Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics
Benidorm (Alicante, Spain), June 1st–6th, 2006

Objective Bayesian Analysis for the Multivariate Normal Model

Dongchu Sun
University of Missouri-Columbia and Virginia Tech, USA
sund@missouri.edu
James O. Berger
Duke University, USA
berger@stat.duke.edu

Summary

Objective Bayesian inference for the multivariate normal distribution is illustrated, using different types of formal objective priors (Jeffreys, invariant, reference and matching), different modes of inference (Bayesian and frequentist), and different criteria involved in selecting optimal objective priors (ease of computation, frequentist performance, marginalization paradoxes, and decision-theoretic evaluation).

In the course of the investigation of the bivariate normal model in Berger and Sun (2006), a variety of surprising results were found, including the availability of objective priors that yield exact frequentist inferences for many functions of the bivariate normal parameters, such as the correlation coefficient. Certain of these results are generalized to the multivariate normal situation.

The prior that most frequently yields exact frequentist inference is the right-Haar prior, which unfortunately is not unique. Two natural proposals are studied for dealing with this non-uniqueness: first, mixing over the right-Haar priors; second, choosing the 'empirical Bayes' right-Haar prior, that which maximizes the marginal likelihood of the data. Quite surprisingly, we show that neither of these possibilities yields a good solution. This is disturbing and sobering. It is yet another indication that improper priors do not behave as do proper priors, and that it can be dangerous to apply 'understandings' from the world of proper priors to the world of improper priors.

Keywords and Phrases: Kullback-Leibler divergence; Jeffreys prior; multivariate normal distribution; matching priors; reference priors; invariant priors.

This research was supported by the National Science Foundation, under grants DMS-0103265 and SES-0351523, and the National Institute of Health, under grants R01-CA100760 and R01-MH071418.

1. INTRODUCTION

Estimating the mean and covariance matrix of a multivariate normal distribution became of central theoretical interest when Stein (1956, 1972) showed that standard estimators had significant problems, including inadmissibility from a frequentist perspective. Most problematical were standard estimators of the covariance matrix; see Yang and Berger (1994) and the references therein.

In the Bayesian literature, the most commonly used prior for a multivariate normal distribution is a normal prior for the normal mean and an inverse Wishart prior for the covariance matrix. Such priors are conjugate, leading to easy computation, but lack flexibility and also lead to inferences of the same structure as those shown to be inferior by Stein. More flexibile and better performing priors for a covariance matrix were developed by Leonard and Hsu (1992), and Brown (2001) (the generalized inverse Wishart prior). In the more recent Bayesian literature, aggressive shrinkage of eigenvalues, correlations, or other features of the covariance matrix are entertained; see, for example, Daniels (1999, 2002), Liechty (2004) and the references therein. These priors may well be successful in practice, but they do not seem to be formal objective priors according to any of the common definitions.

Recently, Berger and Sun (2006) considered objective inference for parameters of the bivariate normal distribution and functions of these parameters, with special focus on development of objective confidence or credible sets. In the course of the study, many interesting issues were explored involving objective Bayesian inference, including different types of objective priors (Jeffreys, invariant, reference and matching), different modes of inference (Bayesian and frequentist), and different criteria involved in deciding on optimal objective priors (ease of computation, frequentist performance and marginalization paradoxes).

In this paper, we first generalize some of the bivariate results to the multivariate normal distribution; Section 2 presents the generalizations of the various objective priors discussed in Berger and Sun (2006). We particularly focus on reference priors, and show that the right-Haar prior is indeed a one-at-a-time reference prior (Berger and Bernardo, 1992) for many parameters and functions of parameters.

Section 3 gives some basic properties of the resulting posterior distributions and gives *constructive posterior distributions* for many of the priors. Constructive posteriors are expressions for the posterior distribution which allow very simply simulation from the posterior. Constructive posteriors are also very powerful for proving results about exact frequentist matching. (Exact frequentist matching means that $100(1-\alpha)\%$ credible sets arising from the resulting posterior are also exact frequentist confidence sets at the specified level.) Results about matching for the right-Haar prior are given in Section 4 for a variety of parameters.

One of the most interesting features of right-Haar priors is that, while they result in exact frequentist matching, they also seem to yield marginalization paradoxes (Dawid, Stone and Zidek, 1973). Thus one is in the philosophical conundrum of having to choose between frequentist matching and avoidance of the marginalization paradox. This is also discussed in Section 4.

Another interesting feature of the right-Haar priors is that they are not unique; they depend on which triangular decomposition of a covariance matrix is employed. In Section 5, two natural proposals are studied to deal with this non-uniqueness. The first is to simply mix over the right-Haar priors. The second is to choose the 'empirical Bayes' right-Haar prior, namely that which maximizes the marginal likelihood of the data. Quite surprisingly, it is shown that both of these solutions gives inferior answers, a disturbing and sobering phenomenon. It is yet another

indication that improper priors do not behave as do proper priors, and that it can be dangerous to apply 'understandings' from the world of proper priors to the world of improper priors.

2. OBJECTIVE PRIORS FOR THE MULTIVARIATE NORMAL DISTRBUTION

Consider the p-dimentional multivariate normal population, $\boldsymbol{x} = (x_1, \cdots, x_p)' \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, whose density is given by

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \quad (1)$$

2.1. *Previously Considered Objective Priors*

Perhaps the most popular prior for the multivariate normal distribution is the Jeffreys (rule) prior (Jeffreys, 1961)

$$\pi_J(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-(p+2)/2}. \quad (2)$$

Another commonly used prior is the independence-Jeffreys prior

$$\pi_{IJ}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-(p+1)/2}. \quad (3)$$

It is commonly thought that either the Jeffreys or independence-Jeffreys priors are most natural, and most likely to yield classical inferences. However, Geisser and Cornfield (1963) showed that the prior which is exact frequentist matching for all means and variances (and which also yields Fisher's fiducial distribution for these parameters) is

$$\pi_{GC}(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-p}. \quad (4)$$

It is simple chance that this prior happens to be the Jeffreys prior for $p = 2$ (and perhaps simple chance that it agrees with π_{IJ} for $p = 1$); these coincidences may have contributed significantly to the popular notion that Jeffreys priors are generally successful.

In spite of the frequentist matching success of π_{GC} for means and variances, the prior seems to be quite bad for correlations, predictions, or other inferences involving a multivariate normal distribution. Thus a variety of other objective priors have been proposed in the literature.

Chang and Eaves (1990) ((7) on page 1605) derived the reference prior for the parameter ordering $(\mu_1, \cdots, \mu_p; \sigma_1, \cdots, \sigma_p; \boldsymbol{\Upsilon})$,

$$\pi_{CE}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma} = \frac{1}{|\boldsymbol{\Sigma}|^{(p+1)/2}|\boldsymbol{I}_p + \boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-1}|^{1/2}}\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma} \quad (5)$$

$$= 2^p \left[\prod_{i=1}^{p} \frac{d\mu_i d\sigma_i}{\sigma_i}\right]\left[\frac{1}{|\boldsymbol{\Upsilon}|^{(p+1)/2}|\boldsymbol{I}_p + \boldsymbol{\Upsilon}^*\boldsymbol{\Upsilon}^{-1}|^{1/2}} \prod_{i<j} d\rho_{ij}\right], \quad (6)$$

where $\boldsymbol{\Upsilon}$ is the correlation matrix and $\boldsymbol{A}^*\boldsymbol{B}$ denotes the Hadamard product of the squared matrices $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$, whose entries are $c_{ij} = a_{ij}b_{ij}$. In the

bivariate normal case, this prior is the same as Lindley's (1965) prior, derived using certain notions of transformation to constant information, and was derived as a reference prior for the correlation coefficient ρ in Bayarri (1981).

Chang and Eaves (1990) also derived the reference prior for the ordering $(\mu_1, \cdots, \mu_p; \lambda_1, \cdots, \lambda_p; \boldsymbol{O})$, where $\lambda_1 > \cdots > \lambda_p$ are the ordered eigenvalues of $\boldsymbol{\Sigma}$, and $\boldsymbol{O}$ is an orthogonal matrix such that $\boldsymbol{\Sigma} = \boldsymbol{O}'\text{diag}(\lambda_1, \cdots, \lambda_p)\boldsymbol{O}$. This reference prior was discussed in detail in Berger and Yang (1994) and has the form,

$$\pi_E(\mu, \boldsymbol{\Sigma})\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma} = \frac{I_{[\lambda_1 > \cdots > \lambda_p]}}{|\boldsymbol{\Sigma}| \prod_{i<j}(\lambda_i - \lambda_j)}\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}\,. \tag{7}$$

Another popular prior is the right Haar prior, which has been extensively studied (see, e.g., Eaton and Sudderth, 2002).

It is most convenient to express this prior in terms of a lower-triangular matrix $\boldsymbol{\Psi}$ with positive diagonal elements and such that

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Psi}'\boldsymbol{\Psi}. \tag{8}$$

(Note that there are many such matrices, so that the right Haar prior is not unique.) The right Haar prior corresponding to this decomposition is given by

$$\pi_H(\mu, \boldsymbol{\Psi})\, d\boldsymbol{\mu}\, d\boldsymbol{\Psi} = \prod_{i=1}^{p} \frac{1}{\psi_{ii}^i}\, d\boldsymbol{\mu}\, d\boldsymbol{\Psi}\,. \tag{9}$$

We will see in the next subsection that, this prior is one-at-a-time reference prior for various parameterizations.

Because $d\boldsymbol{\Sigma} = \prod_{i=1}^{p} \psi_{ii}^{i-2(p+1)} d\boldsymbol{\Psi}$, the independence Jeffreys prior $\pi_{IJ}(\boldsymbol{\mu}, \boldsymbol{\Psi})$ corresponding to the left-haar measure is given by

$$\pi_{IJ}(\boldsymbol{\mu}, \boldsymbol{\Psi}) = \prod_{i=1}^{p} \frac{1}{\psi_{ii}^{p-i+1}}\, d\boldsymbol{\mu}\, d\boldsymbol{\Psi}\,. \tag{10}$$

The right-Haar prior and the independence Jefferys prior are limiting cases of generalized Wishart priors; see Brown (2001) for a review.

We now give the Fisher information matrix and a result about the reference prior for a group ordering related to the right-Haar parameterization; proofs are relegated to the appendix.

Fact 1 *(a) The Fisher information matrix for* $\{\boldsymbol{\mu}, \psi_{11}, (\psi_{21}, \psi_{22}), \cdots, (\psi_{p1}, \cdots, \psi_{pp})\}$ *is*

$$\boldsymbol{J} = -E\left(\frac{\partial^2 \log f}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) = diag(\boldsymbol{\Sigma}^{-1}, \boldsymbol{\Lambda}_1, \cdots, \boldsymbol{\Lambda}_p), \tag{11}$$

where, for $i = 1, \cdots, p$,

$$\boldsymbol{\Lambda}_i \quad = \quad \boldsymbol{\Sigma}_i + \frac{1}{\psi_{ii}^2} \boldsymbol{e}_i \boldsymbol{e}_i'\,,$$

with $\boldsymbol{e}_i$ being the i^{th} unit column vector.
(b) The reference prior of $\boldsymbol{\Psi}$ for the ordered group $\{\mu_1, \cdots, \mu_p, \psi_{11}, (\psi_{21}, \psi_{22}), \cdots, (\psi_{p1}, \cdots, \psi_{pp})\}$ is given by

$$\pi_{R1}(\boldsymbol{\mu}, \boldsymbol{\Psi}) \quad \propto \quad \frac{1}{\prod_{i=1}^{p} \psi_{ii}}. \tag{12}$$

Note that the right-Haar prior, the Jeffreys (rule) prior, the independence Jeffreys prior, the Geisser-Cornfield prior and the reference prior π_{R1} have the form

$$\pi_{\boldsymbol{a}}(\boldsymbol{\mu}, \boldsymbol{\Psi}) d\boldsymbol{\mu} d\boldsymbol{\Psi} \quad = \quad \prod_{i=1}^{p} \frac{1}{\psi_{ii}^{a_i}} d\boldsymbol{\mu} d\boldsymbol{\Psi}, \tag{13}$$

where $\boldsymbol{a} = (a_1, \cdots, a_p)$. This class of priors has also received considerable attention in directed acyclic graphical models. See, for example, Roverato and Consonni (2004). Also, Consonni, Gutiérrez-Peña, and Veronese (2004) and their references for reference priors for exponential families with simple quadratic variance function.

Table 1: Summary of objective priors of $(\boldsymbol{\mu}, \boldsymbol{\Psi})$ as special cases of (13)

prior	form	$(a_1, \cdots, a_p)$
π_H	$\prod_{i=1}^{p} \psi_{ii}^{-i}$	$a_i = i$
π_J	$\prod_{i=1}^{p} \psi_{ii}^{-(p-i)}$	$a_i = p - i$
π_{IJ}	$\prod_{i=1}^{p} \psi_{ii}^{-(p-i+1)}$	$a_i = p - i + 1$
π_{GC}	$\prod_{i=1}^{p} \psi_{ii}^{-(2-i)}$	$a_i = 2 - i$
π_{R1}	$\prod_{i=1}^{p} \psi_{ii}^{-1}$	$a_i = 1$

2.2. *Reference Priors under Alternative Parameterizations*

Pourahmadi (1999) considered another decomposition of $\boldsymbol{\Sigma}^{-1}$. Let $\boldsymbol{T} = (t_{ij})_{p \times p}$ be the $p \times p$ unit lower triangular matrix, where

$$t_{ij} = \begin{cases} 0 & \text{if } i < j, \\ 1 & \text{if } i = j, \\ \frac{\psi_{ij}}{\psi_{ii}} & \text{if } i > j. \end{cases} \tag{14}$$

Pourahmadi (1999) pointed out the statistical interpretations of the below-diagonal entries of $\boldsymbol{T}$ and the diagonal entries of $\boldsymbol{\Psi}$. In fact, $x_1 \sim N(\mu_1, d_1)$, $x_i \sim N(\mu_i - \sum_{j=1}^{j-1} t_{ij}(x_j - \mu_j), \psi_{ii}^{-2})$, $(j \geq 2)$, so the t_{ij} are the negatives of the coefficients of the best linear predictor of x_i based on $(x_1, \cdots, x_{i-1})$, and ψ_{ii}^2 is the precision of the predictive distribution. Write $\widetilde{\boldsymbol{\Psi}} = diag(\psi_{11}, \cdots, \psi_{pp})$. Clearly

$$\boldsymbol{\Psi} \quad = \quad \widetilde{\boldsymbol{\Psi}} \boldsymbol{T}, \tag{15}$$

$$\boldsymbol{\Sigma} \quad = \quad (\boldsymbol{T}' \widetilde{\boldsymbol{\Psi}}^2 \boldsymbol{T})^{-1}. \tag{16}$$

For $i = 2, \cdots, p$, define $\widetilde{\boldsymbol{\Psi}}_i = diag(\psi_{11}, \cdots, \psi_{ii})$ and denote the upper and left $i \times i$ submatrix of $\boldsymbol{T}$ by $\boldsymbol{T}_i$. Then

$$\boldsymbol{\Psi}_i = \widetilde{\boldsymbol{\Psi}}_i \boldsymbol{T}_i \,, \tag{17}$$

$$\boldsymbol{\Sigma}_i = (\boldsymbol{T}_i' \widetilde{\boldsymbol{\Psi}}_i^2 \boldsymbol{T}_i)^{-1}, \quad i = 2, \cdots, p \,. \tag{18}$$

Fact 2 *(a) The Fisher information for* $(\boldsymbol{\mu}, \psi_{11}, (t_{21}, \psi_{22}), (t_{31}, t_{32}, \psi_{33}), \cdots, (t_{p1}, \cdots, t_{p,p-1}, \psi_{ii}))$ *is of the form*

$$\widetilde{\boldsymbol{J}} = diag(\boldsymbol{T}' \widetilde{\boldsymbol{\Psi}}^2 \boldsymbol{T}, \frac{2}{\psi_{11}^2}, \widetilde{\boldsymbol{J}}_2, \cdots, \widetilde{\boldsymbol{J}}_p), \tag{19}$$

where for $i = 2, \cdots, p$,

$$\widetilde{\boldsymbol{J}}_i = diag\left(\psi_{ii}^2 \boldsymbol{T}_{i-1}^{-1} \widetilde{\boldsymbol{\Psi}}_{i-1}^{-2} \boldsymbol{T}_{i-1}'^{-1}, \frac{2}{\psi_{ii}^2} \right). \tag{20}$$

(b) The one-at-a-time reference prior for $\{\mu_1, \cdots, \mu_p, \psi_{11}, \psi_{22}, \cdots, \psi_{ii}, t_{21}, t_{31}, t_{32}, \cdots, t_{p1}, \cdots, t_{p,p-1}\}$*, and with any ordering of parameters, is*

$$\widetilde{\pi}_R(\widetilde{\boldsymbol{\theta}}) = \prod_{i=1}^{p} \frac{1}{\psi_{ii}} \,. \tag{21}$$

(c) The reference prior in (b) is the same as the right-Haar measure for $\boldsymbol{\Psi}$*, given in (9).*

Consider the parameterization $\boldsymbol{D} = \mathrm{diag}(d_1, \cdots, d_p)$ and $\boldsymbol{T}$, where $d_i = 1/\psi_{ii}^2$. Clearly $\boldsymbol{D} = \widetilde{\boldsymbol{\Psi}}^{-2}$ and $\boldsymbol{\Sigma}^{-1} = \boldsymbol{T}' \boldsymbol{D}^{-1} \boldsymbol{T}$. Also write $\boldsymbol{D}_i = \boldsymbol{\Psi}_i^{*-2}$.

