The role of the isotonizing algorithm in Stein's covariance matrix estimator

Brett Naul, Bala Rajaratnam & Dario Vincenzi

Computational Statistics

ISSN 0943-4062
Volume 31
Number 4

Comput Stat (2016) 31:1453-1476
DOI 10.1007/s00180-016-0672-4

ISSN 0943-4062

Computational Statistics 31 · 4

2016

ORIGINAL PAPERS

(Contents continued on outside back cover)

Springer

Springer

Comput Stat (2016) 31:1453–1476
DOI 10.1007/s00180-016-0672-4

ORIGINAL PAPER

The role of the isotonizing algorithm in Stein's covariance matrix estimator

Brett Naul[1] · Bala Rajaratnam[2] · Dario Vincenzi[3]

Received: 4 November 2015 / Accepted: 12 July 2016 / Published online: 10 August 2016

Abstract Covariance matrix estimation is central to many applications in statistics and allied fields. A useful estimator in this context was proposed by Stein which regularizes the sample covariance matrix by shrinking its eigenvalues together. This estimator can sometimes yield estimates of the eigenvalues that are negative or differ in order from the observed eigenvalues. In order to rectify this problem, Stein also proposed an *ad hoc* "isotonizing" procedure which pools together eigenvalue estimates in such a way that the original ordering and positivity of the estimates are enforced. From numerical studies, Stein's "isotonized" estimator is known to have good risk properties in comparison with the maximum likelihood estimator. However, it remains unclear what role is played by the isotonizing procedure in the remarkable risk reductions achieved by Stein's estimator. Through two distinct lines of investigations, it is established that Stein's estimator without the isotonizing algorithm gives only modest risk reductions.

Electronic supplementary material The online version of this article (doi:10.1007/s00180-016-0672-4) contains supplementary material, which is available to authorized users.

✉ Dario Vincenzi
dario.vincenzi@unice.fr

Brett Naul
bnaul78@stanford.edu

Bala Rajaratnam
brajarat@stanford.edu

1 Institute for Computational and Mathematical Engineering, Stanford University, 475 Via Ortega-Huang Building, Stanford, CA 94305, USA

2 Department of Statistics, Stanford University, 390 Serra Mall-Sequoia Hall, Stanford, CA 94305, USA

3 Laboratoire Jean Alexandre Dieudonné, CNRS, Université Nice Sophia Antipolis, Parc Valrose, 06108 Nice, France

In cases where the isotonizing algorithm is frequently used, however, Stein's estimator can lead to significant risk reductions for certain domains of the parameter. In other cases, Stein's estimator can even yield risk increases, such as when (1) the theoretical eigenvalues are well separated, and/or (2) when the sample size is moderate to large, leading to over-shrinkage.

Keywords Covariance matrix estimation · Eigenvalues · Shrinkage · Steinian estimation · Isotonized estimator · Risks

1 Introduction

The estimation of the covariance matrix of a random vector is central to many multivariate statistical procedures, and has found applications in various branches of the sciences and engineering (Ledoit and Wolf 2004; Schäfer and Strimmer 2005; Karoui 2008; Pope and Szapudi 2008; Hamimeche and Lewis 2009; Fisher and Sun 2011; Khare and Rajaratnam 2011; Won et al. 2013; Li et al. 2013; Touloumis 2015). It is well known that the standard estimator, the sample covariance matrix, performs poorly unless the sample size n is much larger than the dimension of the covariance matrix p. To this end, various alternative estimators have been proposed in the literature. Estimators in both the frequentist and Bayesian frameworks have been developed, often by imposing some structure, either implicitly or explicitly, in order to obtain a regularized estimator with good risk properties. The reader is referred to Rajaratnam et al. (2008), Pourahmadi (2011), and references therein for a brief literature review.

A useful covariance matrix estimator was proposed by Stein (1975, 1977, 1986). Stein notes that the sample spectrum is severely distorted unless $n \gg p$, in the sense that there is a much larger spread in the sample spectrum as compared to its population counterpart. He proposes an approach to "shrink" the sample eigenvalues closer together by deriving the so-called unbiased estimator of risk. This approach allows Stein to optimally modify the sample eigenvalues in order to minimize the unbiased estimator of risk. An undesirable feature of this estimator is that the modified eigenvalues can lead to negative eigenvalue estimates or deviation from the original order of the sample eigenvalues, or even to both types of violation simultaneously. To this end an isotonizing algorithm was proposed, the purpose of which is to retain the original order of the sample eigenvalues and maintain positivity. The isotonizing algorithm produces ordered, positive eigenvalue estimates by recursively pooling together estimates for which either the desired order or sign is violated. Although this procedure is guaranteed to produce estimates which satisfy the natural order and sign constraints, it is nevertheless *ad hoc* in the sense that it no longer corresponds to an estimator which minimizes the unbiased estimator of risk, and therefore its effect on the risk properties of the estimator are not well understood. Nevertheless, the resulting estimator, "Stein's covariance matrix estimator," has been found to perform well in many numerical studies and is often used as a benchmark for comparisons with new estimators (Lin and Perlman 1985; Loh 1991; Yang and Berger 1994; Daniels and Kass 2001; Ledoit and Wolf 2004, 2014; Wang et al. 2015).

Despite the desirable risk properties of Stein's estimator, to the best of our knowledge a systematic investigation of Stein's estimator has not been undertaken in the literature. In this paper, we aim to quantify the effect of Stein's isotonizing algorithm on the risk reductions given by Stein's estimator. In particular, we undertake two lines of investigation corresponding to two different ways of isolating Stein's isotonizing algorithm from Stein's "raw" estimator.

The first line of investigation studies a variant of Stein's estimator where Stein's isotonizing algorithm is replaced by the maximum likelihood estimate whenever sign or order violations are encountered. This has the effect of isolating Stein's "raw" estimator from Stein's isotonizing algorithm and thus enables a comparison between the MLE and Stein's "raw" estimator. We examine how the sample size n and parameter Σ affect the isotonized and non-isotonized cases.

In the second line of investigation we calculate the risk reductions for cases/samples that require isotonizing separately from those that do not require isotonizing. We compare the risk reductions in the two cases, as well as the probability that isotonizing is required. We go one step further and quantify how the magnitude of the risk reductions depends on the number of order/sign violations. Interesting properties of the isotonizing algorithm and the important role it plays in risk reductions are elucidated. The effect of isotonizing is compared for various choices of n and Σ, from which we can draw useful conclusions about the isotonizing procedure and how it is influenced by different parameter regimes. Our numerical investigations consider both the small p regime, as considered in Lin and Perlman (1985), as well as high-dimensional analogs.

The outline of the paper is as follows. Section 2 briefly introduces preliminaries. Section 3 isolates the effect of Stein's isotonizing algorithm by replacing it with the MLE in the presence of sign and order violations. Section 4 describes the second component of our simulation study and gives a breakdown of the risk reductions into two scenarios: when isotonizing is required as compared with when it is not. Section 5 extends the analysis performed in Sect. 4 to other classes of covariance matrices. Section 6 concludes by summarizing the results in the paper. A supplementary document is also provided, which serves to give more detail on some of the results in the paper.

2 Preliminaries

The definition of Stein's estimator is briefly recalled in this section; for more details, the reader is referred to Lin and Perlman (1985) and Rajaratnam and Vincenzi (2015). Consider a random sample, $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, from a p-dimensional normal distribution $\mathcal{N}_p(0, \Sigma)$ with $n \geqslant p$. The eigenvalues of the theoretical covariance matrix Σ are denoted as $\lambda_i, i = 1, \ldots, p$. The sample covariance matrix S (up to a multiplicative constant) is given by:

$$S = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^t \tag{1}$$

and satisfies: $S \sim W_p(\Sigma, n)$, where $W_p(\Sigma, n)$ denotes the p-dimensional Wishart distribution with scale matrix Σ and n degrees of freedom. The matrix S admits the following spectral decomposition: $S = H \operatorname{diag}(\boldsymbol{l}) H^t$, where H is orthogonal and $\boldsymbol{l} = (l_1, l_2, ..., l_p)$ with $l_1 \geqslant l_2 \geqslant ... \geqslant l_p > 0$ being the ordered eigenvalues of S. Stein (1975, 1977, 1986) considers the class of orthogonally invariant estimators:

$$\widehat{\Sigma} = H\Phi(\boldsymbol{l})H^t, \tag{2}$$

where $\Phi(\boldsymbol{l}) = \mathrm{diag}(\varphi_1(\boldsymbol{l}), \varphi_2(\boldsymbol{l}), ..., \varphi_p(\boldsymbol{l}))$. The MLE given by S/n corresponds to $\widehat{\varphi}_j^{\mathrm{ml}}(\boldsymbol{l}) = l_j/n$.

