
Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **63**(3), 2021, 517–541 doi: 10.1111/anzs.12339

The Inverse G-Wishart distribution and variational message passing

Luca Maestrini* and Matt P. Wand

University of Technology Sydney

Summary

Message passing on a factor graph is a powerful paradigm for the coding of approximate inference algorithms for arbitrarily large graphical models. The notion of a factor graph fragment allows for compartmentalisation of algebra and computer code. We show that the Inverse G-Wishart family of distributions enables fundamental variational message passing factor graph fragments to be expressed elegantly and succinctly. Such fragments arise in models for which approximate inference concerning covariance matrix or variance parameters is made, and are ubiquitous in contemporary statistics and machine learning.

Key words: approximate Bayesian inference; G-Wishart distribution; mean-field variational Bayes; scalable statistical methodology

1. Introduction

We argue that a very general family of covariance matrix distributions, known as the *Inverse G-Wishart* family, plays a fundamental role in modularisation of variational inference algorithms via variational message passing when a factor graph fragment (Wand 2017) approach is used. A factor graph fragment, or *fragment* for short, is a sub-graph of the relevant factor graph consisting of a factor and all of its neighbouring nodes. Even though use of the Inverse G-Wishart distribution is not necessary, its adoption allows for fundamental factor graph fragment natural parameter updates to be expressed elegantly and succinctly. An essential aspect of this strategy is that the Inverse G-Wishart distribution is the *only* distribution used for covariance matrix and variance parameters. The family includes as special cases the Inverse Chi-Squared, Inverse Gamma and Inverse Wishart distributions. Therefore, just a single distribution is required which leads to savings in notation and code. While similar comments concerning modularity apply to Monte Carlo-based approaches to approximate Bayesian inference, here we focus on variational inference.

Two of the most common contemporary approaches to fast approximate Bayesian inference are mean-field variational Bayes (e.g. Attias 1999) and expectation propagation (e.g. Minka 2001). Minka (2005) explains how each approach can be expressed as message passing on relevant *factor graphs* with *variational message passing* (Winn & Bishop 2005) being the name used for the message passing version of mean-field variational Bayes. Wand (2017) introduces the concept of *factor graph fragments*, or *fragments* for short, for compartmentalisation of variational message passing into atom-like components. Chen &

*Author to whom correspondence should be addressed.
School of Mathematical and Physical Sciences, University of Technology Sydney, P.O. Box 123, Broadway, NSW 2007, Australia. email: luca.maestrini@uts.edu.au.

Wand (2020) demonstrate the use of fragments for expectation propagation. Explanations of factor graph-based variational message passing that match the current exposition are given in Wand (2017 Sections 2.4–2.5).

Wand (2017 Sections 4.1.2–4.1.3) introduces two variational message passing fragments known as the *Inverse Wishart prior* fragment and the *iterated Inverse G-Wishart* fragment. The first of these simply corresponds to imposing an Inverse Wishart prior on a covariance matrix. In the scalar case, this reduces to imposing an Inverse Chi-Squared or, equivalently, an Inverse Gamma prior on a variance parameter. The iterated Inverse G-Wishart fragment facilitates the imposition of arbitrarily non-informative priors on standard deviation parameters such as members of the Half-t family (Gelman 2006; Polson & Scott 2012). An extension to the covariance matrix case, for which there is the option to impose marginal Uniform distribution priors over the interval $(-1, 1)$ on correlation parameters, is elucidated in Huang & Wand (2013). Mulder & Pericchi (2018) provide a different type of extension that is labelled the *Matrix-F* distribution. These two fragments arise in many classes of Bayesian models, such as both Gaussian and generalised response linear mixed models (e.g. McCulloch, Searle & Neuhaus 2008), Bayesian factor models (e.g. Conti *et al.* 2014), vector autoregressive models (e.g. Assaf *et al.* 2019), and generalised additive mixed models and group-specific curve models (e.g. Harezlak, Ruppert & Wand 2018).

Despite the fundamentalness of Inverse G-Wishart-based fragments for variational message passing, the main reference to date, Wand (2017), is brief in its exposition and contains some errors that affect certain cases. In this article, we provide a detailed exposition of the Inverse G-Wishart distribution in the context of variational message passing and list the Inverse Wishart prior and iterated Inverse G-Wishart fragment updates in full ready-to-code forms. R functions (R Core Team 2021) that implement these algorithms are provided as part of the supplementary material of this article. We also explain the errors in Wand (2017).

Section 2 contains relevant definitions and results concerning the G-Wishart and Inverse G-Wishart distributions. Connections with the Huang–Wand and Matrix-F families of marginally noninformative prior distributions for covariance matrices are summarized in Section 3 and in Section 4 we point to background material on variational message passing. In Sections 5 and 6, we provide detailed accounts of the two variational message passing fragments pertaining to variance and covariance matrix parameters, expanding on what is presented in Wand (2017 Sections 4.1.2 and 4.1.3), and making some corrections to what is presented there. In Section 7, we provide explicit instructions on how the two fragments are used to specify different types of prior distributions on standard deviation and covariance matrix parameters in variational message passing-based approximate Bayesian inference. Section 8 contains an data analytic example that illustrates the use of the covariance matrix fragment update algorithms. Some closing discussion is given in Section 9. A web-supplement contains relevant details.

2. The G-Wishart and Inverse G-Wishart distributions

A random matrix $\boldsymbol{X}$ has an Inverse G-Wishart distribution if and only if $\boldsymbol{X}^{-1}$ has a G-Wishart distribution. In this section, we first review the G-Wishart distribution, which has an established literature. Then we discuss the Inverse G-Wishart distribution and list properties that are relevant to its employment in variational message passing.

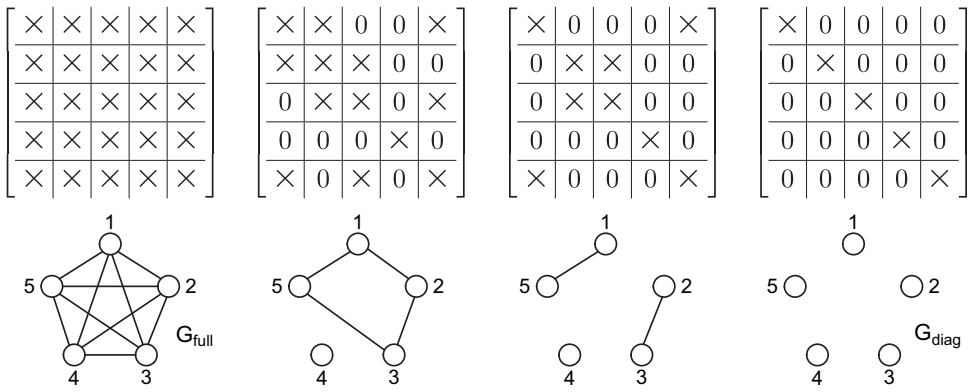

Figure 1. The zero/non-zero entries of four 5×5 symmetric matrices with non-zero entries denoted by $\times$. Underneath each matrix is the 5-node undirected graph that the matrix respects. The nodes are numbered according to the rows and columns of the matrices. A graph edge is present between nodes i and j whenever the (i,j) entry of the matrix is non-zero. The graph respected by the full matrix is denoted by G_{full}. The graph respected by the diagonal matrix is denoted by G_{diag}.

Let G be an undirected graph with d nodes labelled $1,\ldots,d$ and set E consisting of pairs of nodes that are connected by an edge. We say that the symmetric $d \times d$ matrix $\boldsymbol{M}$ *respects* G if

$$\boldsymbol{M}_{ij} = 0, \quad \text{for all} \quad \{i,j\} \notin E.$$

Figure 1 shows the zero/non-zero entries of four 5×5 symmetric matrices. For each matrix, the 5-node graph that the matrix respects is shown underneath.

The first graph in Figure 1 is totally connected and corresponds to the matrix being full. Hence, we denote this graph by G_{full}. At the other end of the spectrum is the last graph of Figure 1, which is totally disconnected. Since this corresponds to the matrix being diagonal, we denote this graph by G_{diag}.

An important concept in G-Wishart and Inverse G-Wishart distribution theory is graph decomposability. An undirected graph G is *decomposable* if and only if all cycles of four or more nodes have an edge that is not part of the cycle but connects two nodes of the cycle. In Figure 1, the first, third and fourth graphs are decomposable. However, the second graph is not decomposable since it contains a four-node cycle that is devoid of edges that connect pairs of nodes within this cycle. Alternative labels for decomposable graphs are *chordal* graphs and *triangulated* graphs.

In Sections 2.1 and 2.2 we define the G-Wishart and Inverse G-Wishart distributions and treat important special cases. This exposition depends on particular notation, which we define here. For a generic proposition $\mathcal{P}$, we define $\mathbf{1}(\mathcal{P})$ to equal 1 if $\mathcal{P}$ is true and zero otherwise. If the random variables x_j, $1 \leqslant j \leqslant d$, are independent such that x_j has distribution $\mathcal{D}_j$ we write $x_j \overset{\text{ind.}}{\sim} \mathcal{D}_j$, $1 \leqslant j \leqslant d$. For a $d \times 1$ vector $\boldsymbol{v}$, let $\text{diag}(\boldsymbol{v})$ be the $d \times d$ diagonal matrix with diagonal comprising the entries of $\boldsymbol{v}$ in order. For a $d \times d$ matrix $\boldsymbol{M}$, let $\text{diagonal}(\boldsymbol{M})$ denote the $d \times 1$ vector comprising the diagonal entries of $\boldsymbol{M}$ in order. The vec and vech matrix operators are well-established (e.g. Gentle 2007). If $\boldsymbol{a}$ is a $d^2 \times 1$ vector

then $\text{vec}^{-1}(\boldsymbol{a})$ is the $d \times d$ matrix such that $\text{vec}\big(\text{vec}^{-1}(\boldsymbol{a})\big) = \boldsymbol{a}$. The matrix $\boldsymbol{D}_d$, known as the *duplication matrix of order* d, is the $d^2 \times \{d(d+1)/2\}$ matrix containing only zeros and ones such that $\boldsymbol{D}_d\,\text{vech}(\boldsymbol{A}) = \text{vec}(\boldsymbol{A})$ for any symmetric $d \times d$ matrix $\boldsymbol{A}$ (Magnus & Neudecker 2019). For example,

$$\boldsymbol{D}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The Moore–Penrose inverse of $\boldsymbol{D}_d$ is $\boldsymbol{D}_d^+ = (\boldsymbol{D}_d^\top \boldsymbol{D}_d)^{-1}\boldsymbol{D}_d^\top$ and is such that $\boldsymbol{D}_d^+\text{vec}(\boldsymbol{A}) = \text{vech}(\boldsymbol{A})$ for a symmetric matrix $\boldsymbol{A}$.

2.1. The G-Wishart distribution

The *G-Wishart distribution* (Atay-Kayis & Massam 2005) is defined as follows:

Definition 1 Let $\boldsymbol{X}$ be a $d \times d$ symmetric and positive definite random matrix and G be a d-node undirected graph such that $\boldsymbol{X}$ respects G. For $\delta > 0$ and a symmetric positive definite $d \times d$ matrix $\boldsymbol{\Lambda}$ we say that $\boldsymbol{X}$ has a G-Wishart distribution with graph G, shape parameter δ and rate matrix $\boldsymbol{\Lambda}$, and write

$$\boldsymbol{X} \sim \text{G-Wishart}(G, \delta, \boldsymbol{\Lambda}),$$

if and only if the non-zero values of the density function of $\boldsymbol{X}$ satisfy

$$p(\boldsymbol{X}) \propto |\boldsymbol{X}|^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\,\text{tr}(\boldsymbol{\Lambda}\boldsymbol{X})\right\}. \tag{1}$$

Obtaining an expression for the normalising factor of a general G-Wishart density function is a challenging problem and recently was resolved by Uhler, Lenkoski & Richards (2018). In the special case where G is a decomposable graph, a relatively simple expression for the normalising factor exists and is given, for example, by (1.4) of Uhler *et al.* (2018). The non-decomposable case is much more difficult and treated in Section 3 of Uhler *et al.* (2018), but the normalising factor does not have a succinct expression for general G. Similar comments apply to expressions for the mean of a G-Wishart random matrix. As discussed in Atay-Kayis & Massam (2005 section 3), the G-Wishart distribution has connections with other distributional constructs such as the hyper Wishart law defined by Dawid & Lauritzen (1993).

