtained the hyperparameters $\mathbf{\Delta}$ and α by maximization over $\mathcal{F} \times [0, 1]$ of the marginal likelihood:

$$\ell(\alpha, \mathbf{\Delta}) = \int_{\mathbf{\Sigma}} p(\boldsymbol{x} \mid \mathbf{\Sigma}) \cdot p(\mathbf{\Sigma} \mid \alpha, \mathbf{\Delta})\, \mathrm{d}\mathbf{\Sigma}. \tag{5}$$

From a computational standpoint, [2] chose the expectation–maximization (EM) algorithm for maximizing $\ell(\alpha, \mathbf{\Delta})$. The obtained values were then introduced into Eq. (4) and the consistency of the resulting estimator was established.

After its introduction, the GC approach has been further studied by [11,25] but has not triggered further interest since then, has not been tested for performance comparison, and has never been implemented in an applied context to our knowledge. In contrast, the more recent LW approach has attracted substantial interest and has been successfully applied in a wide variety of contexts. The discrepancy in the popularity of these two approaches may come from the fact that several criticisms have been raised regarding [2]. First, the use of the inverse Wishart conjugate prior has been challenged as being too limited in its flexibility to model prior information [12], too restrictive [19], or unable to achieve the desirable eigenvalues shrinkage [28]. Second, the computational cost associated to the evaluation of the marginal likelihood required in the empirical Bayesian approach is reputedly high and potentially prohibitive [1,23]. Finally, the very principle of empirical Bayesian analysis is also sometimes questioned [16].

It is argued in this paper that in spite of these criticisms, the GC approach is rather attractive because it provides a simple framework to choose an optimal shrinkage intensity and target within a specified target structure $\mathcal{F}$. Further, it can be implemented identically for any choice of $\mathcal{F}$—a general applicability which is appealing and contrasts with the LW framework, in which calculations have to be adapted for every choice of $\mathcal{F}$ and may often be actually intractable. Accordingly, the purpose of this paper is (i) to revisit, complement and improve the GC approach; (ii) to explore the relevance of its aforementioned drawbacks as well as of its potential advantages; (iii) to evaluate whether or not its advantages overcome its drawbacks, comparatively to the LW approach. In Section 2, we first introduce notations, recall the method, and simultaneously introduce a few extensions. In Section 3, we discuss implementation issues and introduce an approximated procedure. In Section 4, we evaluate its performance based on simulations for the nine target structures of Table 1 and compare it to that of LW estimators—provided they are available. Section 5 discusses our results in perspective with aforementioned criticisms and concludes.

2. Bayesian shrinkage under conjugate prior

2.1. Model definition and covariance estimators

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be a random sample of n independent and identically distributed (i.i.d.) vectors of dimension $p \times 1$ from a multivariate Gaussian with mean zero and covariance matrix $\boldsymbol{\Sigma}$ of dimension $p \times p$. Denote $\mathbf{S} = \boldsymbol{x}\boldsymbol{x}'/n$ as the empirical covariance matrix of dimension $p \times p$. The probability density function (pdf) of $\boldsymbol{x}$ can be written as

$$p(\boldsymbol{x} \mid \boldsymbol{\Sigma}) = 2\pi^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\operatorname{Tr}(n\mathbf{S}\,\boldsymbol{\Sigma}^{-1})\right\}. \tag{6}$$

The parameter of interest here is the covariance $\boldsymbol{\Sigma}$. We wish to perform a Bayesian inference on this parameter. For this purpose, we choose to use the conjugate a priori distribution on $\boldsymbol{\Sigma}$, the so-called inverse Wishart distribution. The usual expression of the inverse Wishart pdf is

$$\mathcal{W}^{-1}(\boldsymbol{\Sigma} \mid \nu, \boldsymbol{\Omega}) = 2^{-\frac{\nu p}{2}} \Gamma_p\left(\frac{\nu}{2}\right)^{-1} |\boldsymbol{\Omega}|^{\frac{\nu}{2}} |\boldsymbol{\Sigma}|^{-\frac{\nu+p+1}{2}} \exp\left\{-\frac{1}{2}\operatorname{Tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1})\right\}, \tag{7}$$

where $\nu > p+1$ and $\boldsymbol{\Omega} \in \mathcal{S}^{+*}$ are two hyperparameters that drive shape, scale and position. For the purpose of the present paper, it is convenient to introduce the following bijective parameter change, using $(\alpha, \boldsymbol{\Delta})$ instead of $(\nu, \boldsymbol{\Omega})$

$$(\alpha, \boldsymbol{\Delta}) = \left(\frac{\nu - p - 1}{n + \nu - p - 1}, \frac{\boldsymbol{\Omega}}{\nu - p - 1}\right) \Leftrightarrow (\nu, \boldsymbol{\Omega}) = \left(\frac{\alpha\, n}{1-\alpha} + p + 1\,,\, \frac{\alpha\, n}{1-\alpha}\,\boldsymbol{\Delta}\right), \tag{8}$$

with $\alpha \in [0, 1]$ and $\boldsymbol{\Delta} \in \mathcal{S}^{+*}$. With this new choice of metaparameters, the expression of the inverse Wishart pdf becomes

$$\mathcal{W}^{-1}(\boldsymbol{\Sigma} \mid \alpha, \boldsymbol{\Delta}) \propto |\boldsymbol{\Sigma}|^{-\frac{\alpha\, n}{2(1-\alpha)} - p - 1} \exp\left\{-\frac{\alpha\, n}{2(1-\alpha)}\operatorname{Tr}(\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1})\right\}, \tag{9}$$

where we omit the normalizing constant for simplicity. Metaparameter $\boldsymbol{\Delta}$ can now be interpreted straightforwardly, as the a priori mean of $\boldsymbol{\Sigma}$

$$\mathbb{E}(\boldsymbol{\Sigma} \mid \alpha, \boldsymbol{\Delta}) = \boldsymbol{\Delta}, \tag{10}$$

whereas on the other hand, metaparameter α drives the magnitude of the a priori variance of the elements of $\boldsymbol{\Sigma}$

$$\mathrm{V}(\boldsymbol{\Sigma}_{ij} \mid \alpha, \boldsymbol{\Delta}) \simeq \frac{1-\alpha}{\alpha n}(\boldsymbol{\Delta}_{ij}^2 + \boldsymbol{\Delta}_{ii}\boldsymbol{\Delta}_{jj}), \tag{11}$$

with zero variance for $\alpha = 1$ and infinite variance for $\alpha = 0$. Finally, for a given choice of metaparameters α and $\boldsymbol{\Delta}$, we denote $\mathcal{M}_{\alpha,\boldsymbol{\Delta}}$ as the model defined by the data distribution of Eq. (6) associated to the a priori distribution of Eq. (9).

In model $\mathcal{M}_{\alpha,\boldsymbol{\Delta}}$, the conjugate prior conveniently yields a closed form expression of the posterior distribution of $\boldsymbol{\Sigma}$, which is again an inverse Wishart distribution

$$p(\boldsymbol{\Sigma} \mid \boldsymbol{x}, \alpha, \boldsymbol{\Delta}) = \mathcal{W}^{-1}(\boldsymbol{\Sigma} \mid (2-\alpha)^{-1}, (1-\alpha)\mathbf{S} + \alpha\boldsymbol{\Delta}). \tag{12}$$

We now choose to use the quadratic loss $\mathcal{L}(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|^2$ for building our estimator of $\boldsymbol{\Sigma}$. It is well known that the estimator obtained by minimizing the a posteriori expected value of this loss is then equal to the posterior mean

$$\widehat{\boldsymbol{\Sigma}} = \mathbb{E}(\boldsymbol{\Sigma} \mid \boldsymbol{x}, \alpha, \boldsymbol{\Delta}) = (1-\alpha)\mathbf{S} + \alpha\boldsymbol{\Delta}. \tag{13}$$

The estimator of $\boldsymbol{\Sigma}$ under the GC approach is thus a weighted average of the empirical covariance $\mathbf{S}$ and the a priori mean covariance $\boldsymbol{\Delta}$. It is thus a linear shrinkage estimator.

