Bayesian Analysis (2013) **8**, Number 2, pp. 439–452

Simple Marginally Noninformative Prior Distributions for Covariance Matrices

Alan Huang * and M. P. Wand †

Abstract. A family of prior distributions for covariance matrices is studied. Members of the family possess the attractive property of all standard deviation and correlation parameters being marginally noninformative for particular hyperparameter choices. Moreover, the family is quite simple and, for approximate Bayesian inference techniques such as Markov chain Monte Carlo and mean field variational Bayes, has tractability on par with the Inverse-Wishart conjugate family of prior distributions. A simulation study shows that the new prior distributions can lead to more accurate sparse covariance matrix estimation.

Keywords: Bayesian inference, Gibbs sampling, Markov chain Monte Carlo, Mean field variational Bayes

1 Introduction

We study a family of prior distributions for covariance matrices in Bayesian hierarchical models that possess the tractability properties of Inverse-Wishart priors, but have better noninformativity properties. The ease with which these priors can be integrated into the approximate Bayesian inference, via both Markov chain Monte Carlo (MCMC) and mean field variational Bayes (MFVB), is demonstrated. Our proposed family lies within the class of matrix densities developed in Mathai (2005), but our Inverse-Wishart scale mixture representation leads to the abovementioned tractability advantages.

Gelman (2006) argued against the use of Inverse-Gamma priors for variance parameters, with shape and rate parameters set to a positive number ε, on the grounds that they impose a degree of informativity for all ε and posterior inferences are sensitive to the choice of ε. Instead, Uniform and Half-t priors on the standard deviation parameters, with large scale parameters, are recommended. Inverse-Wishart priors on covariance matrices have the same drawbacks as their univariate analogues, since they imply that the variance parameters along the diagonal have Inverse-Gamma distributions.

The proposed covariance matrix family involves a multivariate extension of Result 5 of Wand et al. (2011) which revealed that Half-t distributions arise as a scale mixture of Inverse-Gamma distributions. In the multivariate version, we propose a scale mixture involving an Inverse-Wishart distribution and independent Inverse-Gamma distributions for each dimension. The ensuing covariance matrix distribution is such that all standard deviation parameters have Half-t distributions. In addition, the correlation parameters

*School of Mathematical Sciences, University of Technology, Sydney, Broadway, Australia Alan.Huang@uts.edu.au

†School of Mathematical Sciences, University of Technology, Sydney, Broadway, Australia Matt.Wand@uts.edu.au

 DOI:10.1214/13-BA815

have uniform distributions on $(-1, 1)$ for a particular choice of the Inverse-Wishart shape parameter.

The upshot is that it is possible to choose shape and scale parameters that achieve arbitrarily high noninformativity of all standard deviation and correlation parameters. The joint distributions of these parameters are more difficult to characterize, and we make no claims about their noninformativity. We suspect that joint informativity is inescapable for covariance matrix priors.

A particularly attractive feature of the proposed family of covariance matrix priors is the ease with which it can be incorporated into approximate inference methodology, with a relatively small increase in complexity compared with the common Inverse-Wishart prior. For instance, if the MCMC algorithm is Gibbsian with an Inverse-Wishart prior, then the same is true with the proposed family. Analogous results hold for MFVB approaches to approximate inference.

Note that, apart from Mathai (2005), there are some other recently proposed alternatives to Inverse-Wishart priors for covariance matrices. Examples include those in Barnard et al. (2000) and O'Malley and Zaslavsky (2008).

In Section 2, we describe the family of covariance matrix priors and discuss its theoretical properties in Section 3. A convenient conditional conjugacy property is then illustrated in Section 4, which leads to particularly simple implementations of Gibbs sampling and MFVB approaches. The theory is complemented by a data analysis example using the pigs bodyweight data set from Diggle et al. (2002). A simulation study examining the performance of the proposed prior for sparse covariance matrix estimation is carried out in Section 5. We close with a brief summary in Section 6.

2 Proposed Family and Scale Mixture Representation

The main recommendation in Gelman (2006) is to use Half-t priors on standard deviation parameters to achieve arbitrarily high noninformativity. Half-t priors may appear to be computationally harder to work with than classical Inverse-Gamma priors, but a recent discovery by Wand et al. (2011, Result 5) showed that the Half-t distribution can be written as a scale mixture of simpler Inverse-Gamma distributions. The result for the Half-Cauchy case also appears as Proposition 1 in Armagan et al. (2011). Explicitly,

$$\begin{aligned} &\text{if } \sigma^2 \,|\, a \sim \text{Inverse-Gamma}(\nu/2, \nu/a) \quad \text{and} \quad a \sim \text{Inverse-Gamma}(1/2, 1/A^2) \\ &\text{then } \sigma \sim \text{Half-}t(\nu, A), \end{aligned} \tag{1}$$

where $x \sim \text{Inverse-Gamma}(\alpha, \beta)$ if the corresponding density function satisfies

$$p(x) \propto x^{-\alpha-1}\, e^{-\beta/x}, \quad x > 0$$

and $x \sim \text{Half-t}(\nu, A)$ if the corresponding density function is such that

$$p(x) \propto \{1 + (x/A)^2/\nu\}^{-(\nu+1)/2}, \quad x > 0.$$

The hierarchical representation (1) makes MCMC and MFVB particularly easy to carry out due to the conditional conjugacy properties of the Inverse-Gamma distribution.

