The Annals of Statistics
2020, Vol. 48, No. 4, 2381–2403
https://doi.org/10.1214/19-AOS1891

BAYESIAN ANALYSIS OF THE COVARIANCE MATRIX OF A MULTIVARIATE NORMAL DISTRIBUTION WITH A NEW CLASS OF PRIORS

BY JAMES O. BERGER[1], DONGCHU SUN[2] AND CHENGYUAN SONG*[3]

[1] *Department of Statistical Science, Duke University, berger@stat.duke.edu*
[2] *Department of Statistics, University of Nebraska-Lincoln, sund9@unl.edu*
[3] *School of Statistics, East China Normal University, songchengyuanchina@163.com*

Bayesian analysis for the covariance matrix of a multivariate normal distribution has received a lot of attention in the last two decades. In this paper, we propose a new class of priors for the covariance matrix, including both inverse Wishart and reference priors as special cases. The main motivation for the new class is to have available priors—both subjective and objective—that do not "force eigenvalues apart," which is a criticism of inverse Wishart and Jeffreys priors. Extensive comparison of these "shrinkage priors" with inverse Wishart and Jeffreys priors is undertaken, with the new priors seeming to have considerably better performance. A number of curious facts about the new priors are also observed, such as that the posterior distribution will be proper with just three vector observations from the multivariate normal distribution—regardless of the dimension of the covariance matrix—and that useful inference about features of the covariance matrix can be possible. Finally, a new MCMC algorithm is developed for this class of priors and is shown to be computationally effective for matrices of up to 100 dimensions.

1. Introduction. Estimating the unknown covariance matrix of a multivariate normal population has been an important issue for more then half a century. It has a wide range of modern applications including astrophysics ([18, 26]), economics ([23]), the environmental sciences ([11, 13]), climatology ([15]) and genetics ([30]).

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ be a random sample of $k \times 1$ vectors from the $N_k(\mathbf{0}, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\Sigma}$ is the $k \times k$ unknown covariance matrix. (For simplicity, we assume the normal mean is zero, although essentially the same results would hold for a nonzero mean.) Our goal is to find good prior distributions—objective and subjective—for $\boldsymbol{\Sigma}$.

The historical thread of this work starts with efforts to improve upon the "classical" estimator for $\boldsymbol{\Sigma}$,

$$\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{m}\boldsymbol{S} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{y}_i \boldsymbol{y}_i'.$$

This is the maximum likelihood estimate, the unbiased estimator and the posterior mean arising from use of the Jeffreys prior $\pi^J(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-(k+1)/2}$. It has long been known to be a suboptimal estimator. First, (modest) improvements were obtained through use of the best equivariant estimator (see Section 4.3 for definition), with major improvements occurring when [33] discovered the value of shrinking together the eigenvalues of $\hat{\boldsymbol{\Sigma}}_0$. Important follow-up research included [3, 7–9, 16, 17, 24, 25, 29, 34, 35] and [20].

On the Bayesian side, the early priors that were used (and still are used) were the Jeffreys prior and conjugate inverse Wishart (IW) priors, but they have been criticized in much the

Received August 2018; revised June 2019.

*Corresponding author.

MSC2020 subject classifications. Primary 62F15; secondary 62C10, 62H10, 62H86.
Key words and phrases. Covariance, objective priors, shrinkage priors, inverse Wishart prior.

same way that $\hat{\mathbf{\Sigma}}_0$ was criticized. Indeed, when transforming to the eigenvalue-eigenvector parameterization, these priors were seen ([35]) to have a term

$$\prod_{i<j}(\lambda_i - \lambda_j)$$

in the density, where $\lambda_1 > \lambda_2 > \cdots > \lambda_k$ are the ordered eigenvalues of $\mathbf{\Sigma}$. Since this term becomes zero whenever eigenvalues get close together, these common priors have the effect of forcing the eigenvalues of $\mathbf{\Sigma}$ apart. It is thus no surprise that [33] improved on $\hat{\mathbf{\Sigma}}_0$ by shrinking the eigenvalues together; Jeffreys prior (for which $\hat{\mathbf{\Sigma}}_0$ is the posterior mean) had forced the eigenvalues apart in creating the estimate.

The motivation for this work was twofold. First, to develop a class of priors—which we call *shrinkage inverse Wishart (SIW)* priors (including both subjective and objective versions)—that corrects the "forcing eigenvalues apart" problem. Second, to develop computational schemes for the new priors that, while not nearly as simple computationally as the IW priors, allow for computational handling of large dimensional (e.g., $k = 100$) covariance matrices. Note that approaching this problem from the Bayesian side carries a number of benefits, including the fact that the Bayesian estimates will be guaranteed to be positive definite (a problem with non-Bayesian approaches) and, as usual, the availability of measures of accuracy of estimates.

It is important to note here that we are considering the "vanilla" covariance matrix problem; we are assuming no special structure or sparsity for $\mathbf{\Sigma}$. There is a vast modern literature dealing with priors for structured or sparse covariance matrices (see [27] for discussion of some of the early contributions).

Some curiosities were observed in this investigation. One such was that the posteriors for the SIW priors are proper when the sample size, m, is three or more, regardless of the dimension k of the covariance matrix. It is commonly perceived that one needs k observations to "identify" $\mathbf{\Sigma}$, so the situation is interesting. Indeed, we provide some evidence that certain features of $\mathbf{\Sigma}$ (such as its trace) can be learned with much fewer than k observations, which we call *low rank learning*.

In Section 2, the new class of priors for $\mathbf{\Sigma}$ is proposed; interestingly, it is also a conjugate class. Propriety and moment existence results are obtained for both the prior and posterior distributions, and a method for subjectively eliciting the parameters of the prior is developed for an important special case. Computation for the priors and posteriors is considered in Section 3, with the proposed new method being capable of handling large (e.g., $k = 100$) dimensional covariance matrices. Section 4 presents extensive comparisons of the new and old priors, with the new priors appearing to be much better. In this section, we also explore issues surrounding low rank learning.

2. A new class of priors.

2.1. *Definition.* The new class of priors for $\mathbf{\Sigma}$ consists of densities

$$\pi(\mathbf{\Sigma} \mid a, b, \boldsymbol{H}) \propto \frac{\operatorname{etr}(-\frac{1}{2}\mathbf{\Sigma}^{-1}\boldsymbol{H})}{|\mathbf{\Sigma}|^a[\prod_{i<j}(\lambda_i - \lambda_j)]^b}, \tag{2.1}$$

where $\lambda_1 > \cdots > \lambda_k > 0$ are the eigenvalues of $\mathbf{\Sigma}$, a is a real constant, b is a number in $[0, 1]$, $\boldsymbol{H}$ is a positive semidefinite matrix and $\operatorname{etr}(\boldsymbol{A}) = \exp\{\operatorname{trace}(\boldsymbol{A})\}$ for a square matrix $\boldsymbol{A}$. If $b = 0$, this becomes the inverse Wishart density (denoted by $\mathbf{\Sigma} \sim \operatorname{IW}(a, \boldsymbol{H})$)

$$\pi^{\mathrm{IW}}(\mathbf{\Sigma} \mid a, \boldsymbol{H}) = \frac{|\boldsymbol{H}|^{a-(k+1)/2}\operatorname{etr}(-\frac{1}{2}\mathbf{\Sigma}^{-1}\boldsymbol{H})}{2^{(2a-k-1)k/2}\pi^{k(k-1)/4}\prod_{i=1}^k \Gamma(\frac{2a-k-1}{2})|\mathbf{\Sigma}|^a}, \tag{2.2}$$

which is proper if $a > k$.

Common objective priors that are in this class include:

- the constant prior, $\pi^C(\boldsymbol{\Sigma}) = 1$ (corresponding to $a = b = 0$, $\boldsymbol{H} = \boldsymbol{0}$);
- the Jeffreys prior ([22]), $\pi^J(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-(k+1)/2}$ (corresponding to $a = (k+1)/2$, $b = 0$ and $\boldsymbol{H} = \boldsymbol{0}$);
- the [35] reference prior, $\pi^R(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-1}[\prod_{i<j}(\lambda_i - \lambda_j)]^{-1}$ (corresponding to $a = b = 1$ and $\boldsymbol{H} = \boldsymbol{0}$);
- the modified reference prior, $\pi^{\mathrm{MR}}(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-[1-1/(2k)]}[\prod_{i<j}(\lambda_i - \lambda_j)]^{-1}$ (corresponding to $a = 1 - 1/(2k)$, $b = 1$ and $\boldsymbol{H} = \boldsymbol{0}$), which was suggested in [2] for use with covariance matrices that occurred at higher levels of a hierarchical model.

2.1.1. *Shrinkage inverse Wishart priors.* In this paper, we focus on the subclass in (2.1) when $b = 1$:

$$\pi^{\mathrm{SIW}}(\boldsymbol{\Sigma} \mid a, \boldsymbol{H}) \propto \frac{\mathrm{etr}(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{H})}{|\boldsymbol{\Sigma}|^a \prod_{i<j}(\lambda_i - \lambda_j)}. \tag{2.3}$$

This will be called the *shrinkage inverse Wishart* (SIW) class. To see why the label "shrinkage" is attached to this class, consider the one-to-one transformation from $\boldsymbol{\Sigma}$ to $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_k)$ and the orthogonal matrix $\boldsymbol{\Gamma}$ of corresponding eigenvectors; it follows from [12] that the Jacobian is

$$\left|\frac{\partial \boldsymbol{\Sigma}}{\partial(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})}\right| = \prod_{i<j}(\lambda_i - \lambda_j), \tag{2.4}$$

and the prior density (2.1) for $\boldsymbol{\Sigma}$ becomes the density of $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$

$$\pi(\boldsymbol{\Lambda}, \boldsymbol{\Gamma} \mid a, b, \boldsymbol{H}) \propto \frac{\mathrm{etr}(-\frac{1}{2}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}'\boldsymbol{H})}{|\boldsymbol{\Lambda}|^a[\prod_{i<j}(\lambda_i - \lambda_j)]^{b-1}} 1_{\{\lambda_1 > \cdots > \lambda_k\}}, \tag{2.5}$$

with respect to Lebesgue measure on $(\lambda_1, \dots, \lambda_k)$ and the invariant Haar measure over the space of all orthonormal matrices $\boldsymbol{\Gamma}$. The invariant prior on $\boldsymbol{\Gamma}$ (essentially a uniform prior over rotations) is natural and noncontroversial. However, when $b = 0$ (which corresponds to the commonly used priors such as inverse Wishart, Jeffreys and constant), the presence of the term $[\prod_{i<j}(\lambda_i - \lambda_j)]$ in the prior is quite strange; the prior is near zero whenever eigenvalues are close together, so that the prior effectively forces eigenvalues apart. This seems contrary to common intuition and typical prior beliefs.

In contrast, the SIW priors have $b = 1$, so that the questionable term in the density disappears. The SIW priors are essentially neutral as to how the eigenvalues should be spread out; in that sense, calling them "shrinkage" priors is something of a misnomer, but they are shrinkage priors compared to the commonly used $b = 0$ priors.

2.1.2. *SIW priors with $\boldsymbol{H} \propto \boldsymbol{I}_k$.* Unfortunately, working with the SIW priors is not as easy as working with the IW priors, but the special case of SIW$(a, c\boldsymbol{I}_k)$ priors is quite tractable. With IW$(a, \boldsymbol{H})$ priors, $\boldsymbol{H}$ is very often chosen to be a multiple of the identity, so this subclass of SIW priors is important. Note, from (2.5), that the density of $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$ for this subclass is

$$\pi(\boldsymbol{\Lambda}, \boldsymbol{\Gamma} \mid a, c) \propto \prod_{i=1}^{k} \frac{1}{\lambda_i^a} e^{-\frac{c}{2\lambda_i}} 1_{\{\lambda_1 > \lambda_2 > \cdots > \lambda_k\}}.$$

TABLE 1
Different priors for $\boldsymbol{\Sigma}$ *and corresponding priors for* $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$ *where* $\mathcal{A} = \{(\lambda_1, \dots, \lambda_k) \mid \lambda_1 > \cdots > \lambda_k\}$.

Prior for $\boldsymbol{\Sigma}$	Prior for $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$
$\pi(\boldsymbol{\Sigma} \mid a, b, \boldsymbol{H}) \propto \frac{\mathrm{etr}(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{H})}{\lvert\boldsymbol{\Sigma}\rvert^a \prod_{i<j}(\lambda_i - \lambda_j)^b}$	$\pi(\boldsymbol{\Lambda}, \boldsymbol{\Gamma} \mid a, b, \boldsymbol{H}) \propto \frac{\mathrm{etr}(-\frac{1}{2}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}'\boldsymbol{H}) 1_{\mathcal{A}}}{\lvert\boldsymbol{\Lambda}\rvert^a \prod_{i<j}(\lambda_i - \lambda_j)^{b-1}}$
$\pi^{\mathrm{IW}}(\boldsymbol{\Sigma} \mid a, \boldsymbol{H}) \propto \frac{\mathrm{etr}(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{H})}{\lvert\boldsymbol{\Sigma}\rvert^a}$	$\pi^{\mathrm{IW}}(\boldsymbol{\Lambda}, \boldsymbol{\Gamma} \mid a, \boldsymbol{H}) \propto \frac{\prod_{i<j}(\lambda_i - \lambda_j) 1_{\mathcal{A}}}{\lvert\boldsymbol{\Lambda}\rvert^a \,\mathrm{etr}(\frac{1}{2}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}'\boldsymbol{H})}$
$\pi^{\mathrm{SIW}}(\boldsymbol{\Sigma} \mid a, \boldsymbol{H}) \propto \frac{\mathrm{etr}(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{H})}{\lvert\boldsymbol{\Sigma}\rvert^a \prod_{i<j}(\lambda_i - \lambda_j)}$	$\pi^{\mathrm{SIW}}(\boldsymbol{\Lambda}, \boldsymbol{\Gamma} \mid a, \boldsymbol{H}) \propto \frac{\mathrm{etr}(-\frac{1}{2}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}'\boldsymbol{H}) 1_{\mathcal{A}}}{\lvert\boldsymbol{\Lambda}\rvert^a}$
$\pi^J(\boldsymbol{\Sigma}) \propto \frac{1}{\lvert\boldsymbol{\Sigma}\rvert^{(k+1)/2}}$	$\pi^J(\boldsymbol{\Lambda}, \boldsymbol{\Gamma}) \propto \frac{\prod_{i<j}(\lambda_i - \lambda_j)}{\lvert\boldsymbol{\Lambda}\rvert^{(k+1)/2}} 1_{\mathcal{A}}$
$\pi^R(\boldsymbol{\Sigma}) \propto \frac{1}{\lvert\boldsymbol{\Sigma}\rvert \prod_{i<j}(\lambda_i - \lambda_j)}$	$\pi^R(\boldsymbol{\Lambda}, \boldsymbol{\Gamma}) \propto \frac{1}{\lvert\boldsymbol{\Lambda}\rvert} 1_{\mathcal{A}}$
$\pi^{\mathrm{MR}}(\boldsymbol{\Sigma}) \propto \frac{1}{\lvert\boldsymbol{\Sigma}\rvert^{1-1/(2k)} \prod_{i<j}(\lambda_i - \lambda_j)}$	$\pi^{\mathrm{MR}}(\boldsymbol{\Lambda}, \boldsymbol{\Gamma}) \propto \frac{1}{\lvert\boldsymbol{\Lambda}\rvert^{1-1/(2k)}} 1_{\mathcal{A}}$
$\pi^C(\boldsymbol{\Sigma}) \propto 1$	$\pi^C(\boldsymbol{\Lambda}, \boldsymbol{\Gamma}) \propto \prod_{i<j}(\lambda_i - \lambda_j) 1_{\mathcal{A}}$
$\pi^U(\boldsymbol{\Sigma}) \propto \frac{1}{\prod_{i<j}(\lambda_i - \lambda_j)}$	$\pi^U(\boldsymbol{\Lambda}, \boldsymbol{\Gamma}) \propto 1_{\mathcal{A}}$

REMARK 1. The prior for $\boldsymbol{\Gamma}$ is constant, and the marginal prior density of $(\lambda_1, \dots, \lambda_k)$ is

$$\pi(\boldsymbol{\Lambda} \mid a, c) \propto \prod_{i=1}^{k} \frac{1}{\lambda_i^a} e^{-\frac{c}{2\lambda_i}} 1_{\{\lambda_1 > \lambda_2 > \cdots > \lambda_k\}},$$

which will be seen to be equivalent to the eigenvalues, $\lambda_1 > \cdots > \lambda_k$, arising as the order statistics of k observations from the Inverse Gamma$(a - 1, \frac{c}{2})$ distribution.