2.2. Model selection and optimal shrinkage

In Section 2.1 where our a priori knowledge on $\boldsymbol{\Sigma}$ was described by an inverse Wishart distribution with known metaparameters α and $\boldsymbol{\Delta}$, our estimation problem was solved by computing $\mathbb{E}(\boldsymbol{\Sigma} \mid \boldsymbol{x}, \alpha, \boldsymbol{\Delta})$ for model $\mathcal{M}_{\alpha,\boldsymbol{\Delta}}$. Now, we focus on a slightly different situation in which our a priori knowledge about $\boldsymbol{\Sigma}$ still corresponds to imposing a inverse Wishart family, but the information about its mean $\boldsymbol{\Delta}$ is incomplete and the magnitude of its variance is unknown. We only know that the matrix $\boldsymbol{\Delta}$ belongs to the set $\mathcal{F}$ and the group of admissible models is defined as $\mathbf{M}_{\mathcal{F}} = \{\mathcal{M}_{\alpha,\boldsymbol{\Delta}} \mid \alpha \in [0, 1], \boldsymbol{\Delta} \in \mathcal{F}\}$, rather than by the unique model $\mathcal{M}_{\alpha,\boldsymbol{\Delta}}$. In this setting, our estimation problem can therefore be treated in two steps: 1. select the most appropriate model $\mathcal{M}_{\alpha,\boldsymbol{\Delta}}$ given the data and 2. estimate $\boldsymbol{\Sigma}$ in this optimal model $\mathcal{M}_{\alpha,\boldsymbol{\Delta}}$. The second step has already been treated. For the first step, we rely on a classic model selection metric provided by the Bayesian framework,

the so-called Bayes factor [13]. For two models $\mathcal{M}_{\alpha,\mathbf{\Delta}}$ and $\mathcal{M}_{\alpha',\mathbf{\Delta}'} \in \mathbf{M}_{\mathcal{F}}$, the Bayes factor $B(\mathcal{M}_{\alpha,\mathbf{\Delta}},\ \mathcal{M}_{\alpha',\mathbf{\Delta}'})$ evaluates the additional level of evidence in favor of model $\mathcal{M}_{\alpha,\mathbf{\Delta}}$ versus model $\mathcal{M}_{\alpha,\mathbf{\Delta}}$ brought by the data. It has the following expression:

$$B(\mathcal{M}_{\alpha,\mathbf{\Delta}},\ \mathcal{M}_{\alpha',\mathbf{\Delta}'}) = \frac{m(\boldsymbol{x} \mid \alpha, \mathbf{\Delta})}{m(\boldsymbol{x} \mid \alpha', \mathbf{\Delta}')}, \tag{14}$$

where $m(\boldsymbol{x} \mid \alpha, \mathbf{\Delta}) = \int p(\boldsymbol{x} \mid \mathbf{\Sigma}) \cdot \mathcal{W}^{-1}(\mathbf{\Sigma} \mid \alpha, \mathbf{\Delta})\, \mathrm{d}\mathbf{\Sigma}$ is the marginal pdf of x in model $\mathcal{M}$ obtained by integrating out the model parameter $\mathbf{\Sigma}$ based on its prior $\mathcal{W}^{-1}(\mathbf{\Sigma} \mid \nu, \mathbf{\Omega})$ in $\mathcal{M}$. Because of conjugacy, we obtain an exact expression of $m(\boldsymbol{x} \mid \alpha, \mathbf{\Delta})$

$$\begin{aligned} m(\boldsymbol{x} \mid \alpha, \mathbf{\Delta}) &\propto \left| \frac{\alpha}{1-\alpha}\, \mathbf{\Delta} + \mathbf{S} \right|^{-\frac{1}{2}\left(\frac{n}{1-\alpha}+p+1\right)} \\ &\propto \left| \mathbf{I} + \frac{1-\alpha}{\alpha} \boldsymbol{x}' \mathbf{\Delta}^{-1} \boldsymbol{x} \right|^{-\frac{1}{2}\left(\frac{n}{1-\alpha}+p+1\right)}, \end{aligned} \tag{15}$$

where the normalizing constant is omitted, and where $\mathbf{I}$ denotes here the $n \times n$ identity matrix. Under the second expression in Eq. (15), obtained by application of Sylvester's determinant theorem, we can recognize that $m(\boldsymbol{x} \mid \alpha, \mathbf{\Delta})$ is the familiar matrix-variate t distribution. The model selection rule applied over $\mathbf{M}_{\mathcal{F}}$ based on the Bayes factor is equivalent to maximizing the quantity $\ell(\alpha, \mathbf{\Delta})$, interpretable as a matrix-variate t log-likelihood, which after a few algebra can be arranged into

$$\begin{aligned} \ell(\alpha, \mathbf{\Delta}) = &\left(\frac{\alpha\, n}{1-\alpha} + p + 1\right) \log \left| \frac{\alpha}{1-\alpha} \mathbf{\Delta} \right| - \left(\frac{n}{1-\alpha} + p + 1\right) \log \left| \mathbf{S} + \frac{\alpha}{1-\alpha} \mathbf{\Delta} \right| \\ &+ 2 \log \left(\Gamma_p \left\{ \frac{1}{2} \left(\frac{n}{1-\alpha} + p + 1 \right) \right\} \Big/ \Gamma_p \left\{ \frac{1}{2} \left(\frac{\alpha\, n}{1-\alpha} + p + 1 \right) \right\} \right). \end{aligned} \tag{16}$$

Consequently, in the above described situation of imprecise a priori knowledge, our model selection approach narrows down to deriving an optimal shrinkage intensity α^* and an optimal shrinkage target $\mathbf{\Delta}^*$ defined by

$$(\alpha^*, \mathbf{\Delta}^*) = \underset{\alpha \in [0,1],\ \mathbf{\Delta} \in \mathcal{F}}{\operatorname{argmax}}\ \ell(\alpha, \mathbf{\Delta}), \tag{17}$$

from which we obtain the associated optimal shrinkage estimator defined by

$$\widehat{\mathbf{\Sigma}}^* = (1-\alpha^*)\mathbf{S} + \alpha^* \mathbf{\Delta}^*. \tag{18}$$

A naive solution to the maximization of Eq. (17) would be to compute $\ell(\alpha, \mathbf{\Delta})$ over $[0, 1] \times \mathcal{F}$ using the closed form expression of Eq. (16). This is of course inefficient, but it is possible for low dimensional structures $\mathcal{F}$ such as A2. It is useful to do so at this point for the sake of a graphical illustration: several examples of plots of $\ell(\alpha, \mathbf{\Delta})$ in the latter case A2 are presented in Fig. 2. These plots also show that the criterion $\ell(\alpha, \mathbf{\Delta})$ appears to often reach its maximum α^* nearby the perfect oracle intensity α_o.

Beyond this naive solution, the implementation of the maximization of $\ell(\alpha, \mathbf{\Delta})$ deserves a further discussion, which is proposed in Section 3. But leaving aside implementation issues at this stage, the optimal shrinkage estimator of Eq. (18), referred to as the GC estimator, can be computed based on the mere knowledge of the data $\boldsymbol{x}$ and of the structural constraint $\mathcal{F}$. In that sense, the GC estimator resolves the shrinkage problem as formulated in Section 1 under the same conditions as the LW estimator recalled in the same section – provided it is obtainable – does. Their performance can thus be compared and it is the purpose of Section 4 to perform such a comparison, when possible, based on numerical simulation.

2.3. Asymptotic properties

Before moving to implementation aspects, asymptotic properties of the GC estimator in Eq. (18) can be established rather easily. For this purpose, we use the Stirling approximation $\log \Gamma(z) = z \log z - z + \mathcal{O}(\log z)$ for $z \in \mathbb{R}^+,\ z \to +\infty$, and apply it to Eq. (16). After a few algebra, it comes

$$\ell(\alpha, \mathbf{\Delta}) = n \log n - n \left\{ \frac{1}{1-\alpha} \log |\alpha \mathbf{I} + (1-\alpha) \mathbf{\Delta}^{-1} \mathbf{S}| + \log |\mathbf{\Delta}| + p(1 + \log 2) \right\} + \mathcal{O}(\log n), \tag{19}$$

when n goes to $+\infty$ and p is fixed. In this case, for n sufficiently large, the maximization of $\ell(\alpha, \mathbf{\Delta})$ in α for any given $\mathbf{\Delta}$ is thus equivalent to the maximization of $\ell_\infty(\alpha) = (1-\alpha)^{-1} \log |\alpha \mathbf{I} + (1-\alpha) \mathbf{\Delta}^{-1} \mathbf{S}|$ in α. Since ℓ_∞ is monotonously decreasing on $[0, 1]$, the latter maximum is reached in $\alpha = 0$. There thus exists $N \in \mathbb{N}$ such that for any $n > N$, we have $\alpha^* = 0$ and $\widehat{\mathbf{\Sigma}}^* = \mathbf{S}$. As a consequence, the asymptotic properties of $\widehat{\mathbf{\Sigma}}^*$ are the same as those of $\mathbf{S}$—in particular $\widehat{\mathbf{\Sigma}}^*$ is asymptotically consistent and efficient.