Our proposed family arises from an extension of (1) to $p \times p$ random matrices Σ:

$$\begin{aligned} \Sigma|a_1,\ldots,a_p &\sim \text{Inverse-Wishart}\left(\nu+p-1, 2\nu\,\text{diag}(1/a_1,\ldots,1/a_p)\right), \\ a_k &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(1/2, 1/A_k^2\right), \quad k=1,\ldots,p. \end{aligned} \tag{2}$$

Here, $\text{diag}(1/a_1,\ldots,1/a_p)$ denotes a diagonal matrix with $1/a_1,\ldots,1/a_p$ on the diagonal and $\nu, A_1,\ldots,A_p$ are positive scalars. The notation $\overset{\text{ind.}}{\sim}$ stands for "independently distributed as". Lastly, the notation $\Sigma \sim \text{Inverse-Wishart}(\kappa, B)$ means that the density function of Σ is

$$p(\Sigma) \propto |B|^{\kappa/2}|\Sigma|^{-(\kappa+p+1)/2}\exp\{-\tfrac{1}{2}\text{tr}(B\Sigma^{-1})\}, \quad \kappa > 0,\ \Sigma, B \text{ both positive definite.}$$

We show in Section 3.2 that (2) induces Half-$t(\nu, A_k)$ distributions for each standard deviation term σ_k in Σ. Thus, arbitrarily large values of A_k lead to arbitrarily weakly informative priors on the corresponding standard deviation term, as in Gelman (2006). Moreover, in Section 3.3 we show the particular choice $\nu = 2$ leads to marginal uniform distributions for all correlation terms $\rho_{k,k'}$, $k \neq k'$. The distribution characterized by (2) is therefore a matrix generalization of the Half-t prior of Gelman (2006), with the additional property that it can induce marginal uniform distributions for all off-diagonal correlation terms.

It is also possible to write down the density function of Σ in closed form,

$$p(\Sigma) \propto |\Sigma|^{-(\nu+2p)/2}\prod_{k=1}^{p}\left\{\nu\left(\Sigma^{-1}\right)_{kk} + 1/A_k^2\right\}^{-(\nu+p)/2}, \quad \Sigma \text{ positive definite,}$$

from which it can be seen that the proposed family appears implicitly in Mathai (2005). In that paper, an explicit density representation is given for a more general class of matrix distributions and some theoretical properties are derived, but no feasible method to fit such models to data is prescribed. In contrast, the scale mixture representation (2) in this paper leads directly to particularly simple algorithms for fitting the model to data, as we demonstrate in Section 4.

3 Properties

The family of covariance matrices characterized by (2) has some attractive properties, particularly with regard to the marginal distributions of statistically meaningful parameters such as standard deviation and correlation parameters. We lay them out here.

3.1 Distributional Invariance of Sub-Covariance Matrices

An attractive feature of (2) is that any sub-covariance matrix of Σ belongs to the same family of distributions, after an adjustment is made for the reduction in dimension.

This allows for easier derivation of the marginal distributions of the standard deviation and correlation parameters. To see this, partition Σ as

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} ,$$

where Σ_{11} has dimension $p_1 \times p_1$, $p_1 \leq p$. We then have

$$\Sigma_{11}|a_1, \ldots, a_{p_1} \sim \text{Inverse-Wishart}\left(\nu + p_1 - 1, 2\nu \ \text{diag}(1/a_1, \ldots, 1/a_{p_1})\right)$$

by standard Inverse-Wishart distribution theory, with

$$a_k \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(1/2, 1/A_k^2\right), \quad k = 1, \ldots, p_1 \ .$$

Note that this is of the same form as (2). Applying this result to appropriate permutations of the variables underlying Σ leads to the following property:

Property 1 (Distributional invariance of sub-covariance matrices)**.** *The marginal distribution of any sub-covariance matrix in* Σ *has the same distributional form as* Σ *itself.*

Property 1 is desirable because a covariance model should be expandable and collapsible over any set of variables in a self-consistent manner. This property is also nontrivial – for example, the uniform distribution over all correlation matrices, described in Barnard et al. (2000), does not possess this property.