2.1.3. *Summary of priors.* The various priors for $\boldsymbol{\Sigma}$ considered above, and the corresponding priors for $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$, are summarized in Table 1. One additional prior is given therein, namely π^U, labeled as the uniform prior because it corresponds to the constant prior for $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$.

2.1.4. *SIW posteriors.* For a simple random sample, $\boldsymbol{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_m)$, from $N_k(\mathbf{0}, \boldsymbol{\Sigma})$ and using the prior SIW$(a, \boldsymbol{H})$, the posterior is given by

$$\pi(\boldsymbol{\Sigma} \mid \boldsymbol{Y}) \propto \frac{1}{\lvert\boldsymbol{\Sigma}\rvert^r [\prod_{i<j}(\lambda_i - \lambda_j)]} \mathrm{etr}\left(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{H}_0\right), \tag{2.6}$$

where $r = a + m/2$ and $\boldsymbol{H}_0 = \boldsymbol{H} + \boldsymbol{S}$. This is the SIW$(r, \boldsymbol{H}_0)$ distribution, so the SIW priors are a conjugate family.

2.2. *Propriety of* SIW *priors and posteriors.* The following theorem, whose proof is in Appendix A.1, gives sufficient conditions for propriety of the SIW prior distribution.

THEOREM 1. *For the* SIW$(a, \boldsymbol{H})$ *prior for* $\boldsymbol{\Sigma}$, *with* $p = \mathrm{rank}(\boldsymbol{H}) > 0$,

(a) *when* $p = k$, *the prior is proper iff* $a > 1$;
(b) *when* $0 < p < k$, *the prior is proper iff* $1 < a < 1 + p/2$.

It follows from Theorem 1 that, for the priors (2.1) with $p = \mathrm{rank}(H) < k$, $a = 1 + p/2$ is the boundary of impropriety.

For propriety of the posterior, we need the following lemma, whose proof is given in Appendix A.1.

LEMMA 1. *Let $\boldsymbol{H}$ be the prior scale matrix with* $\operatorname{rank}(\boldsymbol{H}) = p$. *Then, with probability one,*

$$p^* = \operatorname{rank}(\boldsymbol{H}_0) = \operatorname{rank}(\boldsymbol{H} + \boldsymbol{S}) = \min\{k, m + p\}. \tag{2.7}$$

Hence the conditions for posterior propriety can be read from Theorem 1 by replacing p with p^* and a with $a + m/2$.

2.3. *Moments of* SIW *priors and posteriors.*

2.3.1. *Existence of prior and posterior moments.* The following theorem gives some necessary and sufficient conditions for the existence of SIW prior and posterior moments.

THEOREM 2. *For the* $\mathrm{SIW}(a, \boldsymbol{H})$ *prior for* $\boldsymbol{\Sigma}$, *with* $p = \operatorname{rank}(\boldsymbol{H}) > 0$,

(a) *when* $p = k$, $E(\boldsymbol{\Sigma}^{-1})$ *exists iff* $a > 1$ *while, for any positive integer* q, $E(\boldsymbol{\Sigma}^q)$ *exists iff* $a > 1 + q$;

(b) *when* $0 < p < k$, $E(\boldsymbol{\Sigma}^{-1})$ *exists iff* $1 < a < p/2$ *while, for any positive integer* q, $E(\boldsymbol{\Sigma}^q)$ *exists iff* $1 + q < a < 1 + p/2$.

Existence results for the posterior moments are found, with probability one, by replacing p by p^* *from* (2.7) *and a by* $a + m/2$.

PROOF. For parts (a1) and (b1), it follows from Lemma 4 (see Appendix A.1) that

$$E(\boldsymbol{\Sigma}^{-1}) = \frac{\int\int \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}'|\boldsymbol{\Lambda}|^{-a}\operatorname{etr}(-\frac{1}{2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}'\boldsymbol{H}\boldsymbol{\Gamma})\,d\boldsymbol{\Gamma}\,d\boldsymbol{\Lambda}}{\int\int |\boldsymbol{\Lambda}|^{-a}\operatorname{etr}(-\frac{1}{2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}'\boldsymbol{H}\boldsymbol{\Gamma})\,d\boldsymbol{\Gamma}\,d\boldsymbol{\Lambda}}. \tag{2.8}$$

Theorem 1 gives a condition when the denominator exists. Let $\boldsymbol{E}_{hj}$ be the $k \times k$ matrix with 1 in the (h, j) entry and 0 elsewhere, and $\boldsymbol{o}_j$ be the k-dimensional vector with 1 in the jth component and 0 elsewhere, for $h, j \le k$. Then the numerator in (2.8) equals

$$\sum_{h=1}^{k}\sum_{j=1}^{k}\boldsymbol{E}_{hj}C_{hj} \quad \text{where } C_{hj} = \int\int \boldsymbol{o}_h'\boldsymbol{\Lambda}^{-1}\boldsymbol{o}_j \prod_{i=1}^{k}\frac{1}{\lambda_i^a}\operatorname{etr}\Big(-\frac{1}{2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}'\boldsymbol{H}\boldsymbol{\Gamma}\Big)\,d\boldsymbol{\Lambda}\,d\boldsymbol{\Gamma}.$$

Clearly, $\lambda_1^{-1} \le |\boldsymbol{o}_h'\boldsymbol{\Lambda}^{-1}\boldsymbol{o}_j| \le \lambda_k^{-1}$. Then

$$C_{hj} \le \int\int \lambda_k^{-(a+1)}\prod_{i=1}^{k-1}\lambda_i^{-a}\operatorname{etr}\Big(-\frac{1}{2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}'\boldsymbol{H}\boldsymbol{\Gamma}\Big)\,d\boldsymbol{\Lambda}\,d\boldsymbol{\Gamma},$$

$$C_{hj} \ge \int\int \lambda_1^{-(a+1)}\prod_{i=2}^{k}\lambda_i^{-a}\operatorname{etr}\Big(-\frac{1}{2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}'\boldsymbol{H}\boldsymbol{\Gamma}\Big)\,d\boldsymbol{\Lambda}\,d\boldsymbol{\Gamma}.$$

Proceeding as in the proof of Theorem 1, one can show that C_{hj} is finite if and only if either $(p = k, a > 1)$ or $(0 < p < k, 1 < a < 1 + p/2)$ holds. Parts (a1) and (b1) are proved. The proofs of parts (a2) and (b2) are similar. □

Recall that the constant prior, π^C, the reference prior, π^R, the modified reference prior, π^{MR}, and the uniform prior for $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$, π^U, are improper. Table 2 presents sample sizes m needed for existence of $\pi(\boldsymbol{\Sigma} \mid \boldsymbol{S})$, $E(\boldsymbol{\Sigma}^{-1} \mid \boldsymbol{S})$ and $E(\boldsymbol{\Sigma} \mid \boldsymbol{S})$ under π^C, π^R, π^{MR}, and π^U. Among the four priors, π^{MR} requires the least number of observations for the posterior to be proper; indeed, it is surprising that a single (vector) observation suffices to yield posterior propriety, since the common perception is that a $k \times k$ covariance matrix needs k observations to be "identifiable." It is equally surprising that the constant prior requires more than $2k$ observations; the suggestion is that the constant prior is way too diffuse. (See [4] for discussion.)

TABLE 2
Sample size, m, needed for the existence of $\pi(\boldsymbol{\Sigma} \mid \boldsymbol{S})$, $E(\boldsymbol{\Sigma}^{-1} \mid \boldsymbol{S})$ and $E(\boldsymbol{\Sigma} \mid \boldsymbol{S})$ under π^C, π^R, π^{MR} and π^U

Prior	Existence of $\pi(\boldsymbol{\Sigma} \mid \boldsymbol{S})$	Existence of $E(\boldsymbol{\Sigma}^{-1} \mid \boldsymbol{S})$	Existence of $E(\boldsymbol{\Sigma} \mid \boldsymbol{S})$
π^C	$m > 2k$	$m > 2k$	$m > 2k + 2$
π^R	$m \ge k$	$m \ge k$	$m \ge \max(k, 3)$
π^{MR}	$m \ge 1$	$m \ge k$	$m \ge 3$
π^U	$m \ge 3$	$m \ge \max(k, 3)$	$m \ge 5$

2.3.2. *Expressions for prior and posterior moments.* To work with the SIW priors as subjective priors, one must assess the parameters a and $\boldsymbol{H}$. As with inverse Wishart priors, this is most naturally done by subjectively specifying prior moments, and then solving for a and $\boldsymbol{H}$. We first give general expressions for the SIW prior and posterior moments (the proof in Appendix A.2), and then specialize to the important special case where $\boldsymbol{H} \propto \boldsymbol{I}_k$.

THEOREM 3. *Consider the priors* $\mathrm{SIW}(a, \boldsymbol{H})$ *for* $\boldsymbol{\Sigma}$ *with* $p = \mathrm{rank}(\boldsymbol{H}) > 0$. *Consider the eigenvalue-eigenvector decomposition* $\boldsymbol{H} = \boldsymbol{Z}\boldsymbol{\Delta}\boldsymbol{Z}'$. *For any integer* $q \ge -1$, *if* (a, p) *satisfies the conditions of Theorem* 2,

$$E(\boldsymbol{\Sigma}^q) = \boldsymbol{Z}\,\mathrm{diag}(\phi_{q,1}, \dots, \phi_{q,k})\boldsymbol{Z}', \tag{2.9}$$

where for $i = 1, \dots, k$,

$$\phi_{q,i}(a, \boldsymbol{\Delta}) = \frac{k\Gamma(a-q-1)}{2^q\Gamma(a-1)} \frac{\int t_{i1}^2 \|\bar{\boldsymbol{t}}_1\|^{2q} \prod_{j=1}^k \|\bar{\boldsymbol{t}}_j\|^{-2(a-1)}\, d\boldsymbol{T}}{\int \prod_{j=1}^k \|\bar{\boldsymbol{t}}_j\|^{-2(a-1)}\, d\boldsymbol{T}}, \tag{2.10}$$

with $\boldsymbol{T} = (t_{ij})$ *being orthogonal and* $\|\bar{\boldsymbol{t}}_j\|^2 = \sum_{h=1}^k \delta_h t_{hj}^2$, *where the* δ_h *are the diagonal elements of* $\boldsymbol{\Delta}$.

The posterior moments are found by replacing (*above*) p *by* p^* *from* (2.7), a *by* $a + m/2$, *and* $\boldsymbol{H}$ *by* $\boldsymbol{H}_0 = \boldsymbol{H} + \boldsymbol{S}$.

COROLLARY 1. *Consider the priors* $\mathrm{SIW}(a, c\boldsymbol{I}_k)$ *for* $\boldsymbol{\Sigma}$.

(a) *If* $a > 1$, $E(\boldsymbol{\Sigma}^{-1}) = \frac{2(a-1)}{c}\boldsymbol{I}_k$.
(b) *If* $a > 2$, *the first moment of* (2.1) *is* $E(\boldsymbol{\Sigma}) = \frac{c}{2(a-2)}\boldsymbol{I}_k$.
(c) *If* $a > 3$, *the second moment of* (2.1) *is* $E(\boldsymbol{\Sigma}^2) = \frac{c^2}{4(a-2)(a-3)}\boldsymbol{I}_k$.

PROOF. When $\boldsymbol{H} = c\boldsymbol{I}_k$, $\|\tilde{\boldsymbol{t}}_j\|^2 = c$ in (2.10) so, for any integer $q \ge -1$ and $a > q$, (2.10) becomes

$$\phi_{q,i} = \frac{k\Gamma(a-q-1)}{2^q\Gamma(a-1)} c^q \int t_{i1}^2\, d\boldsymbol{T} = \frac{c^q\Gamma(a-q-1)}{2^q\Gamma(a-1)}. \tag{2.11}$$

The results hold. □

2.4. *Eliciting prior parameters for* IW *and* SIW *priors.* We confine consideration to the case $\boldsymbol{H} \propto \boldsymbol{I}_k$. First, for $\boldsymbol{\Sigma} \sim \mathrm{IW}(\alpha, \beta\boldsymbol{I}_k)$, a natural way to specify α and β is to subjectively specify the mean, μ, and variance, τ^2 (smaller than μ^2), of a diagonal element of $\boldsymbol{\Sigma}$ (noting that all diagonal elements have the same distribution). Noting that the prior mean and variance of, say, σ_{11} are (for $\alpha > k+2$)

$$E[\sigma_{11}] = \frac{\beta}{2(\alpha-k-1)} \quad \text{and} \quad \mathrm{Var}[\sigma_{11}] = \frac{\beta^2}{4(\alpha-k-1)(\alpha-k-2)},$$

we equate these with the subjectively specified μ and τ^2 and solve to obtain

$$\alpha = k+2-\frac{\mu^2}{\mu^2-\tau^2} \quad \text{and} \quad \beta = 2\mu(\alpha-k-1). \tag{2.12}$$

The variance of σ_{11} from the $\mathrm{SIW}(a, c\boldsymbol{I}_k)$ prior is not readily available, but $E[\boldsymbol{\Sigma}^2]$ is available from Corollary 1, so we can equate the first and second moments of $\boldsymbol{\Sigma}$ for the $\mathrm{SIW}(a, c\boldsymbol{I}_k)$ and $\mathrm{IW}(\alpha, \beta\boldsymbol{I}_k)$ priors, and solve to obtain a and c, in terms of (2.12). The result is in the following lemma.