The risk of $\widehat{\Sigma}$ under the loss function

$$L_1(\widehat{\Sigma}, \Sigma) := \mathrm{tr}(\widehat{\Sigma}\Sigma^{-1}) - \ln\det(\widehat{\Sigma}\Sigma^{-1}) - p \tag{3}$$

is given by

$$R_1(\widehat{\Sigma}, \Sigma) := \mathbb{E}_{\Sigma}[L_1(\widehat{\Sigma}, \Sigma)]. \tag{4}$$

Stein proves the following identity:

$$R_1(\widehat{\Sigma}, \Sigma) = \mathbb{E}_{\Sigma}[F(\boldsymbol{l})], \tag{5}$$

where

$$F(\boldsymbol{l}) := \sum_{j=1}^{p} \left[(n-p-1)\frac{\varphi_j(\boldsymbol{l})}{l_j} + 2\varphi_j(\boldsymbol{l}) \sum_{i \neq j} \frac{1}{l_j - l_i} + 2\frac{\partial \varphi_j}{\partial l_j} - \ln \frac{\varphi_j(\boldsymbol{l})}{l_j} \right] - c_{p,n} \tag{6}$$

with

$$c_{p,n} := \mathbb{E}\left(\sum_{j=1}^{p} \ln \chi^2_{n-j+1} \right) + p = \sum_{j=1}^{p} \frac{\Gamma'\left(\frac{1}{2}(n-j+1)\right)}{\Gamma\left(\frac{1}{2}(n-j+1)\right)} + p\ln 2 + p. \tag{7}$$

Stein observes that $F(\boldsymbol{l})$ is an unbiased estimator of the risk of $\widehat{\Sigma}$ (Stein 1975, 1977, 1986). To obtain a closed-form *bona fide* estimator, Stein disregards $\partial\varphi_j/\partial l_j$ in $F(\boldsymbol{l})$ and minimizes the resulting expression with respect to the φ_j. He thus obtains the following modified estimates of the eigenvalues of Σ:

$$\widehat{\varphi}_j^{\mathrm{St}}(\boldsymbol{l}) := \frac{l_j}{\alpha_j(\boldsymbol{l})}, \qquad j = 1, \ldots, p, \tag{8}$$

where

$$\alpha_j(\boldsymbol{l}) := n - p + 1 + 2l_j \sum_{i \neq j} \frac{1}{l_j - l_i}. \tag{9}$$

The $\widehat{\varphi}_j^{\mathrm{St}}(\boldsymbol{l})$ can yield estimators that violate the original ordering of the sample eigenvalues (as given by $l_1 \geqslant l_2 \geqslant \cdots \geqslant l_p > 0$) and furthermore can also yield negative estimates (Lin and Perlman 1985; Rajaratnam and Vincenzi 2015). Stein thus proposes an isotonizing algorithm which removes such violations by pooling adjacent estimators together (Stein 1975, 1977, 1986). The "pooled estimator" obtained by using $\widehat{\varphi}_j^{\text{St}}(\boldsymbol{l}), \widehat{\varphi}_{j+1}^{\text{St}}(\boldsymbol{l}), \ldots, \widehat{\varphi}_{j+s}^{\text{St}}(\boldsymbol{l})$ is:

$$\widehat{\varphi}_j^{\text{iso}}(\boldsymbol{l}) = \widehat{\varphi}_{j+1}^{\text{iso}}(\boldsymbol{l}) = \cdots = \widehat{\varphi}_{j+s}^{\text{iso}}(\boldsymbol{l}) := \frac{l_j + l_{j+1} + \cdots + l_{j+s}}{\alpha_j(\boldsymbol{l}) + \alpha_{j+1}(\boldsymbol{l}) + \cdots + \alpha_{j+s}(\boldsymbol{l})}\ . \tag{10}$$

Estimates that violate decreasing order or positivity are pooled according to the following procedure. First, negative values α_i are pooled together with previous values α_{j-1} until all estimates are positive. Next, order violations are corrected by pooling together pairs of estimates that are increasing rather than decreasing. The algorithm terminates when the sequence contains no more order or sign violations. For more details on the isotonizing algorithm, we refer the reader to the appendix in Lin and Perlman (1985). To distinguish between Stein's isotonized estimator and the original version, we shall refer to the former as Stein's isotonized estimator and the latter as Stein's "raw" estimator, unless the context is clear. The study of the impact of the isotonizing algorithm on risk reductions obtained when using Stein's isotonized estimator is the subject of the next sections.

3 Isolating the impact of isotonization: I. Substituting the isotonized values with the MLE

Lin and Perlman (1985) perform a numerical experiment comparing the average loss for several covariance matrix estimators including Stein's estimator across a variety of test population covariance structures. The selected covariance matrices have dimension $p = 6$ and are meant to represent a wide range of possible covariance structures, including the equal variance white noise case, matrices with just one large eigenvalue and the rest small and close together, and matrices with all widely spaced eigenvalues.

The test cases in Lin and Perlman (1985) are parametrized by two vectors: first, a p-dimensional vector $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_p)$, which represents the standard deviations of each variable; and second, a symmetric $p \times p$ correlation matrix R with diagonal 1 and $p(p-1)/2$ off-diagonal entries $-1 \leq \rho_{ij} \leq 1$. The covariance matrices considered can then be expressed as $\Sigma = \text{diag}(\boldsymbol{\sigma})R\text{diag}(\boldsymbol{\sigma})$, so that $\sum_{ij} = \sigma_i\sigma_j\rho_{ij}$. Five 6×6 test matrices Σ_α $(1 \leqslant \alpha \leqslant 5)$ have been examined in Lin and Perlman (1985). They are specified below:

$$\begin{aligned}
&\boldsymbol{\sigma}_1 = (1, 1, 1, 1, 1, 1), \quad \boldsymbol{\rho}_1 = (0; 0, 0; \ldots; 0, \ldots, 0);\\
&\boldsymbol{\sigma}_2 = (1, 1, 1, 1, 1, 1),\\
&\boldsymbol{\rho}_2 = (0.9; 0.9, 0.9; \ldots; 0.9, \ldots, 0.9);\\
&\boldsymbol{\sigma}_3 = (3.08, 2.66, 3.00, 2.55, 4.73, 2.93),\\
&\boldsymbol{\rho}_3 = (0.60; -0.38, -0.45; 0.61, 0.43, -0.61;\\
&\qquad 0.09, 0.34, -0.51, 0.63; -0.36, 0.08, 0.36, -0.21, 0.20);\\
&\boldsymbol{\sigma}_4 = (1, 1, 1, 1, 1, 1),
\end{aligned}$$

Table 1 Eigenvalues of the population covariance matrices Σ_α

Σ_α	Eigenvalues	Ratios of adjacent eigenvalues
Σ_1	(1, 1, 1, 1, 1, 1)	(1, 1, 1, 1, 1)
Σ_2	(5.5, 0.1, 0.1, 0.1, 0.1, 0.1)	(0.18, 1, 1, 1, 1)
Σ_3	$(2^5, 2^4, 2^3, 2^2, 2^1, 2^0)$	(0.5, 0.5, 0.5, 0.5, 0.5)
Σ_4	(4.56, 0.71, 0.42, 0.17, 0.08, 0.06)	(0.16, 0.60, 0.40, 0.47, 0.75)
Σ_5	(81.54, 3.25, 2.78, 2.27, 0.65, 0.51)	(0.04, 0.86, 0.82, 0.29, 0.78)

$$\begin{aligned}\boldsymbol{\rho}_4 &= (0.58; 0.61, 0.58; 0.60, 0.53, 0.94; 0.57, 0.53, 0.87, 0.88; 0.60, 0.55,\\ &\quad 0.88, 0.88, 0.92);\\ \boldsymbol{\sigma}_5 &= (1, 2, 3, 4, 5, 6), \quad \boldsymbol{\rho}_5 = \boldsymbol{\rho}_4.\end{aligned}$$

The eigenvalues of the matrices Σ_α and the ratios of the adjacent eigenvalues are given in Table 1.

In addition, the following covariance matrix, not investigated in Lin and Perlman (1985), is also added to the five above:

$$\Sigma_6 = \text{diag}(10^2, 10, 1, 10^{-1}, 10^{-2}, 10^{-3}).$$

The covariance matrices considered above all represent classes of parameters which arise naturally in practice. The first matrix Σ_1 corresponds to the well known case of "sphericity" and is ubiquitous in hypothesis tests of covariance structure. The second matrix Σ_2 describes collections of random variables which are all highly positively correlated with each other. The corresponding population eigenvalue ensemble confirms that much of the variation in the random variables is explained by just one leading principal component, which is indicative of the intrinsic "low dimensional" nature of the covariance matrix. The third covariance matrix Σ_3 describes the setting where there are both positive and negative correlations and is typical of various applications in genomics, environmental sciences, and finance. The population eigenvalue ensemble varies in such a way that the eigenvalues increase by powers of two. The fourth covariance matrix Σ_4 describes collections of random variables with equal variances that are all moderately positively correlated with each other and have just one leading principal component. Though Σ_4 is similar to Σ_2 in terms of correlations, the case where the correlations are only moderately large is more typical in applications. The fifth covariance matrix Σ_5 is similar to Σ_4 except that the variances (i.e., the diagonal terms) are increasing. This added flexibility in Σ_5 (as compared to the homoscedastic assumption of Σ_4) is also more realistic and has the effect of increasing the variation explained by the largest principal component. The matrix Σ_6 is an example of a covariance matrix in which the adjacent eigenvalues are very well separated; thus, shrinkage imposed by Stein's estimator is unlikely to change the ordering of the sample eigenvalues, and so the isotonizing algorithm should be required only infrequently in this case.