Corollary 0.1 *(a) The Fisher information for* $(\boldsymbol{\mu}, d_1, \cdots, \cdots, d_p; t_{21}; t_{31}, t_{32}; \cdots, t_{p1}, \cdots, t_{p,p-1})$ *is of the form*

$$\boldsymbol{J}^{\#} = diag(\boldsymbol{T}' \boldsymbol{D}^{-1} \boldsymbol{T}, \frac{1}{d_1^2}, \cdots, \frac{1}{d_p^2}, \boldsymbol{\Delta}_2, \cdots, \boldsymbol{\Delta}_p), \tag{22}$$

where, for $i = 2, \cdots, p$,

$$\boldsymbol{\Delta}_i = \frac{1}{d_i} \boldsymbol{T}_{i-1}^{-1} \boldsymbol{D}_{i-1} \boldsymbol{T}_{i-1}'^{-1} \,. \tag{23}$$

(b) The one-at-a-time reference prior for $\{\mu_1, \cdots, \mu_p, d_1, \cdots, \cdots, d_p, t_{21}, t_{31}, t_{32}, \cdots, t_{p1}, \cdots, t_{p,p-1}\}$*, and with any ordering, is*

$$\widetilde{\pi}_R(\boldsymbol{\theta}) \propto \prod_{i=1}^{p} \frac{1}{d_i} \,. \tag{24}$$

(c) The reference prior in (b) is the same as the right-Haar measure for $\boldsymbol{\Psi}$*, given in (9).*

Suppose one is interested in the generalized variance $|\boldsymbol{\Sigma}| = \prod_{i=1}^{p} d_i$; the one-at-a-time reference prior is also the right-Haar measure π_H. To see this, define

$$\left\{ \begin{array}{lcl} \xi_1 & = & \frac{d_1}{d_2}\,, \\ \xi_2 & = & \frac{(d_1 d_2)^{1/2}}{d_3}\,, \\ \cdots & & \cdots \\ \xi_{p-1} & = & \frac{(\prod_{j=1}^{p-1} d_j)^{1/(p-1)}}{d_p}\,, \\ \xi_p & = & \prod_{j=1}^{p} d_j\,. \end{array} \right. \tag{25}$$

Fact 3 *(a) The Fisher information matrix for* $(\boldsymbol{\mu}, \xi_1, \cdots, \xi_p; t_{21}, t_{31}, t_{32}, \cdots, t_{p1}, \cdots, t_{p,p-1})$ *is*

$$diag\left(\boldsymbol{\Sigma}^{-1}, \frac{1}{2\xi_1^2}, \frac{2}{3\xi_2^2}, \cdots, \frac{p-1}{p\,\xi_{p-1}^2}, \frac{1}{p\,\xi_p^2}, \boldsymbol{\Delta}_2, \cdots, \boldsymbol{\Delta}_p\right), \tag{26}$$

where $\boldsymbol{\Delta}_i$ *is given by (23).*
(b) The one-at-a-time reference prior of any ordering for $\{\mu_1, \cdots, \mu_p, \xi_1, \cdots, \xi_p; t_{21}, t_{31}, t_{32}, \cdots, t_{p1}, \cdots, t_{p,p-1}\}$ *is*

$$\widetilde{\pi}_R(\boldsymbol{\theta}) \propto \prod_{i=1}^{p} \frac{1}{\xi_i}. \tag{27}$$

(c) The reference prior in (b) is π_H*, given in (9).*

Corollary 0.2 *Since* $\xi_p = |\boldsymbol{\Sigma}|$*, it is immediate that the one-at-a-time reference prior for* ξ_p*, with nuisance parameters* $(\boldsymbol{\mu}, \xi_1, \cdots \xi_{p-1}, t_{21}, t_{31}, t_{32}, \cdots, t_{p1}, \cdots, t_{p,p-1})$*, is the right-Haar prior* π_H*.*

corollary

Corollary 0.3 *One might be interested in* $\eta_i \equiv |\boldsymbol{\Sigma}_i| = \prod_{j=1}^{i} d_i$*, the generalized variance of the upper left* $i \times i$ *submatrix of* $\boldsymbol{\Sigma}$*. Using the same arguments as in Fact 3, the Fisher information for* $(\boldsymbol{\mu}, \xi_1, \cdots, \xi_{i-1}, \eta_i, d_{i+1}, \cdots, d_p; t_{21}, t_{31}, t_{32}, \cdots, t_{p1}, \cdots, t_{p,p-1})$ *is*

$$diag\left(\boldsymbol{\Sigma}^{-1}, \frac{1}{2\xi_1^2}, \cdots, \frac{i-1}{i\,\xi_{i-1}^2}, \frac{1}{i\,\eta_i^2}, \frac{1}{d_{i+1}^2}, \cdots, \frac{1}{d_p^2}, \boldsymbol{\Delta}_2, \cdots, \boldsymbol{\Delta}_p\right). \tag{28}$$

The one-at-a-time reference prior for $|\boldsymbol{\Sigma}_i|$*, with nuisance parameters* $\{\mu_1, \cdots, \mu_p, \xi_1, \cdots, \xi_{i-1}, d_{i+1}, \cdots, d_p; t_{21}, t_{31}, t_{32}, \cdots, t_{p1}, \cdots, t_{p,p-1})$ *and any parameter order, is the right-Haar prior* π_H*.*

3. POSTERIOR DISTRIBUTIONS

Let $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ be a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The likelihood function of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \;=\; (2\pi)^{-np/2} |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{ -\frac{n}{2} (\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) - \frac{1}{2} tr(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) \right\},$$

where

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i, \qquad \boldsymbol{S} = \sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)'.$$

Since all the considered priors are constant in $\boldsymbol{\mu}$, the conditional posterior for $\boldsymbol{\mu}$ will be

$$(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{X}) \sim N_p(\overline{\boldsymbol{x}}, \frac{1}{n}\boldsymbol{\Sigma}). \tag{29}$$

Generation from this is standard, so the challenge of simulation from the posterior distribution requires only sampling from the marginal posterior of $\boldsymbol{\Sigma}$ given $\boldsymbol{S}$. Note that the marginal likelihood of $\boldsymbol{\Sigma}$ based on $\boldsymbol{S}$ is

$$L_1(\boldsymbol{\Sigma}) = \frac{(2\pi)^{-np/2}}{|\boldsymbol{\Sigma}|^{(n-1)/2}} etr\bigg(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\bigg). \tag{30}$$

Throughout the paper, we assume that $\boldsymbol{S}$ is positive definite, as this is true with probability one.

3.1. *Marginal Posteriors of* $\boldsymbol{\Sigma}$ *under* π_J, π_{IJ}, π_{CE} *and* π_E

Marginal Posteriors Under π_J and π_{IJ}: It is immediate that these marginal posteriors for $\boldsymbol{\Sigma}$ are Inverse Wishart $(\boldsymbol{S}^{-1}, n)$ and Inverse Wishart $(\boldsymbol{S}^{-1}, n-1)$, respectively.

Marginal Posterior Under π_{CE}: This marginal posterior distribution is imposing in its complexity. However, rather remarkably there is a simple rejection algorithm that can be used to generate from it:

Step 1. Generate $\boldsymbol{\Sigma} \sim$ Inverse Wishart $(\boldsymbol{S}^{-1}, n-1)$.

Step 2. Simulate $u \sim$ Uniform$(0,1)$. If $u \leq 2^{p/2}\,|\boldsymbol{I}_p + \boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-1}|^{-1/2}$, report $\boldsymbol{\Sigma}$. Otherwise go back to *Step 1.*

Note that the acceptance probability $2^{p/2}\,|\boldsymbol{I}_p + \boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-1}|^{-1/2}$ is equal to one if the proposed $\boldsymbol{\Sigma}$ is diagonal, but is near zero when the proposed $\boldsymbol{\Sigma}$ is nearly singular. That this algorithm is a valid accept-reject algorithm, based on generation of $\boldsymbol{\Sigma}$ from the independence Jeffreys posterior, is established in Berger and Sun (2006).
Marginal Posterior Under π_E: It is possible to generate from this posterior using the following Metropolis-Hastings algorithm from Berger et. al. (2005).

Step 1. Generate $\boldsymbol{\Sigma}^* \sim$ Inverse Wishart $(\boldsymbol{S}^{-1}, n-1)$.

Step 2. Set $\boldsymbol{\Sigma}' = \begin{cases} \boldsymbol{\Sigma}^* & \text{with probability } \alpha, \\ \boldsymbol{\Sigma} & \text{otherwise,} \end{cases}$ where

$$\alpha = \min\left\{1,\ \frac{\prod_{i<j}(\lambda_i^* - \lambda_j^*)}{\prod_{i<j}(\lambda_i - \lambda_j)} \cdot \frac{|\boldsymbol{\Sigma}|^{(p-1)/2}}{|\boldsymbol{\Sigma}^*|^{(p-1)/2}}\right\}.$$

3.2. *Marginal Posterior of* $\mathbf{\Sigma}$ *under* π_a

In the following, we will write $\mathbf{\Sigma} = \mathbf{\Sigma}_p$, including the dimension in order to dervive some useful recursive formulas. Let $\boldsymbol{\psi}_{p,p-1}$ be the $(p-1)\times 1$ vector of the last column of $\mathbf{\Psi}_p'$, excluding ψ_{pp}. Then

$$\mathbf{\Psi}_1 = \psi_{11},\ \mathbf{\Psi}_2 = \begin{pmatrix} \psi_{11} & 0 \\ \psi_{21} & \psi_{22} \end{pmatrix},\ \cdots, \mathbf{\Psi}_p = \begin{pmatrix} \mathbf{\Psi}_{p-1} & \mathbf{0} \\ \boldsymbol{\psi}_{p,p-1}' & \psi_{pp} \end{pmatrix}.$$

We will also write $\boldsymbol{S} = \boldsymbol{S}_p$ and let $\boldsymbol{s}_{p,p-1}$ represent the $(p-1)\times 1$ vector of the last column of $\boldsymbol{S}_p$ excluding s_{pp}. Thus

$$\boldsymbol{S}_1 = s_{11},\ \boldsymbol{S}_2 = \begin{pmatrix} s_{11} & s_{21} \\ s_{21} & s_{22} \end{pmatrix},\ \cdots, \boldsymbol{S}_p = \begin{pmatrix} \boldsymbol{S}_{p-1} & \boldsymbol{s}_{p,p-1} \\ \boldsymbol{s}_{p,p-1}' & s_{pp} \end{pmatrix}. \tag{31}$$

Fact 4 *Under the prior* $\pi_{\boldsymbol{a}}$ *in (13), the marginal posterior of* $\mathbf{\Psi}$ *is given as follows. (a) For given* $(\psi_{11}, \cdots, \psi_{pp})$*, the conditonal posteriors of the off-diagonal vector* $\boldsymbol{\psi}_{i-1,i}$ *are independent normal,*

$$(\boldsymbol{\psi}_{i,i-1} \mid \psi_{ii}, \boldsymbol{S}) \sim N(-\psi_{ii}\boldsymbol{S}_{i-1}^{-1}\boldsymbol{s}_{i,i-1},\ \boldsymbol{S}_{i-1}^{-1}). \tag{32}$$

(b) The marginal posteriors of ψ_{ii}^2 $(1 \le i \le p)$ *are independent gamma* $((n - a_i)/2, w_i/2)$*, where*

$$w_i \ = \ \begin{cases} s_{11}, & \text{if } i = 1, \\ \frac{|\boldsymbol{S}_i|}{|\boldsymbol{S}_{i-1}|} \ = \ s_{ii} - \boldsymbol{s}_{i,i-1}'\boldsymbol{S}_{i-1}^{-1}\boldsymbol{s}_{i-1,i}, & \text{if } i = 2, \cdots, p. \end{cases} \tag{33}$$

(c) The marginal likelihood of $\boldsymbol{Y}$ *(or the normalizing constant) is*

$$\begin{aligned} M \ &\equiv \ \int L(\boldsymbol{\mu}, \mathbf{\Psi})\pi_{\boldsymbol{a}}(\boldsymbol{\mu}, \mathbf{\Psi}) d\boldsymbol{\mu} d\mathbf{\Psi} \\ &= \ \frac{\prod_{i=1}^{p} \Gamma(\frac{1}{2}(n - a_i)) 2^{(n-a_i)/2}}{2^p (2\pi)^{(n-p)p/2} n^{p/2}} \frac{\prod_{i=1}^{p-1} |\boldsymbol{S}_i|^{(a_i - a_{i+1} - 1)/2}}{|\boldsymbol{S}|^{(n-a_p)/2}}. \end{aligned} \tag{34}$$

3.2.1. *Constructive Posteriors*

In the remainder of the paper we use, without further comment, the notation that * appended to a random variable denotes randomness arising from the constructive posterior (i.e., from the random variables used in simulation from the posterior), while a random variable without a * refers to randomness arising from the (frequentist) distribution of a statistic. Also, let Z_{ij} denote standard normal random variables. Whenever several of these occur in an expression, they are all independent (except that random variables of the same type and with the same index refer to the same random variable). Finally, we reserve quantile notation for posterior quantiles, with respect to the * distributions.

Fact 5 *Consider the prior* $\pi_{\boldsymbol{a}}$ *in (13). Let* $\chi^{2*}_{n-a_i}$ *denote independent draws from chi-squared distributions with the indicated degree of freedoms, and* $\boldsymbol{z}^*_{i,i-1}$ *denote*

independent draws from $N_{i-1}(\mathbf{0}, \boldsymbol{I}_{i-1})$. *The constructive posterior of* $(\psi_{11}, \cdots, \psi_{pp}, \boldsymbol{\psi}_{2,1}, \cdots, \boldsymbol{\psi}_{p,p-1})$ *given* $\boldsymbol{X}$ *can be expressed as*

$$\begin{aligned} \psi^*_{ii} &= \sqrt{\frac{\chi^{2*}_{n-a_i}}{w_i}}, \quad i = 1, \cdots, p\,, \qquad (35) \\ \boldsymbol{\psi}^*_{i,i-1} &= \boldsymbol{S}^{-1/2}_{i-1}\boldsymbol{z}^*_{i,i-1} - \psi^*_{ii}\boldsymbol{S}^{-1}_{i-1}\boldsymbol{s}_{i,i-1} \\ &= \boldsymbol{S}^{-1/2}_{i-1}\boldsymbol{z}^*_{i,i-1} - \sqrt{\frac{\chi^{2*}_{n-a_i}}{w_i}}\boldsymbol{S}^{-1}_{i-1}\boldsymbol{s}_{i,i-1}, \quad i = 2, \cdots, p\,. \qquad (36) \end{aligned}$$

Letting $\boldsymbol{\Psi}^* = (\psi^*_{ij})$, *the constructive posterior of* $\boldsymbol{\Sigma}$ *is simply* $\boldsymbol{\Sigma}^* = \boldsymbol{\Psi}^{*-1}(\boldsymbol{\Psi}^{*-1})'$.