3.2 Marginal Distributions of Standard Deviation Parameters

The marginal distributions of the diagonals of Σ are easy to establish using Property 1. For example, the top left entry, σ_1^2, of Σ has a marginal distribution characterized by

$$\begin{aligned} \sigma_1^2|a_1 &\sim \text{Inverse-Wishart}(\nu, 2\nu/a_1) \ , \\ a_1 &\sim \text{Inverse-Gamma}(1/2, 1/A_1^2) \ , \end{aligned}$$

by Property 1. This in turn implies that the positive square-root σ_1 is distributed as a Half-t distribution with ν degrees of freedom and scale parameter A_1 (Wand et al. 2011, Result 5).

Applying this to appropriate permutations of the variables underlying Σ, it follows that the marginal distributions of the positive square-roots σ_k of the diagonal elements of Σ are all Half-t with degrees of freedom ν and scale parameter A_k.

Property 2 (Half-t standard deviations)**.** *The marginal distribution of any standard deviation term* σ_k *in* Σ *is Half-*$t(\nu, A_k)$.

3.3 Marginal Distributions of Correlation Parameters

It suffices to consider only the marginal distribution of the correlation term in a 2×2 covariance matrix. This is because for any (k, k') correlation term in a $p \times p$ matrix Σ,

we can apply Property 1 to the 2×2 sub-covariance matrix involving only variables k and k'.

Hence, consider a 2×2 matrix Σ with distribution

$$\begin{aligned} \Sigma|a_1, a_2 &\sim \text{Inverse-Wishart}\left(\nu + 2 - 1, 2\nu \,\text{diag}(1/a_1, 1/a_2)\right) , \\ a_k &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(1/2, 1/A_k^2), \quad k = 1, 2. \end{aligned}$$

Parametrize Σ by $(\sigma_1, \sigma_2, \rho)$ through

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} ,$$

so that the conditional density of $(\sigma_1, \sigma_2, \rho)$ given (a_1, a_2) is

$$\begin{aligned} p(\sigma_1, \sigma_2, \rho|a_1, a_2) \quad \propto \quad & (a_1 a_2)^{-\frac{\nu+1}{2}} (1-\rho^2)^{-\frac{\nu}{2}-2} \sigma_1^{-\nu-2} \sigma_2^{-\nu-2} \exp\left\{-\frac{\nu}{a_1(1-\rho^2)\sigma_1^2}\right\} \\ & \times \exp\left\{-\frac{\nu}{a_2(1-\rho^2)\sigma_2^2}\right\}, \quad a_1, a_2, \sigma_1, \sigma_2 > 0, \ -1 < \rho < 1, \end{aligned}$$

after applying the Jacobian formula for a change of variables. The random variables a_1 and a_2 are independent, so their joint density is the product of their marginal densities,

$$p(a_1, a_2) \propto a_1^{-\frac{1}{2}-1} \exp\left(-\frac{1}{a_1 A_1^2}\right) a_2^{-\frac{1}{2}-1} \exp\left(-\frac{1}{a_2 A_2^2}\right), \quad a_1, a_2 > 0.$$

The joint density of $(\sigma_1, \sigma_2, \rho, a_1, a_2)$ is therefore

$$\begin{aligned} p(\sigma_1, \sigma_2, \rho, a_1, a_2) \propto \ & (1-\rho^2)^{-\frac{\nu}{2}-2} \sigma_1^{-\nu-2} \sigma_2^{-\nu-2} a_1^{-\frac{\nu+2}{2}-1} \exp\left[-\frac{1}{a_1}\left\{\frac{\nu}{(1-\rho^2)\sigma_1^2} + \frac{1}{A_1^2}\right\}\right] \\ & \times a_2^{-\frac{\nu+2}{2}-1} \exp\left[-\frac{1}{a_2}\left\{\frac{\nu}{(1-\rho^2)\sigma_2^2} + \frac{1}{A_2^2}\right\}\right], \quad a_1, a_2, \sigma_1, \sigma_2 > 0, \ -1 < \rho < 1. \end{aligned}$$

By noting that the terms involving a_1 form the kernel of an Inverse-Gamma distribution with shape parameter $(\nu+2)/2$ and rate parameter $\frac{\nu}{(1-\rho^2)\sigma_1^2} + \frac{1}{A_1^2}$, and similarly for a_2, we have the marginal density of $(\sigma_1, \sigma_2, \rho)$ as

$$\begin{aligned} p(\sigma_1, \sigma_2, \rho) \quad \propto \quad & (1-\rho^2)^{-\frac{\nu}{2}-2} \sigma_1^{-\nu-2} \sigma_2^{-\nu-2} \left\{\frac{\nu}{(1-\rho^2)\,\sigma_1^2} + \frac{1}{A_1^2}\right\}^{-(\nu+2)/2} \\ & \times \left\{\frac{\nu}{(1-\rho^2)\,\sigma_2^2} + \frac{1}{A_2^2}\right\}^{-(\nu+2)/2}, \quad \sigma_1, \sigma_2 > 0, \ -1 < \rho < 1. \end{aligned}$$