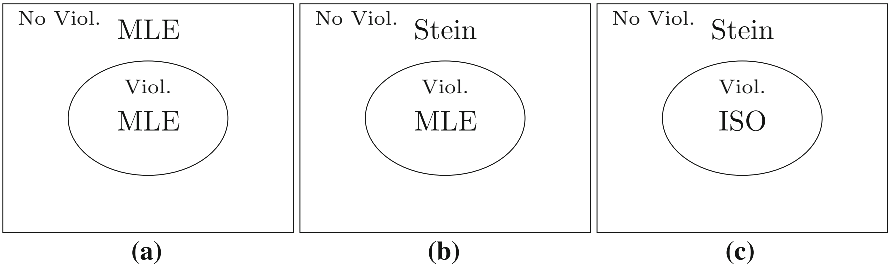

Fig. 1 A schematic representation of the three types of estimators compared in Sect. 3. "No Viol." represents the region of the sample space where sign and order of the sample eigenvalues are preserved by Stein's estimator, and conversely "Viol." represents the region where Stein's estimator yields negative values or deviates from the observed ordering. 1**a** represents the estimator where the MLE is used regardless of whether or not there are sign/order violations (ML). 1**b** represents the estimator where Stein's estimator is used when there are no violations and the MLE is used when sign/order violations are present (ST+ML). 1**c** represents the estimator where Stein's estimator is used when there are no violations and isotonization is used when violations arise (ST+ISO)

In order to study the relationship between the covariance matrix parameter Σ and the magnitude of the isotonizing correction, we introduce the following estimator:

$$\widehat{\varphi}_j^{\,\mathrm{St+ml}}(\boldsymbol{l}) := \begin{cases} \widehat{\varphi}_j^{\,\mathrm{St}}(\boldsymbol{l}) & \text{if } \alpha_j(\boldsymbol{l}) > 0 \text{ and } \frac{l_j}{\alpha_j(\boldsymbol{l})} \leqslant \frac{l_{j-1}}{\alpha_{j-1}(\boldsymbol{l})} \ \forall\, 2 \leqslant j \leqslant p \\ \widehat{\varphi}_j^{\,\mathrm{ml}}(\boldsymbol{l}) & \text{otherwise,} \end{cases} \tag{11}$$

with $j = 1, 2, \ldots, p$. That is, the above estimator $\widehat{\varphi}_j^{\,\mathrm{St+ml}}$ leaves Stein's raw estimator unchanged when the isotonizing correction is not required but isotonizes it by using the MLE when Stein's estimator requires isotonization. This is an alternative way to correct Stein's estimator when the estimated eigenvalues are negative and/or when their order is violated. More importantly, such an approach isolates the role of Stein's isotonizing algorithm in risk reductions. Figure 1 provides a schematic overview of how the role of Stein's isotonizing algorithm in risk reductions can be isolated from Stein's raw estimator.

The performance of the estimators $\widehat{\varphi}_j^{\,\mathrm{St+iso}}$ and $\widehat{\varphi}_j^{\,\mathrm{St+ml}}$ is compared by sampling $N = 1000$ random Wishart matrices from population covariance matrices Σ_α $(1 \leqslant \alpha \leqslant 6)$ and computing the percentage reduction in average loss over the MLE for each estimator,[1] defined as

$$\gamma_{L_1} := \frac{R_1\left(\widehat{\Sigma}^{\mathrm{ml}}, \Sigma\right) - R_1\left(\widehat{\Sigma}, \Sigma\right)}{R_1\left(\widehat{\Sigma}^{\mathrm{ml}}, \Sigma\right)} \times 100. \tag{12}$$

The corresponding results are indicated in Table 2 by "ST+ISO" and "ST+ML", respectively.

[1] Our R package that implements Stein's isotonized estimator is available on CRAN at https://cran.r-project.org/web/packages/stcov/.

Table 2 Percentage reduction in average loss, γ_{L_1}, for $\widehat{\varphi}_j^{\text{St+iso}}$ and $\widehat{\varphi}_j^{\text{St+ml}}$ and for different n and Σ_α. The estimator $\widehat{\varphi}_j^{\text{St+ml}}$ is defined in (11)

Σ	$\widehat{\Sigma}$	$n=6$	$n=15$	$n=30$	$n=60$	$n=100$
Σ_1	ST+ISO	53.3	71.3	74.5	75.7	76.1
	ST+ML	1.0	0.3	0.3	0.2	0.3
Σ_2	ST+ISO	47.0	54.2	53.4	52.5	52.0
	ST+ML	3.0	1.5	1.2	1.1	1.0
Σ_3	ST+ISO	38.7	20.5	7.1	0.2	−0.9
	ST+ML	5.2	6.6	5.3	1.1	−0.6
Σ_4	ST+ISO	37.8	21.4	10.3	3.7	1.4
	ST+ML	7.1	7.4	6.3	3.5	1.8
Σ_5	ST+ISO	39.5	26.5	19.4	15.3	12.2
	ST+ML	5.6	5.3	5.1	5.2	5.4
Σ_6	ST+ISO	23.4	8.8	4.7	2.5	1.6
	ST+ML	19.1	8.8	4.7	2.5	1.6

Several important insights come to light. Generally speaking, when the sample eigenvalues are not well separated, it should be expected that order/sign violations arise more frequently (a more detailed study of this point is performed in Sect. 4). When Σ is such that the probability density function of the corresponding sample eigenvalues attributes a significant weight to the region where order and sign violations occur (as in the case of Σ_1 and Σ_2 for all n, and in the case of Σ_3, Σ_4, and Σ_5 for small n), then the risk reductions given by Stein's isotonized estimator are considerably greater than those for the estimator where the MLE is used whenever violations are present. By contrast, if Σ is such that the density of the sample eigenvalues is concentrated where order and sign violations are not present, then the difference in risk reductions between the two estimators is minimal. In the case of Σ_6, the isotonizing algorithm is expected to play a minor role, since the sample eigenvalues are well separated. The results of the two estimators are indeed indistinguishable except for the case $n = p$, where the isotonizing correction should apply to a relatively larger portion of the domain. It is in the Σ_6 case where one actually sees Stein's intended shrinkage effect in action, as opposed to the effect of the isotonizing algorithm. It gives some modest but non-negligible risk reductions, but not on the same scale as those given by Stein's isotonizing algorithm.

Note that for Σ_3 and $n = 60$, using ST+ISO yields a risk reduction of 0.2% over the MLE, whereas ST+ML yields a risk reduction of 1.1% over the MLE. Since the two estimators coincide when there are no violations, the following calculation reveals that the risk reduction in ST+ISO stems from Stein's isotonizing algorithm yielding lower risk reductions than when the MLE is used. In particular, let

$$\begin{aligned}
\omega_1 &:= \Pr(\text{absence of sign/order violations}),\\
\omega_2 &:= \Pr(\text{presence of sign/order violations}),\\
\gamma_1^{\text{St}} &:= \mathbb{E}[L_1(\Sigma^{\text{St}}, \Sigma)|\text{absence of sign/order violations}],\\
\gamma_2^{\text{iso}} &:= \mathbb{E}[L_1(\Sigma^{\text{St}}, \Sigma)|\text{presence of sign/order violations}],\\
\gamma_1^{\text{ml}} &:= \mathbb{E}[L_1(\Sigma^{\text{ml}}, \Sigma)|\text{absence of sign/order violations}],
\end{aligned}$$

$$\gamma_2^{\text{ml}} := \mathbb{E}[L_1(\Sigma^{\text{ml}}, \Sigma)|\text{presence of sign/order violations}],$$
$$k := \mathbb{E}[L(\Sigma^{\text{ml}}, \Sigma)].$$

Note that

$$k = \omega_1\gamma_1^{\text{ml}} + \omega_2\gamma_2^{\text{ml}},$$
$$\text{Relative risk reduction of } \widehat{\Sigma}^{\text{St+iso}} = \frac{\omega_1\gamma_1^{\text{St}} + \omega_2\gamma_2^{\text{iso}} - k}{k}, \tag{13}$$
$$\text{Relative risk reduction of } \widehat{\Sigma}^{\text{St+ml}} = \frac{\omega_1\gamma_1^{\text{St}} + \omega_2\gamma_2^{\text{ml}} - k}{k}. \tag{14}$$

For the $\Sigma_\alpha = \Sigma_3$ and $n = 60$ case, the risk reduction can be decomposed as

$$1.1 = \frac{\omega_1\gamma_1^{\text{St}} + \omega_2\gamma_2^{\text{ml}} - k}{k} = \frac{\omega_1\gamma_1^{\text{St}} + \omega_2\gamma_2^{\text{ml}} - (\omega_1\gamma_1^{\text{ml}} + \omega_2\gamma_2^{\text{ml}})}{k} = \frac{\omega_1(\gamma_1^{\text{St}} - \gamma_1^{\text{ml}})}{k}.$$

Furthermore,

$$\begin{aligned} 0.2 &= \frac{\omega_1\gamma_1^{\text{St}} + \omega_2\gamma_2^{\text{iso}} - k}{k} = \frac{\omega_1\gamma_1^{\text{St}} + \omega_2\gamma_2^{\text{iso}} - (\omega_1\gamma_1^{\text{ml}} + \omega_2\gamma_2^{\text{ml}})}{k} \\ &= \frac{\omega_1(\gamma_1^{\text{St}} - \gamma_1^{\text{ml}})}{k} + \frac{\omega_2(\gamma_2^{\text{iso}} - \gamma_2^{\text{ml}})}{k} = 1.1 + \frac{\omega_2(\gamma_2^{\text{iso}} - \gamma_2^{\text{ml}})}{k} \\ &\implies \gamma_2^{\text{iso}} - \gamma_2^{\text{ml}} < 0 \text{ since } \omega_2 > 0 \text{ and } k > 0. \end{aligned}$$

As the above calculations demonstrate, when the isotonizing algorithm based on Stein's shrinkage estimator is used to rectify sign/order violations, it can sometimes actually yield lower risk reductions than when using the MLE.