Alternatively, let $\boldsymbol{V}$ be the Cholesky decomposition of $\boldsymbol{S}$, i.e., $\boldsymbol{V}$ is the lower-triangular matrix with positive diagonal elements such that $\boldsymbol{S} = \boldsymbol{V}\boldsymbol{V}'$. It is easy to see that

$$w_i = t^2_{ii}, \quad i = 1, \cdots, p. \qquad (37)$$

We will also write $\boldsymbol{V} = \boldsymbol{V}_p$ and let $\boldsymbol{v}_{p,p-1}$ represent the $(p-1)\times 1$ vector of the last column of $\boldsymbol{V}_p$ excluding v_{pp}. We have

$$\boldsymbol{V}_1 = v_{11}, \ \boldsymbol{V}_2 = \begin{pmatrix} v_{11} & 0 \\ v_{21} & v_{22} \end{pmatrix}, \ \cdots, \boldsymbol{V}_p = \begin{pmatrix} \boldsymbol{V}_{p-1} & 0 \\ \boldsymbol{v}'_{p,p-1} & v_{pp} \end{pmatrix}. \qquad (38)$$

Corollary 0.4 *Under the prior* $\pi_{\boldsymbol{a}}$ *in (13),*

$$\begin{aligned} (\boldsymbol{\psi}_{i,i-1} \mid \psi_{ii}, \boldsymbol{V}) &\sim N(-\psi_{ii}\boldsymbol{V}^{-1\prime}_{i-1}\boldsymbol{v}_{i,i-1}, (\boldsymbol{V}_{i-1}\boldsymbol{V}'_{i-1})^{-1}), \qquad (39) \\ (\psi^2_{ii} \mid \boldsymbol{V}) &\sim \textit{inverse gamma}((n-a_i)/2, v^2_{ii}/2). \qquad (40) \end{aligned}$$

Proof. This follows from Fact 4 and the equality $\boldsymbol{S}^{-1}_{i-1}\boldsymbol{s}_{i,i-1} = \boldsymbol{V}^{-1\prime}_{i-1}\boldsymbol{v}_{i,i-1}$. □

The following fact is immediate.

Fact 6 *Under the assumptions of Fact 5, the constructive posterior of* $(\psi_{11}, \cdots, \psi_{pp}, \boldsymbol{\psi}_{21}, \cdots, \boldsymbol{\psi}_{p,p-1})$ *given* $\boldsymbol{X}$ *can be expressed as*

$$\begin{aligned} \psi^*_{ii} &= \frac{\sqrt{\chi^{2*}_{n-a_i}}}{v_{ii}}, \quad i = 1, \cdots, p, \qquad (41) \\ \boldsymbol{\psi}^*_{i,i-1} &= \boldsymbol{V}^{-1\prime}_{i-1}\boldsymbol{z}^*_{i,i-1} - \frac{\sqrt{\chi^{2*}_{n-a_i}}}{v_{ii}}\boldsymbol{V}^{-1\prime}_{i-1}\boldsymbol{v}_{i,i-1}, \quad i = 2, \cdots, p. \qquad (42) \end{aligned}$$

3.2.2. *Posterior Means of* $\boldsymbol{\Sigma}$ *and* $\boldsymbol{\Sigma}^{-1}$

Fact 7 *(a) If* $n - a_i > 0, \ i = 1, \cdots, p$, *then*

$$E(\boldsymbol{\Sigma}^{-1} \mid \boldsymbol{X}) = E(\boldsymbol{\Psi}'\boldsymbol{\Psi} \mid \boldsymbol{X}) = \boldsymbol{V}^{-1\prime}\mathrm{diag}(g_1, \cdots, g_p)\boldsymbol{V}^{-1}, \qquad (43)$$

where $g_i = n - a_i + p - i$, $i = 1, \cdots, p$.

(b) If $n - a_i > 2$, $i = 1, \cdots, p$, *then*

$$E(\boldsymbol{\Sigma} \mid \boldsymbol{X}) \quad = \quad E((\boldsymbol{\Psi}'\boldsymbol{\Psi})^{-1} \mid \boldsymbol{X}) = \boldsymbol{V}\,\mathrm{diag}(h_1, \cdots, h_p)\boldsymbol{V}', \tag{44}$$

where $h_1 = u_1$, $h_j = u_j \prod_{i=1}^{j-1}(1 + u_i)$, $j = 2, \cdots, p$, *with* $u_i = 1/(n - a_i - 2)$, $i = 1, \cdots, p$.

Proof. Letting $\boldsymbol{Y} = \boldsymbol{\Psi}\boldsymbol{V}$, then $\boldsymbol{Y} = (y_{ij})_{p\times p}$ is still lower-triangular and

$$[\boldsymbol{Y} \mid \boldsymbol{X}] \quad \propto \quad \prod_{i=1}^{p} (y_{ii}^2)^{(n-a_i-1)/2} \exp\Big\{-\frac{1}{2} tr(\boldsymbol{Y}\boldsymbol{Y}')\Big\}. \tag{45}$$

From above, we know that all y_{ij}, $1 \le i \le j \le p$, are independent and

$$\begin{aligned} y_{ij} \quad &\sim \quad N(0,1), \quad 1 \le j < i \le p\,; \\ y_{ii} \quad &\sim \quad (y_{ii}^2)^{(n-a_i-1)/2} \exp(-\frac{1}{2} y_{ii}^2), \quad 1 \le i \le p. \end{aligned}$$

If $n - a_i > 0$, $i = 1, \cdots, p$, then $(y_{ii}^2 \mid \boldsymbol{X}) \sim gamma((n-a_i)/2,\ 1/2)$ and $E(y_{ii}^2 \mid \boldsymbol{X})$ exists. Thus it is straightforward to get (43). For (44), we just need to show $E\{(\boldsymbol{Y}\boldsymbol{Y}')^{-1} \mid \boldsymbol{X}) = \mathrm{diag}(h_1, \cdots, h_p)$. Under the condition $n - a_i > 1$, $E(y_{ii}^{-2} \mid \boldsymbol{X})$ exists and is equal to u_i, $i = 1, \cdots, p$. Thus we obtain the result using the same procedure as in Eaton and Olkin (1987). □

4. FREQUENTIST COVERAGE AND MARGINALIZATION PARADOXES

4.1. *Frequentist Coverage Probabilities and Exact Matching*

In this subsection we compare the frequentist properties of posterior credible intervals for various quantities under the prior $\pi_{\mathbf{a}}$, given in (13). As is customary in such comparisons, we study one-sided intervals $(\theta_L, q_{1-\alpha}(\mathbf{x}))$ of a parameter θ, where θ_L is the lower bound on the parameter θ (e.g., 0 or $-\infty$) and $q_{1-\alpha}(\mathbf{x})$ is the posterior quantile of θ, defined by

$$P(\theta < q_{1-\alpha}(\boldsymbol{x}) \mid \boldsymbol{x}) = 1 - \alpha.$$

Of interest is the frequentist coverage of the corresponding confidence interval, i.e.,

$$P(\theta < q_{1-\alpha}(\mathbf{X}) \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

The closer this coverage is to the nominal $1 - \alpha$, the better the procedure (and corresponding objective prior) is judged to be.

Berger and Sun (2006) showed that, when $p = 2$, the right-Haar prior is exact matching prior for many functions of parameters of the bivariate normal distribution. Here we generalize the results to the multivariate normal distribution.

To prove frequentist matching, note first that $(\boldsymbol{S} \mid \boldsymbol{\Sigma}) \sim Wishart(n - 1, \boldsymbol{\Sigma})$. It is easy to see that the joint density for $\boldsymbol{V}$ (the Chelosky decomposition of $\boldsymbol{S}$), given $\boldsymbol{\Psi}$, is

$$f(\boldsymbol{V} \mid \boldsymbol{\Psi}) \propto \prod_{i=1}^{p} v_{ii}^{n-i-1} etr\Big(-\frac{1}{2}\boldsymbol{\Psi}\boldsymbol{V}\boldsymbol{V}'\boldsymbol{\Psi}'\Big). \tag{46}$$

The following technical lemmas are also needed. The first lemma follows from the expansion

$$tr(\mathbf{\Psi V V'\Psi'}) = \sum_{i=1}^{p} \psi^2 v_{ii}^2 + \sum_{i=1}^{p}\sum_{j=1}^{i-1}\Big(\sum_{k=1}^{i}\psi_{ik}v_{kj})\Big)^2. \tag{47}$$

The proofs for both lemmas are straightforwad and are omitted.

Lemma 1 *For $n \geq p$ and given $\mathbf{\Sigma}^{-1} = \mathbf{\Psi\Psi'}$, the following random variables are independent and have the indicated distributions:*

$$Z_{ij} = \psi_{ii}\Big(v_{ij} + \sum_{k=1}^{i-1} t_{ik}v_{kj}\Big) \sim N(0,1), \tag{48}$$

$$\psi_{ii}v_{ii} = \chi^2_{n-i}. \tag{49}$$

Lemma 2 *Let $Y_{1-\alpha}$ denote the $1-\alpha$ quantile of any random variable Y.*
(a) If $g(\cdot)$ is a monotonically increasing function, $[g(Y)]_{1-\alpha} = g(Y_{1-\alpha})$ for any $\alpha \in (0,1)$.
(b) If W is a positive random variable, $(WY)_{1-\alpha} \geq 0$ if and only if $Y_{1-\alpha} \geq 0$.

Theorem 1 *(a) For any $\alpha \in (0,1)$ and fixed $i = 1, \cdots, p$, the posterior $1-\alpha$ quantile of ψ_{ii} has the expression*

$$(\psi^*_{ii})_{1-\alpha} = \frac{\sqrt{(\chi^{2*}_{n-a_i})_{1-\alpha}}}{v_{ii}}. \tag{50}$$

(b) For any $\alpha \in (0,1)$ and any $(\boldsymbol{\mu}, \mathbf{\Psi})$, the frequentist coverage probability of the credible interval $(0, (\psi^*_{ii})_{1-\alpha})$ is

$$P\Big(\psi_{ii} < (\psi^*_{ii})_{1-\alpha} \mid \boldsymbol{\mu}, \mathbf{\Psi}\Big) = P\Big(\chi^2_{n-i} < (\chi^{2*}_{n-a_i})_{1-\alpha}\Big), \tag{51}$$

which does not depend on $(\boldsymbol{\mu}, \mathbf{\Psi})$ and equals $1-\alpha$ if and only if $a_i = i$.

Corollary 1.1 *For any $\alpha \in (0,1)$, the posterior quantile of $d_i = var(x_i \mid x_1, \cdots, x_{i-1})$ is $(d^*_i)_{1-\alpha} = \frac{v^2_{ii}}{(\chi^{2*}_{n-a_i})_\alpha}$. For any $(\boldsymbol{\mu}, \mathbf{\Sigma})$, the frequentist coverage probability of the credible interval $(0, (d^*_i)_{1-\alpha}) = \Big(\chi^2_{n-i} < (\chi^{2*}_{n-a_i})_{1-\alpha}\Big)$, is a constant $P\Big(\chi^2_{n-i} > (\chi^{2*}_{n-a_i})_\alpha\Big)$, and equals $1-\alpha$ if and only if $a_i = i$.*

Observing that $|\mathbf{\Sigma_i}| = \prod_{\mathbf{j=1}}^{\mathbf{i}} \mathbf{d_j}$ yields the following result.

Theorem 2 *(a) For any α, the posterior $1-\alpha$ quantile of $|\mathbf{\Sigma}_i|$ has the expression*

$$(|\mathbf{\Sigma}_i|)_{1-\alpha} = \frac{\prod_{j=1}^{i} v^2_{jj}}{\Big(\prod_{j=1}^{i}\chi^{2*}_{n-a_j}\Big)_\alpha}. \tag{52}$$

(b) For any $\alpha \in (0,1)$ *and any* $(\boldsymbol{\mu}, \boldsymbol{\Psi})$*, the frequentist coverage probability of the credible interval* $(0, (|\boldsymbol{\Sigma}_i|)_{1-\alpha})$ *is*

$$P(|\boldsymbol{\Sigma}_i| < (|\boldsymbol{\Sigma}_i|)_{1-\alpha} \mid \boldsymbol{\mu}, \boldsymbol{\Psi}) = P\Big(\prod_{j=1}^{i} \chi^2_{n-j} > (\prod_{j=1}^{i} \chi^{2*}_{n-a_j})_\alpha \Big), \tag{53}$$

which is a constant and equals $1-\alpha$ *if and only if* $(a_1, \cdots, a_i)$ *is a permutation of* $(1, \cdots, i)$*.*

For the bivariate normal case, Berger and Sun (2006) showed that the right-Haar measure is the exact matching prior for ψ_{21} and t_{12}. We also expect that, for the multivariate normal distribution, the right-Haar prior is exact matching for all ψ_{ij} and t_{ij}.

4.2. *Marginalization Paradoxes*

While the Bayesian credible intervals for many parameters under the right-Haar measure are exact matching priors, it can be seen that the prior can suffer marginalization paradoxes. The basis for such paradoxes (Dawid, Stone, and Zidek (1973)) is that any proper prior has the property: if the marginal posterior distribution for a parameter θ depends only on a statistic T – whose distribution in turn depends only on θ – then the posterior of θ can be derived from the distribution of T together with the marginal prior for θ. While this is a basic property of any proper Bayesian prior, it can be violated for improper priors, with the result then called a marginalization paradox.

In Berger and Sun (2006), it was shown that, when using the right-Haar prior, the posterior distribution of the correlation coefficient ρ for a bivariate normal distribution depends only on the sample correlation coefficient r. Brillinger (1962) showed that there does not exist a prior $\pi(\rho)$ such that the this posterior density equals $f(r \,|\, \rho)\pi(\rho)$, where $f(r \,|\, \rho)$ is the density of r given ρ. This thus provides an example of a marginalization paradox.

Here is another marginalization paradox in the bivariate normal case. We know from Berger and Sun (2006) that the right-Haar prior π_H is exact matching prior for ψ_{21}. Note that the constructive posterior of ψ_{21} is

$$\frac{Z^*}{\sqrt{s_{11}}} - \frac{\sqrt{\chi^{2*}_{n-2}}}{\sqrt{s_{11}}} \frac{r}{\sqrt{1-r^2}}, \tag{54}$$

which clearly depends only on (s_{11}, r).

It turns out that the joint density of (s_{11}, r) depends only on (σ_{11}, ρ). Note that the posterior of (σ_{11}, ρ) based on the product of $f(s_{11}, r \mid \sigma_{11}, \rho)$ and the marginal prior for (σ_{11}, ρ) based on π_H is different from the marginal posterior of (σ_{11}, ρ) based on π_H. Consequently, the posterior distribution of ψ_{21} from the right haar provides another example of the marginalization paradox.

It is somewhat controversial as to whether violation of the marginalization paradox is a serious problem. For instance, in the bivariate normal problem, there is probably no proper prior distribution that yields a marginal posterior distribution of ρ which depends only on r, so the relevance of an unattainable property of proper priors could be questioned.

In any case, this situation provides an interesting philosophical conundrum of a type that we have not previously seen: a complete objective Bayesian and frequentist unification can be obtained for inference about the usual parameters of the bivariate normal distribution, but only if violation of the marginalization paradox is accepted. The prior π_{CE} does avoid the marginalization paradox for ρ_{12}, but is not exact frequentist matching. We, alas, know of no way to adjudicate between the competing goals of exact frequentist matching and avoidance of the marginalization paradox, and so will simply present both as possible objective Bayesian approaches.

5. ON THE NON-UNIQUENESS OF RIGHT-HAAR PRIORS

While the right-Haar priors seem to have some very nice properties, the fact that they depend on the particular lower triangular matrix decomposition of $\mathbf{\Sigma}^{-1}$ that is used is troubling. In the bivariate case, for instance, both

$$\pi_1(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_2^2(1-\rho^2)} \quad \text{and} \quad \pi_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1^2(1-\rho^2)}$$

are right-Haar priors (expressed with respect to $d\mu_1\, d\mu_2\, d\sigma_1\, d\sigma_2\, d\rho$).

