Statistics and Probability Letters 167 (2020) 108893

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A Stein's approach to covariance matrix estimation using regularization of Cholesky factor and log-Cholesky metric

Olivier Besson [*], François Vincent, Xavier Gendre

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

ARTICLE INFO

Article history:
Received 6 April 2020
Received in revised form 5 June 2020
Accepted 20 July 2020
Available online 30 July 2020

Keywords:
Cholesky factor
Covariance matrix estimation
Regularization
Steins's estimation

ABSTRACT

We consider a Stein's approach to estimate a covariance matrix using regularization of the sample covariance matrix Cholesky factor. We propose a method to estimate accurately the regularization vector which minimizes the risk associated with the recently introduced log-Cholesky metric.

1. Introduction and problem statement

Covariance matrix (CM) estimation is at the core of most methods used to process multichannel data, in a wide variety of applications, including social science, life science, physics, engineering, finance. The estimation of the covariance matrix is indeed needed for the most widely used tools of multivariate analysis, e.g., principal component analysis, adaptive detection, filtering (Scharf, 1991; Pourahmadi, 2013; Srivastava, 2002). Under the Gaussian assumption, the maximum likelihood estimate (MLE) of the covariance matrix is the sample covariance matrix (SCM) $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ where $\mathbf{X}$ is the $p \times n$ data matrix where columns of $\mathbf{X}$ are assumed to be independent and to follow a normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$. Unfortunately, when the number of observations n is not significantly larger than the observation size p, $\mathbf{S}$ has been observed to be much less well-conditioned than $\boldsymbol{\Sigma}$. More precisely, large eigenvalues of $\boldsymbol{\Sigma}$ tend to be over-estimated while small eigenvalues tend to be under-estimated. Therefore, there has been a natural need to somehow regularize the SCM.

The literature about this problem is huge. However, the approach proposed by Stein (1956, 1986), James and Stein (1992) has markedly emerged and influenced a great deal of research, see e.g., Haff (1979, 1980), Dey and Srinivasan (1985, 1986), Perron (1992), Ledoit and Wolf (2004), Ma et al. (2012) and Tsukuma (2016) and references therein. The basic principle of Stein's approach is to define a loss function and to minimize its average value, referred to as the risk, within a given class of estimates $\hat{\boldsymbol{\Sigma}}$. Three main classes of estimates have been considered: regularization of the eigenvalues of the SCM, regularization of its Cholesky factor or shrinkage. The first class involves estimates of the form $\hat{\boldsymbol{\Sigma}} = \mathbf{U}\text{diag}(\boldsymbol{\varphi}(\boldsymbol{\lambda}))\mathbf{U}^T$ where $\mathbf{S} = \mathbf{U}\text{diag}(\boldsymbol{\lambda})\mathbf{U}^T$ denotes the eigenvalue decomposition of the SCM and $\boldsymbol{\varphi}(\boldsymbol{\lambda})$ is some non-linear function of the eigenvalues $\boldsymbol{\lambda}$. For the risk $\mathcal{L}_1(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = \text{Tr}\{\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\} - \log\det(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) - p$, it was proposed to minimize an unbiased estimate of the corresponding risk, this unbiased estimate being obtained from the so-called Stein–Haff identity for Wishart matrices (Haff, 1979), or using the distribution of $\boldsymbol{\lambda}$ (Sheena, 1995). It turns out that this procedure results

[*] Corresponding author.
E-mail addresses: olivier.besson@isae-supaero.fr (O. Besson), francois.vincent@isae-supaero.fr (F. Vincent), xavier.gendre@isae-supaero.fr (X. Gendre).

https://doi.org/10.1016/j.spl.2020.108893
0167-7152/© 2020 Elsevier B.V. All rights reserved.

in some eigenvalues of $\hat{\boldsymbol{\Sigma}}$ being negative and, moreover, it does not preserve the order of the eigenvalues in $\boldsymbol{\lambda}$. This means that it can be improved upon by an estimator which preserves the order (Sheena and Takemura, 1992). In order to overcome these problems, Stein proposed an isotonizing scheme which guarantees that the eigenvalues $\boldsymbol{\varphi}(\boldsymbol{\lambda})$ are all positive and in decreasing order, see Lin and Perlman (1985) for details. Note that Ledoit and Wolf, applying to the theory of large-dimensional asymptotics, proposed a procedure that alleviates all above mentioned problems and minimizes an unbiased estimate of the risk, in the spirit of Stein's approach (Ledoit and Wolf, 2018).

Another class of estimates amounts to shrinkage of the SCM to the identity matrix, leading to estimates of the form $\hat{\boldsymbol{\Sigma}} = a[\mathbf{S} + b\mathbf{I}]$ where a and b are regularizing parameters. One of the fundamental works is due to Haff (1980) who considered $b = g(\mathrm{Tr}\{\mathbf{S}^{-1}\})$ and an empirical Bayes approach but this approach triggered the highest number of studies, see e.g., Ledoit and Wolf (2004), Konno (2009), Chen et al. (2010), Coluccia (2015) and Ikeda et al. (2016) and references therein.

However, the first class of estimates considered by Stein was that based on the regularization of the Cholesky factor of $\mathbf{S}$, i.e., estimates of the form $\hat{\boldsymbol{\Sigma}} = \mathbf{G}_{\mathbf{S}}\mathrm{diag}(\mathbf{d})\mathbf{G}_{\mathbf{S}}^T$ where $\mathbf{G}_{\mathbf{S}}$ is the Cholesky factor of $\mathbf{S}$. Stein showed that, for the loss $\mathcal{L}_1(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})$, the corresponding risk $\mathcal{R}_1(\boldsymbol{\Sigma}, \mathbf{d}) = \mathbb{E}\{\mathcal{L}_1(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})\}$ is minimized when

$$[\mathbf{d}_1]_j = (p - j + \mathbb{E}\{\chi^2_{n-j+1}\})^{-1} \tag{1}$$

The optimal vector, say $\mathbf{d}_2$, which minimizes $\mathcal{R}_2(\boldsymbol{\Sigma}, \mathbf{d}) = \mathbb{E}\{\mathcal{L}_2(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})\}$ where $\mathcal{L}_2(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = \mathrm{Tr}\{(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \mathbf{I}_n)^2\}$ was derived by Selliah in Selliah (1964). In Tsukuma and Kubokawa (2016) extensions of these estimators to the case $p < n$ are given. Using estimates of the type $\hat{\boldsymbol{\Sigma}} = \mathbf{G}_{\mathbf{S}}\mathrm{diag}(\mathbf{d})\mathbf{G}_{\mathbf{S}}^T$ is interesting for some reasons. First, the Cholesky factor is easy to compute. Furthermore, it is of interest when used for whitening purposes: indeed, in order to whiten data, only a triangular system of equations needs to be solved. Whitening is particularly useful when detecting a signal of interest among noise with covariance matrix $\boldsymbol{\Sigma}$ since the optimal detection scheme involves whitening followed by matched filtering (Scharf, 1991). Now, since estimation of the Cholesky factor may be interested *per se*, it is natural to consider risk functions that are expressed in terms of the Cholesky factor. This is what was proposed by Eaton and Olkin (1987) who considered the two following loss functions

$$\mathcal{L}_3(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{G}_{\hat{\boldsymbol{\Sigma}}}) = \mathrm{Tr}\{(\mathbf{G}_{\boldsymbol{\Sigma}}^{-1}\mathbf{G}_{\hat{\boldsymbol{\Sigma}}} - \mathbf{I})(\mathbf{G}_{\boldsymbol{\Sigma}}^{-1}\mathbf{G}_{\hat{\boldsymbol{\Sigma}}} - \mathbf{I})^T\} \tag{2}$$

