Bayesian Analysis (2018) **13**, Number 4, pp. 1193–1214

The Matrix-F Prior for Estimating and Testing Covariance Matrices

Joris Mulder* and Luis Raúl Pericchi†

Abstract. The matrix-F distribution is presented as prior for covariance matrices as an alternative to the conjugate inverted Wishart distribution. A special case of the univariate F distribution for a variance parameter is equivalent to a half-t distribution for a standard deviation, which is becoming increasingly popular in the Bayesian literature. The matrix-F distribution can be conveniently modeled as a Wishart mixture of Wishart or inverse Wishart distributions, which allows straightforward implementation in a Gibbs sampler. By mixing the covariance matrix of a multivariate normal distribution with a matrix-F distribution, a multivariate horseshoe type prior is obtained which is useful for modeling sparse signals. Furthermore, it is shown that the intrinsic prior for testing covariance matrices in non-hierarchical models has a matrix-F distribution. This intrinsic prior is also useful for testing inequality constrained hypotheses on variances. Finally through simulation it is shown that the matrix-variate F distribution has good frequentist properties as prior for the random effects covariance matrix in generalized linear mixed models.

Keywords: matrix-variate F distribution, intrinsic prior, testing inequality constraints, horsehoe prior, hierarchical models.

1 Introduction

In the last decade there has been an increasing development of alternatives for the inverse gamma prior for modeling variance components. An important motivation of this development is the poor performance of vague inverse gamma priors for modeling random effects variances in hierarchical models (Gelman, 2006; Browne and Draper, 2006). As shown by Gelman (2006) a vague inverse gamma prior on the random effects variance can unduly be extremely informative in the case of small samples. Recently the half-t prior is becoming a popular alternative for a standard deviation (Gelman, 2006; Polson and Scott, 2012). This prior is a special case of a univariate F distribution on the variance (as will be shown in this paper). As was shown by Pérez et al. (2017), the F distribution (referred to as the 'scaled beta2' distribution in their paper) can be seen as a robustification of a gamma distribution for a precision parameter by mixing the scale parameter with a gamma distribution. This idea dates back to De Finetti (1961) who considered a scale mixture of normals to obtain the more robust t prior with less prior shrinkage to extreme observations. Pérez et al. further showed various attractive properties of this distribution in robust Bayesian analyses. Applications of the F distribution can be found in variable selection problems (e.g. Liang et al., 2008;

*Tilburg University, Tilburg, The Netherlands, j.mulder3@uvt.nl
†University of Puerto Rico, Rio Piedras, Puerto Rico, luis.pericchi@upr.edu

© 2018 International Society for Bayesian Analysis https://doi.org/10.1214/17-BA1092

Maruyama and George, 2011), multiple testing problems (Scott and Berger, 2006), random effects testing (Westfall and Gönen, 1996; Wang and Sun, 2013), modeling nonnormal data (Bradlow et al., 2002), and as an intrinsic prior in objective Bayesian hypothesis when testing the scale of an exponential distribution (Pericchi, 2005). These references illustrate the broad applicability of the F distribution.

In this paper a matrix-variate generalization of the F distribution is presented which is obtained by robustifying the inverse Wishart distribution (the current default prior distribution for covariance matrices). This is achieved by mixing the scale matrix of an inverse Wishart distribution with a Wishart distribution. The resulting distribution will be referred to as the matrix-variate F distribution, or the matrix-F distribution for short. The work of Dawid (1981) is very relevant for the current paper. Dawid proposed a convenient parameterization of a matrix-F distribution with a unity scale matrix which allows an extension to infinite dimensions. In this paper, we adopt this parameterization, and derive the matrix-F distribution with a free scale matrix by following different routes. Furthermore we highlight various probabilistic and robust properties of the matrix-F distribution and show its potential for tackling challenging statistical modeling problems.

As will be shown, the matrix-F distribution abides the concept of "reciprocity". This implies that when a covariance matrix has a matrix-F distribution, its inverse, the precision matrix, also belongs to this family of distributions. It is flexible to model different behaviors at the origin and in the tails. Furthermore, it can straightforwardly be implemented in a Gibbs sampler as a Wishart mixture of inverse Wishart distributions or a Wishart mixture of Wishart distributions.

Another useful property is that the matrix-F distribution can be used to construct horseshoe type priors which are useful for estimating location parameters in the case of sparse signals (Carvalho et al., 2009; Polson and Scott, 2011). This can be achieved by mixing the covariance matrix of a multivariate normal distribution with a matrix-F distribution. The resulting marginal distribution has the desired pole at zero, a key property of the horseshoe.

The matrix-F distribution is also quite tractable and it arises naturally as an objective prior for hypothesis testing via intrinsic prior methodology (Berger and Pericchi, 1996; Moreno et al., 1998; Berger and Pericchi, 2004). Additionally the matrix-F prior is suitable for testing nonnested hypotheses with inequality constraints on the variances. Finally the usefulness of the matrix F prior is shown for covariance matrices of random effects in hierarchical models, a challenging part of a Bayesian analysis (Browne and Draper, 2006). Empirical Bayes choices (Kass and Natarajan, 2006) are also considered resulting in excellent frequency properties.

It is important to note that other matrix-variate distributions have been proposed for covariance matrices, including Barnard et al. (2000), Mathai (2005), O'Malley and Zaslavsky (2008), Huang and Wand (2013), Gelman et al. (2014) (Chapter 15.4), and Chung et al. (2015). In particular we shall compare the marginally noninformative prior of Huang and Wand with the matrix-F prior by looking at the marginal priors for the standard deviations and correlations and by investigating the performance of these priors for the random effect covariance matrices in generalized linear mixed models.

The paper is organized as follows. In Section 2, we introduce the matrix-variate F distribution and discuss specific properties, such as the role of the hyperparameters and its implementation in a Gibbs sampler. Subsequently in Section 3 we show how to construct horseshoe type distributions using the matrix-F distribution. Section 4 presents Bayesian tests of a precise hypothesis, where the matrix-F prior serves as an intrinsic prior, and an inequality constrained hypothesis test on a covariance matrix. Section 5 compares various matrix-F priors for the random effects covariance matrix with other proposals from the literature. We end the paper with some concluding remarks.

2 The matrix-variate F distribution

To obtain the matrix-variate F distribution, we start with the univariate case. The univariate F with additional scale parameter can be obtained by mixing the scale parameter of an inverse gamma distribution with a gamma distribution, i.e.,

$$\begin{aligned} F(\sigma^2;\nu,\delta,b) &= \int IG(\sigma^2;\tfrac{\delta}{2},\psi^2)\times G(\psi^2;\tfrac{\nu}{2},b^{-1})d\psi^2 \\ &= \int \tfrac{(\psi^2)^{\frac{\delta}{2}}}{\Gamma(\frac{\delta}{2})}(\sigma^2)^{-\frac{\delta}{2}-1}\exp\left\{-\tfrac{\psi^2}{\sigma^2}\right\} \\ &\quad \times \tfrac{b^{-\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\psi^2)^{\frac{\nu}{2}-1}\exp\left\{-\tfrac{\psi^2}{b}\right\}d\psi^2 \\ &= \frac{\Gamma\left(\frac{\delta+\nu}{2}\right)}{\Gamma(\frac{\nu}{2})\Gamma\left(\frac{\delta}{2}\right)b^{\frac{\nu}{2}}}\left(\sigma^2\right)^{\frac{\nu}{2}-1}\left(1+\sigma^2/b\right)^{-\frac{\nu+\delta}{2}}, \end{aligned} \tag{1}$$

where the degrees of freedom ν controls the behavior near the origin, the degrees of freedom δ controls the tail behavior, and b is a scale parameter. In (1), $IG(\sigma^2;\alpha,\beta)$ denotes an inverse gamma distribution for σ^2 with shape parameter α and scale parameter β and $G(\psi^2;\alpha,\beta)$ denotes a gamma distribution for ψ^2 with shape parameter α and rate parameter β. A similar construction was presented by Pérez et al. (2017) who showed that the F distribution (equivalent to their 'scaled beta2' distribution) can be constructed as a gamma mixture of gamma distributions. Setting $b=1$, we obtain the standard F distribution. Interestingly a F distribution on a variance results in the following distribution on the standard deviation,

$$p(\sigma;\nu,\delta,b) = \frac{2\Gamma\left(\frac{\delta+\nu}{2}\right)}{\Gamma(\frac{\nu}{2})\Gamma\left(\frac{\delta}{2}\right)b^{\frac{\nu}{2}}}\sigma^{\nu-1}\left(1+\sigma^2/b\right)^{-\frac{\nu+\delta}{2}}. \tag{2}$$

When setting $\nu=1$, we obtain a half-t distribution for σ with scale $\sqrt{b}$ and degrees of freedom δ. The half-t prior is becoming increasingly popular for modeling scale parameters (Gelman, 2006; Polson and Scott, 2012). Other choices for ν are also of interest (Pérez et al., 2017).