It is clear from the simulation study described in Table 2 that Stein's isotonized algorithm cannot be solely credited with the risk reductions in Stein's estimator. In fact, as the previous example shows, Stein's isotonizing algorithm can even lead to relative risk increases. The relationship between relative risk reductions and the isotonizing algorithm is complicated by the confounding factor of the sample size n — in the sense that in general the need for the isotonizing algorithm decreases as sample size n increases, but Stein's estimator also tends to the MLE as n increases. There are exceptions, however. For example, for the case Σ_1, the true eigenvalues are all the same, and therefore $l_i/n \to \lambda_i = \lambda \ \forall i$ almost surely as n increases. In this case, the need for the isotonizing algorithm increases as n increases. Here it is clear that the bulk of the sample values will require isotonizing and the substantial risk reductions arise from using Stein's isotonizing algorithm, even for large n.

In summary, Stein's estimator makes extensive use of the isotonizing algorithm in two scenarios: when some of the sample eigenvalues are close to each other, and/or when the sample size n is comparable to the dimension p. In both scenarios, Stein's isotonized estimator undoubtedly outperforms the MLE, and therefore a significant part of the risk reduction seen in Stein's estimator should be attributed specifically to the isotonizing algorithm (and not the Steinian shrinkage effect that comes from the raw estimator). Nevertheless, there also exist situations in which the isotonizing algorithm based on the shrinkage estimator does not appreciably improve Stein's estimator or in which isotonizing Stein's estimator with the MLE results in even better performance.

4 Isolating the impact of isotonization: II. Breakdown of risk between the violations and no violations scenarios

4.1 Lin–Perlman test cases

The last section compared the performance of Stein's estimator under two scenarios: when either Stein's isotonizing algorithm or the MLE is used when order/sign violations arise. Such an analysis is only able to study the type of shrinkage given by the specific form of Stein's raw estimator when no violations appear. Hence the analysis in Sect. 3 can be most useful when violations are relatively few and far between. However, there are several contexts in which violations occur for a large fraction of samples where the isotonizing algorithm is invoked extensively, such as in the full/partial multiplicity case or the low sample setting (which is quite common in contemporary applications). In such cases, the analysis in Sect. 3 does not fully explain the role played by the isotonizing algorithm. We now undertake a more direct comparison of the cases where isotonizing is either required or not.

In order to assess comprehensively whether the isotonizing algorithm really improves the performance of Stein's original estimator, we need to compare the risk reductions from Stein's raw estimator directly to those of the isotonized version. A novel approach to achieving this goal is to split the sample space into two regions: first into a region where there are no order/sign violations, and second into a region where there are order/sign violations. Risk reduction over the MLE in the first region quantifies the performance of Stein's raw estimator, while that in the second region quantifies contributions from the isotonizing algorithm. Figure 2 shows the two cases that are compared. A comparison of these two conditional risk reductions paints a more accurate picture of the role of the isotonizing algorithm. Before we undertake the aforementioned analysis, a few remarks are in order.

At face value the isotonizing algorithm simply appears to be an order preserving algorithm, but it is important to note that its basic ingredients originate from Stein's estimator itself. In this sense, the isotonized Stein's estimator still retains features of the raw version. The isotonized version does however deviate relatively more from Stein's original estimator when many sign/order violations are present. Second, in order to understand the effect of the isotonizing algorithm, it is first important to understand when it is applied. The isotonizing algorithm "kicks in" relatively more frequently when the sample eigenvalues l_i are close to one another. This is because when the l_i are close, terms of the form $1/(l_i - l_j)$ in Stein's estimator become unbounded and can lead to sign and order violations. Holding all else constant, l_i are close to one another when the true eigenvalues λ_i are either identical or close to one another. The problem

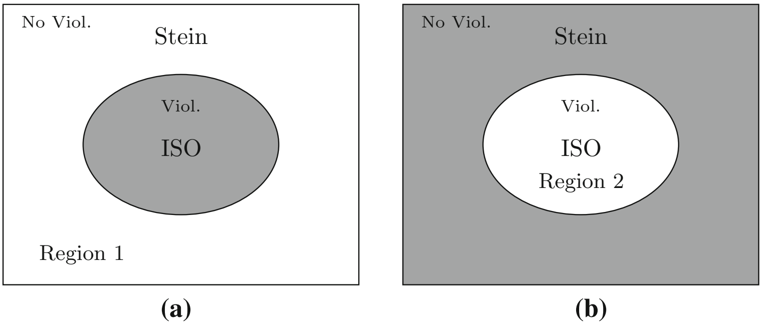

Fig. 2 Breakdown of the isotonizing algorithm's contribution to risk reductions into two regions: **a** reductions over the MLE when no order/sign violations occur (*Region* 1); **b** reductions over the MLE in the presence of order/sign violations (*Region* 2). The *lighter area* indicates the region where the risk is calculated in each case

Table 3 Percentage reduction in average loss for covariance matrices studied in Lin and Perlman (1985): comparison between cases with order/sign violation and those without

Σ	Order/Sign Violations	$n = 6$	$n = 15$	$n = 30$	$n = 60$	$n = 100$
Σ_1	No Violations	42.0	59.9	63.0	70.1	64.4
	Any Violations	54.9	71.9	74.7	76.0	76.3
Σ_2	No Violations	37.8	50.1	50.9	51.4	50.8
	Any Violations	48.2	54.3	53.5	52.8	52.4
Σ_3	No Violations	37.2	26.5	10.4	1.2	−0.6
	Any Violations	38.5	17.4	3.2	−5.5	−12.7
Σ_4	No Violations	36.1	25.5	12.9	5.1	2.3
	Any Violations	38.0	18.8	7.1	0.5	−3.0
Σ_5	No Violations	37.4	30.9	21.5	17.9	15.1
	Any Violations	40.1	23.7	16.4	12.2	7.6
Σ_6	No Violations	21.4	9.8	5.2	2.7	1.6
	Any Violations	4.4	–	–	–	–

of order violation is also exacerbated in small sample sizes due to the inherently higher variability of the l_i in such settings.

Table 3 provides an analysis of the cases considered in Sect. 3, broken down between cases where there are no violations (i.e., without using the isotonizing algorithm), and where there is at least one sign or order violation (i.e., when the isotonizing algorithm is used). The relative frequencies of both cases are given in Table 4 in order to highlight the probability of sign/order violations for each of the different covariance matrices. The quantities in Tables 3 and 4 are based on a simulation size of $N = 10^5$; the width of the normal 95 % confidence interval for each value is at most 0.02 %.

The role that the isotonizing algorithm plays in the risk reductions observed in Stein's estimator is rather complex. As expected, Table 4 clearly demonstrates for the cases Σ_1 and Σ_2 that Stein's raw estimator requires some type of sign or order

Table 4 Probabilities (in %) of order/sign violations for covariance matrices studied in Lin and Perlman (1985): comparison between cases with order/sign violation and those without

Σ	Order/Sign Violations	$n = 6$	$n = 15$	$n = 30$	$n = 60$	$n = 100$
Σ_1	No Violations	1.6	0.4	0.4	0.2	0.3
	Any Violations	98.4	99.6	99.6	99.8	99.7
Σ_2	No Violations	5.1	1.9	1.8	1.8	1.7
	Any Violations	94.9	98.1	98.2	98.2	98.3
Σ_3	No Violations	11.4	24.0	54.3	87.7	97.9
	Any Violations	88.6	76.0	45.7	12.3	2.1
Σ_4	No Violations	15.8	26.8	50.1	71.1	84.3
	Any Violations	84.2	73.2	49.9	28.9	15.7
Σ_5	No Violations	12.1	14.6	21.3	26.9	33.7
	Any Violations	87.9	85.4	78.7	73.1	66.3
Σ_6	No Violations	97.7	100.0	100.0	100.0	100.0
	Any Violations	2.3	0.0	0.0	0.0	0.0

correction in a vast majority of samples. Furthermore, for Σ_1 and Σ_2 the relative risk reductions over the MLE are higher for the cases where there is at least one sign or order violation (see Table 3). This pattern is evident regardless of the sample size n. Hence it is clear that the isotonizing algorithm tends to yield higher risk reductions in cases with at least one violation as compared to Stein's original estimator. Having said this, it should also be mentioned that for Σ_1 and Σ_2 Stein's raw form still gives high relative risk reductions, but not as high as when the isotonizing algorithm is used. The fact that sign/order violations are so prevalent implies that the benefits of Stein's raw form are rarely featured in the estimator. Thus the superior performance of Stein's estimator in the Σ_1 and Σ_2 cases can in part be attributed to the isotonizing algorithm. In this sense, Stein's estimator, without the isotonizing algorithm, can only lead to moderate risk reductions. It is however important to note that Σ_1 and Σ_2 exhibit multiplicity in the true eigenvalues. Hence the pooling performed by the isotonizing algorithm, though seemingly artificial at first, in fact reflects and enforces the multiplicity present in the true eigenvalues. It is therefore an appropriate procedure to use in such cases and can clearly lead to higher risk reductions.

The above assertions do not necessarily imply that Stein's risk reductions stem only from the isotonizing algorithm. We can see this in two ways: first by noting that the isotonizing algorithm itself is based on Stein's original estimator, and second by considering the results in Table 4 in the Σ_3 and Σ_4 cases for $n = 60$ and 100. Indeed, a different story emerges when we consider Σ_3 and Σ_4. For both these cases the relative frequency of sign/order violations decreases rapidly as the sample size increases. Besides the case when the sample size $n = 6$, the relative risk reductions over the MLE are higher for the cases where there are no order or sign violations (see Table 3). There is thus evidence that the isotonizing algorithm can also diminish the performance of Stein's original estimator. Regardless, the role of the isotonizing algorithm is more subtle in the sense that if the sample size is very low it can still lead to higher risk reductions (see $n = 6$ for the Σ_3 and Σ_4 cases). Plots for Σ_1–Σ_6 of (1) the relative risk reduction for the violations/no violations cases and (2) frequencies of violations are given in Fig. 3.