The marginal distribution of ρ can then be obtained by integrating out σ_1 and σ_2, which leads to

$$p(\rho) \propto (1-\rho^2)^{\frac{\nu}{2}-1}, \ -1 < \rho_{ij} < 1 \ .$$

We have established the following two properties:

Property 3 (Explicit marginal density function of correlation parameters)**.** *The marginal distribution of any correlation parameter ρ_{ij} in Σ has density*

$$p(\rho_{ij}) \propto (1-\rho_{ij}^2)^{\frac{\nu}{2}-1}, \quad -1 < \rho_{ij} < 1 \ .$$

Property 4 (Uniformly distributed correlation parameters)**.** *For the particular choice $\nu = 2$, the marginal distribution of each correlation parameter ρ_{ij} in Σ is uniform on $(-1, 1)$.*

4 Conditional Conjugacy and Gibbs Sampling

A family of prior distributions for a parameter is called *conditionally conjugate* if the conditional posterior distribution, given the data and all other parameters in the model, is also in that class (Gelman 2006). From a computational point of view, conditional conjugacy allows Gibbs sampling to be performed for the posterior distribution, provided the prior can be sampled from. From a statistical point of view, conditional conjugacy allows a prior distribution to be interpreted in terms of equivalent data (e.g. Box and Tiao 1973). Furthermore, conditional conjugacy is preserved when the model is expanded hierarchically (Gelman 2006). We illustrate the conditional conjugacy property of our model using two linear mixed models examples.

4.1 Marginal Longitudinal Regression Model

Consider the following marginal longitudinal regression model for n repeated observations on m subjects,

$$\begin{aligned} &\boldsymbol{y}|\boldsymbol{\beta}, \Sigma \sim N(X\boldsymbol{\beta}, I_m \otimes \Sigma), \quad \boldsymbol{\beta} \sim N(0, \sigma_\beta^2 I_b), \\ &\text{and the prior distribution of } \Sigma \text{ given by (2)} \ , \end{aligned} \tag{3}$$

which assumes an unstructured covariance structure for observations within an individual and independence of observations across individuals. The hyperparameters are $\sigma_\beta, A_1, \ldots, A_n$ and $\nu > 0$, and $\otimes$ denotes a Kronecker product. Note that b is the length of the vector $\boldsymbol{\beta}$.

It is straightforward to show that the full conditionals for the parameters in (3) are given by

$$\begin{aligned} \boldsymbol{\beta}|\text{rest} \ \sim \ & N\Big(\Big\{ X^T(I_m \otimes \Sigma^{-1})X + \sigma_\beta^{-2} I_b \Big\}^{-1} X^T(I_m \otimes \Sigma^{-1})\boldsymbol{y}, \\ & \qquad \Big\{ X^T(I_m \otimes \Sigma^{-1})X + \sigma_\beta^{-2} I_b \Big\}^{-1} \Big), \\ \Sigma|\text{rest} \ \sim \ & \text{Inverse-Wishart}\Big(\nu + m + n - 1, \\ & \qquad \sum_{i=1}^{m} (\boldsymbol{y}_i - X_i\boldsymbol{\beta})(\boldsymbol{y}_i - X_i\boldsymbol{\beta})^T + 2\nu \, \text{diag}(1/a_1, \ldots, 1/a_n) \Big) \quad \text{and} \end{aligned}$$

$$a_k|\text{rest} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\frac{\nu+n}{2}, \nu(\Sigma^{-1})_{kk} \ + 1/A_k^2\right), \ k = 1, \ldots, n \ ,$$

where $(\Sigma^{-1})_{kk}$ denotes the (k,k) entry of Σ^{-1}. The full conditionals each have standard forms and MCMC reduces to Gibbs sampling.

In the same vein, the MFVB approximation:

$$p(\boldsymbol{\beta}, \Sigma, \boldsymbol{a}|\boldsymbol{y}) \approx q(\boldsymbol{\beta})\, q(\Sigma)\, q(a_1, \ldots, a_n) \tag{4}$$

involves standard distributions and simple closed form coordinate ascent updates. Relevant derivations and results are given in Menictas and Wand (2013) for more general semiparametric regression models. For (3), these lead to

$q(\boldsymbol{\beta})$ has a $N(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \Sigma_{q(\boldsymbol{\beta})})$ distribution,

$q(\Sigma)$ has a Inverse-Wishart$(A_{q(\Sigma)}, B_{q(\Sigma)})$ distribution

and $q(a_1, \ldots, a_n)$ is a product of Inverse-Gamma$(A_{q(a_k)}, B_{q(a_k)})$ density functions.