Following a similar line of argument, a $k\times k$ covariance matrix $\boldsymbol{\Sigma}$ can be obtained by mixing the scale matrix of an inverse Wishart distribution with a Wishart distribution, i.e.,

$$
\begin{aligned}
F(\boldsymbol{\Sigma};\nu,\delta,\mathbf{B}) &= \int IW(\boldsymbol{\Sigma};\delta+k-1,\boldsymbol{\Psi})\times W(\boldsymbol{\Psi};\nu,\mathbf{B})d\boldsymbol{\Psi}. \\
&= \int 2^{-\frac{k(\delta+k-1)}{2}}\Gamma_k\left(\tfrac{\delta+k-1}{2}\right)^{-1}|\boldsymbol{\Psi}|^{\frac{\delta+k-1}{2}}|\boldsymbol{\Sigma}|^{-\frac{\delta+2k}{2}}\exp\left\{-\tfrac{1}{2}\mathrm{tr}\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right\} \\
&\quad\times\ 2^{-\frac{k\nu}{2}}|\mathbf{B}|^{-\frac{\nu}{2}}\Gamma_k\left(\tfrac{\nu}{2}\right)^{-1}|\boldsymbol{\Psi}|^{\frac{\nu-k-1}{2}}\exp\left\{-\tfrac{1}{2}\mathrm{tr}\boldsymbol{\Psi}\mathbf{B}^{-1}\right\}d\boldsymbol{\Psi} \\
&= 2^{-\frac{k(\nu+\delta+k-1)}{2}}|\mathbf{B}|^{-\frac{\nu}{2}}\Gamma_k\left(\tfrac{\nu}{2}\right)^{-1}\Gamma_k\left(\tfrac{\delta+k-1}{2}\right)^{-1}|\boldsymbol{\Sigma}|^{-\frac{\delta+2k}{2}} \\
&\quad\times\int|\boldsymbol{\Psi}|^{\frac{\nu+\delta-2}{2}}\exp\left\{-\tfrac{1}{2}\mathrm{tr}\boldsymbol{\Psi}\left[\boldsymbol{\Sigma}^{-1}+\mathbf{B}^{-1}\right]\right\}d\boldsymbol{\Psi} \\
&= \frac{\Gamma_k\left(\frac{\nu+\delta+k-1}{2}\right)}{\Gamma_k\left(\frac{\nu}{2}\right)\Gamma_k\left(\frac{\delta+k-1}{2}\right)|\mathbf{B}|^{\frac{\nu}{2}}}\ |\boldsymbol{\Sigma}|^{\frac{\nu-k-1}{2}}|\mathbf{I}_k+\boldsymbol{\Sigma}\mathbf{B}^{-1}|^{-\frac{\nu+\delta+k-1}{2}} \qquad (3)
\end{aligned}
$$

for degrees of freedom $\nu > k-1$ and $\delta > 0$, and a positive definite scale matrix $\mathbf{B}$. By setting $\mathbf{B} = \mathbf{I}_k$, we obtain the same distribution as proposed by Dawid (1981), which was termed the standard matrix-variate F distribution. Similarly, we shall refer to scaled version in (3) as the matrix-variate F distribution, or matrix-F for short. Note that for the univariate case with $k = 1$, the matrix-F distribution in (3) corresponds to the univariate F distribution in (1).

The first degrees of freedom ν controls the behavior near the origin where the diagonal elements of $\boldsymbol{\Sigma}$ are close to zero. To see this, note that the kernel of the matrix-variate F distribution is

$$
\begin{aligned}
F(\boldsymbol{\Sigma};\nu,\delta,\mathbf{B}) &\propto |\boldsymbol{\Sigma}|^{\frac{\nu-k-1}{2}}|\mathbf{I}_k+\boldsymbol{\Sigma}\mathbf{B}^{-1}|^{-\frac{\nu+\delta+k-1}{2}} \\
&\sim |\boldsymbol{\Sigma}|^{\frac{\nu-k-1}{2}},
\end{aligned}
$$

if $\boldsymbol{\Sigma}$ approximates a $k\times k$ matrix of zeros, in the sense that the elements of $\boldsymbol{\Sigma}\mathbf{B}^{-1}$ go zero, for fixed $\delta > 0$, and $\nu > k-1$. The second degrees of freedom δ controls the behavior in the tails. To see this, note that

$$
\begin{aligned}
F(\boldsymbol{\Sigma};\nu,\delta,\mathbf{B}) &\propto |\boldsymbol{\Sigma}|^{\frac{\nu-k-1}{2}}|\mathbf{I}_k+\boldsymbol{\Sigma}\mathbf{B}^{-1}|^{-\frac{\nu+\delta+k-1}{2}} \\
&\sim |\boldsymbol{\Sigma}|^{\frac{\nu-k-1}{2}}|\boldsymbol{\Sigma}\mathbf{B}^{-1}|^{-\frac{\nu+\delta+k-1}{2}} \\
&\propto |\boldsymbol{\Sigma}|^{-\frac{\delta+2k}{2}}
\end{aligned}
$$

if the diagonal elements of $\boldsymbol{\Sigma}$ go to ∞, in the sense that the elements of $\boldsymbol{\Sigma}\mathbf{B}^{-1}$ go ∞, for fixed $\delta > 0$, and $\nu > k-1$. Notice that the matrix-F distribution has thicker tails than the inverse Wishart distribution. This illustrates that the matrix-F distribution is more robust as a prior for covariance matrices. To see that $\mathbf{B}$ serves as scale matrix note that if $\boldsymbol{\Sigma} \sim F(\boldsymbol{\Sigma};\nu,\delta,\mathbf{I}_k)$, and if $\mathbf{B}$ is a positive definite matrix, it holds that $\mathbf{B}\boldsymbol{\Sigma} \sim F(\boldsymbol{\Sigma};\nu,\delta,\mathbf{B})$.

Remark 1. *The matrix-F distribution satisfies the reciprocity property which implies that the inverse of a matrix-F distributed covariance matrix also has a matrix-variate F distribution. To be precise, if* $\boldsymbol{\Sigma} \sim F(\nu,\delta,\boldsymbol{B})$*, then* $\boldsymbol{\Sigma}^{-1} \sim F(\delta+k-1,\nu-k+1,\boldsymbol{B}^{-1})$*. A proof is given in Appendix A (Mulder and Pericchi, 2018).*

Remark 2. *The matrix-F distribution is consistent under marginalization. This is a consequence of the parameterization of the inverse Wishart distribution, which is similar as in Dawid (1981). To see that this property holds we partition the covariance matrix as* $\boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_{11}\ \boldsymbol{\Sigma}_{12}; \boldsymbol{\Sigma}_{21}\ \boldsymbol{\Sigma}_{22}]$*, where* $\boldsymbol{\Sigma}_{ij}$ *are* $k_i \times k_j$*, for* $i, j = 1, 2$*, with* $k_1 + k_2 = k$ *and* $1 \leq k_1, k_2 \leq k - 1$*, and* $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}'$*, and we let* $\boldsymbol{\Sigma} \sim F(\nu, \delta, \boldsymbol{B})$*. Due to (3), it holds that* $\boldsymbol{\Sigma}_{11} \sim IW(\delta + k_1 - 1, \boldsymbol{\Psi}_{11})$ *and* $\boldsymbol{\Psi}_{11} \sim W(\nu, \boldsymbol{B}_{11})$*. After integrating out* $\boldsymbol{\Psi}_{11}$*, we obtain* $\boldsymbol{\Sigma}_{11} \sim F(\nu, \delta, \boldsymbol{B}_{11})$*.*

Remark 3. *The mean matrix of a matrix-variate F distribution equals* $\frac{\nu}{\delta - 2}\boldsymbol{B}$*, for* $\delta > 2$*. A derivation is given in Appendix B (Mulder and Pericchi, 2018). An expression for the (co)variances of the elements of a matrix-variate F covariance matrix can also be found there.*

Remark 4. *The matrix-variate F distribution can also be obtained as a Wishart mixture of Wisharts,*

$$F(\boldsymbol{\Sigma}; \nu, \delta, \boldsymbol{B}) \quad = \quad \int W(\boldsymbol{\Sigma}; \nu, \boldsymbol{\Psi}^{-1}) \times W(\boldsymbol{\Psi}; \delta + k - 1, \boldsymbol{B}) d\boldsymbol{\Psi}, \tag{4}$$

as well as an inverse Wishart mixture of Wisharts (as pointed out by an anonymous reviewer),

$$F(\boldsymbol{\Sigma}; \nu, \delta, \boldsymbol{B}) \quad = \quad \int W(\boldsymbol{\Sigma}; \nu, \boldsymbol{\Psi}) \times IW(\boldsymbol{\Psi}; \delta + k - 1, \boldsymbol{B}) d\boldsymbol{\Psi}. \tag{5}$$

The derivations are similar to (3). These parameter expansions are useful when modeling a precision matrix with a matrix-F prior.