$$\mathcal{L}_4(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{G}_{\hat{\boldsymbol{\Sigma}}}) = \mathrm{Tr}\{(\mathbf{G}_{\hat{\boldsymbol{\Sigma}}}^{-1}\mathbf{G}_{\boldsymbol{\Sigma}} - \mathbf{I})(\mathbf{G}_{\hat{\boldsymbol{\Sigma}}}^{-1}\mathbf{G}_{\boldsymbol{\Sigma}} - \mathbf{I})^T\} \tag{3}$$

and associated risks $\mathcal{R}_3(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d}) = \mathbb{E}\{\mathcal{L}_3(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{G}_{\hat{\boldsymbol{\Sigma}}})\}$ and $\mathcal{R}_4(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d}) = \mathbb{E}\{\mathcal{L}_4(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{G}_{\hat{\boldsymbol{\Sigma}}})\}$. They showed that the optimal $\mathbf{d}$, which minimize these risks, are respectively

$$[\mathbf{d}_3]_j^{1/2} = \frac{\mathbb{E}\{\sqrt{\chi^2_{n-j+1}}\}}{p - j + \mathbb{E}\{\chi^2_{n-j+1}\}} \tag{4}$$

$$[\mathbf{d}_4]_j^{1/2} = \frac{n-1}{(n-j)(n-j-1)} \frac{1}{\mathbb{E}\{(\chi^2_{n-j+1})^{-1/2}\}} \tag{5}$$

In this letter, we investigate estimates of the form $\hat{\boldsymbol{\Sigma}} = \mathbf{G}_{\mathbf{S}}\mathrm{diag}(\mathbf{d})\mathbf{G}_{\mathbf{S}}^T$ and we focus on a loss function that depends on the Cholesky factor. More precisely, we consider a recently proposed distance in the set of lower triangular matrices with positive diagonal entries. We show that the optimal regularization vector depends on $\boldsymbol{\Sigma}$ and we propose a procedure to find an accurate approximation. Finally, we evaluate the four approaches mentioned above as well as our new method on a relevant metric which is related to the natural distance between covariance matrices.

Notations. The jth entry of a vector $\mathbf{d}$ is denoted d_j or $[\mathbf{d}]_j$ and we sometimes use $\mathbf{d} = \mathrm{vect}(d_j)$. The (i, j)-th entry of a $p \times p$ matrix $\mathbf{M}$ is either denoted by $\mathbf{M}_{ij}$ or $[\mathbf{M}]_{ij}$. We let $\mathrm{diag}(\mathbf{M})$ be the $p \times 1$ vector whose entries are $\mathbf{M}_{jj}$. Conversely, for any vector $\mathbf{d}$, $\mathrm{diag}(\mathbf{d})$ is a diagonal matrix whose diagonal entries are d_j. We will sometimes note $\mathrm{diag}(d_j)$. $\mathrm{ddiag}(\mathbf{M})$ is defined as $\mathrm{ddiag}(\mathbf{M}) = \mathrm{diag}(\mathrm{diag}(\mathbf{M}))$. $\odot$ is the Hadamard product, i.e., element-wise product. The Cholesky factor of matrix $\boldsymbol{\Sigma}$ will be denoted $\mathbf{G}_{\boldsymbol{\Sigma}}$, i.e., $\mathbf{G}_{\boldsymbol{\Sigma}}$ is lower-triangular with positive diagonal entries and $\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\boldsymbol{\Sigma}}^T = \boldsymbol{\Sigma}$. $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes the normal distribution with mean μ and variance σ^2. χ^2_q stands for the chi-square distribution with q degrees of freedom and $\mathcal{W}_p\left(n, \boldsymbol{\Sigma}\right)$ stands for the Wishart distribution with n degrees of freedom and parameter matrix $\boldsymbol{\Sigma}$. $\overset{d}{=}$ means "is distributed as".

In the paper, we will need the following result on some statistics associated with $\mathbf{G}_{\bar{\mathbf{W}}}$ where $\bar{\mathbf{W}} \overset{d}{=} \mathcal{W}_p\left(n, \mathbf{I}\right)$. This result stems from the fact that all entries of $\mathbf{G}_{\bar{\mathbf{W}}}$ are independent with $[\mathbf{G}_{\bar{\mathbf{W}}}]_{ij} \overset{d}{=} \mathcal{N}\left(0, 1\right)$ for $i > j$ and $[\mathbf{G}_{\bar{\mathbf{W}}}]_{jj} \overset{d}{=} \sqrt{\chi^2_{n-j+1}}$ (Muirhead, 1982; Gupta and Nagar, 2000).

Result 1. *For any matrix* $\mathbf{M}$, $\mathbb{E}\{\mathbf{M}\mathbf{G}_{\bar{\mathbf{W}}}\} = \mathbf{M}\mathrm{diag}(\mathbb{E}\{\sqrt{\chi^2_{n-j+1}}\})$ *which implies that* $\mathbb{E}\{\left[\mathbf{M}\mathbf{G}_{\bar{\mathbf{W}}}\right]_{jj}\} = [\mathbf{M}]_{jj}\mathbb{E}\{\sqrt{\chi^2_{n-j+1}}\}$. *Additionally* $\mathbb{E}\{\mathbf{G}_{\bar{\mathbf{W}}}^T\mathbf{M}\mathbf{G}_{\bar{\mathbf{W}}}\}$ *is a diagonal matrix given by*

$$\left[\mathbb{E}\{\mathbf{G}_{\bar{\mathbf{W}}}^T\mathbf{M}\mathbf{G}_{\bar{\mathbf{W}}}\}\right]_{jj} = \mathbb{E}\{\chi^2_{n-j+1}\}\mathbf{M}_{jj} + \sum_{i=j+1}^{p} \mathbf{M}_{ii} \tag{6}$$

2. Minimization of risk function associated with log-Cholesky metric

As stated above, we consider here a new Riemannian metric, termed log-Cholesky metric, defined on the set of lower triangular matrices with positive diagonal entries as (Lin, 2019):

$$\begin{aligned}\mathcal{L}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{G}_{\hat{\boldsymbol{\Sigma}}}) &= \sum_{i>j} \left([\mathbf{G}_{\hat{\boldsymbol{\Sigma}}}]_{ij} - [\mathbf{G}_{\boldsymbol{\Sigma}}]_{ij}\right)^2 + \sum_{j=1}^{p} \left(\log[\mathbf{G}_{\hat{\boldsymbol{\Sigma}}}]_{jj} - \log[\mathbf{G}_{\boldsymbol{\Sigma}}]_{jj}\right)^2 \\ &= \left\|\mathbf{G}_{\hat{\boldsymbol{\Sigma}}} - \mathbf{G}_{\boldsymbol{\Sigma}}\right\|_F^2 - \left\|\mathrm{diag}(\mathbf{G}_{\hat{\boldsymbol{\Sigma}}}) - \mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}})\right\|^2 \\ &+ \left\|\log\mathrm{diag}(\mathbf{G}_{\hat{\boldsymbol{\Sigma}}}) - \log\mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}})\right\|^2 \end{aligned} \tag{7}$$

