Factor Graph Fragmentization of Expectation Propagation

By Wilson Y. Chen and Matt P. Wand

University of Technology Sydney

15th January, 2018

Abstract

Expectation propagation is a general approach to fast approximate inference for graphical models. The existing literature treats models separately when it comes to deriving and coding expectation propagation inference algorithms. This comes at the cost of similar, long-winded algebraic steps being repeated and slowing down algorithmic development. We demonstrate how *factor graph fragmentization* can overcome this impediment. This involves adoption of the message passing on a factor graph approach to expectation propagation and identification of factor graph sub-graphs, which we call *fragments*, that are common to wide classes of models. Key fragments and their corresponding messages are catalogued which means that their algebra does not need to be repeated. This allows compartmentalization of coding and efficient software development.

Keywords: Approximate Bayesian inference; Generalized linear mixed models; Graphical models; Kullback-Leibler projection; Message passing.

1 Introduction

Expectation propagation (e.g. Minka, 2005) is gaining popularity as a general approach to fitting and inference for large graphical models, including those that arise in statistical contexts such as Bayesian generalized linear mixed models (e.g. Gelman *et al.*, 2014; Kim & Wand, 2017). Compared with Markov chain Monte Carlo approaches, expectation propagation has the attractions of speed and parallelizability of the computing across multiple processors making it more amenable to high volume/velocity data applications. One price to be paid is inferential accuracy since expectation propagation uses product density simplifications of joint posterior density functions. Another is algebraic overhead: as demonstrated by Kim & Wand (2016) several pages of algebra are required to derive explicit programmable expectation propagation algorithms for even very simple Bayesian models. This article alleviates the latter cost. Using the notions of *message passing* and *factor graph fragments* we demonstrate the compartmentalization of expectation propagation algebra and coding. The resultant infrastructure and updating formulae lead to much more efficient expectation propagation fitting and inference and allows extension to arbitrarily large Bayesian models.

Expectation propagation and *mean field variational Bayes* are the two most common paradigms for obtaining fast approximate inference algorithms for graphical models (e.g. Bishop, 2006; Wainwright & Jordan, 2008; Murphy, 2012). Each is driven by minimum Kullback-Leibler divergence considerations. As explained in Minka (2005), they can both be expressed as message passing algorithms on factor graphs. The alternative appellation *variational message passing* is used for mean field variational Bayes when such an approach is used. The software platform Infer.NET (Minka *et al.*, 2014) uses both expectation propagation and variational message passing to perform fast approximate inference for graphical models. Recently Wand (2017) introduced factor graph fragmentization to streamline variational message passing for semiparametric regression analysis. Semiparametric regression (e.g. Ruppert *et al.*, 2009) is a big class of flexible regression models that includes generalized linear mixed models, generalized additive models and varying-coefficient models as special cases. Nolan & Wand (2017) and McLean & Wand (2018) built on Wand (2017) for more elaborate likelihood fragments.

The crux of this article is to show how the factor graph fragment idea also can be used to streamline expectation propagation. We focus on semiparametric regression models. However, the approach is quite general and applies to other graphical models for which expectation propagation is feasible. The fragment updating algorithms presented and derived here cover a wide range of semiparametric models and pave the way for future derivations of the same type.

Section 2 provides the background material needed for factor graph fragmentization of expectation propagation. This includes exponential family and Kullback-Leibler projection theory, as well as the notions of factor graphs and their fragment sub-graphs. The article's centerpiece is Section 3 in which several key fragments are identified and have message updates derived and catalogued. Such cataloguing implies that updates for a particular fragment never have to be derived again and only need to be implemented once in an expectation propagation software suite. An illustration involving generalized additive mixed model analysis of data from a longitudinal public health study is provided in Section 4. Section 5 contains some commentary of fragmentization of expectation propagation for more elaborate models.

2 Background Material

Factor graph fragmentization of expectation propagation relies on definitions and results concerning both distribution theory and graph theory, not all of which are commonplace in the statistics literature. We provide the necessary background material in this section.

2.1 Exponential Family Distributions

A random variable x has an exponential family distribution if its probability mass function or density function admits the form

$$p(x) = \exp\{\boldsymbol{T}(x)^T\boldsymbol{\eta} - A(\boldsymbol{\eta})\}h(x), \quad x \in \mathbb{R},\ \boldsymbol{\eta} \in H.$$

The vectors $\boldsymbol{T}(x)$ and $\boldsymbol{\eta}$ are called, respectively, the *sufficient statistic* and *natural parameter*. The set H is the space of allowable natural parameter values. The function $A(\boldsymbol{\eta})$ is called the *log-partition function* and $h(x)$ is the *base measure*. A key exponential family distributional result is that

$$E\{\boldsymbol{T}(x)\} = \nabla A(\boldsymbol{\eta}) \tag{1}$$

where $\nabla A(\boldsymbol{\eta})$ is the column vector of partial derivatives of $A(\boldsymbol{\eta})$ with respect to each of the components of $\boldsymbol{\eta}$.

Table 1 lists each of the exponential families distributions arising in this article, along with their defining functions and parameter spaces. The Normal and Inverse Chi-Squared exponential families are well known. The Moon Rock exponential family is less established, and is given this name in McLean & Wand (2018). In Table 1 and elsewhere we use the following indicator function notation: $I(\mathcal{P}) = 1$ if the proposition $\mathcal{P}$ is true and $I(\mathcal{P}) = 0$ if $\mathcal{P}$ is false.

Note also that the Inverse Chi-Squared exponential family is equivalent to the *Inverse Gamma* exponential family. The two families differ in their common parametrizations as explained in, for example, Section S.1.3 of the online supplement of Wand (2017). The Inverse Chi-Squared distribution has the advantage of being the special case of the Inverse Wishart distribution for 1×1 random matrices. Throughout this article we write

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

name	$\boldsymbol{T}(x)$	$A(\boldsymbol{\eta})$	$h(x)$	H
Normal	$\begin{bmatrix} x \\ x^2 \end{bmatrix}$	$-\frac{1}{4}(\eta_1^2/\eta_2)$ $-\frac{1}{2}\,\log(-2\eta_2)$	1	$\{(\eta_1,\eta_2):$ $\eta_1\in\mathbb{R},\eta_2<0\}$
Inverse Chi-Squared	$\begin{bmatrix} \log(x) \\ 1/x \end{bmatrix}$	$(\eta_1+1)\log(-\eta_2)$ $+\log\Gamma(-\eta_1-1)$	$I(x>0)$	$\{(\eta_1,\eta_2):$ $\eta_1<-1,\eta_2<0\}$
Moon Rock	$\begin{bmatrix} \{x\log(x) \\ -\log\Gamma(x)\} \\ x \end{bmatrix}$	$\log\Big[\int_0^\infty\{t^t/\Gamma(t)\}^{\eta_1}$ $\times\exp(\eta_2\,t)\,dt\Big]$	$I(x>0)$	$\{(\eta_1,\eta_2):$ $\eta_1>0,\eta_1+\eta_2<0\}$

Table 1: *Sufficient statistics, log-partition functions, base measures and natural parameter spaces of three exponential families.*

to denote a $d\times d$ random matrix $\boldsymbol{X}$ having density function

$$p(\boldsymbol{X})=\mathcal{C}_{d,\kappa}^{-1}|\boldsymbol{\Lambda}|^{\kappa/2}\,|\boldsymbol{X}|^{-(\kappa+d+1)/2}\exp\{-\tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}\boldsymbol{X}^{-1})\}\,I(\boldsymbol{X}\text{ symmetric and positive definite})$$

where $\kappa>d-1$, $\boldsymbol{\Lambda}$ is a $d\times d$ symmetric positive definite matrix and

$$\mathcal{C}_{d,\kappa}\equiv 2^{d\kappa/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{\kappa+1-j}{2}\right). \tag{2}$$

For the special case of $d=1$ we write

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

2.2 Kullback-Leibler Projection

If p_1 and p_2 are two univariate density functions then the *Kullback-Leibler divergence* of p_2 from p_1 is

$$\text{KL}(p_1\|p_2)\equiv\int_{-\infty}^{\infty}p_1(x)\log\{p_1(x)/p_2(x)\}\,dx.$$

If $\mathcal{Q}$ is a family of univariate density functions then the projection of the univariate density function p onto $\mathcal{Q}$ is

$$\text{proj}_{\mathcal{Q}}[p]\equiv\underset{q\in\mathcal{Q}}{\text{argmin}}\ \text{KL}(p\|q). \tag{3}$$

A core aspect of expectation propagation is projection of an arbitrary *input* density function onto a particular exponential family. This corresponds to (3) with

$$\mathcal{Q}=\big\{q(\,\cdot\,;\boldsymbol{\eta}):q(x;\boldsymbol{\eta})=\exp\{\boldsymbol{T}(x)^T\boldsymbol{\eta}-A(\boldsymbol{\eta})\}h(x),\ \ \boldsymbol{\eta}\in H\big\}.$$

As explained in Section 2.3 of Kim & Wand (2016), the exponential family Kullback-Leibler problem

$$\boldsymbol{\eta}^*=\underset{\boldsymbol{\eta}\in H}{\text{argmin}}\,\text{KL}\big(p\|q(\cdot;\boldsymbol{\eta})\big)$$

is equivalent to the sufficient statistic moment matching problem

$$\int_{-\infty}^{\infty} \boldsymbol{T}(x)\, \exp\{\boldsymbol{T}(x)^T\boldsymbol{\eta}^* - A(\boldsymbol{\eta}^*)\}h(x)\, dx = \int_{-\infty}^{\infty} \boldsymbol{T}(x)\, p(x)\, dx. \quad (4)$$

Because of (1) we can re-write (4) as

$$(\nabla A)(\boldsymbol{\eta}^*) = \int_{-\infty}^{\infty} \boldsymbol{T}(x)\, p(x)\, dx.$$

Then, assuming that the inverse of ∇A is well-defined,

$$\boldsymbol{\eta}^* = (\nabla A)^{-1}\left(\int_{-\infty}^{\infty} \boldsymbol{T}(x)\, p(x)\, dx\right). \quad (5)$$

Hence, given the $\boldsymbol{T}$ moments, Kullback-Leibler projection of a density function p onto an exponential family boils down to inversion of ∇A. Section 3 of Wainwright & Jordan (2008) provides a detailed study of exponential families including properties of A and ∇A. An exponential family distribution with the sufficient statistic $\boldsymbol{T}(x)$ being a $d \times 1$ vector is said to be *regular* if H is an open set in $\mathbb{R}^d$ and *minimal* if there is no $d \times 1$ vector $\boldsymbol{a}$ and constant $b \in \mathbb{R}$ such that $\boldsymbol{a}^T\boldsymbol{T}(x) = b$ almost surely. Each of the exponential families in Table 1 are regular and minimal. Result 1 provides a summary of results from Section 3 of Wainwright & Jordan (2008) that is relevant to (5). It depends on:

Definition 1. *Consider a function $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^d$. Then the set of realizable expectations of $\boldsymbol{f}$ is the set of points $[\tau_1, \ldots, \tau_d]^T \in \mathbb{R}^d$ such that there exists a univariate random variable x for which $E\{\boldsymbol{f}(x)\} = [\tau_1, \ldots, \tau_d]^T$.*

To illustrate the notion of the set of realizable expectations, consider the functions $\boldsymbol{f}_1 : \mathbb{R} \to \mathbb{R}^2$ and $\boldsymbol{f}_2 : \mathbb{R} \to \mathbb{R}^2$ given by

$$\boldsymbol{f}_1(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{f}_2(x) = \begin{bmatrix} x \\ x^3 \end{bmatrix}.$$

The sets of all realizable expectations of $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ are, respectively

$$\mathfrak{M}_1 \equiv \left\{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} : x_2 \geq x_1^2\right\} \quad \text{and} \quad \mathfrak{M}_2 \equiv \left\{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} : \text{sign}(x_1)\, x_2 \geq |x_1|^3\right\}.$$

To show that $\mathfrak{M}_1$ is the set of all realizable expectations of $\boldsymbol{f}_1$ note that $\mathfrak{M}_1 = \mathfrak{M}_{11} \cup \mathfrak{M}_{12}$ where $\mathfrak{M}_{11} \equiv \{[x_1\ x_2]^T : x_2 = x_1^2\}$ and $\mathfrak{M}_{12} \equiv \{[x_1\ x_2]^T : x_2 > x_1^2\}$. Then for any $[x_1\ x_2] \in \mathfrak{M}_{11}$ we can take x to be the degenerate random variable with probability mass function $p(x) = I(x = x_1)$. For such x, $E\{\boldsymbol{f}_1(x)\} = [x_1\ x_1^2]^T = [x_1\ x_2]^T \in \mathfrak{M}_{11}$ which shows that all elements of $\mathfrak{M}_{11}$ are realizable expectations of $\boldsymbol{f}_1$. For any $[x_1\ x_2] \in \mathfrak{M}_{12}$ taking $x \sim N(x_1, x_2 - x_1^2)$ leads to $E\{\boldsymbol{f}_1(x)\} = [x_1\ x_2]^T$ verifying that all elements of $\mathfrak{M}_{12}$ are realizable by $E\{\boldsymbol{f}_1(x)\}$ for some x. Hence, all entries of $\mathfrak{M}_1$ are realizable by $E\{\boldsymbol{f}_1(x)\}$ for some x.. Values $[x_1\ x_2]^T \notin \mathfrak{M}_1$ are not realizable because Jensen's inequality implies that $E(x^2) \geq \{E(x)\}^2$ for any random variable x. Similar arguments can be used to establish that $\mathfrak{M}_2$ is the set of all realizable expectations of $\boldsymbol{f}_2$. Figure 1 shows the sets $\mathfrak{M}_1$ and $\mathfrak{M}_2$.