Let G_{full} be the totally connected d-node undirected graph and G_{diag} be the totally disconnected d-node undirected graph. The special cases of $G = G_{\text{full}}$ and $G = G_{\text{diag}}$ are such that the normalising factor and mean do have simple closed-form expressions. Since these cases arise in fundamental variational message passing algorithms, we now turn our attention to them.

2.1.1. The $G = G_{\text{full}}$ special case

In the case where G is a fully connected graph we have:

Result 1 If the $d \times d$ random matrix $\boldsymbol{X}$ is such that $\boldsymbol{X} \sim \text{G-Wishart}(G_{\text{full}}, \delta, \boldsymbol{\Lambda})$ then

$$p(\boldsymbol{X}) = \frac{|\boldsymbol{\Lambda}|^{(\delta+d-1)/2}}{2^{d(\delta+d-1)/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{\delta+d-j}{2}\right)}|\boldsymbol{X}|^{(\delta-2)/2}\exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{\Lambda}\boldsymbol{X})\right\}$$
$$\times \mathbf{1}(\boldsymbol{X} \textit{ a symmetric and positive definite } d \times d \text{ matrix}). \tag{2}$$

The mean of $\boldsymbol{X}$ is

$$\text{E}(\boldsymbol{X}) = (\delta + d - 1)\boldsymbol{\Lambda}^{-1}.$$

Result 1 is not novel at all since the $G = G_{\text{full}}$ case corresponds to $\boldsymbol{X}$ having a Wishart distribution. In other words, (2) is simply the density function of a Wishart random matrix. However, it is worth pointing out the shape parameter used here is different from that commonly used for the Wishart distribution. For example, in Gelman *et al.* (2014 table A.1) the shape parameter is denoted by ν and is related to the shape parameter of (2) according to

$$\nu = \delta + d - 1$$

and therefore are the same only in the special case of $\boldsymbol{X}$ being scalar. Also, note that Definition 1 and Result 1 use the rate matrix parameterisation, whereas Gelman *et al.* (2014) Table A.1) use the scale matrix parameterisation for the Wishart distribution. The scale matrix is $\boldsymbol{\Lambda}^{-1}$.

2.1.2. The $G = G_{\text{diag}}$ special case

Before treating the $\boldsymbol{X} \sim \text{G-Wishart}(G_{\text{diag}}, \delta, \boldsymbol{\Lambda})$ situation, we define the notation

$$x \sim \text{Gamma}(\alpha, \beta) \tag{3}$$

to mean that the scalar random variable x has a Gamma distribution with shape parameter α and rate parameter β. The density function corresponding to (3) is

$$p(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)\mathbf{1}(x > 0).$$

The G-Wishart$(G_{\text{diag}}, \delta, \boldsymbol{\Lambda})$ distribution is tied intimately to the Gamma distribution, as Result 2 shows.

Result 2 Suppose that the $d \times d$ random matrix $\boldsymbol{X}$ is such that $\boldsymbol{X} \sim \text{G-Wishart}(G_{\text{diag}}, \delta, \boldsymbol{\Lambda})$. Then the non-zero entries of $\boldsymbol{X}$ satisfy

$$X_{jj} \overset{\text{ind.}}{\sim} \text{Gamma}\left(\frac{1}{2}\delta, \frac{1}{2}\Lambda_{jj}\right), \quad 1 \leqslant j \leqslant d,$$

where Λ_{jj} is the jth diagonal entry of $\boldsymbol{\Lambda}$. The density function of $\boldsymbol{X}$ is

$$p(\boldsymbol{X}) = \frac{|\boldsymbol{\Lambda}|^{\delta/2}}{2^{d\delta/2}\Gamma(\delta/2)^d}|\boldsymbol{X}|^{(\delta-2)/2}\exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{\Lambda}\boldsymbol{X})\right\}\prod_{j=1}^{d}\mathbf{1}(X_{jj} > 0)$$
$$= \frac{\prod_{j=1}^{d}\Lambda_{jj}^{\delta/2}}{2^{d\delta/2}\Gamma(\delta/2)^d}\prod_{j=1}^{d}X_{jj}^{(\delta-2)/2}\exp\left(-\frac{1}{2}\sum_{j=1}^{d}\Lambda_{jj}X_{jj}\right)\prod_{j=1}^{d}\mathbf{1}(X_{jj} > 0).$$

The mean of $\boldsymbol{X}$ is

$$\mathrm{E}(\boldsymbol{X}) = \delta\boldsymbol{\Lambda}^{-1} = \delta\,\mathrm{diag}(1/\Lambda_{11},\ldots,1/\Lambda_{dd}).$$

We now make some remarks concerning Result 2.

(i) When $G = G_{\mathrm{diag}}$ the off-diagonal entries of $\boldsymbol{\Lambda}$ have no effect on the distribution of $\boldsymbol{X}$. In other words, the declaration $\boldsymbol{X} \sim \text{G-Wishart}(G_{\mathrm{diag}}, \delta, \boldsymbol{\Lambda})$ is equivalent to the declaration $\boldsymbol{X} \sim \text{G-Wishart}\big(G_{\mathrm{diag}}, \delta, \mathrm{diag}\{\mathrm{diagonal}(\boldsymbol{\Lambda})\}\big)$.

(ii) The declaration $\boldsymbol{X} \sim \text{G-Wishart}(G_{\mathrm{diag}}, \delta, \boldsymbol{\Lambda})$ is equivalent to the diagonal entries of $\boldsymbol{X}$ being independent Gamma random variables with shape parameter $(1/2)\delta$ and rate parameters equalling the diagonal entries of $(1/2)\boldsymbol{\Lambda}$.

(iii) Even though statements concerning the distributions of independent random variables may seem simpler than a statement of the form $\boldsymbol{X} \sim \text{G-Wishart}(G_{\mathrm{diag}}, \delta, \boldsymbol{\Lambda})$, the major thrust of this article is the elegance provided by key variational message passing fragment updates being expressed in terms of a single family of distributions.

2.1.3. Exponential family form and natural parameterisation

Suppose that $\boldsymbol{X} \sim \text{G-Wishart}(G, \delta, \boldsymbol{\Lambda})$. Then for $\boldsymbol{X}$ such that $p(\boldsymbol{X}) > 0$ we have

$$p(\boldsymbol{X}) \propto \exp\left\{ \begin{bmatrix} \log|\boldsymbol{X}| \\ \mathrm{vech}(\boldsymbol{X}) \end{bmatrix}^{\top} \begin{bmatrix} \frac{1}{2}(\delta-2) \\ -\frac{1}{2}\boldsymbol{D}_d^{\top}\,\mathrm{vec}(\boldsymbol{\Lambda}) \end{bmatrix} \right\} = \exp\{\boldsymbol{T}(\boldsymbol{X})^{\top}\boldsymbol{\eta}\}, \tag{4}$$

where

$$\boldsymbol{T}(\boldsymbol{X}) = \begin{bmatrix} \log|\boldsymbol{X}| \\ \mathrm{vech}(\boldsymbol{X}) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(\delta-2) \\ -\frac{1}{2}\boldsymbol{D}_d^{\top}\,\mathrm{vec}(\boldsymbol{\Lambda}) \end{bmatrix}$$

are respectively sufficient statistic and natural parameter vectors. The inverse of the natural parameter mapping is

$$\begin{cases} \delta = 2(\eta_1 + 1), \\ \boldsymbol{\Lambda} = -2\,\mathrm{vec}^{-1}(\boldsymbol{D}_d^{+\top}\boldsymbol{\eta}_2). \end{cases}$$

Note that, throughout this article, we use $\mathrm{vech}(\boldsymbol{X})$ rather than $\mathrm{vec}(\boldsymbol{X})$ since the former is more compact and avoids duplications. Section S.1 in the web-supplement has further discussion on this matter.

2.2. The Inverse G-Wishart distribution

Suppose that $\boldsymbol{X} \sim \text{G-Wishart}(G, \delta, \boldsymbol{\Lambda})$, where $\boldsymbol{X}$ is $d \times d$, and $\boldsymbol{Y} = \boldsymbol{X}^{-1}$. Let the density functions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ be denoted by $p_{\boldsymbol{X}}$ and $p_{\boldsymbol{Y}}$ respectively. Then the density function of $\boldsymbol{Y}$ is

$$p_{\boldsymbol{Y}}(\boldsymbol{Y}) = p_{\boldsymbol{X}}(\boldsymbol{Y}^{-1})|J(\boldsymbol{Y})|, \tag{5}$$

where

$$J(\boldsymbol{Y}) \equiv \text{the determinant of } \frac{\partial \mathrm{vec}(\boldsymbol{Y}^{-1})}{\partial \mathrm{vec}(\boldsymbol{Y})^{\top}}$$

is the Jacobian of the transformation.

An important observation is that the form of $J(\boldsymbol{Y})$ is dependent on the graph G. In the case of G being a decomposable graph an expression for $J(\boldsymbol{Y})$ is given by (2.4) of Letac & Massam (2007), with credit given to Roverato (2000). Therefore, if G is decomposable, the density function of an Inverse G-Wishart random matrix can be obtained by substitution of (2.4) of Letac & Massam (2007) into (5). However, depending on the complexity of G, simplification of the density function expression may be challenging.

With variational message passing in mind, we now turn to the $G = G_{\text{full}}$ and $G = G_{\text{diag}}$ special cases. The $G = G_{\text{diag}}$ case is simple since it involves products of univariate density functions and we have

$$\text{if } G = G_{\text{diag}} \text{ then } |J(\boldsymbol{Y})| = |\boldsymbol{Y}|^{-2}, \text{ for any } d \in \mathbb{N}. \tag{6}$$

The $G = G_{\text{full}}$ case is more challenging and is the focus of Muirhead (1982 Theorem 2.1.8):

$$\text{if } G = G_{\text{full}}, \text{ then } |J(\boldsymbol{Y})| = |\boldsymbol{Y}|^{-(d+1)}. \tag{7}$$

This result is also stated as Lemma 2.1 in Letac & Massam (2007).

Combining (5), (6) and (7) we have:

Result 3 Suppose that $\boldsymbol{Y} = \boldsymbol{X}^{-1}$, where $\boldsymbol{X} \sim \text{G-Wishart}(G, \delta, \boldsymbol{\Lambda})$ and $\boldsymbol{X}$ is $d \times d$.

(i) If $G = G_{\text{full}}$ then $p(\boldsymbol{Y}) \propto |\boldsymbol{Y}|^{-(\delta+2d)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{Y}^{-1})\right\}$.
(ii) If $G = G_{\text{diag}}$ then $p(\boldsymbol{Y}) \propto |\boldsymbol{Y}|^{-(\delta+2)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{Y}^{-1})\right\}$.

While Result 3 only covers $G = G_{\text{full}}$ or $G = G_{\text{diag}}$ it shows that, in these special cases, the density function of an Inverse G-Wishart random matrix $\boldsymbol{Y}$ is proportional to a power of $|\boldsymbol{Y}|$ multiplied by an exponentiated trace of a matrix multiplied by $\boldsymbol{Y}^{-1}$. This form does not necessarily arise for $G \notin \{G_{\text{full}}, G_{\text{diag}}\}$. Since the motivating variational message passing fragment update algorithms only involve the $G \in \{G_{\text{full}}, G_{\text{diag}}\}$ cases we focus on them for the remainder of this section.

2.2.1. The Inverse G-Wishart distribution when $G \in \{G_{\text{full}}, G_{\text{diag}}\}$

For succinct statement of variational message passing fragment update algorithms involving variance and covariance matrix parameters it is advantageous to have a single Inverse G-Wishart distribution notation for the $G \in \{G_{\text{full}}, G_{\text{diag}}\}$ cases.

Definition 2 Let $\boldsymbol{X}$ be a $d \times d$ symmetric and positive definite random matrix and G be a d-node undirected graph such that $\boldsymbol{X}^{-1}$ respects G. Let $\xi > 0$ and $\boldsymbol{\Lambda}$ be a symmetric positive definite $d \times d$ matrix $\boldsymbol{\Lambda}$.