Since $\mathbf{G}_{\hat{\boldsymbol{\Sigma}}} = \mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}}\mathbf{D}^{1/2}$, one has

$$\begin{aligned}\left\|\mathbf{G}_{\hat{\boldsymbol{\Sigma}}} - \mathbf{G}_{\boldsymbol{\Sigma}}\right\|_F^2 &= \mathrm{Tr}\{(\mathbf{G}_{\hat{\boldsymbol{\Sigma}}} - \mathbf{G}_{\boldsymbol{\Sigma}})(\mathbf{G}_{\hat{\boldsymbol{\Sigma}}} - \mathbf{G}_{\boldsymbol{\Sigma}})^T\} \\ &= \mathrm{Tr}\{\mathbf{G}_{\boldsymbol{\Sigma}}(\mathbf{G}_{\bar{\mathbf{W}}}\mathbf{D}^{1/2} - \mathbf{I})(\mathbf{G}_{\bar{\mathbf{W}}}\mathbf{D}^{1/2} - \mathbf{I})^T\mathbf{G}_{\boldsymbol{\Sigma}}^T\} \\ &= \mathrm{Tr}\{\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}}\mathbf{D}\mathbf{G}_{\bar{\mathbf{W}}}^T\mathbf{G}_{\boldsymbol{\Sigma}}^T\} - 2\mathrm{Tr}\{\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}}\mathbf{D}^{1/2}\mathbf{G}_{\boldsymbol{\Sigma}}^T\} + \mathrm{Tr}\{\boldsymbol{\Sigma}\} \\ &= \sqrt{\mathbf{d}}^T \mathrm{ddiag}(\mathbf{G}_{\bar{\mathbf{W}}}^T\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}})\sqrt{\mathbf{d}} - 2\sqrt{\mathbf{d}}^T \mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}}) + \mathrm{Tr}\{\boldsymbol{\Sigma}\}\end{aligned} \tag{8}$$

Moreover, since $[\mathbf{G}_{\hat{\boldsymbol{\Sigma}}}]_{jj} = d_j^{1/2}[\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}}]_{jj}$, one has

$$\begin{aligned}&\left\|\mathrm{diag}(\mathbf{G}_{\hat{\boldsymbol{\Sigma}}}) - \mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}})\right\|^2 = \sum_{j=1}^{p}(d_j^{1/2}[\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}}]_{jj} - [\mathbf{G}_{\boldsymbol{\Sigma}}]_{jj})^2 \\ &= \sqrt{\mathbf{d}}^T \mathrm{ddiag}(\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}} \odot \mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}})\sqrt{\mathbf{d}} - 2\sqrt{\mathbf{d}}^T \mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}} \odot \mathbf{G}_{\boldsymbol{\Sigma}}) + \|\mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}})\|^2\end{aligned} \tag{9}$$

and

$$\begin{aligned}&\left\|\log\mathrm{diag}(\mathbf{G}_{\hat{\boldsymbol{\Sigma}}}) - \log\mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}})\right\|^2 = \sum_{j=1}^{p}(\log d_j^{1/2} + \log[\mathbf{G}_{\bar{\mathbf{W}}}]_{jj})^2 \\ &= (\log\sqrt{\mathbf{d}})^T(\log\sqrt{\mathbf{d}}) + 2(\log\sqrt{\mathbf{d}})^T \log\mathrm{diag}(\mathbf{G}_{\bar{\mathbf{W}}}) + \left\|\log\mathrm{diag}(\mathbf{G}_{\bar{\mathbf{W}}})\right\|^2\end{aligned} \tag{10}$$

Gathering the previous equations yields the following expression for $\mathcal{L}_5(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})$:

$$\begin{aligned}\mathcal{L}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{G}_{\hat{\boldsymbol{\Sigma}}}) &= \sqrt{\mathbf{d}}^T \left[\mathrm{ddiag}(\mathbf{G}_{\bar{\mathbf{W}}}^T\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}}) - \mathrm{ddiag}(\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}} \odot \mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}})\right]\sqrt{\mathbf{d}} \\ &- 2\sqrt{\mathbf{d}}^T \left[\mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}}) - \mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}} \odot \mathbf{G}_{\boldsymbol{\Sigma}})\right] \\ &+ (\log\sqrt{\mathbf{d}})^T(\log\sqrt{\mathbf{d}}) + 2(\log\sqrt{\mathbf{d}})^T \log\mathrm{diag}(\mathbf{G}_{\bar{\mathbf{W}}}) \\ &+ \mathrm{Tr}\{\boldsymbol{\Sigma}\} - \|\mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}})\|^2 + \left\|\log\mathrm{diag}(\mathbf{G}_{\bar{\mathbf{W}}})\right\|^2\end{aligned} \tag{11}$$

It ensues that the corresponding risk is given by

$$\begin{aligned}\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d}) &= \mathbb{E}\{\mathcal{L}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{G}_{\hat{\boldsymbol{\Sigma}}})\} \\ &= \sqrt{\mathbf{d}}^T \mathrm{diag}(\mathbf{a})\sqrt{\mathbf{d}} - 2\sqrt{\mathbf{d}}^T\mathbf{b} + (\log\sqrt{\mathbf{d}})^T(\log\sqrt{\mathbf{d}}) + 2(\log\sqrt{\mathbf{d}})^T\mathbf{c} \\ &+ \mathrm{Tr}\{\boldsymbol{\Sigma}\} - \|\mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}})\|^2 + \mathbb{E}\{\left\|\log\mathrm{diag}(\mathbf{G}_{\bar{\mathbf{W}}})\right\|^2\}\end{aligned} \tag{12}$$

with

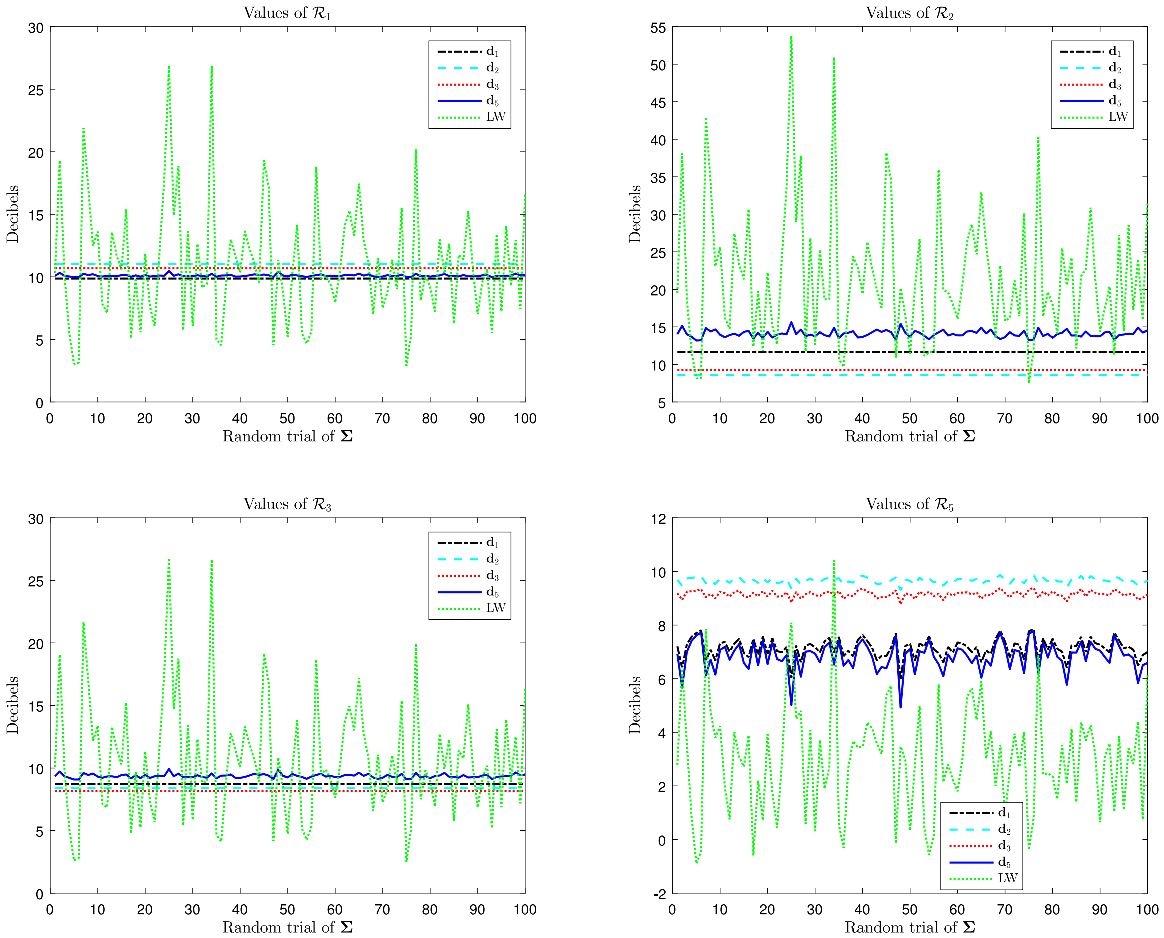
$$\begin{aligned}\mathbf{a} &= \mathbb{E}\{\mathrm{diag}(\mathbf{G}_{\bar{\mathbf{W}}}^T\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}})\} - \mathbb{E}\{\mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}} \odot \mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}})\} \\ \mathbf{b} &= \mathbb{E}\{\mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}})\} - \mathbb{E}\{\mathrm{diag}(\mathbf{G}_{\boldsymbol{\Sigma}}\mathbf{G}_{\bar{\mathbf{W}}} \odot \mathbf{G}_{\boldsymbol{\Sigma}})\} \\ \mathbf{c} &= \mathbb{E}\{\log\mathrm{diag}(\mathbf{G}_{\bar{\mathbf{W}}})\}\end{aligned} \tag{13}$$