We are now ready to give the pivotal:

Result 1 (Wainwright & Jordan, 2008). *Consider a regular and minimal exponential family with d-dimensional sufficient statistic $\boldsymbol{T}(x)$ and corresponding natural parameter vector $\boldsymbol{\eta}$. Then*

(a) *H is a strictly convex subset of $\mathbb{R}^d$.*

(b) *A is a strictly convex and infinitely differentiable function on H.*

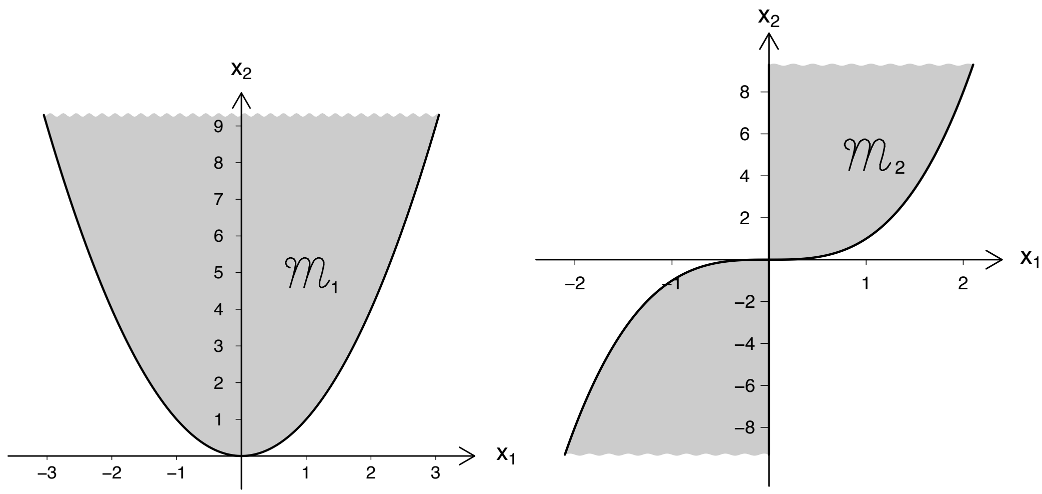


Figure 1: *Left panel: The shaded region is $\mathfrak{M}_1$, the set of realizable expectations of $\boldsymbol{f}_1$. Right panel: The shaded region is $\mathfrak{M}_2$, the set of realizable expectations of $\boldsymbol{f}_2$.*

(c) *∇A is a one-to-one function.*

(d) *The image of ∇A, which we denote by T, is the interior of the set of all realizable expectations of $\boldsymbol{T}$.*

Result 1 guarantees that $\nabla A : H \to T$ is a bijective map and that $(\nabla A)^{-1} : T \to H$ is well-defined.

2.2.1 Normal Distribution Special Case

The Normal distribution is the one of simplest exponential families since ∇A and $(\nabla A)^{-1}$ admit simple closed forms. Firstly, we have

$$\nabla A(\boldsymbol{\eta}) = \begin{bmatrix} -\eta_1/(2\eta_2) \\ (\eta_1^2 - 2\eta_2)/(4\eta_2^2) \end{bmatrix}.$$

It is straightforward to show that the image of H under ∇A is

$$T = \{(\tau_1, \tau_2) : \tau_2 > \tau_1^2\}$$

and the inverse of ∇A is

$$(\nabla A)^{-1}(\boldsymbol{\tau}) = \begin{bmatrix} \tau_1/(\tau_2 - \tau_1^2) \\ -1/\{2(\tau_2 - \tau_1^2)\} \end{bmatrix}.$$

2.2.2 Inverse Chi-Squared Distribution Special Case

For the Inverse Chi-Squared distribution we have

$$\nabla A(\boldsymbol{\eta}) = \begin{bmatrix} \log(-\eta_2) - \text{digamma}(-\eta_1 - 1) \\ (\eta_1 + 1)/\eta_2 \end{bmatrix}$$

where $\text{digamma}(x) \equiv \frac{d}{dx} \log \Gamma(x)$. Determination of the image of H under ∇A is more challenging for the Inverse Chi-Squared distribution. It is aided by Theorem 1 of Kim &

Wand (2016) which establishes that $\log -$digamma is a bijective map between $\mathbb{R}_+$ and $\mathbb{R}_+$. This leads to

$$T = \{(\tau_1, \tau_2) : \tau_2 > e^{-\tau_1}\}.$$

The inverse or ∇A is

$$(\nabla A)^{-1}(\boldsymbol{\tau}) = \begin{bmatrix} -(\log -\text{digamma})^{-1}\big(\tau_1 + \log(\tau_2)\big) - 1 \\ -(\log -\text{digamma})^{-1}\big(\tau_1 + \log(\tau_2)\big)/\tau_2 \end{bmatrix}.$$

Theorem 1 of Kim & Wand (2016) implies that $(\log -\text{digamma})^{-1}$ is well-defined.

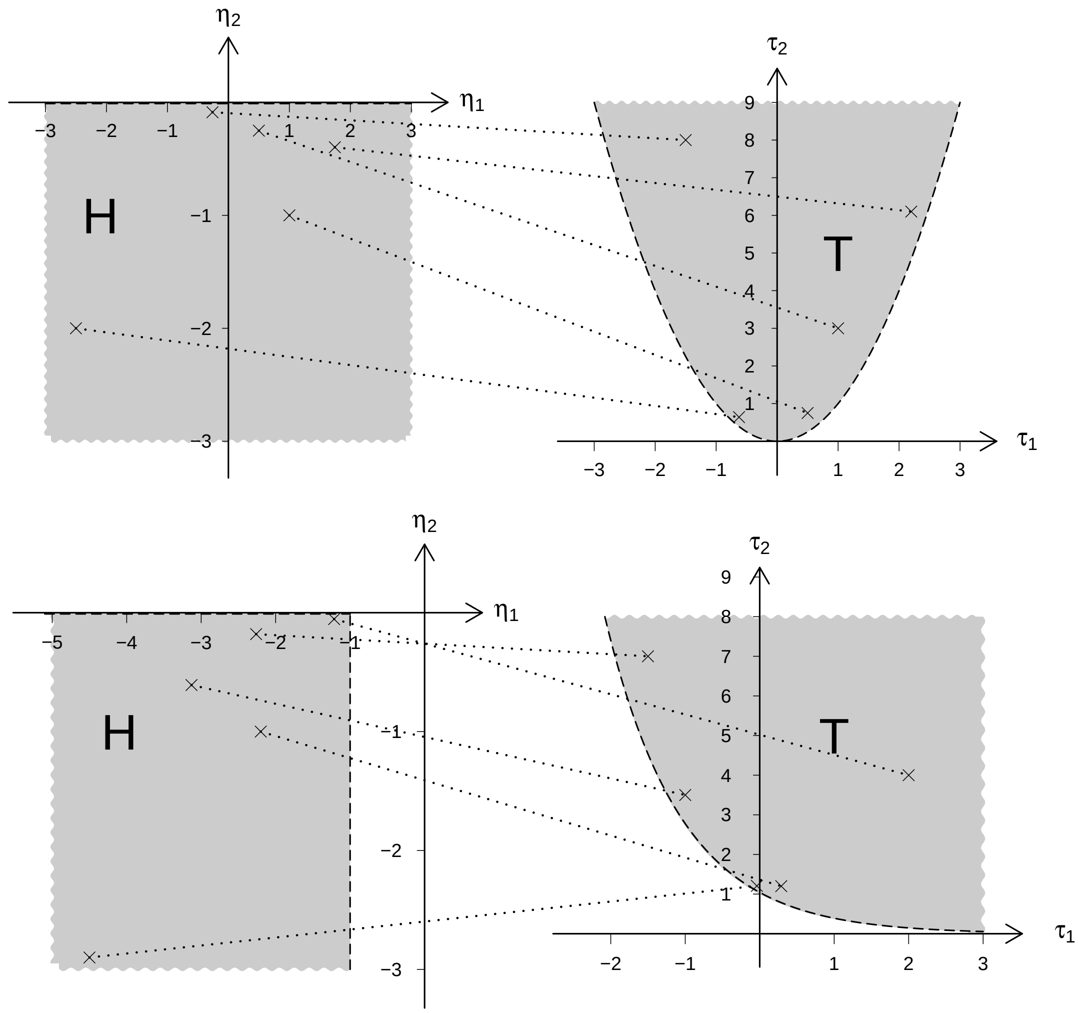


Figure 2: *Upper panel: Illustration of the bijective maps between H and T for the Normal exponential family. The crosses and dotted lines depict five example $\boldsymbol{\eta} \in H$ and $\boldsymbol{\tau} = \nabla A(\boldsymbol{\eta}) \in T$ pairs. Since ∇A is a bijective map, the crosses and dotted lines equivalently depict five example $\boldsymbol{\tau} \in T$ and $\boldsymbol{\eta} = (\nabla A)^{-1}(\boldsymbol{\tau}) \in H$ pairs. Lower panel: Similar illustration for the Inverse Chi-Squared exponential family.*

Figure 2 depicts the ∇A and $(\nabla A)^{-1}$ bijective maps between H and T for both the Normal and Inverse Chi-Squared exponential family distributions.

2.3 Factor Graphs and Factor Graph Fragments

A *factor graph* is a graphical representation of the factor/argument dependencies of a multivariate function. Even though the concept applies to functions in general, the relevant functions are joint density functions in the context of expectation propagation. As an illustration, consider the Bayesian linear model

$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\,\boldsymbol{I}),$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of responses, with the following prior distributions on the regression coefficients and error standard deviation:

$$\boldsymbol{\beta} \sim N(\boldsymbol{0},\sigma_{\boldsymbol{\beta}}^2\boldsymbol{I}) \quad \text{and} \quad \sigma \sim \text{Half-Cauchy}(A),$$

The second prior specification means that σ has prior density function $p(\sigma) = 2/[A\pi\{1+(\sigma/A)^2\}]$ for $\sigma > 0$. An equivalent representation of the model, involving the auxiliary variable a, is

$$\begin{gathered}\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0},\sigma_{\boldsymbol{\beta}}^2\boldsymbol{I}),\\ \sigma^2|a \sim \text{Inverse-}\chi^2(1,1/a), \quad a \sim \text{Inverse-}\chi^2(1,1/A^2).\end{gathered} \tag{6}$$

We work with this auxiliary variable representation since it aids tractability of expectation propagation. The joint density function of the random variables and random vectors in (6) admits the following factorized form:

$$p(\boldsymbol{y},\boldsymbol{\beta},\sigma^2,a) = p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)p(\sigma^2|a)p(a).$$

Now let $\boldsymbol{x}_i^T$ be the ith row of $\boldsymbol{X}$ for $1 \le i \le n$. Then a further breakdown of $p(\boldsymbol{y},\boldsymbol{\beta},\sigma^2,a)$ is

$$p(\boldsymbol{y},\boldsymbol{\beta},\sigma^2,a) = p(\boldsymbol{\beta})\left\{\prod_{i=1}^n \int_{-\infty}^{\infty} \delta(\alpha_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\,p(y_i|\alpha_i,\sigma^2)\,d\alpha_i\right\}p(\sigma^2|\,a)\,p(a) \tag{7}$$

where δ is the Dirac delta function and $p(y_i|\alpha_i,\sigma^2) \equiv (2\pi\sigma^2)^{-1/2}\exp\{-(y_i-\alpha_i)^2/(2\sigma^2)\}$. Figure 3 is a factor graph representation of $p(\boldsymbol{y},\boldsymbol{\beta},\sigma^2,a)$ according to the factors that appear in (7). At this point we note that we are not using the conventional factor graph definition here since some of the factors appear inside the integrals in (7). Kim & Wand (2017) introduced the term *derived variable factor graph* to make this distinction. We will simply call it a *factor graph* from now onwards. The circles are called *stochastic nodes* and the rectangles are called *factors*. Both circles and rectangles are *nodes* of the factor graph. We say that a two nodes are *neighbors* of each other if they are joined by an edge.

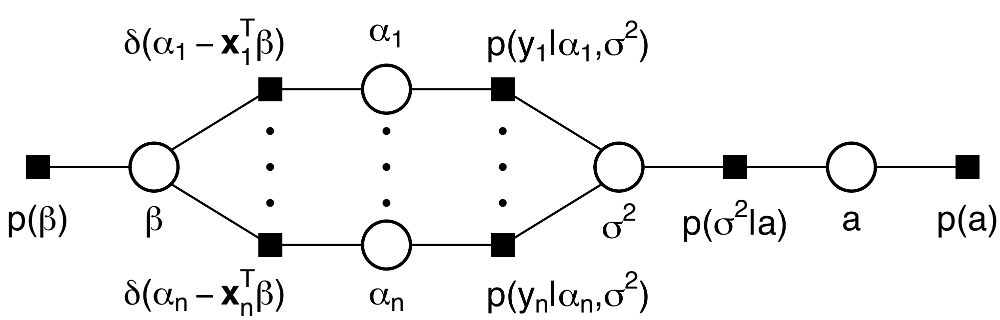


Figure 3: *Factor graph representation of (7).*

Figure 4 is a representation of Figure 3 with factor graph *fragments* of the same type identified via color-coding and numbering of the factors. As defined in Wand (2017), a fragment is a sub-graph of a factor graph consisting of a factor and each of its neighboring stochastic nodes.