The covariance matrix Σ_5 presents a slightly different situation from the $\Sigma_1 - \Sigma_4$ cases in the sense that the relative risk reductions over the MLE are higher for the cases where there is no order or sign violation, though it would appear that the isotonizing algorithm is often required since the probability of sign/order violations remains relatively high even for moderately large sample sizes. For the matrix Σ_6, violations arise only for n extremely small; for $n = 6$, we observe that the risk reductions are much greater when no violations are present.

Further insights into the role of the isotonizing algorithm (in terms of risk reductions) can be gained by separating the samples according to the number of "poolings" that the isotonizing algorithm makes. The number of poolings aims to measure the extent to which the isotonized estimator deviates from Stein's raw estimator. Figure 4 provides the relative risk reductions for each of Σ_1 to Σ_6 broken down into four groups: "no violations", "1 pooling", "2 poolings" and "3+ poolings" for sample sizes $n = 6, 15, 30, 60, 100$. For Σ_1 and Σ_2, higher risk reductions are recorded in the cases where there is a greater number of poolings. This should be expected since a larger number of poolings means that more eigenvalues have been brought together, reflecting the structure of the population parameters Σ_1 and Σ_2. This effect, which is monotonic in the number of poolings, is more pronounced in small sample sizes. The pattern of risk reduction observed for Σ_1 and Σ_2 is reversed for Σ_3 to Σ_5. In the latter cases, more pooling tends to diminish the risk reductions, especially for large n. The reversal is also to be expected since the separated eigenvalue cases do not warrant as much isotonizing, especially for large n. For Σ_6, the probability of multiple violations is close to zero, so it is not possible to evaluate the effect of the total number of poolings without a much greater number of Monte Carlo samples.

Yet another in-depth analysis of the direct risk reductions due to the isotonizing algorithm can be undertaken by further separating the violations into two types, either sign or order violations, and thereafter quantifying their relative risk reductions. This type of breakdown gives further insights into the workings of the isotonizing algorithm. The risk gains incurred when rectifying a sign violation are consistently similar to those when rectifying an order violation. The exception is the $n = 6$ case, where the risk gains when rectifying sign violations are much higher than those for order violations. Specific details are found in the supplementary document (see Sect. A).

In order to confirm the general validity of our results for small p, we have performed additional numerical simulations for $p = 3$ and $p = 10$ (see Sect. B in the supplementary document). The results in the $p = 3$ and $p = 10$ cases confirm our understanding of the isotonizing algorithm.

4.2 Higher-dimensional risk comparisons

In many modern day applications, data sets often contain a very large number of variables. The dimension of the covariance matrix parameter being estimated can therefore be much larger than the $p = 6$ case considered in Lin and Perlman (1985). By

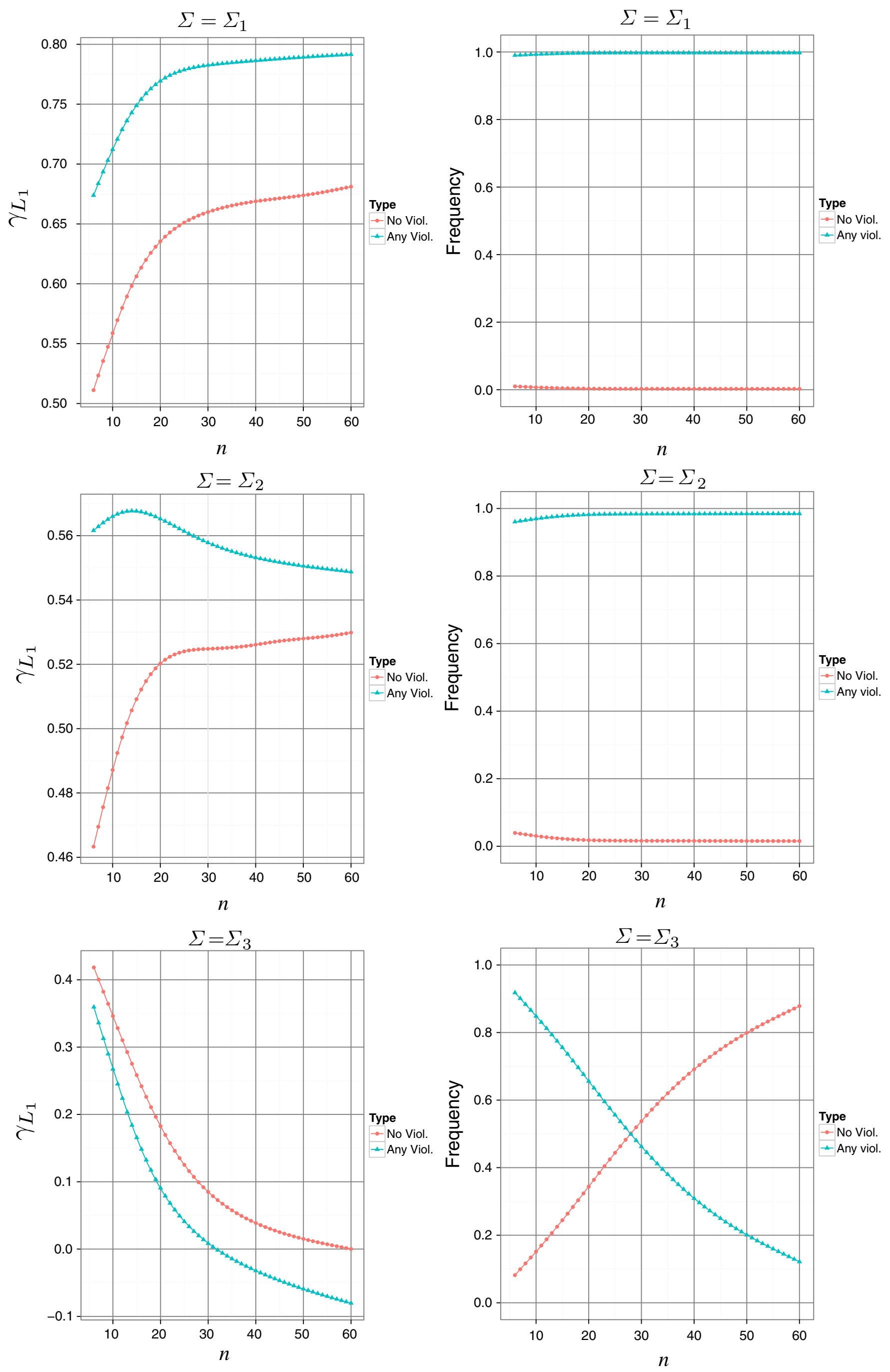

Fig. 3 Percentage reduction in average loss and relative frequency of violations for various Σ_α