LEMMA 2. *$\boldsymbol{\Sigma}$ has the same first two moments for the* $\mathrm{SIW}(a, c\boldsymbol{I}_k)$ *and* $\mathrm{IW}(\alpha, \beta\boldsymbol{I}_k)$ *priors, when $\alpha > k+2$, if*

$$a = 2+\frac{(2\alpha-k-2)(\alpha-k-1)}{\alpha(k+1)-k(k+2)} \quad \text{and} \quad c = \frac{\beta(2\alpha-k-2)}{\alpha(k+1)-k(k+2)}.$$

PROOF. The first two moments of the $\mathrm{SIW}(a, c\boldsymbol{I}_k)$ prior are given in Corollary 1. The first two moments of the inverse Wishart distribution are (from Section 5 of [28])

$$E(\boldsymbol{\Sigma}) = \frac{\beta}{2(\alpha-k-1)}\boldsymbol{I}_k \quad \text{if } \alpha > k+1,$$

$$E(\boldsymbol{\Sigma}^2) = \frac{\beta^2(2\alpha-k-2)}{4(2\alpha-2k-1)(\alpha-k-1)(\alpha-k-2)}\boldsymbol{I}_k \quad \text{if } \alpha > k+2.$$

Equating the moments and solving for a and c gives the result. □

2.5. *Bayes estimation under loss functions.* The most common loss function for estimating $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Sigma}}$ is the entropy loss ([32])

$$L_1(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = \mathrm{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) - \log|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}| - k. \tag{2.13}$$

Sinha and Ghosh [31] studied the similar entropy loss,

$$L_2(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = \mathrm{tr}(\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}) - \log|\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}| - k, \tag{2.14}$$

which we utilize herein because the Bayesian estimator is the posterior mean

$$\hat{\boldsymbol{\Sigma}}_{B2} = E(\boldsymbol{\Sigma} \mid \boldsymbol{Y}). \tag{2.15}$$

A third common loss is the quadratic loss ([35]),

$$L_3(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = \mathrm{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}_k)^2. \tag{2.16}$$

In the risk analyses, all three losses gave similar results so we only present the results for L_2 in this paper; the results for L_1 and L_3 are in the Supplementary Material [5].

For the $\mathrm{SIW}(a, c\boldsymbol{I})$ prior (including π^R, π^{MR} and π^U), Theorem 3 immediately gives expressions for the Bayes estimates under L_2; just choose $q = 1$. Those expressions can also be used to prove the following lemma.

LEMMA 3. *For a* SIW$(a, c\boldsymbol{I})$ *prior, the frequentist risks* (*letting* $\hat{\boldsymbol{\Sigma}}_{Bj}$ *denote the Bayes estimator under* L_j) $R_j(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{Bj}) = E[L_j(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{Bj})]$, $j = 1, 2, 3$, *are equivariant, that is, they depend only on the eigenvalues of* $\boldsymbol{\Sigma}$.

PROOF. For any orthogonal transformation of the data, $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{\Gamma}}\boldsymbol{y}$, the loss L_j and its Bayesian estimate $\hat{\boldsymbol{\Sigma}}_{Bj}(\boldsymbol{S})$ are both equivariant, where $j = 1, 2$. In fact, it follows from (2.10) that $\tilde{\boldsymbol{\Gamma}}'\hat{\boldsymbol{\Sigma}}_{Bj}(\boldsymbol{S})\tilde{\boldsymbol{\Gamma}} = \hat{\boldsymbol{\Sigma}}_{Bj}(\tilde{\boldsymbol{\Gamma}}'\boldsymbol{S}\tilde{\boldsymbol{\Gamma}})$; also $L_j(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{Bj}(\boldsymbol{S})) = L_j(\tilde{\boldsymbol{\Gamma}}'\boldsymbol{\Sigma}\tilde{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{\Gamma}}'\hat{\boldsymbol{\Sigma}}_{Bj}(\boldsymbol{S})\boldsymbol{\Gamma})$. If one chooses $\tilde{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}$, it yields

$$\begin{aligned} R_j(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{Bj}(\boldsymbol{S})) &= E^{\boldsymbol{S}|\boldsymbol{\Sigma}}[L_j(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{Bj}(\boldsymbol{S})] = E^{\boldsymbol{S}|\boldsymbol{\Sigma}}[L_j(\boldsymbol{\Gamma}'\boldsymbol{\Sigma}\boldsymbol{\Gamma}, \boldsymbol{\Gamma}'\hat{\boldsymbol{\Sigma}}_{Bj}(\boldsymbol{S})\boldsymbol{\Gamma})] \\ &= E^{\boldsymbol{S}|\boldsymbol{\Sigma}}[L_j(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Sigma}}_{Bj}(\boldsymbol{\Gamma}'\boldsymbol{S}\boldsymbol{\Gamma})] = R_j(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Sigma}}_{Bj}(\boldsymbol{\Gamma}'\boldsymbol{S}\boldsymbol{\Gamma})). \end{aligned}$$

Since the distribution of $\boldsymbol{\Gamma}'\boldsymbol{S}\boldsymbol{\Gamma}$ depends only on $\boldsymbol{\Lambda}$, the result holds. □

The value of this lemma is that, in the extensive later simulations, it suffices to consider only diagonal covariance matrices $\boldsymbol{\Sigma}$.

3. Computation with the SIW posterior.

The posterior distribution for the SIW$(a, \boldsymbol{H})$ prior can be written

$$\pi(\boldsymbol{\Sigma} \mid \boldsymbol{Y}) \propto \frac{1}{|\prod_{i=1}^k \lambda_i|^r [\prod_{i<j}(\lambda_i - \lambda_j)]} \operatorname{etr}\left(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{H}_0\right), \tag{3.1}$$

where $r = a + m/2$ and $\boldsymbol{H}_0 = \boldsymbol{H} + \boldsymbol{S}$. In this section, methods for simulating from this posterior are discussed. Of course, the methods can also be used to simulate from the prior.

3.1. *Previously suggested sampling methods for* $\boldsymbol{\Sigma}$. *Method 1. Metropolis–Hastings Algorithm* ([19]). Let $\boldsymbol{\Sigma}_0$ be some starting point (e.g., the marginal maximum likelihood estimate or just $\boldsymbol{I}_k$). At iteration $t = 0, 1, 2, \dots,$

Step 1. Generate $\boldsymbol{\Sigma}^* \sim$ Inverse Wishart $(\frac{m+k+1}{2}, \boldsymbol{H}_0)$.

Step 2. Let λ_i^* and λ_i^t be the eigenvalues of $\boldsymbol{\Sigma}^*$ and $\boldsymbol{\Sigma}_t$, respectively. Define

$$\alpha = \min\left\{1, \prod_{i<j} \frac{\lambda_i^t - \lambda_j^t}{\lambda_i^* - \lambda_j^*} \cdot \prod_{i=1}^k \left(\frac{\lambda_i^*}{\lambda_i^t}\right)^{\frac{k+1-2a}{2}}\right\}.$$

Step 3. Let

$$\boldsymbol{\Sigma}_{t+1} = \begin{cases} \boldsymbol{\Sigma}^* & \text{with probability } \alpha, \\ \boldsymbol{\Sigma}_t & \text{otherwise.} \end{cases}$$

Method 2. Hit-and-Run (see [6, 35]). Define $\boldsymbol{\Sigma}^* = \log(\boldsymbol{\Sigma})$, or $\boldsymbol{\Sigma} = \exp(\boldsymbol{\Sigma}^*)$, in the sense that

$$\boldsymbol{\Sigma} = \sum_{i=1}^{\infty} \frac{(\boldsymbol{\Sigma}^*)^i}{i!}.$$

By Lemma 2 of [35], the posterior density of $\boldsymbol{\Sigma}^* = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^*\boldsymbol{\Gamma}'$, where $\boldsymbol{\Lambda}^* = \operatorname{diag}(\lambda_1^*, \dots, \lambda_k^*)$, $\lambda_1^* \geq \cdots \geq \lambda_k^*$, and $\boldsymbol{\Gamma}$ is orthogonal is then

$$\pi^*(\boldsymbol{\Sigma}^* \mid \boldsymbol{H}_0) \propto \frac{1}{\prod_{i<j}(\lambda_i^* - \lambda_j^*)} \operatorname{etr}\left\{-\sum_{i=1}^k (r-1)\lambda_i^* - \frac{1}{2}\boldsymbol{\Gamma}\exp(-\boldsymbol{\Lambda}^*)\boldsymbol{\Gamma}'\boldsymbol{H}_0\right\}.$$

The sampling procedure proceeds as follows:

Step 1. Select a starting p.d. matrix $\mathbf{\Sigma}_0$, set $\mathbf{\Sigma}_0^* = \log \mathbf{\Sigma}_0$ and $t = 0$.

Step 2. At iteration t, simulate a random direction symmetric matrix $\boldsymbol{U}_0 = (u_{ij})_{k\times k}$, whose elements are $u_{ij} = g_{ij}/\sqrt{\sum_{1\le l\le h\le k} g_{lh}^2}$, where $g_{lh} \overset{\text{i.i.d.}}{\sim} N(0,1)$, $1 \le l < h \le k$.

Step 3. Generate $x \sim N(0,1)$. Set $\boldsymbol{X} = \mathbf{\Sigma}_t^* + x\boldsymbol{U}_0$ and

$$\mathbf{\Sigma}_{t+1}^* = \begin{cases} \boldsymbol{X} & \text{with the probability } \min\bigl(1, \pi^*(\boldsymbol{X})/\pi^*(\mathbf{\Sigma}_t^*)\bigr), \\ \mathbf{\Sigma}_t^* & \text{otherwise.} \end{cases}$$

Step 4. Set $\mathbf{\Sigma}_{t+1} = \exp(\mathbf{\Sigma}_{t+1}^*)$.

3.2. *A new method.* The Metropolis and hit-and-run methods work only for small or moderate dimensional covariance matrices. Here, we consider a new Gibbs sampling method (drawing heavily on [21]) that has considerable promise for much higher dimensions.

From (2.4) and Lemma 4 (in the Appendix), (3.1) can be transformed to

$$\pi(\mathbf{\Lambda}, \mathbf{\Gamma} \mid \boldsymbol{H}_0) \propto \frac{1}{\prod_{i=1}^k \lambda_i^r} \operatorname{etr}\Bigl(-\frac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}'\boldsymbol{H}_0\mathbf{\Gamma}\Bigr), \tag{3.2}$$

with the understanding that the λ_i are to be ordered after they are drawn from this distribution.

3.2.1. *Simulating* $\mathbf{\Lambda}$ *given* $(\mathbf{\Gamma}, \boldsymbol{Y})$. To sample $\mathbf{\Lambda}$ from the full conditional given $\mathbf{\Gamma}$, note that

$$\frac{1}{2}\operatorname{tr}\bigl(\mathbf{\Lambda}^{-1}\mathbf{\Gamma}'\boldsymbol{H}_0\mathbf{\Gamma}\bigr) = \sum_{i=1}^k \frac{c_i}{\lambda_i},$$

where c_i is the (i,i) element of $\mathbf{\Gamma}'\boldsymbol{H}_0\mathbf{\Gamma}/2$. Thus

$$\pi(\mathbf{\Lambda} \mid \mathbf{\Gamma}, \boldsymbol{H}_0) \propto \prod_{i=1}^k \frac{1}{\lambda_i^r} e^{-c_i/\lambda_i}. \tag{3.3}$$

For given $\mathbf{\Gamma}$, we can directly sample λ_i independently from Inverse Gamma $(r-1, c_i)$. Then rearrange λ_i, so that $\lambda_1 \ge \cdots \ge \lambda_k$.

3.2.2. *Simulating* $\mathbf{\Gamma}$ *given* $(\mathbf{\Lambda}, \boldsymbol{Y})$. To sample from $\pi(\mathbf{\Gamma} \mid \mathbf{\Lambda}, \boldsymbol{H}_0)$, note that

$$\pi(\mathbf{\Gamma} \mid \mathbf{\Lambda}, \boldsymbol{H}_0) \propto \operatorname{etr}\Bigl(-\frac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}'\boldsymbol{H}_0\mathbf{\Gamma}\Bigr). \tag{3.4}$$

Here, without loss of generality, assume that $\boldsymbol{H}_0 = \operatorname{diag}(h_1, \dots, h_k)$.

Hoff [21] proposed a Gibbs sampler for simulating $\mathbf{\Gamma}$ from (3.4). His method was to randomly select two columns $i < j$ of $\mathbf{\Gamma}$, and then does a Gibbs update of the columns.

We use a slight modification of his method, namely updating the rows of $\mathbf{\Gamma}$. When m is large, there is no real difference in the speed of the methods but recall that, for SIW, m can be much less than k, in which case sampling the rows can be much faster; see Remark 2 for an explanation.