(i) If $G = G_{\text{full}}$ and ξ is restricted such that $\xi > 2d - 2$, then we say that $\boldsymbol{X}$ has an Inverse G-Wishart distribution with graph G, shape parameter ξ and scale matrix $\boldsymbol{\Lambda}$, and write

$$\boldsymbol{X} \sim \text{Inverse G-Wishart}(G, \xi, \boldsymbol{\Lambda}),$$

if and only if the non-zero values of the density function of $\boldsymbol{X}$ satisfy

$$p(\boldsymbol{X}) \propto |\boldsymbol{X}|^{-(\xi+2)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{X}^{-1})\right\}.$$

(ii) If $G = G_{\text{diag}}$, then say that $\boldsymbol{X}$ has an Inverse G-Wishart distribution with graph G, shape parameter ξ and scale matrix $\boldsymbol{\Lambda}$, and write

$$\boldsymbol{X} \sim \text{Inverse G-Wishart}(G, \xi, \boldsymbol{\Lambda}),$$

if and only if the non-zero values of the density function of $\boldsymbol{X}$ satisfy

$$p(\boldsymbol{X}) \propto |\boldsymbol{X}|^{-(\xi+2)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{X}^{-1})\right\}.$$

(iii) If $G \notin \{G_{\text{full}}, G_{\text{diag}}\}$, then $\boldsymbol{X} \sim \text{Inverse G-Wishart}(G, \xi, \boldsymbol{\Lambda})$ is not defined.

The shape parameter ξ used in Definition 2 is a reasonable compromise between various competing parameterisation choices for the Inverse G-Wishart distribution for $G \in \{G_{\text{full}}, G_{\text{diag}}\}$ and for use in variational message passing algorithms. It has the following attractions:

- The exponent of the determinant in the density function expression is $-(\xi+2)/2$ regardless of whether $G = G_{\text{full}}$ or $G = G_{\text{diag}}$, which is consistent with the G-Wishart distributional notation used in Definition 1.
- In the $d = 1$ case ξ matches the shape parameter in the most common parameterisation of the Inverse Chi-Squared distribution such as that used in Gelman *et al.* (2014, Table A.1).

In the case where $\boldsymbol{X} \sim \text{Inverse G-Wishart}(G_{\text{full}}, \xi, \boldsymbol{\Lambda})$ we have the following:

Result 4 If the $d \times d$ random matrix $\boldsymbol{X}$ is such that $\boldsymbol{X} \sim \text{Inverse G-Wishart}(G_{\text{full}}, \xi, \boldsymbol{\Lambda})$, then

$$p(\boldsymbol{X}) = \frac{|\boldsymbol{\Lambda}|^{(\xi-d+1)/2}}{2^{d(\xi-d+1)/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{\xi-d-j}{2}+1\right)}|\boldsymbol{X}|^{-(\xi+2)/2}\exp\left\{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{X}^{-1})\right\}$$
$$\times \mathbf{1}(\boldsymbol{X} \text{ a symmetric and positive definite } d \times d \text{ matrix}).$$

The mean of $\boldsymbol{X}^{-1}$ is

$$\mathrm{E}(\boldsymbol{X}^{-1}) = (\xi - d + 1)\boldsymbol{\Lambda}^{-1}.$$

Result 4 follows directly from the fact that $\boldsymbol{X} \sim \text{Inverse G-Wishart}(G_{\text{full}}, \xi, \boldsymbol{\Lambda})$ if and only if $\boldsymbol{X}$ has an Inverse Wishart distribution and established results for the density function and mean of this distribution given in, for example, Gelman *et al.* (2014, Table A.1).

We now deal with the $G = G_{\text{diag}}$ case.

Definition 3 Let x be a random variable. For $\delta > 0$ and $\lambda > 0$ we say that the random variable x has an Inverse Chi-Squared distribution with shape parameter δ and rate parameter λ, and write

$$x \sim \text{Inverse-}\chi^2(\delta, \lambda),$$

if and only if $1/x \sim \chi^2(\delta, \lambda)$. If $x \sim \text{Inverse-}\chi^2(\delta, \lambda)$, then the density function of x is

$$p(x) = \frac{(\lambda/2)^{\delta/2}}{\Gamma(\delta/2)}x^{-(\delta+2)/2}\exp\left\{-(\lambda/2)/x\right\}\mathbf{1}(x > 0).$$

Result 5 Suppose that the $d \times d$ random matrix $\boldsymbol{X}$ is such that $\boldsymbol{X} \sim$ Inverse-G-Wishart(G_{diag}, $\xi, \boldsymbol{\Lambda}$). Then the non-zero entries of $\boldsymbol{X}$ satisfy

$$X_{jj} \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2(\xi, \Lambda_{jj}), \quad 1 \leqslant j \leqslant d,$$

where Λ_{jj} is the jth diagonal entry of $\boldsymbol{\Lambda}$. The density function of $\boldsymbol{X}$ is

$$\begin{aligned} p(\boldsymbol{X}) &= \frac{|\boldsymbol{\Lambda}|^{\xi/2}}{2^{d\xi/2}\Gamma(\xi/2)^d} |\boldsymbol{X}|^{-(\xi+2)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{X}^{-1})\right\} \prod_{j=1}^{d} \mathbf{1}(X_{jj} > 0) \\ &= \frac{\prod_{j=1}^{d} \Lambda_{jj}^{\xi/2}}{2^{d\xi/2}\Gamma(\xi/2)^d} \prod_{j=1}^{d} X_{jj}^{-(\xi+2)/2} \exp\left\{-\frac{1}{2}\sum_{j=1}^{d}(\Lambda_{jj}/X_{jj})\right\} \prod_{j=1}^{d} \mathbf{1}(X_{jj} > 0). \end{aligned}$$

The mean of $\boldsymbol{X}^{-1}$ is

$$\mathrm{E}(\boldsymbol{X}^{-1}) = \xi\boldsymbol{\Lambda}^{-1} = \xi \operatorname{diag}(1/\Lambda_{11}, \ldots, 1/\Lambda_{dd}).$$

2.2.2. Natural parameter forms and sufficient statistic expectations

Suppose that $\boldsymbol{X} \sim$ Inverse-G-Wishart($G, \xi, \boldsymbol{\Lambda}$) where $G \in \{G_{\text{full}}, G_{\text{diag}}\}$. Then for $\boldsymbol{X}$ such that $p(\boldsymbol{X}) > 0$,

$$p(\boldsymbol{X}) \propto \exp\left\{\begin{bmatrix} \log|\boldsymbol{X}| \\ \operatorname{vech}(\boldsymbol{X}^{-1}) \end{bmatrix}^{\top} \begin{bmatrix} -(\xi+2)/2 \\ -\frac{1}{2}\boldsymbol{D}_d^{\top}\operatorname{vec}(\boldsymbol{\Lambda}) \end{bmatrix}\right\} = \exp\{\boldsymbol{T}(\boldsymbol{X})^{\top}\boldsymbol{\eta}\},$$

where

$$\boldsymbol{T}(\boldsymbol{X}) = \begin{bmatrix} \log|\boldsymbol{X}| \\ \operatorname{vech}(\boldsymbol{X}^{-1}) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}(\xi+2) \\ -\frac{1}{2}\boldsymbol{D}_d^{\top}\operatorname{vec}(\boldsymbol{\Lambda}) \end{bmatrix} \tag{8}$$

are respectively sufficient statistic and natural parameter vectors. The inverse of the natural parameter mapping is

$$\begin{cases} \xi = -2\eta_1 - 2, \\ \boldsymbol{\Lambda} = -2\operatorname{vec}^{-1}(\boldsymbol{D}_d^{+\top}\boldsymbol{\eta}_2). \end{cases} \tag{9}$$

As explained in Section S.1 of the web-supplement, alternatives to (8) are those that use vec($\boldsymbol{X}$) instead of vech($\boldsymbol{X}$). Throughout this article, we use the more compact 'vech' form.

The following result is fundamental to succinct formulation of updates of covariance and variance parameter fragment updates for variational message passing:

Result 6. If $\boldsymbol{X}$ is a $d \times d$ random matrix that has an Inverse G-Wishart distribution with graph $G \in \{G_{\text{full}}, G_{\text{diag}}\}$ and natural parameter vector $\boldsymbol{\eta}$. Then

$$\mathrm{E}(\boldsymbol{X}^{-1}) = \begin{cases} \{\eta_1 + \frac{1}{2}(d+1)\}\{\operatorname{vec}^{-1}(\boldsymbol{D}_d^{+\top}\boldsymbol{\eta}_2)\}^{-1}, & \text{if } G = G_{\text{full}}, \\ (\eta_1 + 1)\{\operatorname{vec}^{-1}(\boldsymbol{D}_d^{+\top}\boldsymbol{\eta}_2)\}^{-1}, & \text{if } G = G_{\text{diag}}. \end{cases}$$

2.2.3. Relationships with the Hyper Inverse Wishart distributions

Throughout this article we follow the G-Wishart nomenclature as used by, for example, Atay-Kayis & Massam (2005), Letac & Massam (2007) and Uhler *et al.* (2018) in our naming of the Inverse G-Wishart family. Some earlier articles, such as Roverato (2000), use the term *Hyper Inverse Wishart* for the same family of distributions. The naming used here is in keeping with the more recent literature concerning Wishart distributions with graphical restrictions.

3. Connections with some recent covariance matrix distributions

Recently, Huang & Wand (2013) and Mulder & Pericchi (2018) developed covariance matrix distributional families that have attractions in terms of the types of marginal prior distributions that can be imposed on interpretable parameters within the covariance matrix. Mulder & Pericchi (2018) referred to their proposal as the *Matrix-F* family of distributions.

3.1. The Huang–Wand family of distributions

A major motivation for working with the Inverse G-Wishart distribution is the fact that the family of marginally non-informative priors proposed in Huang & Wand (2013) can be expressed succinctly in terms of the Inverse-G-Wishart$(G, \xi, \boldsymbol{\Lambda})$ family where $G \in \{G_{\text{full}}, G_{\text{diag}}\}$. This means that variational message fragments that cater for Huang–Wand prior specification, as well as Inverse-Wishart prior specification, only require natural parameter vector manipulations within a single distributional family.

If $\boldsymbol{\Sigma}$ is a $d \times d$ symmetric positive definite matrix then, for $\nu_{\text{HW}} > 0$ and $s_1, \ldots, s_d > 0$, the specification

$$\begin{aligned} \boldsymbol{\Sigma} \,|\, \boldsymbol{A} &\sim \text{Inverse-G-Wishart}\Big(G_{\text{full}}, \nu_{\text{HW}} + 2d - 2, \boldsymbol{A}^{-1}\Big), \\ \boldsymbol{A} &\sim \text{Inverse-G-Wishart}\Big(G_{\text{diag}}, 1, \big\{\nu_{\text{HW}}\,\text{diag}(s_1^2, \ldots, s_d^2)\big\}^{-1}\Big) \end{aligned} \tag{10}$$

places a distribution of the type given in Huang & Wand (2013) on $\boldsymbol{\Sigma}$ with shape parameter ν_{HW} and scale parameters $s_1, \ldots, s_d$.

The specification (10) matches (2) of Huang & Wand (2013) but with some differences in notation. First, d is used for matrix dimension here rather than p in Huang & Wand (2013). Also, the s_j, $1 \leqslant j \leqslant d$, scale parameters are denoted by A_j in Huang & Wand (2013). The a_j auxiliary variables in (2) of Huang & Wand (2013) are related to the matrix $\boldsymbol{A}$ via the expression $\text{diag}(a_1, \ldots, a_d) = 2\nu_{\text{HW}}\boldsymbol{A}$.