There are several natural proposals for dealing with this non-uniqueness. One is to mix over the right-Haar priors. Another is to choose the 'empirical Bayes' right-Haar prior, that which maximizes the marginal likelihood of the data. These proposals are developed in the next two subsections. The last subsection shows, quite surprisingly, that neither of these solutions works! For simplicity, we restrict attention to the bivariate normal case.

5.1. *Symmetrized Right-Haar Priors*

Consider the symmetrized right-Haar prior

$$\begin{aligned} \tilde{\pi}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) &= \pi_1(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) + \pi_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \\ &= \frac{1}{\sigma_1^2(1-\rho^2)} + \frac{1}{\sigma_2^2(1-\rho^2)}. \end{aligned} \tag{55}$$

This can be thought of as a 50-50 mixture of the two right-Haar priors.

Fact 8 *The joint posterior of* $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ *under the prior* $\tilde{\pi}$ *is given by*

$$\begin{aligned} \tilde{\pi}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho \mid \boldsymbol{X}) &= C\pi_1(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho \mid \boldsymbol{X}) \\ &+ (1-C)\pi_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho \mid \boldsymbol{X}), \end{aligned} \tag{56}$$

where

$$C = \frac{s_{11}^{-1}}{s_{11}^{-1} + s_{22}^{-1}}. \tag{57}$$

and $\pi_1(\cdot \mid \boldsymbol{X})$ *and* $\pi_2(\cdot \mid \boldsymbol{X})$ *are the posteriors under the priors* π_1 *and* π_2*, respectively.*

Proof. Let $p = 2$ and $(a_1, a_2) = (1, 2)$ in (34). We get

$$\begin{aligned} C_j &= \int L(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)\pi_j(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)d\mu_1 d\mu_2 d\sigma_1 d\sigma_2 d\rho \\ &= \frac{\Gamma(\frac{n-1}{2})\Gamma(\frac{n-2}{2})2^{(n-2)/2}s_{jj}^{-1}}{\pi^{(n-3)/2}|\boldsymbol{S}|^{(n-2)/2}}, \end{aligned} \tag{58}$$

for $j = 1, 2$. The result is immediate. □

For later use, note that, under the prior $\tilde{\pi}$, the posterior mean of $\boldsymbol{\Sigma}$ has the form

$$\widehat{\boldsymbol{\Sigma}}_S = \mathrm{E}(\boldsymbol{\Sigma} \mid \boldsymbol{X}) \quad \equiv \quad \mathrm{E}(\boldsymbol{\Sigma} \mid \boldsymbol{X}) = C\,\widehat{\boldsymbol{\Sigma}}_1 + (1 - C)\,\widehat{\boldsymbol{\Sigma}}_2, \tag{59}$$

where $\widehat{\boldsymbol{\Sigma}}_i$ is the posterior mean under $\pi_i(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, given by

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n-3}(\boldsymbol{S} + \boldsymbol{G}_i), \tag{60}$$

where

$$\boldsymbol{G}_1 = \begin{pmatrix} 0 & 0 \\ 0 & \frac{2}{n-4}\left(s_{22} - \frac{s_{12}^2}{s_{11}}\right) \end{pmatrix}, \quad \boldsymbol{G}_2 = \begin{pmatrix} \frac{2}{n-4}\left(s_{11} - \frac{s_{12}^2}{s_{22}}\right) & 0 \\ 0 & 0 \end{pmatrix}. \tag{61}$$

Here $\boldsymbol{\Sigma}_1$ is a special case of (44) when $p = 2$ and $(a_1, a_2) = (1, 2)$.

5.2. *The Empirical Bayes Right-Haar Prior*

The right-Haar priors above were essentially just obtained by coordinate permutation. More generally, one can obtain other right-Haar priors by orthonormal transformations of the data. In particular, define the orthonormal matrix

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{pmatrix},$$

where the $\boldsymbol{\gamma}_i$ are orthonormal row vectors. Consider the transformation of the data $\boldsymbol{\Gamma x}$, so that the resulting sample covariance matrix is

$$\boldsymbol{S}^* = \boldsymbol{\Gamma S \Gamma}' = \begin{pmatrix} s_{11}^* & s_{12}^* \\ s_{12}^* & s_{22}^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma}_1 \boldsymbol{S} \boldsymbol{\gamma}_1' & \boldsymbol{\gamma}_1 \boldsymbol{S} \boldsymbol{\gamma}_2' \\ \boldsymbol{\gamma}_2 \boldsymbol{S} \boldsymbol{\gamma}_1' & \boldsymbol{\gamma}_2 \boldsymbol{S} \boldsymbol{\gamma}_2' \end{pmatrix}. \tag{62}$$

The right-Haar prior can be defined in this transformed problem, so that each $\boldsymbol{\Gamma}$ defines a different right-Haar prior.

A commonly employed technique when facing a class of priors, as here, is to choose the 'empirical Bayes' prior, that which maximizes the marginal likelihood of the data. This is given in the following lemma.

Lemma 3 *The empirical Bayes right-Haar prior is given by that* $\boldsymbol{\Gamma}$ *for which*

$$\begin{aligned} s_{11}^* &= \frac{1}{2}(s_{11} + s_{22}) - \frac{1}{2}\sqrt{(s_{11} - s_{22})^2 + 4s_{12}^2}\,, \\ s_{12}^* &= 0\,, \\ s_{22}^* &= \frac{1}{2}(s_{11} + s_{22}) + \frac{1}{2}\sqrt{(s_{11} - s_{22})^2 + 4s_{12}^2}\,. \end{aligned}$$

(Note that the two eigenvalues of $\boldsymbol{S}$ are s^*_{11} and s^*_{22}. Thus this is the orthonormal transformation such that the sample variance of the first coordinate is the smallest eigenvalue.)

Proof. Noting that $|\boldsymbol{S}^*| = |\boldsymbol{S}|$, it follows from (58) that the marginal likelihood of $\boldsymbol{\Gamma}$ is proportional to ${s^*_{11}}^{-1}$. Hence we simply want to find an orthonormal $\boldsymbol{\Gamma}$ to minimize $\boldsymbol{\gamma}_1 \boldsymbol{S} \boldsymbol{\gamma}_1'$. It is standard matrix theory that the minimum is the smallest eigenvalue of $\boldsymbol{S}$, with $\boldsymbol{\gamma}_1$ being the associated eigenvector. Since $\boldsymbol{\Gamma}$ is orthonormal, the remainder of the lemma also follows directly. □

Lemma 4 *Under the empirical Bayes right-Haar prior, the posterior mean of $\boldsymbol{\Sigma}$ is $\widehat{\boldsymbol{\Sigma}}_E = \mathrm{E}(\boldsymbol{\Sigma} \mid \boldsymbol{X})$ and given by*

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_E &= \frac{1}{n-3}\left(\boldsymbol{S} + \frac{s^*_{22}}{n-4}\left(\boldsymbol{I} + \frac{1}{s^*_{22}-s^*_{11}}\begin{pmatrix} s_{11}-s_{22} & 2s_{12} \\ 2s_{12} & s_{22}-s_{11}\end{pmatrix}\right)\right) \\ &= \frac{1}{n-3}\left(\boldsymbol{S} + \frac{s^*_{22}}{n-4}\left(\boldsymbol{I} + \frac{1}{s^*_{22}-s^*_{11}}\boldsymbol{S} - \frac{1}{\frac{1}{s^*_{11}}-\frac{1}{s^*_{22}}}\boldsymbol{S}^{-1}\right)\right).\end{aligned}$$

Proof. Under the empirical Bayes right-Haar prior, the posterior mean of $\boldsymbol{\Sigma}^* = \boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma}'$ is

$$\mathrm{E}(\boldsymbol{\Sigma}^* \mid \boldsymbol{X}) = \frac{1}{n-3}(\boldsymbol{S}^* + \boldsymbol{G}^*),$$

where

$$\boldsymbol{G}^* = \begin{pmatrix} 0 & 0 \\ 0 & g^*\end{pmatrix}, \quad g^* = \frac{2}{n-4}\left(s^*_{22} - \frac{s^{*2}_{12}}{s^*_{11}}\right) = \frac{2s^*_{22}}{n-4}.$$

So the corresponding estimate of $\boldsymbol{\Sigma}$ is

$$\mathrm{E}(\boldsymbol{\Sigma} \mid \boldsymbol{X}) = \boldsymbol{\Gamma}' \,\mathrm{E}(\boldsymbol{\Sigma}^* \mid \boldsymbol{X})\, \boldsymbol{\Gamma} = \frac{1}{n-3}(\boldsymbol{S} + \boldsymbol{\Gamma}'\boldsymbol{G}^*\boldsymbol{\Gamma}).$$

Computation yields that the eigenvector $\boldsymbol{\gamma}_2$ is such that

$$\begin{aligned}\gamma^2_{21} &= \frac{1}{2} + \frac{1}{2}\frac{s_{11}-s_{22}}{\sqrt{(s_{11}-s_{22})^2 + 4s^2_{12}}}, \\ \gamma^2_{22} &= \frac{1}{2} - \frac{1}{2}\frac{s_{11}-s_{22}}{\sqrt{(s_{11}-s_{22})^2 + 4s^2_{12}}}, \\ \gamma_{21}\gamma_{22} &= \frac{s_{12}}{\sqrt{(s_{11}-s_{22})^2 + 4s^2_{12}}}.\end{aligned}$$

Thus

$$\begin{aligned}\boldsymbol{\Gamma}'\boldsymbol{G}^*\boldsymbol{\Gamma} &= g^*\boldsymbol{\gamma}_2'\boldsymbol{\gamma}_2 \\ &= \frac{s^*_{22}}{n-4}\left(\boldsymbol{I} + \frac{1}{\sqrt{(s_{11}-s_{22})^2 + 4s^2_{12}}}\begin{pmatrix} s_{11}-s_{22} & 2s_{12} \\ 2s_{12} & s_{22}-s_{11}\end{pmatrix}\right). \\ &= \frac{s^*_{22}}{n-4}\left(\boldsymbol{I} + \frac{1}{s^*_{22}-s^*_{11}}\begin{pmatrix} s_{11}-s_{22} & 2s_{12} \\ 2s_{12} & s_{22}-s_{11}\end{pmatrix}\right).\end{aligned}$$

The last expression in the lemma follows from algebra. □

5.3. *Decision-Theoretic Evaluation*

To study the effectiveness of the symmetrized right-Haar prior and the empirical Bayes right-Haar prior, we turn to a decision theoretic evaluation, utilizing a natural invariant loss function.

For a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with unknown $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, a natural loss to consider is the entropy loss, defined by

$$\begin{aligned} L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= 2 \int \log \left\{ \frac{f(\boldsymbol{X} \mid \boldsymbol{\mu}, \Sigma)}{f(\boldsymbol{X} \mid \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})} \right\} f(\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{X} \\ &= (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})' \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) + tr(\hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}) - \log |\hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}| - p. \quad (63) \end{aligned}$$

Clearly, the entropy loss has two parts, one is related to the means $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}$ (with $\hat{\boldsymbol{\Sigma}}$ as the weight matrix), and the other is related to $\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}$. The last three terms of this expression are related to 'Stein's loss,' and is the most commonly used losses for estimation of a covariance matrix (cf. James and Stein (1961) and Haff (1977)).

Lemma 5 *Under the loss (63) and for any of the priors considered in this paper, the generalized Bayesian estimator of* $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *is*

$$\hat{\boldsymbol{\mu}}_B = \mathrm{E}(\boldsymbol{\mu} \mid \boldsymbol{X}) = (\bar{x}_1, \bar{x}_2)', \quad (64)$$

$$\hat{\boldsymbol{\Sigma}}_B = \mathrm{E}(\boldsymbol{\Sigma} \mid \boldsymbol{X}) + \mathrm{E}\{(\hat{\boldsymbol{\mu}}_B - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}}_B - \boldsymbol{\mu}) \mid \boldsymbol{X}\} = \frac{n+1}{n} \mathrm{E}(\boldsymbol{\Sigma} \mid \boldsymbol{X}). \quad (65)$$

Proof. For the priors we consider in the paper,

$$[\boldsymbol{\mu} \,|\, \boldsymbol{\Sigma}, \boldsymbol{X}] \sim N_2 \left((\bar{x}_1, \bar{x}_2)', \frac{1}{n} \boldsymbol{\Sigma} \right), \quad (66)$$

so that (64) is immediate. Furthermore, it follows that

$$\mathrm{E}((\hat{\boldsymbol{\mu}}_B - \boldsymbol{\mu})' \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_B - \boldsymbol{\mu}) \mid \boldsymbol{X}) = \frac{1}{n} tr(\hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}) \quad (67)$$

so that the remaining goal is to choose $\hat{\boldsymbol{\Sigma}}$ so as to minimize

$$\begin{aligned} &\mathrm{E}\Big((1 + \frac{1}{n}) tr(\hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}) - \log |\hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}| - p \,|\, \boldsymbol{X} \Big) \\ &= \mathrm{E}\left(tr(\hat{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\Sigma}}) - \log |\hat{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\Sigma}}| - p \mid \boldsymbol{X} \right) + \log(1 + \frac{1}{n}), \quad (68) \end{aligned}$$

where $\tilde{\boldsymbol{\Sigma}} = (1 + \frac{1}{n}) \boldsymbol{\Sigma}$. It is standard (see, e.g., Eaton, 1989) that the first term on the right hand side of the last expression is minimized at

$$\hat{\boldsymbol{\Sigma}} = \mathrm{E}(\tilde{\boldsymbol{\Sigma}} \mid \boldsymbol{X}) = (1 + \frac{1}{n}) \mathrm{E}(\boldsymbol{\Sigma} \mid \boldsymbol{X}), \quad (69)$$

from which the result is immediate. □

We now turn to frequentist decision-theoretic evaluation of the various posterior estimates that arise from the reference priors considered in the paper. Thus we now change perspective and consider $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to be given, and consider the frequentist risk of the posterior estimates $\hat{\boldsymbol{\mu}}_B(\boldsymbol{X})$ and $\widehat{\boldsymbol{\Sigma}}_B(\boldsymbol{X})$, now considered as functions of $\boldsymbol{X}$. Thus we evaluate the frequentist risk

$$R(\hat{\boldsymbol{\mu}}_B, \widehat{\boldsymbol{\Sigma}}_B; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{E} L(\hat{\boldsymbol{\mu}}_B(\boldsymbol{X}), \widehat{\boldsymbol{\Sigma}}_B(\boldsymbol{X}); \boldsymbol{\mu}, \boldsymbol{\Sigma}) \,, \tag{70}$$

where the expectation is over $\boldsymbol{X}$ given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The following lemma states that we can reduce the frequentist risk comparison to a comparison of the frequentist risks of the various posterior means for $\boldsymbol{\Sigma}$ under Stein's loss. It's proof is virtually identical to that of Lemma 5, and is omitted.

Lemma 6 *For frequentist comparison of the various Bayes estimators considered in the paper, it suffices to compare the frequentist risks of the* $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{X}) = \mathrm{E}(\boldsymbol{\Sigma} \mid \boldsymbol{X})$*, with respect to*

$$R(\widehat{\boldsymbol{\Sigma}}(\boldsymbol{X}); \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{E}\left(tr(\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X})\boldsymbol{\Sigma}) - \log|\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X})\boldsymbol{\Sigma}| - p \right) , \tag{71}$$

where the expectation is with respect to $\boldsymbol{X}$*.*

Lemma 7 *Under the right haar prior* π_H*, the risk function (71) is a constant, given by*

$$R(\widehat{\boldsymbol{\Sigma}}(\boldsymbol{X}); \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{j=1}^{p} \log(h_j) + \sum_{j=1}^{p} \mathrm{E} \log(\chi^2_{n-j}). \tag{72}$$

where h_j *is given by (44).*

The proof of this can be found in Eaton (1989). If $p = 2$, it follows that the risk for the two right-Haar priors is

$$\log(n-2) - 2\log(n-3) - \log(n-4) + \mathrm{E}\log(\chi^2_{n-1}) + \mathrm{E}\log(\chi^2_{n-2}).$$

For instance, when $n = 10$, this risk is approximately 0.4271448.