The MFVB coordinate ascent updates are :

$$\begin{aligned}
&\Sigma_{q(\boldsymbol{\beta})} \leftarrow \{X^T(I_m \otimes M_{q(\Sigma^{-1})})X + \sigma_\beta^{-2} I_b\}^{-1} \ ; \ \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \leftarrow \Sigma_{q(\boldsymbol{\beta})} X^T (I_m \otimes M_{q(\Sigma^{-1})})\boldsymbol{y} \\
&B_{q(\Sigma)} \leftarrow \textstyle\sum_{i=1}^m \left\{(\boldsymbol{y}_i - X_i\mu_q(\boldsymbol{\beta}))(\boldsymbol{y}_i - X_i\mu_q(\boldsymbol{\beta}))^T + X_i \Sigma_{q(\boldsymbol{\beta})} X_i^T\right\} \\
&\qquad\qquad + 2\nu \,\text{diag}\left(\mu_{q(1/a_1)}, \ldots, \mu_{q(1/a_n)}\right) \\
&M_{q(\Sigma^{-1})} \leftarrow (\nu + m + n - 1)\, B_{q(\Sigma)}^{-1} \quad ; \quad B_q(a_k) \leftarrow \nu \times \{M_{q(\Sigma^{-1})}\}_{kk} + 1/A_k^2 \\
&\mu_{q(1/a_k)} \leftarrow \tfrac{1}{2}(\nu + n)/B_{q(a_k)}, \ 1 \le k \le n.
\end{aligned} \tag{5}$$

Figure 1 displays graphs corresponding to model (3) and the mean field approximation (4). Graph (a) is the directed acyclic graph representation of (3). Its moral graph (b), obtained by connecting with an edge nodes with a common offspring, shows the full product structure of the joint posterior density function: $p(\boldsymbol{\beta}, \Sigma, \boldsymbol{a}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \Sigma)p(\Sigma|\boldsymbol{a})$. Graph (c), corresponding to the more rigid product restriction (4), is formed by removal of two edges from (b).

4.2 Random Effects Regression Model

Consider a general linear mixed model for longitudinal data on m subjects,

$$\begin{aligned}
&\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{u}_i, \sigma_\varepsilon \overset{\text{ind.}}{\sim} N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{u}_i, \sigma_\varepsilon^2 I) \ , \quad i = 1, \ldots, m, \\
&\boldsymbol{\beta} \sim N(0, \sigma_\beta^2 I_b), \quad \boldsymbol{u}_i \overset{\text{ind.}}{\sim} N(0, \Sigma), \\
&\text{the prior distribution of } \Sigma \text{ is (2) with hyperparameters } A_1, \ldots, A_r \text{ and } \nu > 0, \\
&\text{and the prior distribution of } \sigma_\varepsilon^2 \text{ is (2) with hyperparameters } A_\varepsilon \text{ and } \nu_\varepsilon > 0 \ .
\end{aligned} \tag{6}$$

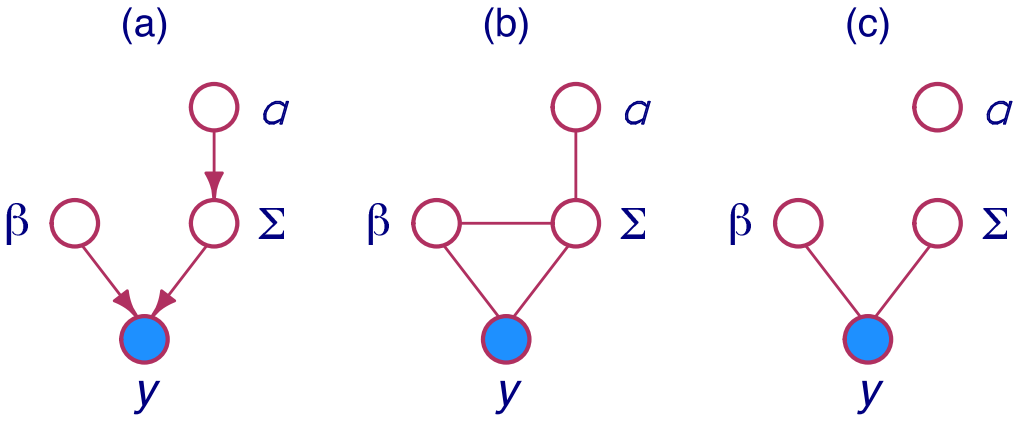

Figure 1: *Graphs relevant to the marginal longitudinal regression model (3) and the mean field variational approximation that gives rise to (5): (a) directed acyclic graph for (3), (b) moral graph for (3), (c) modification of (b) with 2 edges removed to impose product restriction (4). In each graph, shading is used to signify the observed data.*

The number n_i of observations for each individual can be different, so in general, $\boldsymbol{y}_i$ is $n_i \times 1$, X_i is $n_i \times b$, Z_i is $n_i \times r$, $\boldsymbol{\beta}$ is $b \times 1$ and $\boldsymbol{u}_i$ is $r \times 1$. The hyperparameters are $\sigma_\beta, A_1, \ldots, A_r, A_\varepsilon, \nu$ and $\nu_\varepsilon > 0$.