The five different colors in Figure 4 correspond to five different fragment types. Some of the fragment types, such as that corresponding to the $p(\boldsymbol{\beta})$ factor, only appear once in this factor graph. Other types, such as those corresponding to $\delta(\alpha_i - \boldsymbol{x}_i^T\boldsymbol{\beta})$, $1 \le i \le n$,

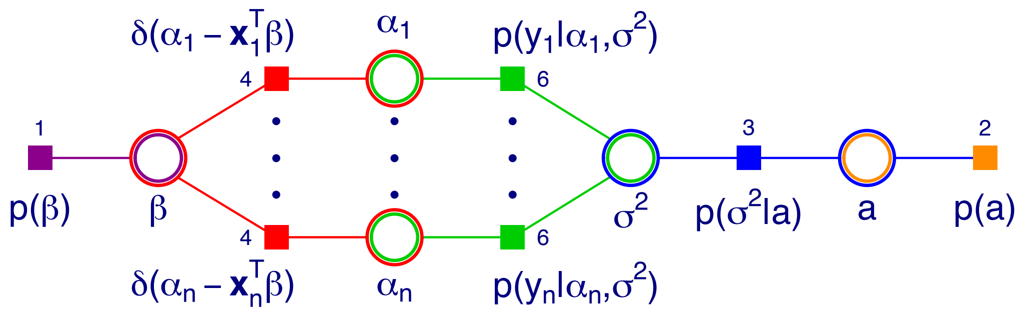


Figure 4: *Fragmentization of the Figure 3 factor graph. Different colors signify fragments of the same type, and are included in Table 2.*

appear multiple times. Recognition of the recurrence of fragments of the same type in this factor graph and factor graphs for other models is at the core of extension to arbitrarily large models. Wand (2017) demonstrated factor graph fragmentization of variational message passing. Our goal here is to do the same for expectation propagation.

2.4 Expectation Propagation

Recent summaries of expectation propagation are provided in Kim & Wand (2016, 2017). We briefly cover the main points here. The function neighbors$(\cdot)$ plays an important role in the algebraic description of the message updates. Consider the illustrative generic form factor graph shown in Figure 5, corresponding to the joint density function of random vectors $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_5$ according to a particular Bayesian model. Then neighbours$(1) = \{1,2,5\}$ since the factor f_1 is connected by edges to each of $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_5$. Similarly, neighbours$(2) = \{2,3,4\}$, neighbours$(3) = \{3\}$, neighbours$(4) = \{4,5\}$ and neighbours$(5) = \{1,5\}$. For general factor graphs with the $\boldsymbol{\theta}_i$ and f_j labeling, neighbours(j) is the set of indices of the $\boldsymbol{\theta}_i$ that are connected to f_j by an edge. Based on (54) of Minka (2005), the

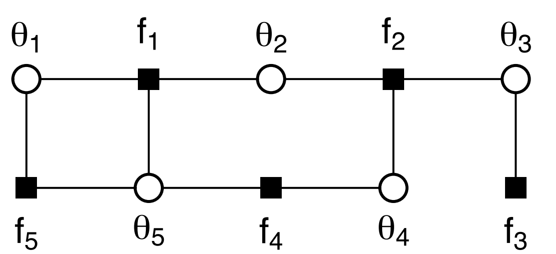


Figure 5: *An illustrative generic form factor graph.*

stochastic node to factor messages are updated according to

$$m_{\boldsymbol{\theta}_i\to f_j}(\boldsymbol{\theta}_i) \longleftarrow \prod_{j'\neq j:\, i\in\text{neighbours}(j')} m_{f_{j'}\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \tag{8}$$

and, based on (83) of Minka (2005), the factor to stochastic node messages updates are

$$m_{f_j\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \longleftarrow \frac{\text{proj}\left[m_{\boldsymbol{\theta}_i\to f_j}(\boldsymbol{\theta}_i)\int f_j(\boldsymbol{\theta}_{\text{neighbours}(j)})\prod_{i'\in\text{neighbours}(j)\backslash\{i\}} m_{\boldsymbol{\theta}_{i'}\to f_j}(\boldsymbol{\theta}_{i'})\,d\boldsymbol{\theta}_{\text{neighbours}(j)\backslash\{i\}}/Z\right]}{m_{\boldsymbol{\theta}_i\to f_j}(\boldsymbol{\theta}_i)}, \tag{9}$$

where Z is the normalizing factor that ensures that the function of $\boldsymbol{\theta}_i$ inside the proj$[\,\cdot\,]$ is a density function. The normalizing factor in (9) involves summation if some of the

$\boldsymbol{\theta}_{i'}$ have discrete components. The proj[$\cdot$] in (9) denotes Kullback-Leibler projection onto an appropriate exponential family of density functions. However, in Kim & Wand (2016) illustration was done only via a simple example in which all of the stochastic nodes were univariate. In the case of linear models, in which vector parameters are present, some adjustments are necessary to avoid intractable multivariate integrals. The first one is an intrinsically important convention and is now spelt out:

Convention 1. *Derived variable factor graphs are treated as ordinary factor graphs when it comes to applying the message passing expressions (8) and (9).*

In practice, iteration involving (8) and (9) may require some tweaking to achieve convergence. Minka (2005) recommends the *damping* adjusment

$$m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \longleftarrow m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)^{\varepsilon} \times \{\text{right-hand side of (9)}\}^{1-\varepsilon}. \tag{10}$$

for some $0 \leq \varepsilon < 1$. Kim and Wand (2017) noted that setting ε to a small positive number such as $\varepsilon = 0.1$ aided convergence for their expectation propagation algorithms for fitting linear models. Therefore, we build this adjustment into the fragment updates in the next section.

The full expectation propagation iterative algorithm is:

Initialize all factor to stochastic node messages.

Cycle until all factor to stochastic node messages converge:

For each factor:

Compute the messages passed to the factor using (8).
Compute the messages passed from the factor using (9) and (10).

Upon convergence the expectation propagation-approximate posterior density function of $\boldsymbol{\theta}_i$ is obtained from

$$q^*(\boldsymbol{\theta}_i) \propto \prod_{j: i \in \text{neighbours}(j)} m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i).$$

3 Fragmentization for Generalized, Linear and Mixed Models

Each of the generalized, linear and mixed models dealt with in Kim & Wand (2017) can be handled with nine distinct fragment types, which are listed in Table 2. The message updates for each fragment type only needs to be derived once. Each subsection deals with the required derivation and summarizes the updates as an algorithm. For a software suite that uses expectation propagation to fit generalized, linear and mixed models the fragment only needs to be implemented once. We now work through each of the Table 2 fragments in turn.

The algorithms use the matrix functions vec and its inverse vec^{-1} which we define here. If $\boldsymbol{A}$ is $d \times d$ matrix then $\text{vec}(\boldsymbol{A})$ is the $d^2 \times 1$ vector obtained by stacking the columns of $\boldsymbol{A}$ underneath each other in order from left to right. if $\boldsymbol{a}$ is a $d^2 \times 1$ vector then $\text{vec}^{-1}(\boldsymbol{a})$ is the $d \times d$ matrix formed from listing the entries of $\boldsymbol{a}$ in a column-wise fashion in order from left to right and is the usual function inverse when the domain of vec is restricted to square matrices.

The following shorthand is used throughout this section:

$$a \xleftarrow{\varepsilon} b \quad \text{denotes} \quad a \longleftarrow \varepsilon\, a + (1-\varepsilon)\, b.$$

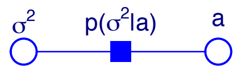
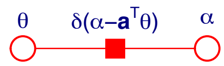
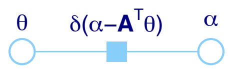
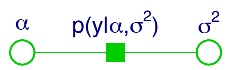
Fragment name	Diagram	Distributional statement
1. Gaussian prior		$\boldsymbol{\theta} \sim N(\boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta})$
2. Inverse Wishart prior		$\boldsymbol{\Theta} \sim \text{Inverse-Wishart}(\kappa_{\boldsymbol{\Theta}}, \boldsymbol{\Lambda_\Theta})$
3. Iterated Inverse Chi-Squared		$\sigma^2 \| a \sim \text{Inverse-}\chi^2(\nu, 1/a)$
4. Linear combination derived variable		$\alpha \equiv \boldsymbol{a}^T\boldsymbol{\theta}$
5. Multivariate linear combination derived variable		$\boldsymbol{\alpha} \equiv \boldsymbol{A}^T\boldsymbol{\theta}$
6. Gaussian		$y\| \, \alpha, \sigma^2 \sim N(\alpha, \sigma^2)$
7. Logistic likelihood		$y\| \, \alpha \sim \text{Bernoulli}(\text{logit}^{-1}(\alpha))$
8. Probit likelihood		$y\| \, \alpha \sim \text{Bernoulli}(\Phi(\alpha))$
9. Poisson likelihood		$y\| \, \alpha \sim \text{Poisson}(\exp(\alpha))$

Table 2: *Fundamental factor graph fragments for expectation propagation fitting of generalized, linear and mixed models.*

3.1 Gaussian Prior Fragment

The Gaussian prior fragment arises from the following prior distribution specification:

$$\boldsymbol{\theta} \sim N(\boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta})$$

for user-specified hyperparameters $\boldsymbol{\mu_\theta}$ and $\boldsymbol{\Sigma_\theta}$. The fragment factor is

$$p(\boldsymbol{\theta}) = (2\pi)^{-d_\theta/1}|\boldsymbol{\Sigma_\theta}|^{-1/2}\exp\big\{-\tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu_\theta})^T\boldsymbol{\Sigma_\theta}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu_\theta})\big\}.$$

We assume that

$$\text{all messages passed to } \boldsymbol{\theta} \text{ from factors outside of the fragment are in the Multivariate Normal family.} \tag{11}$$

The message from $p(\boldsymbol{\theta})$ to $\boldsymbol{\theta}$ takes the form

$$m_{p(\boldsymbol{\theta})\to\boldsymbol{\theta}}(\boldsymbol{\theta}) = \exp\left\{\begin{bmatrix}\boldsymbol{\theta} \\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{\theta})\to\boldsymbol{\theta}}\right\}.$$

Algorithm 1 provides the natural parameter update for this simple fragment. The derivation of Algorithm 1 is given in Section S.2.1 of the online supplement.

Hyperparameter Inputs: $\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}$.

Update:

$$\boldsymbol{\eta}_{p(\boldsymbol{\theta})\,\rightarrow\,\boldsymbol{\theta}} \longleftarrow \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}} \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}) \end{bmatrix}$$

Parameter Output: $\boldsymbol{\eta}_{p(\boldsymbol{\theta})\,\rightarrow\,\boldsymbol{\theta}}$.

Algorithm 1: *The input, update and output of the Gaussian prior fragment.*

3.2 Inverse Wishart Prior Fragment

Let $\boldsymbol{\Theta}$ be a $d^{\Theta} \times d^{\Theta}$ symmetric positive definite random matrix. The prior specification

$$\boldsymbol{\Theta} \sim \text{Inverse-Wishart}\big(\kappa_{\boldsymbol{\Theta}}, \boldsymbol{\Lambda}_{\boldsymbol{\Theta}}\big)$$

leads to a factor graph fragment with factor

$$p(\boldsymbol{\Theta}) = \mathcal{C}_{d^{\Theta},\kappa_{\boldsymbol{\Theta}}}^{-1} |\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}|^{\kappa_{\boldsymbol{\Theta}}/2}\, |\boldsymbol{\Theta}|^{-(\kappa_{\boldsymbol{\Theta}}+d^{\Theta}+1)/2} \exp\{-\tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}\boldsymbol{\Theta}^{-1})\}\, I(\boldsymbol{\Theta}\text{ symmetric and positive definite}).$$

where $\mathcal{C}_{d^{\Theta},\kappa_{\boldsymbol{\Theta}}}$ is defined via (2). The message from $p(\boldsymbol{\Theta})$ to $\boldsymbol{\Theta}$ takes the form

$$m_{p(\boldsymbol{\Theta})\,\rightarrow\,\boldsymbol{\Theta}}(\boldsymbol{\Theta}) = \exp\left\{ \begin{bmatrix} \log|\boldsymbol{\Theta}| \\ \text{vec}(\boldsymbol{\Theta}^{-1}) \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{\Theta})\,\rightarrow\,\boldsymbol{\Theta}} \right\}.$$

Algorithm 2 gives the $\boldsymbol{\eta}_{p(\boldsymbol{\Theta})\,\rightarrow\,\boldsymbol{\Theta}}$ update based on hyperparameter inputs $\kappa_{\boldsymbol{\Theta}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}$.

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

Update:

$$\boldsymbol{\eta}_{p(\boldsymbol{\Theta})\,\rightarrow\,\boldsymbol{\Theta}} \longleftarrow \begin{bmatrix} -\frac{1}{2}(\kappa_{\boldsymbol{\Theta}} + d^{\Theta} + 1) \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}) \end{bmatrix}$$

Parameter Output: $\boldsymbol{\eta}_{p(\boldsymbol{\Theta})\,\rightarrow\,\boldsymbol{\Theta}}$.

Algorithm 2: *The input, update and output of the Inverse Wishart prior fragment.*

A derivation of Algorithm 2 is given in Section S.2.2 of the online supplement.

3.3 Iterated Inverse Chi-Squared Fragment

This fragment arises from the following distributional fact (e.g. Wand *et al.* (2011), Result 5):

$$\begin{gathered} \sigma^2|a \sim \text{Inverse-}\chi^2(\nu,\nu/a) \quad\text{and}\quad a \sim \text{Inverse-}\chi^2(1, 1/A^2) \\ \text{implies}\quad \sigma \sim \text{Half-t}(A,\nu) \end{gathered} \qquad (12)$$

where $x \sim \text{Half-t}(A,\nu)$ if and only if

$$p(x) = \frac{2\Gamma\left(\frac{\nu+1}{2}\right) I(x>0)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)A\{1+(x/A)^2/\nu\}^{(\nu+1)/2}}.$$

The advantage of fact (12) is that non-informative priors within the Half-t family can be imposed on standard deviation parameters using messages within the Inverse Chi-Squared family.