From (3.4), the conditional density of $\mathbf{\Gamma}$, given $\mathbf{\Lambda}$ and $\boldsymbol{H}_0$, can be rewritten

$$\pi(\mathbf{\Gamma} \mid \mathbf{\Lambda}; \boldsymbol{H}_0) \propto \operatorname{etr}\Bigl(-\frac{1}{2}\boldsymbol{H}_0\mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}'\Bigr). \tag{3.5}$$

To update the first two rows of $\mathbf{\Gamma}$, we write $\mathbf{\Gamma} = \operatorname{diag}(\boldsymbol{X}, \boldsymbol{I}_{k-2})(\boldsymbol{T}_{12}', \boldsymbol{T}_{-12}')'$, where $\boldsymbol{T}_{12}$ is the first 2 rows of $\mathbf{\Gamma}$, $\boldsymbol{T}_{-12}$ is the remaining $k-2$ rows of $\mathbf{\Gamma}$, and

$$\boldsymbol{X} = \boldsymbol{D}_\epsilon \boldsymbol{X}_\theta \equiv \begin{pmatrix} \epsilon_1 & 0 \\ 0 & \epsilon_2 \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Here, $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2}]$ and $\epsilon_i = \pm 1$ for $i = 1, 2$. Write $\boldsymbol{H}_1 = \mathrm{diag}(h_1, h_2)$ and $\boldsymbol{H}_2 = \mathrm{diag}(h_3, \ldots, h_k)$. Then the conditional posterior of θ is

$$\pi(\theta \mid \boldsymbol{T}_{12}, \boldsymbol{T}_{-12}, \boldsymbol{\Lambda}; \boldsymbol{H}_0)$$
$$\propto \mathrm{etr}\left\{-\frac{1}{2}\begin{pmatrix} \boldsymbol{H}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_2 \end{pmatrix}\begin{pmatrix} \boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{k-2} \end{pmatrix}\begin{pmatrix} \boldsymbol{T}_{12} \\ \boldsymbol{T}_{-12} \end{pmatrix}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}'_{12}, \boldsymbol{T}'_{-12})\begin{pmatrix} \boldsymbol{X}' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{k-2} \end{pmatrix}\right\}$$
$$\propto \mathrm{etr}\left\{-\frac{1}{2}\boldsymbol{H}_1\boldsymbol{X}\boldsymbol{T}_{12}\boldsymbol{\Lambda}^{-1}\boldsymbol{T}'_{12}\boldsymbol{X}'\right\} = \mathrm{etr}\left\{-\frac{1}{2}\boldsymbol{H}_1\boldsymbol{X}_\theta\boldsymbol{T}_{12}\boldsymbol{\Lambda}^{-1}\boldsymbol{T}'_{12}\boldsymbol{X}'_\theta\right\}.$$

Write

$$\boldsymbol{T}_{12}\boldsymbol{\Lambda}^{-1}\boldsymbol{T}'_{12} = \begin{pmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{pmatrix}\begin{pmatrix} s_1 & 0 \\ 0 & s_2 \end{pmatrix}\begin{pmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{pmatrix},$$

where $\omega \in (-\pi/2, \pi/2]$, and $s_1 > s_2$. Then the conditional posterior of θ is

$$\pi(\theta \mid \boldsymbol{T}_{12}, \boldsymbol{T}_{-12}, \boldsymbol{\Lambda}; \boldsymbol{H}_0) \propto \exp\{c_0\cos^2(\theta + \omega)\}, \quad \theta \in (-\pi/2, \pi/2], \tag{3.6}$$

where $c_0 = -\frac{1}{2}(s_1 - s_2)(h_1 - h_2) \le 0$. Let $\alpha = \cos^2(\theta + \omega)$. Then the full conditional density of α is

$$\pi(\alpha \mid \boldsymbol{O}_{-12}, \boldsymbol{D}; \boldsymbol{H}_0) \propto e^{c_0\alpha}\alpha^{-\frac{1}{2}}(1 - \alpha)^{-\frac{1}{2}}, \quad \alpha \in (0, 1). \tag{3.7}$$

As [21] discussed, sampling $\alpha \in (0, 1)$ can be done by rejection sampling, with the proposal being the Beta$(\frac{1}{2}, \frac{1}{2})$ distribution.

For updating any other i and j rows, the corresponding conditional density of θ has a similar formula as in (3.6), with $c_0 = -\frac{1}{2}(s_1 - s_2)(h_i - h_j)$ and (s_1, s_2) being the eigenvalues of $\boldsymbol{T}_{ij}\boldsymbol{\Lambda}^{-1}\boldsymbol{T}'_{ij}$, and $\boldsymbol{T}_{ij}$ consists of the (i, j)th rows of $\boldsymbol{\Gamma}$.

REMARK 2. When $p \equiv \mathrm{rank}(\boldsymbol{H}_0) < k$, $h_{p+1} = \cdots = h_k = 0$. To update (i, j) rows with $i \le p$ and $j > p$, the conditional density of θ is of the form (3.6) with $c_0 = -\frac{1}{2}(s_1 - s_2)h_i$. Furthermore, to update (i, j) rows with $i > p$ and $j > p$, the conditional density of θ is then simply Uniform$[-\pi/2, \pi/2]$! Therefore, updating the rows of $\boldsymbol{\Gamma}$ are more efficient than Hoff's method of updating the columns of $\boldsymbol{\Gamma}$, when $p \ll k$. This will be the case, for the reference prior and modified reference prior of $\boldsymbol{\Sigma}$ and the uniform prior for $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$, when m is small compared to k.

3.3. *Comparing the three sampling methods.* To compare the three simulation methods, we choose $\boldsymbol{\Sigma}_0$ and obtain a sample $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$ from $N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$. Then we simulate $\boldsymbol{\Sigma}$ from the posterior under the modified reference prior, $\pi^{\mathrm{MR}}(\boldsymbol{\Sigma})$.

Instead of looking at the convergence of the posterior for various components of $\boldsymbol{\Sigma}$, we monitor the convergence of $L_2(\hat{\boldsymbol{\Sigma}}_{B1}, \boldsymbol{\Sigma}_0)$, namely the entropy loss in (2.13) evaluated at the Bayes estimate, $\hat{\boldsymbol{\Sigma}}_{B2}$, computed by simulation. This provides an overall assessment of convergence of the simulation. Convergence was judged using the criterion in [14].

The following four cases of (k, m) and $\boldsymbol{\Sigma}_0$ are considered:

Case I: $(k, m) = (5, 15)$ and $\boldsymbol{\Sigma}_0 = \mathrm{diag}(16, 8, 4, 2, 1)$. The observed $\boldsymbol{S}$ has the eigenvalues $(286, 223, 39, 16, 15)$.

Case II: $(k, m) = (10, 40)$ and $\boldsymbol{\Sigma}_0 = \mathrm{diag}(512, 256, 128, 64, 32, 16, 8, 4, 2, 1)$. The observed $\boldsymbol{S}$ has the eigenvalues (15,255, 11,170, 5185, 3447, 1085, 577, 159, 128, 49, 43).

Case III: $(k, m) = (50, 100)$ and $\boldsymbol{\Sigma}_0 = \mathrm{diag}(50, \ldots, 2, 1)$. The first and last five eigenvalues of $\boldsymbol{S}$ are $(8083, 7903, 7104, 6871, 6352)$ and $(207, 181, 131, 98, 60)$, respectively.

Case IV: $(k, m) = (100, 300)$ and $\boldsymbol{\Sigma}_0 = \mathrm{diag}(100, \ldots, 2, 1)$. The first and last five eigenvalues of $\boldsymbol{S}$ are (46,293, 45,558, 44,045, 42,976, 41,887) and $(827, 684, 529, 440, 221)$, respectively.

TABLE 3
Comparison of computing time for the three methods

	Method	Time (seconds) for 10^3 cirles	# of iterations to convergence	Total time (seconds)
Case I	Metropolis	1.61	2.0×10^7	3.22×10^4
($k = 5$)	Hit-and-Run	3.01	2.0×10^7	6.02×10^4
	New Method	2.88	1.5×10^5	4.32×10^2
Case II	Metropolis	1.79	4.0×10^7	7.16×10^4
($k = 10$)	Hit-and-Run	4.73	3.5×10^7	1.65×10^5
	New Method	5.85	7.0×10^5	4.09×10^3
Case III	Metropolis	4.68	stop at 1.5×10^8	$>7.02 \times 10^5$
($k = 50$)	Hit-and-Run	18.20	4.0×10^7	7.28×10^5
	New Method	43.96	1.0×10^6	4.39×10^4
Case IV	Metropolis	23.84	stop at 1.2×10^8	$>2.86 \times 10^6$
($k = 100$)	Hit-and-Run	80.42	stop at 1.2×10^8	$>9.65 \times 10^6$
	New Method	262.03	1.5×10^6	3.93×10^5

The results are given in Table 3. While the new method can be substantially more expensive per iteration, it requires many fewer iterations for convergence (i.e., mixes much better), so its overall computational time is less. For the $k = 5$ case, the new method was 1000 times faster. But its real benefit was in the high dimensional cases: for $k = 50$, Metropolis simply failed to converge, and both Metropolis and Hit-and-Run failed for $k = 100$.

The story is told, perhaps even more clearly, by looking at the trace plots of $L_2(\hat{\mathbf{\Sigma}}_{B1}, \mathbf{\Sigma}_0)$, for the three methods and the four cases; these are plotted in Figure 1. The much faster convergence of the new method (blue curves) is clear, as is the poor performance of Metropolis (black curves) and its utter failure in higher dimensions. Hit-and-Run (red curves) does better, converging for $k = 50$ and getting reasonably close for $k = 100$; but a much longer running time would be needed for actual convergence.

4. Comparing the IW and SIW priors and posteriors.

In this section, we compare the IW and SIW priors and posteriors. In all sections, the two priors will have been matched to two moments, using Lemma 2.

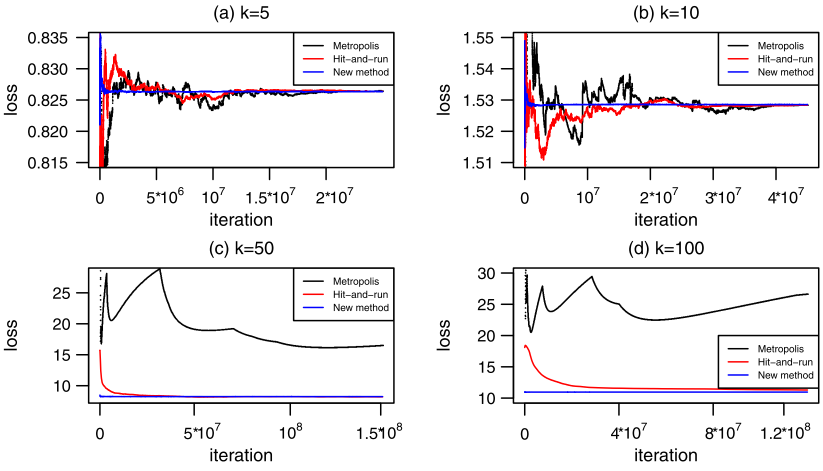

FIG. 1. *Trace plots of the three computational algorithms for* $k = 5, 10, 50, 100$.

In Section 4.1, we look at contours of the largest and smallest eigenvalues of $\mathbf{\Sigma}$, from the prior and posterior distributions. The point is to see if the eigenvalues are, indeed, considerably more spread for the IW priors and posteriors than for those of SIW.

In Section 4.2, we compare Bayes risks arising in estimating $\mathbf{\Sigma}$, using either IW or SIW as the true parameter-generating prior, and then doing the Bayesian analysis under both priors. Of course, the Bayesian analysis using the true prior will be optimal, but it is instructive to see how much worse the results are under the other prior.

In Section 4.3, we consider the usual $m > k$ situation and compare the frequentist risks of IW and SIW in a variety of situations. We also include in the risk comparison the other priors that were earlier discussed.

In Section 4.4, we investigate the $m < k$ scenario, calling this *low rank learning*. That many fewer observations than the dimension of the covariance matrix can result in proper posteriors was a surprise, and studying the resulting posteriors is of considerable interest; can we learn anything useful about $\mathbf{\Sigma}$ in this situation and do certain priors result in a better job of low rank learning than others?

In all four comparisons, SIW does considerably better than IW. This is—at the same time—puzzling and expected. It is puzzling because the IW priors that are considered are proper, and hence, cannot uniformly be improved upon. Yet we seem to be unable to construct a scenario in which they do better (except that of generating from the IW prior and using it for analysis). But this is perhaps expected, in that we began the paper by saying that the eigenvalue inflation property of IW priors is counterintuitive, and our inability to create scenarios where the IW priors are better reflects this.

4.1. *Comparing eigenvalue contours for the priors and posteriors.* It is of interest to see the extent to which IW priors spread eigenvalues apart more so than do SIW priors. One way to look at this is to match the two priors via Lemma 2, and then examine contour plots of the resulting largest and smallest eigenvalues, λ_1 and λ_k.

For the SIW$(a, c\boldsymbol{I}_k)$ prior, let $f(x)$ and $F(x)$ be the probability density and cumulative distribution function of IG$(a-1, c/2)$, respectively. It follows from Remark 1 that the joint density of (λ_1, λ_k) is of the form

$$\pi(\lambda_1, \lambda_k) = k(k-1)\big[F(\lambda_1) - F(\lambda_k)\big]^{k-2} f(\lambda_1) f(\lambda_k), \tag{4.1}$$

where $\lambda_1 > \lambda_k > 0$. For the comparison, we chose

$$E(\mathbf{\Sigma}) = \boldsymbol{I}_k \quad \text{and} \quad E(\mathbf{\Sigma}^2) = 2\boldsymbol{I}_k. \tag{4.2}$$

From Corollary 1, we get, as the matching SIW parameters, $(a, c) = (4, 4)$ (regardless of the dimension k). The corresponding matching IW parameters (solving in Lemma 2) are

$$\alpha = \frac{3}{2} + \frac{5}{4}k + \frac{1}{4}\sqrt{(2+k)^2 + 16} \quad \text{and} \quad \beta = 2(\alpha - k - 1). \tag{4.3}$$

We consider both $k = 5$ and $k = 50$ dimensional covariance matrices, for which, respectively,

$$(\alpha, \beta) = (9.7656, 7.5311) \quad \text{and} \quad (\alpha, \beta) = (77.0384, 52.0768). \tag{4.4}$$

Figure 2 shows the contour plots of the moment matched IW$(\alpha, \beta\boldsymbol{I}_k)$ and SIW$(a, c\boldsymbol{I}_k)$ prior densities of (λ_1, λ_k), for $k = 5$ and 50. The contours for SIW$(a, c\boldsymbol{I}_k)$ are based on (4.1), while those for IW$(\alpha, \beta\boldsymbol{I}_k)$ are based on 100,000 simulated values of $\mathbf{\Sigma}$ from the prior. For $k = 5$ (the left two plots), it is clear that the smallest eigenvalue, λ_k, is typically much smaller for IW than for SIW. The largest eigenvalue for IW is clearly typically larger than that for SIW when $k = 5$. For $k = 50$ (the right two plots), the situation is somewhat muddied, because apparently the SIW prior has fatter tails that the IW prior; but note that for, say, the

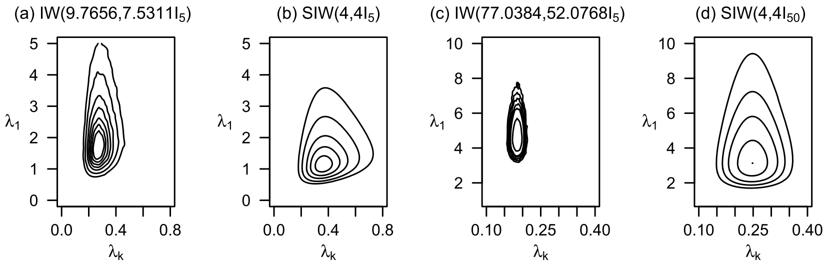

FIG. 2. *Contour plots of moment matched* $\mathrm{IW}(\alpha, \beta \boldsymbol{I}_k)$ *and* $\mathrm{SIW}(a, c\boldsymbol{I}_k)$ *prior densities of* (λ_1, λ_k), *when* $E(\boldsymbol{\Sigma}) = \boldsymbol{I}_k$ *and* $E(\boldsymbol{\Sigma}^2) = 2\boldsymbol{I}_k$ *for* $k = 5$ (*left two plots*) *and* $k = 50$ (*right two plots*).

central level 2 contour, it is clear that the IW largest eigenvalue tends to be much larger than that from SIW. So the figure clearly supports the presumption that IW will force the eigenvalues apart more so than will SIW.