The jth elements of these vectors are given by

$$\begin{aligned}a_j &= \mathbb{E}\{\chi^2_{n-j+1}\}\left([\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}]_{jj} - [\mathbf{G}_{\boldsymbol{\Sigma}}]_{jj}^2\right) + \sum_{i=j+1}^{p}[\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}]_{ii} \\ b_j &= \mathbb{E}\{\sqrt{\chi^2_{n-j+1}}\}\left([\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}]_{jj} - [\mathbf{G}_{\boldsymbol{\Sigma}}]_{jj}^2\right) \\ c_j &= \mathbb{E}\{\log\sqrt{\chi^2_{n-j+1}}\}\end{aligned} \tag{14}$$

Table 1
Average value of $\text{RRI}_5(\boldsymbol{\Sigma}, \mathbf{d}) = [\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d}) - \mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d}_5(\boldsymbol{\Sigma}))]/\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d}_5(\boldsymbol{\Sigma}))$.

p, n	$\mathbf{d} = \mathbf{d}_5(\mathbf{I}_p)$	$\mathbf{d} = \hat{\mathbf{d}}_5$
$p = 16$, $n = p + 2$	13.32%	0.56%
$p = 16$, $n = 2p$	11.08%	0.09%
$p = 64$, $n = p + 2$	17.38%	0.18%
$p = 64$, $n = 2p$	12.45%	0.03%


Values of $\mathcal{R}_1$

Values of $\mathcal{R}_2$

Values of $\mathcal{R}_3$

Values of $\mathcal{R}_5$

Fig. 1. Risks of the various estimators over 100 random trials of $\boldsymbol{\Sigma}$. $p = 16$ and $n = p + 2$.

It can be observed that $\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d})$ depends on $\mathbf{G}_{\boldsymbol{\Sigma}}$. Note also that, in order to minimize this risk, we need to minimize over $\mathbb{R}^+$ functions of the form

$$f_j(x) = a_j x^2 - 2b_j x + \log^2 x + 2c_j \log x$$

We have

$$\begin{aligned} f_j'(x) &= 2a_j x - 2b_j + 2x^{-1} \log x + 2c_j x^{-1} \\ &= 2x^{-1}[a_j x^2 - b_j x + \log x + c_j] \\ &= 2x^{-1} g_j(x) \end{aligned} \tag{15}$$

Differentiating $g_j(x)$ yields

$$g_j'(x) = x^{-1}(2a_j x^2 - b_j x + 1)$$

It can be readily shown that $g_j'(x) > 0$ given the values of a_j and b_j, which implies that $g_j(x)$ is monotonically increasing from $-\infty$ (when $x \to 0$) to $+\infty$ (when $x \to +\infty$). Therefore, there exists a unique $x_j^\star$ for which $g_j(x_j^\star) = 0$ and hence

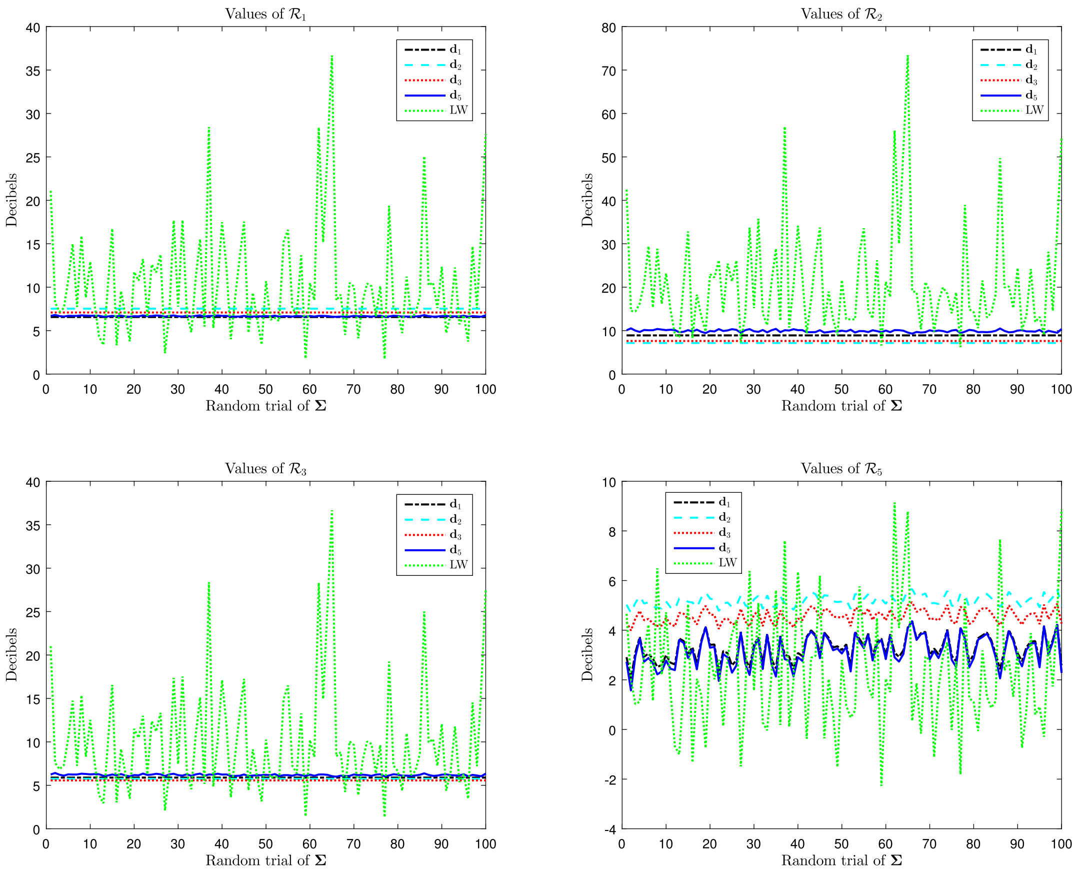


Fig. 2. Risks of the various estimators over 100 random trials of $\boldsymbol{\Sigma}$. $p = 16$ and $n = 2p$.