The fragment factor is

$$p(\sigma^2|a) = \frac{\{\nu/(2a)\}^{\nu/2}}{\Gamma(\nu/2)}\,(\sigma^2)^{-(\nu/2)-1}\exp\{-\nu/(2a\sigma^2)\}\,I(\sigma^2>0)\,I(a>0)$$

and it is assumed that:

$$\begin{array}{l}\text{all messages passed to } \sigma^2 \text{ from factors outside of}\\ \text{the fragment are in the Inverse Chi-Squared family and}\\ \text{all messages passed to } a \text{ from factors outside of}\\ \text{the fragment are also in the Inverse Chi-Squared family.}\end{array} \tag{13}$$

The messages from the factor to its neighboring stochastic nodes are

$$m_{p(\sigma^2|a)\,\rightarrow\,\sigma^2}(\sigma^2) = \exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma^2\end{bmatrix}^T \boldsymbol{\eta}_{p(\sigma^2|a)\,\rightarrow\,\sigma^2}\right\}$$

and

$$m_{p(\sigma^2|a)\,\rightarrow\,a}(a) = \exp\left\{\begin{bmatrix}\log(a)\\ 1/a\end{bmatrix}^T \boldsymbol{\eta}_{p(\sigma^2|a)\,\rightarrow\,a}\right\}.$$

Algorithm 3 provides the updates of the natural parameters of these messages given messages from outside the fragment. The function G^{IG3} is defined in Section S.1.3.

Data Input: $\nu > 0,\quad 0 \le \varepsilon < 1.$

Parameter Inputs: $\boldsymbol{\eta}_{p(\sigma^2|a)\,\rightarrow\,\sigma^2}, \boldsymbol{\eta}_{p(\sigma^2|a)\,\rightarrow\,a}, \boldsymbol{\eta}_{\sigma^2\,\rightarrow\,p(\sigma^2|a)}, \boldsymbol{\eta}_{a\,\rightarrow\,p(\sigma^2|a)}.$

Updates:

$$\boldsymbol{\eta}_{p(\sigma^2|a)\,\rightarrow\,\sigma^2} \xleftarrow{\varepsilon} G^{\text{IG3}}\left(\boldsymbol{\eta}_{\sigma^2\,\rightarrow\,p(\sigma^2|a)}, \boldsymbol{\eta}_{a\,\rightarrow\,p(\sigma^2|a)}; \nu+2, \nu\right)$$

$$\boldsymbol{\eta}_{p(\sigma^2|a)\,\rightarrow\,a} \xleftarrow{\varepsilon} G^{\text{IG3}}\left(\boldsymbol{\eta}_{a\,\rightarrow\,p(\sigma^2|a)}, \boldsymbol{\eta}_{\sigma^2\,\rightarrow\,p(\sigma^2|a)}; \nu, \nu\right)$$

Parameter Outputs: $\boldsymbol{\eta}_{p(\sigma^2|a)\,\rightarrow\,\sigma^2}, \boldsymbol{\eta}_{p(\sigma^2|a)\,\rightarrow\,a}.$

Algorithm 3: *The inputs, updates and outputs of the iterated Inverse Chi-Squared fragment.*

3.4 Linear Combination Derived Variable Fragment

The linear combination derived variable fragment corresponds to equating a scalar variable α with a linear combination $\boldsymbol{a}^T\boldsymbol{\theta}$. If g is a general function that depends on the linear combination form $\boldsymbol{a}^T\boldsymbol{\theta}$ and other variables, denoted by $\boldsymbol{o}$, then the derived variable α arises from the equality:

$$g(\boldsymbol{a}^T\boldsymbol{\theta};\boldsymbol{o}) = \int_{-\infty}^{\infty}\delta(\alpha-\boldsymbol{a}^T\boldsymbol{\theta})\,g(\alpha;\boldsymbol{o})\,d\alpha. \tag{14}$$

where δ is the Dirac delta function. Under Convention 1 given in Section 2.4, the integral sign is ignored when it comes to applying the expectation propagation updates (8) and (9). We assume that:

$$\begin{array}{l}\text{all messages passed to } \alpha \text{ from factors outside of}\\ \text{the fragment are in the Univariate Normal family}\\ \text{and all messages passed to } \boldsymbol{\theta} \text{ from factors outside}\\ \text{of the fragment are in the Multivariate Normal family.}\end{array} \tag{15}$$

The function $\delta(\alpha-\boldsymbol{a}^T\boldsymbol{\theta})$ is the factor for this fragment. According to conjugacy restrictions, messages passed from $\delta(\alpha-\boldsymbol{a}^T\boldsymbol{\theta})$ to α and $\boldsymbol{\theta}$ take the forms

$$m_{\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})\,\rightarrow\,\alpha}(\alpha) = \exp\left\{\left[\begin{array}{c}\alpha\\ \alpha^2\end{array}\right]^T \boldsymbol{\eta}_{\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})\,\rightarrow\,\alpha}\right\} \tag{16}$$

and

$$m_{\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})\,\rightarrow\,\boldsymbol{\theta}}(\boldsymbol{\theta}) = \exp\left\{\left[\begin{array}{c}\boldsymbol{\theta}\\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{array}\right]^T \boldsymbol{\eta}_{\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})\,\rightarrow\,\boldsymbol{\theta}}\right\}.$$

Algorithm 4 provides the updates to the natural parameter vectors

$$\boldsymbol{\eta}_{\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})\,\rightarrow\,\alpha} \quad \text{and} \quad \boldsymbol{\eta}_{\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})\,\rightarrow\,\boldsymbol{\theta}}$$

given inputs

$$\boldsymbol{\eta}_{\alpha\,\rightarrow\,\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})} \quad \text{and} \quad \boldsymbol{\eta}_{\boldsymbol{\theta}\,\rightarrow\,\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})}.$$

It uses the notation

$$\begin{array}{rl}\left(\boldsymbol{\eta}_{\boldsymbol{\theta}\,\rightarrow\,\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})}\right)_1 \equiv & \text{vector containing the first } d^{\boldsymbol{\theta}} \text{ entries of } \boldsymbol{\eta}_{\boldsymbol{\theta}\,\rightarrow\,\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})}\\ \text{and } \left(\boldsymbol{\eta}_{\boldsymbol{\theta}\,\rightarrow\,\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})}\right)_2 \equiv & \text{vector containing the remaining } (d^{\boldsymbol{\theta}})^2 \text{ entries of } \boldsymbol{\eta}_{\boldsymbol{\theta}\,\rightarrow\,\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})}\end{array}$$

where $d^{\boldsymbol{\theta}}$ is the number of entries in $\boldsymbol{\theta}$. The derivations of these updates are given in Section S.2.5 of the online supplement.

3.5 Multivariate Linear Combination Derived Variable Fragment

Now consider the following bivariate extension of (14):

$$g\big(\boldsymbol{a}_1^T\boldsymbol{\theta}, \boldsymbol{a}_2^T\boldsymbol{\theta}; \boldsymbol{o}\big) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(\alpha_1 - \boldsymbol{a}_1^T\boldsymbol{\theta})\delta(\alpha_2 - \boldsymbol{a}_2^T\boldsymbol{\theta})\, g(\alpha_1, \alpha_2; \boldsymbol{o})\, d\alpha_1\, d\alpha_2, \tag{17}$$

where the primary argument of the function g is now bivariate. The established result for the Dirac delta function applied to bivariate arguments leads to the equivalent form for the right-hand side of (17) taking the form:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta\left(\left[\begin{array}{c}\alpha_1\\ \alpha_2\end{array}\right] - \boldsymbol{A}^T\boldsymbol{\theta}\right) g(\alpha_1, \alpha_2; \boldsymbol{o})\, d\alpha_1\, d\alpha_2 \quad \text{where} \quad \boldsymbol{A} \equiv [\boldsymbol{a}_1\ \boldsymbol{a}_2].$$

It follows that

$$\boldsymbol{\alpha} \equiv \left[\begin{array}{c}\alpha_1\\ \alpha_2\end{array}\right]$$

is a bivariate derived variable corresponding to the multivariate linear combination $\boldsymbol{A}^T\boldsymbol{\theta}$.

Data Input: $\boldsymbol{a}$ (vector having the same dimension as $\boldsymbol{\theta}$), $\quad 0 \le \varepsilon < 1$.

Parameter Inputs: $\boldsymbol{\eta}_{\delta(\alpha - \boldsymbol{a}^T\boldsymbol{\theta}) \to \alpha}$, $\boldsymbol{\eta}_{\delta(\alpha - \boldsymbol{a}^T\boldsymbol{\theta}) \to \boldsymbol{\theta}}$, $\boldsymbol{\eta}_{\alpha \to \delta(\alpha - \boldsymbol{a}^T\boldsymbol{\theta})}$, $\boldsymbol{\eta}_{\boldsymbol{\theta} \to \delta(\alpha - \boldsymbol{a}^T\boldsymbol{\theta})}$.

Updates:

$$\boldsymbol{\omega} \longleftarrow \left\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{\boldsymbol{\theta} \to \delta(\alpha - \boldsymbol{a}^T\boldsymbol{\theta})}\big)_2\Big)\right\}^{-1}\boldsymbol{a}$$

$$\boldsymbol{\eta}_{\delta(\alpha - \boldsymbol{a}^T\boldsymbol{\theta}) \to \alpha} \xleftarrow{\varepsilon} \frac{1}{\boldsymbol{\omega}^T\boldsymbol{a}}\begin{bmatrix} \boldsymbol{\omega}^T\big(\boldsymbol{\eta}_{\boldsymbol{\theta} \to \delta(\alpha - \boldsymbol{a}^T\boldsymbol{\theta})}\big)_1 \\ 1 \end{bmatrix}$$

$$\boldsymbol{\eta}_{\delta(\alpha - \boldsymbol{a}^T\boldsymbol{\theta}) \to \boldsymbol{\theta}} \xleftarrow{\varepsilon} \begin{bmatrix} \boldsymbol{a}\big(\boldsymbol{\eta}_{\alpha \to \delta(\alpha - \boldsymbol{a}^T\boldsymbol{\theta})}\big)_1 \\ \text{vec}(\boldsymbol{a}\boldsymbol{a}^T)\big(\boldsymbol{\eta}_{\alpha \to \delta(\alpha - \boldsymbol{a}^T\boldsymbol{\theta})}\big)_2 \end{bmatrix}$$

Parameter Outputs: $\boldsymbol{\eta}_{\delta(\alpha - \boldsymbol{a}^T\boldsymbol{\theta}) \to \alpha}$, $\boldsymbol{\eta}_{\delta(\alpha - \boldsymbol{a}^T\boldsymbol{\theta}) \to \boldsymbol{\theta}}$.

Algorithm 4: *The inputs, updates and outputs of the linear combination derived variable fragment.*

In the most general case, $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are, respectively, $d^{\boldsymbol{\theta}} \times 1$ and $d^{\boldsymbol{\alpha}} \times 1$ vectors and $\boldsymbol{A}$ is a $d^{\boldsymbol{\theta}} \times d^{\boldsymbol{\alpha}}$ matrix. The fragment factor is $\delta(\boldsymbol{\alpha} - \boldsymbol{A}^T\boldsymbol{\theta})$ and the message given in (16) generalizes to

$$\boldsymbol{\eta}_{\delta(\boldsymbol{\alpha} - \boldsymbol{A}^T\boldsymbol{\theta}) \to \boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \exp\left\{\begin{bmatrix} \boldsymbol{\alpha} \\ \text{vec}(\boldsymbol{\alpha}\boldsymbol{\alpha}^T) \end{bmatrix}^T \boldsymbol{\eta}_{\delta(\boldsymbol{\alpha} - \boldsymbol{A}^T\boldsymbol{\theta}) \to \boldsymbol{\alpha}}\right\}.$$

Algorithm 5 lists the natural parameter updates. Their derivations are given in Section S.2.5 of the online supplement.

Note that Algorithm 5 is a generalization of Algorithm 4. Therefore, from a strict mathematical standpoint, Algorithm 4. However, since ordinary linear combinations are common in expectation propagation fitting of linear models we feel that it is worth having a separate fragment and algorithm for this special case.

3.6 Gaussian Fragment

The Gaussian fragment corresponds to the specification

$$y|\alpha, \sigma^2 \sim N(\alpha, \sigma^2).$$

The fragment's factor is

$$p(y|\alpha, \sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\{-(y-\alpha)^2/(2\sigma^2)\}$$

which, as a function of α, is in the Normal family and, as a function of σ^2, is in the Inverse Chi-Squared family. Exponential family constraint considerations then lead to the following assumption for the Gaussian fragment:

$$\begin{array}{l}\text{all messages passed to } \alpha \text{ from factors outside of} \\ \text{the fragment are in the Univariate Normal family} \\ \text{and all messages passed to } \sigma^2 \text{ from factors outside} \\ \text{of the fragment are in the Inverse Chi-Squared family.}\end{array} \tag{18}$$

Data Input: $\boldsymbol{A}$ (matrix with number of columns matching the dimension of $\boldsymbol{\theta}$), $0 \leq \varepsilon < 1$.

Parameter Inputs: $\boldsymbol{\eta}_{\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})\to\boldsymbol{\alpha}}$, $\boldsymbol{\eta}_{\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})\to\boldsymbol{\theta}}$, $\boldsymbol{\eta}_{\boldsymbol{\alpha}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}$, $\boldsymbol{\eta}_{\boldsymbol{\theta}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}$.