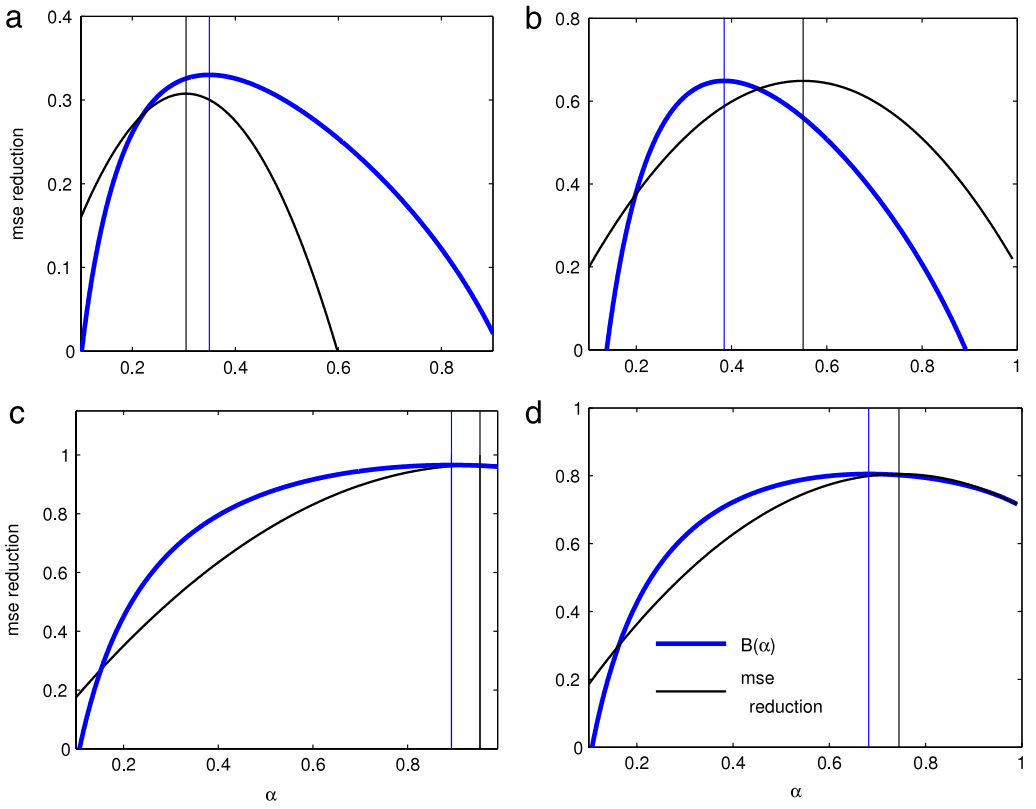

Fig. 2. Function $\ell(\alpha, \mathbf{I})$ (thick blue line) and % reduction in the squared estimation error $\|(1-\alpha)\mathbf{S} + \alpha\boldsymbol{I} - \boldsymbol{\Sigma}\|^2$ (light dark line) for $\alpha \in [0, 1]$ in four different situations. (a): $p = 200$, $n = 180$. (b): $p = 20$, $n = 10$. (c): $p = 20$, $n = 15$. (d): $p = 20$, $n = 15$. The vertical lines indicate the values of α maximizing $\ell(\alpha, \mathbf{I})$ (light blue line) and % reduction in squared estimation error (light dark line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Implementation

3.1. Exact derivation in three particular cases

We detail three particular cases of interest which allow a straightforward maximization of $\ell(\alpha, \mathbf{\Delta})$. First, when $\mathcal{F}$ is the singleton $\{\mathbf{A}\}$, $\mathbf{A} \in \mathcal{S}^{+*}$, then α^* can be obtained by maximizing $\ell(\alpha, \mathbf{A})$ unidimensionally in α over $[0, 1]$. Although this maximum cannot be obtained under a closed form even in simple cases such as $\mathbf{A} = \mathbf{I}$, its numerical implementation presents no particular technical difficulty and can be conducted based on the classic Newton–Raphson algorithm [20]. Initializing the algorithm in $\alpha^{(0)} = 1/2$, the following recurrence is applied:

$$\alpha^{(k+1)} = \alpha^{(k)} - \left(\ell''(\alpha^{(k)}, \mathbf{A})\right)^{-1} \cdot \ell'(\alpha^{(k)}, \mathbf{A}), \tag{20}$$

where ℓ' and ℓ'' are the first and second order derivatives of $\ell(\alpha, \mathbf{A})$ w.r.t. α, which are also available under a closed form expression (Appendix B). The recurrence is stopped when $|\alpha^{(k+1)} - \alpha^{(k)}|$ is lower than a chosen precision level (e.g. 10^{-3}). When running this algorithm, if p is large and n is small compared to p, note that it is computationally cheaper to replace $\log|\mathbf{S} + \frac{\alpha}{1-\alpha}\mathbf{A}|$ in $\ell(\alpha, \mathbf{A})$ using

$$\begin{aligned}\log\left|\mathbf{S} + \frac{\alpha}{1-\alpha}\mathbf{A}\right| &= \log|\mathbf{A}| + \log\left|\frac{\alpha}{1-\alpha}\mathbf{I} + \boldsymbol{x}'\mathbf{A}^{-1}\boldsymbol{x}\right| \\ &= \log|\mathbf{A}| + \sum_{i=1}^{n}\log\left(\frac{\alpha}{1-\alpha} + a_i\right),\end{aligned} \tag{21}$$

where $(a_1, \ldots, a_n)$ are the eigenvalues of the $n \times n$ matrix $\boldsymbol{x}'\mathbf{A}^{-1}\boldsymbol{x}$.

Second, when $\mathcal{F}$ is unspecified but the sample size n approaches zero, then we have

$$\lim_{n\to 0} \ell(\alpha, \mathbf{\Delta}) = -(p+1)\log\left|\mathbf{I} + \frac{1-\alpha}{\alpha}\mathbf{\Delta}^{-1}\mathbf{S}\right| = -(p+1)\sum_{i=1}^{n}\log\left(1+\frac{1-\alpha}{\alpha}\delta_i\right) \tag{22}$$

where $(\delta_1, \ldots, \delta_n)$ are the eigenvalues of $\boldsymbol{x}'\mathbf{\Delta}^{-1}\boldsymbol{x}$. Since the latter are all positive, the right hand side of Eq. (22) is maximized in $\alpha^* = 1$ for any $\mathbf{\Delta} \in \mathcal{F}$. Now, a first order Taylor expansion of $\log|\alpha\mathbf{\Delta} + (1-\alpha)\mathbf{S}|$ in the vicinity of $\alpha = 1$ yields

$$\lim_{\alpha\to 1} \ell(\alpha, \mathbf{\Delta}) = n\{\log|\mathbf{\Delta}| + \mathrm{Tr}(\mathbf{S}\mathbf{\Delta}^{-1})\}. \tag{23}$$

The right hand side of Eq. (23) is the Gaussian likelihood function. This is of course unsurprising because the prior variance of $\mathbf{\Sigma}$ in model $\mathcal{M}_{\alpha,\mathbf{\Delta}}$ is zero when $\alpha = 1$, i.e. it is a Dirac mass in $\mathbf{\Sigma} = \mathbf{\Delta}$. The marginal distribution of the data in $\mathcal{M}_{\alpha,\mathbf{\Delta}}$ is thus the Gaussian distribution with covariance $\mathbf{\Delta}$. The optimal target matrix $\mathbf{\Delta}^*$ – and consequently the optimal shrinkage estimator $\mathbf{\Sigma}^*$ which is equal to $\mathbf{\Delta}^*$ for $\alpha^* = 1$ – is therefore the Gaussian mle over $\mathcal{F}$.

Third, when the target structure is unconstrained, i.e. $\mathcal{F} = \mathcal{S}^{+*}$, then we can also obtain an exact expression of the GC estimator. Deriving $\ell(\alpha, \mathbf{\Delta})$ w.r.t. $\mathbf{\Delta}$ and using the equality $\partial\log|\mathbf{A}|/\partial\mathbf{A} = \mathbf{A}^{-1}$ that holds for any $\mathbf{A} \in \mathcal{S}^{+*}$, we have

$$\partial\ell(\alpha, \mathbf{\Delta})/\partial\mathbf{\Delta} = \left(\frac{\alpha n}{1-\alpha} + p + 1\right)\mathbf{\Delta}^{-1} - \left(\frac{n}{1-\alpha} + p + 1\right)\left(\frac{1-\alpha}{\alpha}\mathbf{S} + \mathbf{\Delta}\right)^{-1}. \tag{24}$$

The first order condition $\partial\ell(\alpha, \mathbf{\Delta})/\partial\mathbf{\Delta} = 0$ implies $\mathbf{\Delta} = \{1 + (p+1)(1-\alpha)/\alpha n\}\,\mathbf{S}$. The next step is thus to maximize in α

$$\ell\left(\alpha, \left(1 + \frac{(p+1)(1-\alpha)}{\alpha n}\right)\mathbf{S}\right) = p\,(\nu\log\nu - (\nu+n)\log(\nu+n)) + 2\log\left(\Gamma_p\left\{\frac{1}{2}(\nu+n)\right\}\Big/\Gamma_p\left\{\frac{1}{2}\nu\right\}\right) - n\log\left|\frac{1}{n}\mathbf{S}\right|, \tag{25}$$

where we use here $\nu = \alpha n/(1-\alpha) + p + 1$ for convenience. After a few calculations, we can show that $\partial\ell(\alpha, (1 + \frac{(p+1)(1-\alpha)}{\alpha n})\,\mathbf{S})/\partial\alpha$ is always positive for any α, n and p and thus the quantity in Eq. (25) reaches its maximum in $\alpha^* = 1$. Therefore, $\mathbf{\Delta}^* = \mathbf{S}$ and consequently we find that for $\mathcal{F} = \mathcal{S}^{+*}$, $\mathbf{\Sigma}^* = \mathbf{S}$. As is coherent, the GC estimator obtained when no structural constraint is imposed on the target matrix is the empirical covariance $\mathbf{S}$.