As discussed in Huang & Wand (2013), special cases of (10) correspond to marginally noninformative prior specification of the covariance matrix $\boldsymbol{\Sigma}$ in the sense that the standard deviation parameters $\sigma_j = (\boldsymbol{\Sigma})_{jj}^{1/2}$, $1 \leqslant j \leqslant d$, can have Half-$t$ priors with arbitrarily large-scale parameters, controlled by the s_j values. This is in keeping with the advice given in Gelman (2006). Moreover, correlation parameters $\rho_{jj'} = (\boldsymbol{\Sigma})_{jj'}^{1/2}/(\sigma_j \sigma_{j'})$, for each $j \neq j'$ pair, have a Uniform distribution over the interval $(-1, 1)$ when $\nu_{\text{HW}} = 2$. We refer to this special case as the *Huang–Wand* marginally non-informative prior distribution with scale parameters $s_1, \ldots, s_d$ and write

$$\boldsymbol{\Sigma} \sim \text{Huang–Wand}(s_1, \ldots, s_d) \tag{11}$$

as a shorthand for (10) with $\nu_{\text{HW}} = 2$.

3.2. The Matrix-*F* family of distributions

For $\nu_{\text{MP}} > d-1$, $\delta_{\text{MP}} > 0$ and $\boldsymbol{B}_{\text{MP}}$ a $d \times d$ symmetric positive definite matrix, Mulder & Pericchi (2018) defined a $d \times d$ random matrix $\boldsymbol{\Sigma}$ to have a Matrix-F distribution, written

$$\boldsymbol{\Sigma} \sim F(\nu_{\text{MP}}, \delta_{\text{MP}}, \boldsymbol{B}_{\text{MP}}), \tag{12}$$

if its density function has the form

$$p(\boldsymbol{\Sigma}) \propto \frac{|\boldsymbol{\Sigma}|^{(\nu_{\text{MP}}-d-1)/2}\, \mathbf{1}(\boldsymbol{\Sigma} \text{ symmetric and positive definite})}{|\boldsymbol{I}_d + \boldsymbol{\Sigma}\boldsymbol{B}_{\text{MP}}^{-1}|^{(\nu_{\text{MP}}+\delta_{\text{MP}}+d-1)/2}}.$$

However, standard manipulations of results given in Mulder & Pericchi (2018) show that specification (12) is equivalent to

$$\begin{aligned} \boldsymbol{\Sigma} \,|\, \boldsymbol{A} &\sim \text{Inverse-G-Wishart}\Big(G_{\text{full}}, \delta_{\text{MP}} + 2d - 2, \boldsymbol{A}^{-1}\Big), \\ \boldsymbol{A} &\sim \text{Inverse-G-Wishart}\Big(G_{\text{full}}, \nu_{\text{MP}} + d - 1, \boldsymbol{B}_{\text{MP}}^{-1}\Big) \end{aligned} \tag{13}$$

in the notation used in the current paper. An important difference between (10) and (13) is that the former involves $\boldsymbol{A}$ having an Inverse G-Wishart distribution with the restriction $G = G_{\text{diag}}$, while the latter has $G = G_{\text{full}}$. Section 2.4 of Mulder & Pericchi (2018) compares the two specifications in terms of the types of prior distributions that can be imposed on standard deviation and correlation parameters.

4. Variational message passing background

The overarching goal of this article was to identify and specify algebraic primitives for flexible imposition of covariance matrix priors within a variational message passing framework. In Wand (2017), these algebraic primitives are organised into fragments. This formalism is also used in Nolan & Wand (2017), Maestrini & Wand (2018) and McLean & Wand (2019).

Despite it being a central theme of this article, we will not provide a detailed description of the variational message passing here. Instead, we refer the reader to Wand (2017 Sections 2–4) for the relevant variational message passing background material.

Since the notational conventions for messages used in this section's references are used in the remainder of this article, we summarize them here. If f denotes a generic factor and θ denotes a generic stochastic variable that is a neighbour of f in the factor graph then the message passed from f to θ and the message passed from θ to f are both functions of θ and are denoted by respectively,

$$m_{f\to\theta}(\theta) \quad \text{and} \quad m_{\theta\to f}(\theta).$$

Typically, the messages are proportional to an exponential family density function with sufficient statistic $\boldsymbol{T}(\theta)$, and we have

$$m_{f\to\theta}(\theta) \propto \exp\Big\{\boldsymbol{T}(\theta)^\top \boldsymbol{\eta}_{f\to\theta}\Big\} \quad \text{and} \quad m_{\theta\to f}(\theta) \propto \exp\Big\{\boldsymbol{T}(\theta)^\top \boldsymbol{\eta}_{\theta\to f}\Big\},$$

where $\boldsymbol{\eta}_{f\to\theta}$ and $\boldsymbol{\eta}_{\theta\to f}$ are the message natural parameter vectors. Such vectors play a central role in variational message passing iterative algorithms. We also adopt the notation

$$\boldsymbol{\eta}_{f\leftrightarrow\theta} = \boldsymbol{\eta}_{f\to\theta} + \boldsymbol{\eta}_{\theta\to f}.$$

Figure 2. Diagram of the Inverse G-Wishart prior fragment.

5. The Inverse G-Wishart prior fragment

The Inverse G-Wishart prior fragment corresponds to the following prior imposition on a $d \times d$ covariance matrix $\boldsymbol{\Theta}$:

$$\boldsymbol{\Theta} \sim \text{Inverse-G-Wishart}(G_{\boldsymbol{\Theta}}, \xi_{\boldsymbol{\Theta}}, \boldsymbol{\Lambda}_{\boldsymbol{\Theta}}),$$

for a d-node undirected graph $G_{\boldsymbol{\Theta}}$, scalar shape parameter $\xi_{\boldsymbol{\Theta}}$ and scale matrix $\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}$. The fragment's factor is

$$p(\boldsymbol{\Theta}) \propto |\boldsymbol{\Theta}|^{-(\xi_{\boldsymbol{\Theta}}+2)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}\boldsymbol{\Theta}^{-1})\right\} \times \mathbf{1}(\boldsymbol{\Theta} \text{ is symmetric and positive definite and } \boldsymbol{\Theta}^{-1} \text{ respects } G_{\boldsymbol{\Theta}}).$$

Figure 2 is a diagram of the fragment, which shows that its only factor to stochastic node message is

$$m_{p(\boldsymbol{\Theta})\to\boldsymbol{\Theta}}(\boldsymbol{\Theta}) \propto p(\boldsymbol{\Theta}),$$

which leads to

$$m_{p(\boldsymbol{\Theta})\to\boldsymbol{\Theta}}(\boldsymbol{\Theta}) = \exp\left\{\begin{bmatrix}\log|\boldsymbol{\Theta}| \\ \operatorname{vech}(\boldsymbol{\Theta}^{-1})\end{bmatrix}^{\top}\begin{bmatrix}-\frac{1}{2}(\xi_{\boldsymbol{\Theta}}+2) \\ -\frac{1}{2}\boldsymbol{D}_d^{\top}\operatorname{vec}(\boldsymbol{\Lambda}_{\boldsymbol{\Theta}})\end{bmatrix}\right\}.$$

Therefore, the natural parameter update is

$$\boldsymbol{\eta}_{p(\boldsymbol{\Theta})\to\boldsymbol{\Theta}} \longleftarrow \begin{bmatrix}-\frac{1}{2}(\xi_{\boldsymbol{\Theta}}+2) \\ -\frac{1}{2}\boldsymbol{D}_d^{\top}\operatorname{vec}(\boldsymbol{\Lambda}_{\boldsymbol{\Theta}})\end{bmatrix}.$$

Apart from passing the natural parameter vector out of the fragment, we should also pass the graph out of the fragment. This entails the update:

$$G_{p(\boldsymbol{\Theta})\to\boldsymbol{\Theta}} \longleftarrow G_{\boldsymbol{\Theta}}.$$

Algorithm 1 provides the inputs, updates and outputs for the Inverse G-Wishart prior fragment.

Algorithm 1 *The inputs, updates and outputs for the Inverse G-Wishart prior fragment.*

Hyperparameter Inputs: $G_{\boldsymbol{\Theta}}$, $\xi_{\boldsymbol{\Theta}}$, $\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}$.

Updates:

$$\boldsymbol{\eta}_{p(\Theta)\to\Theta} \longleftarrow \begin{bmatrix}-\frac{1}{2}(\xi_{\boldsymbol{\Theta}}+2) \\ -\frac{1}{2}\boldsymbol{D}_d^{\top}\operatorname{vec}(\boldsymbol{\Lambda}_{\boldsymbol{\Theta}})\end{bmatrix}; \; G_{p(\Theta)\to\Theta} \longleftarrow G_{\boldsymbol{\Theta}}.$$

Outputs: $G_{p(\Theta)\to\Theta}$, $\boldsymbol{\eta}_{p\Theta\to\Theta}$.

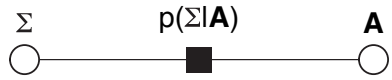

Figure 3. Diagram of the iterated Inverse G-Wishart fragment.

6. The Iterated Inverse G-Wishart fragment

The iterated Inverse G-Wishart fragment corresponds to the following specification involving a $d \times d$ covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} \,|\, \boldsymbol{A} \sim \text{Inverse-G-Wishart}(G, \xi, \boldsymbol{A}^{-1}),$$

where G is a d-node undirected graph such that $G \in \{G_{\text{full}}, G_{\text{diag}}\}$ and ξ is a particular deterministic value of the Inverse G-Wishart shape parameter according to Definition 2. Figure 3 is a diagram of this fragment, showing that it has a factor $p(\boldsymbol{\Sigma}|\boldsymbol{A})$ connected to two stochastic nodes $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$.

The factor of the iterated Inverse G-Wishart fragment is, as a function of both $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$,

$$p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A}) \propto \begin{cases} |\boldsymbol{A}|^{-(\xi-d+1)/2} |\boldsymbol{\Sigma}|^{-(\xi+2)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{-1})\right\}, & \text{if } G = G_{\text{full}}, \\ |\boldsymbol{A}|^{-\xi/2} |\boldsymbol{\Sigma}|^{-(\xi+2)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{-1})\right\}, & \text{if } G = G_{\text{diag}}. \end{cases}$$

As shown in Section S.2.1 of the web-supplement both of the factor to stochastic node messages of this fragment,

$$m_{p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A}) \to \boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) \quad \text{and} \quad m_{p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A}) \to \boldsymbol{A}}(\boldsymbol{A}),$$

are proportional to Inverse G-Wishart density functions with graph $G \in \{G_{\text{full}}, G_{\text{diag}}\}$. We assume the following conjugacy constraints:

All messages passed to $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$ from outside the fragment are proportional to Inverse G-Wishart density functions with graph $G \in \{G_{\text{full}}, G_{\text{diag}}\}$. The Inverse G-Wishart messages passed between $\boldsymbol{\Sigma}$ and $p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A})$ have the same graph. The Inverse G-Wishart messages passed between $\boldsymbol{A}$ and $p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A})$ have the same graph.

Under these constraints, and in view of for example, (7) of Wand (2017), the message passed from $\boldsymbol{\Sigma}$ to $p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A})$ has the form

$$m_{\boldsymbol{\Sigma} \to p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A})}(\boldsymbol{\Sigma}) = \exp\left\{ \begin{bmatrix} \log|\boldsymbol{\Sigma}| \\ \operatorname{vech}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix}^{\top} \boldsymbol{\eta}_{\boldsymbol{\Sigma} \to p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A})} \right\}$$

and the message passed from $\boldsymbol{A}$ to $p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A})$ has the form

$$m_{\boldsymbol{A} \to p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A})}(\boldsymbol{A}) = \exp\left\{ \begin{bmatrix} \log|\boldsymbol{A}| \\ \operatorname{vech}(\boldsymbol{A}^{-1}) \end{bmatrix}^{\top} \boldsymbol{\eta}_{\boldsymbol{A} \to p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A})} \right\}.$$

Algorithm 2 gives the full set of updates of the message natural parameter vectors and graphs for the iterated Inverse-G-Wishart fragment. The derivation of Algorithm 2 is given in Section S.2 of the web-supplement.