Next, we look at posterior contour plots for the above priors. For economy of space, we only present the $k = 50$ results; the results for $k = 5$ exhibited the same pattern and are available in the Supplementary Material [5]. The sample sizes used are $m = 50$ (first and fourth columns) and $m = 200$ (second and third columns). To obtain the posteriors, the data was generated from $\boldsymbol{I}_{50}$ (compatible with the prior distributions) and $\boldsymbol{\Sigma}_{50} = \mathrm{diag}(50, \dots, 1)$ (incompatible with the priors) by sampling $\boldsymbol{S}$ from $\mathrm{Wishart}_{50}(m, \boldsymbol{I}_k)$ and $\mathrm{Wishart}_{50}(m, \boldsymbol{\Sigma}_k)$. For each of the cases, we simulate 10^6 values of $\boldsymbol{\Sigma}$ from the posteriors, $\pi(\boldsymbol{\Sigma} \mid \boldsymbol{S})$, under the two priors.

Figure 3 presents the contour plots of posterior densities of $(\lambda_1, \lambda_{50})$ based on these 10^6 points. The first row of each table presents the results for IW; the second row, those for SIW. The first two columns are with $\boldsymbol{I}_{50}$; the last two are with $\boldsymbol{\Sigma}_{50}$. These figures clearly show the expected pattern: λ_1 tends to be much larger—and λ_{50} much smaller—for the IW posteriors than for the SIW posteriors.

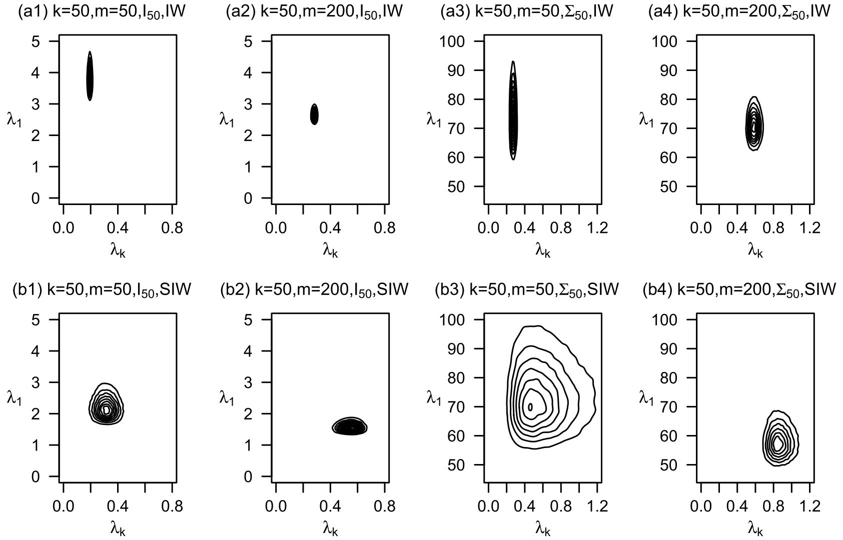

FIG. 3. *Posterior contour plots for* (λ_1, λ_k) *of moment matched* $\mathrm{IW}(\alpha, \beta \boldsymbol{I}_{50})$ (*top*) *and* $\mathrm{SIW}(a, c\boldsymbol{I}_{50})$ (*bottom*) *prior densities*, *when* $E(\boldsymbol{\Sigma}) = \boldsymbol{I}_{50}$ *and* $E(\boldsymbol{\Sigma}^2) = 2\boldsymbol{I}_{50}$. *Here*, $k = 50$, $m = 50$ *or* $m = 200$, *and the true* $\boldsymbol{\Sigma}$ *was either* $\boldsymbol{I}_{50}$ *or* $\boldsymbol{\Sigma}_{50} = \mathrm{diag}(50, \dots, 1)$.

4.2. *Comparing Bayes risks under* IW *and* SIW *priors*. We choose $k = 5, 20$, $m = 2k + 3, 5k$. For the $\mathrm{IW}_k(\alpha, \beta \boldsymbol{I}_k)$ and $\mathrm{SIW}(a, c\boldsymbol{I}_k)$ priors, we match

$$E(\boldsymbol{\Sigma}) = \boldsymbol{I}_k \quad \text{and} \quad E(\boldsymbol{\Sigma}^2) = 3\boldsymbol{I}_k. \tag{4.5}$$

Corollary 1 yields $(a, c) = (3.5, 3)$ and, solving in (Lemma 2), yields $(\alpha, \beta) = (8.3228, 4.6457)$ for $k = 5$ and $(\alpha, \beta) = (26.8776, 11.7552)$ for $k = 20$.

To compute the Bayes risks, we repeatedly draw $\boldsymbol{\Sigma}$ from the true prior (either IW or SIW), simulate $\boldsymbol{S} \mid \boldsymbol{\Sigma} \sim \mathrm{Wishart}_k(m, \boldsymbol{\Sigma})$, compute the Bayes estimators for both priors for the loss L_2, from the expressions in Section 2.5 and finally compute the actual losses of the Bayes estimators. This is repeated 10,000 times, and the losses averaged to obtain an estimate of the Bayes risks, which are denoted $r_2(\pi_T, \hat{\boldsymbol{\Sigma}}_{B2}^{\pi})$, π_T being the true prior, π_W being the other prior, and π being the prior used to compute the Bayes estimate $\hat{\boldsymbol{\Sigma}}_{B2}^{\pi}$ under that loss. Of course, the smallest risks are obtained when the true prior is used to compute the Bayes estimates, but it is useful to look at the ratio

$$\frac{r_2(\pi_T, \hat{\boldsymbol{\Sigma}}_{B2}^{\pi_W})}{r_2(\pi_T, \hat{\boldsymbol{\Sigma}}_{B2}^{\pi_T})},$$

where the denominator is this optimal risk for the true prior and the numerator is the Bayes risk when the wrong prior is used.

These ratios are presented in Table 4. The first entry, 1.133, shows that the SIW prior's performance is 13.3% worse than that of the IW prior, when the IW prior is the true prior. The interesting feature of this table is that the IW performance, when SIW is true, is considerably worse than the SIW performance, when IW is true. The most extreme case is when $k = 20$ and $m = 43$, in which case the SIW risk is only 7.5% worse when IW is the true prior, but the IW risk is 44.8% worse when SIW is the true prior. This asymmetry strongly suggests that SIW is the more robust prior.

4.3. *Comparing risk functions for* $m \geq k$. We compare the frequentist risk (expected loss), $R_2(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{B2})$, of the Bayes estimates under the seven priors, $\pi^R, \pi^{\mathrm{MR}}, \pi^U, \pi^J, \pi^C$ and $\mathrm{IW}_k(\alpha, \beta \boldsymbol{I}_k)$ and $\mathrm{SIW}(a, c\boldsymbol{I}_k)$ as in the previous section, under loss L_2 when $m \geq k$. The results for losses L_1 and L_3 exhibit the same pattern, and are available in the Supplementary Material [5]. We also consider the risk function of the best equivariant estimator, $\hat{\boldsymbol{\Sigma}}_{E2}$, of $\boldsymbol{\Sigma}$. From [10], the best equivariant estimator of $\boldsymbol{\Sigma}$, under the loss function L_2, utilizing the lower triangular Cholesky decomposition $\boldsymbol{S} = \boldsymbol{K}\boldsymbol{K}'$, is $\hat{\boldsymbol{\Sigma}}_{E2} = \boldsymbol{K}\boldsymbol{\Lambda}_{E2}\boldsymbol{K}'$, where $\boldsymbol{\Lambda}_{E2}$ is a diagonal matris with elements $\lambda_{2i} = (m-1)/[(m-i-1)(m-i)]$, $i = 1, \ldots, k$.

TABLE 4
Comparison of Bayes risks for matched SIW *and* IW *priors* (*matching to* $E(\boldsymbol{\Sigma}) = \boldsymbol{I}_k$, $E(\boldsymbol{\Sigma}^2) = 3\boldsymbol{I}_k$), *using first one and then the other as the truth, under* L_2

		$k = 5$		$k = 20$	
		$(\alpha, \beta) = (8.322, 4.645)$		$(\alpha, \beta) = (26.877, 11.755)$	
Loss	Bayes risk ratio	$m = 13$	$m = 25$	$m = 43$	$m = 100$
L_2	$\frac{r_2(\mathrm{IW}, \hat{\boldsymbol{\Sigma}}_{B2}^{\mathrm{SIW}})}{r_2(\mathrm{IW}, \hat{\boldsymbol{\Sigma}}_{B2}^{\mathrm{IW}})}$	1.133	1.109	1.075	1.033
	$\frac{r_1(\mathrm{SIW}, \hat{\boldsymbol{\Sigma}}_{B2}^{\mathrm{IW}})}{r_1(\mathrm{SIW}, \hat{\boldsymbol{\Sigma}}_{B2}^{\mathrm{SIW}})}$	1.254	1.243	1.448	1.316

TABLE 5
Risks (*expected losses under* L_2) *of Bayes estimates under the indicated priors and the equivariant estimator* E, *for* $k = 5, 20$, $m = 2k + 3, 5k, 10k$, $\boldsymbol{\Sigma} = \boldsymbol{I}_k$, $\boldsymbol{\Sigma}_{k1} = \mathrm{diag}(8k - 7, \dots, 9, 1)$, *and* $\boldsymbol{\Sigma}_{k2} = \mathrm{diag}(\lfloor\frac{k+1}{2}\rfloor, \lfloor\frac{k+1}{2}\rfloor - 1, \dots, 1, \frac{1}{2}, \dots, \lfloor\frac{k+1}{2}\rfloor^{-1})$

Loss	k	$(m, \boldsymbol{\Sigma})$	π^{IW}	π^{SIW}	π^{R}	π^{MR}	π^{U}	π^{J}	π^{C}	E
L_2	5	$(13, \boldsymbol{I}_5)$	0.78	0.23	0.48	0.47	0.50	1.74	7.19	1.51
		$(25, \boldsymbol{I}_5)$	0.54	0.16	0.23	0.22	0.23	0.72	1.04	0.68
		$(50, \boldsymbol{I}_5)$	0.30	0.09	0.10	0.10	0.11	0.33	0.38	0.32
		$(13, \boldsymbol{\Sigma}_{k1})$	3.13	1.51	1.12	1.11	1.24	1.75	7.19	1.51
		$(25, \boldsymbol{\Sigma}_{k1})$	1.07	0.64	0.55	0.54	0.57	0.73	1.04	0.68
		$(50, \boldsymbol{\Sigma}_{k1})$	0.41	0.30	0.28	0.28	0.29	0.33	0.38	0.32
		$(13, \boldsymbol{\Sigma}_{k2})$	0.94	0.85	1.11	1.09	1.06	1.76	7.19	1.51
		$(25, \boldsymbol{\Sigma}_{k2})$	0.55	0.51	0.59	0.58	0.57	0.73	1.05	0.68
		$(50, \boldsymbol{\Sigma}_{k2})$	0.29	0.27	0.30	0.29	0.29	0.33	0.38	0.32
		$(43, \boldsymbol{I}_{20})$	4.27	0.39	0.56	0.55	0.50	7.48	5.20	6.10
		$(100, \boldsymbol{I}_{20})$	2.17	0.19	0.21	0.21	0.20	2.45	3.32	2.29
		$(200, \boldsymbol{I}_{20})$	1.09	0.10	0.10	0.10	0.09	1.13	1.28	1.09
	20	$(43, \boldsymbol{\Sigma}_{k1})$	16.99	3.61	3.25	3.24	3.24	7.49	5.22	6.10
		$(100, \boldsymbol{\Sigma}_{k1})$	3.66	1.63	1.60	1.60	1.60	2.45	3.32	2.29
		$(200, \boldsymbol{\Sigma}_{k1})$	1.40	0.89	0.88	0.88	0.88	1.13	1.28	1.09
		$(43, \boldsymbol{\Sigma}_{k2})$	5.67	4.85	5.14	5.14	4.93	7.48	20.19	6.10
		$(100, \boldsymbol{\Sigma}_{k2})$	2.33	2.06	2.12	2.11	2.09	2.46	3.32	2.29
		$(200, \boldsymbol{\Sigma}_{k2})$	1.12	1.03	1.05	1.04	1.04	1.13	1.28	1.09

We chose $k = 5, 20$, $m = 2k + 3, 5k, 10k$, and evaluated the risks at the three covariance matrices $\boldsymbol{\Sigma} = \boldsymbol{I}_k$, $\boldsymbol{\Sigma}_{k1} = \mathrm{diag}(8k - 7, \dots, 9, 1)$ and $\boldsymbol{\Sigma}_{k2} = \mathrm{diag}(\lfloor\frac{k+1}{2}\rfloor, \lfloor\frac{k+1}{2}\rfloor - 1, \dots, 1, \frac{1}{2}, \dots, \frac{1}{\lfloor\frac{k+1}{2}\rfloor})$, where $\lfloor\cdot\rfloor$ is a floor function. The identity covariance matrix is completely compatible with the IW and SIW priors, but might be thought to favor SIW, since the eigenvalues are the same. The second covariance matrix is completely incompatible with the IW and SIW priors, and serves to measure the robustness of the priors to misspecification of their inputs. The third covariance matrix is reasonably compatible with the IW and SIW priors, but has spread eigenvalues and so was thought to perhaps favor the IW prior.