3.2. Exact derivation in the general case

To handle the general case, it is both convenient and natural to formulate the target structure $\mathcal{F}$ in a parametric manner, as in the illustrations given in Section 1. We assume in the remainder of this article that $\mathcal{F} = \{\mathbf{\Delta}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \mathbf{\Theta}\}$ where $\boldsymbol{\theta}$ is a r-dimensional vector of parameters, and $\mathbf{\Theta} \subset \mathbb{R}^r$ is such that $\mathcal{F} \subset \mathcal{S}^{+*}$. Under this parametric formulation, the maximization of Eq. (17) is equivalent to

$$(\alpha^*, \boldsymbol{\theta}^*) = \underset{\alpha\in[0,1],\,\boldsymbol{\theta}\in\mathbf{\Theta}}{\mathrm{argmax}}\ \ell(\alpha, \boldsymbol{\theta}), \tag{26}$$

where $\ell(\alpha, \boldsymbol{\theta}) \equiv \ell(\alpha, \mathbf{\Delta}(\boldsymbol{\theta}))$. From the expressions of $\partial\ell/\partial\alpha$ and $\partial\ell/\partial\boldsymbol{\theta}$ given in Appendix B, we can see that the maximization cannot be obtained under a closed form because the resolution of the first order conditions $\partial\ell/\partial\alpha = 0$ and $\partial\ell/\partial\boldsymbol{\theta} = 0$ are inextricable. Thus the maximization problem must be solved numerically. To do so, we propose to use once again the Newton–Raphson algorithm, now in its multidimensional version

$$\begin{bmatrix}\alpha^{(k+1)}\\ \boldsymbol{\theta}^{(k+1)}\end{bmatrix} = \begin{bmatrix}\alpha^{(k)}\\ \boldsymbol{\theta}^{(k)}\end{bmatrix} - \begin{bmatrix}\left.\frac{\partial^2\ell}{\partial\alpha^2}\right|_{(k)} & \left.\frac{\partial^2\ell}{\partial\alpha\partial\boldsymbol{\theta}}\right|'_{(k)}\\ \left.\frac{\partial^2\ell}{\partial\alpha\partial\boldsymbol{\theta}}\right|_{(k)} & \left.\frac{\partial^2\ell}{\partial\boldsymbol{\theta}^2}\right|_{(k)}\end{bmatrix}^{-1} \times \begin{bmatrix}\left.\frac{\partial\ell}{\partial\alpha}\right|_{(k)}\\ \left.\frac{\partial\ell}{\partial\boldsymbol{\theta}}\right|_{(k)}\end{bmatrix}, \tag{27}$$

where $|_{(k)}$ means that the value is taken in $(\alpha^{(k)}, \boldsymbol{\theta}^{(k)})$ and where the closed form expressions of the two first derivatives of ℓ in α and $\boldsymbol{\theta}$ are given in Appendix B.

For convergence to occur, the Newton–Raphson algorithm requires to be initialized near the solution. We propose to initialize the algorithm in $(\alpha^{(0)}, \boldsymbol{\theta}^{(0)}) = (1, \widetilde{\boldsymbol{\theta}}^*)$ where $\widetilde{\boldsymbol{\theta}}^*$ is the Gaussian mle of $\boldsymbol{\theta}$

$$\widetilde{\boldsymbol{\theta}}^* = \underset{\boldsymbol{\theta}\in\mathbf{\Theta}}{\mathrm{argmax}}\ \log|\mathbf{\Delta}(\boldsymbol{\theta})| + \mathrm{Tr}\{\mathbf{S}\mathbf{\Delta}^{-1}(\boldsymbol{\theta})\}. \tag{28}$$

The rationale for this choice of initial values is as follows. For a fixed $\alpha \in [0, 1]$ let us consider $\mathcal{L}_\alpha(\mathbf{\Delta}, \mathbf{S}) = \ell(\alpha, \mathbf{\Delta}) - \ell(\alpha, \mathbf{S})$, a quantity which can be viewed as a loss function. For $\alpha = 1$, we find that this loss function simplifies to $\mathcal{L}_1(\mathbf{\Delta}, \mathbf{S}) = -\log|\mathbf{S}\mathbf{\Delta}^{-1}| + \mathrm{Tr}\{\mathbf{S}\mathbf{\Delta}^{-1}\} - p$ and we recognize the so-called Stein's loss function, which measures the degree of similarity between $\mathbf{\Delta}$ and $\mathbf{S}$ in a Gaussian likelihood sense. For α equal to $\alpha^* < 1$, the loss function $\mathcal{L}_{\alpha^*}(\mathbf{\Delta}, \mathbf{S})$ has a different expression, but it still measures the degree of similarity between $\mathbf{\Delta}$ and $\mathbf{S}$—this time in a Student's t likelihood sense. It is intuitive that because the two functions $\mathcal{L}_{\alpha^*}(\mathbf{\Delta}, \mathbf{S})$ and $\mathcal{L}_1(\mathbf{\Delta}, \mathbf{S})$ both measure a degree of similarity between $\mathbf{\Delta}$ and $\mathbf{S}$, their respective minimizers in $\mathbf{\Delta}$ over a given set $\mathcal{F}$ should be closed. To illustrate this intuition, these minimizers were actually found to be identical when $\mathcal{F} = \mathcal{S}^{+*}$ (since they both coincide with $\mathbf{S}$). As a consequence, $\widetilde{\boldsymbol{\theta}}^*$ appears to be a fair choice of proxy for $\boldsymbol{\theta}^*$, which can thus be used for initializing the maximization algorithm.

It is important to emphasize that the practical interest of this proxy choice is of course to take advantage of the fact that by contrast with $\boldsymbol{\theta}^*$, the value of $\widetilde{\boldsymbol{\theta}}^*$ is often either readily available in the literature or easily obtainable under a closed or approached form. But when this is not the case, this practical interest is lost. Then, another easily obtainable proxy of $\boldsymbol{\theta}^*$—which may be based either on the minimization of a different loss $\mathcal{L}(\mathbf{\Delta}, \mathbf{S})$ such as the quadratic loss, or on a mere initial best guess for $\boldsymbol{\theta}$—should be used instead of $\widetilde{\boldsymbol{\theta}}^*$.

3.3. Approximation in the general case

When the dimension r of $\boldsymbol{\theta}$ is small – e.g. $r = 1$ (A2, B1, C1) or $r = 2$ (B2, C2) – then the numerical maximization of Section 3.2 can be treated at an affordable computational cost. Nevertheless, it may be computationally too expensive when r and p are large, e.g. $r = \mathcal{O}(p)$ (A3, B3, C3). In this situation, based on the considerations of Section 3.2, we propose instead to approximate $\boldsymbol{\theta}^*$ by $\widetilde{\boldsymbol{\theta}}^*$. Subsequently, α^* is approximated by $\widetilde{\alpha}^*$ defined by

$$\widetilde{\alpha}^* = \underset{\alpha \in [0,1]}{\mathrm{argmax}}\ \ell(\alpha, \widetilde{\boldsymbol{\theta}}^*), \tag{29}$$

which can be obtained as in the singleton $\mathcal{F}$ case of Section 3.1 based on the much lighter unidimensional maximization algorithm. Finally, $\widehat{\mathbf{\Sigma}}^*$ is naturally approximated here by $\widetilde{\mathbf{\Sigma}}^* = (1 - \widetilde{\alpha}^*)\mathbf{S} + \widetilde{\alpha}^*\mathbf{\Delta}(\widetilde{\boldsymbol{\theta}}^*)$.

In this approximation, the computations of the optimal target matrix and then of the optimal intensity of the shrinkage estimator are thus performed sequentially, similarly to the procedure used in [16,23,3]. In terms of computational complexity, the immediate implication of this approximation is that we turn a $r + 1$-dimensional maximization in $(\alpha, \boldsymbol{\theta})$ into a cheaper r-dimensional maximization in $\boldsymbol{\theta}$ followed by a one-dimensional maximization in α. However, this gain is marginal. As in Section 3.2, the main practical interest of this approximation is rather that the value of $\widetilde{\boldsymbol{\theta}}^*$ is often much easier to obtain than that of $\boldsymbol{\theta}^*$. This is the case in A2, B2 where $\widetilde{\lambda}^* = \mathrm{Tr}(\mathbf{S})/p$ and B3, C3 where $\widetilde{\mathbf{\Lambda}}^* = \mathrm{diag}(\mathbf{S}_{11}, \ldots, \mathbf{S}_{pp})$. Should not the Gaussian mle $\widetilde{\boldsymbol{\theta}}^*$ be more easily obtainable than the Student's t mle $\boldsymbol{\theta}^*$ – e.g. in C1, C2 – then this approximation does not present any interest as compared to the exact maximization – the latter then being more relevant.

Finally, the performance of $\widetilde{\mathbf{\Sigma}}^*$ is compared to that of $\widehat{\mathbf{\Sigma}}^*$ to gauge the relevance of the approximation. Asymptotically, it is immediate to show that $\widetilde{\alpha}^* = 0$ by using the same argument as for α^*. Therefore, $\lim_{n\to+\infty} \|\widetilde{\mathbf{\Sigma}}^* - \widehat{\mathbf{\Sigma}}^*\| = 0$ and the approximation is asymptotically relevant. Secondly for finite sample size n, we compared $\widetilde{\mathbf{\Sigma}}^*$ and $\widehat{\mathbf{\Sigma}}^*$ based on simulations under the B2 structure using $p = 100$ and $n = 20$. Fig. 3 shows the result of this comparison and in particular that the approximated estimator $\widetilde{\mathbf{\Sigma}}^*$ appears to perform usually as well as, yet on average – surprisingly so – better than, the exact estimator $\widehat{\mathbf{\Sigma}}^*$.

4. Numerical simulations

The purpose of this section is to assess the performance of the GC estimator by implementing it on simulated observations, and to compare it to the performance of the LW estimator—when available. We conduct this performance evaluation for the nine target structures described in Section 1 (Table 1).