Updates:

$$\boldsymbol{\Omega} \longleftarrow \left\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{\boldsymbol{\theta}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}\big)_2\Big)\right\}^{-1}\boldsymbol{A}$$

$$\boldsymbol{\eta}_{\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})\to\boldsymbol{\alpha}} \xleftarrow{\varepsilon} \begin{bmatrix} (\boldsymbol{\Omega}^T\boldsymbol{A})^{-1}\boldsymbol{\Omega}^T\Big(\boldsymbol{\eta}_{\boldsymbol{\theta}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}\Big)_1 \\ \mathrm{vec}\big((\boldsymbol{\Omega}^T\boldsymbol{A})^{-1}\big) \end{bmatrix}$$

$$\boldsymbol{\eta}_{\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})\to\boldsymbol{\theta}} \xleftarrow{\varepsilon} \begin{bmatrix} \boldsymbol{A}\Big(\boldsymbol{\eta}_{\boldsymbol{\alpha}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}\Big)_1 \\ (\boldsymbol{A}\otimes\boldsymbol{A})\Big(\boldsymbol{\eta}_{\boldsymbol{\alpha}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}\Big)_2 \end{bmatrix}$$

Parameter Outputs: $\boldsymbol{\eta}_{\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})\to\boldsymbol{\alpha}}$, $\boldsymbol{\eta}_{\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})\to\boldsymbol{\theta}}$.

Algorithm 5: *The inputs, updates and outputs of the multivariate linear combination derived variable fragment.*

The messages from $p(y|\alpha,\sigma^2)$ take the forms

$$m_{p(y|\alpha,\sigma^2)\to\alpha}(\alpha) = \exp\left\{\begin{bmatrix}\alpha\\ \alpha^2\end{bmatrix}^T \boldsymbol{\eta}_{p(y|\alpha,\sigma^2)\to\alpha}\right\}$$

and

$$m_{p(by|\alpha,\sigma^2)\to\sigma^2}(\sigma^2) = \exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma^2\end{bmatrix}^T \boldsymbol{\eta}_{p(y|\alpha,\sigma^2)\to\sigma^2}\right\}$$

with natural parameters updated according to Algorithm 6. The functions G^{N} and G^{IG3} are defined in Section S.1.3. Algorithm 6's derivation is given in Section S.2.6.

3.7 Logistic Likelihood Fragment

The logistic likelihood fragment corresponds to the specification

$$y|\alpha \sim \text{Bernoulli}\{\text{logit}^{-1}(\alpha)\}.$$

The factor of the fragment is

$$p(y|\alpha) = \exp\{y\alpha - \log(1+e^{\alpha})\}.$$

We assume that:

$$\begin{array}{l}\text{all messages passed to } \alpha \text{ from other factors are}\\ \text{within the Univariate Normal exponential family.}\end{array} \tag{19}$$

Conjugacy then dictates that

$$m_{p(y|\alpha)\to\alpha}(\alpha) = \exp\left\{\begin{bmatrix}\alpha\\ \alpha^2\end{bmatrix}^T \boldsymbol{\eta}_{p(y|\alpha)\to\alpha}\right\}. \tag{20}$$

Data Input: $y \in \mathbb{R}, \quad 0 \leq \varepsilon < 1.$

Parameter Inputs: $\boldsymbol{\eta}_{p(y|\alpha,\sigma^2) \to \alpha}$, $\boldsymbol{\eta}_{p(y|\alpha,\sigma^2) \to \sigma^2}$, $\boldsymbol{\eta}_{\alpha \to p(y|\alpha,\sigma^2)}$, $\boldsymbol{\eta}_{\sigma^2 \to p(y|\alpha,\sigma^2)}$.

Update:

$$\boldsymbol{\eta}_{p(y|\alpha,\sigma^2) \to \alpha} \xleftarrow{\varepsilon} G^{\text{N}}\left(\boldsymbol{\eta}_{\alpha \to p(y|\alpha,\sigma^2)}, \boldsymbol{\eta}_{\sigma^2 \to p(y|\alpha,\sigma^2)}; \begin{bmatrix} 1 \\ y \\ y^2 \end{bmatrix}\right)$$

$$\boldsymbol{\eta}_{p(y|\alpha,\sigma^2) \to \sigma^2} \xleftarrow{\varepsilon} G^{\text{IG1}}\left(\boldsymbol{\eta}_{\sigma^2 \to p(y|\alpha,\sigma^2)}, \boldsymbol{\eta}_{\alpha \to p(y|\alpha,\sigma^2)}; \begin{bmatrix} 1 \\ y \\ y^2 \end{bmatrix}\right)$$

Parameter Outputs: $\boldsymbol{\eta}_{p(y|\alpha,\sigma^2) \to \alpha}$, $\boldsymbol{\eta}_{p(y|\alpha,\sigma^2) \to \sigma^2}$.

Algorithm 6: *The inputs, updates and outputs of the Gaussian fragment.*

Algorithm 7 provides the update to the natural parameter vector

$$\boldsymbol{\eta}_{p(y|\alpha) \to \alpha} \quad \text{based on input} \quad \boldsymbol{\eta}_{\alpha \to p(y|\alpha)}$$

and depends on the function H_{logistic} defined at (S.3) in the online supplement.
Its derivation is given in Section S.2.7 of the online supplement.

Data Input: $y \in \{0,1\}, \quad 0 \leq \varepsilon < 1.$

Parameter Inputs: $\boldsymbol{\eta}_{p(y|\alpha) \to \alpha}$, $\boldsymbol{\eta}_{\alpha \to p(y|\alpha)}$.

Update:

$$\boldsymbol{\eta}_{p(y|\alpha) \to \alpha} \xleftarrow{\varepsilon} H_{\text{logistic}}(\boldsymbol{\eta}_{\alpha \to p(y|\alpha)}; y)$$

Parameter Output: $\boldsymbol{\eta}_{p(y|\alpha) \to \alpha}$.

Algorithm 7: *The inputs, updates and outputs of the logistic likelihood fragment.*

3.8 Probit Likelihood Fragment

The probit likelihood fragment corresponds to the specification

$$y|\alpha \sim \text{Bernoulli}\big(\Phi(\alpha)\big).$$

The factor of the fragment is

$$p(y|\alpha) = \exp\left[y\log\{\Phi(\alpha)\} + (1-y)\log\{1-\Phi(\alpha)\}\right].$$

As for the logistic likelihood fragment, we also assume (19) which implies that $m_{p(y|\alpha) \to \alpha}(\alpha)$ also takes the form (20). The fragment update is given in Algorithm 8, with justification deferred to Section S.2.8 of the online supplement. The function H_{probit} is defined in Section S.1.3 of the online supplement. Note that H_{probit} has the advantage of admitting a closed form expression. This is not the case for H_{logistic} and numerical integration is required for its evaluation.

Data Input: $y \in \{0,1\}, \quad 0 \le \varepsilon < 1.$

Parameter Inputs: $\boldsymbol{\eta}_{p(y|\alpha) \to \alpha}$, $\boldsymbol{\eta}_{\alpha \to p(y|\alpha)}$.

Update:

$$\boldsymbol{\eta}_{p(y|\alpha) \to \alpha} \xleftarrow{\varepsilon} H_{\text{probit}}(\boldsymbol{\eta}_{\alpha \to p(y|\alpha)}; y)$$

Parameter Outputs: $\boldsymbol{\eta}_{p(y|\alpha) \to \alpha}$.

Algorithm 8: *The inputs, updates and outputs of the probit likelihood fragment.*

3.9 Poisson Likelihood Fragment

The Poisson likelihood fragment matches

$$y|\alpha \sim \text{Poisson}\{\exp(\alpha)\}$$

and the factor of the fragment is

$$p(y|\alpha) = \exp\{y\alpha - e^{\alpha} - \log(y!)\}.$$

As for the logistic and Poisson likelihood fragments, we also assume (19) which implies that $m_{p(y|\alpha) \to \alpha}(\alpha)$ also takes the form (20). The fragment update is given in Algorithm 9 with the H_{Poisson} function defined at (S.3)

Section S.2.9 of the online supplement contains justification of Algorithm 9.

Data Input: $y \in \{0,1,2,\ldots\}, \quad 0 \le \varepsilon < 1.$

Parameter Inputs: $\boldsymbol{\eta}_{p(y|\alpha) \to \alpha}$, $\boldsymbol{\eta}_{\alpha \to p(y|\alpha)}$.

Update:

$$\boldsymbol{\eta}_{p(y|\alpha) \to \alpha} \xleftarrow{\varepsilon} H_{\text{Poisson}}(\boldsymbol{\eta}_{\alpha \to p(y|\alpha)}; y)$$

Parameter Outputs: $\boldsymbol{\eta}_{p(y|\alpha) \to \alpha}$.

Algorithm 9: *The inputs, updates and outputs of the Poisson likelihood fragment.*

4 Illustration

We now provide illustration via a generalized additive mixed model analysis. The data are from the Indonesian Children's Health Study (Sommer, 1982), corresponding to a cohort of 275 Indonesian children who are repeatedly examined. The response variable is

$$y_{ij} = \begin{cases} 1, & \text{if respiratory infection present in the } i\text{th child at the } j\text{th examination,} \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

for $1 \le i \le m$ and $1 \le j \le n_i$. For these data note that $m = 275$ and the $n_i \in \{1,\ldots,6\}$. Potential predictor variables are age, indicator of vitamin A deficiency, indicator of being female, height, indicator of being stunted and indicators for the number of clinic visits for each child. We let a_{ij} denote the age in years of the ith child at the jth examination. Consider the following Bayesian generalized additive mixed model:

$$
\begin{gathered}
y_{ij}|\beta_0,\boldsymbol{\beta_x},\beta_{\text{spl}},\boldsymbol{u}_{\text{grp}},\boldsymbol{u}_{\text{spl}} \stackrel{\text{ind.}}{\sim} \text{Bernoulli}\Big(\text{logit}^{-1}\big(\beta_0+u_{\text{grp},i}+\boldsymbol{\beta}_{\boldsymbol{x}}^T\boldsymbol{x}_{ij}+f(a_{ij})\big)\Big),\\[1ex]
f(a_{ij})\equiv\beta_{\text{spl}}\,a_{ij}+\sum_{k=1}^{K}u_{\text{spl},k}\,z_k(a_{ij})\text{ is a low-rank smoothing spline in }a_{ij},\\
\text{where }\{z_k(\cdot):1\le k\le K\}\text{ is a suitable spline basis,}\\[1ex]
\boldsymbol{\beta}\equiv\left[\begin{array}{c}\beta_0\\ \boldsymbol{\beta_x}\\ \beta_{\text{spl}}\end{array}\right]\sim N(\boldsymbol{\mu_\beta},\boldsymbol{\Sigma_\beta}),\quad \boldsymbol{u}\equiv\left[\begin{array}{c}\boldsymbol{u}_{\text{grp}}\\ \boldsymbol{u}_{\text{spl}}\end{array}\right]\Bigg|\sigma^2_{\text{grp}},\sigma^2_{\text{spl}}\sim N\left(\left[\begin{array}{c}\boldsymbol{0}\\ \boldsymbol{0}\end{array}\right],\left[\begin{array}{cc}\sigma^2_{\text{grp}}\boldsymbol{I}_m & \boldsymbol{0}\\ \boldsymbol{0} & \sigma^2_{\text{spl}}\boldsymbol{I}_K\end{array}\right]\right),\\[1ex]
\sigma^2_{\text{grp}}|A_{\text{grp}}\sim\text{Inverse-}\chi^2(1,1/A_{\text{grp}}),\quad \sigma^2_{\text{spl}}|a_{\text{spl}}\sim\text{Inverse-}\chi^2(1,1/a_{\text{spl}}),\\[1ex]
A_{\text{grp}}\sim\text{Inverse-}\chi^2(1,1/A^2_{\text{grp}}),\quad a_{\text{spl}}\sim\text{Inverse-}\chi^2(1,1/A^2_{\text{spl}}).
\end{gathered}
\tag{22}
$$

The 'grp' and 'spl' subscripting indicates whether the random effect vector and corresponding variance parameter is for the random subject intercept or for spline coefficients in the non-linear function of age. Let $\boldsymbol{y}$ denote the $N\times 1$ vector containing the y_{ij}, where $N\equiv\sum_{i=1}^m n_i$. Despite the common use of double subscript notation as in (21), it is more convenient to label the entries of $\boldsymbol{y}$ with a single subscript when it comes to fitting via expectation propagation. To avoid a notational clash we use y^s_ℓ, $1\le\ell\le N$, to denote the ℓth entry of $\boldsymbol{y}$. Let $d^{\boldsymbol{\beta}}$ be the number of rows in $\boldsymbol{\beta}$. For the Indonesian Children's Health Study Data application $d^{\boldsymbol{\beta}}=11$. Then let $\boldsymbol{X}$ by the $N\times d^{\boldsymbol{\beta}}$ matrix containing the predictor data. The random effects design matrix is $\boldsymbol{Z}=[\boldsymbol{Z}_{\text{grp}}\ \boldsymbol{Z}_{\text{spl}}]$ where

$$
\boldsymbol{Z}_{\text{grp}}\equiv\underset{1\le i\le m}{\text{blockdiag}}(\boldsymbol{1}_{n_i})\quad\text{and}\quad \boldsymbol{Z}_{\text{spl}}\equiv\left[\underset{1\le k\le K}{z_k(a_{ij})}\right]_{1\le j\le n_i,\ 1\le i\le m}.
$$

Then the likelihood can be written as

$$
y^s_\ell|\boldsymbol{\beta},\boldsymbol{u}\stackrel{\text{ind.}}{\sim}\text{Bernoulli}\Big(\text{logit}^{-1}\big((\boldsymbol{X\beta}+\boldsymbol{Zu})_\ell\big)\Big),\quad 1\le\ell\le N.
$$