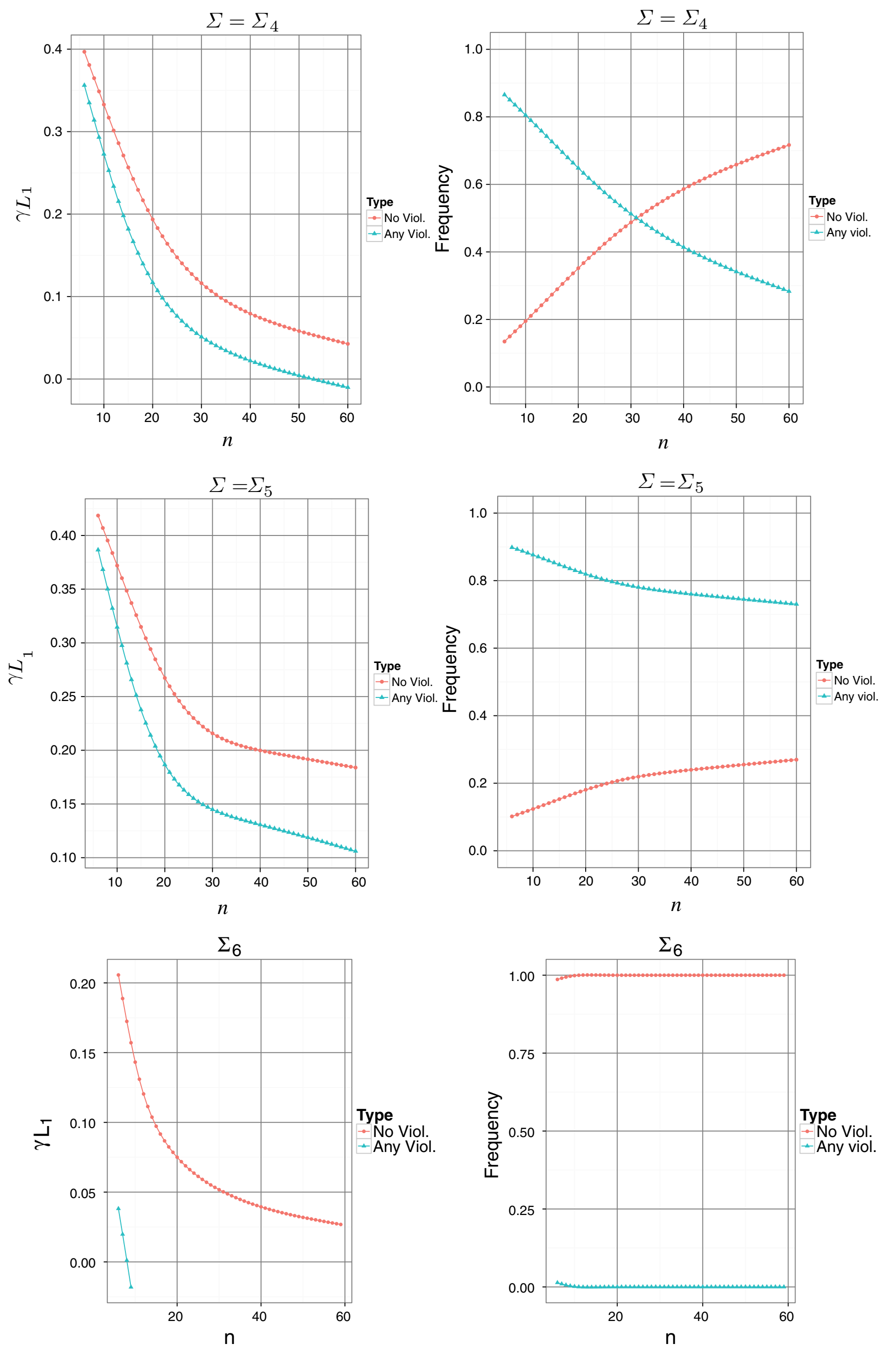

Fig. 3 continued

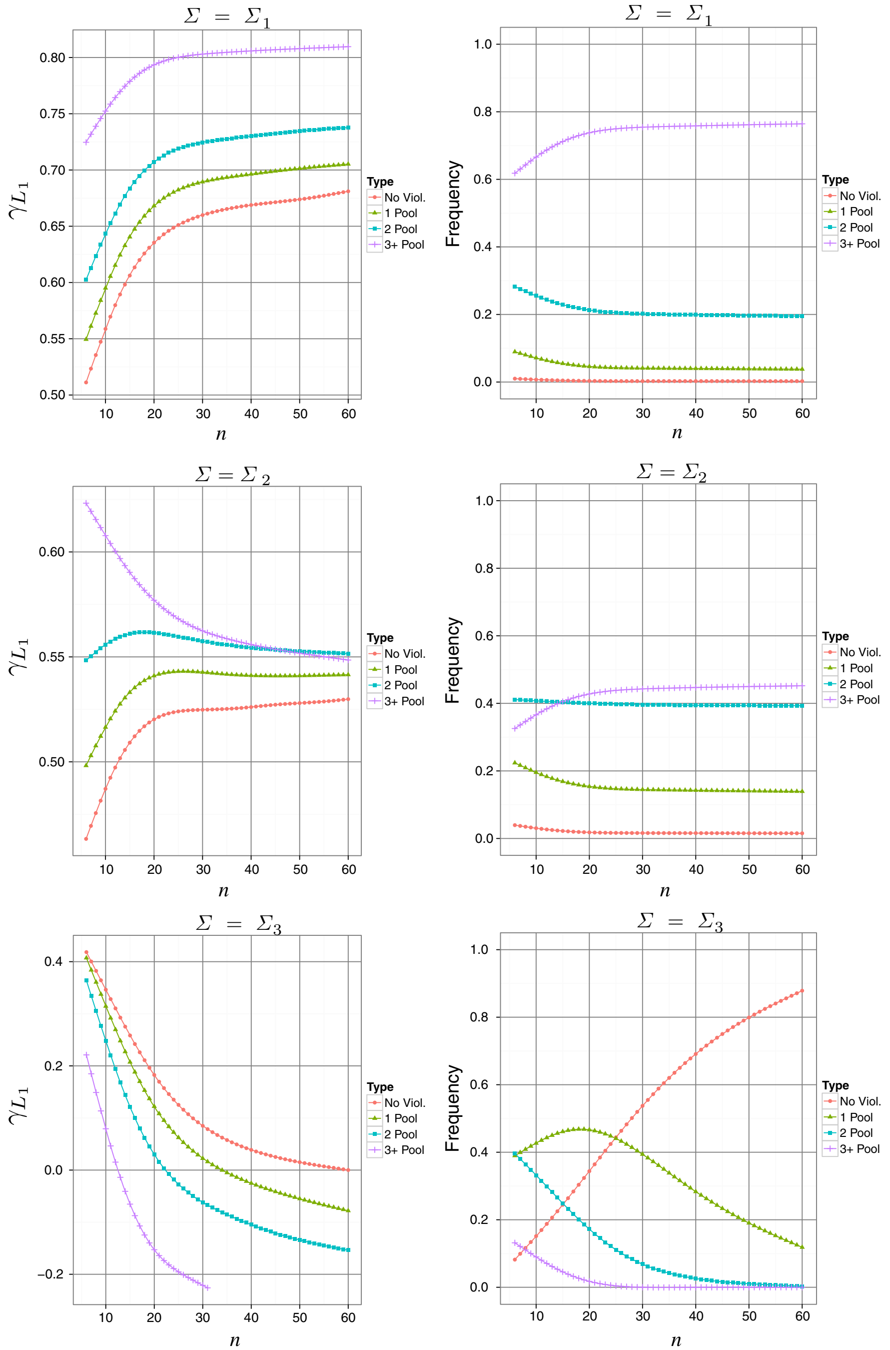

Fig. 4 Percentage reduction in average loss and relative frequency of violations by number of poolings for various Σ_α

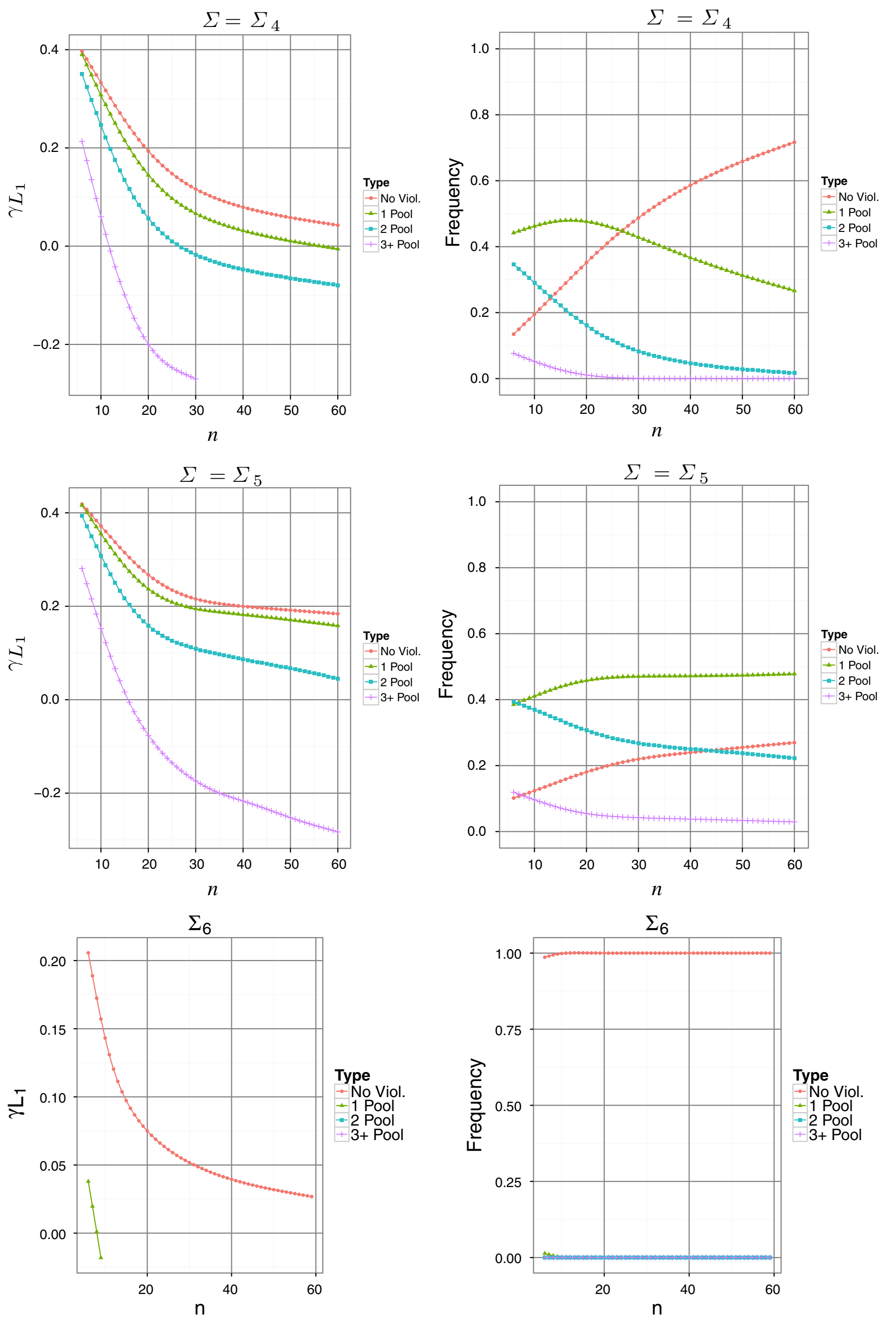

Fig. 4 continued

carefully studying a variety of distinct types of covariance matrices in the $p = 6$ case, we do gain important qualitative insights into the general behavior of the isotonizing algorithm. However, it is also important to examine quantitatively how these results translate to higher dimensions. To this end, we extended the Lin–Perlman simulations by generalizing the parameters Σ_1 to Σ_6 for larger values of p. Below we examine the $p = 200$ case in detail. The eigenvalues of the matrices Σ_α in this case are:

$$\begin{aligned}
&\Sigma_1 : (1, 1, \ldots, 1, 1);\\
&\Sigma_2 : (180.1, 1, \ldots, 1, 1);\\
&\Sigma_3 : (2^{10}, 2^{9.95}, \ldots, 2^{0.15}, 2^{0.1});\\
&\Sigma_4 : (72.64, 1.27, \ldots, 0.033, 0.021);\\
&\Sigma_5 : (145.5, 19.44, \ldots, 1.30, 1.21);\\
&\Sigma_6 : (10^5, 10^{4.95}, \ldots, 10^{-4.95}, 10^{-5}).
\end{aligned}$$