The risks of the seven Bayesian estimators, together with that of $\hat{\boldsymbol{\Sigma}}_{E2}$, are given in Table 5. They were computed by averaging the losses over 3000 draws of $\boldsymbol{S} \sim \mathrm{Wishart}_k(m, \boldsymbol{\Sigma})$ and using $2 * 10^5$ posterior draws to compute $\hat{\boldsymbol{\Sigma}}_{B2}$. Some observations from the table:

- SIW always has smaller risk than IW—often by a large margin—even for $\boldsymbol{\Sigma}_{k2}$, which was included as a guess of a covariance matrix that would be better for IW.
- The constant prior is almost always the worst; this is yet another warning that this commonly used prior is problematical.
- Of the objective priors, π^{MR} and π^{R} are very close, with π^{MR} being slightly better. But, surprisingly, π^{U} is a strong competitor, being modestly better and worse than π^{MR}, in roughly equal proportions.
- The Jeffreys prior was decidedly inferior.
- While SIW could be expected to be optimal (and was) in estimating the identity matrix, its strong performance for the other two covariance matrices was unexpected. Especially surprising were the results for $\boldsymbol{\Sigma}_{k1}$, which was far out in the tails of the SIW prior, and yet the risks using SIW were only moderately higher than those using π^{MR}.

4.4. *Low rank learning.* When the sample size m is smaller than k, we saw that the posterior distributions of $\boldsymbol{\Sigma}$ could be proper under the priors $\mathrm{IW}_k(\alpha, \beta\boldsymbol{I}_k)$, $\mathrm{SIW}(a, c\boldsymbol{I}_k)$, π^{MR}

and π^U. Clearly, such posteriors cannot illuminate all of $\boldsymbol{\Sigma}$ (we know that at least m observations are needed to "identify" $\boldsymbol{\Sigma}$), but the posteriors might be able to illuminate some features of $\boldsymbol{\Sigma}$. This is explored in Section 4.4.1.

The low rank case is also an interesting domain in which to investigate estimation risks of the various priors, as low rank might be expected to exacerbate problems with a prior. This is studied in Section 4.4.2.

4.4.1. *Posterior distributions of features of* $\boldsymbol{\Sigma}$. We consider here the largest eigenvalue, λ_1, and the trace, $\mathrm{tr}(\boldsymbol{\Sigma})$; these were chosen as often being of interest and because they were representative of two extremes involving low rank learning. To challenge the possibility of learning features of the posteriors under the four priors, we chose $k = 100$ with much smaller sample sizes. For the $\mathrm{IW}_k(\alpha, \beta \boldsymbol{I}_k)$ and $\mathrm{SIW}(a, c\boldsymbol{I}_k)$ priors, we follow (4.5) in Section 4.3, and set $(a, c) = (3.5, 3)$ and (solving in Lemma 2) $(\alpha, \beta) = (126.78, 51.56)$.

For the true $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_k$ (compatible with IW and SIW) and $\boldsymbol{\Sigma}_k = \mathrm{diag}(k, \dots, 1)$ (not compatible with IW and SIW), we sample $\boldsymbol{S}$ from $\mathrm{Wishart}_k(m, \boldsymbol{\Sigma}_j)$, for $j = 1, 2$ and $m = 5, 20, 100$. Then we generate $2 * 10^5$ posterior samples of $\boldsymbol{\Sigma}$ given $\boldsymbol{S}$ under the IW, SIW, π^{MR} and π^U priors. Table 6 summarizes the corresponding posterior means and standard deviations of λ_1 and $\mathrm{tr}(\boldsymbol{\Sigma})$, under the four priors.

None of the posteriors for λ_1 are accurate, in the sense of the mean being close to the true value and the mean plus or minus two standard deviations covering the true value. For $\mathrm{tr}(\boldsymbol{\Sigma})$, however, many of the posteriors are reasonably accurate especially those from SIW, covering the true value in four of the six cases, and only missing moderately in the other two cases.

We repeated this exercise with numerous other features of $\boldsymbol{\Sigma}$, and the above results seemed to generalize. One cannot use low rank learning effectively with individual elements of $\boldsymbol{\Sigma}$—such as variances, covariances or eigenvalues—but overall properties of $\boldsymbol{\Sigma}$—such as the trace or determinant—are approachable with low rank learning.

4.4.2. *Comparison of estimation risks under loss* L_2. We next compare the risks, $R_2(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{B2})$, in lower rank learning under loss L_2 under the five priors $\mathrm{IW}_k(\alpha, \beta \boldsymbol{I}_k)$, $\mathrm{SIW}(3.5, 3\boldsymbol{I}_k)$, π^R, π^{MR} and π^U. We fix $m = 5$ and choose $k = 5$ (for which the matching

TABLE 6
Posterior means and standard deviations of λ_1 *and* $\mathrm{tr}(\boldsymbol{\Sigma})$ *under* $\mathrm{IW}(126.78, 51.56\boldsymbol{I}_k)$, $\mathrm{SIW}(3.5, 3\boldsymbol{I}_k)$ π^{MR} *and* π^U *priors, for* $k = 100$, $m = 5, 20, 100$ *and* $\boldsymbol{\Sigma} = \boldsymbol{I}_k$ *and* $\boldsymbol{\Sigma}_k = \mathrm{diag}(k, k-1, \dots, 1)$

			IW		SIW		π^{MR}		π^U	
$\boldsymbol{\Sigma}$	Feature of $\boldsymbol{\Sigma}$	m	Mean	sd	Mean	sd	Mean	sd	Mean	sd
$\boldsymbol{I}_k$	$\lambda_1 = 1$	5	5.2	0.47	5.9	2.7	22.0	23.7	147.6	1112
		20	5.0	0.43	4.2	1.1	7.5	1.9	6.4	1.9
		100	4.1	0.27	2.2	0.3	2.3	0.3	2.4	0.3
	$\mathrm{tr}(\boldsymbol{\Sigma}) = 100$	5	100.2	1.93	101.8	6.4	172.8	33.6	517	1122
		20	99.9	1.80	99.7	3.5	110.7	6.4	124	6.1
		100	99.4	1.31	98.9	1.5	100.9	1.6	103	1.6
$\boldsymbol{\Sigma}_k$	$\lambda_1 = 100$	5	80	9.55	997	438	2339	2470	7567	40,000
		20	216	28.78	401	87	395	105	342	100
		100	161	12.39	167	24	173	25	176	26
	$\mathrm{tr}(\boldsymbol{\Sigma}) = 5050$	5	344	16.93	3015	711	8923	3160	26,668	41,231
		20	1511	68.60	4491	301	5726	339	6437	318
		100	2509	42.21	4781	82	5070	89	5175	90

TABLE 7

Risks (expected losses using L_2) for π^{IW}, π^{SIW}, π^{MR}, π^{R} and π^{U}, when $m = 5$, $k = 5, 20$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_k$ or $\boldsymbol{\Sigma}_{1k} \equiv \mathrm{diag}(8k - 7, \ldots, 9, 1)$

$(k, \boldsymbol{\Sigma})$	π^{IW}	π^{SIW}	π^{R}	π^{MR}	π^{U}
$(5, \boldsymbol{I}_5)$	0.7457	0.2422	29.6682	3.7089	3.6429
$(5, \boldsymbol{\Sigma}_{1k})$	6.6057	2.3119	39.3877	5.9465	5.5438
$(20, \boldsymbol{I}_{20})$	1.9201	0.4720	NA	6.9858	15.3567
$(20, \boldsymbol{\Sigma}_{1k})$	157.7758	37.0476	NA	12.9270	22.6516

$(\alpha, \beta) = (8.3228, 4.6457)$) and $k = 20$ ($(\alpha, \beta) = (26.8776, 11.755)$) and $\boldsymbol{\Sigma} = \boldsymbol{I}_k$ (compatible with the prior information) and $\boldsymbol{\Sigma}_{1k} = \mathrm{diag}(8k - 7, \ldots, 9, 1)$ (not compatible).

The risks of the Bayes estimators are given in Table 7 and were obtained by averaging the expected losses over 3000 draws of $\boldsymbol{S} \sim \mathrm{Wishart}_k(m, \boldsymbol{\Sigma})$, and utilizing $2 * 10^5$ posterior draws, for each given $\boldsymbol{S}$, to compute the Bayes estimate. The performance of SIW continues to impress, soundly beating IW, and even beating the objective priors, except when $k = 20$ and $\boldsymbol{\Sigma}_{1k}$ (a matrix far in the tail of the SIW prior) is the true covariance matrix.

5. Generalizations. While we were just considering the vanilla covariance matrix problem here, there have been many generalizations of the vanilla IW prior to structured IW priors, especially in high dimensions ([29] and [27] being two examples). A major difficulty in similar extensions of the SIW prior is that marginal and some conditional distributions are considerably more difficult to work with for the SIW priors; in particular, the marginal distribution of a diagonal block of $\boldsymbol{\Sigma}$ does not have a SIW distribution (a diagonal block of an IW distribution is IW).

At one level, generalizing a highly structured IW prior to a highly structured SIW prior is trivial; just replace every IW component in the IW prior with a SIW component. But important interconnections between the IW components or important update considerations could be destroyed by this. Thus one would need to go through each structured IW scenario carefully, determining the extent to which IW could be replaced by SIW. Such an exploration is beyond the scope of this paper.

There is also much more work that could be done involving low rank learning. The key, however, is finding a structured way to investigate the problem. For instance, one could consider the $k > m$ scenario, with m growing, but the focus now being on low rank learning without sparsity assumptions. Perhaps consistency results (e.g., for the trace of $\boldsymbol{\Sigma}$) are available.

APPENDIX

A.1. Proofs of Theorem 1 and Lemma 1. We will need the following lemmas.

LEMMA 4. *Let $\boldsymbol{R} = (r_{ij})$ be $I \times J$ random matrix, whose element r_{ij} is a function of $\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}'$ with final integral with respect to $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$. Then*

$$\int\int \boldsymbol{R} 1_{\{\lambda_1 > \cdots > \lambda_k\}}\, d\boldsymbol{\Lambda}\, d\boldsymbol{\Gamma} = \frac{1}{k!}\int\int \boldsymbol{R}\, d\boldsymbol{\Lambda}\, d\boldsymbol{\Gamma}.$$

PROOF. For any given (i, j), by Lemma 3.4 in [2],

$$\int\int r_{ij} 1_{\{\lambda_1 > \cdots > \lambda_k\}}\, d\boldsymbol{\Lambda}\, d\boldsymbol{\Gamma} = \frac{1}{k!}\int\int r_{ij}\, d\boldsymbol{\Lambda}\, d\boldsymbol{\Gamma}.$$

The proof is complete. □

LEMMA 5. *Let X be a random variable and f_i be a nonnegative function satisfying $\sum_i^k f_i(x) = C$ for any $x \in \Omega$. If all $g_i's$ are monotone increasing or decreasing on* $[0, C]$,

$$E\left(\prod_{i=1}^{k} g_i(f_i(X))\right) \le \prod_{i=1}^{k} E(g_i(f_i(X))). \tag{A.1}$$

PROOF. It is enough to show the result for $k = 2$. In fact, for any x_1 and $x_2 \in \mathbf{\Omega}$, we have $[g_1(f_1(x_1)) - g_1(f_1(x_2))][g_2(f_2(x_1)) - g_2(f_2(x_2))] \le 0$. Therefore,$E[g_1(f_1(X_1)) - g_1(f_1(X_2))][g_2(f_2(X_1)) - g_2(f_2(X_2))] \le 0$, which implies that

$$E[g_1(f_1(X))g_2(f_2(X))] \le E[g_1(f_1(X))]E[g_2(f_2(X))]. \tag{A.2}$$

This proves the result for $k = 2$. □

Proof of Theorem 1. First, we consider the sufficient conditions of parts (a) and (b). From (2.4), we get

$$\int \pi^{\mathrm{SIW}}(\mathbf{\Sigma})\,d\mathbf{\Sigma} = \int\int |\mathbf{\Lambda}|^{-a} \operatorname{etr}\left(-\frac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}'\boldsymbol{H}\mathbf{\Gamma}\right) 1_{\{\lambda_1 > \cdots > \lambda_k\}}\,d\mathbf{\Lambda}\,d\mathbf{\Gamma}.$$

It follows from Lemma 4 that

$$\int \pi^{\mathrm{SIW}}(\mathbf{\Sigma})\,d\mathbf{\Sigma} = \frac{1}{k!}\int\int |\mathbf{\Lambda}|^{-a} \operatorname{etr}\left(-\frac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}'\boldsymbol{H}\mathbf{\Gamma}\right) d\mathbf{\Lambda}\,d\mathbf{\Gamma}.$$

Write $\boldsymbol{H} = \boldsymbol{Z}\mathbf{\Delta}\boldsymbol{Z}'$, where $\boldsymbol{Z}$ is an orthogonal matrix and $\mathbf{\Delta} = \operatorname{diag}\{\delta_1, \ldots, \delta_p, 0, \ldots, 0\}$ and $\delta_1 \ge \cdots \ge \delta_p > 0$. We define $\boldsymbol{T} = (t_{ij}) = \mathbf{\Gamma}'\boldsymbol{Z}$. Clearly, $\operatorname{tr}(\mathbf{\Lambda}^{-1}\mathbf{\Gamma}'\boldsymbol{H}\mathbf{\Gamma}) = \operatorname{tr}(\mathbf{\Lambda}^{-1}\mathbf{\Gamma}'\boldsymbol{Z}\mathbf{\Delta}\boldsymbol{Z}'\mathbf{\Gamma}) = \operatorname{tr}(\mathbf{\Lambda}^{-1}\boldsymbol{T}\mathbf{\Delta}\boldsymbol{T}')$. Then

$$\int \pi^{\mathrm{SIW}}(\mathbf{\Sigma})\,d\mathbf{\Sigma} = \frac{1}{k!}\int\int |\mathbf{\Lambda}|^{-a} \operatorname{etr}\left(-\frac{1}{2}\mathbf{\Lambda}^{-1}\boldsymbol{T}\mathbf{\Delta}\boldsymbol{T}'\right) d\mathbf{\Lambda}\,d\boldsymbol{T}.$$

For $i = 1, \ldots, k$, let $\tilde{\boldsymbol{t}}_i = (t_{i1}, \ldots, t_{ip})'$ be the first p components of the ith row of $\boldsymbol{T}$. Then

$$\begin{aligned}\int \pi^{\mathrm{SIW}}(\mathbf{\Sigma})\,d\mathbf{\Sigma} &= \frac{1}{k!}\int\int |\mathbf{\Lambda}|^{-a} \exp\left\{-\frac{1}{2}\sum_{i=1}^{k}\frac{\sum_{j=1}^{p}\delta_j t_{ij}^2}{\lambda_i}\right\} d\mathbf{\Lambda}\,d\boldsymbol{T}\\ &\le \frac{1}{k!}\int\int |\mathbf{\Lambda}|^{-a} \exp\left\{-\frac{\delta_p}{2}\sum_{i=1}^{k}\frac{\|\tilde{\boldsymbol{t}}_i\|^2}{\lambda_i}\right\} d\mathbf{\Lambda}\,d\boldsymbol{T}.\end{aligned} \tag{A.3}$$

If $p = k$, $\|\tilde{\boldsymbol{t}}_i\| = 1$ and $\int d\boldsymbol{T} = 1$. From (A.3), it yields

$$\int \pi^{\mathrm{SIW}}(\mathbf{\Sigma})\,d\mathbf{\Sigma} \le \frac{1}{k!}\int \prod_{i=1}^{k}\lambda_i^{-a} \exp\left\{-\frac{\delta_k}{2}\sum_{i=1}^{k}\frac{1}{\lambda_i}\right\} d\mathbf{\Lambda}, \tag{A.4}$$

which is proper if $a > 1$.