Remark 5. *The standard matrix-F distribution was originally derived by Olkin and Rubin (1964) via* $\boldsymbol{\Sigma} = \boldsymbol{\Phi}_2^{-\frac{1}{2}} \boldsymbol{\Phi}_1 \boldsymbol{\Phi}_2^{-\frac{1}{2}}$*, where* $\boldsymbol{\Phi}_1 \sim W(\nu, \boldsymbol{I})$ *and* $\boldsymbol{\Phi}_2 \sim W(\boldsymbol{\Phi}_2; \delta + k - 1, \boldsymbol{I})$ *(with a slightly different parameterization). This distribution was referred to as the multivariate beta II distribution by Tan (1969).*

2.1 A minimally informative default prior

For an inverse Wishart prior with degrees of freedom $\delta + k - 1$ and scale matrix $\boldsymbol{\Psi}$, conventional wisdom dictates that a reasonable default prior is obtained by setting a small value for δ, such as the smallest allowed integer 1, and to set $\boldsymbol{\Psi}$ equal to a "minimally informative" prior guess divided by $\delta + k - 1$ (e.g., Kass and Natarajan, 2006). This can be used to specify the hyperparameters of a matrix-variate F prior in a minimally informative setting. In the matrix-variate F distribution in (3), a Wishart distribution with degrees of freedom of ν and scale matrix $\mathbf{B}$ is used for the scale matrix $\boldsymbol{\Psi}$ of the inverse Wishart distribution. Since the mean of this Wishart distribution equals $\nu\mathbf{B}$, it seems reasonable to let $\frac{\nu}{\delta + k - 1}\mathbf{B}$ be equal to our prior guess. Consequently, a minimally informative matrix-variate F prior can be obtained by setting the prior degrees of freedom equal to $\nu = k$, $\delta = 1$ and $\mathbf{B}$ equal to the prior guess.

When prior information is weak, an empirical Bayes choice could be specified for $\mathbf{B}$ instead. Kass and Natarajan (2006) proposed an inverse Wishart prior for the random

effects covariance matrix with minimal information and an empirical Bayes scale matrix, which was denoted by $\mathbf{R}^*$. Due to the known problems of the inverse gamma prior for random effects variances however (e.g., Gelman, 2006), an inverse Wishart prior also does not seem recommendable in general. The matrix-variate F distribution with $\nu = k$, $\delta = 1$, and $\mathbf{B} = \mathbf{R}^*$ seems a promising alternative empirical Bayes solution for the random effects covariance matrix in hierarchical models. The performance of this prior will investigated in Section 5.

2.2 Proper neighboring priors

Traditionally objective Bayesian analyses are performed using noninformative improper priors (Berger, 2006). A "good" improper prior is typically characterized by good frequency properties of the resulting posterior, such as accurate coverage rates of credibility intervals. A known problem of improper prior is however that they may result in improper posteriors. A well-known example is when using the improper prior σ^{-2} for the random effects variance σ^2 in a Bayesian hierarchical model (e.g., Gelman, 2006). It has been argued that the improper prior $(\sigma^2)^{-\frac{1}{2}}$ is a better choice for the random effects variance (Berger, 2006; Berger and Strawderman, 1996).

To perform an approximate objective Bayesian analysis with a proper prior, one can approximate an objective prior with a *proper neighboring prior* (see also Gelman, 2006, for a related discussion). We shall use this term when the posterior based on the proper neighbor can approximate the posterior based on the improper prior to any precision. For example the objective prior $(\sigma^2)^{-\frac{1}{2}}$ can be approximated by a univariate F prior in (1) because $F(\sigma^2; \nu = 1, \delta, b) \sim (\sigma^2)^{-\frac{1}{2}}$ as $b \to \infty$, for any fixed δ.

Similarly, the matrix-F prior can serve as a proper neighbor for improper priors for covariance matrices. For example note that $F(\mathbf{\Sigma}; \nu, \delta, \mathbf{B}) \sim |\mathbf{\Sigma}|^{\frac{\nu-k-1}{2}}$, when letting $\mathbf{B} = b\mathbf{I}_k$ and $b \to \infty$, for fixed δ. Thus by setting $\nu = k$, which is the smallest allowed integer, the matrix-F prior would be a proper neighbor of $|\mathbf{\Sigma}|^{-\frac{1}{2}}$. The coverage rates and classical risk outcomes will be investigated for this choice in generalized linear mixed models in Section 5.

2.3 Implementation in a Gibbs sampler

The matrix-variate F prior can be implemented in a Gibbs sampler using a parameter expansion to ensure efficient Bayesian computation. The parameter expansion follows from the fact that the matrix-F distribution can be written as a Wishart mixture of the scale matrix in an inverse Wishart prior, as noted in (3). Thus, instead of working with $\mathbf{\Sigma} \sim F(\nu, \delta, \mathbf{B})$ directly, one can model $\mathbf{\Sigma} \sim IW(\delta + k - 1, \mathbf{\Psi})$ with $\mathbf{\Psi} \sim W(\nu, \mathbf{B})$, which follows directly from (3). In this parameter expansion, the conditional prior for $\mathbf{\Sigma}|\mathbf{\Psi}$ has an inverse Wishart distribution with $\delta + k - 1$ degrees of freedom and scale matrix $\mathbf{\Psi}$. This is a conjugate prior for a covariance matrix of multivariate normal observations. Furthermore, the conditional prior for $\mathbf{\Psi}|\mathbf{\Sigma}$ has a Wishart distribution with $\nu + \delta + k - 1$ degrees of freedom and scale matrix $(\mathbf{\Sigma}^{-1} + \mathbf{B}^{-1})^{-1}$. Because no information is directly available for $\mathbf{\Psi}$, the conditional posterior for $\mathbf{\Psi}|\mathbf{\Sigma}$ also has a Wishart distribution with

$\nu + \delta + k - 1$ degrees of freedom and scale matrix $(\mathbf{\Sigma}^{-1} + \mathbf{B}^{-1})^{-1}$. Alternatively when modeling a precision matrix, the matrix-F distribution can be implemented in a Gibbs sampler as a Wishart (or inverse Wishart) mixture of Wishart's via (4) or (5).

2.4 Marginal distributions of standard deviations and correlations: A comparison with Huang and Wand (2013)

The matrix-F distribution is related to the distribution proposed by Huang and Wand (2013). Their proposal was to start with an inverse Wishart distribution with a diagonal scale matrix, and mix the diagonal elements of the scale matrix with independent gamma distributions. The main selling point of the prior of Huang and Wand is that the corresponding marginal priors for the standard deviations have half-t distributions (e.g., half-Cauchy), as recommended by Gelman (2006), and the marginal priors for the correlations have beta distributions in the interval $(-1, 1)$ with equal shape parameters (e.g., uniform), as recommended by Barnard et al. (2000). This prior therefore gives some flexibility to tune the marginal distributions of the standard deviations and correlations. It is important to note however that the marginal priors of the correlations are always centered at zero, in the sense that $P(\rho_{ij} < 0) = P(\rho_{ij} > 0) = \frac{1}{2}$, for all $i \neq j \in \{1, \ldots, k\}$ (due to the diagonal scale matrix in the inverse Wishart distribution). This property may not be flexible enough for practical situations. For example in multitrait-multimethod applications the interest is in the correlations between multiple traits (e.g., abilities) measured using different methods (e.g., raters) (Campbell and Fiske, 1959). The correlations of particular interest, i.e., the correlations between the measurements of the same trait using different methods and the correlations between different traits using the same methods, are generally expected to be positive (e.g., Lievens and Conway, 2001; Muis et al., 2007; Mulder, 2016). Further note that the expected magnitude of a correlation generally varies across different areas of research (e.g., medical, social science, or education) (Cohen, 1988). Therefore a prior that is concentrated around zero may be too restrictive for general usage.

When considering a $k \times k$ covariance matrix with a matrix-$F(\nu, \delta, \mathbf{B})$ distribution, the marginal distribution of the j-th variance, σ^2_{jj}, is univariate $F(\nu, \delta, b_{jj})$, with $\nu > k - 1$, $\delta > 0$, $b_{jj} > 0$. This is a consequence of the marginalization property (Remark 2). Thus even though the univariate F distribution on a variance component is equivalent to a half-t on a standard deviation, the marginal distribution of the standard deviation σ_j in a covariance matrix with a matrix-F distribution is not half-t as ν cannot be set to 1 for $k \geq 2$ (which is necessary for the prior to be proper). Another fundamental difference is that the marginal priors for the correlations of a covariance matrix having a matrix-F distribution are not centered at zero. The reason is that the matrix-F distribution is constructed from an unrestricted scale matrix, unlike the prior of Huang and Wand (2013) which is constructed from a diagonal scale matrix. In that sense the matrix-F distribution can be viewed as more flexible for modeling the correlations in a covariance matrix. In the following sections the flexibility of the matrix-F distribution is highlighted in different modeling situations. The matrix-F distribution can therefore be viewed as a natural generalization of the univariate F distribution. In Section 5.2 we come back to the prior of Huang and Wand by comparing its frequency properties

with the matrix-F prior when modeling the covariance matrix of the random effects in a mixed logistic regression model.

3 A multivariate horseshoe prior

There is an increasing interest in the development of horseshoe priors which are known to perform well in sparse situations (Carvalho et al., 2009; Polson and Scott, 2011; Pérez et al., 2017). A key feature of the horseshoe prior is that it has a pole at the origin resulting in heavy shrinkage of small noisy signals. A second key feature of a horseshoe prior is that it has heavy tails to ensure that large observed effects remain large in the posterior. Here we show that the matrix-F distribution is useful for constructing multivariate horseshoe type priors for location parameters.