Each of the above matrices Σ_α was chosen so as to resemble the same type of matrix as the corresponding Lin–Perlman Σ_α matrix. The first case, Σ_1, is the identity matrix, which represents a "white noise" Wishart model, i.e., independent random variables with equal variance. Similarly, Σ_2 also has unit diagonal (representing homoscedastic random variables), but with very strong positive correlations (0.9) between every variable. The matrix Σ_3 has eigenvalues that increase geometrically and are relatively well separated; the off-diagonals contain both positive and negative terms. The fourth case, Σ_4, has unit diagonal and positive off-diagonal elements like Σ_2, but the values are chosen to be smaller ($\rho_{ij} \in [0.6, 0.8]$ rather than $\rho_{ij} = 0.9$). The matrix Σ_5 has the same correlation structure as Σ_4, but has unequal variance terms evenly spaced between 10 and 1. Finally, Σ_6 is diagonal matrix with logarithmically spaced eigenvalues between 10^5 and 10^{-5}, which should result in relatively few order and sign violations.

As in the previous section, it is possible to explore the role of the isotonizing algorithm by breaking down the sample space according to the number of order and sign violations that arise. Unlike the $p = 6$ case, for sufficiently large p there will almost always be at least some violations, since the number of pairs of eigenvalues which must maintain their original ordering and sign increases rapidly with p. Thus, the isotonizing algorithm plays an even greater role in the behavior of Stein's (isotonized) estimator in the large-p setting. We give a breakdown of the sample space according to the number of eigenvalues that are pooled together by the isotonizing algorithm. For $p = 200$, we consider four groupings: between 1 and 100 poolings are performed, between 101 and 150 poolings, between 151 and 180 poolings, and 181 or more poolings.

Table 5 shows the risk reductions in the $p = 200$ setting, broken down by the number of eigenvalues pooled together by the isotonizing algorithm. Table 6 shows how often each amount of pooling occurs. The simulation size is once again taken to be $N = 10^5$. The quantities being estimated are deterministic; however, when a certain violation type is extremely rare, the estimate of the reduction in average loss can be poor. For the sake of accuracy, we filter our reported estimates in the following

Table 5 Percentage reduction in average loss for $p = 200$ Lin–Perlman style test cases grouped by number of poolings

Σ	No. of poolings	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 3333$
Σ_1	No Violations	–	–	–	–	–
	1–100 Pool	–	–	–	–	–
	101–150 Pool	–	–	–	–	–
	151–180 Pool	88.7	–	–	–	–
	181+ Pool	92.7	99.7	99.8	99.8	99.8
	Overall	92.7	99.7	99.8	99.8	99.8
Σ_2	No Violations	–	–	–	–	–
	1–100 Pool	–	–	–	–	–
	101–150 Pool	–	–	–	–	–
	151–180 Pool	89.0	–	–	–	–
	181+ Pool	92.4	99.0	98.9	98.9	98.9
	Overall	92.4	99.0	98.9	98.9	98.9
Σ_3	No Violations	–	–	–	–	–
	1–100 Pool	–	–	10.3	5.1	3.0
	101–150 Pool	39.1	17.9	10.0	4.9 ± 6.6	–
	151–180 Pool	33.8	–	–	–	–
	181+ Pool	–	–	–	–	–
	Overall	37.0	17.9	10.0	5.1	3.0
Σ_4	No Violations	–	–	–	–	–
	1–100 Pool	–	–	–	–	–
	101–150 Pool	–	–	33.8	24.3	17.9
	151–180 Pool	50.5	25.5	30.4	22.8	–
	181+ Pool	44.6	–	–	–	–
	Overall	50.1	25.5	30.6	24.3	17.9
Σ_5	No Violations	–	–	–	–	–
	1–100 Pool	–	–	–	–	–
	101–150 Pool	–	–	28.6	20.9	15.4
	151–180 Pool	43.7	18.1	25.9	19.9 ± 2.1	–
	181+ Pool	36.4	–	–	–	–
	Overall	43.5	18.1	26.7	20.9	15.4
Σ_6	No Violations	–	–	–	3.1	1.8
	1–100 Pool	34.1	12.4	6.3	3.1	1.8
	101–150 Pool	30.4 ± 3.3	–	–	–	–
	151–180 Pool	–	–	–	–	–
	181+ Pool	–	–	–	–	–
	Overall	34.1	12.4	6.3	3.1	1.8

Table 6 Probabilities (in %) of numbers of poolings for $p = 200$ Lin–Perlman style test cases

Σ	No. of poolings	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 3333$
Σ_1	No Violations	0.0	0.0	0.0	0.0	0.0
	1–100 Pool	0.0	0.0	0.0	0.0	0.0
	101–150 Pool	0.0	0.0	0.0	0.0	0.0
	151–180 Pool	0.02	0.0	0.0	0.0	0.0
	181+ Pool	99.99	100.0	100.0	100.0	100.0
Σ_2	No Violations	0.0	0.0	0.0	0.0	0.0
	1–100 Pool	0.0	0.0	0.0	0.0	0.0
	101–150 Pool	0.0	0.0	0.0	0.0	0.0
	151–180 Pool	0.04	0.0	0.0	0.0	0.0
	181+ Pool	99.96	100.0	100.0	100.0	100.0
Σ_3	No Violations	0.0	0.0	0.0	0.0	0.0
	1–100 Pool	0.0	0.0	7.8	99.98	100.0
	101–150 Pool	59.7	100.0	92.2	0.02	0.0
	151–180 Pool	40.3	0.0	0.0	0.0	0.0
	181+ Pool	0.0	0.0	0.0	0.0	0.0
Σ_4	No Violations	0.0	0.0	0.0	0.0	0.0
	1–100 Pool	0.0	0.0	0.0	0.0	0.0
	101–150 Pool	0.0	0.0	5.4	97.8	100.0
	151–180 Pool	94.3	100.0	94.6	2.2	0.0
	181+ Pool	5.7	0.0	0.0	0.0	0.0
Σ_5	No Violations	0.0	0.0	0.0	0.0	0.0
	1–100 Pool	0.0	0.0	0.0	0.0	0.0
	101–150 Pool	0.0	0.0	27.5	99.9	100.0
	151–180 Pool	97.0	100.0	72.5	0.1	0.0
	181+ Pool	3.0	0.0	0.0	0.0	0.0
Σ_6	No Violations	0.0	0.0	0.0	0.1	32.8
	1–100 Pool	99.9	100.0	100.0	99.9	67.2
	101–150 Pool	0.1	0.0	0.0	0.0	0.0
	151–180 Pool	0.0	0.0	0.0	0.0	0.0
	181+ Pool	0.0	0.0	0.0	0.0	0.0

ways. First, when the estimated probability of a violation type is too small ($< 20\,\%$ of the width of the corresponding multinomial 95 % confidence interval), we omit the estimated risk reduction and replace the probability with "**". Furthermore, in cases where the normal 95 % confidence interval for the average loss is wider than 5 % of the absolute value of the estimated risk, the confidence interval is also included along with the average risk reduction. The same technique is used for other simulation results presented in the supplementary document.

Broadly speaking, the effect of the isotonizing algorithm on risk reductions here appears to mirror what was observed in the $p = 6$ case. In the Σ_1 and Σ_2 cases, the isotonizing algorithm pools together large numbers of eigenvalues even for large sample sizes. Since there is multiplicity in the population parameters, pooling the sample eigenvalues together better reflects the underlying eigenstructure, and as a result the risk reductions increase with the number of poolings. In this sense, the isotonizing algorithm contributes greatly to the risk reductions achieved by Stein's isotonized estimator in the Σ_1 and Σ_2 cases. For Σ_3–Σ_6, the population eigenvalues are all distinct, and accordingly we see that on average relatively fewer poolings are performed by the isotonizing algorithm. In almost every case, for these matrices more poolings lead to diminished risk reductions, and in some cases they are non-negligible (see, for example, the difference in risk reductions for Σ_4 when $n = 1000$ between the 101–150 poolings and 151–180 poolings cases). Thus, in these cases we conclude that the form of Stein's raw estimator can yield risk reductions, though these risk reductions are modest in relative terms. Overall, the breakdown of risk reductions by number of poolings reinforces the same conclusions drawn in previous sections. When the true parameter value contains many eigenvalues that are equal or close together, more frequent application of the isotonizing algorithm results in significantly higher risk reductions. Conversely, when most of the population eigenvalues are all well separated, pooling together too many of the sample eigenvalues can result in substantially lower risk reductions.

Additional simulation results for moderate-dimensional ($p = 50$) analogs of the Lin–Perlman test cases Σ_1–Σ_6 are presented in the supplementary document (Sect. C). The same overall patterns described above are evident in the moderate-dimensional regime.

5 Other population covariance models

In addition to test cases based on those from Lin and Perlman (1985), for completeness we also present simulation results in Sects. D and E of the supplementary document for two other classes of covariance matrices. First, Sect. D of the supplementary document tests so-called "spiked" covariance models, in which the population covariance matrix has a few (or one) large eigenvalues and the rest are equal to the same small value. Such a matrix corresponds to a model where a small number of factors are responsible for the majority of the variation in the data, with the rest of the variability coming from random white noise (see Johnstone 2001, for details). In Sect. D of the supplementary document, we consider the cases $p = 6, 50$, and 200 for various numbers and magnitude of spikes, so that the population eigenvalues are given by $\boldsymbol{\lambda} = (M, \ldots, M, 1, \ldots, 1)$ for some large values M.