Table 2 gives the risks for the estimates arising from the two right-Haar priors, $\widehat{\boldsymbol{\Sigma}}_1$ and $\widehat{\boldsymbol{\Sigma}}_2$, the estimate $\widehat{\boldsymbol{\Sigma}}_S$ arising from the symmetrized right-Haar prior, the estimate $\widehat{\boldsymbol{\Sigma}}_E$ arising from the empirical Bayes right-Haar prior, $\widehat{\boldsymbol{\Sigma}}_{R\rho}$ arising from the reference prior for ρ, and an estimate in the spirit of Dey and Srinivasan (1985) and Dey (1988) that will be discussed shortly.

The simulated risks are given in the Table 2 for $\widehat{\boldsymbol{\Sigma}}_1$ and $\widehat{\boldsymbol{\Sigma}}_2$ instead of the exact risks, because the comparisons between estimates is then more meaningful (the simulation errors being highly correlated since the estimates were all based on common realizations of sample covariance matrices).

The first surprise here is that the risk of $\widehat{\boldsymbol{\Sigma}}_S$ is actually worse than the risk of the right-Haar prior estimates. This is in contradiction to the usual belief that,

Table 2: *Frequentist risks of various estimates of* $\boldsymbol{\Sigma}$ *when* $n = 10$ *and for various choices of* $\boldsymbol{\Sigma}$*. These were computed by simulation, using 10,000 generated values of* $\boldsymbol{S}$

$(\sigma_1, \sigma_2, \rho)$	$R(\widehat{\boldsymbol{\Sigma}}_1)$	$R(\widehat{\boldsymbol{\Sigma}}_2)$	$R(\widehat{\boldsymbol{\Sigma}}_S)$	$R(\widehat{\boldsymbol{\Sigma}}_E)$	$R(\widehat{\boldsymbol{\Sigma}}_D)$	$R(\widehat{\boldsymbol{\Sigma}}_{R\rho})$
$(1, 1, 0)$	.4287	.4288	.4452	.6052	.3833	.4095
$(1, 2, 0)$	.4278	.4270	.4424	.5822	.3859	.4174
$(1, 5, 0)$	.4285	.4287	.4391	.5404	.3989	.4334
$(1, 50, 0)$	.4254	.4250	.4272	.5100	.4194	.4427
$(1, 1, .1)$	.4255	.4266	.4424	.5984	.3810	.4241
$(1, 1, .5)$	.4274	.4275	.4403	.5607	.3906	.3936
$(1, 1, .9)$	.4260	.4255	.4295	.5159	.4134	.4206
$(1, 1, -.9)$	.4242	.4243	.4280	.5119	.4118	.4219

if considering alternate priors, utilization of a mixture of the two priors will give superior performance.

This would also seem to be in contradiction to the known fact for a convex loss function (such as Stein's loss) that, if two estimators $\widehat{\boldsymbol{\Sigma}}_1$ and $\widehat{\boldsymbol{\Sigma}}_2$ have equal risk functions, then an average of the two estimators will have lower risk. But this refers to a constant average of the two estimators, not a data-weighted average as in $\widehat{\boldsymbol{\Sigma}}_S$. What is particularly striking is that the data-weighted average arises from the posterior marginal likelihoods corresponding to the two different priors, so the posterior seems to be 'getting it wrong,' weighting the 'bad' prior more than the 'good' prior.

This is indicated in even more dramatic fashion by $\widehat{\boldsymbol{\Sigma}}_E$, the empirical Bayes version, which is based on that right-Haar prior which is 'most likely' for given data. In that the risk of $\widehat{\boldsymbol{\Sigma}}_E$ is much worse than even the risk of $\widehat{\boldsymbol{\Sigma}}_S$, it seems that empirical Bayes has selected the worst of all of the right-Haar priors!

The phenomenon arising here is disturbing and sobering. It is yet another indication that improper priors do not behave as do proper priors, and that it can be dangerous to apply 'understandings' from the world of proper priors to the world of improper priors. (Of course, the same practical problems could arise from use of vague proper priors, so use of such is not a solution to the problem.)

From a formal objective Bayesian position (e.g., the viewpoint from the reference prior perspective), there is no issue here. The various reference priors we considered are (by definition) the correct objective priors for the particular contexts (choice of parameterization and parameter of interest) in which they were derived. It is use of these priors – or modifications of them based on 'standard tricks' – out of context that is being demonstrated to be of concern.

APPENDIX A: PROOFS

Proof of Fact 1. The likelihood function of $(\boldsymbol{\mu}, \boldsymbol{\Psi})$ is

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Psi}) \quad \propto \quad |\boldsymbol{\Psi}'\boldsymbol{\Psi}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Psi}'\boldsymbol{\Psi}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

and the log-likelihood is then

$$\log f \quad = \quad const + \sum_{i=1}^{p} \log(\psi_{ii}) - \frac{1}{2}\sum_{i=1}^{p}\left(\sum_{j=1}^{i}\psi_{ij}(x_j-\mu_j)\right)^2.$$

For any fixed $i = 1, \cdots, p$, let $\boldsymbol{\Sigma}_i$ be the variance and covariance matrix of $(x_1, \cdots, x_i)'$. Also, let $\boldsymbol{e}_i$ be the $i \times 1$ vector whose ith element is 1 and 0 otherwise. The Fisher information matrix of $\boldsymbol{\theta}$ is then (11).

Note that $Var(\boldsymbol{x}) = \boldsymbol{\Sigma} = \boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}'^{-1}$. Let $\boldsymbol{\Psi}_i$ be the $i \times i$ left and top sub-matrix of $\boldsymbol{\Psi}$. It is easy to versify that $\boldsymbol{\Sigma}_i = \boldsymbol{\Psi}_i^{-1}\boldsymbol{\Psi}_i'^{-1}$. Using the fact that $|\boldsymbol{B} + \boldsymbol{a}\boldsymbol{a}'| = |\boldsymbol{B}|(1 + \boldsymbol{a}'\boldsymbol{B}^{-1}\boldsymbol{a})$ where $\boldsymbol{B}$ is invertible and $\boldsymbol{a}$ is a vector, we can show that

$$|\boldsymbol{\Lambda}_i| \quad = \quad 2\prod_{j=1}^{i}\frac{1}{\psi_{jj}^2}. \tag{73}$$

From (11) and (73), the reference prior of $\boldsymbol{\Psi}$ for the ordered group $\{\boldsymbol{\mu}, \psi_{11}, (\psi_{21}, \psi_{22}), \cdots, (\psi_{p1}, \cdots, \psi_{pp})\}$, is easy to obtain as (12) according to the algorithm in Berger and Bernardo (1992).

Proof of Fact 2. For $i = 2, \cdots, p$, denote $\boldsymbol{t}_{i,i-1} = \boldsymbol{\psi}_{i,i-1}/\psi_{ii}$. Clearly, the Jacobian from $(\boldsymbol{\psi}_{i,i-1}, \psi_{ii})$ to $(\boldsymbol{t}_{i,i-1}, \psi_{ii})$ is

$$\boldsymbol{J}_i = \frac{\partial(\boldsymbol{\psi}_{i,i-1}, \psi_{ii})}{\partial(\boldsymbol{t}_{i,i-1}, \psi_{ii})} = \begin{pmatrix} \psi_{ii}\boldsymbol{I}_{i-1} & \boldsymbol{t}_{i,i-1} \\ \boldsymbol{0}' & 1 \end{pmatrix}. \tag{74}$$

The Fisher information for $\widetilde{\boldsymbol{\theta}}$ has the form (19), where

$$\widetilde{\boldsymbol{\Lambda}}_i = \boldsymbol{J}_i'\boldsymbol{\Lambda}_i\boldsymbol{J}_i = \boldsymbol{J}_i'\left(\boldsymbol{\Psi}_i^{-1}\boldsymbol{\Psi}_i'^{-1} + \frac{1}{\psi_{ii}^2}\boldsymbol{e}_i\boldsymbol{e}_i'\right)\boldsymbol{J}_i. \tag{75}$$

Note that

$$\boldsymbol{\Psi}_i = \begin{pmatrix} \boldsymbol{\Psi}_{i-1} & \boldsymbol{0} \\ \psi_{ii}\boldsymbol{t}_{i,i-1}' & \psi_{ii} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Psi}_i^{-1} = \begin{pmatrix} \boldsymbol{\Psi}_{i-1}^{-1} & \boldsymbol{0} \\ -\boldsymbol{t}_{i,i-1}'\boldsymbol{\Psi}_{i-1}^{-1} & \frac{1}{\psi_{ii}} \end{pmatrix}.$$

We have that

$$\boldsymbol{J}_i'\boldsymbol{\Psi}_i^{-1} = \begin{pmatrix} \psi_{ii}\boldsymbol{I}_{i-1} & \boldsymbol{t}_{i,i-1} \\ \boldsymbol{0}' & 1 \end{pmatrix}\begin{pmatrix} \boldsymbol{\Psi}_{i-1}^{-1} & \boldsymbol{0} \\ -\boldsymbol{t}_{i,i-1}'\boldsymbol{\Psi}_{i-1}^{-1} & \frac{1}{\psi_{ii}} \end{pmatrix} = \begin{pmatrix} \psi_{ii}\boldsymbol{\Psi}_{i-1}^{-1} & \boldsymbol{0} \\ \boldsymbol{0}' & \frac{1}{\psi_{ii}} \end{pmatrix}. \tag{76}$$

Substituting (76) into (75) and using the fact that $\boldsymbol{J}_i'\boldsymbol{e}_i = \boldsymbol{e}_i$,

$$\widetilde{\boldsymbol{J}}_i = \begin{pmatrix} \psi_{ii}^2\boldsymbol{\Psi}_{i-1}^{-1}\boldsymbol{\Psi}_{i-1}'^{-1} & \boldsymbol{0} \\ \boldsymbol{0}' & \frac{2}{\psi_{ii}^2} \end{pmatrix} = \begin{pmatrix} \psi_{ii}^2\boldsymbol{T}_{i-1}^{-1}\widetilde{\boldsymbol{\Psi}}_{i-1}^{-2}\boldsymbol{T}_{i-1}'^{-1} & \boldsymbol{0} \\ \boldsymbol{0}' & \frac{2}{\psi_{ii}^2} \end{pmatrix}. \tag{77}$$

Part (a) holds. It is easy to see that the upper and left $i \times i$ submatrix of $\mathbf{\Lambda}_i^*$ does not depend on $\boldsymbol{t}_{i,i-1}$. Part (b) can be proved using the algorithm of Berger and Bernardo (1992b). Furthermore, part (c) holds because of $|\boldsymbol{J}_i| = \psi_{ii}^{i-1}$.

$$\widetilde{\pi}_R(\widetilde{\boldsymbol{\theta}}) \prod_{i=2}^{p} \frac{1}{|\boldsymbol{J}_i|} = \prod_{i=1}^{p} \frac{1}{\psi_{ii}^i} = \pi_H(\boldsymbol{\Psi}).$$

Proof of Fact 3. Note that (25) is equivalent to

$$\left\{ \begin{array}{lcl} d_1 & = & \xi_1^{\frac{1}{2}} \ \xi_2^{\frac{1}{3}} \ \cdots\cdots \ \xi_{p-2}^{\frac{1}{p-1}} (\xi_{p-1}\xi_p)^{\frac{1}{p}}, \\ d_2 & = & \xi_1^{-\frac{1}{2}} \xi_2^{\frac{1}{3}} \ \cdots\cdots \ \xi_{p-2}^{\frac{1}{p-1}} (\xi_{p-1}\xi_p)^{\frac{1}{p}}, \\ d_3 & = & \quad \xi_2^{-\frac{2}{3}} \cdots\cdots \ \xi_{p-2}^{\frac{1}{p-1}} (\xi_{p-1}\xi_p)^{\frac{1}{p}}, \\ \cdots & & \qquad\qquad\quad \cdots\cdots \\ d_{p-1} & = & \qquad\qquad\quad \xi_{p-2}^{-\frac{p-2}{p-1}} (\xi_{p-1}\xi_p)^{\frac{1}{p}}, \\ d_p & = & \qquad\qquad\qquad\quad \xi_{p-1}^{-\frac{p-2}{p}} \xi_p^{\frac{1}{p}}. \end{array} \right.$$

Then, the Hessian is

$$\begin{aligned} \boldsymbol{H} &= \frac{\partial(d_1, \cdots, d_p)}{\partial(\xi_1, \cdots, \xi_p)} \\ &= \begin{pmatrix} \frac{d_1}{2\xi_1} & \frac{d_1}{3\xi_2} & \frac{d_1}{4\xi_3} & \cdots & \frac{d_1}{(p-1)\xi_{p-2}} & \frac{d_1}{p\xi_{p-1}} & \frac{d_1}{p\xi_p} \\ -\frac{d_2}{2\xi_1} & \frac{d_2}{3\xi_2} & \frac{d_2}{4\xi_3} & \cdots & \frac{d_2}{(p-1)\xi_{p-2}} & \frac{d_2}{p\xi_{p-1}} & \frac{d_2}{p\xi_p} \\ 0 & \frac{d_3}{3\xi_2} & \frac{d_3}{4\xi_3} & \cdots & \frac{d_3}{(p-1)\xi_{p-2}} & \frac{d_3}{p\xi_{p-1}} & \frac{d_3}{p\xi_p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{(p-2)d_{p-1}}{(p-1)\xi_{p-2}} & \frac{d_{p-1}}{p\xi_{p-1}} & \frac{d_{p-1}}{p\xi_p} \\ 0 & 0 & 0 & \cdots & 0 & -\frac{(p-1)d_{p-1}}{p\xi_{p-1}} & \frac{d_{p-1}}{p\xi_p} \end{pmatrix} \\ &= \boldsymbol{DQ\Xi}, \end{aligned}$$

where $\boldsymbol{\Xi} = diag(\xi_1, \cdots, \xi_p)$ and

$$\boldsymbol{Q} = \begin{pmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{p-1} & \frac{1}{p} & \frac{1}{p} \\ -\frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{p-1} & \frac{1}{p} & \frac{1}{p} \\ 0 & -\frac{2}{3} & \frac{1}{4} & \cdots & \frac{1}{p-1} & \frac{1}{p} & \frac{1}{p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{p-2}{p-1} & \frac{1}{p} & \frac{1}{p} \\ 0 & 0 & 0 & \cdots & 0 & -\frac{p-1}{p} & \frac{1}{p} \end{pmatrix}.$$

Note that the Fisher information matrix for $(d_1, \cdots, d_p)$ is $(\boldsymbol{D}^2)^{-1}$. The Fisher information matrix for $(\xi \cdots, \xi_p)$ is then

$$\boldsymbol{H}'\boldsymbol{D}^{-2}\boldsymbol{H} \ = \ \boldsymbol{\Xi}'\boldsymbol{Q}'\boldsymbol{D}\boldsymbol{D}^{-1}\boldsymbol{D}^{-1}\boldsymbol{DQ\Xi} \ = \ \boldsymbol{\Xi}'\boldsymbol{Q}'\boldsymbol{Q\Xi}.$$

It is easy to verify that

$$\boldsymbol{Q}'\boldsymbol{Q} = diag\Big(\frac{1}{2}, \frac{2}{3}, \cdots, \frac{p-1}{p}, \frac{1}{p}\Big).$$

We have that

$$\boldsymbol{H}'\boldsymbol{D}^{-2}\boldsymbol{H} \quad = \quad diag\Big(\frac{1}{2\xi_1^2}, \frac{2}{3\xi_2^2}, \cdots, \frac{p-1}{p\,\xi_{p-1}^2}, \frac{1}{p\,\xi_p^2}\Big).$$

This proves part (a). Parts (b) and (c) are immediate.