4.1. Description of the set-up

Two pools of 135 experiments were created, each combining the 9 target structures with 15 values of n ranging from $n = 2$ to $n = 100$. We used $p = 100$ in all experiments. The approximated procedure of Section 3.3 was run on the first pool. A light procedure was run on the second pool by assuming that $\mathbf{\Delta}$ is known beforehand, i.e. $\mathcal{F} = \{\mathbf{\Delta}\}$. The latter procedure reduces the problem to evaluating the optimal shrinkage intensity only, as opposed to evaluating both the optimal intensity and target matrix. The interest of doing so will appear further in this section.

Two thousand simulations were performed for each experiment. For each simulation, we first generate a correlation matrix: for correlation structures 2 and 3, ρ is simulated from the uniform distribution over [0.2, 0.8]. Then, variances are generated: for the choice of variance B, we use $\lambda = 1$ without loss of generality; for the choice of variance C, $\lambda_1, \ldots, \lambda_n$ are simulated i.i.d. based on the log-normal distribution with mean equal to one and standard deviation equal to five, following [17]. The target covariance matrix $\mathbf{\Delta}$ is then built out of the correlation matrix and the variances. The covariance

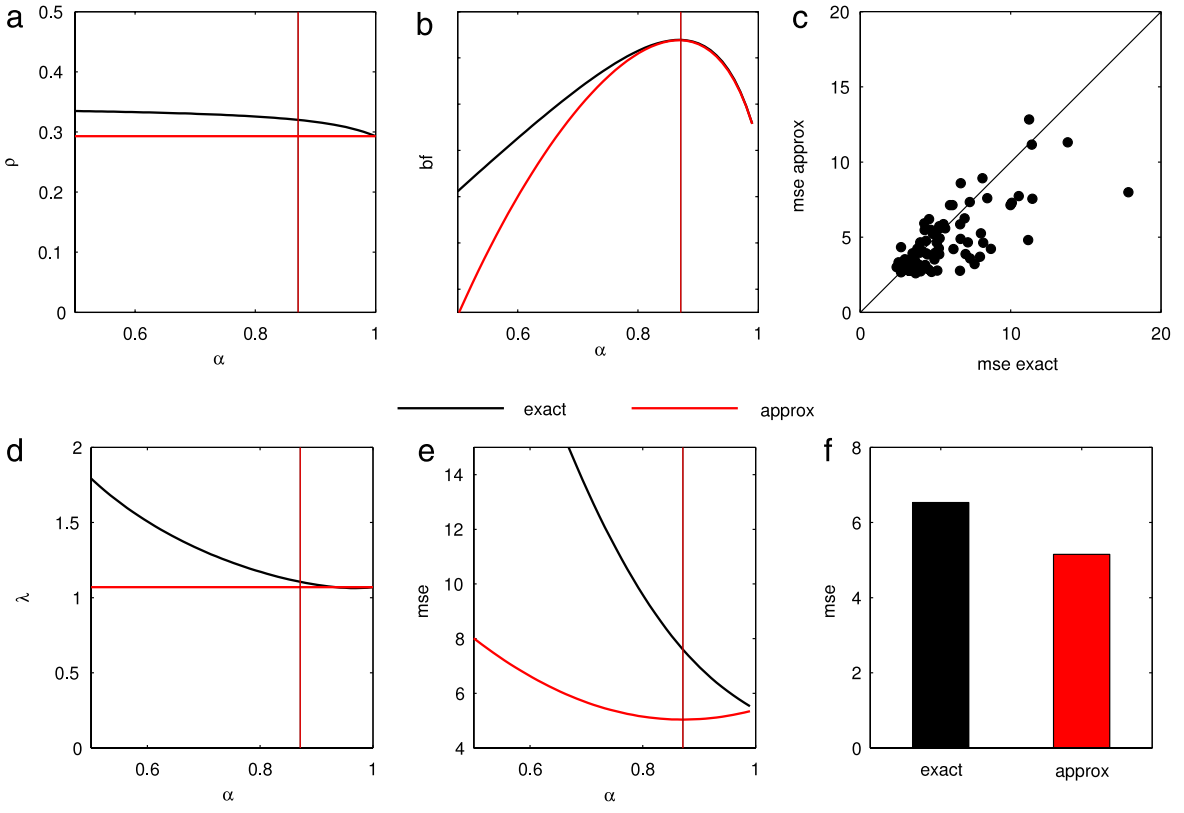

Fig. 3. Plots were obtained in situation B2, with $p = 100$ and $n = 20$, based on one or 100 simulations. (a) Plot of the exact and approximated estimators of ρ as a function of α: exact estimator $\rho^* = \text{argmax}_\rho\, \ell(\alpha, \lambda^*, \rho)$ (dark line) and approximated estimator $\widetilde{\rho}^* = \text{argmax}_\rho\, \ell(1, \widetilde{\lambda}^*, \rho)$ (red line). (d) Same for the exact estimator $\lambda^* = \text{argmax}_\lambda\, \ell(\alpha, \lambda, \rho^*)$ (dark line) and approximated estimator $\widetilde{\lambda}^* = \text{argmax}_\rho\, \ell(1, \lambda, \widetilde{\rho}^*)$ (red line). (b) Same for $\ell(\alpha, \lambda^*, \rho^*)$ (dark line) and $\ell(1, \widetilde{\lambda}^*, \widetilde{\rho}^*)$ (red line). (e) Same for the mse $\|\widehat{\mathbf{\Sigma}}^* - \mathbf{\Sigma}\|^2$ (dark line) and $\|\widetilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}\|^2$ (red line). (c) Scatterplot of the mse of the exact and approximated estimators for 100 simulations. (f) Average mse of the exact and approximated estimators over 100 simulations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

matrix $\mathbf{\Sigma}$ is randomly generated from an inverse Wishart distribution having mean $\mathbf{\Delta}$ and parameter $\nu = 1.2p + 1$. This choice yields values of $\mathbf{\Sigma}$ that are distributed around $\mathbf{\Delta}$ with a level of spread such that $\mathbb{E}(\|\mathbf{\Sigma} - \mathbf{\Delta}\|)/\|\mathbf{\Delta}\| \simeq 0.3$. The data $\boldsymbol{x}$ is finally generated from the multivariate Gaussian distribution with covariance $\mathbf{\Sigma}$.

The GC estimator was derived for each simulation of each experiment. The same was done for the LW estimator(s) when available. For target structure A2, three distinct LW estimators are available [17,3]; all three were computed. For target structures A1, A3, B1, B2, B3, only one LW estimator is available [23] whereas for target structures C1, C2, C3, no LW estimator is available to our knowledge. Besides, we also derived the oracle estimator for each simulation of each experiment. The oracle estimator is of course not applicable under real conditions where $\mathbf{\Sigma}$ is not known. Yet the interest of computing it is to obtain a performance benchmark, since the oracle estimator reaches the best possible performance (measured by mse) achievable by a linear shrinkage estimator.

Performance was measured based on mean squared error $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2$ for each estimator (GC, LW and oracle), in each simulation of each experiment. Performance was then averaged by experiment and estimator.

4.2. Results

We first emphasize that, as expected, the 'shrinkage magic' happens: both the LW estimator and the GC estimator perform better than the empirical estimator $\mathbf{S}$ and the target estimator $\mathbf{\Delta}$, i.e. a well-chosen linear combination of $\mathbf{S}$ and $\mathbf{\Delta}$ does prove to outperform any of these two estimators individually (Fig. 4). By construction, LW and GC also both perform worse than the oracle estimator. Unsurprisingly, when n is large performances, of the GC and LW estimators converge towards that of the empirical estimator $\mathbf{S}$.

In terms of performance comparison between GC and LW, we find that the GC estimator performs on average better than the LW estimator in all our experiments (Figs. 4, 5).