The final class of population covariance matrices we consider is that of exponentially decaying eigenvalues, for which simulation results are presented in Sect. E of the supplementary document. In particular, we consider matrices of the form $\Sigma = \mathrm{diag}(M^k, M^{k-k/p}, \ldots, M^{2k/p}, M^{k/p})$; the parameter k can be thought of as controlling the rate of decay of the eigenvalues, and the parameter M determines the magnitude of the largest eigenvalue. Such a covariance structure provides a continuous generalization of the spiked covariance model: instead of a few large values and many much smaller entries, an exponentially decaying set of eigenvalues produces a continuous spectrum with support everywhere between the largest value M^k

and 1. Whereas a spiked model represents a few strong signals added to white noise, a decaying-eigenvalue model corresponds to p different sources of noise (one for each observed variable), each with a distinct noise level.

For the most part, these additional test cases reinforce the conclusions drawn from the analyses of the Lin–Perlman style test cases: Stein's estimator yields the greatest risk reductions over the MLE when the true eigenvalues have high multiplicity, but the improvement decreases when the number of poolings becomes farther away from the true multiplicity of the underlying parameter. However, the spiked covariance model, in particular for Σ_6 (i.e., many large spikes), also illuminates the potential for very large errors. This increase in risk is closely related to the number of poolings performed by the isotonizing algorithm. For example, in the $\Sigma = \Sigma_6$ case, the true multiplicities of the population eigenvalues 200 and 1 are 25 and 175, respectively; when n is large, there is a high probability that eigenvalues from the two clusters are pooled together erroneously, leading to either a huge over- or underestimate of many of the eigenvalues. In such cases, the average loss of Stein's isotonized estimator can be more than an order of magnitude greater than that of the MLE.

6 Summary and concluding remarks

This paper undertakes a numerical investigation of the risk reductions achieved by Stein's estimator and the complex role that Stein's isotonizing algorithm plays in this regard. In particular, we quantified the effect of Stein's isotonizing algorithm via two sets of numerical simulations. Our first line of investigation demonstrates that the significant risk reductions in Stein's estimator cannot be solely attributed to the form of the raw estimator. One way to see this is to replace Stein's estimator by the MLE whenever the isotonizing algorithm would be employed. This approach isolates Stein's isotonizing algorithm from Stein's raw estimator and allows one to compare Stein's estimator in its raw form to the MLE. Such an investigation reveals that Stein's raw estimator leads to only modest risk reductions (and not anywhere close to the significant risk reductions reported in the literature). It therefore appears that the isotonizing algorithm plays a crucial role in the risk reductions reported in the literature with regards to Stein's estimator. In cases where the isotonizing is often employed, it can potentially lead to significant risk reductions, as is seen in our second line of investigation. This is evident from the difference in risk reductions between the cases when there are no sign/order violations and when there is at least one sign/order violation. However, these insights need to be interpreted in context, as the isotonized estimator still retains features of Stein's original estimator, though by itself it has no decision theoretic basis. Furthermore, for some parameter values the use of the isotonizing algorithm can also diminish risk reductions. Hence one can verify that in certain settings Stein's isotonizing algorithm can be beneficial, while in other cases it is not as useful. The significant risk reductions are attributable to the use of the isotonizing algorithm (as compared to the form of Stein's raw estimator) when pooling is appropriate for the underlying parameter values. Such cases arise (1) when there is multiplicity in the population eigenvalues, and/or (2) when the sample size is low resulting in many order violations. Rajaratnam and Vincenzi (2015) recently undertook a detailed study of the theoretical properties of Stein's covariance matrix estimator. This study allows a practitioner to understand in which regimes Stein's estimator is expected to perform better than the MLE.

The preceding analysis provides a deep understanding of Stein's estimator, including the type of shrinkage given by Stein's raw covariance matrix estimator and the complex role played by Stein's isotonizing algorithm in risk reductions. The analysis brings us to a more philosophical question: Why is Stein's estimator not performing as well as originally perceived?

We offer two ways to look at this. First, we feel that it is not that Stein's estimator is performing poorly, but rather that the risk reductions in the general case are modest compared to those observed in the full multiplicity case. Aiming for the risk reductions seen in the full/partial multiplicity case is misleading because (1) this parameter setting is unique and is therefore not representative, and (2) the isotonizing algorithm is optimal when there is multiplicity in the eigenvalues, hence corresponding risk reductions paints a misleading picture that such reductions are possible in general in other settings. The use of the isotonizing algorithm is not as applicable in other settings and there is no similar "regularization" mechanism that can be used to shrink the estimator towards the (unknown) true parameter. The reasoning above can be supplemented with another explanation: first note that Stein's estimator modifies only the eigenvalues but retains the original eigenvectors of the sample covariance matrix. The number of parameters encoded by the eigenvectors is of $O(p^2)$. It is well known that these parameters could also benefit from Steinian shrinkage (Daniels and Kass 2001).

Second, note that Stein's estimator is "optimal" since the form of the estimator arises out of minimizing an objective function. The minimization however disregards a derivative term in order to solve the optimization problem (see Rajaratnam and Vincenzi 2015). Hence Stein's raw form does not fully attain its potential for risk reductions since the optimization is inexact, but this is unavoidable in order to obtain a closed form *bona fide* estimator. Moreover, the optimization being only approximate has two "unfortunate" consequences: (a) lowering of risk reductions that could have been achieved, and (b) singularities leading to sign and order reversal. The isotonizing algorithm of course provides a means to rectify the latter problem. Though isotonization has no formal decision theoretic basis, it is appropriate in certain parameter regimes. It also retains some features of Stein's raw form, leading to a complex pattern of risk reductions.

Acknowledgments The authors would like to gratefully acknowledge funding from the France–Stanford Center for Interdisciplinary Studies which facilitated their collaborations. We thank Prof. Charles Stein for his encouragement and enthusiasm when this work was initiated. BR was supported in part by the National Science Foundation under Grant Nos. DMS 0906392, DMS-CMG 1025465, AGS-1003823, DMS-1106642 and Grants NSA H98230-11-1-0194, DARPA-YFA N66001-11-1-4131, and SUWIEVP10-SUFSC10-SMSCVISG0906. BN was supported in part by grant DARPA-YFA N66001-11-1-4131.

References

Daniels M, Kass R (2001) Shrinkage estimators for covariance matrices. Biometrics 57(4):1173–1184

El Karoui N (2008) Spectrum estimation for large dimensional covariance matrices using random matrix theory. Ann Stat 36(6):2757–2790

Fisher TJ, Sun XQ (2011) Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. Comp Stat Data Anal 55(5):1909–1918

Hamimeche S, Lewis A (2009) Properties and use of CMB power spectrum likelihoods. Phys Rev D. 79(8):083012

Johnstone IM (2001) On the distribution of the largest eigenvalue in principal component analysis. Ann Stat 29(2):295–327

Khare K, Rajaratnam B (2011) Wishart distributions for decomposable covariance graph models. Ann Stat 39(1):514–555

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

Ledoit O, Wolf M (2014) Optimal estimation of a large-dimensional covariance matrix under Stein's loss. Working Paper N. 122, Department of Economics, University of Zurich

Li W, Chen J, Qin Y, Bai Z, Yao J (2013) Estimation of the population spectral distribution from a large dimensional sample covariance matrix. J Stat Plan Inference 143(11):1887–1897

Lin S, Perlman M (1985) A Monte Carlo comparison of four estimators of a covariance matrix. In: Krishnaiah P (ed) Multivariate Analysis, vol 6. North Holland, Amsterdam, pp 411–429

Loh WL (1991) Estimating covariance matrices. Ann Stat 19(1):283–296

Pope A, Szapudi I (2008) Shrinkage estimation of the power spectrum covariance matrix. Mon Not R Astron Soc 389(2):766–774

Pourahmadi M (2011) Covariance estimation: The GLM and regularization perspectives. Stat Sci 26(3):369–387

Rajaratnam B, Vincenzi D (2015) A theoretical study of Stein's covariance estimator. Biometrika (in press) https://hal.archives-ouvertes.fr/hal-01243388

Rajaratnam B, Massam H, Carvalho C (2008) Flexible covariance estimation in graphical Gaussian models. Ann Stat 36(6):2818–2849

Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4(1):32

Stein C (1975) Estimation of a covariance matrix. Rietz Lecture

Stein C (1977) Lectures on the theory of estimation of many parameters (in Russian). In: Ibragimov IA, Nikulin MS (eds) Studies in the Statistical Theory of Estimation, Part I, Proceedings of Scientific Seminars of the Steklov Institute, Leningrad Division, vol 74, pp 4–65

Stein C (1986) Lectures on the theory of estimation of many parameters. J Math Sci 34(1):1373–1403

Touloumis A (2015) Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. Comput Stat Data Anal 83:251–261

Wang C, Pan G, Tong T, Zhu L (2015) Shrinkage estimation of large dimensional precision matrix using random matrix theory. Stat Sin 25:993–1008

Won J, Lim J, Kim S, Rajaratnam B (2013) Condition-number-regularized covariance estimation. J R Stat Soc Ser B Stat Methodol 75(3):427–450

Yang R, Berger JO (1994) Estimation of a covariance matrix using the reference prior. Ann Stat 22(3):1195–1211