Consider a multivariate normal prior for $\boldsymbol{\theta}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with a matrix-F distribution on $\boldsymbol{\Sigma}$. Then the marginal prior for $\boldsymbol{\theta}$ with $\nu = k$ and $\delta = 1$ is a horseshoe prior. To see this, note that

$$\begin{aligned}
\pi(\boldsymbol{\theta}) &= \int N(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \times F(\boldsymbol{\Sigma}; \nu, \delta, \mathbf{B}) d\boldsymbol{\Sigma} \\
&= C \int |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right\} |\boldsymbol{\Sigma}|^{\frac{\nu-k-1}{2}} |\mathbf{I}_k + \boldsymbol{\Sigma}\mathbf{B}^{-1}|^{-\frac{\nu+\delta+k-1}{2}} d\boldsymbol{\Sigma} \\
&= C \int \exp\left\{-\tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right\} |\boldsymbol{\Sigma}|^{\frac{\nu-k-2}{2}} |\mathbf{I}_k + \boldsymbol{\Sigma}\mathbf{B}^{-1}|^{-\frac{\nu+\delta+k-1}{2}} d\boldsymbol{\Sigma},
\end{aligned}$$

where the constant C is the product of the normalizing constants of the multivariate normal density and matrix-F density. To see that this distribution has a pole at $\boldsymbol{\mu}$, note that the marginal density at $\boldsymbol{\theta} = \boldsymbol{\mu}$ is given by

$$\begin{aligned}
\pi(\boldsymbol{\theta} = \boldsymbol{\mu}) &= C \int |\boldsymbol{\Sigma}|^{\frac{\nu-k-2}{2}} |\mathbf{I}_k + \boldsymbol{\Sigma}\mathbf{B}^{-1}|^{-\frac{\nu+\delta+k-1}{2}} d\boldsymbol{\Sigma} \\
&= \tilde{C} \int F(\boldsymbol{\Sigma}; \nu - 1, \delta + 1, \mathbf{B}) d\boldsymbol{\Sigma}, \qquad (6)
\end{aligned}$$

where $\tilde{C}$ is equal to C divided by the normalizing constant of $F(\nu - 1, \delta + 1, \mathbf{B})$. The matrix F distribution in (6) is improper when setting $\nu = k$ because the first degrees of freedom would then be equal to $k - 1$. This implies that the density at $\boldsymbol{\theta} = \boldsymbol{\mu}$ in (6) is ∞ for $\nu = k$, resulting in a marginal prior for $\boldsymbol{\theta}$ with a pole at $\boldsymbol{\mu}$.

To get some insights about the thickness of the tails, note that the marginal distribution can be written as follows

$$\begin{aligned}
\pi(\boldsymbol{\theta}) &= \int N(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \times F(\boldsymbol{\Sigma}; k, \delta, \mathbf{B}) d\boldsymbol{\Sigma} \\
&= \iint N(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \times IW(\boldsymbol{\Sigma}; \delta + k - 1, \boldsymbol{\Psi}) \times W(\boldsymbol{\Psi}; k, \mathbf{B}) d\boldsymbol{\Psi} d\boldsymbol{\Sigma} \\
&= \int t(\boldsymbol{\theta}; \boldsymbol{\mu}, \delta^{-1}\boldsymbol{\Psi}, \delta) \times W(\boldsymbol{\Psi}; k, \mathbf{B}) d\boldsymbol{\Psi},
\end{aligned}$$

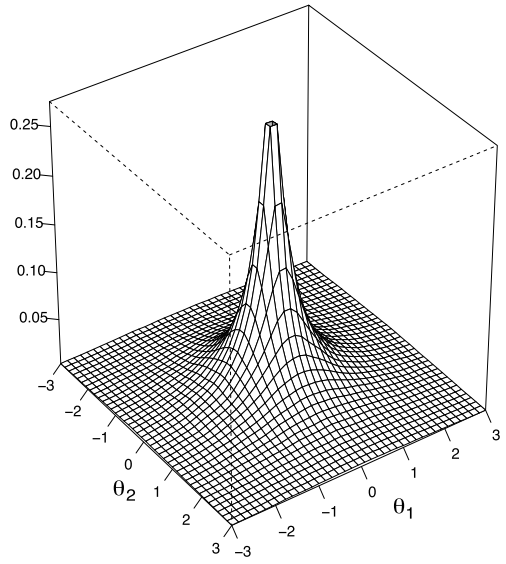

Figure 1: Surface plot of a bivariate horseshoe prior obtained by mixing the normal covariance matrix of $N(\boldsymbol{\theta}; \mathbf{0}, \boldsymbol{\Sigma})$ distribution with a $F(\boldsymbol{\Sigma}; 2, 1, \mathbf{I}_2)$ distribution. The marginal distribution of $\boldsymbol{\theta}$ has a pole at $\mathbf{0}$.

where $t(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Psi}, \nu)$ denotes the multivariate Student t distribution with location parameter $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Psi}$, and degrees of freedom ν. By setting $\delta = 1$, we obtain a multivariate Cauchy prior where the scale matrix is mixed with a Wishart distribution. This implies that the marginal distribution of $\boldsymbol{\theta}$ has heavier tails than a multivariate Cauchy distribution.

Due to the pole at $\boldsymbol{\mu}$ and the heavy tails, the multivariate normal distribution with a matrix-F distribution with $\nu = k$ and $\delta = 1$ on the normal covariance matrix can be seen as a horseshoe type distribution. In Figure 2 a surface plot is presented for the bivariate case with $\mathbf{B} = \mathbf{I}_2$.

4 Testing covariance matrices

4.1 Precise hypothesis testing of a covariance matrix

There has been a considerable interest in the development of intrinsic priors when evaluating statistical models using default Bayes factors. Default Bayes factors are characterized by the fact that the marginal likelihoods are computed in an automatic fashion by splitting the data in a minimal subset referred to as a minimal training sample that is used to construct a (implicit) proper default prior and a remaining set that is used for computing the marginal likelihoods. When a default Bayes factor is approximately equivalent to a Bayes factor based on a certain pair of priors, these priors are

referred to as the intrinsic priors. For this reason the intrinsic prior methodology is of considerable interest because it provides a formal means to construct (proper, or at least well-calibrated or predictively matching; Pericchi, 2005) priors when comparing statistical models using the Bayes factor (which is known to be sensitive to the prior). Interesting references on this topic are Berger and Pericchi (1996), Moreno et al. (1998), Pérez and Berger (2002), Berger and Pericchi (2004), and the references therein.

When testing a null model with a fixed covariance matrix against an unrestricted alternative, the intrinsic prior of the covariance matrix under the unrestricted model has a matrix-variate F distribution.

Theorem 1. *When testing* $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ *versus* $H_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$ *using iid* k*-variate data with* $\boldsymbol{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, for* $i = 1, \ldots, n$*, the intrinsic prior under* H_1 *is given by* $\pi_1^I(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = F(\boldsymbol{\Sigma}; k, 1, \boldsymbol{\Sigma}_0)$ *based on the noninformative improper priors* $\pi_1^N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-\frac{k+1}{2}}$ *and* $\pi_0^N(\boldsymbol{\mu}) = 1$*, and a minimal training sample of size* $m = k + 1$*.*

Proof. Appendix C of Mulder and Pericchi (2018). □

Given the interpretation of the hyperparameters, the intrinsic prior for $\boldsymbol{\Sigma}$ under H_1 contains minimal information because the prior degrees of freedom ν and δ are equal to the smallest allowed integer. Furthermore the intrinsic prior is concentrated around the null value as can be seen from the prior scale matrix which equals $\mathbf{B} = \boldsymbol{\Sigma}_0$. Hence, the intrinsic prior satisfies Jeffreys' heuristic argument that when testing a null value, the prior distribution of the parameter under the unrestricted alternative H_1 should be concentrated around the null value. The argument is that if the null is false it would be reasonable to expect that the covariance matrix is close to the assumed covariance matrix under the null.

For the univariate test, with $k = 1$, the intrinsic prior for the variance equals $F(\sigma^2; 1, 1, \sigma_0^2)$, which corresponds to a half-Cauchy prior with scale σ_0 for the standard deviation. The matrix-variate test has an intrinsic prior with a $F(\boldsymbol{\Sigma}; k, 1, \boldsymbol{\Sigma}_0)$ distribution which does not correspond to half-t priors for the standard deviations (see Section 2.4). This suggests that a natural matrix-variate generalization of the univariate F prior does not have half-t priors for the standard deviations.

It is also interesting to note that the intrinsic prior is the same in the case of a known mean $\boldsymbol{\mu}$. The derivation is similar to Appendix C. Note that in general the intrinsic priors differ in the case of known or unknown nuisance parameters. For example when testing the mean μ of a univariate normal population with unknown variance σ^2, the intrinsic prior is different when the variance is known than when it is unknown (Moreno and Pericchi, 2014). The fact that the intrinsic prior for $\boldsymbol{\Sigma}$ in both testing problems results in the same matrix-F distribution illustrates the broad applicability of this family of prior distributions, and it hints at a possible existence of a unifying approach for modeling variance components.

The intrinsic prior can directly be used to compute the intrinsic Bayes factors when testing a fixed null covariance matrix against the unrestricted alternative. This can be done as follows.