It is again straightforward to compute the full conditionals for all parameters in (6). For notational convenience, define matrices C and M_Σ by

$$C = \begin{bmatrix} X_1 & Z_1 & 0 & \cdots & 0 \\ X_2 & 0 & Z_2 & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ X_m & 0 & \cdots & 0 & Z_m \end{bmatrix}, \quad M_\Sigma = \begin{bmatrix} \sigma_\beta^{-2} I_b & 0 \\ 0 & I_m \otimes \Sigma^{-1} \end{bmatrix},$$

and let $n = n_1 + \ldots + n_m$ be the total sample size. We then have

$$\begin{aligned}
a_\varepsilon|\text{rest} &\sim \text{Inverse-Gamma}\left(\frac{\nu_\varepsilon + 1}{2}, \nu_\varepsilon/\sigma_\varepsilon^2 + 1/A_\varepsilon^2\right), \\
a_k|\text{rest} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\frac{\nu + r}{2}, \nu(\Sigma^{-1})_{kk} + 1/A_k^2\right), \ k = 1, \ldots, r, \\
\sigma_\varepsilon^2|\text{rest} &\sim \text{Inverse-Wishart}\left(\nu_\varepsilon + n, \left\|\boldsymbol{y} - C(\boldsymbol{\beta}, \boldsymbol{u})^T\right\|^2 + 2\nu_\varepsilon/a_\varepsilon\right), \\
\Sigma|\text{rest} &\sim \text{Inverse-Wishart}\left(\nu + m + r - 1, \sum_{i=1}^m \boldsymbol{u}_i \boldsymbol{u}_i^T + 2\nu\,\text{diag}(1/a_1, \ldots, 1/a_r)\right) \\
\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}\Bigg|\text{rest} &\sim N\left((\sigma_\varepsilon^{-2} C^T C + M_\Sigma)^{-1} \sigma_\varepsilon^{-2} C^T \boldsymbol{y}, (\sigma_\varepsilon^{-2} C^T C + M_\Sigma)^{-1}\right).
\end{aligned}$$

The full conditions are again of standard forms, so Gibbs sampling MCMC and

MFVB are straightforward to implement.

Illustration for pig bodyweights data

Diggle et al. (2002) describe a data set consisting of weekly bodyweight measurements (kg) on 48 pigs over a period of 9 weeks. A plot of the data suggests a linear mixed model, with a random intercept and slope for each pig, to be a suitable model for the data. That is, we consider the model

$$\text{weight}_{ij} = \beta_0 + U_i + (\beta_1 + V_i)\text{week}_{ij} + \varepsilon_{ij}, \qquad i = 1, \ldots, 48,\ j = 1, \ldots, 9, \tag{7}$$

with $\varepsilon_{ij} \overset{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2)$ and

$$\begin{bmatrix} U_i \\ V_i \end{bmatrix} \overset{\text{ind.}}{\sim} N(0, \Sigma), \qquad \text{where} \qquad \Sigma = \begin{bmatrix} \sigma_U^2 & \rho_{UV} \\ \rho_{UV} & \sigma_V^2 \end{bmatrix}.$$

To fit this model, we first rescale the response and the covariate to the unit interval $[0, 1]$. We then use a normal prior for $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ and our proposed prior for σ_ε and Σ, as in Section 4.2, with the values of the hyperparameters, $\sigma_\beta, A_\varepsilon, A_1$ and A_2, all set to 10^5 and the shape parameters, ν_ε and ν, set to 1 and 2, respectively. This corresponds to highly noninformative priors for β, σ_ε and Σ. More precisely, the priors for the intercept β_0 and slope β_1 are independent $N(0, 10^5)$, the prior for σ_ε is Half-Cauchy(10^5), the priors for both standard deviation parameters σ_U and σ_V in Σ are Half-$t(2, 10^5)$, and the prior on the correlation parameter ρ_{UV} in Σ is uniform on $(-1, 1)$.

The posterior distributions for parameters in model (7) are straightforward to estimate using the Gibbs sampling procedure prescribed earlier in this section. In our simulations, samples were thinned by keeping every 5th simulation after an initial burn-in period of 5000 simulations, with the total number of samples kept being 1000.