Algorithm 2 *The inputs, updates and outputs for the iterated Inverse G-Wishart fragment.*

Graph Input: $G \in \{G_{\text{full}}, G_{\text{diag}}\}$.
Shape Parameter Input: $\xi > 0$.
Message Graph Input: $G_{A \to p(\Sigma|A)} \in \{G_{\text{full}}, G_{\text{diag}}\}$.
Natural Parameter Inputs: $\boldsymbol{\eta}_{\Sigma \to p(\Sigma|A)}$, $\boldsymbol{\eta}_{p(\Sigma|A) \to \Sigma}$, $\boldsymbol{\eta}_{A \to p(\Sigma|A)}$, $\boldsymbol{\eta}_{p(\Sigma|A) \to A}$.
Updates:

$G_{p(\Sigma|A)\to\Sigma} \longleftarrow G$; $G_{p(\Sigma|A)\to A} \longleftarrow G_{A\to p(\Sigma|A)}$

$\boldsymbol{\eta}_{p(\Sigma|A)\leftrightarrow\Sigma} \longleftarrow \boldsymbol{\eta}_{p(\Sigma|A)\to\Sigma} + \boldsymbol{\eta}_{\Sigma\to p(\Sigma|A)}$

$\boldsymbol{\eta}_{p(\Sigma|A)\leftrightarrow A} \longleftarrow \boldsymbol{\eta}_{p(\Sigma|A)\to A} + \boldsymbol{\eta}_{A\to p(\Sigma|A)}$

If $G_{p(\Sigma|A)\to A} = G_{\text{full}}$, then $\omega_1 \longleftarrow (d+1)/2$

If $G_{p(\Sigma|A)\to A} = G_{\text{diag}}$, then $\omega_1 \longleftarrow 1$

$$\text{E}_q(\boldsymbol{A}^{-1}) \longleftarrow \left\{\left(\boldsymbol{\eta}_{p(\Sigma|A)\leftrightarrow A}\right)_1 + \omega_1\right\}\left\{\text{vec}^{-1}\left(\boldsymbol{D}_d^{+\top}\left(\boldsymbol{\eta}_{p(\Sigma|A)\leftrightarrow A}\right)_2\right)\right\}^{-1}$$

$$\text{If } G_{p(\Sigma|A)\to\Sigma} = G_{\text{diag}}, \text{ then } \text{E}_q(\boldsymbol{A}^{-1}) \longleftarrow \text{diag}\left\{\text{diagonal}\left(\text{E}_{\text{q}}(\boldsymbol{A}^{-1})\right)\right\}$$

$$\boldsymbol{\eta}_{p(\Sigma|A)\to\Sigma} \longleftarrow \begin{bmatrix} -\frac{1}{2}(\xi+2) \\ -\frac{1}{2}\boldsymbol{D}_d^{\top}\,\text{vec}\left(\text{E}_q(\boldsymbol{A}^{-1})\right)\end{bmatrix}$$

If $G_{p(\Sigma|A)\to\Sigma} = G_{\text{full}}$, then $\omega_2 \longleftarrow (d+1)/2$

If $G_{p(\Sigma|A)\to\Sigma} = G_{\text{diag}}$, then $\omega_2 \longleftarrow 1$

$$\text{E}_q(\boldsymbol{\Sigma}^{-1}) \longleftarrow \left\{\left(\boldsymbol{\eta}_{p(\Sigma|A)\leftrightarrow\Sigma}\right)_1 + \omega_2\right\}\left\{\text{vec}^{-1}\left(\boldsymbol{D}_d^{+\top}\left(\boldsymbol{\eta}_{p(\Sigma|A)\leftrightarrow\Sigma}\right)_2\right)\right\}^{-1}$$

$$\text{If } G_{p(\Sigma|A)\to A} = G_{\text{diag}}, \text{ then } \text{E}_q(\boldsymbol{\Sigma}^{-1}) \longleftarrow \text{diag}\left\{\text{diagonal}\left(\text{E}_q(\boldsymbol{\Sigma}^{-1})\right)\right\}$$

$$\boldsymbol{\eta}_{p(\Sigma|A)\to A} \longleftarrow \begin{bmatrix} -(\xi+2-2\omega_2)/2 \\ -\frac{1}{2}\boldsymbol{D}_d^{\top}\text{vec}\left(\text{E}_q(\boldsymbol{\Sigma}^{-1})\right)\end{bmatrix}$$

Outputs: $G_{p(\Sigma|A)\to\Sigma}$, $G_{p(\Sigma|A)\to A}$, $\boldsymbol{\eta}_{p(\Sigma|A)\to\Sigma}$, $\boldsymbol{\eta}_{p(\Sigma|A)\to A}$.

6.1. Corrections to Section 4.1.3 of Wand (2017)

The iterated Inverse G-Wishart fragment was introduced in Wand (2017) Section 4.1.3) and it is one of the five fundamental fragments of semiparametric regression given in Table 1. However, there are some errors due to the author of Wand (2017) failing to recognise particular subtleties regarding the Inverse G-Wishart distribution, as discussed in Section 2.2. We now point out misleading or erroneous aspects in Wand (2017) Section 4.1.3).

Firstly, in Wand (2017) $\boldsymbol{\Theta}_1$ plays the role of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Theta}_2$ plays the role of $\boldsymbol{A}$. The dimension of $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ is denoted by d^{Θ}. The first displayed equation of Wand (2017 Section 4.1.3) is

$$\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2 \sim \text{Inverse-G-Wishart}(G, \kappa, \boldsymbol{\Theta}_2^{-1}), \tag{14}$$

for $\kappa > d^{\Theta} - 1$ but it is only in the $G = G_{\text{full}}$ case that such a statement is reasonable for general $d^{\Theta} \in \mathbb{N}$. When $G = G_{\text{full}}$ then $\kappa = \xi - d^{\Theta} + 1$ according to the notation used in the current article. Therefore, (14) involves a different parameterisation to that used throughout this article. Therefore, our first correction is to replace the first displayed equation of Wand (2017 Section 4.1.3) by:

$$\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2 \sim \text{Inverse-G-Wishart}(G, \xi, \boldsymbol{\Theta}_2^{-1}),$$

Table 1. Specifications of inputs of Algorithms 1 and 2 for several variance, standard deviation and covariance matrix prior impositions.

Prior specification	Algorithm 1			Algorithm 2		
	$G_{\boldsymbol{\Theta}}$	$\xi_{\boldsymbol{\Theta}}$	$\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}$	ξ	G	$G_{\boldsymbol{A}\to p(\boldsymbol{\Sigma}\mid\boldsymbol{A})}$
$\sigma^2 \sim$ Inverse-$\chi^2(\delta_{\sigma^2}, \lambda_{\sigma^2})$	G_{full}	δ_{σ^2}	λ_{σ^2}	N.A.	N.A.	N.A.
$\sigma^2 \sim$ Inv:-Gamma$(\alpha_{\sigma^2}, \beta_{\sigma^2})$	G_{full}	$2\alpha_{\sigma^2}$	$2\beta_{\sigma^2}$	N.A.	N.A.	N.A.
$\boldsymbol{\Sigma} \sim$ Inv:-Wishart$(\kappa_{\boldsymbol{\Sigma}}, \boldsymbol{\Lambda}_{\boldsymbol{\Sigma}})$	G_{full}	$\kappa_{\boldsymbol{\Sigma}} + d - 1$	$\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}$	N.A.	N.A.	N.A.
$\sigma \sim$ Half-$t(s_\sigma, \nu_\sigma)$	G_{diag}	1	$(\nu_\sigma s_\sigma^2)^{-1}$	ν_σ	G_{full}	G_{diag}
$\sigma \sim$ Half-Cauchy(s_σ)	G_{diag}	1	$(s_\sigma^2)^{-1}$	1	G_{full}	G_{diag}
$\boldsymbol{\Sigma} \sim$ Huang–Wand $(s_{\boldsymbol{\Sigma},1}, \ldots, s_{\boldsymbol{\Sigma}d})$	G_{diag}	1	$\left\{2\,\text{diag}(s^2_{\boldsymbol{\Sigma},1}, \ldots, s^2_{\boldsymbol{\Sigma}d})\right\}^{-1}$	$2d$	G_{full}	G_{diag}
$\boldsymbol{\Sigma} \sim$ Matrix-F $(\nu_{\text{MP}}, \delta_{\text{MP}}, \boldsymbol{B}_{\text{MP}})$	G_{full}	$\nu_{\text{MP}} + d - 1$	$\boldsymbol{B}_{\text{MP}}^{-1}$	$\delta_{\text{MP}} + 2d - 2$	G_{full}	G_{full}

Note: The abbreviation N.A. stands for not applicable since Algorithm 2 is not needed for the first three prior impositions.

where $\xi > 0$ if $G = G_{\text{diag}}$ and $\xi > 2d^{\Theta} - 2$ if $G = G_{\text{full}}$.

The following sentence in Wand (2017) Section 4.1.3): 'The fragment factor is of the form

$$p(\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2) \propto |\boldsymbol{\Theta}_2|^{-\kappa/2}|\boldsymbol{\Theta}_1|^{-(\kappa+d^{\Theta}+1)/2} \exp\left\{-\frac{1}{2}\,\text{tr}(\boldsymbol{\Theta}_1^{-1}\boldsymbol{\Theta}_2^{-1})\right\},$$

should instead be 'The fragment factor is of the form

$$p(\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2) \propto \begin{cases} |\boldsymbol{\Theta}_2|^{-(\xi-d^{\Theta}+1)/2}|\boldsymbol{\Theta}_1|^{-(\xi+2)/2} \exp\left\{-\frac{1}{2}\,\text{tr}(\boldsymbol{\Theta}_1^{-1}\boldsymbol{\Theta}_2^{-1})\right\} & \text{if } G = G_{\text{full}}, \\ |\boldsymbol{\Theta}_2|^{-\xi/2}|\boldsymbol{\Theta}_1|^{-(\xi+2)/2} \exp\left\{-\frac{1}{2}\,\text{tr}(\boldsymbol{\Theta}_1^{-1}\boldsymbol{\Theta}_2^{-1})\right\} & \text{if } G = G_{\text{diag}}. \end{cases}'$$

In (31) of Wand (2017), the first entry of the vector on the right-hand side of the $\longleftarrow$ should be

$$-(\xi+2)/2 \quad \text{rather than} \quad -(\kappa+d^{\Theta}+1)/2.$$

To match the correct parameterisation of the Inverse G-Wishart distribution, as used in the current article, (32) of Wand (2017) should be

$$\text{'E}(\boldsymbol{X}^{-1}) \quad \text{where} \quad \boldsymbol{X} \sim \text{Inverse-G-Wishart}(G, \xi, \boldsymbol{\Lambda})\text{'}.$$

The equation in Wand (2017) Section 4.1.3):

$$\text{'}\boldsymbol{\eta}_{p(\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2)\to\boldsymbol{\Theta}_2} \longleftarrow \begin{bmatrix} -\kappa/2 \\ -\frac{1}{2}\,\text{vec}\left(\text{E}_{p(\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2)\to\boldsymbol{\Theta}_2}(\boldsymbol{\Theta}_1^{-1})\right) \end{bmatrix},$$

should be replaced by

$$\text{'}\boldsymbol{\eta}_{p(\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2)\to\boldsymbol{\Theta}_2} \longleftarrow \begin{bmatrix} -(\xi+2-2\omega_2)/2 \\ -\frac{1}{2}\,\text{vec}\left(\text{E}_{p(\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2)\to\boldsymbol{\Theta}_2}(\boldsymbol{\Theta}_1^{-1})\right) \end{bmatrix},$$

where ω_2 depends on the graph of the Inverse G-Wishart distribution corresponding to $\text{E}_{p(\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2)\to\boldsymbol{\Theta}_2}$. If the graph is G_{full} then $\omega_2 = (d^{\Theta}+1)/2$ and if the graph is G_{diag} then $\omega_2 = 1$.'

Lastly the iterated Inverse G-Wishart fragment natural parameter updates given by (36) and (37) of Wand (2017) are affected by the oversights described in the preceding paragraphs. They should be replaced by the updates given in Algorithm 2 with $\boldsymbol{\Theta}_1 = \boldsymbol{\Sigma}$ and $\boldsymbol{\Theta}_2 = \boldsymbol{A}$.