Proposition 1. *The intrinsic Bayes factor in favor of* $H_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$ *against* $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ *using the intrinsic priors* $\pi_1^I(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = F(\boldsymbol{\Sigma}; k, 1, \boldsymbol{\Sigma}_0)$ *and* $\pi_0^I(\boldsymbol{\mu}) = 1$ *equals*

$$\begin{aligned} B_{10} &= \frac{\Gamma_k(k)}{\Gamma_k\left(\frac{k}{2}\right)\Gamma_k\left(\frac{k}{2}\right)}|\boldsymbol{\Sigma}_0|^{\frac{n-k-1}{2}} \exp\left\{\tfrac{1}{2} tr \boldsymbol{\Sigma}_0^{-1}\boldsymbol{S}\right\} \\ &\quad \times \int |\boldsymbol{I}_k + \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}|^{-k}|\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{-\tfrac{1}{2} tr \boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right\} d\boldsymbol{\Sigma}. \end{aligned} \tag{7}$$

Proof. Appendix C of Mulder and Pericchi (2018). □

The integral in (7) can easily be computed via importance sampling, for instance, using an inverse Wishart proposal distribution with $\max(k, n-k-1)$ degrees of freedom and scale matrix $\mathbf{S}$. Below we show that the intrinsic Bayes factor (7) is consistent.

Proposition 2. *The intrinsic Bayes factor in* (7) *of* $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ *versus* $H_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$ *is consistent.*

Proof: Appendix D of Mulder and Pericchi (2018).

4.2 Inequality-constrained hypothesis testing of a covariance matrix

We consider a multivariate normal model for balanced data of K repeated measurement of n individuals, i.e., $\mathbf{y}_i \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, for $i = 1, \ldots, n$, where $\boldsymbol{\theta}$ is the vector of repeated measures means and $\boldsymbol{\Sigma}$ is an unstructured repeated measures covariance matrix. By working with an unstructured covariance matrix we do not have to make any assumptions about a specific multilevel structure of the data. In repeated measures studies researchers are often interested in testing whether individuals tend to become more heterogenous or more homogeneous over time (Böing-Messing and Mulder, 2016; Böing-Messing et al., 2017). For example, Aunola et al. (1994) argued that the variance of math ability of children either increase or decrease over grades. This can be translated to the following inequality-constrained hypothesis test,

$$\begin{aligned} H_1 &: \sigma_1^2 < \ldots < \sigma_k^2 \\ H_2 &: \sigma_1^2 > \ldots > \sigma_k^2 \\ H_3 &: \text{not } H_1, H_2, \end{aligned}$$

where $\sigma_{k'}^2$ is the variance of the k'-th measurement (e.g., grade k') for $k' = 1, \ldots, k$ and the k'-diagonal element of $\boldsymbol{\Sigma}$. We shall write the parameter space under H_t as Σ_t, e.g., $\Sigma_1 = \{\boldsymbol{\Sigma} | \sigma_1^2 < \ldots < \sigma_k^2\}$. Hypothesis H_1 assumes that a strict destabilization occurs over time, H_2 assumes that a strict stabilization occurs, and H_3 assumes that neither a strict destabilization nor a strict stabilization occurs over time.

To our knowledge no criterion has yet been proposed for testing multiple nonnested hypotheses with inequality constraints on the variances in a multivariate normal model. One might consider looking at the posterior probabilities that the inequality constraints hold under a larger unconstrained model. A potential issue of this approach would be

that we would ignore the differences in model complexity of the three hypotheses resulting in a bias towards the larger hypothesis H_3 (Mulder, 2014). Therefore we consider a Bayes factor approach. When testing nonnested hypotheses using the Bayes factor, it is natural to specify an unconstrained prior, denoted by $\pi_u(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, under a larger unconstrained hypothesis, $H_u : \boldsymbol{\sigma}^2 \in \Sigma_u = \Sigma_1 \cup \Sigma_2 \cup \Sigma_3 = (\mathbb{R}^+)^k$, in which H_1, H_2, and H_3 are nested (e.g., Berger and Mortera, 1999). Subsequently truncations of the unconstrained prior are specified under the constrained hypotheses H_t, for $t = 1,\ 2$, and 3, i.e., $\pi_t(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \pi_u(\boldsymbol{\theta}, \boldsymbol{\Sigma}) 1_{\Sigma_t}(\boldsymbol{\Sigma}) / Pr(\boldsymbol{\Sigma} \in \Sigma_t | H_u)$. It has been shown that for this prior choice the Bayes factor of a hypothesis H_t with inequality constraints on means against an unconstrained hypothesis H_u can be written as the ratio of the posterior and prior probability that the constraints of H_t hold under H_u (Klugkist et al., 2005). This is also the case for the Bayes factor of hypotheses with inequality constraints on variance components, i.e.,

$$\begin{aligned} B_{tu} &= \frac{\int\int_{\Sigma_t} f(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) \pi_t(\boldsymbol{\theta}, \boldsymbol{\Sigma}) d\boldsymbol{\theta} d\boldsymbol{\Sigma}}{\int\int_{\Sigma_u} f(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) \pi_u(\boldsymbol{\theta}, \boldsymbol{\Sigma}) d\boldsymbol{\theta} d\boldsymbol{\Sigma}} \\ &= \frac{Pr(\boldsymbol{\Sigma} \in \Sigma_t | \mathbf{Y}, H_u)}{Pr(\boldsymbol{\Sigma} \in \Sigma_t | H_u)}. \end{aligned} \tag{8}$$

The derivation can be found in Appendix A (Mulder and Pericchi, 2018). Because the matrix-variate F distribution with $\nu = k$ and $\delta = 1$ served as an intrinsic prior when testing a precise null with a flat prior for the common nuisance parameters $\boldsymbol{\theta}$, we shall also use this distribution as unconstrained prior. For the scale matrix $\mathbf{B}$ of the matrix-variate F distribution we consider a diagonal matrix with equal diagonal elements, so that the marginal distribution of each variance is equal. Furthermore the prior probability of each of the $k!$ possible orderings of K variances is equal to $(k!)^{-1}$ under H_u, a desirable property when testing inequality-constrained hypotheses (Mulder et al., 2010). Consequently, $Pr(\boldsymbol{\Sigma} \in \Sigma_1 | H_u) = Pr(\boldsymbol{\Sigma} \in \Sigma_2 | H_u) = (k!)^{-1}$, and $Pr(\boldsymbol{\Sigma} \in \Sigma_3 | H_u) = 1 - 2(k!)^{-1}$. The posterior probabilities can be obtained as the proportion of unconstrained draws that satisfy the constraints of H_t. Unconstrained posterior draws can be obtained using a Gibbs sampler by writing the matrix-F prior as a Wishart mixture of inverse Wisharts as in (3). As an alternative we also consider a (default) inverse Wishart distribution with k degrees of freedom with the same scale matrix.

To illustrate the difference between the two prior approaches we consider $k = 3$ repeated measures, and sums of squares of $\mathbf{S} = \text{diag}(1, s, s^2)$ for a data set of $n = 20$ observations. Note that when $s > 1$ the estimates of the variances (ML or Bayesian) satisfy the inequality constraints of the destabilization hypothesis H_1. Figure 2 (left panel) displays the logarithm of the Bayes factor of H_1 versus H_2 (solid lines) and the logarithm of the Bayes factor of H_1 versus H_3 (dashed lines) based on an unconstrained prior with a matrix-variate $F(\boldsymbol{\Sigma}; k, 1, \mathbf{I}_k)$ distribution (black lines) and the Bayes factor based on an unconstrained prior with a $IW(\boldsymbol{\Sigma}; k, \mathbf{I}_k)$ distribution (red lines), as a function of the scale matrix $\mathbf{S} = \text{diag}(1, s, s^2)$, while letting s go from $\exp(0)$ to $\exp(2)$. As can be seen the Bayes factor based on the matrix-variate F distribution results in more evidence for the inequality constrained hypothesis that is supported by the data, H_1, against the other two hypotheses H_2 and H_3, in comparison to the Bayes factors based

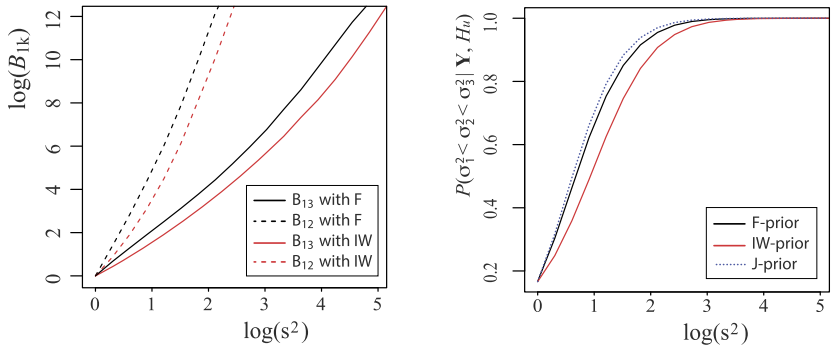

Figure 2: Left panel. Logarithm of the Bayes factor of $H_1 : \sigma_1^2 < \sigma_2^2 < \sigma_2^2$ versus $H_2 : \sigma_1^2 > \sigma_2^2 > \sigma_3^2$ (solid lines) and of H_1 versus the complement hypothesis H_3 (dashed lines) using an encompassing prior with a $F(3, 1, \mathbf{I}_3)$ distribution (black lines) and a $IW(3, \mathbf{I}_3)$ distribution (red lines), as function of the scale matrix $\mathbf{S} = \text{diag}(1, s, s^2)$, while letting s^2 go from exp(0) to exp(5). Right panel. The posterior probability that the constraints of H_1 hold under the unconstrained model using the $F(3, 1, \mathbf{I}_3)$-prior (black solid line), the $IW(3, \mathbf{I}_3)$-prior (red solid line), and the improper Jeffreys prior $|\boldsymbol{\Sigma}|^{-\frac{k+1}{2}}$ (blue dotted line).

on the inverse Wishart prior. As Figure 2 (right panel) shows that this is a consequence of the posterior probability $P(\sigma_1^2 < \sigma_2^2 < \sigma_3^2 | \mathbf{Y}, H_u)$ in the numerator in (8), which is larger when using the matrix-F prior (black solid line) in comparison to the inverse Wishart prior (red solid line). This can be explained by the thicker tails of the matrix-F distribution resulting in less prior shrinkage, an important property of robust priors. As a comparison, the posterior probability based on the matrix-F prior is only slightly smaller than the posterior probability based on the improper Jeffreys' prior $|\boldsymbol{\Sigma}|^{-\frac{k+1}{2}}$ (blue dotted line) with no prior shrinkage.