$f_j'(x_j^\star) = 0$. Since $g_j(x) < 0$ for $x < x_j^\star$, it follows that $f_j(x)$ achieves its unique minimum at $x_j^\star$. Therefore, the risk is minimized for $d_j^{1/2} = x_j^\star$. We let $\mathbf{d}_5(\boldsymbol{\Sigma})$ be the optimal vector for this log-Cholesky distance, where we emphasize that this vector depends on $\boldsymbol{\Sigma}$.

Since the optimal regularizing matrix depends on $\boldsymbol{\Sigma}$ which is unknown, the usual procedure is to resort to Stein unbiased risk estimation (SURE), that is to find a function $\hat{\mathcal{L}}_5(\mathbf{d}, \mathbf{S})$ such that $\mathbb{E}\{\hat{\mathcal{L}}_5(\mathbf{d}, \mathbf{S})\} = \mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d})$, or at least such that $\mathbb{E}\{\hat{\mathcal{L}}_5(\mathbf{d}, \mathbf{S})\}$ coincides with the part of $\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d})$ that depends on $\mathbf{d}$, and to minimize $\hat{\mathcal{L}}_5(\mathbf{d}, \mathbf{S})$. In our case, this amounts to obtain unbiased estimates of $\mathbf{a}$ and $\mathbf{b}$. Since one has

$$\begin{aligned}\mathbb{E}\{[\mathbf{G}_\mathbf{S}^T\mathbf{G}_\mathbf{S}]_{jj}\} &= \mathbb{E}\{\chi^2_{n-j+1}\}[\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}]_{jj} + \sum_{i=j+1}^{p}[\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}]_{ii}\\ \mathbb{E}\{[\mathbf{G}_\mathbf{S}]_{jj}^2\} &= \mathbb{E}\{\chi^2_{n-j+1}\}[\mathbf{G}_{\boldsymbol{\Sigma}}]_{jj}^2\end{aligned}$$

it is clear that $\hat{a}_j = [\mathbf{G}_\mathbf{S}^T\mathbf{G}_\mathbf{S}]_{jj} - [\mathbf{G}_\mathbf{S}]_{jj}^2$ is an unbiased estimate of a_j. However, finding an unbiased estimate of b_j turns out to be more problematic because an unbiased estimate of $[\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}]_{jj}$ does not seem feasible to obtain because of the term $\sum_{i=j+1}^{p}[\mathbf{G}_{\boldsymbol{\Sigma}}^T\mathbf{G}_{\boldsymbol{\Sigma}}]_{ii}$ in $\mathbb{E}\{[\mathbf{G}_\mathbf{S}^T\mathbf{G}_\mathbf{S}]_{jj}\}$. Moreover, minimizing an estimate $\hat{\mathcal{L}}_5(\mathbf{d}, \mathbf{S})$ instead of $\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d})$ unavoidably leads to performance loss. Therefore, we investigate alternative solutions here.

A first straightforward method comes from the observation that, for all other losses $\mathcal{R}_i(\boldsymbol{\Sigma}, \mathbf{d})$, $i = 1, \ldots, 4$, the optimal vector $\mathbf{d}_i(\boldsymbol{\Sigma})$ does not depend on $\boldsymbol{\Sigma}$ and can be computed as $\mathbf{d}_i(\mathbf{I}_p)$. Therefore, one can choose $\mathbf{d}_5(\mathbf{I}_p)$ as the regularization vector. This solution is very simple, yet one needs to study how far is the risk associated with $\mathbf{d}_5(\mathbf{I}_p)$ from the risk obtained with the optimal solution $\mathbf{d}_5(\boldsymbol{\Sigma})$. In case the risk increase is not very important, using $\mathbf{d}_5(\mathbf{I}_p)$ is much simpler than an approximate SURE approach and may perform as well.

However, one can anticipate some loss of performance of $\mathbf{d}_5(\mathbf{I}_p)$ compared to $\mathbf{d}_5(\boldsymbol{\Sigma})$. To remedy this problem, we use the fact that $\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d})$ and its minimizer depend on $\boldsymbol{\Sigma}$ and we suggest an alternative approach which consists in finding $\mathbf{d}$ as the minimizer of $\mathcal{R}_5(\mathbf{G}_{\hat{\boldsymbol{\Sigma}}}, \mathbf{d})$ where $\hat{\boldsymbol{\Sigma}}$ is some estimate of $\boldsymbol{\Sigma}$. For instance, one could use $\hat{\boldsymbol{\Sigma}} = n^{-1}\mathbf{S}$. Then, one obtains $\hat{\mathbf{d}} = \arg\min_{\mathbf{d}} \mathcal{R}_5(\mathbf{G}_{\hat{\boldsymbol{\Sigma}}}, \mathbf{d})$, which provides the estimate $\mathbf{G}_\mathbf{S}\mathrm{diag}(\hat{\mathbf{d}})\mathbf{G}_\mathbf{S}^T$. In fact, the process can be iterated as follows. Let

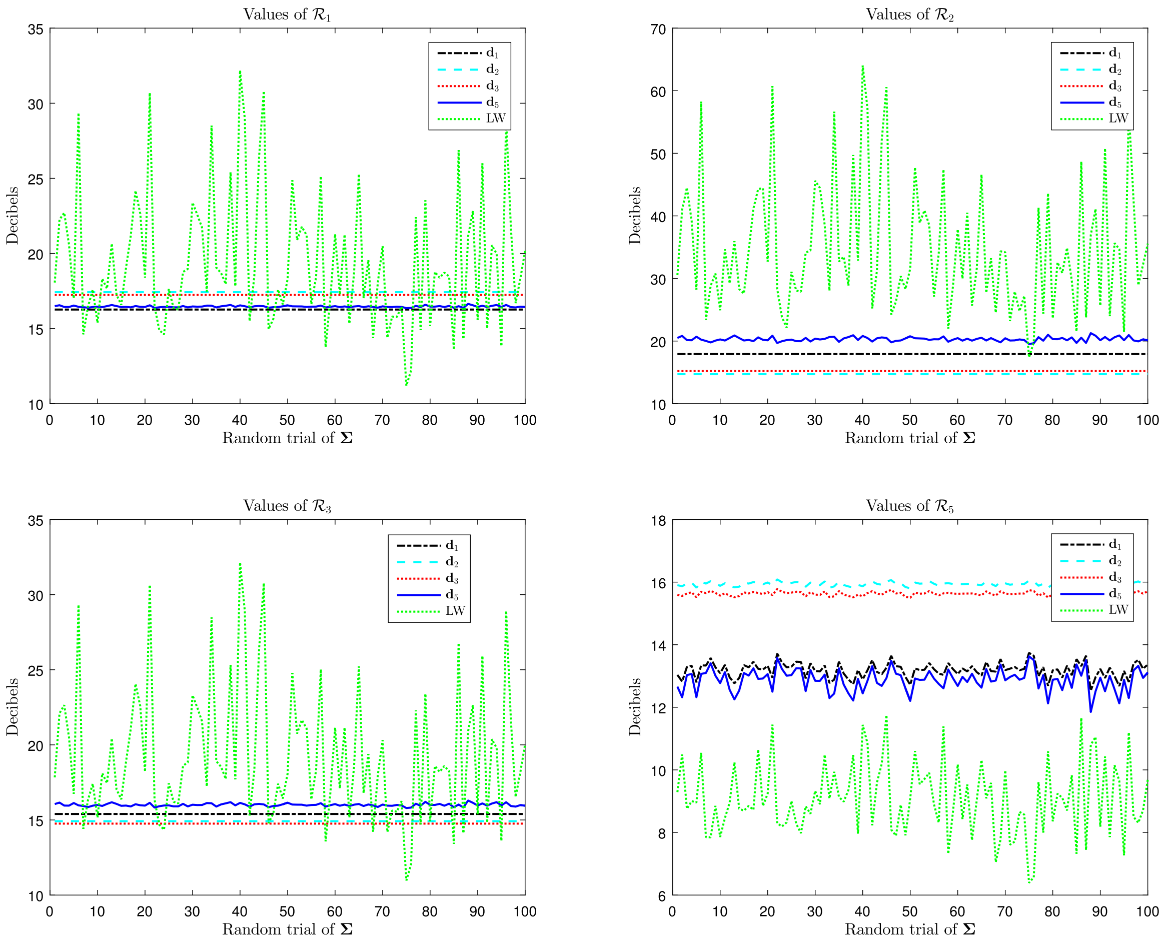


Fig. 3. Risks of the various estimators over 100 random trials of $\boldsymbol{\Sigma}$. $p = 64$ and $n = p + 2$.