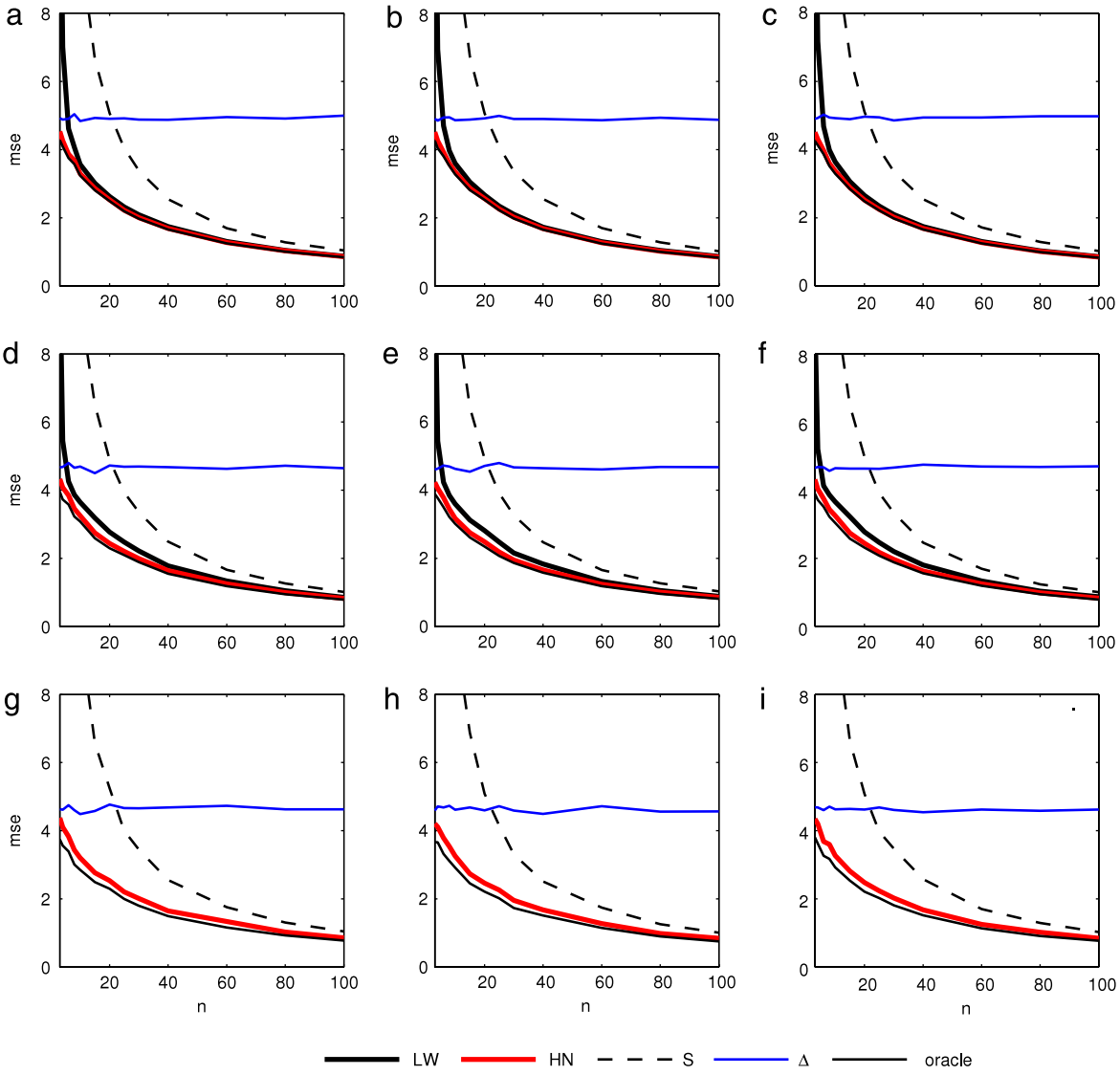

Fig. 4. Average mse of the five estimators of $\mathbf{\Sigma}$ for n ranging from 2 to 100 and $p = 100$: LW (thick dark line), GC (thick red line), $\mathbf{S}$ (dashed dark line), $\mathbf{\Delta}$ (thin blue line), oracle (thin dark line). (a) A1, (b), A2, (c) A3, (d) B1, (e) B2, (f) B3, (g) C1, (h) C2, (i) C3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The mse of the GC estimator for the target B2 is approximately 2%–3% higher than the mse of the oracle whereas the mse of the LW estimator is approximately 10% higher (Fig. 5). Further, these performance ratios are not very sensitive to n, except for the steep degradation of the performance of the LW estimator observed for very small values of n ($n \leq 5$–10) for which by contrast the GC performance ratio w.r.t. the oracle maintains. Roughly similar results are found for the other five targets for which the LW estimator is available (A1, A2, A3, B1, B3).

In an attempt to understand the reason for this gap in performance, we computed the average over each experiment of the shrinkage intensities α^*_{lw} and α^* associated to estimators LW and GC, and we compared them to the average 'perfect' shrinkage intensity α_o associated to the oracle. We find that in situation B2, the average intensity of the GC estimator matches almost perfectly the average oracle intensity: their two curves are almost indistinguishable (Fig. 5(c)). In that sense, it can be said that the GC intensity is unbiased. By contrast, the LW intensity is biased: it slightly overestimates the oracle intensity for intermediate to high values of n and steeply underestimates it for small values of n. Besides, in addition to this bias, we find that the variance of the LW intensity is also higher than the variance of the GC intensity, resulting overall in a higher mse on intensity for the LW estimator (Fig. 5(d)).

It thus appears that, despite the fact the LW estimator aims by construction to estimate the optimal oracle intensity α_o, this estimation is biased and also quite noisy as compared to the GC estimator—this, when applying the straightforward LW approach with direct empirical substitution as recalled in Section 1. This point was actually already spotted by [3] in the

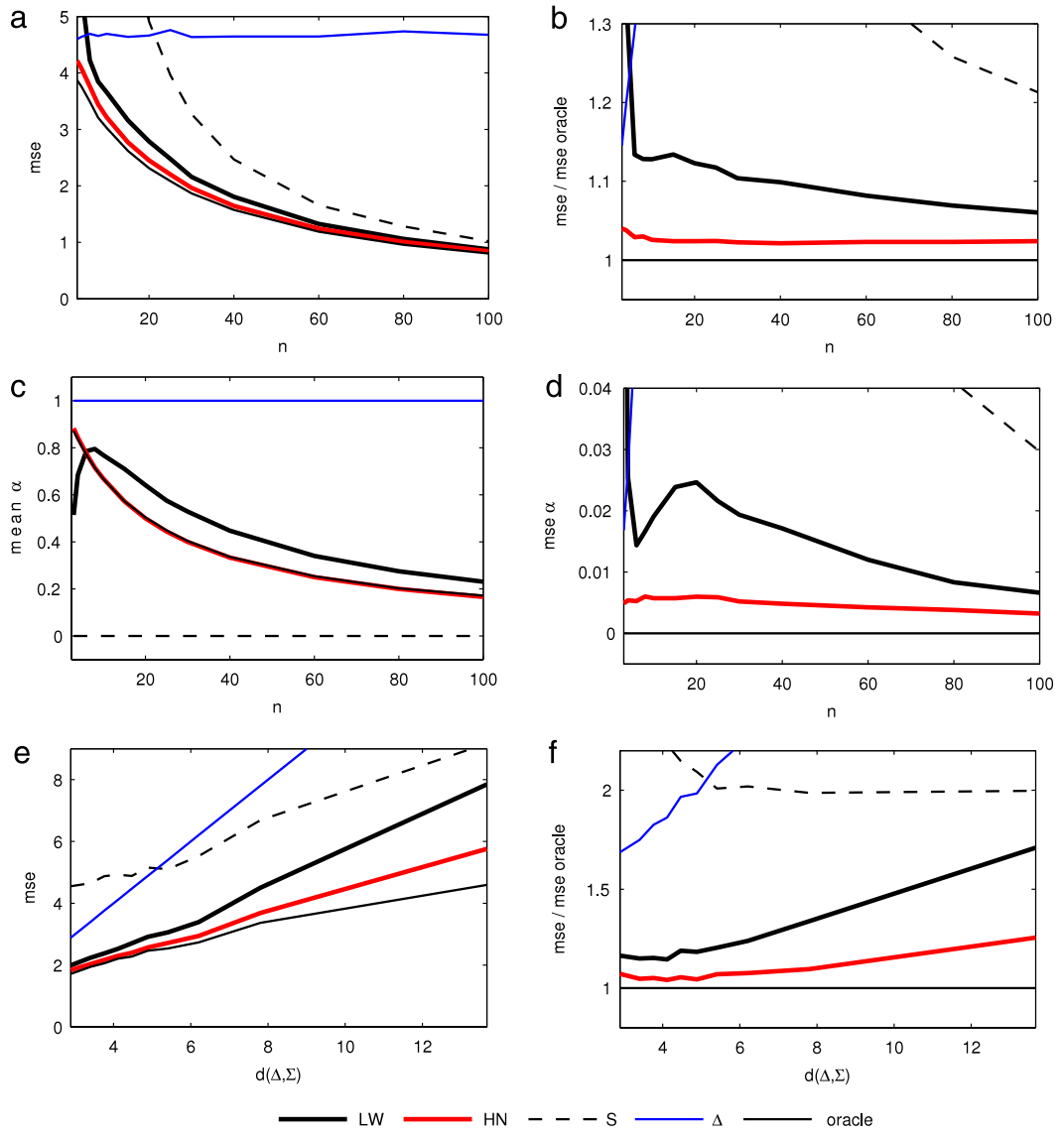

Fig. 5. (a) Average mse of the following five estimators of $\mathbf{\Sigma}$ for n ranging from 2 to 100 and $p = 100$: LW (thick dark line), GC (thick red line), $\mathbf{S}$ (dashed dark line), $\mathbf{\Delta}$ (thin blue line), oracle (thin dark line) in situation B2. (b) Same for the average ratio of mse to mse of the oracle. (c) Same for the average shrinkage intensity. (d) Same for the average mse of the intensity. (e) Same for the average mse of the estimator as a function of the distance between $\mathbf{\Delta}$ and $\mathbf{\Sigma}$. (f) Same for the average ratio of mse to mse of the oracle as a function of the distance between $\mathbf{\Delta}$ and $\mathbf{\Sigma}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

case of A2, and was the motivation of their work. These authors proposed two improved LW estimators in this case, which we implemented. We find that one of the two proposed correction (OAS) successfully resolves the issue as the corrected intensity estimator α_{oas} is now found to be unbiased (Fig. 6(a), (c)). The performance of the corresponding estimator of $\mathbf{\Sigma}$ is significantly improved as a result—yet the GC estimator still outperforms it slightly (Fig. 6(b)).