5 The matrix-F prior for estimating hierarchical models

Prior specification of the covariance matrix of the random effects in a hierarchical Bayesian model is an important but challenging aspect of a Bayesian analysis. In this section we investigate the matrix-F prior for the covariance matrix of the random effects in generalized linear mixed models. To evaluate its performance relative to other proposed priors, we reran various simulations from the literature using the matrix-F prior for the covariance matrix (Natarajan and Kass, 1999; Gelman, 2006; Kass and Natarajan, 2006; Polson and Scott, 2012). We considered minimally informative matrix-variate F priors with $\nu = k$ and $\delta = 1$. Different choices were considered for the scale matrix $\mathbf{B}$. First, we considered an empirical Bayes scale matrix for $\mathbf{B}$, denoted by $\mathbf{R}^*$, as suggested by Kass and Natarajan (2006) (Section 2.1). Second, we considered a proper neighboring prior of the improper prior $|\boldsymbol{\Sigma}|^{-\frac{1}{2}}$, by setting $\mathbf{B} = 10^3\mathbf{I}_k$ (Section 2.2). We also considered the improper prior $|\boldsymbol{\Sigma}|^{-\frac{1}{2}}$ as it is a generalization of a univariate improper

prior $(\sigma^2)^{-\frac{1}{2}}$ that is recommended in the literature for a variance σ^2 (e.g., Berger and Strawderman, 1996; Berger, 2006). Finally we also included the prior of Huang and Wand (2013) which has uniform marginal priors on $(-1, 1)$ for the correlations in the covariance matrix and half-$t(2, 10^5)$ marginal priors for the standard deviations.

5.1 Mixed Poisson regression

Kass and Natarajan (2006) proposed an empirical Bayes prior with an inverse Wishart distribution with minimal degrees of freedom k and an empirical "prior guess", denoted by $\mathbf{R}^*$. Unlike the ML estimate of the covariance matrix of the random effects, the positive definite scale matrix $\mathbf{R}^*$ is always positive definite. The performance of the prior was investigated in a mixed Poisson regression model, where $y_i|b_i \sim \text{Poisson}(\mu_i^b)$, with univariate random intercepts $b_i \sim N(0, \sigma^2)$, and $\mu_i^b = \exp\{\beta_0 + \beta_1 \log(x_i + 10) + \beta_2 x_i + b_i\}$, for $i = 1, \ldots, 18$, with six dosage levels each on three different plates (i.e., $\mathbf{x} = (0 \cdot \mathbf{1}_3', 10 \cdot \mathbf{1}_3', 33 \cdot \mathbf{1}_3', 100 \cdot \mathbf{1}_3', 333 \cdot \mathbf{1}_3', 1000 \cdot \mathbf{1}_3')$), where σ^2 is the random effects variance. Following Kass and Natarajan, the unknown model parameters were set to $\beta_0 = 2.203$, $\beta_1 = .311$, $\beta_2 = -.001$, and $\sigma^2 = .040$.

We reran their simulation using a minimally informative $F(1, 1, B^*)$-prior with empirical Bayes scale R^*, the improper prior $(\sigma^2)^{-\frac{1}{2}}$, the proper neighboring $F(1, 1, 10^3)$ of $(\sigma^2)^{-\frac{1}{2}}$, and the univariate version of the prior of Huang and Wand, i.e., $F(1, 2, 10^{10})$ for the variance, which can also be viewed as a proper neighboring prior of $(\sigma^2)^{-\frac{1}{2}}$. The estimated risk of the fixed effects, i.e., $E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$, the risk based on the entropy loss function for the variance, i.e., $L(\sigma^2, \hat{\sigma}^2) = \hat{\sigma}^2/\sigma^2 - \log(\hat{\sigma}^2/\sigma^2) - 1$, and noncoverage rates are presented in Table 1. The results of the prior of Hwang and Wand were omitted as the results were virtually the same as the proper neighboring prior $F(1, 1, 10^3)$.

The results for $IW(1, R^*)$ and π_{us} (the uniform shrinkage prior of Natarajan and Kass (1999)) were taken from Kass and Natarajan (2006), where harmonic posterior means were used for σ^2 for the computation of the entropy loss function $L(\sigma^2, \hat{\sigma}^2) = \hat{\sigma}^2/\sigma^2 - \log(\hat{\sigma}^2/\sigma^2) - 1$. In our simulations the arithmetic posterior means were used to compute the risk. The results show that the coverage rates for the improper prior σ^{-1} and both F priors are accurate which is not the case for the other priors. The uniform shrinkage prior also shows competitive results for the coverage. Regarding the classical risk, the F priors performed slightly worse than the inverse Wishart prior with empirical Bayes scale.

5.2 Mixed logistic regression model

Natarajan and Kass (1999) presented a simulation of a mixed logistic regression model with a random intercept and random slope, inspired by the work of Zeger and Karim (1991). Conditionally independent Bernoulli responses y_{ij} were generated for $n = 30$ clusters with mean $\text{logit}(\mu_{ij}^{\mathbf{b}}) = \beta_0 + \beta_1 t_j + \beta_3 x_i + \beta_4 x_i t_j + b_{i0} + b_{i1} t_j$, where $x_i = 1$ for half of the samples and 0 elsewhere, and $t_j = j - 4$, for $j = 1, \ldots, 7$. Furthermore, the fixed effects were set to $\boldsymbol{\beta} = (-.625, .25, -.25, .125)'$, and the random effects were generated according to $\mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = [\sigma_1^2 \ \sigma_{12}; \sigma_{12} \ \sigma_2^2] = [.50 \ 0; 0 \ .25]$.

	$IW(1, R^*)$	π_{us}	$F(1, 1, R^*)$	$F(1, 1, 10^3)$	$(\sigma^2)^{-\frac{1}{2}}$
Risk					
$\boldsymbol{\beta}$	$.01 \pm .00$	$.01 \pm .00$	$.11 \pm .00$	$.10 \pm .00$	$.11 \pm .00$
σ^2	$.12 \pm .00$	$.62 \pm .02$	$.23 \pm .01$	$.28 \pm .01$	$.27 \pm .01$
Noncoverage					
β_0	$.056 \pm .007$	$.070 \pm .008$	$.064 \pm .007$	$.047 \pm .007$	$.048 \pm .008$
β_1	$.059 \pm .007$	$.067 \pm .008$	$.065 \pm .007$	$.048 \pm .007$	$.049 \pm .007$
β_2	$.060 \pm .007$	$.075 \pm .008$	$.053 \pm .007$	$.058 \pm .007$	$.051 \pm .007$
σ^2	$.007 \pm .003$	$.037 \pm .006$	$.048 \pm .007$	$.050 \pm .007$	$.045 \pm .007$

Table 1: Risk and noncoverage rates for the (fixed) regression parameters $\boldsymbol{\beta}$ and the random intercept variance σ^2 using different priors.

Again we considered the minimally informative matrix-F prior with empirical Bayes scale matrix $\mathbf{B} = \mathbf{R}^*$, the improper prior $|\boldsymbol{\Sigma}|^{-\frac{1}{2}}$, a proper neighboring matrix-F prior with large scale matrix $\mathbf{B} = 10^3\mathbf{I}_2$, and the prior of Huang and Wand (2013). The models were estimated using the Gibbs sampler of Kinney and Dunson (2007) by approximating the logistic distribution with a mixture of normals. Importance weights were applied to correct for the very small approximation errors.

The simulation of Natarajan and Kass (1999) showed that the approximate uniform shrinkage prior, denoted by π_{us}, performed best in this setting. For this reason, the performance of the other priors is compared with this prior. Tables 2, 3, and 4 presents the classical risk ($E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$, $E[\text{tr}((\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \mathbf{I}_2)^2)]$, $\sum_i E[\hat{b}_{i0} - b_{i0}]$, and $\sum_i E[\hat{b}_{i1} - b_{i1}]$, using posterior means as Bayesian estimates), and the noncoverage rates and average width of the 95% credibility intervals.