$\hat{\mathbf{d}}_5^{(0)}$ be some initial vector, for instance $\hat{\mathbf{d}}_5^{(0)} = \mathbf{d}_5(\mathbf{I}_p)$ or $\hat{\mathbf{d}}_5^{(0)} = n^{-1/2}\mathbf{1}_p$ where $\mathbf{1}_p$ is a length-p vector whose elements are all equal to one, and $\hat{\boldsymbol{\Sigma}}^{(0)} = \mathbf{G}_\mathbf{S}\text{diag}(\hat{\mathbf{d}}_5^{(0)})\mathbf{G}_\mathbf{S}^T$. Then, for $n = 1, \ldots,$ one can compute $\hat{\mathbf{d}}_5^{(n)} = \arg\min_\mathbf{d} \mathcal{R}_5(\mathbf{G}_{\hat{\boldsymbol{\Sigma}}^{(n-1)}}, \mathbf{d})$ and $\hat{\boldsymbol{\Sigma}}^{(n)} = \mathbf{G}_\mathbf{S}\text{diag}(\hat{\mathbf{d}}_5^{(n)})\mathbf{G}_\mathbf{S}^T$. We let $\hat{\mathbf{d}}_5$ be the vector at the end of the iterations. It is our experience that these iterations converge rather fast and that, typically, 5 iterations are sufficient.

In order to evaluate the difference between $\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d}_5(\boldsymbol{\Sigma}))$, $\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d}_5(\mathbf{I}))$ and $\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \hat{\mathbf{d}}_5)$ we conducted the following experiment. A large number of matrices $\boldsymbol{\Sigma}$ were drawn at random as $\boldsymbol{\Sigma} = \mathbf{U}\text{diag}(\boldsymbol{\lambda})\mathbf{U}^T$ where $\mathbf{U}$ is uniformly distributed over the set of unitary matrices, and λ_j are independent random variables drawn uniformly on $]0, 1]$. For each matrix $\boldsymbol{\Sigma}$, the optimal $\mathbf{d}_5(\boldsymbol{\Sigma})$ was computed along with the corresponding risk. The relative risk increase $\text{RRI}_5(\boldsymbol{\Sigma}, \mathbf{d}) = [\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d}) - \mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d}_5(\boldsymbol{\Sigma}))]/\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d}_5(\boldsymbol{\Sigma}))$ was evaluated for $\mathbf{d} = \mathbf{d}_5(\mathbf{I}_p)$ and $\mathbf{d} = \hat{\mathbf{d}}_5$, and then averaged over the 10^3 experiments. For $\hat{\mathbf{d}}_5$, the iterative scheme was initialized with $\mathbf{d}_5(\mathbf{I}_p)$ and 5 iterations were used. The results are given in Table 1. It can be observed that the risk increase incurred when using $\mathbf{d}_5(\mathbf{I})$ instead of $\mathbf{d}_5(\boldsymbol{\Sigma})$ is about $10 - 13\%$, which is acceptable. However, the iterative scheme is seen to perform very well and incurs almost no loss compared to the optimal $\mathbf{d}_5(\boldsymbol{\Sigma})$, which makes it a rather optimal solution to minimize $\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d})$.

3. Numerical illustrations

In this section we first evaluate the performance of each vector $\mathbf{d}_q$ not only for $\mathcal{R}_q(\boldsymbol{\Sigma}, \mathbf{d})$ – for which it is optimal – but also for all other risks, with a view to figure out the performance of $\mathbf{d}_q$ over a wider range of losses. Through preliminary simulations it appeared that $\mathbf{d}_4$ provided very poor performance and thus it is not considered in the sequel, nor the corresponding risk. Furthermore, since all remaining $\mathbf{d}$ perform better than the SCM, the latter is not shown in the figures below. On the other hand, we compare the above schemes, based on regularization of the Cholesky factor of the SCM, to a reference method, namely the Ledoit–Wolf (LW) estimator (Ledoit and Wolf, 2004) which corresponds to shrinkage of the SCM. Similarly to the previous section, we draw a large number of matrices $\boldsymbol{\Sigma} = \mathbf{U}\text{diag}(\boldsymbol{\lambda})\mathbf{U}^T$ where $\mathbf{U}$

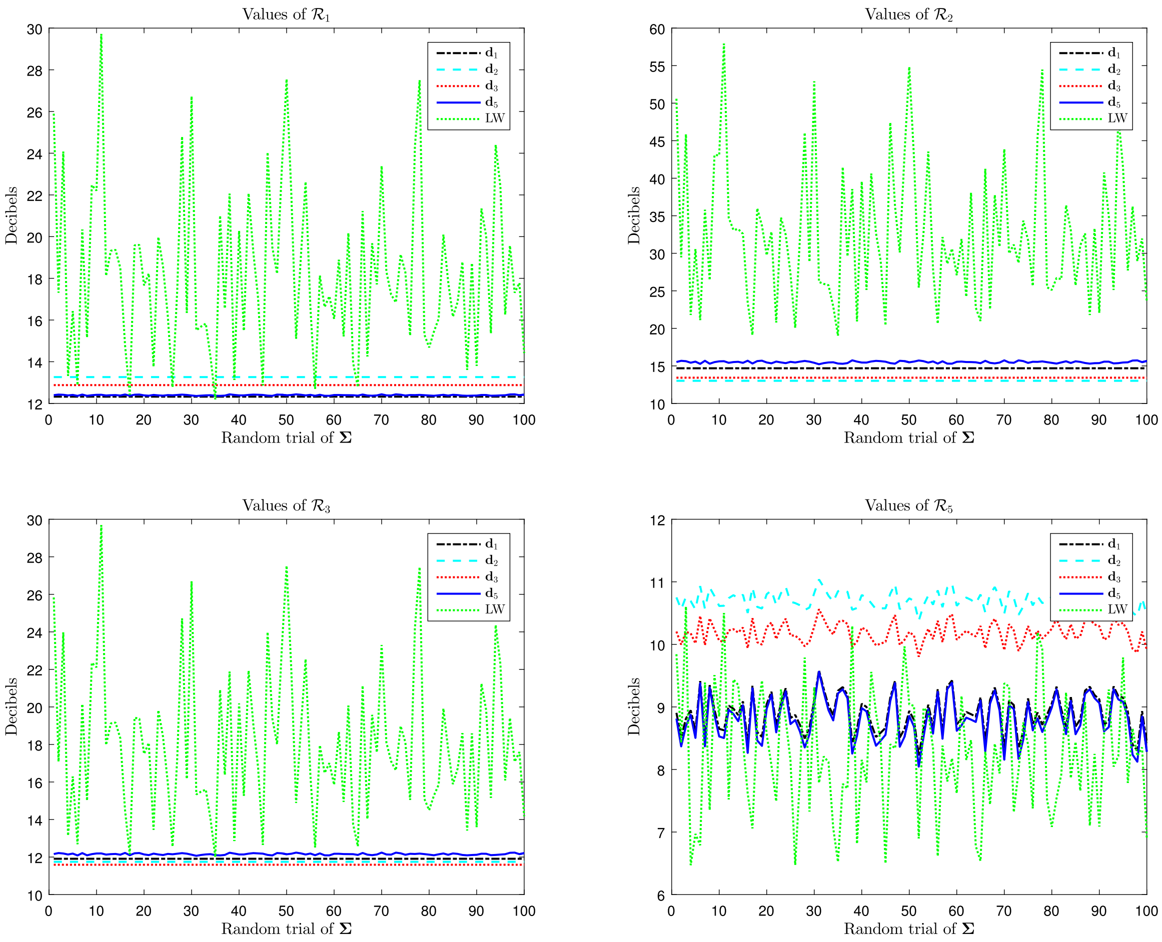


Fig. 4. Risks of the various estimators over 100 random trials of $\boldsymbol{\Sigma}$. $p = 64$ and $n = 2p$.