Next, let $\boldsymbol{C}\equiv[\boldsymbol{X}\ \boldsymbol{Z}]$ so that

$$
\boldsymbol{X\beta}+\boldsymbol{Zu}=\boldsymbol{C}\left[\begin{array}{c}\boldsymbol{\beta}\\ \boldsymbol{u}\end{array}\right]
$$

and let $\boldsymbol{c}_\ell^T$ be the ℓth row of $\boldsymbol{C}$. Let $\boldsymbol{e}_r$ be the $(m+K)\times 1$ vector with rth entry equal to 1 and zeroes elsewhere for $1\le r\le m+K$. Lastly, let $\boldsymbol{E}_{d^{\boldsymbol{\beta}}}$ be the $(d^{\boldsymbol{\beta}}+m+K)\times d^{\boldsymbol{\beta}}$ matrix with the $d^{\boldsymbol{\beta}}\times d^{\boldsymbol{\beta}}$ identity matrix at the top and all other entries equal to zero. The joint density function of all random variables in the model is

$$
\begin{aligned}
&p(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u},\sigma^2_{\text{grp}},\sigma^2_{\text{spl}},A_{\text{grp}},a_{\text{spl}})\\
&\quad = p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u})p(\boldsymbol{\beta})p(\boldsymbol{u}|\sigma^2_{\text{grp}},\sigma^2_{\text{spl}})p(\sigma^2_{\text{grp}}|A_{\text{grp}})p(\sigma^2_{\text{spl}}|a_{\text{spl}})p(A_{\text{grp}})p(a_{\text{spl}})\\
&\quad = p(\boldsymbol{\beta})\left\{\prod_{\ell=1}^{N}p(y^s_\ell|\boldsymbol{\beta},\boldsymbol{u})\right\}\left\{\prod_{i=1}^{m}p(u_{\text{grp},i}|\sigma^2_{\text{grp}})\right\}\left\{\prod_{k=1}^{K}p(u_{\text{spl},k}|\sigma^2_{\text{spl}})\right\}p(\sigma^2_{\text{grp}}|A_{\text{grp}})\\
&\qquad\quad \times p(\sigma^2_{\text{spl}}|a_{\text{spl}})p(A_{\text{grp}})p(a_{\text{spl}})\\
&\quad = \left\{\int_{\mathbb{R}^{d^\beta}}p(\widetilde{\boldsymbol{\beta}})\,\delta\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{E}^T_{d^\beta}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\right)\,d\widetilde{\boldsymbol{\beta}}\right\}\left\{\prod_{\ell=1}^{N}\int_{-\infty}^{\infty}p(y^s_\ell|\,\alpha_\ell)\delta\left(\alpha_\ell-\boldsymbol{c}^T_\ell\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\right)\,d\alpha_\ell\right\}\\
&\qquad \times\left\{\prod_{i=1}^{m}\int_{-\infty}^{\infty}p(\widetilde{u}_{\text{grp},i}|\,\sigma^2_{\text{grp}})\delta\left(\widetilde{u}_{\text{grp},i}-\boldsymbol{e}^T_{d^\beta+i}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\right)\,d\widetilde{u}_{\text{grp},i}\right\}p(\sigma^2_{\text{grp}}|A_{\text{grp}})p(A_{\text{grp}})\\
&\qquad \times\left\{\prod_{k=1}^{K}\int_{-\infty}^{\infty}p(\widetilde{u}_{\text{spl},k}|\,\sigma^2_{\text{grp}})\delta\left(\widetilde{u}_{\text{spl},k}-\boldsymbol{e}^T_{d^\beta+m+k}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\right)\,d\widetilde{u}_{\text{spl},k}\right\}p(\sigma^2_{\text{spl}}|a_{\text{spl}})p(a_{\text{spl}}).
\end{aligned}
\tag{23}
$$

Figure 6 is the derived variable factor graph corresponding to the representation of the joint density function given in (23). All of the fragments in Figure 6 are versions of fundamental fragments listed in Table 2 and are color-coded and numbered accordingly. Expectation propagation inference for this model and data involves iteratively passing messages between neighboring nodes on the Figure 6 factor graph. The parameter updates for the factor to stochastic node messages are given by the relevant algorithms in Section 3. The stochastic node to factor message parameter updates are a simple consequence of (8).

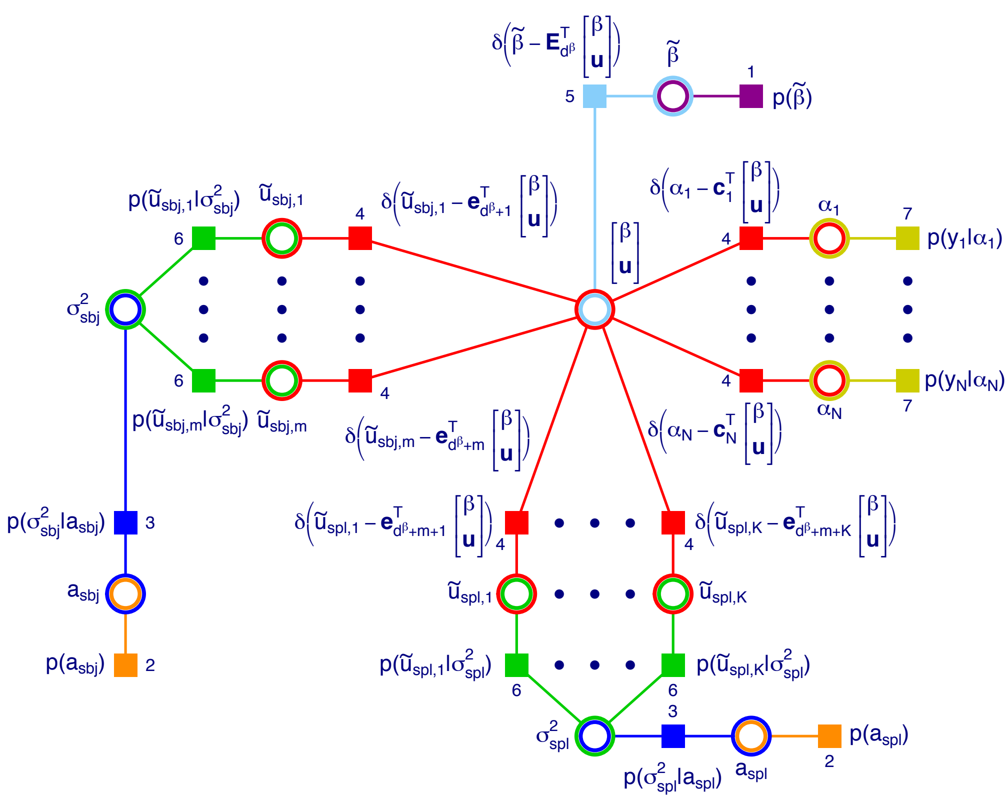


Figure 6: *Derived variable factor graph corresponding to the representation of the joint density function of random variables in the generalized additive mixed model (22) given by (23).*

We fit (22) using 1,000 iterations of expectation propagation message passing on the factor graph of Figure 6. We also conducted Markov chain Monte Carlo fitting via the func-

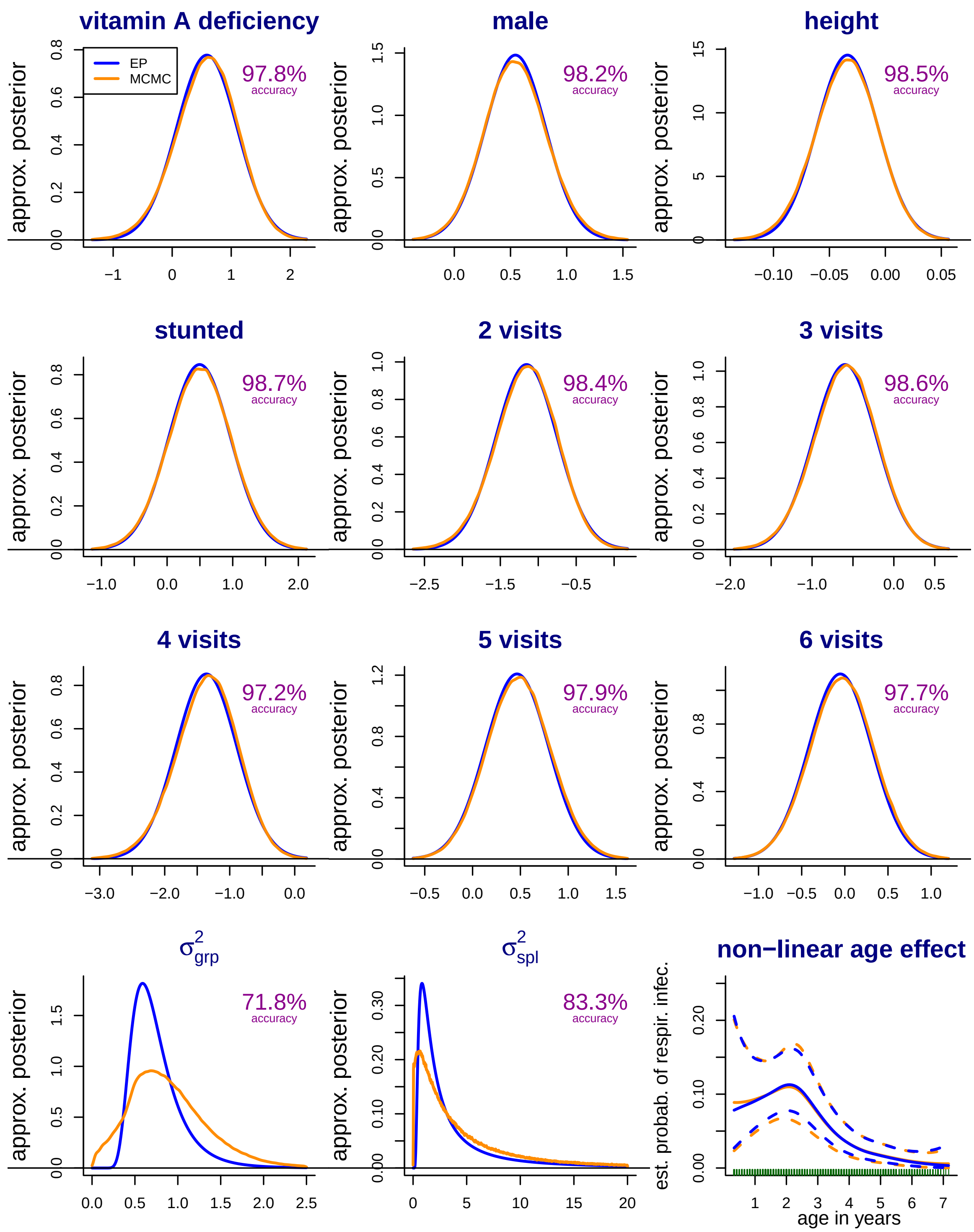


Figure 7: *Comparison of two approximate Bayesian inference methods, expectation propagation and Markov chain Monte Carlo, for model (22) applied to the Indonesian Children's Health Study Data. The first three rows compares approximate posterior density functions for the fixed effects parameters. The heading at the top of the panel is the corresponding predictor. The first two panels in the fourth row compares approximate posterior density functions for the two variance parameters. The accuracy percentages correspond to the definition at (24). The bottom right panel compares the low-rank smoothing spline fits for the non-linear age effect on the probability of respiratory infection with all other predictors set to their averages. In this panel, the dashed curves indicate pointwise 95% credible intervals and the tick marks show the age data.*

tion `stan()` in the R package rstan (Guo, Gabry & Goodrich, 2017), which interfaces the Stan language (Carpenter *et al.*, 2017), with a warmup size of 50,000 and a retained sample size of 1,000,000. The hyperparameters were set to $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = 10^{10}\boldsymbol{I}$, $\sigma_{\text{grp}} = \sigma_{\text{spl}} = 10^5$ with continuous variables standardized for the analyses and then results transformed to correspond to the original units. Figure 7 compares the Bayesian inference arising from the two approaches. The first three rows compare the expectation and Markov chain Monte Carlo approximate posterior density functions for the fixed effects parameters. The last row contains similar comparisons for the variance parameters and the low-rank smoothing spline fits for the non-linear age effect. The estimated probability functions are such that all other predictors are set at their average values, and are accompanied by pointwise 95% credible intervals.

The posterior density function comparisons are accompanied by accuracy percentages. For a generic parameter θ, the accuracy of the approximation $q(\theta)$ to the posterior density function $p(\theta|\boldsymbol{y})$ is given by

$$\text{accuracy} \equiv\ 100\left\{1 - \tfrac{1}{2}\int_{-\infty}^{\infty}\big|q(\theta) - p(\theta|\boldsymbol{y})\big|\,d\theta\right\}\%. \tag{24}$$

The Markov chain Monte Carlo-based posterior density functions, as well as the accuracy percentages on which they depend, are binned kernel density estimates obtained using the R function `bkde()` in the package KernSmooth (Wand & Ripley, 2015) with direct plug-in bandwidth selection via the function `dpik()`. The density estimates should be very close to the actual posterior density functions since they are based on one million posterior draws.

We see from Figure 7 that expectation propagation achieves excellent accuracy for the fixed effect parameters, in keeping with the simulation studies of Kim & Wand (2017). The variance parameter posterior density estimates are not as good for this particular example with accuracy scores of about 72% and 83%. Such mediocre accuracy was not apparent in the Kim & Wand (2017) simulations although their Figures 9 and 11 show accuracies for variance parameters being substantially lower than that those for fixed effect parameters. We ran the code that produced Figure 7 on some simulated data and got accuracy scores in the 85%-95% range for the variance parameters. Further research is needed to gain a fuller understanding of the accuracy of expectation propagation in the generalized additive mixed model context corresponding to this example.

5 More Elaborate Expectation Propagation Fragments

The fragments listed in Table 2 and covered in Section 3 are the most fundamental ones for generalized, linear and mixed models. Whilst these fragments support expectation propagation fitting of a wide range of models, additional fragments are needed for various elaborations. We now illustrate this fact by investigating fragments needed for (a) the extension to multivariate random effects, and (b) models where the response variable is modeled according to the t distribution. As we will see, expectation propagation is quite numerically challenging for such extensions.