Next, we consider the case when $0 < p < k$. From (A.3), if $a > 1$, note that

$$\begin{aligned}\int \pi^{\mathrm{SIW}}(\mathbf{\Sigma})\,d\mathbf{\Sigma} &\le \frac{1}{k!}\int\left[\prod_{i=1}^{k}\int_0^\infty \lambda_i^{-a} \exp\left\{-\frac{\delta_p}{2}\frac{\|\tilde{\boldsymbol{t}}_i\|^2}{\lambda_i}\right\} d\lambda_i\right] d\boldsymbol{T}\\ &\le C\int \prod_{i=1}^{k}\|\tilde{\boldsymbol{t}}_i\|^{-2(a-1)}\,d\boldsymbol{T},\end{aligned} \tag{A.5}$$

where $C = \frac{1}{k!}(\delta_p/2)^{-k(a-1)}[\Gamma(a-1)]^k$. By Lemma 5, since $\|\tilde{\boldsymbol{t}}_1\|^2 + \cdots + \|\tilde{\boldsymbol{t}}_k\|^2 = p$, it yields

$$\int \pi^{\mathrm{SIW}}(\boldsymbol{\Sigma})\,d\boldsymbol{\Sigma} \le C \prod_{i=1}^{k} \int \|\tilde{\boldsymbol{t}}_i\|^{-2(a-1)}\,d\boldsymbol{T} = C\biggl(\int \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)}\,d\boldsymbol{T}\biggr)^k. \tag{A.6}$$

It suffices to verify that $\int \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)}\,d\boldsymbol{T}$ is finite. Note that

$$\boldsymbol{T} = (\boldsymbol{T}_{12}\boldsymbol{T}_{13}\cdots\boldsymbol{T}_{1k})(\boldsymbol{T}_{23}\cdots\boldsymbol{T}_{2k})\cdots(\boldsymbol{T}_{k-1,k})\boldsymbol{\Lambda}_\epsilon,$$

where $\boldsymbol{T}_{ij}$ is a simple orthogonal matrix such as

$$\boldsymbol{T}_{ij} = \boldsymbol{T}_{ij}(\theta_{ij}) = \begin{pmatrix} \boldsymbol{I} & 0 & 0 & 0 & 0 \\ 0 & \cos\theta_{ij} & 0 & -\sin\theta_{ij} & 0 \\ 0 & 0 & \boldsymbol{I} & 0 & 0 \\ 0 & \sin\theta_{ij} & 0 & \cos\theta_{ij} & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{I} \end{pmatrix}. \tag{A.7}$$

Here, $-\pi/2 < \theta_{ij} \le \pi/2$ and $\boldsymbol{\Lambda}_\epsilon$ being a diagonal matrix with diagonal elements 1 or -1 (see [1]). It follows from [1] that the Jacobian is

$$\left|\frac{\partial \boldsymbol{T}}{\prod_{i<j}\partial\theta_{ij}}\right| = \prod_{i=1}^{k-1}\prod_{j=i+1}^{p} \cos^{j-i-1}\theta_{ij}.$$

Define $\boldsymbol{\Omega} = \{-\frac{\pi}{2} \le \theta_{ij} \le \pi/2, i < j\}$. Therefore, we get

$$\int \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)}\,d\boldsymbol{T} = \int_\Omega \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)}\Biggl(\prod_{i=1}^{k-1}\prod_{j=i+1}^{k} \cos^{j-i-1}\theta_{ij}\,d\theta_{ij}\Biggr) \tag{A.8}$$

$$\le \int_\Omega \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)}\Biggl(\prod_{i=1}^{k-1}\prod_{j=i+1}^{k} d\theta_{ij}\Biggr). \tag{A.9}$$

For $i < j < l < k$, it is easy to verify $\boldsymbol{T}_{ik}\boldsymbol{T}_{jl} = \boldsymbol{T}_{jl}\boldsymbol{T}_{ik} = \boldsymbol{T}_{ik} + \boldsymbol{T}_{jl} - \boldsymbol{I}_k$. Using the relationship, it yields

$$\boldsymbol{T} = \Biggl(\prod_{j=1}^{k-2}\prod_{l=j+1}^{k-1}\boldsymbol{T}_{jl}\Biggr)\Biggl(\prod_{i=1}^{k}\boldsymbol{T}_{ik}\Biggr)\boldsymbol{\Lambda}_\epsilon.$$

Note that, the kth row of $\prod_{j=1}^{k-2}\prod_{l=j+1}^{k-1}\boldsymbol{T}_{jl}$ is $(0, 0, \ldots, 0, 1)$ and the kth row of $\prod_{i=1}^{k}\boldsymbol{T}_{ik}$ is

$$\Biggl(\sin\theta_{1k}, \sin\theta_{2k}\cos\theta_{1k}, \ldots, \sin\theta_{k-1,k}\prod_{i=1}^{k-2}\cos\theta_{ik}, \prod_{i=1}^{k-1}\cos\theta_{ik}\Biggr). \tag{A.10}$$

Therefore, $\|\tilde{\boldsymbol{t}}_k\|^2$ is the sum of squares of the first p components of (A.10), that is,

$$\begin{aligned} \|\tilde{\boldsymbol{t}}_k\|^2 &= \sin^2\theta_{1k} + \cos^2\theta_{1k}\sin^2\theta_{2k} + \cdots + \Biggl(\prod_{i=1}^{p-1}\cos^2\theta_{ik}\Biggr)\sin^2\theta_{pk} \\ &= 1 - \prod_{i=1}^{p}\cos^2\theta_{ik}. \end{aligned} \tag{A.11}$$

Defining $\boldsymbol{\Omega}_k = \{(\theta_{1k}, \ldots, \theta_{pk}) : 0 \le \theta_{ik} \le \pi/2, i = 1, \ldots, p\}$, (A.9) implies

$$\int \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)}\,d\boldsymbol{T} \le 2^p \int_{\boldsymbol{\Omega}_k} \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)} \prod_{i=1}^{p} d\theta_{ik}.$$

From (A.11), it yields

$$\|\tilde{\boldsymbol{t}}_k\|^2 \ge 1 - \left(\prod_{i=1}^p \cos^2\theta_{ik}\right)^{\frac{1}{p}} \ge \frac{p}{p} - \frac{1}{p}(\cos^2\theta_{1k} + \cdots + \cos^2\theta_{pk}) \tag{A.12}$$
$$= \frac{1}{p}(\sin^2\theta_{1k} + \sin^2\theta_{2k} + \cdots + \sin^2\theta_{pk}).$$

The second step follows the mean value inequality. Therefore, it yields that

$$\int_{\boldsymbol{\Omega}_k} \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)} \prod_{i=1}^p d\theta_{ik} \le 2^p p^{a-1} \int_{\boldsymbol{\Omega}_0} \frac{\prod_{i=1}^p d\theta_{ik}}{(\sin^2\theta_{1k} + \cdots + \sin^2\theta_{pk})^{a-1}},$$

where $\boldsymbol{\Omega}_0 = \{0 \le \theta_{ik} \le \pi/4, i \le p\}$. If $p = 1$, define $x = \sin\theta_{1k}$, so $d\theta_{1k} = \sqrt{1-x^2}\,dx$, and

$$\int_{\boldsymbol{\Omega}_k} \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)}\, d\theta_{1k} \le 2\int_0^{\frac{\pi}{4}} (\sin\theta_{1k})^{-2(a-1)}\, d\theta_{1k} \le \int_0^1 x^{-2(a-1)}(1-x)^{-1/2}\, dx,$$

which is finite if $a < 3/2$. For $p > 1$, we make the transformations $x_i = \sin\theta_{ik}$ for $i = 1, \dots, p$, and get

$$\int_{\boldsymbol{\Omega}_k} \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)} \prod_{i=1}^p d\theta_{ik} \tag{A.13}$$
$$\le 2^{\frac{3p}{2}} p^{a-1} \int_{\{x_1^2+\cdots+x_p^2 \le p, x_i \ge 0, i \le p\}} \frac{1}{(x_1^2 + \cdots + x_p^2)^{a-1}} \prod_{i=1}^p dx_i.$$

Let $z = x_1^2 + \cdots + x_p^2$ and $y_i = x_i^2/z$ for $i < p$. From (A.13), we have

$$\int_{\boldsymbol{\Omega}_k} \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)} \prod_{i=1}^p d\theta_{ik}$$
$$\le 2^{\frac{3p}{2}} p^a \int_0^p z^{-a+\frac{p}{2}}\, dz \int_{\{y_i > 0, y_1+\cdots+y_{p-1}<1\}} \left(1 - \sum_{i=1}^{p-1} y_i\right)^{-\frac{1}{2}} \prod_{i=1}^{p-1} y_i^{-\frac{1}{2}}\, dy_i,$$

which is finite if $a < \frac{p}{2} + 1$. The proof of the sufficient condition is completed.

For the necessary condition of parts (a) and (b), note that

$$\int \pi^{\mathrm{SIW}}(\boldsymbol{\Sigma})\, d\boldsymbol{\Sigma} \ge \frac{1}{k!} \int \left[\prod_{i=1}^k \int_0^\infty \lambda_i^{-a} \exp\left\{-\frac{\delta_1}{2} \frac{\|\tilde{\boldsymbol{t}}_i\|^2}{\lambda_i}\right\} d\lambda_i\right] d\boldsymbol{T},$$

the integration with respect to $d\boldsymbol{\Lambda}$ is infinite if $a \le 1$. Furthermore, if $0 < p < k$, note that $\|t_i\|^2 \le 1$ and $a > 1$. Then we get

$$\int \pi^{\mathrm{SIW}}(\boldsymbol{\Sigma})\, d\boldsymbol{\Sigma} \ge C_1 \int \prod_{i=1}^k \|\tilde{\boldsymbol{t}}_i\|^{-2(a-1)}\, d\boldsymbol{T} \ge C_1 \int \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)}\, d\boldsymbol{T},$$

where $C_1 = (k!)^{-1}(\delta_1/2)^{-k(a-1)}\Gamma^k(a-1)$. Define $\boldsymbol{\Omega}_1 = \{0 \le \theta_{ij} \le \pi/4, i < j\}$, from (A.8), we have

$$\int \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)}\, d\boldsymbol{T} \ge 2^{-k(k-1)} \int_{\boldsymbol{\Omega}_1} \|\tilde{\boldsymbol{t}}_k\|^{-2(a-1)} \left(\prod_{i=1}^{k-1} \prod_{j=i+1}^k d\theta_{ij}\right).$$

From (A.11), we have $\|\tilde{\boldsymbol t}_k\|^2 \le \sum_{i=1}^p \sin^2\theta_{ik}$. Therefore,

$$\text{(A.14)}\qquad \begin{aligned}\int \pi^{\mathrm{SIW}}(\boldsymbol\Sigma)\,d\boldsymbol\Sigma &\ge 2^{-k(k-1)}C_1\int_{\boldsymbol\Omega_1}\frac{\prod_{i=1}^{k-1}\prod_{j=i+1}^{k}d\theta_{ij}}{(\sin^2\theta_{1k}+\cdots+\sin^2\theta_{pk})^{a-1}}\\ &\ge 2^{-k(k-1)}C_1\int_{\boldsymbol\Omega_0}\frac{\prod_{i=1}^{p}d\theta_{ik}}{(\sin^2\theta_{1k}+\cdots+\sin^2\theta_{pk})^{a-1}}.\end{aligned}$$

By an argument similar to that for proving sufficiency, (A.14) is infinite if $a \ge 1 + p/2$.

PROOF OF LEMMA 1. If $p = 0$, or $p = k$, Lemma 1 holds. For $1 < p < k$, write $\boldsymbol H = \boldsymbol O'\boldsymbol D\boldsymbol O$, where $\boldsymbol O$ is orthogonal matrix, and $\boldsymbol D$ is diagonal matrix with last p diagonal elements greater than 0, and the rest 0. Clearly, $r_0 = \operatorname{rank}(\boldsymbol D + \boldsymbol O\boldsymbol S\boldsymbol O')$. The matrices $\boldsymbol D$, $\boldsymbol O\boldsymbol S\boldsymbol O'$ and $\boldsymbol O\boldsymbol\Sigma\boldsymbol O'$ can be partitioned as

$$\boldsymbol D = \begin{pmatrix}\boldsymbol D_1 & \mathbf 0\\ \mathbf 0 & \boldsymbol D_2\end{pmatrix},\qquad \boldsymbol O\boldsymbol S\boldsymbol O' = \begin{pmatrix}\tilde{\boldsymbol S}_{11} & \tilde{\boldsymbol S}_{12}\\ \tilde{\boldsymbol S}'_{12} & \tilde{\boldsymbol S}_{22}\end{pmatrix},\qquad \boldsymbol O\boldsymbol\Sigma\boldsymbol O' = \begin{pmatrix}\tilde{\boldsymbol\Sigma}_{11} & \tilde{\boldsymbol\Sigma}_{12}\\ \tilde{\boldsymbol\Sigma}'_{12} & \tilde{\boldsymbol\Sigma}_{22}\end{pmatrix},$$

where $\boldsymbol D_1$, $\tilde{\boldsymbol S}_{11}$ and $\tilde{\boldsymbol\Sigma}_{11}$ are $m \times m$ diagonal matrices. $\tilde{\boldsymbol S}_{11} \sim \mathrm{Wishart}_m(m, \tilde{\boldsymbol\Sigma}_{11})$, so $\operatorname{rank}(\tilde{\boldsymbol S}_{11}) = m$ with probability one. Note that

$$\boldsymbol D + \boldsymbol O\boldsymbol S\boldsymbol O' = \begin{pmatrix}\boldsymbol D_1 + \tilde{\boldsymbol S}_{11} & \tilde{\boldsymbol S}_{12}\\ \boldsymbol S'_{12} & \boldsymbol D_2 + \tilde{\boldsymbol S}_{22}\end{pmatrix},$$