Proof of Fact 4. We have

$$\begin{aligned} M &\equiv \int L(\boldsymbol{\mu}, \boldsymbol{\Psi})\pi_{\boldsymbol{a}}(\boldsymbol{\mu}, \boldsymbol{\Psi}) d\boldsymbol{\mu} d\boldsymbol{\Psi} \\ &= \int \frac{\prod_{i=1}^p (\psi_{ii}^2)^{(n-a_i-1)/2}}{(2\pi)^{(n-1)p/2} n^{p/2}} \exp\Big(-\frac{1}{2} tr(\boldsymbol{\Psi}\boldsymbol{S}\boldsymbol{\Psi}')\Big) d\boldsymbol{\Psi}. \end{aligned} \tag{78}$$

Note that $w_i = s_{ii} - \boldsymbol{s}'_{i,i-1}\boldsymbol{S}_{i-1}^{-1}\boldsymbol{s}_{i,i-1} > 0$ for $i = 2, \cdots, p$. Also let $\boldsymbol{g}_i = -\psi_{ii}\boldsymbol{S}_{i-1}^{-1}\boldsymbol{s}_{i,i-1}$. We then have a recursive formula,

$$\begin{aligned} tr(\boldsymbol{\Psi}_p \boldsymbol{S}_p \boldsymbol{\Psi}'_p) &= tr(\boldsymbol{\Psi}_{p-1}\boldsymbol{S}_{p-1}\boldsymbol{\Psi}'_{p-1}) + (\boldsymbol{\psi}_{p,p-1} - \boldsymbol{g}_p)'\boldsymbol{S}_{p-1}(\boldsymbol{\psi}_{p,p-1} - \boldsymbol{g}_p) \\ &= \sum_{i=1}^p \psi_{ii}^2 w_i + \sum_{i=2}^p (\boldsymbol{\psi}_{i,i-1} - \boldsymbol{g}_i)'\boldsymbol{S}_{i-1}(\boldsymbol{\psi}_{i,i-1} - \boldsymbol{g}_i). \end{aligned}$$

Then

$$M = \int \frac{\prod_{i=1}^p (\psi_{ii}^2)^{(n-a_i-1)/2}}{(2\pi)^{(n-1)p/2-(p-1)p/2} n^{p/2} \prod_{j=1}^{p-1} |\boldsymbol{S}_j|^{1/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^p \psi_{ii}^2 w_i\Big) \prod_{i=1}^p d\psi_{ii}. \tag{79}$$

Let $\delta_i = \psi_{ii}^2$. The right hand side of (79) is equal to

$$\begin{aligned} &\frac{\prod_{j=1}^{p-1} |\boldsymbol{S}_j|^{-1/2}}{2^p (2\pi)^{(n-p)p/2} n^{p/2}} \prod_{i=1}^p \int_0^\infty \delta_i^{(n-a_i)/2-1} \exp\Big(-\frac{w_i}{2}\delta_i\Big) d\delta_i \\ = \quad &\frac{\prod_{i=1}^p \Gamma(\frac{1}{2}(n-a_i)) 2^{(n-a_i)/2}}{2^p (2\pi)^{(n-p)p/2} n^{p/2}} \prod_{j=1}^{p-1} |\boldsymbol{S}_j|^{-1/2} \frac{1}{s_{11}^{(n-a_1)/2}} \prod_{i=2}^p \Big(\frac{|\boldsymbol{S}_{i-1}|}{|\boldsymbol{S}_i|}\Big)^{(n-a_i)/2}. \end{aligned}$$

The fact holds.

ACKNOWLEDGMENTS

The authors would like to thank Susie Bayarri and Jose Bernardo for helpful comments and discussions throughout the period when we were working on the project.

REFERENCES

Bayarri, M. J. (1981). Inferencia bayesiana sobre el coeficiente de correlación de una población normal bivariante', *Trabajos de Estadistica e Investigacion Operativa* **32**, 18–31.

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors (with discussion), *in* 'Bayesian Statistics 4', Oxford Univ. Press: London, 35–60.

Berger, J. O. and Bernardo, J. M. (1992). Reference priors in a variance components problem, *in* 'Bayesian analysis in statistics and econometrics', 177–194.

Berger, J. O., Strawderman, W. and Tang, D. (2005). Posterior propriety and admissibility of hyperpriors in normal hierarchical models, *Ann. Statist.* **33**, 606–646.

Berger, J. O. and Sun, D. (2006). Objective priors for a bivariate normal model. Submitted.

Brillinger, D. R. (1962). Examples bearing on the definition of fiducial probability with a bibliography', *Ann. Math. Statist.* **33**, 1349–1355.

Brown, P. J. (2001). The generalized inverted wishart distribution, *in* 'Encylopedia of Environmetrics'.

Chang, T. and Eaves, D. (1990). Reference priors for the orbit in a group model, *Ann. Statist.* **18**, 1595–1614.

Consonni, G., Gutiérrez-Peña, E. and Veronese, P. (2004). Reference priors for exponential families with simple quadratic variance function. *J. Multivariate Analysis* **88**, 335-364.

Daniels, M. and Kass, R. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models, *J. Amer. Statist. Assoc.* **94**, 1254–1263.

Daniels, M. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data, *Biometrika*, **89**, 553–566.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion), *Journal of the Royal Statistical Society, Series B* **35**, 189–233.

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

Dey, D. (1988). Simultaneous estimation of eigenvalues. *Ann. Inst. Statist. Math.* **40**, 137-147.

Eaton, M. L. (1989). *Group invariance applications in statistics*, Institute of Mathematical Statistics.

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

Eaton, M.L. and Sudderth, W. (2002). Group invariant inference and right haar measure, *J. Statist. Planning and Inference* **103**, 87–99.

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters, *J. Roy. Statist. Soc. B* **25**, 368–376.

Haff, L. (1977). Minimax estimators for a multivariate precision matrix, *J. Multivariate Analasys* **7**, 374–385.

James, W. and Stein, C. (1961). Estimation with quadratic loss, *in* 'Proc Fourth Berkely Symp. Math. Statist. Probability, 1', University of California Press, pp. 361–380.

Jeffreys, H. (1961). *Theory of Probability*, Oxford University Press, London.

Leonard, T. and Hsu, J.S.J. (1992). Bayesian inference for a covariance matrix. *Ann. Statist.* **20**, 1669-1696.

Liechty, J., Liechty, M. and Müller, P. (2004). Bayesian correlation estimation, *Biometrika* **91**, 1–14.

Lindley (1965). The use of prior probability distributions in statistical inference and decisions. 453–468.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation, *Biometrika* **86**, 677–690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87**, 425-435.

Roverato, A. and Consonni, G. (2004). Compatible prior distributions for dag models, *Journal of the Royal Statistical Society, Series B, Methodological* **66**, 47–61.

Stein, C. (1956). Some problems in multivariate analysis. part i., Technical Report 6, Department of Statistics, Stanford University.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior, *Ann. Statist.* **22**, 1195–1211.

DISCUSSION

BRUNERO LISEO (*Università di Roma "La Sapienza", Italy*)

General Comments

I have enjoyed reading this authoritative paper by Jim Berger and Dongchu Sun. The paper is really thought provoking, rich of new ideas, new proposals, and useful technical results about the most used and popular statistical model.

I assume that the role of the discussion leader in a conference is different from that of a journal referee, and his/her main goal should be to single out the particular features of the paper that deserve attention, and to provide a personal perspective and opinion on the subject.

This paper deserves discussion in at least three important aspects:

(i) technical results;

(ii) philosophical issues;

(iii) guidance on the choice among available "objective" prior distributions.

Among the technical results I would like to remark first the fact that almost all the proposed "objective posterior" are given in a constructive form, and it is usually nearly immediate to obtain a sample from the marginal posterior of the parameter of interest. This simple fact facilitates the reader's task, since it is straightforward to perform other numerical comparisons and explore different aspects of the problem.

There are also many philosophical issues raised by the Authors. The paper, as many others written by Berger and Sun, stands on the interface between frequentist and Bayesian statistics. From this perspective, the main contributions of the paper are, in my opinion, the following:

- to provide a classical interpretation of some objective Bayes procedures;
- to provide an objective Bayesian interpretation of some classical procedures;
- to derive optimal and nearly optimal classical/fiducial/objective Bayesian procedures for many inferential problems related to the bivariate and multivariate normal distribution.

Before discussing these issues, I would like to linger over a tentative definition of what is the main goal of an objective Bayesian approach, which seems, in some respects, lacking. I have asked many Bayesian statisticians the question: "Could you please provide, in a line, a definition of the Objective Bayesian approach?". I report the most common answers, together with some of the contributions illustrated in Berger (2006) and O'Hagan (2006):

(i) A formal Bayesian analysis using some *conventional* prior information, which is largely "accepted" by researchers.

(ii) The easiest way to obtain good frequentist procedures.

(iii) A Bayesian analysis where the prior is obtained from the sampling distribution; it is the only feasible approach when there is no chance of getting genuine prior information for our model and we do not want to abandon that model.

(iv) The optimal Bayesian strategy under some specific frequentist criterion (frequentist matching philosophy)

(v) A cancerous oxymoron.

(vi) A convention we should adopt in scenarios in which a subjective analysis is not tenable.

(vii) A collection of convenient and useful techniques for approximating a genuine Bayesian analysis.

While preferring option (iii), I must admit that option (ii) captures important aspects of the problem, in that it provides a way to validate procedures and to facilitate communication among statisticians. On the other hand, option (iv) might be a dangerous goal to pursue, as the Authors, perhaps indirectly, illustrate in the paper.

Indeed the Authors present many examples where

- If the prior is chosen to be "exact" in terms of frequentist coverage, then the resulting posterior will suffer from some pathologies like, for example, the marginalization paradox.
- To be "optimal" one has to pay the price of introducing unusual (albeit, it is the right Haar prior!) default priors. For instance, in the bivariate normal example, when ρ is the parameter of interest, the "exact" frequentist matching prior is

$$\pi_H(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \propto \frac{1}{\sigma_1^2(1-\rho^2)}, \tag{80}$$

  in which, in any possible "objective sense", the a priori discrimination between the two standard deviations are, at least, disturbing for the practitioner.

I personally do not consider the marginalization paradox so troublesome. After all, it is a potential consequence of using improper priors (Regazzini, 1983). Here the main issue seems to be: once that the myth of optimality has been abandoned, a complete agreement between frequentist, Bayesian and fiducial approaches cannot be achieved. It is time, I believe, to decide whether this is a problem or not. My personal view is that there exist, nowadays, many examples in which frequentist reasonable behaviour of objective Bayesian procedures is simply impossible and "objective" Bayesians should not consider that as the main "objective" of a Bayesian analysis. Many examples of irreconcilability between classical and Bayesian analysis arise when the parameter space is constrained, but even in the regular bivariate normal problem there are functions of the parameters for which frequentist matching

$(\rho, \sigma_1, \sigma_2)$	π_H	π_{21}	$\pi_{R\rho}$
$(0, 1, 1)$	**0.154**	0.229	0.188
$(0.5, 10, 1)$	**0.143**	0.176	0.180
$(0.5, 1, 10)$	**0.132**	0.170	0.178
$(0.9, 1, 10)$	0.045	**0.034**	0.056
$(-0.9, 10, 1)$	0.046	0.035	**0.028**
$(-0.5, 10, 1)$	**0.142**	0.166	0.148
$(-0.1, 10, 1)$	**0.155**	0.212	0.176
$(0.1, 10, 10)$	0.160	**0.150**	0.182
$(0.8, 10, 10)$	0.084	**0.065**	0.186

Table 3: Mean square error for different possible priors. Here we assume that the estimation of ρ is the goal of inference and the three priors are compared in terms of the mean squared errors of the resulting estimators. For each simulation, the bold character indicates the best.The right Haar prior is the winner except when $\mid \rho \mid$ is close to one, and/or σ_1 and σ_2 are both large.

behaviour cannot be achieved by Bayesian solutions: perhaps the Fieller's problem is the most well known example (Gleser and Hwang, 1987). Another example is the problem of estimating the coefficient of variation of a scalar Normal distribution (Berger et al., 1999).

In the Fieller's problem one has to estimate the ratio of two Normal means, that is $\theta = \mu_1/\mu_2$; Gleser and Hwang (1987) showed that any confidence procedure of level $1 - \alpha < 1$ produces infinite sets with positive sampling probability; on the other hand, any "reasonable" objective prior like Jeffreys' or reference prior always provides finite HPD sets. Also, the frequentist coverage of one side credible sets derived from such objective priors is always far from the nominal level. In cases like this one, the mathematical structure of the model (in which a sort of *local* unidentifiability occurs) simply prevents from deriving an exact and non trivial confidence procedure. What should an objective Bayesian decide to do here? what his/her goal should be in these cases?

There is another compelling reason to be against option (iv) as a possible *manifesto* of objective Bayesian inference: the choice of the optimality criterion is crucial and it sounds like another way to re-introduce subjectivity in the problem. To explore this issue I have performed a small simulation in the bivariate case, when ρ is the parameter of interest. I have compared the three most "promising" priors in terms of frequentist quadratic loss. Table 1 shows that the results are sensibly different from those obtained from a coverage matching perspective. Notice that in Table 1 π_H is the right Haar prior, the one giving exact frequentist coverage, while π_{IJ} is the Independence-Jeffreys' prior and $\pi_{R\rho}$ is the reference prior when ρ is the parameter of interest.

Improper priors

Another result provided by the Authors, that I have found particularly stimulating, is the problem of sub-optimality of the mixed Haar prior. Since, in the multivariate

case, the right Haar prior is not unique and a default choice cannot be made, Berger and Sun propose to consider a mixture of the possible priors. Then, in the bivariate case, this procedure produces, for the covariance matrix, an estimator with a frequentist risk higher than those one would have obtained by using a single right Haar prior. I believe that this surprising result can be explained by the fact that a convex combination of two improper priors arbitrarily depends on the weights we tacitly attach to the two priors. In other words, as the Authors notice, it seems an example where one extends the use of elementary probability rules to a scenario (the use of improper priors) where they could easily fail. To this end, Heath and Sudderth (1978) state that

> *... many "obvious" mathematical operations become just formal devices when used in the presence of improper priors...*

This is exactly what happens when the marginalization paradox shows up. There was an interesting debate in the 80's about the connections between improper priors, non conglomerability and the marginalization paradox: see for example, Akaike (1980), Heath and Sudderth (1978), Jaynes (1980), Sudderth (1980) and Regazzini (1983). In particular, Regazzini (1987) clearly showed that the marginalization paradox is nothing more than a consequence of the fact that the (improper) prior distribution may be non conglomerable. While this phenomenon never happens with proper priors, its occurrence is theoretically conceivable when σ-additivity is not taken for granted, as is the case with the use of improper priors. Interestingly, in one of the examples discussed by Regazzini (1983) namely the estimation of the ratio of two exponential means (Stone and Dawid, 1972), the marginalization paradox is produced by an improper prior distribution that is obtained as a solution of a Bayesian/Frequentist agreement problem. More precisely, the question was: let $(x_1, \cdots, x_n)$ be an i.i.d. sample from an Exponential distribution with parameter $\phi\,\theta$ and let $(y_1, \cdots, y_n)$ be an i.i.d. sample from an Exponential distribution with parameter ϕ and suppose we are interested on θ: does it exist a prior distribution such that the posterior mean of θ is exactly equal to the "natural" classical point estimate $\bar{y}/\bar{x}$? In that example, it is also possible to show that the only (improper) prior which satisfies the above desideratum, namely $\pi(\theta, \phi) \propto 1/\theta$, produces the marginalization paradox.

Today we all "live in the sin" of improper priors, but it is still important to discriminate among them. One can achieve this goal in different ways: for example it is possible to check if a given improper prior is coherent (that is, if it has a finitely additive interpretation). However, especially in the applied world, this road might be, admittedly, hard to follow.

Alternatively, one could try to figure out where do these distributions put the prior mass. Sudderth (1980) interestingly stresses the fact that, for example, the uniform improper prior on the real line and the finitely additive limit of uniform priors on a sequence of increasing compact sets show a sensibly different behavior: indeed,

> *... the odds on compact sets versus the whole space are* $0:1$ *for the finitely additive prior and finite: infinite for the improper prior.*

My personal view is that a reasonable way to discriminate among priors might be to measure, in terms of some Kullback-Leibler's type index, the discrepancy of the induced posteriors with respect to some *benchmark* posterior. The reference prior

algorithm has its own benchmark in a sort of asymptotic maximization of missing information and it works remarkably well in practice; I believe that other possible benchmarks might be envisaged, perhaps in terms of prediction.