7. Use of the fragments for covariance matrix prior specification

The underlying rationale for the Inverse G-Wishart prior and iterated Inverse G-Wishart fragments is their ability to facilitate the specification of a wide range of covariance matrix priors within the variational message passing framework. In the $d = 1$ special case, covariance matrix parameters reduce to variance parameters and their square roots are standard deviation parameters. In this section, we spell out how the fragments, and their natural parameter updates in Algorithms 1 and 2, can be used for prior specification in important special cases.

7.1. Imposing an Inverse Chi-Squared prior on a variance parameter

Let σ^2 be a variance parameter and consider the prior imposition

$$\sigma^2 \sim \text{Inverse-}\chi^2(\delta_{\sigma^2}, \lambda_{\sigma^2}),$$

for hyperparameters $\delta_{\sigma^2}, \lambda_{\sigma^2} > 0$, within a variational message passing scheme. Then Algorithm 1 should be called with inputs set to:

$$G_{\boldsymbol{\Theta}} = G_{\text{full}}, \quad \xi_{\boldsymbol{\Theta}} = \delta_{\sigma^2}, \quad \boldsymbol{\Lambda}_{\boldsymbol{\Theta}} = \lambda_{\sigma^2}.$$

7.2. Imposing an Inverse Gamma prior on a variance parameter

Let σ^2 be a variance parameter and consider the prior imposition

$$\sigma^2 \sim \text{Inverse-Gamma}(\alpha_{\sigma^2}, \beta_{\sigma^2}), \tag{15}$$

for hyperparameters $\alpha_{\sigma^2}, \beta_{\sigma^2} > 0$. The density function corresponding to (15) is

$$p(\sigma^2; \alpha_{\sigma^2}, \beta_{\sigma^2}) \propto (\sigma^2)^{-\alpha_{\sigma^2}-1} \exp\{-\beta_{\sigma^2}/(\sigma^2)\}\mathbf{1}(\sigma^2 > 0).$$

Note that the Inverse Chi-Squared and Inverse Gamma distributions are simple reparameterisations of each other since

$$x \sim \text{Inverse-}\chi^2(\delta, \lambda) \quad \text{if and only if} \quad x \sim \text{Inverse-Gamma}\left(\frac{1}{2}\,\delta, \frac{1}{2}\,\lambda\right).$$

To achieve (15) Algorithm 1 should be called with inputs set to:

$$G_{\boldsymbol{\Theta}} = G_{\text{full}}, \quad \xi_{\boldsymbol{\Theta}} = 2\alpha_{\sigma^2}, \quad \boldsymbol{\Lambda}_{\boldsymbol{\Theta}} = 2\beta_{\sigma^2}.$$

7.3. Imposing an Inverse Wishart prior on a covariance matrix parameter

A random matrix $\boldsymbol{X}$ is defined to have an Inverse Wishart distribution with shape parameter κ and scale matrix $\boldsymbol{\Sigma}$, written $\boldsymbol{X} \sim \text{Inverse-Wishart}(\kappa, \boldsymbol{\Lambda})$, if and only if the density function of $\boldsymbol{X}$ is

$$p(\boldsymbol{X}) = \frac{|\boldsymbol{\Lambda}|^{\kappa/2}}{2^{\kappa d/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{\kappa+1-j}{2}\right)}|\boldsymbol{X}|^{-(\kappa+d+1)/2}\exp\left\{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{X}^{-1})\right\} \times \mathbf{1}(\boldsymbol{X} \text{ a symmetric and positive definite } d\times d \text{ matrix}). \tag{16}$$

Note that this is the common parameterisation of the Inverse Wishart distribution (e.g. Gelman *et al.* 2014, Table A.1). Crucially, (16) uses a *different* shape parametrisation from that used for the Inverse G-Wishart distribution in Definition 2 when $G = G_{\text{full}}$ with the relationship between the two shape parameters given by $\kappa = \xi - d + 1$. Even though the more general Inverse G-Wishart family is important for the internal workings of variational message passing, the ordinary Inverse Wishart distribution, with the parameterisation as given in (16), is more common when imposing a prior on a covariance matrix.

Let $\boldsymbol{\Sigma}$ be a $d \times d$ matrix and consider the prior imposition

$$\boldsymbol{\Sigma} \sim \text{Inverse-Wishart}(\kappa_{\boldsymbol{\Sigma}}, \boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}) \tag{17}$$

for hyperparameters $\kappa_{\boldsymbol{\Sigma}}, \boldsymbol{\Lambda}_{\boldsymbol{\Sigma}} > 0$, within a variational message passing scheme. Then Algorithm 1 should be called with inputs set to:

$$G_{\boldsymbol{\Theta}} = G_{\text{full}}, \quad \xi_{\boldsymbol{\Theta}} = \kappa_{\boldsymbol{\Sigma}} + d - 1, \quad \boldsymbol{\Lambda}_{\boldsymbol{\Theta}} = \boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}.$$

7.4. Imposing a Half-*t* prior on a standard deviation parameter

Consider the prior imposition

$$\sigma \sim \text{Half-}t(s_\sigma, \nu_\sigma), \tag{18}$$

for a scale parameter $s_\sigma > 0$ and a degrees of freedom parameter $\nu_\sigma > 0$. The density function corresponding to (18) is such that $p(\sigma) \propto \{1 + (\sigma/s_\sigma)^2/\nu_\sigma\}^{-(\nu_\sigma+1)/2}\mathbf{1}(\sigma > 0)$. This is equivalent to

$$\sigma^2|a \sim \text{Inverse-}\chi^2(\nu_\sigma, 1/a) \quad \text{and} \quad a \sim \text{Inverse-}\chi^2(1, 1/s_\sigma^2). \tag{19}$$

Since $d = 1$, the graphs G_{full} and G_{diag} are the same – a single node graph. Treating σ^2 and a as 1×1 matrices we can re-write (19) as

$$\sigma^2|a \sim \text{Inverse-G-Wishart}(G_{\text{full}}, \nu_\sigma, a^{-1}) \quad \text{and} \quad a \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, (\nu_\sigma s_\sigma^2)^{-1})$$

(e.g. Armagan, Dunson & Clyde 2011). The specification

$$a \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, (\nu_\sigma s_\sigma^2)^{-1})$$

involves calling Algorithm 1 with

$$G_{\boldsymbol{\Theta}} = G_{\text{diag}}, \quad \xi_{\boldsymbol{\Theta}} = 1 \quad \text{and} \quad \boldsymbol{\Lambda}_{\boldsymbol{\Theta}} = (\nu_\sigma s_\sigma^2)^{-1}.$$

The output is the single node graph $G_{p(\boldsymbol{\Theta})\to\boldsymbol{\Theta}}$ and the 2×1 natural parameter vector

$$\boldsymbol{\eta}_{p(\boldsymbol{\Theta})\to\boldsymbol{\Theta}} = \boldsymbol{\eta}_{p(a)\to a}.$$

The specification

$$\sigma^2|a \sim \text{Inverse-G-Wishart}(G_{\text{full}}, \nu_\sigma, a^{-1})$$

implies that Algorithm 2 is called with graph input $G = G_{\text{full}}$, shape parameter input $\xi = \nu_\sigma$ and message parameter inputs

$$\boldsymbol{\eta}_{p(\boldsymbol{\Sigma}\,|\,\boldsymbol{A})\to\boldsymbol{\Sigma}} = \boldsymbol{\eta}_{p(\sigma^2|a)\to\sigma^2}, \quad \boldsymbol{\eta}_{\boldsymbol{\Sigma}\to p(\boldsymbol{\Sigma}\,|\,\boldsymbol{A})} = \boldsymbol{\eta}_{\sigma^2\to p(\sigma^2|a)},$$

and

$$G_{p(\boldsymbol{\Sigma}\,|\,\boldsymbol{A})\to\boldsymbol{A}} = G_{\text{diag}}, \quad \boldsymbol{\eta}_{p(\boldsymbol{\Sigma}\,|\,\boldsymbol{A})\to\boldsymbol{A}} = \boldsymbol{\eta}_{p(\sigma^2|a)\to a} \quad \text{and} \quad \boldsymbol{\eta}_{\boldsymbol{A}\to p(\boldsymbol{\Sigma}\,|\,\boldsymbol{A})} = \boldsymbol{\eta}_{a\to p(\sigma^2|a)}.$$

Note that in this $d = 1$ special case G_{full} and G_{diag} are both the single node graph.

7.4.1. The Half-Cauchy special case

The special case of

$$\sigma \sim \text{Half-Cauchy}(s_\sigma) \tag{20}$$

corresponds to $\nu_\sigma = 1$. The density function corresponding to (20) is such that $p(\sigma) \propto \{1 + (\sigma/s_\sigma)^2\}^{-1}\mathbf{1}(\sigma > 0)$. Therefore, one should set $\xi = 1$ in the call to Algorithm 2.

7.5. Imposing a Huang–Wand prior on a covariance matrix

To impose the Huang–Wand prior

$$\boldsymbol{\Sigma} \sim \text{Huang–Wand}(s_{\boldsymbol{\Sigma},1},\ldots,s_{\boldsymbol{\Sigma},d})$$

in a variational message passing framework we should have the inputs to Algorithm 1 being as follows:

$$G_{\boldsymbol{\Theta}} = G_{\text{diag}}, \quad \xi_{\boldsymbol{\Theta}} = 1 \quad \text{and} \quad \boldsymbol{\Lambda}_{\boldsymbol{\Theta}} = \left\{2\,\text{diag}(s^2_{\boldsymbol{\Sigma},1},\ldots,s^2_{\boldsymbol{\Sigma},d})\right\}^{-1}.$$

The graph parameter input to Algorithm 2 should be $G = G_{\text{full}}$ and the shape parameter input should be $\xi = 2d$.

7.6. Imposing a Matrix-*F* prior on a covariance matrix

To impose the Matrix-F prior

$$\boldsymbol{\Sigma} \sim F(\nu_{\text{MP}}, \delta_{\text{MP}}, \boldsymbol{B}_{\text{MP}})$$

in a variational message passing framework the inputs to Algorithm 1 should be as follows:

$$G_{\boldsymbol{\Theta}} = G_{\text{full}}, \quad \xi_{\boldsymbol{\Theta}} = \nu_{\text{MP}} + d - 1 \quad \text{and} \quad \boldsymbol{\Lambda}_{\boldsymbol{\Theta}} = \boldsymbol{B}^{-1}_{\text{MP}}.$$

The graph parameter input to Algorithm 2 should be $G = G_{\text{full}}$ and the shape parameter input should be $\xi = \delta_{\text{MP}} + 2d - 2$.

7.7. Tabular summary of fragment-based prior specification

Table 1 summarizes the results of this section and is a crucial reference for placing priors of covariance matrix, variance and standard deviation parameters in variational message passing schemes that make use of Algorithms 1 and 2.

8. Illustrative example

We illustrate the use of Algorithms 1 and 2 for the case of Bayesian linear mixed models with t distribution responses. Such t-based models impose a form of robustness in situations where the responses are susceptible to having outlying values (e.g. Lange, Little & Taylor 1989). The notation $y \sim \text{t}(\mu, \sigma, \nu)$ indicates that the random variable y has a t distribution with location parameter μ, scale parameter $\sigma > 0$ and degrees of freedom parameter $\nu > 0$. The corresponding density function of y is

$$p(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\pi\nu}\Gamma(\nu/2)[1 + \{(y-\mu)/\sigma\}^2/\nu]^{\frac{\nu+1}{2}}}.$$

Now suppose that the response data consists of repeated measures within each of m groups. Let

$$y_{ij} = \text{the } j\text{th response for the } i\text{th group}, \quad 1 \leqslant j \leqslant n_i, \quad 1 \leqslant i \leqslant m,$$

and then let $\boldsymbol{y}_i$, $1 \leqslant i \leqslant m$, be the $n_i \times 1$ vectors containing y_{ij} data for the ith group. For each $1 \leqslant i \leqslant m$, let $\boldsymbol{X}_i$ be $n_i \times p$ design matrices corresponding to the fixed effects and $\boldsymbol{Z}_i$ be $n_i \times q$ design matrices corresponding to the random effects. Next put

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix} \quad \text{and} \quad \boldsymbol{Z} = \underset{1 \leqslant i \leqslant m}{\text{blockdiag}}(\boldsymbol{Z}_i), \tag{21}$$

and define $N = n_1 + \ldots + n_m$ to be the number of rows in each of $\boldsymbol{y}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$. Let y_ℓ be the ℓth entry of $\boldsymbol{y}$, $1 \leqslant \ell \leqslant N$. The family of Bayesian t response linear mixed models that we consider is

$$\begin{aligned} &y_\ell | \boldsymbol{\beta}, \boldsymbol{u}, \sigma \overset{\text{ind.}}{\sim} \text{t}\big((\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})_\ell, \sigma, \nu\big), \quad 1 \leqslant \ell \leqslant N, \quad \boldsymbol{u} | \boldsymbol{\Sigma} \sim \text{N}(\boldsymbol{0}, \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}), \\ &\boldsymbol{\beta} \sim \text{N}(0, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{I}), \quad \sigma \sim \text{Half-Cauchy}(s_\sigma), \quad \frac{1}{2}\nu \sim \text{Moon-Rock}(0, \lambda_\nu), \\ &\boldsymbol{\Sigma} \sim \text{Huang–Wand}\big(s_{\boldsymbol{\Sigma},1}, \ldots, s_{\boldsymbol{\Sigma},q}\big), \end{aligned} \tag{22}$$

for hyperparameters $\sigma_{\boldsymbol{\beta}}, s_\sigma, \lambda_\nu, s_{\boldsymbol{\Sigma},1}, \ldots, s_{\boldsymbol{\Sigma},q} > 0$.