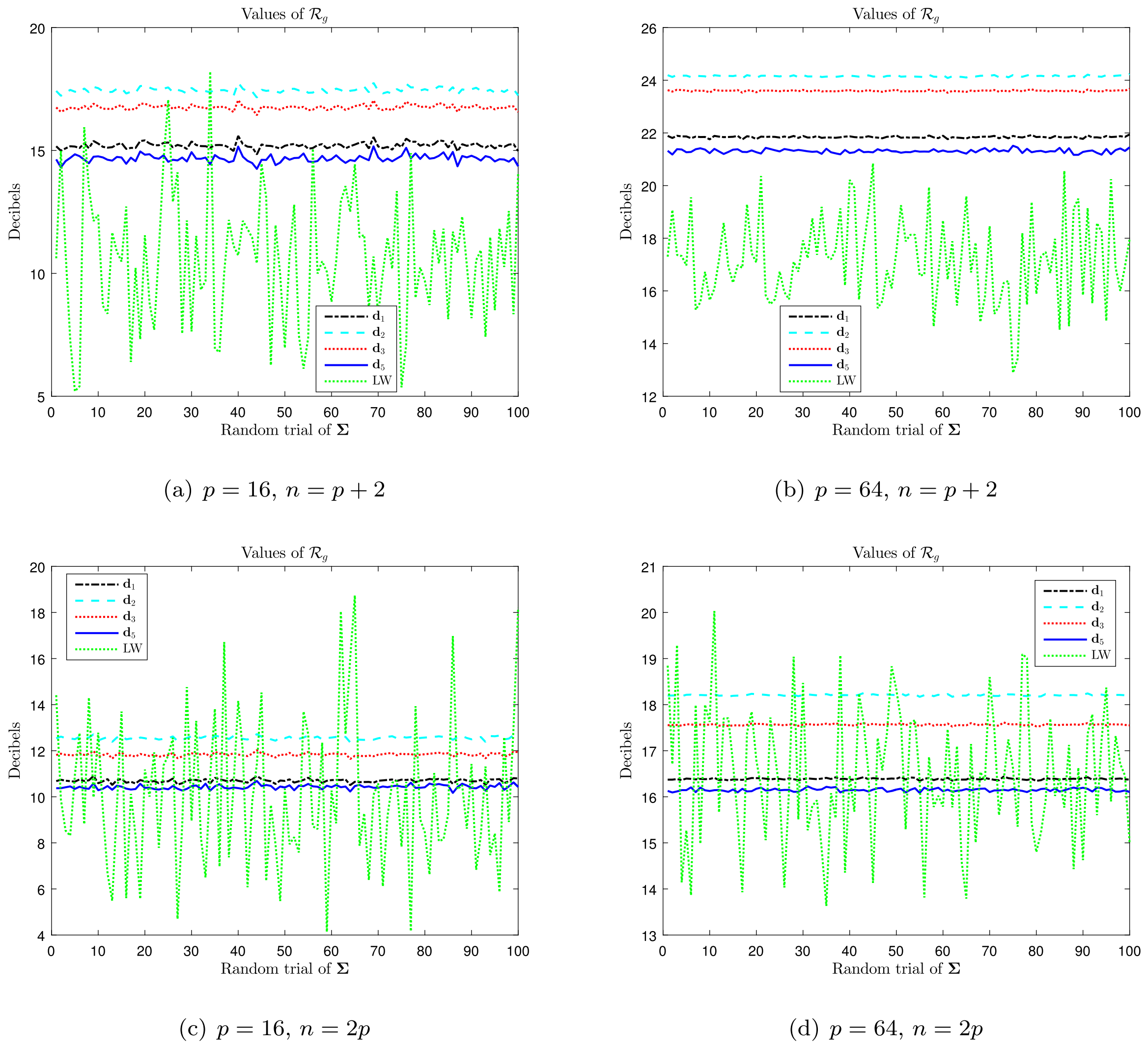
is uniformly distributed over the set of unitary matrices, and λ_j are independent random variables drawn uniformly on $]0, 1]$. For each $\boldsymbol{\Sigma}$, the risks $\mathcal{R}_q(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})$ $(q = 1 - 3, 5)$ are evaluated for vectors $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_3$ and $\hat{\mathbf{d}}_5$, and for the LW estimator.

The results are reported in Figs. 1–2 for $p = 16$ and in Figs. 3–4 for $p = 64$. These curves show rather interesting results. First note that $\mathbf{d}_1$ performs well for $\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d})$ and, vice versa, $\hat{\mathbf{d}}_5$ is very good for $\mathcal{R}_1(\boldsymbol{\Sigma}, \mathbf{d})$. Additionally both of them are quite accurate with respect to $\mathcal{R}_3(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d})$. Therefore, it seems that the new estimator $\hat{\mathbf{d}}_5$ bears some resemblance with Stein's initial method. A similarity is also observed between $\mathbf{d}_2$ and $\mathbf{d}_3$ which performs well only on $\mathcal{R}_2(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d})$ and $\mathcal{R}_3(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d})$. As for the LW estimator, it performs very poorly on $\mathcal{R}_2(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})$, poorly on $\mathcal{R}_1(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})$ and $\mathcal{R}_3(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})$ but is the best method for $\mathcal{R}_5(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})$, at least when $n = p + 2$. For $n = 2p$, $\mathbf{d}_1$ and $\hat{\mathbf{d}}_5$ achieve the same risks $\mathcal{R}_1(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})$ and $\mathcal{R}_5(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})$ as the LW estimator. However a striking fact is that *the risks associated with the LW estimator are highly variable* when $\boldsymbol{\Sigma}$ varies. In contrast, the estimators based on $\mathbf{d}_{1-3}$ have constant risks $\mathcal{R}_{1-3}(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d})$ and their risk $\mathcal{R}_5(\mathbf{G}_{\boldsymbol{\Sigma}}, \mathbf{d}_{1-3})$ varies very weakly. Similarly, $\hat{\mathbf{d}}_5$ offers weakly varying risks $\mathcal{R}_{1-3,5}(\mathbf{G}_{\boldsymbol{\Sigma}}, \hat{\mathbf{d}}_5)$. Therefore, while the LW has often a lower risk $\mathcal{R}_{1,5}(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})$, it exhibits some instability as the values of the risks are highly dependent on $\boldsymbol{\Sigma}$, which is a drawback in practice.