5.1 Multivariate Random Effects

The fragments in Table 2 can handle the univariate random effects structure

$$u|\sigma^2 \sim N(0, \sigma^2)$$

but they do not cover the *multivariate* random effects extension:

$$\boldsymbol{u}|\boldsymbol{\Sigma} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is a unstructured $d^{\boldsymbol{u}} \times d^{\boldsymbol{u}}$ matrix.

The fragment corresponding to the factor

$$p(\boldsymbol{u}|\boldsymbol{\Sigma}) = (2\pi)^{-d^{\boldsymbol{u}}/2}|\boldsymbol{\Sigma}|^{-1/2}\exp(-\tfrac{1}{2}\,\boldsymbol{u}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{u})$$

is shown in Figure 8.

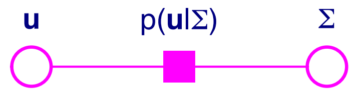


Figure 8: *The factor graph fragment corresponding the factor* $p(\boldsymbol{u}|\boldsymbol{\Sigma})$.

Under the usual conjugacy constraints, the message from $p(\boldsymbol{u}|\boldsymbol{\Sigma})$ to $\boldsymbol{\Sigma}$ is

$$m_{p(\boldsymbol{u}|\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) = \frac{\text{proj}_{\text{IW}}\big[m_{\boldsymbol{\Sigma}\,\rightarrow\,p(\boldsymbol{u}|\boldsymbol{\Sigma})}(\boldsymbol{\Sigma})\int_{\mathbb{R}^{d^{\boldsymbol{u}}}}p(\boldsymbol{u}|\boldsymbol{\Sigma})m_{\boldsymbol{u}\,\rightarrow\,p(\boldsymbol{u}|\boldsymbol{\Sigma})}(\boldsymbol{u})\,d\boldsymbol{u}/Z\big]}{m_{\boldsymbol{\Sigma}\,\rightarrow\,p(\boldsymbol{u}|\boldsymbol{\Sigma})}(\boldsymbol{\Sigma})} \tag{25}$$

where proj_{IW} denotes projection onto the $d^{\boldsymbol{u}}$-dimensional Inverse Wishart family. The messages on the right-hand side of (25) have the form

$$m_{\boldsymbol{\Sigma}\,\rightarrow\,p(\boldsymbol{u}|\boldsymbol{\Sigma})}(\boldsymbol{\Sigma}) = \exp\left\{\begin{bmatrix}\log|\boldsymbol{\Sigma}|\\ \text{vec}(\boldsymbol{\Sigma}^{-1})\end{bmatrix}^T\boldsymbol{\eta}_{\boldsymbol{\Sigma}\,\rightarrow\,p(\boldsymbol{u}|\boldsymbol{\Sigma})}\right\}$$

and

$$m_{\boldsymbol{u}\,\rightarrow\,p(\boldsymbol{u}|\boldsymbol{\Sigma})}(\boldsymbol{u}) = \exp\left\{\begin{bmatrix}\boldsymbol{u}\\ \text{vec}(\boldsymbol{u}\boldsymbol{u}^T)\end{bmatrix}^T\boldsymbol{\eta}_{\boldsymbol{u}\,\rightarrow\,p(\boldsymbol{u}|\boldsymbol{\Sigma})}\right\}$$

Introducing the shorthand

$$\boldsymbol{\eta}^{\heartsuit} \equiv \boldsymbol{\eta}_{\boldsymbol{\Sigma}\,\rightarrow\,p(\boldsymbol{u}|\boldsymbol{\Sigma})} \quad \text{and} \quad \boldsymbol{\eta}^{\diamondsuit} \equiv \boldsymbol{\eta}_{\boldsymbol{u}\,\rightarrow\,p(\boldsymbol{u}|\boldsymbol{\Sigma})},$$

arguments analogous to those given in Appendix 6 lead to the function of $\boldsymbol{\Sigma}$ inside the $\text{proj}_{\text{IW}}[\,\cdot\,]$ in (25) being proportional to

$$\begin{aligned}&|\boldsymbol{\Sigma}|^{\eta_1^{\heartsuit}}|2\boldsymbol{\Sigma}\text{vec}^{-1}(\boldsymbol{\eta}_2^{\diamondsuit})-\boldsymbol{I}|^{-1/2}\text{tr}\{\boldsymbol{\Sigma}^{-1}\text{vec}^{-1}(\boldsymbol{\eta}_2^{\heartsuit})\}\\ &\quad\times\exp\Big[-\tfrac{1}{4}(\boldsymbol{\eta}_1^{\diamondsuit})^T\{2\boldsymbol{\Sigma}\text{vec}^{-1}(\boldsymbol{\eta}_2^{\diamondsuit})-\boldsymbol{I}\}\{\text{vec}^{-1}(\boldsymbol{\eta}_2^{\diamondsuit})\}^{-1}\boldsymbol{\eta}_1^{\diamondsuit}\Big].\end{aligned} \tag{26}$$

The next step is to compute $E\{\log|\boldsymbol{\Sigma}|\}$ and $E(\boldsymbol{\Sigma}^{-1})$ with expectation with respect to the density function obtained by normalizing (26). This is a particularly challenging numerical problem since it involves numerical integration of the cone of $d^{\boldsymbol{u}} \times d^{\boldsymbol{u}}$ symmetric positive definite matrices. Then

$$\boldsymbol{\eta}_{p(\boldsymbol{u}|\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{\Sigma}} = (\nabla A)_{\text{IW}}^{-1}\left(\begin{bmatrix}E\{\log|\boldsymbol{\Sigma}|\}\\ E\{\text{vech}(\boldsymbol{\Sigma}^{-1})\}\end{bmatrix}\right). \tag{27}$$

Note that the function $(\nabla A)_{\text{IW}}$ admits the explicit form

$$(\nabla A)_{\text{IW}}\left(\begin{bmatrix}\eta_1\\ \boldsymbol{\eta}_2\end{bmatrix}\right) = \begin{bmatrix}\log\Big|-\text{vech}^{-1}(\boldsymbol{\eta}_2)\Big| - \displaystyle\sum_{j=1}^{d^{\boldsymbol{u}}}\text{digamma}\big(-\eta_1-\tfrac{1}{2}(d^{\boldsymbol{u}}+j)\big)\\ \{\eta_1+\tfrac{1}{2}(d^{\boldsymbol{u}}+1)\}\text{vech}[\{\text{vech}^{-1}(\boldsymbol{\eta}_2)\}^{-1}]\end{bmatrix}$$

where $[\eta_1\ \boldsymbol{\eta}_2^T]^T$ is the partition of the natural parameter vector into the first entry (η_1) and the remaining $\frac{1}{2}d^{\boldsymbol{u}}(d^{\boldsymbol{u}}+1)$ entries ($\boldsymbol{\eta}_2$). However, evaluation of (27) involves numerical inversion of $(\nabla A)_{\text{IW}}$ in $\{1+\frac{1}{2}d^{\boldsymbol{u}}(d^{\boldsymbol{u}}+1)\}$-dimensional space.

In conclusion, literal application of expectation propagation for multivariate random effects is quite daunting and effective implementation for even $2 \le d^{\boldsymbol{u}} \le 5$ is a very challenging numerical problem.

5.2 t Likelihood

The Gaussian fragment, treated in Section 3.6, corresponds to the specification $y|\alpha,\sigma^2 \sim N(\alpha,\sigma^2)$. Now consider the extension to the t distribution:

$$y|\alpha,\sigma^2,\nu \sim t(\alpha,\sigma^2,\nu) \tag{28}$$

where $\nu > 0$ is the degrees of freedom parameter. Low values of ν correspond to heavy-tailed distributions. The Gaussian likelihood is the $\nu \to \infty$ limiting case. The density function corresponding to (28) is

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

One could work with this density function in the expectation propagation message equations (8) and (9), but trivariate numerical integration is required. In other Bayesian computation contexts such as Markov chain Monte Carlo (e.g. Verdinelli & Wasserman, 1991) and variational message passing (e.g. McLean & Wand, 2018) it is common to replace (28) by the auxiliary variable representation

$$y|\alpha,\sigma^2,a \sim N(\alpha,a\sigma^2), \quad a|\nu \sim \text{Inverse-}\chi^2(\nu,\nu) \tag{29}$$

to aid tractability. Expectation propagation also benefits from this representation of the t-likelihood specification. The fragments corresponding to the factor product

$$p(y|\alpha,\sigma^2,a)\,p(a|\nu)\,p(\nu) \tag{30}$$

are shown in Figure 9.

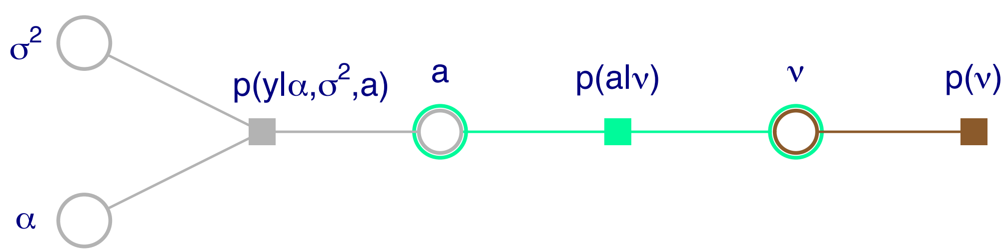


Figure 9: *Color-coded fragments corresponding to the factors $p(y|\alpha,\sigma^2,a)$, $p(a|\nu)$ and $p(\nu)$ appearing in (30).*

None of these fragments are among those treated in Section 3. Therefore, extension to t likelihood models requires expectation propagation updates for these three new fragments. Unfortunately, as we will see, difficult numerical challenges arise for these updates. We now focus on each one in turn.

5.2.1 The $p(y|\alpha,\sigma^2,a)$ Fragment

The factor for this fragment is

$$p(y|\alpha,\sigma^2,a) = (2\pi\,a\sigma^2)^{-1/2}\exp\left\{-\frac{(y-\alpha)^2}{2\,a\sigma^2}\right\}.$$

Conjugacy considerations dictate the assumption:

> all messages passed to α from factors outside of the fragment are in the Univariate Normal family and all messages passed to either a of σ^2 from factors outside of the fragment are in the Inverse Chi-Squared family. (31)

This leads to the factor to stochastic node messages taking the forms:

$$m_{p(y|\alpha,\sigma^2,a)\,\rightarrow\,\alpha}(\alpha) = \exp\left\{\begin{bmatrix}\alpha\\ \alpha^2\end{bmatrix}^T \boldsymbol{\eta}_{p(y|\alpha,\sigma^2,a)\,\rightarrow\,\alpha}\right\},$$

$$m_{p(y|\alpha,\sigma^2,a)\,\rightarrow\,\sigma^2}(\sigma^2) = \exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma^2\end{bmatrix}^T \boldsymbol{\eta}_{p(y|\alpha,\sigma^2,a)\,\rightarrow\,\sigma^2}\right\}$$

and

$$m_{p(y|\alpha,\sigma^2,a)\,\rightarrow\,a}(a) = \exp\left\{\begin{bmatrix}\log(a)\\ 1/a\end{bmatrix}^T \boldsymbol{\eta}_{p(y|\alpha,\sigma^2,a)\,\rightarrow\,a}\right\}.$$

The derivations of the natural parameter updates are similar in nature to those given in Appendix S.2.6 for the Gaussian fragment. However, the form $a\sigma^2$ (rather than σ^2) in the variance means that the natural parameter updates require evaluation of the bivariate integral-defined function

$$\mathcal{B}_2(p,q_1,q_2,r_1,r_2,s,t,u) \equiv$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{x_1^p\exp\{q_1x_1+q_2x_2-r_1e^{x_1}-r_2e^{x_2}-se^{x_1+x_2}/(t+e^{x_1+x_2})\}}{(t+e^{x_1+x_2})^u}\,dx_1\,dx_2$$

for

$$p\ge 0,\ q_1,q_2\in\mathbb{R},\ r_1,r_2>0,\ s\ge 0,\ t>0,\ u>0$$

rather than the univariate integral-defined function $\mathcal{B}(p,q,r,s,t,u)$ given by (S.1).