5. Conclusion

The Gaussian Conjugate (GC) approach to linear shrinkage for covariance estimation was first introduced decades ago by [2] and has apparently been somewhat forgotten since then. In particular, it had never been implemented in real applications nor tested and compared in performance to recent, state-of-the-art shrinkage estimators of the LW class so far.

In this paper, we proposed a new look at the GC approach and we provided two main extensions to its seminal introduction. First, we derived a closed form expression of the criterion $\ell(\alpha, \mathbf{\Delta})$ to be maximized in α and $\mathbf{\Delta}$ which is interpretable as a matrix-variate t likelihood. Surprisingly and remarkably, this expression does not appear in [2] who

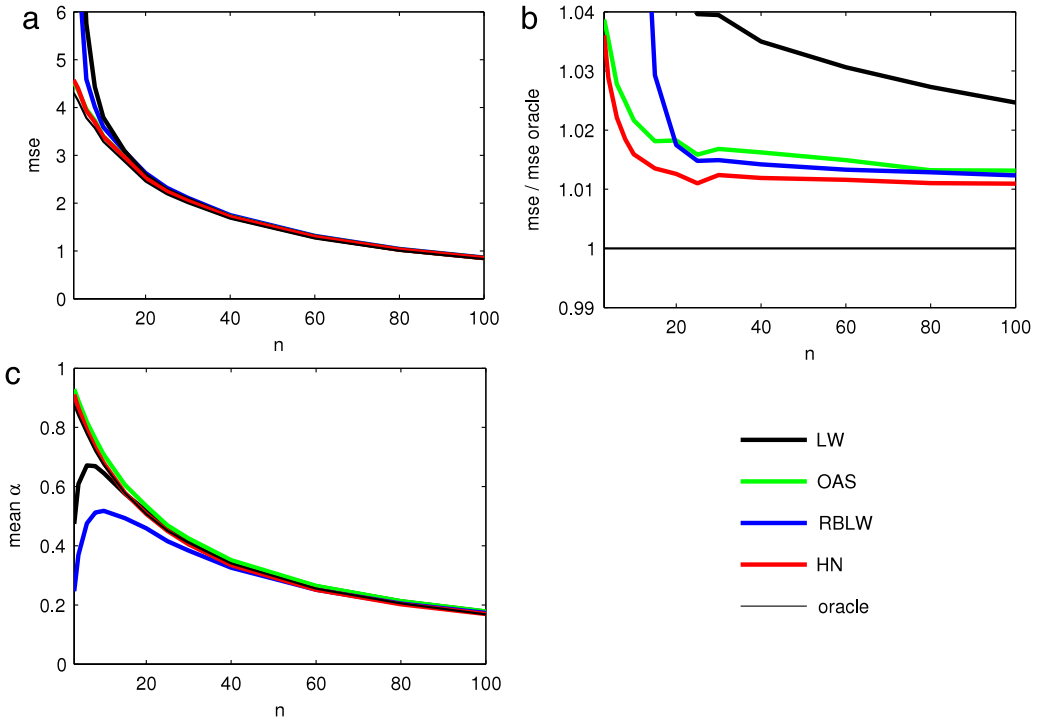

Fig. 6. (a) Average mse of the following five estimators of $\mathbf{\Sigma}$ for n ranging from 2 to 100 and $p = 100$: LW (thick dark line), OAS (thick green line), RBLW (thick blue line), GC (thick red line), oracle (thin dark line) in situation A2. (b) Same for the average ratio of mse to mse of the oracle. (c) Same for the average shrinkage intensity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

preferred an indirect EM maximization, this despite the fact that its derivation is straightforward. Second, since the maximization of $\ell(\alpha, \mathbf{\Delta})$ cannot be obtained in a closed form, we proposed an approximation which proceeds in two steps: first, the Gaussian mle $\widetilde{\mathbf{\Delta}}^*$ over the target structure set $\mathcal{F}$ is derived as an approximation for the optimal target matrix $\mathbf{\Delta}^*$, and second an approximated optimal intensity $\widetilde{\alpha}^*$ is obtained by maximizing $\ell(\alpha, \widetilde{\mathbf{\Delta}}^*)$ in α. The first step is often immediate since the mle may have a closed form or can be derived easily, and the second step is a univariate maximization over $[0, 1]$ which is computationally cheap and technically easy to implement. These extensions may thus help mitigate one of the main critiques of the approach of [2], i.e. its high computational cost.

A performance comparison based on numerical simulations showed that the approximated GC estimator outperforms the LW estimator, based on implementation of both on the six generic target structures for which the latter is available. On the other hand, the modified LW estimator obtained by [3] in the case A2 appears to be able to reach a similar level of performance than the GC estimator, but it is applicable only to the case A2. These findings thus challenge another critique of the approach of [2], i.e. its claimed inability to achieve the desirable eigenvalues shrinkage as well as its reputed lack of flexibility, associated to the use of the inverse Wishart conjugate prior.

In addition to its performance advantage, perhaps the main advantage of the GC approach w.r.t. the LW approach is its general applicability. As was underlined in Section 1, whereas the LW estimator must be at best redesigned and recomputed for each and every different target structure $\mathcal{F}$ considered, and on the other hand may not be systematically obtainable, the GC estimator can be derived by maximizing the exact same criterion $\ell(\alpha, \mathbf{\Delta})$ over each different $\mathcal{F}$. A natural extension of this work would therefore be to apply the GC approach in distinct applicational contexts involving ad-hoc structures $\mathcal{F}$. For example in portfolio management, the single-index structure of [16] could be tested in the GC approach, and compared to the corresponding LW estimator. Similarly, in atmospheric science and climate science, ACP analysis (or EOF analysis) is commonly used to identify variability modes, often in a situation of large p and small n. An ad-hoc spatial target structure such as the exponential, squared exponential or Matérn structure, could be used here for shrinkage estimation of the covariance, yielding regularized spatial eigenvectors that may be smoother and easier to interpret. Various other important applications originating from the same fields (e.g. data assimilation, detection and attribution) also require an estimate of the covariance matrix and could similarly benefit from the GC estimator.

Finally, there has been much progress made in recent years regarding testing for specific structures of high dimensional covariance matrices. This includes the tests for identity and sphericity (i.e. A1 and A2 respectively) of [4] and the test for $\mathbf{\Sigma}$ being banded (i.e. A3 in the particular case where the bandwidth is equal to one) of [21]. Such tests may be valuable options to select a suitable target structure $\mathcal{F}$.

Appendix A. Examples of target structures

One may assume a general structure on the correlation matrix $\mathbf{R}$ and decline it over three possible structures for the covariance matrix: the unit variance ($\mathcal{F} = \{\mathbf{R}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$), the common variance ($\mathcal{F} = \{\lambda\mathbf{R}(\boldsymbol{\theta}) \mid \lambda > 0, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$), or the unequal variance structure ($\mathcal{F} = \{\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{R}(\boldsymbol{\theta})\boldsymbol{\Lambda}^{\frac{1}{2}} \mid \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p), \lambda_i > 0, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$). Simple correlation structures are for instance the identity ($\mathbf{R} = \mathbf{I}$), the constant correlation structure ($\mathbf{R}_{ij} = \rho$ for $i \neq j$), or the autoregressive correlation structure ($\mathbf{R}_{ij} = \rho^{|i-j|}$) with $\rho \in]0, 1[$. For the above examples, three correlation structures associated to three variance choices yield nine possible combinations of target structure that are recalled in Table 1 and are used for illustration across this article.

On the other hand, the stock return model of [24] leads to the following specific structure: $\mathcal{F} = \{\sigma_0^2\beta\beta' + \boldsymbol{\Lambda} \mid \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p), \lambda_i > 0, \sigma_0 > 0, \beta \in \mathbb{R}^p\}$.

Appendix B. First and second derivatives of $\ell(\alpha, \boldsymbol{\theta})$

We denote ψ_p as the p-multivariate digamma function; ψ_p' as the p-multivariate trigamma function; and $(\delta_i)_{i=1}^n$ as the n eigenvalues of $\boldsymbol{x}'\boldsymbol{\Delta}^{-1}\boldsymbol{x}$. From Eq. (16) we successively obtain after some tedious algebra:

$$\begin{aligned}
\partial\ell(\alpha, \boldsymbol{\Delta})/\partial\alpha \;=\; & \frac{np}{(1-\alpha)^2}\left(1 + \log\left(\frac{\alpha}{1-\alpha}\right)\right) + \frac{p(p+1)}{\alpha(1-\alpha)} - \frac{n}{(1-\alpha)^2}\sum_{i=1}^{n}\log\left(\frac{\alpha}{1-\alpha} + \delta_i\right) \\
& - \frac{\alpha n + (p+1)(1-\alpha)}{(1-\alpha)^2}\sum_{i=1}^{n}(\alpha(1-\delta_i) + \delta_i)^{-1} + \frac{n}{(1-\alpha)^2}\left(\psi_p\left(\frac{\nu}{2}\right) - \psi_p\left(\frac{\nu+n}{2}\right)\right) \\
\partial^2\ell(\alpha, \boldsymbol{\Delta})/\partial\alpha^2 \;=\; & \frac{2np}{(1-\alpha)^3}\left(1 + \frac{1}{2\alpha} + \log\left(\frac{\alpha}{1-\alpha}\right)\right) + \frac{p(p+1)(2\alpha-1)}{\alpha(1-\alpha)^2} \\
& - \frac{2n}{(1-\alpha)^3}\sum_{i=1}^{n}\log\left(\frac{\alpha}{1-\alpha} + \delta_i\right) - \frac{2n(1+\alpha) + (p+1)(1-\alpha)}{(1-\alpha)^3}\sum_{i=1}^{n}(\alpha(1-\delta_i) + \delta_i)^{-1} \\
& + \frac{\alpha n + (p+1)(1-\alpha)}{(1-\alpha)^3}\sum_{i=1}^{n}(\alpha(1-\delta_i) + \delta_i)^{-2} + \frac{2n}{(1-\alpha)^3}\left(\psi_p\left(\frac{\nu}{2}\right) - \psi_p\left(\frac{\nu+n}{2}\right)\right) \\
& + \frac{n^2}{2(1-\alpha)^4}\left(\psi_p'\left(\frac{\nu}{2}\right) - \psi_p'\left(\frac{\nu+n}{2}\right)\right).
\end{aligned} \tag{B.1}$$