BERTRAND CLARKE (*University of Brithsh Columbia, Canada*)

Sun and Berger (2006) examine a series of priors in various objective senses, but the focus is always on the priors and how well they perform inferentially. While these questions are reasonable, in fact it is the process of obtaining the prior, not the prior so obtained, that makes for objectivity and compels inquiry.

Indeed, the term objective prior is a misnomer. The hope is merely that they do not assume a lot of information that might bias the inferences in misleading directions. However, to be more precise, consider the following which tries to get at the process of obtaining the prior.

Definition: A prior is objective if and only if the information it represents is of specified, known provenance.

This means the prior is objective if its information content can be transmitted to a recipient who would then derive the same prior. The term information is somewhat vague; it is meant to be the propositions that characterize the origin of the prior, in effect systematizing its obtention and specifying its meaning. As a generality, the proliferation of objective priors in the sense of this definition seems to stem from the various ways to express the concept of absence of information. Because absence of information can be formulated in so many ways, choosing one formulation is so informative as to uniquely specify a prior fairly often. Thus it is not the priors themselves that should be compared so much as the assumptions in their formulation.

The information in the prior may stem from a hunch, an expression of conservatism, or of optimism. Or, better, from modeling the physical context of the problem. What is distinctive about this definition of objective priors is that their information content is unambiguously identified. It is therefore ideationally complex enough to admit agreement or disagreement on the basis of rational argument, modeling, or extra-experimental empirical verification while remaining a separate source of information from the likelihood or data.

By this definition, one could specify Jeffreys prior in expression (2) in several equivalent ways. It can be regarded as 1) the asymptotically least favorable prior in an entropy sense, 2) a transformation invariant prior provided an extra condition is invoked so it is uniquely specified, 3) the frequentist matching prior (for p=2). Likewise, the information content of other priors in Sun and Berger (2006) can be specified since they are of known provenance and hence objective.

Jeffreys prior is justly regarded as noninformative in the sense that it changes the most on average upon receipt of the data under a relative entropy criterion. This concept of noninformativitity is appropriate if one is willing to model the data being collected as having been transmitted from a source, and to model the parameter value as a message to be decoded. Jeffreys prior is the logical consequence of this modeling strategy. So, if an experimenter is unhappy with the performance of the Jeffreys prior, the experimenter must model the experiment differently. It is not logically consistent to adopt the data transmission model that leads to Jeffreys prior but then reject the Jeffreys prior.

However, if some information theoretic model is thought appropriate, a natural alternative to the Jeffreys' prior could be Rissanen's (1983) prior which has the plus of being proper on the real line. Rissanen's prior also has an information theoretic

interpretation but it's in a signal-to-noise ratio sense rather than a transmission sense. Indeed, there are many ways to motivate physically the choice of prior; several similar information theoretic criteria are used in Clarke and Yuan (2004) resulting in ratios of variances. By contrast, if it is the tail behavior of the prior that matters more than the information interpretation, then the relative entropy might not be the right distance, obviating information theory. Other distances such as the Chi-square lead to other powers of the Fisher information, see Clarke and Sun (1997).

Essentially, the argument here is to turn prior selection into an aspect of modeling. Thus, the issue is not whether a given prior gives the best performance, for instance in the decision theoretic sense of Table 2, but whether the information implicit in the prior is appropriate for the problem. In particular, although one might be mathematically surprised that the risk of $\hat{\Sigma}_S$ is higher than the risk of right Haar priors, this is not the point. The point is whether the assumptions undergirding one or another of these priors is justified. For instance, if the model for the experiment includes the belief that minimal risk will be achieved, one would not be led to the mixture of two priors. On the other hand, if the risk is not part of the physical model then one is not precluded from using the mixture of priors if it is justified.

In the same spirit, the issues of the marginalization paradox andfrequentist matching can be seen as generally irrelevent. The marginalization paradox does not exist for proper priors and it is relatively straightforward to derive priors that satisfy a noninformativity principle and are proper. Rissanen's prior is only one example. Indeed, the information for prior construction may include 'use the first k data points to upgrade an improper prior chosen in such-and-such a way to propriety and then proceed with $n - k$ data points for inference'. Similarly, matching priors merely represent the desire to replicate frequentist analysis. If the two match closely enough, the models may be indistinguiable, otherwise they are different and can't both be right. Moreover, the fact that matching frequentist analysis and avoiding the marginalization paradox often conflict is just a fact: The models from which the priors derive cannot satisfy both criteria, perhaps for purely mathematical reasons, and little more can be made of it without examining classes of models.

A reasonable implication from this definition is that subjective priors have no place in direct inference because their provenance cannot be evaluated. The main role remaining to subjective priors may be some aspects of robustness analysis. It remains perfectly reasonable to say 'I have consulted my predilections and impressions and think they are well represented if I draw this shape of density, announce that analytic form and then see what the inferential consequences are'. However, since the information in such a prior is not of known provenance, it cannot be subject to inquiry or validation and so without further justification does not provide a firm basis for deriving inferences. Of course, being able to argue that n is large enough that the prior information is irrelevant would be adequate, and in some cases, an extensive robustness analysis around a subjective prior (especially if the robustness included 'objective' priors, like powers of the Fisher information) might lend the subjective approach credence.

GUIDO CONSONNI and PIERO VERONESE
(*University of Pavia, Italy and Bocconi University, Italy*)

The paper by Sun and Berger presents an impressive collection of results on a wide range of objective priors for the multivariate normal model. Some of them are closely related to previous results of ours, which were derived using a unified

approach based on the theory for conditionally reducible exponential families. In the sequel we briefly present and discuss some of the main aspects.

In Consonni and Veronese (2001) we discuss Natural Exponential Families (NEFs) having a particular recursive structure, that we called conditional reducibility, and the allied parameterization ϕ of the sampling family. Each component of ϕ is the canonical parameter of the corresponding conditional NEF. One useful feature of ϕ is that it allows a direct construction of "enriched" conjugate families. Furthermore the parameterization ϕ is especially suitable for the construction of reference priors, see Consonni, Gutiérrez-Peña, and Veronese (2004) within the framework of NEFs having a simple quadratic variance function. Interestingly, we show that reference priors for different groupings of ϕ belong to the corresponding enriched conjugate family.

In Consonni and Veronese (2003), henceforth CV03, the notion of conditional reducibility is applied to NEFs having a homogeneous quadratic variance function, which correspond to the Wishart family on symmetric cones, i.e the set of symmetric and positive definite matrices in the real case. In this setting we construct the enriched conjugate family, which is shown to coincide with the Generalized Inverse Wishart (GIW) prior on the expectation Σ of the Wishart, and provide structural distributional properties including expressions for the expectation of Σ and Σ^{-1}. We also obtain a grouped-reference prior for the ϕ-parameterization, as well as for $\Sigma = (\sigma_{ij},\, i,j = 1,\ldots,p)$ according to the grouping $\{\sigma_{11}, (\sigma_{21}, \sigma_{22}),\, \ldots,\, (\sigma_{p1},\ldots,\sigma_{pp})\}$, showing that the reference prior belongs to the enriched conjugate family in this case, too. This in turn allows to prove directly that the reference posterior is always proper and to compute exact expressions for the posterior expectation of Σ and Σ^{-1}.

There is a close connection between our ϕ-parameterization and the Cholesky decomposition of $\Sigma^{-1} = \Psi'\Psi$ in terms of the triangular matrix Ψ of formula (8): in particular ϕ and Ψ are related through a block lower-triangular transformation, (see CV03, where further connections between ϕ and other parameterizations are explored). As a consequence the reference prior on Ψ can be obtained from that of ϕ through a change-of-variable.

Our results on the Wishart family are directly applicable to a random sample of size n from a multivariate normal $N_p(0,\Sigma)$, since the Wishart family corresponds to the distribution of the sample cross-products divided by n (a sufficient statistic).

In the paper by Sun and Berger the multivariate normal model $N_p(\mu,\Sigma)$ is considered, and several objective priors for (μ,Ψ), and other parameterizations, are discussed. Special emphasis is devoted to the class of priors π_a, see (13), that includes a collection of widely used distributions such as the reference prior π_{R1}, see (12), corresponding to the grouping $\{\mu, \psi_{11}, (\psi_{21},\psi_{22}), \ldots, (\psi_{p1},\ldots,\psi_{pp})\}$. We remark that the right-hand-side of (12), $\prod_{i=1}^{p} \psi_{ii}^{-1}$, coincides with the reference on Ψ derived from our reference prior on ϕ. To see why this occurs, notice that the Fisher information matrix for $\{\mu, \psi_{11}, (\psi_{21},\psi_{22}),\, \ldots,\, (\psi_{p1},\ldots,\psi_{pp})\}$ is block-diagonal, with the first block constant w.r.t. μ, while the remaining blocks do not involve μ and are equal to those that hold under the $N_p(0,\Sigma)$ case. Incidentally, while our reference on Ψ is actually of the GIW type, this is clearly not the case for (12), which is a prior on (μ,Ψ), so that the support is not the space of real and symmetric positive definite matrices.

As the Authors point out at the beginning of Section 3, since the conditional posterior for μ given Ψ is normal, interest centers on the marginal posterior of Ψ, which depends on the data only through the sample variance $S/(n-1)$, whose

distribution is Wishart, so that our results are still relevant. Specifically the marginal on Ψ under $\pi_a(\mu, \Psi)$, as well as the posterior, belongs to the enriched conjugate/GIW family, whence items (a) and (b) of Fact 5, and Fact 8, are readily available from Corollary 1, Proposition 1 and Proposition 2 of CV03.

Our last point concerns the marginalization paradox in the multivariate setting. Partition Σ into four blocks Σ_{ij}, $i, j = 1, 2$; then the reference prior for ϕ, which corresponds to the prior π_{R1}, does not incur the marginalization paradox for the marginal variance Σ_{11}, the conditional variance $\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, and the pair $\Sigma_{11}, \Sigma_{2|1}$, see CV03.

A. PHILIP DAWID (*University College London, UK*)

SUMMARY

By relating the problem to one of fiducial inference, I present a general argument as to when we can expect a marginal posterior distribution based on a formal right-Haar prior to be frequency calibrated.

Keywords and Phrases: FIDUCIAL DISTRIBUTION; GROUP INVARIANCE; STRUCTURAL MODEL

The following problem formulation and analysis closely follow Dawid, Stone and Zidek (1973) and Dawid and Stone (1982), which should be consulted for full details.

Stuctural model

Consider a simple "structural model" (Fraser, 1968):

$$X = \Theta E \tag{81}$$

where the *observable* X, the *parameter* Θ, and the *error variable* E all take values in a group G, and the right-hand side of 81 involves group multiplication. Morevover, E has known distribution P, independently of the value θ of Θ. Letting P_θ denote the implied distribution of $X = \theta E$, we obtain an induced *statistical model* $\mathcal{P} = \{P_\theta : \theta \in G\}$ for X. (Note however that distinct structural models can induce the same statistical model.)

Now assign to Θ the formal right-Haar prior over G. It then follows (Hora and Buehler, 1966) that the resulting formal posterior Π_x for Θ, based on model $\mathcal{P}$ and data $X = x$, will be identical with the *structural (fiducial)* distribution for Θ based on 81, which is contructively represented by the *fiducial model*:

$$\Theta = xE^{-1} \tag{82}$$

with the distribution of E still taken to be P.

Non-invariant estimation

It easily follows that a suitably invariant level-γ posterior credible set will also be a level-γ confidence interval. Less obviously, such a confidence interpretation of a posterior interval also holds for inference about suitable one-dimensional functions of Θ, even though no invariance properties are retained.

Thus let H be a subgroup of G, and $w : G \to \mathbb{R}$ a maximal invariant function under left-multiplication by H. Suppose further that w is *monotone*:

$$w(\theta_1) \leq w(\theta_2) \Leftrightarrow w(\theta_1 e) \leq w(\theta_2 e) \tag{83}$$

for all $\theta_1, \theta_2, e \in G$. Define $Z := w(X)$ and $\Lambda = w(\Theta)$. It now follows (Dawid and Stone, 1982, §4.2) that, under mild continuity conditions:

- Z is a function of Λ and E: say $Z = f(\Lambda, E)$
- the marginal sampling distribution of Z depends only on Λ
- the marginal posterior distribution of Λ (which is the same as its marginal fiducial distribution) depends only on Z: say Π_z for $Z = z$
- Π_z is represented constructively by $\Lambda = g(z, E)$, with E having distribution P; here $g(z, e)$ represents the solution, for λ, of $z = f(\lambda, e)$.
- Π_z agrees with Fisher's fiducial distribution, obtained by differentiating with respect to λ the cumulative distribution function P_λ for Z given $\Lambda = \lambda$
- if, for each z, λ_z is a posterior level-γ upper credible limit for $\Lambda = w(\Theta)$, *i.e.*, $\Pi_z(\Lambda \leq \lambda_z) = \gamma$, then λ_Z will also be an exact level-γ upper confidence limit for Λ, *i.e.*, $P_\theta(w(\theta) \leq \lambda_Z) = \gamma$, all θ.

The application to the case of the correlation coefficient Λ in a bivariate normal model (albeit with zero means) was treated explicitly in §3 of Dawid and Stone (1982), with G the triangular group and H its diagonal subgroup. The results of Berger and Sun (2006) follow directly.

As noted in §2.4 of Dawid, Stone and Zidek (1973), use of the right-Haar prior on G will typically entail a marginalization paradox for Λ. For discussion of such logical issues, see Dawid, Stone and Zidek (2006).

Ancillaries

The above theory generalizes readily to problems where E and X live in a general sample space $\mathcal{X}$, and $\Theta \in G$, an exact transformation group on $\mathcal{X}$: the relevant theory is in Dawid, Stone and Zidek (1973). Let $a(\cdot)$ be a maximal invariant function on $\mathcal{X}$ under the action of G. Then defining $A = a(X)$, it is also the case that $A = a(E)$, and A is thus ancillary. In particular, on observing $X = x$ we learn that $E \in \mathcal{E}_x := \{e : a(e) = a(x)\}$. In the fiducial model 82 we must now restrict E to $\mathcal{E}_x$, defining xe^{-1} as the then unique solution for θ of $x = \theta e$, and assigning to E the distribution over $\mathcal{E}_x$ obtained by conditioning its initial distribution P on $a(E) = a(x)$. We must also evaluate sampling performance conditional on the ancillary statistic $a(X)$. Then all the above results go through (indeed, we obtain the coverage property conditional on A, which is stronger than unconditional coverage).

JAYANTA GHOSH (*Purdue University, USA*)

I have a couple of general as well as specific comments on objective priors inspired by the paper of Sun and Berger and its discussion by Liseo, both of which are very interesting.

I focus on two major problems. Objective priors for point or interval estimates, generated by standard algorithms, are almost always improper and unique. They are improper because they are like a uniform distribution on a non-compact parameter space. They are not unique because different algorithms lead to different priors. Typically they may not have the properties one usually wants, namely, coherence, absence of marginalization paradox, some form of probability matching,

and some intuitive motivation. Sun and Berger (2006) show no prior for the bivariate normal can avoid marginalization paradox and also be probability matching for the correlation coefficient. Brunero points out that probability matching or any other Frequentist matching need not be desirable and the marginalization paradox is a consequence of the prior being improper. Incidentally, for one or two dimensions, probability matching priors are like reference priors, for higher dimensions probability matching priors are not well-understood.

Brunero asks why one should try to do Frequentist matching at all to justify an objective prior. He suggests a better way of comparing two objective priors might be to examine where they put most of their mass.

It is worth pointing out that coherence could help us decide which one to choose. A stronger notion, namely, admissibility, has been used by Jim in a number of cases. Probability matching seems to provide a very weak and asymptotic form of coherence. A detailed study of coherence of improper priors is contained in Kadane et al (1999).

Another way of comparing two objective priors, which addresses impropriety directly, could explore how well the posteriors for the improper priors are approximated by the posteriors for the approximating truncated priors on a sequence of compact sets. This is usually not checked but is related to one of the basic requirements for construction of a reference prior, vide Berger and Bernardo (1992). As far as non-uniqueness is concerned, a mitigating factor is that a moderate amount of data lead to very similar posteriors for different objective priors, even though the data are not large enough to wash away most priors and make the posteriors approximately normal. I wonder if this holds for some or all the objective priors for the bivariate normal.