Tables 2, 3, and 4 show that the uniform shrinkage prior and the minimally informative empirical Bayes matrix-variate F prior performed best. When estimating the random effects covariance matrix $\boldsymbol{\Sigma}$, the uniform shrinkage prior resulted in a lower risk than the empirical Bayes-F prior. The coverage rates for the covariance σ_{12} on the other hand seems more accurate when using the empirical Bayes matrix-F prior. Regarding the estimation of the fixed and random effects, the performance of the matrix-variate F prior with empirical Bayes scale matrix and the uniform shrinkage prior were similar, with slightly wider intervals for the uniform shrinkage prior. The results of the prior of Huang and Wand (2013), the improper prior and proper neighboring prior were considerably worse than the other two priors.

5.3 Standard random intercept model

Polson and Scott (2012) investigated the classical risk of hypergeometric inverted-beta priors with different combinations of the hyperparameters for the standard deviation of a random intercept. These authors recommended a special case of the inverted-beta prior corresponding to the half-Cauchy prior for the random intercept standard deviation. This half-Cauchy prior corresponds to a univariate ($p = 1$) F distribution on the random intercept variance with hyperparameters $\nu = \delta = 1$ and $b = 1$.

Prior	Risk	Noncoverage			Interval width		
		σ_1^2	σ_{12}	σ_2^2	σ_1^2	σ_{12}	σ_2^2
$\lvert\boldsymbol{\Sigma}\rvert^{-\frac{1}{2}}$	$3.90 \pm .11$	.166	.047	.202	4.05	2.00	1.74
$F(\boldsymbol{\Sigma}; 2, 2, 10^3 \times \mathbf{I}_2)$	$3.84 \pm .11$	.134	.048	.198	3.97	1.99	1.74
$F(\boldsymbol{\Sigma}; 2, 2, \mathbf{R}^*)$	$3.32 \pm .18$	.034	.045	.043	2.11	1.07	.90
π_{us}	$3.10 \pm .19$	.035	.029	.041	2.12	1.05	.88
HW-prior	$7.64 \pm .50$	.070	.009	.110	2.89	1.08	1.28

Table 2: Risk function ($\pm$ standard errors), noncoverage probabilities, and average interval width for $\boldsymbol{\Sigma}$.

Prior	Risk		Noncoverage		Interval width	
	b_0	b_1	b_0	b_1	b_0	b_1
$\lvert\boldsymbol{\Sigma}\rvert^{-\frac{1}{2}}$	$14.99 \pm .25$	$6.08 \pm .13$	.034	.035	3.20	1.99
$F(\boldsymbol{\Sigma}; 2, 2, 10^3 \times \mathbf{I}_2)$	$14.88 \pm .28$	$6.07 \pm .12$	.033	.035	3.18	1.99
$F(\boldsymbol{\Sigma}; 2, 2, \mathbf{R}^*)$	$11.65 \pm .13$	$4.67 \pm .05$	.058	.057	2.54	1.60
π_{us}	$11.51 \pm .12$	$4.51 \pm .05$	.045	.048	2.67	1.63
HW-prior	$12.46 \pm .17$	$5.20 \pm .08$	.049	.046	2.80	1.77

Table 3: Risk function ($\pm$ standard errors), noncoverage probabilities, and average interval width for predictors of the random intercept (b_0) and slope (b_1).

Prior	Risk	Noncoverage				Interval width			
		β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
$\lvert\boldsymbol{\Sigma}\rvert^{-\frac{1}{2}}$	$.58 \pm .02$	.025	.034	.027	.036	1.65	1.03	2.33	1.46
$F(\boldsymbol{\Sigma}; 2, 2, 10^3 \times \mathbf{I}_2)$	$.64 \pm .02$	.046	.050	.043	.034	1.64	1.04	2.31	1.46
$F(\boldsymbol{\Sigma}; 2, 2, \mathbf{R}^*)$	$.44 \pm .01$	.052	.048	.055	.045	1.33	.81	1.89	1.15
π_{us}	$.46 \pm .02$	.033	.058	.044	.045	1.44	.83	2.12	1.19
HW-prior	$.51 \pm .02$	.061	.046	.055	.044	1.45	.91	2.05	1.28

Table 4: Risk function ($\pm$ standard errors), noncoverage probabilities, and average interval width for $\boldsymbol{\beta}$.

We look at a similar example as Polson and Scott (2012). Let $y_i = b_i + \epsilon_i$, with random intercept $b_i \sim N(0, \sigma^2)$ and $\epsilon_i \sim N(0, 1)$, for $i = 1, \ldots, 7$. The priors were evaluated by computing the classical risk given by the mean squared error $||\mathbf{b} - \hat{\mathbf{b}}||^2$. Posterior means were used as Bayesian estimates for the random effects $\mathbf{b}$. Besides the univariate F prior with $\nu = \delta = 1$ and $b = 1$, as recommended by Polson and Scott (2012), we also considered the improper prior of Berger and Strawderman (1996), i.e., $(\sigma^2)^{-1}$; a proper neighboring prior of σ^{-1} having a F distribution with $\nu = 1$, $\delta = .2$, and $b = 10^3$, the univariate version of the prior of Huang and Wand (2013), i.e., $F(1, 2, 10^{10})$ for the variance, another proper neighbor of $(\sigma^2)^{-1}$, and an empirical Bayes F prior with $\nu = \delta = 1$ and scale R^*. We also included the nonhierarchical ML estimate solution, and the James–Stein solution, similar as in Polson and Scott (2012).

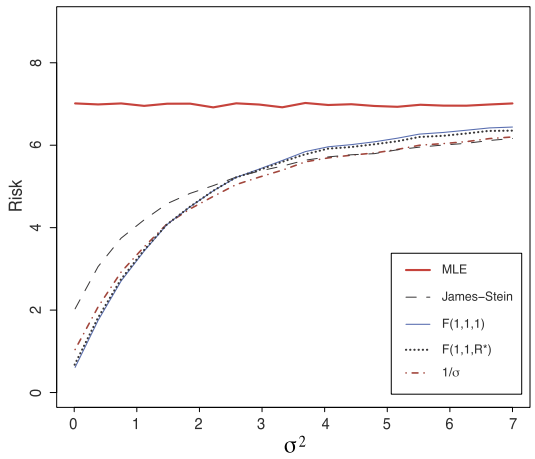

Figure 3: The risk $E(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2)$ as function of the true random effects variance σ^2. Note that the $F(1, 1, 1)$-prior corresponds to a standard half-Cauchy prior on the random effects standard deviation.

The classical risk for each estimate is plotted in Figure 3 as a function of the true variance σ^2. As can be seen in the figure, for very small values of σ^2, the $F(1, 1, 1)$ prior (i.e., a standard half-Cauchy prior for the standard deviation) results in the smallest risk. For medium to large values of σ^2, the $F(1, 1, 1)$ prior performs considerably worse than the improper prior σ^{-1} and its proper neighbors $F(1, .2, 10^3)$ and $F(1, 2, 10^{10})$ (which were virtually identical to σ^{-1} and therefore omitted in the figure), as well as the James–Stein estimate. The empirical Bayes $F(1, 1, R^*)$ prior performs slightly better than the $F(1, 1, 1)$ prior with only slightly higher risk for small values of σ^2 and lower risks for medium to larger values. Overall the improper prior σ^{-1} and its proper neighboring prior $F(1, .2, 10^3)$ seem to have the lowest risk overall. For small values of σ^2 these priors clearly outperform the James–Stein estimate and only do a bit worse than the $F(1, 1, 1)$ prior, for medium values of σ^2 these priors result in the lowest risk, and for larges values the risk based on these priors is only slightly higher than the James–Stein estimate.

6 Summary

In this paper we investigated the potential of the matrix-variate F prior for modeling a covariance matrix in different contexts. Based on our analyses we highlight the following attractive properties.

- The matrix-F prior can straightforwardly be implemented in a Gibbs sampler as a Wishart mixture of inverse Wishart distributions (for modeling a covariance matrix) or as a Wishart or inverse Wishart mixture of Wishart distributions (for modeling a precision matrix).

- If a prior guess for the covariance matrix can be elicited from existing knowledge, this prior guess can be used to specify the scale matrix in the matrix-F prior while the prior degrees of freedom can be tuned depending on the amount of confidence about the prior guess (e.g., $\nu = k$ and $\delta = 1$ in the case of a minimally informative prior guess).

- The matrix-F prior can be used for constructing multivariate horseshoe type priors useful for estimating sparse signals.

- The matrix-F prior serves as an intrinsic prior when testing a covariance matrix of multivariate normal data. This intrinsic prior contains minimal information (i.e., $\nu = k$ and $\delta = 1$) with a scale matrix equal to the null value of the covariance matrix. Interestingly, the intrinsic prior is identical in the case of a known population mean as well as an unknown population mean. The resulting intrinsic Bayes factor is consistent for a precise hypothesis test of a covariance matrix.

- The matrix-F prior is useful as encompassing prior when testing inequality constrained hypotheses on variances. Overall the F prior results in more evidence for an inequality constrained hypothesis that is supported by the data in comparison to its inverse Wishart counterpart.

- The matrix-F prior is promising for modeling the random effects covariance matrix in generalized linear mixed models. A minimally informative matrix-variate F prior with an empirical Bayes scale matrix results in accurate coverage rates and reasonably low risk.