and

$$\begin{aligned}&\begin{pmatrix}\boldsymbol I_m & \mathbf 0\\ -\boldsymbol S'_{12}(\boldsymbol D_1+\tilde{\boldsymbol S}_{11})^{-1} & \boldsymbol I_{k-m}\end{pmatrix}\begin{pmatrix}\boldsymbol D_1 + \tilde{\boldsymbol S}_{11} & \tilde{\boldsymbol S}_{12}\\ \boldsymbol S'_{12} & \boldsymbol D_2 + \tilde{\boldsymbol S}_{22}\end{pmatrix}\begin{pmatrix}\boldsymbol I_m & -(\boldsymbol D_1+\tilde{\boldsymbol S}_{11})^{-1}\tilde{\boldsymbol S}_{12}\\ \mathbf 0 & \boldsymbol I_{k-m}\end{pmatrix}\\ &\quad= \begin{pmatrix}\boldsymbol D_1 + \tilde{\boldsymbol S}_{11} & \mathbf 0\\ \mathbf 0 & \boldsymbol D_2 + \boldsymbol\Phi\end{pmatrix},\end{aligned}$$

where $\boldsymbol\Phi = \tilde{\boldsymbol S}_{22} - \tilde{\boldsymbol S}'_{12}(\boldsymbol D_1 + \tilde{\boldsymbol S}_{11})^{-1}\tilde{\boldsymbol S}_{12}$. Since $\operatorname{rank}(\boldsymbol O\boldsymbol S\boldsymbol O') = \operatorname{rank}(\tilde{\boldsymbol S}_{11}) = m$, $\operatorname{rank}(\tilde{\boldsymbol S}_{22} - \tilde{\boldsymbol S}'_{12}\tilde{\boldsymbol S}_{11}^{-1}\tilde{\boldsymbol S}_{12}) = \operatorname{rank}(\boldsymbol O\boldsymbol S\boldsymbol O') - \operatorname{rank}(\tilde{\boldsymbol S}_{11}) = 0$ with probability one. Therefore, $\boldsymbol\Phi$ is nonnegative definite, and

$$r_0 = \operatorname{rank}(\boldsymbol D_1 + \tilde{\boldsymbol S}_{11}) + \operatorname{rank}(\boldsymbol D_2 + \boldsymbol\Phi) \ge m + \operatorname{rank}(\boldsymbol D_2).$$

Since $\operatorname{rank}(\boldsymbol D_2) = \min(k - m, p)$, $r_0 \ge \min\{k, m + p\}$. It is clear that $r_0 \le \min\{k, \operatorname{rank}(\boldsymbol H) + \operatorname{rank}(\boldsymbol S)\} = \min\{k, m + p\}$. The lemma is proved. □

A.2. Proof of Theorem 3. From Theorem 2, $E(\boldsymbol\Sigma^q \mid \boldsymbol Y)$ exists. From Lemma 4, we have

$$E(\boldsymbol\Sigma^q) = \frac{\int\int \boldsymbol\Gamma\boldsymbol\Lambda^q\boldsymbol\Gamma'|\boldsymbol\Lambda|^{-a}\operatorname{etr}(-\frac12\boldsymbol\Lambda^{-1}\boldsymbol\Gamma'\boldsymbol H\boldsymbol\Gamma)\,d\boldsymbol\Lambda\,d\boldsymbol\Gamma}{\int\int |\boldsymbol\Lambda|^{-a}\operatorname{etr}(-\frac12\boldsymbol\Lambda^{-1}\boldsymbol\Gamma'\boldsymbol H\boldsymbol\Gamma)\,d\boldsymbol\Lambda\,d\boldsymbol\Gamma}.$$

Recall $\boldsymbol H = \boldsymbol Z\boldsymbol\Delta\boldsymbol Z'$, we define $\boldsymbol T = \boldsymbol Z'\boldsymbol\Gamma$. Then $\boldsymbol\Gamma\boldsymbol\Lambda^q\boldsymbol\Gamma' = \boldsymbol Z\boldsymbol T\boldsymbol\Lambda^q\boldsymbol T'\boldsymbol Z'$ and $E(\boldsymbol\Sigma^q) = \boldsymbol Z\boldsymbol\Phi\boldsymbol Z'$, where

$$\boldsymbol\Phi = \frac{\int\int \boldsymbol T\boldsymbol\Lambda^q\boldsymbol T'\prod_{i=1}^k\lambda_i^{-a}\operatorname{etr}(-\frac12\boldsymbol\Lambda^{-1}\boldsymbol T'\boldsymbol\Delta\boldsymbol T)\,d\boldsymbol\Lambda\,d\boldsymbol T}{\int\int \prod_{i=1}^k\lambda_i^{-a}\operatorname{etr}(-\frac12\boldsymbol\Lambda^{-1}\boldsymbol T'\boldsymbol\Delta\boldsymbol T)\,d\boldsymbol\Lambda\,d\boldsymbol T}.$$

We now show that $\boldsymbol\Phi$ is diagonal. In fact, for $i \ne i'$, we have

$$\boldsymbol\Phi(i, i') = \sum_{h=1}^k\frac{\int\int \lambda_h^q t_{ih}t_{i'h}\prod_{i=1}^k\lambda_i^{-a}\exp\{-\sum_{j=1}^k\frac{\|\tilde{\boldsymbol t}_j\|^2}{2\lambda_j}\}\,d\boldsymbol\Lambda\,d\boldsymbol T}{\int\int \prod_{i=1}^k\lambda_i^{-a}\exp\{-\sum_{j=1}^k\frac{\|\tilde{\boldsymbol t}_j\|^2}{2\lambda_j}\}\,d\boldsymbol\Lambda\,d\boldsymbol T}.$$

For any given $\boldsymbol{\Lambda}$, we have

$$\begin{aligned}&\int t_{ih}t_{i'h}\prod_{i=1}^{k}\exp\left\{-\sum_{j=1}^{k}\frac{\|\bar{\boldsymbol{t}}_j\|^2}{2\lambda_j}\right\}d\boldsymbol{T}\\&\quad=\left\{\int_{t_{ih}t_{i'h}>0}+\int_{t_{ih}t_{i'h}<0}\right\}t_{ih}t_{i'h}\prod_{i=1}^{k}\exp\left\{-\frac{\|\bar{\boldsymbol{t}}_j\|^2}{2\lambda_j}\right\}d\boldsymbol{T}=0.\end{aligned}$$

Thus the off diagonal elements are 0. For the (i,i)th diagonal element of $\boldsymbol{\Phi}$,

$$\begin{aligned}\boldsymbol{\Phi}(i,i)&=\sum_{h=1}^{k}\frac{\int\int\lambda_h^q t_{ih}^2\prod_{i=1}^{k}\lambda_i^{-a}\exp\{-\sum_{j=1}^{k}\frac{\|\bar{\boldsymbol{t}}_j\|^2}{2\lambda_j}\}\,d\boldsymbol{\Lambda}\,d\boldsymbol{T}}{\int\int\prod_{i=1}^{k}\lambda_i^{-a}\exp\{-\sum_{j=1}^{k}\frac{\|\bar{\boldsymbol{t}}_j\|^2}{2\lambda_j}\}\,d\boldsymbol{\Lambda}\,d\boldsymbol{T}}\\&=\frac{\Gamma(a-q-1)}{2^q\Gamma(a-1)}\sum_{h=1}^{k}\frac{\int t_{ih}^2\|\bar{\boldsymbol{t}}_h\|^{2q}\prod_{j=1}^{k}\|\bar{\boldsymbol{t}}_j\|^{-2(a-1)}\,d\boldsymbol{T}}{\int\prod_{j=1}^{k}\|\bar{\boldsymbol{t}}_j\|^{-2(a-1)}\,d\boldsymbol{T}}=\phi_{q,i}.\end{aligned}$$

The last equality holds since the integration in the numerator is equal for each $h=1,\dots,k$. The theorem is proved.

Acknowledgments. J. O. Berger is supported by US NSF Grant DMS-1407775. J. O. Berger and C. Sun are supported in part by Chinese 111 Project B14019 and Chinese NSF Grant 11671147. C. Sun is also a professor at East China Normal University.

Dongchu Sun is also affiliated with the School of Statistics, East China Normal University.

SUPPLEMENTARY MATERIAL

Supplement to "Bayesian analysis of the covariance matrix of a multivariate normal distribution with a new class of priors" (DOI: 10.1214/19-AOS1891SUPP; .pdf). The results for L_1 and L_3 are in the Supplementary Material.

REFERENCES

[1] ANDERSON, T. W., OLKIN, I. and UNDERHILL, L. G. (1987). Generation of random orthogonal matrices. *SIAM J. Sci. Statist. Comput.* **8** 625–629. MR0892309 https://doi.org/10.1137/0908055

[2] BERGER, J. O., STRAWDERMAN, W. and TANG, D. (2005). Posterior propriety and admissibility of hyperpriors in normal hierarchical models. *Ann. Statist.* **33** 606–646. MR2163154 https://doi.org/10.1214/009053605000000075

[3] BERGER, J. O. and SUN, D. (2008). Objective priors for the bivariate normal model. *Ann. Statist.* **36** 963–982. MR2396821 https://doi.org/10.1214/07-AOS501

[4] BERGER, J. O., SUN, D. and SONG, C. (2020). An objective prior for hyperparameters in normal hierarchical models. *J. Multivariate Anal.* To appear.

[5] BERGER, J. O., SUN, D. and SONG, C. (2020). Supplement to "Bayesian analysis of the covariance matrix of a multivariate normal distribution with a new class of priors." https://doi.org/10.1214/19-AOS1891SUPP.

[6] CHEN, M.-H. and SCHMEISER, B. (1993). Performance of the Gibbs, hit-and-run, and Metropolis samplers. *J. Comput. Graph. Statist.* **2** 251–272. MR1272394 https://doi.org/10.2307/1390645

[7] DANIELS, M. J. and KASS, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *J. Amer. Statist. Assoc.* **94** 1254–1263. MR1731487 https://doi.org/10.2307/2669939

[8] DANIELS, M. J. and KASS, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57** 1173–1184. MR1950425 https://doi.org/10.1111/j.0006-341X.2001.01173.x

[9] DEY, D. K. and SRINIVASAN, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* **13** 1581–1591. MR0811511 https://doi.org/10.1214/aos/1176349756

[10] EATON, M. L. and OLKIN, I. (1987). Best equivariant estimators of a Cholesky decomposition. *Ann. Statist.* **15** 1639–1650. MR0913579 https://doi.org/10.1214/aos/1176350615

[11] EGUCHI, N., SAITO, R., SAEKI, T., NAKATSUKA, Y., BELIKOV, D. and MAKSYUTOV, S. (2010). A priori covariance estimation for CO2 and CH4 retrievals. *J. Geophys. Res.* **115** Art. ID D10215.

[12] FARRELL, R. H. (1985). *Multivariate Calculation: Use of the Continuous Groups. Springer Series in Statistics*. Springer, New York. MR0770934 https://doi.org/10.1007/978-1-4613-8528-8

[13] FREI, M. and KUNSCH, H. R. (2012). Sequential state and observation noise covariance estimation using combined ensemble Kalman and particle filters. *Mon. Weather Rev.* **140** 1476–1495.

[14] GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

[15] GUILLOT, D., RAJARATNAM, B. and EMILE-GEAY, J. (2015). Statistical paleoclimate reconstructions via Markov random fields. *Ann. Appl. Stat.* **9** 324–352. MR3341118 https://doi.org/10.1214/14-AOAS794

[16] HAFF, L. R. (1979). Estimation of the inverse covariance matrix: Random mixtures of the inverse Wishart matrix and the identity. *Ann. Statist.* **7** 1264–1276. MR0550149

[17] HAFF, L. R. (1991). The variational form of certain Bayes estimators. *Ann. Statist.* **19** 1163–1190. MR1126320 https://doi.org/10.1214/aos/1176348244

[18] HAMIMECHE, S. and LEWIS, A. (2009). Properties and use of CMB power spectrum likelihoods. *Phys. Rev. D* **79** Art. ID 83012.

[19] HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109. MR3363437 https://doi.org/10.1093/biomet/57.1.97

[20] HOFF, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 971–992. MR2750253 https://doi.org/10.1111/j.1467-9868.2009.00716.x

[21] HOFF, P. D. (2009). Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *J. Comput. Graph. Statist.* **18** 438–456. MR2749840 https://doi.org/10.1198/jcgs.2009.07177

[22] JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford. MR0187257

[23] LEDOIT, O. and WOLF, M. (2004). Honey, I shrunk the sample covariance matrix. *J. Portf. Manag.* **4** 110–119.

[24] LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339 https://doi.org/10.1016/S0047-259X(03)00096-4

[25] LIN, S. P. and PERLMAN, M. D. (1985). A Monte Carlo comparison of four estimators of a covariance matrix. In *Multivariate Analysis VI* (*Pittsburgh*, *Pa*., 1983) 411–429. North-Holland, Amsterdam. MR0822310

[26] POPE, A. C. and SZAPUDI, I. (2005). Shrinkage estimation of the power spectrum covariance matrix. *Mon. Not. R. Astron. Soc.* **389** 766–774.

[27] POURAHMADI, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statist. Sci.* **26** 369–387. MR2917961 https://doi.org/10.1214/11-STS358

[28] PRESS, S. J. (2012). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. Courier Corporation, North Chelmsford, MA.

[29] RAJARATNAM, B., MASSAM, H. and CARVALHO, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.* **36** 2818–2849. MR2485014 https://doi.org/10.1214/08-AOS619

[30] SCHÄFER, J. and STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. ID 32. MR2183942 https://doi.org/10.2202/1544-6115.1175

[31] SINHA, B. K. and GHOSH, M. (1987). Inadmissibility of the best equivariant estimators of the variance–covariance matrix, the precision matrix, and the generalized variance under entropy loss. *Statist. Decisions* **5** 201–227. MR0905238

[32] STEIN, C. (1956). Some problems in multivariate analysis. Part I. Technical Report 6, Dept. Statistics, Stanford Univ.

[33] STEIN, C. (1975). Estimation of a covariance matrix, Rietz Lecture. In 39*th Annual Meeting IMS*, *Atlanta*, *GA*.

[34] SUN, D. and BERGER, J. O. (2007). Objective Bayesian analysis for the multivariate normal model. In *Bayesian Statistics* 8. *Oxford Sci. Publ.* 525–562. Oxford Univ. Press, Oxford. MR2433206

[35] YANG, R. and BERGER, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22** 1195–1211. MR1311972 https://doi.org/10.1214/aos/1176325625