Using the notation $\nu = \frac{\alpha}{1-\alpha}n + p + 1$ for parsimony and considering the equality $\partial \log|\mathbf{A}|/\partial\mathbf{A} = \mathbf{A}^{-1}$ it becomes:

$$\partial\ell(\alpha, \boldsymbol{\Delta})/\partial\boldsymbol{\Delta} = \nu\boldsymbol{\Delta}^{-1} - (\nu+n)\left(\frac{1-\alpha}{\alpha}\mathbf{S} + \boldsymbol{\Delta}\right)^{-1}. \tag{B.2}$$

Using Eq. (B.2) as well as the equalities $\partial f(\mathbf{A})/\partial x = \mathrm{Tr}\{\partial f(\mathbf{A})/\partial\mathbf{A} \times \partial\mathbf{A}/\partial x\}$ and $\partial\mathbf{A}^{-1}/\partial x = -\mathbf{A}^{-1} \times \partial\mathbf{A}/\partial x \times \mathbf{A}^{-1}$ we successively obtain after additional tedious algebra:

$$\begin{aligned}
\partial\ell(\alpha, \boldsymbol{\theta})/\partial\theta_k &= \nu\,\mathrm{Tr}\left\{\boldsymbol{\Delta}^{-1} \times \partial\boldsymbol{\Delta}/\partial\theta_k\right\} + (\nu+n)\mathrm{Tr}\left\{\left(\frac{1-\alpha}{\alpha}\mathbf{S} + \boldsymbol{\Delta}\right)^{-1} \times \partial\boldsymbol{\Delta}/\partial\theta_k\right\} \\
\partial^2\ell(\alpha, \boldsymbol{\theta})/\partial\alpha\partial\theta_k \;=\; & \frac{n}{(1-\alpha)^2}\,\mathrm{Tr}\left\{\boldsymbol{\Delta}^{-1} \times \partial\boldsymbol{\Delta}/\partial\theta_k\right\} + \frac{n}{(1-\alpha)^2}\,\mathrm{Tr}\left\{\left(\frac{1-\alpha}{\alpha}\mathbf{S} + \boldsymbol{\Delta}\right)^{-1} \times \partial\boldsymbol{\Delta}/\partial\theta_k\right\} \\
& - \frac{\nu+n}{\alpha^2}\,\mathrm{Tr}\left\{\left(\frac{1-\alpha}{\alpha}\mathbf{S} + \boldsymbol{\Delta}\right)^{-1} \times \mathbf{S} \times \left(\frac{1-\alpha}{\alpha}\mathbf{S} + \boldsymbol{\Delta}\right)^{-1} \times \partial\boldsymbol{\Delta}/\partial\theta_k\right\} \\
\partial^2\ell(\alpha, \boldsymbol{\theta})/\partial\theta_k\partial\theta_l \;=\; & \nu\,\mathrm{Tr}\left\{\boldsymbol{\Delta}^{-1} \times \partial^2\boldsymbol{\Delta}/\partial\theta_k\partial\theta_l\right\} + (n+\nu)\mathrm{Tr}\left\{\left(\frac{1-\alpha}{\alpha}\mathbf{S} + \boldsymbol{\Delta}\right)^{-1} \times \partial^2\boldsymbol{\Delta}/\partial\theta_k\partial\theta_l\right\} \\
& - \nu\,\mathrm{Tr}\left\{\boldsymbol{\Delta}^{-1} \times \partial\boldsymbol{\Delta}/\partial\theta_l \times \boldsymbol{\Delta}^{-1} \times \partial\boldsymbol{\Delta}/\partial\theta_k\right\} \\
& - (\nu+n)\mathrm{Tr}\left\{\left(\frac{1-\alpha}{\alpha}\mathbf{S} + \boldsymbol{\Delta}\right)^{-1} \times \partial\boldsymbol{\Delta}/\partial\theta_l \times \left(\frac{1-\alpha}{\alpha}\mathbf{S} + \boldsymbol{\Delta}\right)^{-1} \times \partial\boldsymbol{\Delta}/\partial\theta_k\right\}.
\end{aligned} \tag{B.3}$$

References

[1] J. Bai, S. Shi, Estimating high dimensional covariance matrices and its applications, Ann. Econ. Finance 12 (2) (2011) 199–215.
[2] C.F. Chen, Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis, J. R. Stat. Soc. Ser. B 41 (1979) 235–248.
[3] 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.
[4] S.X. Chen, L.X. Zhang, P.-S. Zhong, Tests for high-dimensional covariance matrices, J. Amer. Statist. Assoc. 105 (2010) 810–819.
[5] M.J. Daniels, R.E. Kass, Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models, J. Amer. Statist. Assoc. 94 (1999) 1254–1263.
[6] M.J. Daniels, R.E. Kass, Shrinkage estimators for covariance matrices, Biometrics 57 (2001) 1173–1184.
[7] D.K. Dey, C. Srinivasan, Estimation of a covariance matrix under Stein's loss, Ann. Statist. 13 (4) (1985) 1581–1591.
[8] N. El Karoui, On the largest eigenvalue of Wishart matrices with identity covariance when n, p and n/p tend to infinity. Unpublished Manuscript., 2011.
[9] I.G. Evans, Bayesian estimation of parameters of a multivariate normal distribution, J. R. Stat. Soc. Ser. B 27 (1965) 279–283.
[10] A. Gelman, J.B. Clarin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, Chapman and Hall, London, 1997.
[11] L.R. Haff, Empirical Bayes estimation of the multivariate normal covariance matrix, Ann. Statist. 8 (1980) 586–597.
[12] C.W. Hsu, M.S. Sinay, J.S.J. Hsu, Bayesian estimation of a covariance matrix with flexible prior specification, Ann. Inst. Statist. Math. 64 (2) (2012) 319–342.
[13] R.E. Kass, A.E. Raftery, Bayes factors, J. Amer. Statist. Assoc. 90 (430) (1994) 773–795.
[14] K. Khare, B. Rajaratnam, Wishart distributions for covariance graph models, Ann. Statist. 39 (1) (2011) 514–555.
[15] O. Ledoit, S. Péché, Eigenvectors of some large sample covariance matrix ensembles, Probab. Theory Related Fields 151 (2012) 233–264.
[16] O. Ledoit, M. Wolf, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, J. Empir. Finance 10 (2003) 603–621.
[17] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal. 88 (2004) 365–411.
[18] O. Ledoit, M. Wolf, Nonlinear shrinkage estimation of large-dimensional covariance matrices, Ann. Statist. 40 (2) (2012) 1024–1060.
[19] T. Leonard, J.S.J. Hsu, Bayesian inference for a covariance matrix, Ann. Statist. 20 (1992) 1669–1696.
[20] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes: The Art of Scientific Computing, third ed., Cambridge University Press, New York, USA, 2007.
[21] Y.M. Qiu, S.X. Chen, Test for bandedness of high dimensional covariance matrices with bandwidth estimation, Ann. Statist. 40 (2012) 1285–1314.
[22] B. Rajarratnam, H. Massam, C. Carvahlo, Flexible covariance estimation in graphical Gaussian models, Ann. Statist. 36 (2008) 2818–2849.
[23] J. Schäfer, K. Strimmer, A Shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, Stat. Appl. Genet. Mol. Biol. 4 (1) (2005) 32.
[24] W.F. Sharpe, A simplified model for portfolio analysis, Manag. Sci. 9 (1963) 277–293.
[25] D. Sharma, K. Krishnamoorthy, Empirical Bayes estimators of normal covariance matrix, Indian J. Statist. Ser. A 47 (2) (1985) 247–254.
[26] C. Stein, Estimation of a covariance matrix, Rietz Lecture, in: 39th Annual Meeting IMS, Atlanta, GA., 1975.
[27] J. Won, S.J. Kim, J. Lim, B. Rajaratnam, Condition number-regularized covariance estimation, J. R. Stat. Soc. Ser. B 75 (3) (2013) 427–450.
[28] R. Yang, J.O. Berger, Estimation of a covariance matrix using the reference prior, Ann. Statist. 22 (1994) 1195–1211.