Supplementary Material

Supplementary material for "The matrix-F prior for estimating and testing covariance matrices" (DOI: 10.1214/17-BA1092SUPP; .pdf). The Supplementary Material for "The matrix-F prior for estimating and testing covariance matrices" contains a proof that the matrix-F distribution has the reciprocity property (Section 1); a derivation of the means and (co)variances of the elements of a random matrix having a matrix-F distribution (Section 2); the derivation of the intrinsic prior for a precise hypothesis test of a covariance matrix and the resulting intrinsic Bayes factor (Section 3); a proof that the intrinsic Bayes factor is consistent (Section 4); and a derivation of the Bayes factor of an inequality-constrained covariance matrix against an unconstrained covariance matrix (Section 5).

References

Aunola, K., Leskinen, E., Lerkkanen, M.-K., and Nurmi, J.-E. (1994). "Developmental dynamics of math performance from preschool to grade 2." *Journal of Educational Psychology*, 96: 699–713. 1203

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). "Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage." *Statistica Sinica*, 10: 1282–1311. 1194, 1199

Berger, J. O. (2006). "The case for objective Bbayesian analysis." *Bayesian Analysis*, 1: 385–402. 1198, 1206

Berger, J. O. and Mortera, J. (1999). "Default Bayes factors for nonnested hypothesis testing." *Journal of American Statistical Association*, 94: 542–554. MR1702325. doi: https://doi.org/10.2307/2670175. 1204

Berger, J. O. and Pericchi, L. R. (1996). "The intrinsic Bayes factor for model selection and prediction." *Journal of the American Statistical Association*, 91: 109–122. MR1394065. doi: https://doi.org/10.2307/2291387. 1194, 1202

Berger, J. O. and Pericchi, L. R. (2004). "Training Samples in Objective Bayesian Model Selection." *The Annals of Statistics*, 32(3): 841–869. 1194, 1202

Berger, J. O. and Strawderman, W. E. (1996). "Choice of hierarchical priors: Admissibility in estimation of normal means." *Annals of Statistics*, 24: 931–951. 1198, 1206, 1208

Böing-Messing, F., van Assen, M. A. L. M., Hofman, A. D., Hoijtink, H., and Mulder, J. (2017). "Bayesian Evaluation of Constrained Hypotheses on Variances of Multiple Independent Groups." *Psychological Methods*, 22: 262–287. 1203

Böing-Messing, F. and Mulder, J. (2016). "Automatic Bayes factors for testing variances of two independent normal distributions." *Journal of Mathematical Psychology*, 72: 158–170. 1203

Bradlow, E., Hardie, B., and Faber, P. (2002). "Closed-Form Bayesian Inference for the Negative Binomial Distribution." *Journal of Computational and Graphical Statistics*, 11: 189–202. 1194

Browne, W. J. and Draper, D. (2006). "A comparison of Bayesian and likelihood-based methods for fitting multilevel models." *Bayesian Analysis*, 1: 473–514. 1193, 1194

Campbell, D. T. and Fiske, D. W. (1959). "Convergent and discriminant validation by the multitrait-multimethod matrix." *Psychological Bulletin*, 56: 81–105. 1199

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Sparsity via the horseshoe." *Journal of Machine Learning Research: Workshops and Case Proceedings*, 5: 73–80. 1194, 1200

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., and Dorie, V. (2015). "Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models." *Journal of Educational and Behavioral Statistics*, 40: 136–157. 1194

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum, second edition. 1199

Dawid, A. P. (1981). "Some matrix-variate distribution theory: Notational considerations and a Bayesian application." *Biometrika*, 68: 265–274. 1194, 1196, 1197

De Finetti, B. (1961). *The Bayesian Approach to the Rejection of Outliers*, 199–210. Berkeley, CA: University of California Press. 1193

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1: 515–534. 1193, 1195, 1198, 1199, 1205

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Boca Raton: Chapman & Hall/CRC, third edition. 1194

Huang, A. and Wand, M. P. (2013). "Simple Marginally Noninformative Prior Distributions for Covariance Matrices." *Bayesian Analysis*, 8: 439–452. 1194, 1199, 1206, 1207, 1208

Kass, R. E. and Natarajan, R. (2006). "A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1: 535–542. 1194, 1197, 1205, 1206

Kinney, S. and Dunson, D. B. (2007). "Fixed and Random Effects Selection in Linear and Logistic Models." *Biometrics*, 63: 690–698. 1207

Klugkist, I., Laudy, O., and Hoijtink, H. (2005). "Inequality constrained analysis of variance: A Bayesian approach." *Psychological Methods*, 10: 477–493. 1204

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g priors for Bayesian variable selection." *Journal of American Statistical Association*, 103(481): 410–423. 1193

Lievens, F. and Conway, J. M. (2001). "Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies." *Journal of Applied Psychology*, 86: 1202–1222. 1199

Maruyama, Y. and George, E. (2011). "Fully Bayes factors with a generalized g-prior." *The Annals of Statistics*, 39: 2740–2765. 1193

Mathai, A. M. (2005). "A pathway to matrix-variate gamma and normal densities." *Linear Algebra and Its Applications*, 396: 317–328. 1194

Moreno, E., Bertolino, F., and Racugno, W. (1998). "An intrinsic limiting procedure for model selection and hypotheses testing." *Journal of the American Statistical Association*, 93: 1451–1460. 1194, 1202

Moreno, E. and Pericchi, L. (2014). "Intrinsic Priors for Objective Bayesian Model Selection." *Advances in Econometrics*, 34: 279–300. 1202

Muis, K. R., Winne, P. H., and Jamieson-Noel, D. (2007). "Using a multitrait-multimethod analysis toexamine conceptual similarities of threeself-regulated learning inventories." *British Journal of Educational Psychology*, 77: 177–195. 1199

Mulder, J. (2014). "Bayes factors for testing inequality constrained hypotheses: Issues with Prior Specification." *British Journal of Statistical and Mathematical Psychology*, 67: 153–171. 1204

Mulder, J. (2016). "Bayes factors for testing order-constrained hypotheses on correlations." *Journal of Mathematical Psychology*, 72: 104–115. 1199

Mulder, J., Hoijtink, H., and Klugkist, I. (2010). "Equality and Inequality Constrained Multivariate Linear Models: Objective Model Selection Using Constrained Posterior Priors." *Journal of Statistical Planning and Inference*, 140: 887–906. 1204

Mulder, J. and Pericchi, L. R. (2018). "Supplementary material for "The matrix-F prior for estimating and testing covariance matrices"." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1092SUPP. 1196, 1197, 1202, 1203, 1204

Natarajan, R. and Kass, R. E. (1999). "Reference Bayesian Methods for Generalized Linear Mixed Models." *Journal of the American Statistical Association*, 95: 227–237. MR1803151. doi: https://doi.org/10.2307/2669540. 1205, 1206, 1207

Olkin, I. and Rubin, H. (1964). "Multivariate Beta Distributions and Independence Properties of the Wishart Distribution." *The Annals of Mathematical Statistics*, 35: 261–269. 1197

O'Malley, A. and Zaslavsky, A. (2008). "Domain-level covariance analysis for multilevel survey data with structured nonresponse." *Journal of the American Statistical Association*, 103: 1405–1418. 1194

Pérez, J. M. and Berger, J. O. (2002). "Expected Posterior Prior Distributions for Model Selection." *Biometrika*, 89: 491–511. 1202

Pérez, M. E., Pericchi, L. R., and Ramirez, I. C. (2017). "The Scaled Beta2 Distribution as a Robust Prior for Scales." *Bayesian Analysis*, 12. 1193, 1195, 1200

Pericchi, L. R. (2005). "Model selection and hypothesis testing based on objective probabilities and Bayes factors." *Handbook of Statistics*, 25: 115–149. 1194, 1202

Polson, N. G. and Scott, J. G. (2011). *Shrink globally, act locally: sparse Bayesian regularization and prediction*. Oxford University Press. 1194, 1200

Polson, N. G. and Scott, J. G. (2012). "On the Half-Cauchy Prior for a Global Scale Parameter." *Bayesian Analysis*, 7. 1193, 1195, 1205, 1207, 1208

Scott, J. G. and Berger, J. O. (2006). "An exploration of aspects of Bayesian multiple testing." *Journal of Statistical Planning and Inference*, 136: 2144–2162. 1194

Tan, W. Y. (1969). "Note on the multivariate and the generalized multivariate beta distributions." *Journal of American Statistical Association*, 64: 230–41. MR0240899. 1197

Wang, M. and Sun, X. (2013). "Bayes Factor Consistency for One-way Random Effects Model." *Statistics: A Journal of Theoretical and Applied Statistics*, 47: 1104–1115. 1194

Westfall, P. and Gönen, M. (1996). "Asymptotic properties of anova Bayes factors." *Communications in Statistics: Theory and Methods*, 25: 3101–3123. 1194

Zeger, S. L. and Karim, M. R. (1991). "Generalized Linear Models With Random Effects; A Gibbs Sampling Approach." *Journal of the American Statistical Association*, 86: 79–86. 1206

Acknowledgments

We would like to acknowledge many colleagues for useful discussions, especially Phil Brown, Phil Dawid, Maria Eglee Pérez, Jean-Paul Fox, Maurits Kaptain, and Isabel Ramirez. Furthermore we would like to thank two anonymous reviewers whose remarks greatly contributed to this paper. The first author was supported by a Veni grant of the Netherlands Organization for Scientific Research (NWO).