As explained in McLean & Wand (2019), the Moon Rock family of distributions is conjugate for the parameter $\frac{1}{2}\nu$, with the notation $x \sim \text{Moon-Rock}(\alpha, \beta)$ indicating that the corresponding density function satisfies $p(x) \propto \{x^x/\Gamma(x)\}^\alpha \exp(-\beta x)\mathbf{1}(x > 0)$. In the variational message passing treatment of the degrees of freedom parameter it is simpler to work with

$$\upsilon = \frac{1}{2}\nu, \quad \text{so that} \quad \upsilon \sim \text{Moon-Rock}(0, \lambda_\nu).$$

After the approximate posterior density function of υ is obtained via variational message passing, it is trivial to then obtain the same for ν. Hence, we work with υ, rather than ν, in the upcoming description of variational message passing-based fitting and inference for (22).

Next note that

$$\boldsymbol{y}_\ell | \boldsymbol{\beta}, \boldsymbol{u}, \sigma \overset{\text{ind.}}{\sim} \text{t}\Big((\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})_\ell, \sigma, 2\upsilon\Big), \quad 1 \leqslant \ell \leqslant N,$$

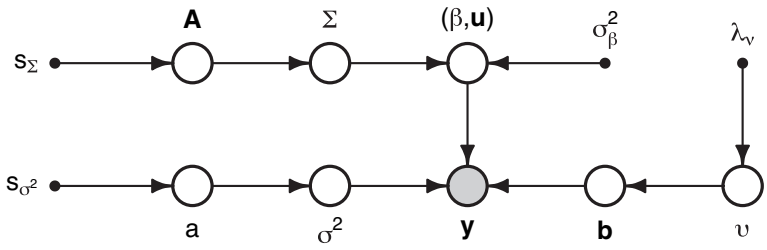

Figure 4. Directed acyclic graph corresponding to the t response linear mixed model (22) with auxiliary variable representations (23)–(25). The shaded circle corresponds to the observed data. The unshaded circles correspond to model parameters and auxiliary variables. The small solid circles correspond to hyperparameters.

is equivalent to

$$\boldsymbol{y}_\ell|\boldsymbol{\beta},\boldsymbol{u},\sigma^2,b_\ell \sim \mathrm{N}\big((\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})_\ell, b_\ell\sigma^2\big), \quad b_\ell|\upsilon \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2\left(2\upsilon,2\upsilon\right), \tag{23}$$

$\sigma \sim \text{Half-Cauchy}(s_\sigma)$ is equivalent to

$$\sigma^2|a \sim \text{Inverse-}\chi^2(1,1/a) \quad \text{and} \quad a \sim \text{Inverse-}\chi^2(1,1/s_\sigma^2) \tag{24}$$

and $\boldsymbol{\Sigma} \sim \text{Huang–Wand}\big(s_{\boldsymbol{\Sigma},1},\ldots,s_{\boldsymbol{\Sigma},q}\big)$ is equivalent to

$$\begin{aligned} \boldsymbol{\Sigma}\,|\,\boldsymbol{A} &\sim \text{Inverse-G-Wishart}(G_{\text{full}},2q,\boldsymbol{A}^{-1}),\\ \boldsymbol{A} &\sim \text{Inverse-G-Wishart}\Big(G_{\text{diag}},1,\big\{2\,\text{diag}(s^2_{\boldsymbol{\Sigma},1},\ldots,s^2_{\boldsymbol{\Sigma},q})\big\}^{-1}\Big). \end{aligned} \tag{25}$$

Substitution of (23), (24) and (25) into (22) leads to the hierarchical Bayesian model depicted as a directed acyclic graph in Figure 4 with $\boldsymbol{b} = (b_1,\ldots,b_N)$. The unshaded circles in Figure 4 correspond to model parameters and auxiliary variables and will be referred to as *hidden nodes*.

Consider the following mean-field approximation of the joint posterior of the hidden nodes in Figure 4

$$p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\upsilon,\boldsymbol{\Sigma},a,\boldsymbol{A},\boldsymbol{b}|\boldsymbol{y}) \approx q(\boldsymbol{\beta},\boldsymbol{u},a,\boldsymbol{A},\boldsymbol{b})q(\sigma^2,\boldsymbol{\Sigma},\upsilon), \tag{26}$$

where q denotes the approximate posterior density functions of the relevant parameters. Application of induced factor results (e.g. Bishop 2006, Section 10.2.5) leads to the additional factorisations

$$q(\boldsymbol{\beta},\boldsymbol{u},a,\boldsymbol{A},\boldsymbol{b}) = q(\boldsymbol{\beta},\boldsymbol{u})q(a)q(\boldsymbol{A})\prod_{\ell=1}^{N} q(b_\ell) \quad \text{and} \quad q(\sigma^2,\boldsymbol{\Sigma},\upsilon) = q(\sigma^2)q(\boldsymbol{\Sigma})q(\upsilon),$$

and so the restriction given in (26) is equivalent to

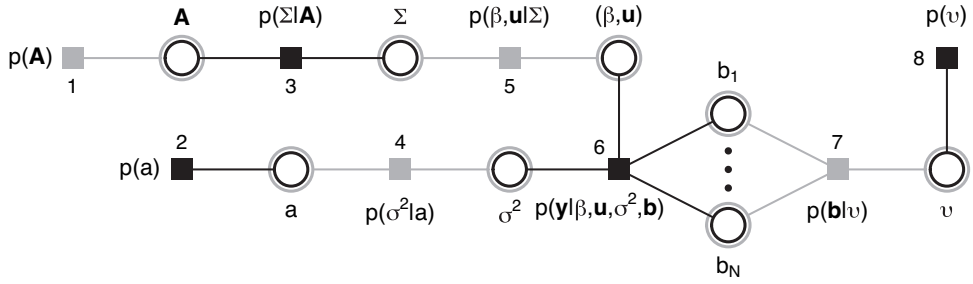

Figure 5. Factor graph corresponding to the t response linear mixed model (22) with auxiliary variable representations (23)–(25). The circular nodes correspond to stochastic nodes in the q-density factorisation in (27). The rectangular nodes correspond to the factors on the right-hand side of (28). The fragments are numbered 1 to 8 according to appearance from left to right. Shading is used to show the distinction between adjacent fragments.

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, \upsilon, \boldsymbol{\Sigma}, a, \boldsymbol{A}, \boldsymbol{b}|\boldsymbol{y}) \approx q(\boldsymbol{\beta}, \boldsymbol{u}) q(\sigma^2) q(\upsilon) q(\boldsymbol{\Sigma}) q(a) q(\boldsymbol{A}) \prod_{\ell=1}^{N} q(b_\ell). \tag{27}$$

Figure 5 is a factor graph representation of the joint density function of all random variables and vectors, or stochastic nodes, in Figure 4 hierarchical model, with unshaded circles for each stochastic node according to the q-density factorisation given in (27) and filled-in rectangles corresponding to factors on the right-hand side of

$$\begin{aligned} &p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, \upsilon, \boldsymbol{\Sigma}, a, \boldsymbol{A}, \boldsymbol{b}) \\ &\quad = p(\boldsymbol{A}) p(a) p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A}) p(\sigma^2|a) p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, \boldsymbol{b}) p(\boldsymbol{b}|\upsilon) p(\upsilon). \end{aligned} \tag{28}$$

Edges join each factor to a stochastic node that appears in the factor. To aid upcoming discussion, the fragments are numbered 1 to 8 according to appearance from left to right. Recall that a fragment is a sub-graph consisting of a factor and all of its neighbouring nodes. Figure 5 uses shading to show the distinction between adjacent fragments.

Note that (e.g. Minka 2005; Wand 2017) the variational message passing iteration loop has the following generic steps:

1. Choose a factor.
2. Update the parameter vectors of the messages passed from the factor's neighbouring stochastic nodes to the factor.
3. Update the parameter vectors of the messages passed from the factor to its neighbouring stochastic nodes.

Step 2. is very simple and has generic form given by, for example, (7) of Wand (2017). In the Figure 5 factor graph an example of Step 2. is:

$$\begin{aligned} &\text{the message passed from } \boldsymbol{\Sigma} \text{ to } p(\boldsymbol{\beta}, \boldsymbol{u} \,|\, \boldsymbol{\Sigma}) \\ &\quad = \text{the message passed from } p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A}) \text{ to } \boldsymbol{\Sigma} \text{ in the previous iteration.} \end{aligned} \tag{29}$$

In terms of natural parameter vector updates, (29) corresponds to:

$$\boldsymbol{\eta}_{\boldsymbol{\Sigma} \to p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A})} \longleftarrow \boldsymbol{\eta}_{p(\boldsymbol{\Sigma} \,|\, \boldsymbol{A}) \to \boldsymbol{\Sigma}}.$$

Most of the other stochastic node to factor updates in Figure 5 have an analogous form. The exception are the messages passed within fragments 6 and 7, which require use of the slightly more complicated form as given by, for example, (7) of Wand (2017).

It remains to discuss Step 3, corresponding to the factor to stochastic node updates:

- Fragments 1 and 2 are Inverse G-Wishart prior fragments and the factor to stochastic node parameter vector updates are performed according to Algorithm 1. In view of Table 1, the graph and shape hyperparameter inputs are $G_{\boldsymbol{\Theta}} = G_{\text{diag}}$ and $\xi_{\boldsymbol{\Theta}} = 1$. For fragment 1 the rate hyperparameter is $\boldsymbol{\Lambda}_{\boldsymbol{\Theta}} = \{2\,\text{diag}\left(s^2_{\boldsymbol{\Sigma},1},\ldots,s^2_{\boldsymbol{\Sigma},q}\right)\}^{-1}$. For fragment 2 the rate hyperparameter is $\boldsymbol{\Lambda}_{\boldsymbol{\Theta}} = (s^2_\sigma)^{-1}$.
- Fragments 3 and 4 are iterated Inverse G-Wishart prior fragments and the factor to stochastic node parameter vector updates are performed according to Algorithm 2. As shown in Table 1, the graph inputs should be

$$G_{\boldsymbol{A}\to p(\boldsymbol{\Sigma}\,|\,\boldsymbol{A})} = G_{\text{diag}}, \quad G_{a\to p(\sigma^2|a)} = G_{\text{diag}},$$
$$G_{\boldsymbol{\Sigma}\to p(\boldsymbol{\Sigma}\,|\,\boldsymbol{A})} = G_{\text{full}}, \quad \text{and } G_{\sigma^2\to p(\sigma^2|a)} = G_{\text{full}}.$$

  The first two of these are imposed by the messages passed from fragments 1 and 2. For fragment 3, the shape parameter input is $\xi = 2q$. For fragment 4, the shape parameter input is $\xi = 1$.
- Fragment 5 is the Gaussian penalisation fragment described in Section 4.1.4 of Wand (2017) with, in the notation given there, $L = 1$, $\boldsymbol{\mu}_{\boldsymbol{\theta}_0} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0} = \sigma^2_{\boldsymbol{\beta}}\boldsymbol{I}$.
- Fragments 6 and 7 correspond to the t likelihood fragment. Its natural parameter updates are provided by McLean & Wand (2019) Algorithm 2).
- Fragment 8 corresponds to the imposition of a Moon Rock prior distribution on a shape parameter. This is a very simple fragment for which the only inputs are the Moon Rock prior specification hyperparameters and the output is the natural parameter vector of the Moon Rock prior density function. Since this fragment is not listed as an algorithm in this article or elsewhere, we provide further details in the paragraph after the next one.