Figure 2 summarizes the Gibbs output after back-transforming to the original units. Rudimentary diagnostic plots of the Gibbs samples given in the first few columns indicate excellent mixing. The kernel density estimated marginal posterior densities of all model parameters, Bayes estimates and 95% credible sets, based on the Gibbs samples, are also shown. They indicate, for example, significance of the random slope component since the lower endpoint of the 95% credible set for σ_V is strictly positive.

5 Practical Implications for Sparse Covariance Matrix Estimation

The high level of noninformativeness and relatively simple computational implementation of the proposed model would be moot points if there are no discernible improvements in practical performance over the classical Inverse-Wishart prior. Here, we carry out simulations to study the performance of the two approaches in the specific context of sparse covariance matrix estimation. Sparse covariance matrix estimation is becoming

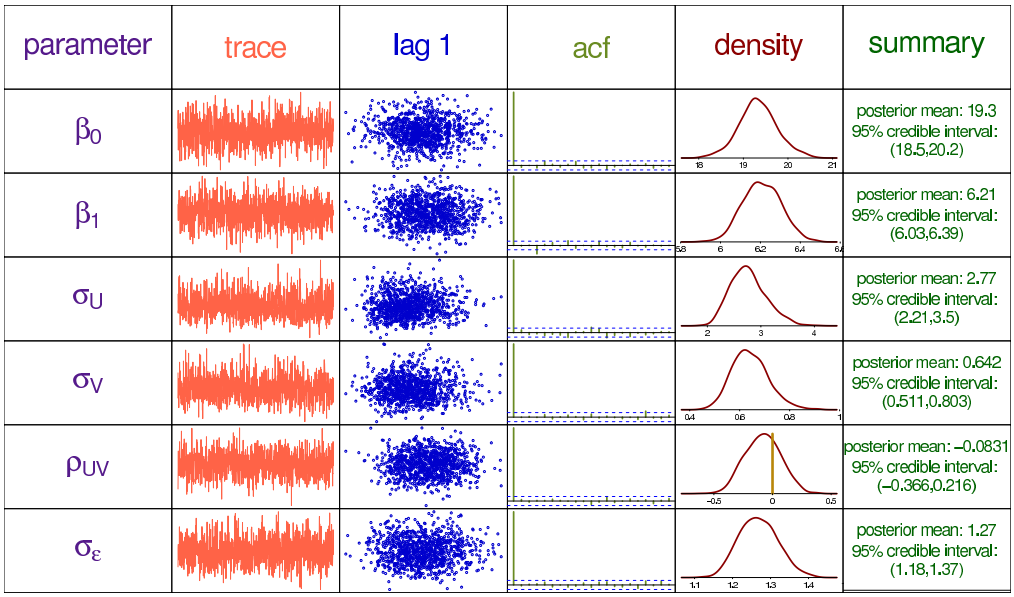

Figure 2: *Summary of MCMC samples for fitting the linear mixed model (7) to the pig bodyweights data. The columns are: (1) parameter name, (2) trace plot of the Gibbs sample, (3) plot of Gibbs sample against its lag 1 sample, (4) sample autocorrelation function, (5) kernel density estimate of posterior density function, (6) numerical summaries of posterior density function.*

increasingly important in high-dimensional data settings, where complex data structures can be substantially simplified through correct identification of independencies between variables (e.g., Bien and Tibshirani 2011).

The synthetic datasets used in our simulations are constructed in the following way. For each simulation, a 10×10 sparse covariance matrix Σ_{true} is constructed by first randomly selecting 10 of its 45 correlation terms to be nonzero. The nonzero correlations are then randomly chosen uniformly between -1 and 1, subject to the condition that the resulting correlation matrix is positive-definite, with the diagonal variance terms all set to 1. Conditional on Σ_{true}, datasets consisting of m independent and identically distributed random vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ are simulated from the multivariate normal distribution $N(\mathbf{0}, \Sigma_{\text{true}})$. This procedure is repeated 1000 times for each sample size. The sample sizes of $m = 20, 40$ and 60 we consider are small compared to the number of free parameters in an unconstrained 10×10 covariance matrix. It is in such situations that any shrinkage effects may be useful.

For each dataset, the proposed model (2) is used to fit the data. That is, we consider

sample size	zero correlations	nonzero correlations	diagonal entries
20	26.7%	11.4%	47.3%
40	13.8%	4.9%	22.1%
60	9.3%	3.1%	15.0%

Table 1: *Average percentage reduction in the posterior mean absolute errors for the proposed prior over the Inverse-Wishart prior when estimating (i) truly zero correlations, (ii) truly nonzero correlations and (iii) diagonal entries, in a* 10×10 *sparse covariance matrix.*

the model

$$\boldsymbol{y}_i | \Sigma \overset{\text{ind.}}{\sim} N(0, \Sigma), \quad i = 1, \ldots, m \ ,$$

the prior distribution of Σ is (2) with hyperparameters $A_1, \ldots, A_{10}$ and $\nu > 0$.