To close this section, we now evaluate the respective merits of the above estimates on another loss function, which is highly relevant for the case of interest where covariance matrices are concerned, since it is the (square of the) natural distance between covariance matrices defined as

$$\mathcal{L}_g(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = \sum_{k=1}^{p} \log^2 \lambda_k(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}) \tag{16}$$

where the λ_ks are the generalized eigenvalues of $(\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma})$. The corresponding risk is defined as $\mathcal{R}_g(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = \mathbb{E}\{\mathcal{L}_g(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})\}$. The loss in (16) corresponds to the natural distance between $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ in the set of positive definite matrices (Bhatia, 2007). It can also be derived as the geodesic distance induced by Rao's metric which measures dissimilarity between two



(a) $p = 16$, $n = p + 2$

(b) $p = 64$, $n = p + 2$

(c) $p = 16$, $n = 2p$

(d) $p = 64$, $n = 2p$

Fig. 5. Risk corresponding to the natural distance of the various estimators over 100 random trials of $\mathbf{\Sigma}$. $p = 16$ (left panel) and $p = 64$ (right panel).

zero-mean Gaussian distributions with different covariance matrices (Amari et al., 1987; Atkinson and Mitchell, 1981). It is thus a very relevant metric and there is interest in comparing $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_3$ and $\hat{\mathbf{d}}_5$ on this third-party, meaningful criterion. As before, the risks were computed for 100 different $\mathbf{\Sigma}$. The results are given in Fig. 5. They show that the LW estimator offers the smallest risk, followed by $\hat{\mathbf{d}}_5$, which is the best among estimators based on regularization of the Cholesky factor. Note that the difference between $\hat{\mathbf{d}}_5$ and LW is rather small for $n = 2p$. Therefore, given that $\hat{\mathbf{d}}_5$ results in much smaller risk variability, it constitutes an interesting trade-off.

4. Conclusions

In this paper, we considered estimation of a covariance matrix using Stein's approach. We focused on estimates of the form $\hat{\mathbf{\Sigma}} = \mathbf{G_S}\text{diag}(\mathbf{d})\mathbf{G_S}^T$ where $\mathbf{G_S}$ is the Cholesky factor of the sample covariance matrix. This problem was addressed by Stein, Selliah, Eaton and Olkin for various loss functions. We extended this kind of approach to a recently introduced Riemannian metric on the set of lower triangular matrices with positive diagonal elements. The optimal regularization vector was shown to depend on $\mathbf{\Sigma}$ but we proposed an iterative scheme that incurs a very small loss and offers a good trade-off over various loss functions, including the natural distance between covariance matrices. Moreover, the risks associated with this new estimator are very weakly dependent on $\mathbf{\Sigma}$.

CRediT authorship contribution statement

Olivier Besson: Conceptualization, Methodology, Validation, Software, Writing - original draft, Writing - review & editing. **François Vincent:** Conceptualization, Methodology, Validation, Writing - original draft, Writing - review & editing. **Xavier Gendre:** Conceptualization, Methodology, Validation, Writing - original draft, Writing - review & editing.

References

Amari, S.-I., Barndorff-Nielsen, O.E., Kass, R.E., Lauritzen, S.L., Rao, C.R., 1987. Chapter 5: Differential metrics in probability spaces. In: Gupta, S.S. (Ed.), Differential Geometry in Statistical Inference. In: Lecture Notes–Monograph Series, vol. 10, Institute of Mathematical Statistics, Hayward, CA, pp. 217–240.

Atkinson, C., Mitchell, F.S., 1981. Rao's distance measure. Sankhya 43 (3), 345–365.

Bhatia, R., 2007. Positive Definite Matrices. Princeton University Press.

Chen, Y., Wiesel, A., Eldar, Y.C., Hero, A.O., 2010. Shrinkage algorithms for MMSE covariance estimation. IEEE Trans. Signal Process. 58 (10), 5016–5029.

Coluccia, A., 2015. Regularized covariance matrix estimation via empirical Bayes. IEEE Signal Process. Lett. 22 (11), 2127–2131.

Dey, D.K., Srinivasan, C., 1985. Estimation of a covariance matrix under Stein's loss. Ann. Statist. 13 (4), 1581–1591.

Dey, D.K., Srinivasan, C., 1986. Trimmed minimax estimator of a covariance matrix. Ann. Inst. Statist. Math. 38, 101–108.

Eaton, M.L., Olkin, I., 1987. Best equivariant estimators of a Cholesky decomposition. Ann. Statist. 15 (4), 1639–1650.

Gupta, A.K., Nagar, D.K., 2000. Matrix Variate Distributions. Chapman & Hall/CRC, Boca Raton, FL.

Haff, L.R., 1979. An identity for the Wishart distribution with applications. J. Multivariate Anal. 9, 531–544.

Haff, L.R., 1980. Empirical Bayes estimation of the multivariate normal covariance matrix. Ann. Statist. 8 (3), 586–597.

Ikeda, Y., Kubokawa, T., Srivastava, M.S., 2016. Comparison of linear shrinkage estimators of a large covariance matrix in normal and non-normal distributions. Comput. Statist. Data Anal. 95, 95–108.

James, W., Stein, C., 1992. Estimation with quadratic loss. In: Kotz, S., Johnson, N. (Eds.), Breakthroughs in Statistics. In: Springer Series in Statistics (Perspectives in Statistics), Springer, pp. 443–460.

Konno, Y., 2009. Shrinkage estimators for large covariance matrices in multivariate real and complex normal distributions under an invariant quadratic loss. J. Multivariate Anal. 100 (10), 2237–2253.

Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivariate Anal. 88 (2), 365–411.

Ledoit, O., Wolf, M., 2018. Optimal estimation of a large dimensional covariance matrix under Stein's loss. Bernoulli 24 (4B), 3791–3832.

Lin, Z., 2019. Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. SIAM J. Matrix Anal. Appl. 40 (4), 1353–1370.

Lin, S., Perlman, M., 1985. A Monte Carlo comparison of four estimators of a covariance matrix. In: Krishnaiah, P.R. (Ed.), Multivariate Analysis VI. North Holland, Amsterdam, pp. 411–429.

Ma, T., Jia, L., Su, Y., 2012. A new estimator of covariance matrix. J. Statist. Plann. Inference 142 (2), 529–536.

Muirhead, R.J., 1982. Aspects of Multivariate Statistical Theory. John Wiley & Sons, Hoboken, NJ.

Perron, F., 1992. Minimax estimators of a covariance matrix. J. Multivariate Anal. 43 (1), 16–28.

Pourahmadi, M., 2013. High Dimensional Covariance Estimation. In: Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ.

Scharf, L.L., 1991. Statistical Signal Processing: Detection, Estimation and Time Series Analysis. Addison Wesley, Reading, MA.

Selliah, J.B., 1964. Estimation and Testing Problems in a Wishart Distribution. Technical report no. 10, Department of Statistics, Stanford University.

Sheena, Y., 1995. Unbiased estimator of risk for an orthogonally invariant estimator of covariance matrix. J. Japan Statist. Soc. 25 (1), 35–48.

Sheena, Y., Takemura, A., 1992. Inadmissibility of non-order preserving orthogonally invariant estimators of the covariance matrix in the case of Stein's loss. J. Multivariate Anal. 41, 117–131.

Srivastava, M.S., 2002. Methods of Multivariate Statistics. John Wiley & Sons., New York.

Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate distribution. In: Proceedings 3rd Berkeley Symposium on Mathematical Statistics and Probability. pp. 197–206.

Stein, C., 1986. Lectures on the theory of estimation of many parameters. J. Math. Sci. 34, 1373–1403.

Tsukuma, H., 2016. Estimation of a high-dimensional covariance matrix with the Stein loss. J. Multivariate Anal. 148, 1–17.

Tsukuma, H., Kubokawa, T., 2016. Unified improvements in estimation of a normal covariance matrix in high and low dimensions. J. Multivariate Anal. 143, 233–248.