VICTOR RICHMOND R. JOSE, KENNETH C. LICHTENDAHL, JR., ROBERT F. NAU, AND ROBERT L. WINKLER (*Duke University, USA*)

This paper is a continuation of the work by Berger and Sun to increase our understanding of diffuse priors. Bayesian analyses with diffuse priors can be very useful, although we find the term "objective" inappropriate and misleading, especially when divergent rules for generating "objective" priors can yield different results for the same situation. How can a prior be "objective" when we are given a list of different "objective" priors, found using different criteria, from which to choose? The term diffuse, which has traditionally been used, is fine, as is weakly informative (O'Hagan, 2006). In the spirit of Savage's (1962) "precise measurement," diffuse priors can be chosen for convenience as long as they satisfy the goal of providing good approximations to the results that would be obtained with more carefully assessed priors. Issues surrounding so-called "objective Bayesian analysis" have been discussed at length in papers by Berger (2006) and Goldstein (2006) and the accompanying entertaining commentary. There is neither need nor space to go over all of that ground here, but our views are in line with the comments of Fienberg (2006), Kadane (2006), Lad (2006), and O'Hagan (2006).

Over the years, much time and effort has been spent pointing out basic differences between frequentist and Bayesian methods and indicating why Bayesian methods are fundamentally more sound (they condition appropriately, they address the right questions, they provide fully probabilistic statements about both parameters and observables, etc.). Jim Berger has participated actively in this endeavor. Given this effort, why would we want to unify the Bayesian and frequentist approaches? Why should we be interested in "the prior that most frequently yields

exact frequentist inference," which just leads us to the same place as using frequentist methods, many of which have been found lacking from a Bayesian viewpoint? Why should we care about the frequentist performance of Bayesian methods? Is it not preferable to focus more on how well Bayesian analyses (including those using diffuse priors) perform in a *Bayesian sense*? That means, for example, using scoring rules to evaluate predictive probabilities.

One aim is apparently to market Bayesian methods to non-Bayesians. As Lad (2006) notes, "The marketing department has taken over from the production department." Our sense is that, since Bayesian methods are inherently more sound, sensible, and satisfying than frequentist methods, the use of Bayesian methods will continue to increase. Rather than trying to stimulate this by marketing Bayesian procedures as "objective" (which neither they nor frequentist procedures can be) and meeting frequentist criteria (which is not what they are intended to do), let's invest more effort toward continuing to apply Bayesian methods to important problems and toward making Bayesian methods more accessible. The development of more user-friendly Bayesian software for modeling both prior and likelihood and for handling Bayesian computations would be a big step in the right direction.

Sun and Berger's concern about improper priors is well-founded, since improper diffuse priors are commonly encountered. If we think of the Bayesian framework in terms of a big joint distribution of parameters and observables, improper priors leave us without a proper joint distribution and leave us unable to take advantage of the full Bayesian menu of options. First, it is often argued that if the posteriors following improper diffuse priors are themselves proper, all is well (e.g., Berger, 2006, p. 393). However, posteriors following improper diffuse priors are not always proper. For example, in a normal model with high dimensional data and small sample sizes (large p, small n), the posteriors that follow from many of the priors considered by Sun and Berger are not proper. Second, even though improper priors may (eventually) yield proper posteriors, they do not, in general, provide proper predictive distributions. Although decision-making problems are typically expressed in terms of parameters rather than observables in Bayesian statistics, which means that posterior distributions are of interest, the primary focus in decision analysis is often on observables and predictive distributions. For important preposterior decisions such as the design of optimal sampling plans and for value of information calculations, proper predictive distributions are needed. As a result, many so-called "objective" priors, including priors proposed by Sun and Berger, leave us unable to make formal preposterior decisions and force us to resort to ad hoc procedures instead.

As an illustration, consider optimal sampling in clinical drug trials. Should we sample at all, and how big should the initial sample be? With improper "objective" priors, we are unable to provide a formal analysis of such important decisions, which have serious life-and-death implications. As Bayesians, and more generally as scientists, we should actively promote the use of tools that are more conducive to good decisions.

F. J. O'REILLY (*Mexico*)

The authors ought to be congratulated for providing yet another piece of research where rather than looking for differences in statistical procedures, they try to identify closeness between what do theories provide. This paper is very much in line with the objective Bayesian point of view which as mentioned in Berger (2006), was present in science long before the subjective Bayesian approach made its formal arrival.

One wonders why there has been such an extreme position in stressing differences between Bayesian and classical results. Exact coincidences between reference posterior densities and fiducial densities exist, not too often as shown in Lindley (1958), but in many cases, practical differences are small. And in trying to understand these differences, a compromise between procedures, in this case, with exact coverage probabilities or procedures for which the marginalization paradox does not arise, must be faced. The authors mention this fact very clearly. For some, there is no compromise to be done; they have made their point and should be respected, but for those exploring this compromise, we believe they too, should be respected.

We would like to refer just to one aspect of the paper on the correlation coefficient ρ in the bivariate normal distribution which in our opinion stands out. On the one hand, right Haar priors are elicited following invariance considerations but unfortunately (inherited from the non-uniqueness of the factorization) there are two right Haar measures which appear in a nonsymmetrical fashion despite the symmetrical role that the standard deviations σ_1 and σ_2 "should" have. The authors explore mixing these two Haar measures in an effort, it seems, to get rid of the uncomfortable asymmetry and they do recognize that the effort does not produce a reasonable answer.

On the other hand, the symmetric reference prior $\pi(\rho)$ given in Bayarri (1982), is mentioned as not yielding exact coverage probabilities for the corresponding posterior, but certainly free from the marginalization paradox. Two questions arise at this point. The first one is how different is the coverage probability from the exact one? The second question has to do with the interesting relationship between the two Haar measures and the reference prior, which in this case is the geometric mean.

Do we have to mix with convex linear combinations $(\alpha, 1-\alpha)$? In the discussion, it was mentioned that with improper priors, the weights placed on a convex linear combination have little interpretation, except if one seeks symmetry ($\alpha = 0.5$), as it seems to be the case. A priori weights α and $1 - \alpha$ for the two priors mean a different set of weights for the posteriors when representing the posterior obtained from the mixed prior as a mixture of the posteriors. Why not simply work with the prior obtained by "mixing" both prior measures using their geometric mean? The result, in general, would provide a proper posterior if the two associated posteriors are proper (Schwarz inequality).

It would have been interesting to see some graphs of the various posteriors discussed for ρ in a few cases with small and medium values for the sample size n when the sampling correlation coefficient r is varied along its range. In some examples we have been doing in non-location scale families, even with very small sample sizes, the graphs of the fiducial and the reference posterior densities are almost indistinguishable. That is the case among others, of the truncated exponential in O'Reilly and Rueda (2006).

REPLY TO THE DISCUSSION

We are grateful to all discussants for their considerable insights; if we do not mention particular points in the discussions, it is because we agree with those points.

Dr. Liseo: We agree with essentially all of Dr. Liseo's illuminating comments, and will mention only two of them.

The survey and comments concerning the definition of the Objective Bayesian approach were quite fun, and we feel that there is some truth in all the definitions (except one), reinforcing the difficulty of making the definition precise.

Dr. Liseo reminds us that there are many situations where objective Bayesians simply cannot reach agreement with frequentists, thus suggesting that frequentist performance should be a secondary criterion for objective Bayesians. He also comments, however, that he does not consider the marginalization paradox to be particularly troublesome, and shows that there is no clear winner when looking at estimation of ρ (which, interestingly, is also the case when entropy loss is used). Hence we are left with the uneasy situation of not having a clear recommendation between the right-Haar and reference prior for ρ.

Dr. Clarke gives a strong argument for the appealing viewpoint that choice of objective priors is simply a choice of the communicable information that is used in its construction, giving examples of priors arising from information transmission arguments. One attractive aspect of this is that it also would apply to a variety of priors that are based on well-accepted understanding, for instance various hierarchical priors or even scientific priors: if all scientists agreed (based on common information they hold) that the uncertainty about μ was reflected by a $N(0,1)$ distribution, shouldn't that be called an objective prior?

There are two difficulties with implementation of this viewpoint, however. The first is simply a matter of communication; while calling the above $N(0,1)$ prior objective might well be logical, the name 'objective prior' has come to be understood as something quite different, and there is no use fighting history when it comes to names.

A more troubling difficulty is that this formal view of objective priors does not seem to be practically implementable. For instance, Dr. Clarke suggests that the Jeffreys prior is always suitable, assuming we choose to view data transmission under relative entropy to be the goal. We know, of course, that the Jeffreys prior can give nonsensical answers for multivariate parameters (e.g., inconsistency in the Neyman-Scott problem). Dr. Clarke's argument is that, if we feel the resulting answer to be nonsensical, then we cannot really have had the data transmission problem as the goal. While this is perhaps a logically sound position, it leaves us in the lurch when it comes to practice; when can we use the Jeffreys prior and when should we not? Until an approach to determination of objective priors provides answers to such questions, we strongly believe in the importance of actually evaluating the performance of the prior that results from applying the approach.

Drs. Consonni and Veronese give a nice discussion of the relationship of some of the priors considered in our paper with their very interesting results on reference priors for natural exponential families with a particular recursive structure, namely conditional reducibility; for these priors, some of the facts listed in our paper are indeed immediate from results of Consonni and Veronese. (Of course, the right Haar prior and those in (5) and (7) are not in this class.) It is also indeed interesting that this class of priors does not lead to a marginalization paradox.

Dr. Dawid's result is quite fascinating, because it shows a wide class of problems in which objective priors can be exact frequentist matching (and fiducial), even though the parameter of interest is not itself fully invariant in the problem. Also, the posterior distribution can be given constructively in these situations. Such results have long been known for suitably invariant parameters, but this is a significant step forward for situations in which full invariance is lacking.

In Berger and Sun (2006), several of the results concerning exact frequentist matching and constructive posteriors could have been obtained using this result of Dr. Dawid. However, some of the results in that paper (e.g. those concerning the

parameters $\eta_3 = -\rho/[\sigma_1\sqrt{1-\rho^2}]$ and $\xi_1 = \mu_1/\sigma_1$) required a more difficult analysis; these may suggest further generalizations of Dr. Dawid's result.

Dr. Ghosh: We certainly agree with the insightful comments of Dr. Ghosh. We have not studied the sample size needed for different objective priors to give essentially the same posterior; presumably this would be a modest sample size for the bivariate case, but the multivariate case is less clear.

Drs. Jose, Lichtendahl, Nau, and Winkler give a nice discussion of some of the concerns involving objective priors. We agree with much of what they say; for instance, they point out that there are numerous situations – including experimental design – where use of objective priors may not be sensible. Yet we also are strong believers that objective priors can be of great use in much of statistics. The discussion papers in Bayesian Analysis are a good source for seeing all sides of this debate.

The property that improper priors need a sufficiently large sample size to yield a proper posterior can be viewed as a strength of objective priors: one realizes whether or not there is enough information in the data to make the unknown parameters identifiable. A proper prior masks the issue by always yielding a proper posterior, which can be dangerous in the non-identifiable case: essentially, for some of the parameters (or functions thereof) there is then no information provided by the data and the posterior is just the prior, which can be quite dangerous if extreme care was not taken in developing the prior.

Dr. O'Reilly makes the very interesting suggestion that one should symmetrize the right-Haar priors by geometric mixing, rather than arithmetic mixing, and observes that this indeed yields the reference prior for ρ. This is an appealing suggestion and strengthens the argument for using the reference prior.

Dr. O'Reilly asks if the reference prior for ρ yields credible sets with frequentist coverage that differ much from nominal. We did conduct limited simulations on this and the answer appears to be no; for instance, even with a minimal sample size of 5, the coverage of a nominal 95% set appears to be no worse than 93%.

REFERENCES

Akaike, H. (1980). The interpretation of improper prior distributions as limits of data dependent proper prior distributions. *J. Roy. Statist. Soc. B* **42**, 46–52.

Bayarri, M. J. (1981). Inferencia Bayesiana sobre el coeficiente de correlación de una población normal bivariante. *Trab. Estadist.* **32**, 18-31.

Berger, J.O. (2006). The case for Objective Bayesian Analysis. *Bayesian Analysis* **1**, 3, 385-402.

Berger, J., O. and Bernardo, J., M.(1992). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 35–60.

Berger, J.O., Liseo, B. and Wolpert, R.L. (1999) Integrated Likelihood Methods for Eliminating Nuisance Parameters. *Statist. Science* **14**, 1, 1–28.

Clarke, B. and Sun, D. (1997). "Reference Priors under the Chi-Squared Distance." *Sankhya* A, **59**, Part II, 215-231.

Clarke, B. and Yuan, A., (2004) " Partial Information Reference priors. *J. Stat. Planning Inference*, **123**, 2, 313-345.

Consonni, G. and Veronese, P. (2001). Conditionally reducible natural exponential families and enriched conjugate priors. *Scandinavian J. Statist.* **28**, 377-406.

Consonni, G. and Veronese, P. (2003). Enriched conjugate and reference priors for the wishart family on symmetric cones. *Ann. Statist.* **31**, 1491-1516.

Consonni, G., Gutiérrez-Peña, E. and Veronese, P. (2004). Reference priors for exponential families with simple quadratic variance function. *J. Multivariate Analysis* **88**, 335-364.

Dawid, A. P. and Stone, M. (1982). The functional-model basis of fiducial inference (with discussion). *Ann. Statist.* **10**, 1054–74.

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. Roy. Statist. Soc. B* **35**, 189–233.

Dawid, A. P., Stone, M., and Zidek, J. V. (2006). The marginalization paradox revisited. In preparation.

Fienberg, S. E. (2006). Does it make sense to be an "Objective Bayesian"? *Bayesian Analysis* **1**, 3, 429-432.

Fraser, D. A. S. (1968). *The Structure of Inference*. New York: Wiley.

Gleser, L.J. and Hwang, J.T. (1987). The non-existence of $100(1-\alpha)\%$ confidence sets of finite expected diameter in errors-in-variables and related models. *Ann. Statist.* **15**, 4, 1351–1362.

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis* **1**, 3, 403-420.

Heath, D. and Sudderth, W. (1978). On finitely additive priors, coherence and extended admissibility. *Ann. Statist.* **6**, 333–345.

Hora, R. B. and Buehler, R. J. (1966). Fiducial theory and invariant estimation. *Ann. Math. Statist.* **37**, 643–56.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.) Oxford: University Press.

Kadane, J. B. (2006). Is "Objective Bayesian Analysis" objective, Bayesian, or wise? *Bayesian Analysis* **1**, 3, 433-436.

Kadane, J., B., Schervish, M. J. and Seidenfeld, T. (1999). *Cambridge Studies in Probability, Induction and Decision Theory*. Cambridge: University Press.

Lad, F. (2006). Objective Bayesian statistics ... Do you buy it? Should we sell it? *Bayesian Analysis* **1**, 3, 441-444.

Lindley, D.V. (1958), Fiducial distributions and Bayes theorem. *J. Roy. Statist. Soc. B* **20**, 102-107.

O'Hagan, A. (2006) Science, subjectivity and software. Comments on the papers by Berger and by Goldstein. *Bayesian Analysis* **1**, 3, 445–450.

O'Reilly, F. and Rueda, R. (2006), "Inferences in the truncated exponential distribution" *Serie Preimpresos*, IIMAS, UNAM, México.

Regazzini, E. (1983). Non conglomerabilità e paradosso di marginalizzazione. In *Sulle probabilità coerenti nel senso di de Finetti*. Clueb, Bologna, Italy.

Regazzini, E. (1987) de Finetti coherence and statistical inference. *Ann. Statist.* **15**, 2, 845–864.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **15**, 2, 416–431.

Savage, L. J., et al. (1962). The Foundations of Statistical Inference: A Discussion. London: Methuen.

Stone, M. and Dawid, A.P. (1972). Un-Bayesian implications of improper Bayes inference in routine statistical problems. *Biometrika* **59**, 369–375.

Sudderth, W. (1980). Finitely additive priors, coherence and the marginalization paradox. *J. Roy. Statist. Soc. B* **42**, 3, 339–341.