As in Section 4.2, the hyperparameters $A_1, \ldots, A_{10}$ are set to 10^5 and the shape parameter ν set to 2, corresponding to highly noninformative Half-t priors on each standard deviation term and uniform priors on each correlation term. The posterior distributions of $\Sigma | \boldsymbol{y}$ are then estimated using the Gibbs sampling procedure prescribed in Section 4.2, with a burn-in period of 5000 simulations and the total number of samples kept being 1000. We then used each Gibbs sample to approximate the posterior mean absolute error (MAE) for each entry of Σ, defined by

$$\text{MAE}(\Sigma_{kk'}) = E\Big\{ |\Sigma_{kk'} - (\Sigma_{\text{true}})_{kk'}| \,\Big|\, \boldsymbol{y} \Big\}, \quad 1 \le k, k' \le 10.$$

In Table 1, we present the average reduction in the MAEs of the proposed prior over the classical Inverse-Wishart$(2\,\varepsilon, 2\varepsilon I)$ prior for estimating (i) the truly zero correlations, (ii) the truly nonzero correlations, and (iii) the standard deviation terms on the diagonal. In the 1×1 case, the latter prior corresponds to the commonly used Inverse-Gamma$(\varepsilon, \varepsilon)$ prior for variance parameters (e.g., Spiegelhalter et al. 2003). We set $\varepsilon = 0.01$. The averages are taken over both the 1000 replicates and either (i) the 35 truly zero correlations, (ii) the 10 truly nonzero correlations or (iii) the 10 standard deviations, respectively.

We see from Table 1 that the proposed model offers substantial reductions in MAE for correlation terms that are truly zero, with the effect being more pronounced for smaller sample sizes. There appears to be a reasonably strong shrinkage effect for the truly zero correlation terms. For the nonzero correlations, the proposed model still offers a general reduction in MAE, although the effect here is much less pronounced. We also see that the proposed model offers substantial reductions in MAE for the standard deviation terms. These results suggest that the high level of noninformativity offered by our proposed prior, through inducing marginal uniform and Half-t priors on the correlation and standard deviation parameters respectively, translates into an overall increase in practical performance over the classical Inverse-Wishart prior for sparse covariance matrix estimation.

6 Summary

The arguments in Gelman (2006) for using Half-t priors on standard deviation parameters to achieve arbitrarily high noninformativity are compelling. In this paper, we extend this idea to the multivariate setting by introducing a family of covariance matrix priors that induces Half-t priors on each standard deviation term and uniform priors on each correlation. It is demonstrated through simulations that the proposed family can offer better practical performance over the classical Inverse-Wishart prior in the context of sparse covariance matrix estimation. The approach is particularly easy to implement computationally, with a conditional conjugacy property leading to simple, exact Gibbs sampling for the posterior distribution. Moreover, this conditional conjugacy property is preserved when the model is expanded hierarchically, allowing the proposed prior to be used in more complicated Bayesian hierarchical models with minimal additional computational burden.

References

Armagan, A., Dunson, D. B., and Clyde, M. (2011). "Generalized Beta Mixtures of Gaussians." In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems 24*, 523–531. 440

Barnard, J., McCulloch, R., and Meng, X. L. (2000). "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage." *Statistica Sinica*, 10: 1281–1311. 440, 442

Bien, J. and Tibshirani, R. J. (2011). "Sparse estimation of a covariance matrix." *Biometrika*, 98: 807–820. 448

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley. 444

Diggle, P., Heagerty, P., Liang, K.-L., and Zeger, S. (2002). *Analysis of Longitudinal Data*. New York: Cambridge University Press, 2 edition. 440, 447

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1: 515–533. 439, 440, 441, 444, 450

Mathai, A. M. (2005). "A pathway to matrix-variate gamma and normal densities." *Linear Algebra and Its Applications*, 396: 317–328. 439, 440, 441

Menictas, M. and Wand, M. (2013). "Variational inference for marginal longitudinal semiparametric regression." *Stat*, 2: 61–71. 445

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. 440

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R., and Lunn, D. (2003). *BUGS: Bayesian inference using Gibbs sampling*. Medical Research Council Biostatistics Unit, Cambridge, UK.
URL http://www.mrc-bsu.cam.ac.uk/bugs 449

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Frühwirth, R. (2011). "Mean field variational Bayes for elaborate distributions." *Bayesian Analysis*, 6: 847–900. 439, 440, 442

Acknowledgments

We are grateful to Vincent Dorie, Andrew Gelman, Ben Goodrich, Jim Hodges, the Associate Editor and two anonymous referees for their comments. This research was partially supported by Australian Research Council Discovery Project DP110100061.