For Fragments 5, 6 and 7 simple conversions between two different versions of natural parameter vectors need to be made. Section S.1 of the web-supplement explains these conversions.

The most general Moon Rock prior specification for a generic parameter θ is

$$\theta \sim \text{Moon-Rock}(\alpha_\theta, \beta_\theta).$$

This corresponds to the prior density function having exponential family form

$$p(\theta) \propto \exp\left\{ \begin{bmatrix} \theta\log(\theta) - \log\Gamma(\theta) \\ \theta \end{bmatrix}^\top \begin{bmatrix} \alpha_\theta \\ -\beta_\theta \end{bmatrix} \right\}.$$

The inputs of the Moon Rock prior fragment are $\alpha_\theta \geqslant 0$ and $\beta_\theta > 0$ and the output is the natural parameter vector

$$\boldsymbol{\eta}_{p(\theta)\to\theta} \longleftarrow \begin{bmatrix} \alpha_\theta \\ -\beta_\theta \end{bmatrix}.$$

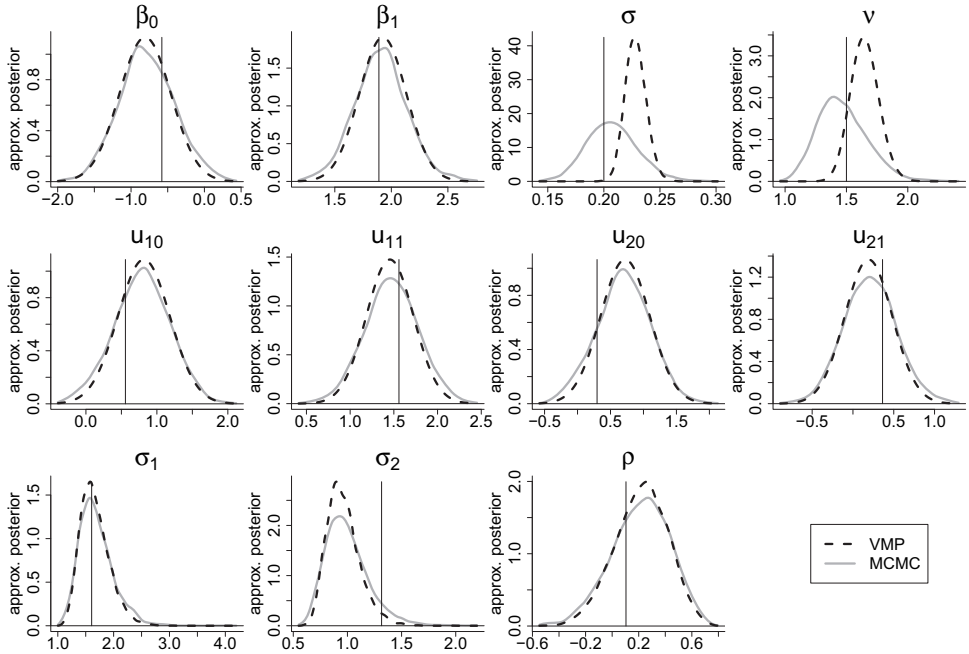

Figure 6. Approximate posterior density functions for the parameters in model (22) based on both variational message passing (VMP) and Markov chain Monte Carlo (MCMC) algorithms applied to data simulated according to the true values (30) and sample sizes, predictor values and hyperparameter values as described in the text. The vertical lines indicate true parameter values.

Since, for the t response mixed model illustrative example, we have the prior imposition $\upsilon \sim \text{Moon-Rock}(0, \lambda_v)$ we simply call the Moon Rock prior fragment with $(\alpha_\theta, \beta_\theta)$ set to $(0, \lambda_v)$.

To demonstrate variational message passing for fitting and inference for model (22), we simulated data according to the dimension values $p = q = 2$ and the true parameter values

$$\boldsymbol{\beta}_{\text{true}} = \begin{bmatrix} -0.58 \\ 1.89 \end{bmatrix}, \quad \sigma^2_{\text{true}} = 0.2, \quad \boldsymbol{\Sigma}_{\text{true}} = \begin{bmatrix} 2.58 & 0.22 \\ 0.22 & 1.73 \end{bmatrix} \quad \text{and} \quad \nu_{\text{true}} = 1.5. \tag{30}$$

The sample sizes were $m = 20$, with $n_i = 15$ observations per group, and the predictor data were generated from the Uniform distribution on the unit interval. The hyperparameter values were set at

$$\sigma_{\boldsymbol{\beta}} = s_\sigma = s_{\boldsymbol{\Sigma},1} = s_{\boldsymbol{\Sigma},2} = 10^5 \quad \text{and} \quad \lambda_v = 0.01.$$

We ran the variational message passing algorithm as described above until the relative change the variational parameters is below 10^{-10}. as well as Markov chain Monte Carlo via the R language (R Core Team 2021) package rstan (Stan Development Team 2020). For Markov chain Monte Carlo fitting, a warmup of size 1000 was used, followed by chains of size 5000 retained for inference.

Figure 6 compares the approximate posterior density functions based on both variational message passing (VMP) and Markov chain Monte Carlo (MCMC). The middle row performs

the comparison for the random intercept and slope parameters, u_{i0} and u_{i1}, for $i = 1, 2$. The parameters in the third row of Figure 6 are for the standard deviation and correlation parameters in the $\boldsymbol{\Sigma}$ matrix, according to the notation $(\boldsymbol{\Sigma})_{11} = \sigma_1$, $(\boldsymbol{\Sigma})_{22} = \sigma_2$ and $(\boldsymbol{\Sigma})_{12} = \sigma_1\sigma_2\rho$. For most of the stochastic nodes, the accuracy of variational message passing is seen to be very good. For σ and v, some under-approximation of the spread and locational shift is apparent. A likely root cause is the imposition of the product restriction $q(\sigma, v) = q(\sigma)q(v)$ even though these two parameters have a significant amount of posterior dependence.

We have prepared a bundle of R language code that carries out variational message passing for this illustrative example, including use of Algorithms 1 and 2 for the imposition of Half Cauchy and Huang–Wand priors. This code is part of the web-supplement for this article.

Lastly, we point out that this illustrative example does not involve matrix algebraic streamlining for random effects models. This relatively new area for variational message passing research, which streamlines calculations involving sparse matrix forms that arise in linear mixed models, is described in Nolan, Menictas & Wand (2020).

9. Closing remarks

Algorithms 1 and, especially, Algorithm 2 and their underpinnings are quite involved and dependent upon a careful study of particular special cases of the inverses of G-Wishart random matrices. The amount of detail provided by this article is tedious, but necessary, to ensure that the fragment updates based on a single distributional structure, the Inverse G-Wishart distribution with $G \in \{G_{\text{full}}, G_{\text{diag}}\}$, are correct. The good news is that these algorithms only need to be derived once. Their implementations, within a suite of computer programmes for carrying out variational message passing for models containing variance and covariance matrix parameters, can be isolated into subroutines which, once working as intended, do not have to be revisited ever again. Given the quintessence of variance and covariance parameters in throughout statistics and machine learning, Algorithms 1 and 2 are important and fundamental contributions to variational message passing.

Acknowledgements

We are grateful to two referees for their comments and suggestions. This research was supported by Australian Research Council Discovery Project DP140100441.

References

Armagan, A., Dunson, D.B. & Clyde, M. (2011). Generalized beta mixtures of Gaussians. In *Advances in Neural Information Processing Systems 24*, eds. J. Shawe-Taylor, R.S. Zamel, P. Bartlett, F. Pereira and K.Q. Weinberger, pp. 523–531. La Jolla, CA: The Neural Information Processing Systems (NIPS) Foundation.

Assaf, A.G., Li, G., Song, H. & Tsionas, M.G. (2019). Modeling and forecasting regional tourism demand using the Bayesian global vector autoregressive (BGVAR) model. *Journal of Travel Research* **58**, 383–397.

Atay-Kayis, A. & Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika* **92**, 317–335.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, eds. K.B. Laskey and H. Prade, pp. 21–30. San Francisco: Morgan Kaufmann.

BISHOP, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.
CHEN, W.Y. & WAND, M.P. (2020). Factor graph fragmentization of expectation propagation. *Journal of the Korean Statistical Society* **49**, 722–756.
CONTI, G., FRÜHWIRTH-SCHNATTER, S., HECKMAN, J.J. & PIATEK, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics* **183**, 31–57.
DAWID, A.P. & LAURITZEN, S.L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* **21**, 1272–1317.
GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1**, 515–534.
GELMAN, A., CARLIN, J.B., STERN, H.S., DUNSON, D.B., VEHTARI, A. & RUBIN, D.B. (2014). *Bayesian Data Analysis*, 3rd edn. Boca Raton, Florida: CRC Press.
GENTLE, J.E. (2007). *Matrix Algebra*. New York: Springer.
HAREZLAK, J., RUPPERT, D. & WAND, M.P. (2018). *Semiparametric Regression with R*. New York: Springer.
HUANG, A. & WAND, M.P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis* **8**, 439–452.
LANGE, K.L., LITTLE, R.J. & TAYLOR, J.M. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84**, 881–896.
LETAC, G. & MASSAM, H. (2007). Wishart distributions for decomposable graphs. *The Annals of Statistics* **35**, 1278–1323.
MAESTRINI, L. & WAND, M.P. (2018). Variational message passing for skew t regression. *Stat* **7**, e196, 1–11 .
MAGNUS, J.R. & NEUDECKER, H. (2019). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 3rd edn. Chichester U.K.: John Wiley & Sons.
MCCULLOCH, C.E., SEARLE, S.R. & NEUHAUS, J.M. (2008). *Generalized, Linear, and Mixed Models*, 2nd edn. New York: John Wiley & Sons.
MCLEAN, M. & WAND, M. (2019). Variational message passing for elaborate response regression models. *Bayesian Analysis* **14**, 371–398.
MINKA, T.P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 362–369. Burlington, Massachusetts: Morgan Kaufmann.
MINKA, T.P. (2005). Divergence measures and message passing. *Microsoft Research Technical Report Series* **MSR-TR-2005-173**, 1–17.
MUIRHEAD, R.J. (1982). *Aspects of Multivariate Statistical Theory*. New York: John Wiley & Sons.
MULDER, J. & PERICCHI, L.R. (2018). The matrix-F prior for estimating and testing covariance matrices. *Bayesian Analysis* **13**, 1193–1214.
NOLAN, T.H., MENICTAS, M. & WAND, M.P. (2020). Streamlined computing for variational inference with higher level random effects. *Journal of Machine Learning Research* **21**, 1–62.
NOLAN, T.H. & WAND, M.P. (2017). Accurate logistic variational message passing: algebraic and numerical details. *Stat* **6**, 102–112.
POLSON, N.G. & SCOTT, J.G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis* **7**, 887–902.
R CORE TEAM (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from URL: https://www.R-project.org/.
ROVERATO, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika* **87**, 99–112.
STAN DEVELOPMENT TEAM (2020). RStan: the R interface to Stan. Available from URL: http://mc-stan.org/. R package version 2.21.2.
UHLER, C., LENKOSKI, A. & RICHARDS, D. (2018). Exact formulas for the normalizing constants of Wishart distributions for graphical models. *The Annals of Statistics* **46**, 90–118.
WAND, M.P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *Journal of the American Statistical Association* **112**, 137–168.
WINN, J. & BISHOP, C.M. (2005). Variational message passing. *Journal of Machine Learning Research* **6**, 661–694.