5.2.2 The $p(a|\nu)$ Fragment

The relevant factor is

$$p(a|\nu) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\,a^{-(\nu/2)-1}\,\exp\{-(\nu/2)/a\},\quad a,\nu>0.$$

Let $\upsilon\equiv\nu/2$ be a simple linear transformation of ν. For the remainder of this section we work with υ, rather than ν, since it leads to a simpler exposition. Now note that

$$p(a|\upsilon)\propto\begin{cases}\exp\left\{\begin{bmatrix}\log(a)\\ \\ 1/a\end{bmatrix}^T\begin{bmatrix}-\upsilon\\ \\ -\upsilon-1\end{bmatrix}\right\} & \text{as a function of } a,\\ \\ \exp\left\{\begin{bmatrix}\upsilon\log(\upsilon)-\log\{\Gamma(\upsilon)\}\\ \\ \upsilon\end{bmatrix}^T\begin{bmatrix}1\\ \\ -1/a-\log(a)\end{bmatrix}\right\} & \text{as a function of } \upsilon.\end{cases}$$

To ensure conjugacy we should then impose the restriction:

all messages passed to a from factors outside of the fragment are in the Inverse Chi-Squared family and all messages passed to either υ from factors outside of the fragment are in the Moon Rock family. (32)

The definition of the Moon Rock family is given in Table 1. The messages passed from $p(a|\upsilon)$ are then of the form

$$m_{p(a|\upsilon)\,\rightarrow\,a}(a) = \exp\left\{\begin{bmatrix}\log(a)\\ \\ 1/a\end{bmatrix}^T \boldsymbol{\eta}_{p(a|\upsilon)\,\rightarrow\,a}\right\}$$

and

$$m_{p(a|\upsilon)\,\rightarrow\,\upsilon}(\upsilon) = \exp\left\{\left[\begin{array}{c}\upsilon\log(\upsilon)-\log\{\Gamma(\upsilon)\}\\ \upsilon\end{array}\right]^T \boldsymbol{\eta}_{p(a|\upsilon)\,\rightarrow\,\upsilon}\right\}.$$

The message $m_{p(a|\upsilon)\,\rightarrow\,a}(a)$ has a treatment similar to that for $m_{p(\sigma^2|a)\,\rightarrow\,\sigma^2}(\sigma^2)$ and $m_{p(\sigma^2|a)\,\rightarrow\,a}(a)$ in Appendix S.2.3 and $m_{p(by|\alpha,\sigma^2)\,\rightarrow\,\sigma^2}(\sigma^2)$ in Appendix S.2.6 with projection onto the Inverse Chi-Squared family, although bivariate numerical integration is required. On the other hand,

$$m_{p(a|\upsilon)\,\rightarrow\,\upsilon}(\upsilon) = \frac{\text{proj}_{\text{MR}}\Big[m_{\upsilon\,\rightarrow\,p(a|\upsilon)}(\upsilon)\int_0^\infty p(a|\upsilon)\,m_{p(a|\upsilon)\,\rightarrow\,a}(a)\,da/Z\Big]}{m_{\upsilon\,\rightarrow\,p(a|\upsilon)}(\upsilon)}.$$

where proj_{MR} denotes projection onto the Moon Rock family. The function of υ inside the $\text{proj}_{\text{MR}}[\,]$ is proportional to

$$h(\upsilon)\equiv\{\upsilon^\upsilon/\Gamma(\upsilon)\}^{\eta_1^\flat+1}e^{\eta_2^\flat\upsilon}\Gamma(\upsilon-\eta_1^\sharp)/(\upsilon+\eta_2^\sharp)^{\upsilon-\eta_1^\sharp}$$

where

$$\boldsymbol{\eta}^\sharp\equiv\boldsymbol{\eta}_{p(a|\upsilon)\,\rightarrow\,a}\quad\text{and}\quad\boldsymbol{\eta}^\flat\equiv\boldsymbol{\eta}_{\upsilon\,\rightarrow\,p(a|\upsilon)}.$$

Then

$$\boldsymbol{\eta}_{p(a|\upsilon)\,\rightarrow\,\upsilon} = (\nabla A_{\text{MR}})^{-1}\left(\left[\begin{array}{c}\int_0^\infty\{\upsilon\,\log(\upsilon)-\log\Gamma(\upsilon)\}h(\upsilon)\,d\upsilon\big/\int_0^\infty h(\upsilon)\,d\upsilon\\ \int_0^\infty\,\upsilon\,h(\upsilon)\,d\upsilon\big/\int_0^\infty h(\upsilon)\,d\upsilon\end{array}\right]\right)$$

where

$$A_{\text{MR}}(\boldsymbol{\eta})\equiv\log\left[\int_0^\infty\{t^t/\Gamma(t)\}^{\eta_1}\exp(\eta_2\,t)\,dt\right]$$

is the log-partition function of the Moon Rock exponential family. This implies that

$$(\nabla A_{\text{MR}})(\boldsymbol{\eta}) = \left[\begin{array}{c}\int_0^\infty\{t\log(t)-\log\Gamma(t)\}\{t^t/\Gamma(t)\}^{\eta_1}\exp(\eta_2\,t)\,dt\Big/\int_0^\infty\{t^t/\Gamma(t)\}^{\eta_1}\exp(\eta_2\,t)\,dt\\ \int_0^\infty\,t\{t^t/\Gamma(t)\}^{\eta_1}\exp(\eta_2\,t)\,dt\Big/\int_0^\infty\{t^t/\Gamma(t)\}^{\eta_1}\exp(\eta_2\,t)\,dt\end{array}\right].$$

This particular exponential family is not well-studied and we are not aware of any published theory concerning the properties of ∇A_{MR} and $(\nabla A_{\text{MR}})^{-1}$. Standard analytic arguments can be used to show that the domain of ∇A_{MR} is

$$H=\left\{\left[\begin{array}{c}\eta_1\\ \eta_2\end{array}\right]:\eta_1\ge 0,\ \eta_1+\eta_2<0\right\}.$$

It is conjectured that the image of H under ∇A_{MR} is

$$T=\left\{\left[\begin{array}{c}\tau_1\\ \tau_2\end{array}\right]:\tau_1<\tau_2\log(\tau_2)-\log\Gamma(\tau_2)\right\}.$$

Figure 10 shows the domain of ∇A_{MR} and the conjectured domain of $(\nabla A_{\text{MR}})^{-1}$ as well as some example mappings between the two spaces.

Evaluation of $(\nabla A_{\text{MR}})^{-1}$ is a non-trivial problem. It requires numerical inversion techniques such as Newton-Raphson iteration. Moreover, each of the iterative updates involves evaluation of ∇A_{MR} and, possibly, its first partial derivatives. None of these functions are available in closed form and require numerical integration.

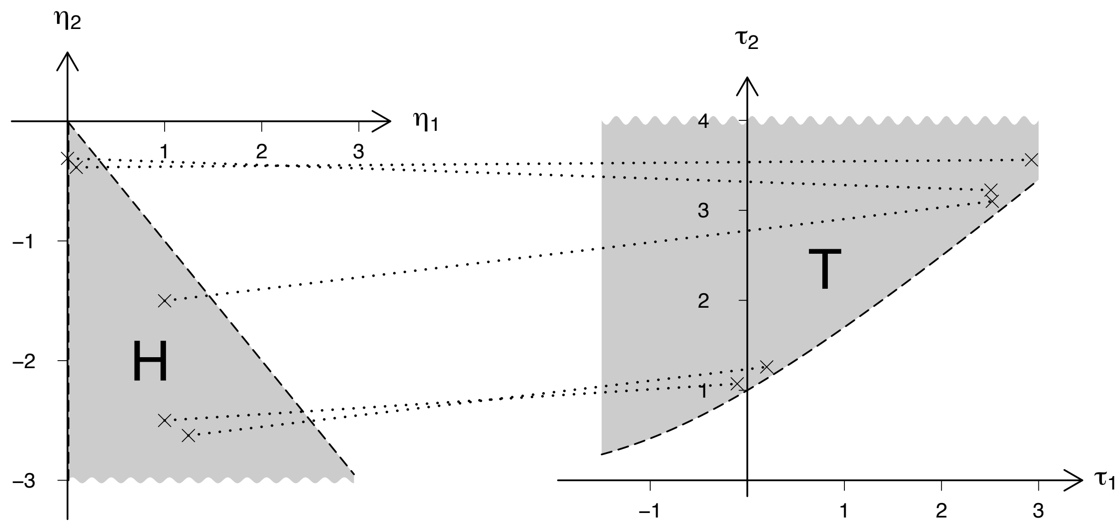


Figure 10: *Illustration of the bijective maps between H and T for the Moon Rock exponential family. The crosses and dotted lines depict five example $\boldsymbol{\eta} \in H$ and $\boldsymbol{\tau} = \nabla A_{MR}(\boldsymbol{\eta}) \in T$ pairs. Since ∇A_{MR} is a bijective map, the crosses and dotted lines equivalently depict five example $\boldsymbol{\tau} \in T$ and $\boldsymbol{\eta} = (\nabla A_{MR})^{-1}(\boldsymbol{\tau}) \in H$ pairs.*

5.2.3 The $p(\nu)$ Fragment

This simple fragment has the factor to stochastic node message

$$m_{p(\nu)\,\rightarrow\,\nu}(\nu) \propto p(\nu)$$

corresponding to the prior distribution on ν. The conjugate family of prior density functions is

$$p(\nu) \propto \{(\nu/2)^{\nu/2}/\Gamma(\nu/2)\}^{A_\nu}\ \exp(-\tfrac{1}{2}\,B_\nu\nu), \quad \nu > 0.$$

for hyperparameters $A_\nu \geq 0$ and $B_\nu > A_\nu$.

In terms of $\upsilon = \nu/2$, the relevant message is

$$m_{p(\upsilon)\,\rightarrow\,\upsilon}(\upsilon) = \exp\left\{\begin{bmatrix} \upsilon\log(\upsilon) - \log\{\Gamma(\upsilon)\} \\ \upsilon \end{bmatrix}^T \begin{bmatrix} A_\nu \\ -B_\nu \end{bmatrix}\right\}.$$

5.3 Summary of Numerical Challenges

The previous two subsections make it clear that elaborations such as multivariate random effects and fancier likelihoods involve profound numerical challenges for the expectation propagation paradigm. Table 3 summarizes the numerical challenges of all of the non-trivial fragments treated in this article.

The first ten fragments in Table 3 have the attraction of requiring only numerical evaluation of univariate integral within the families given by (S.1) and (S.2). The probit likelihood fragment stands out as a special case of a likelihood that does not require any numerical methods for expectation propagation message passing.

The last three fragments of Table 3 are considerably more demanding in terms of numerical analysis. In a recent article, Gelman *et al.* (2017) discuss the possibility of adopting Monte Carlo methods to deal with difficult computational problems in expectation propagation, but we are not aware of any existing methodology of this type.

fragment name	numeric. integrat. demands	Kull.-Leib. projec. demands
Gaussian prior	none	none
Inverse Wishart prior	none	none
Moon Rock prior	none	none
Iterated Inverse Chi Squared	univariate quadrature	inversion of $\log -$digamma
Linear comb. deriv. var.	none	none
Multiv. lin. comb. deriv. var.	none	none
Gaussian	univariate quadrature	inversion of $\log -$digamma
Logistic likelihood	univariate quadrature	trivial
Probit likelihood	none	trivial
Poisson likelihood	univariate quadrature	trivial
Multiple random effects	multivariate quadrature	inversion of a multivariate function
t likelihood direct	trivariate quadrature	inversion of $\log -$digamma and a non-explicit bivariate function
t likelihood aux. var.	bivariate quadrature	inversion of $\log -$digamma and a non-explicit bivariate function

Table 3: *The numerical integration and Kullback-Leibler projection demands of the non-trivial fragments discussed in this article.*

Acknowledgments

This research was supported by Australian Research Council Discovery Project DP180100597.

References

Azzalini, A. (2017). The R package `sn`: The skew-normal and skew-t distributions (version 1.5). `http://azzalini.stat.unipd.it/SN`

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, **76**, Issue 1, 1–32.

Gelman, A., Vehtari, A., Jylänki, P., Sivula, T., Tran, D., Suhai, S., Blomstedt, P. Cunningham, J.P., Schiminovic, D. and Robert, C. (2017). Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data. Unpublished manuscript (`arXiv:1412.4869v2`).

Guo, B.-N. and Qi, F. (2013). Refinements of lower bounds for polygamma functions. *Proceedings of the American Mathematical Society*, **141**, 1007–1015.

Guo, J., Gabry, J. and Goodrich, B. (2017). The R package rstan: R interface to Stan. R package (version 2.17.2). `http://mc-stan.org`.

Harville, D.A. (2008). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Kim, A.S.I. and Wand, M.P. (2016). The explicit form of expectation propagation for a simple statistical model. *Electronic Journal of Statistics*, **10**, 550–581.

Kim, A.S.I. and Wand, M.P. (2017). On expectation propagation for generalized, linear and mixed models. *Australian and New Zealand Journal of Statistics*, **59**, in press.

Minka, T. (2005). Divergence measures and message passing. *Microsoft Research Technical Report Series*, **MSR-TR-2005-173**, 1–17.

McLean, M.W. and Wand, M.P. (2018). Variational message passing for elaborate response regression models. *Bayesian Analysis*, under revision.

Nolan, T.H. and Wand, M.P. (2017). Accurate logistic variational message passing: algebraic and numerical details. *Stat*, **6**, 102–112.

Smyth, G. (2015). The R package `statmod`: Statistical modeling (version 1.4). `http://http://cran.r-project.org`

Sommer, A. (1982). *Nutritional Blindness*. New York: Oxford University Press.

Wainwright, M.J. and Jordan, M.I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, **1**, 1–305.

Wand, M.P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing (with discussion). *Journal of the American Statistical Association*, **112**, 137–168.

Wand, M.P. and Jones, M.C. (1993). Comparison of smoothing parameterizations in bivariate density estimation. *Journal of the American Statistical Association*, **88**, 520–528.

Wand, M.P. and Ripley, B.D. (2015). The R package `KernSmooth`. Functions for kernel smoothing supporting Wand & Jones (1995) (version 2.23). `https://cran.R-project.org`.

Supplement for:

Factor Graph Fragmentization of Expectation Propagation

BY WILSON Y. CHEN AND MATT P. WAND

University of Technology Sydney

S.1 Function Definitions

Expectation propagation algorithms that are based on the fragments in Section 3 have straightforward implementation once a few key functions are identified. Many of the functions are simple but long-winded. However, they only have to be implemented once and after that all fragment updates are simple.

The functions can be divided into three types:

1. functions defined via non-analytic integral families.
2. a function defined via inversion of an established function
3. functions that are explicit given function types 1. and 2.

We now give details of each of these types in turn.

S.1.1 Functions Defined via Non-Analytic Integral Families

Two fundamental families of integrals for expectation propagation in linear model contexts are:

$$
\begin{aligned}
\mathcal{A}(p,q,r,s,t,u) &\equiv \int_{-\infty}^{\infty} \frac{x^p\,\exp(qx-rx^2)\,dx}{(x^2+sx+t)^u},\\
&\qquad p\geq 0,\ q\in\mathbb{R}, r>0, s\in\mathbb{R}, t>\tfrac{1}{4}\,s^2, u>0\\
\text{and}\quad \mathcal{B}(p,q,r,s,t,u) &\equiv \int_{-\infty}^{\infty} \frac{x^p\,\exp\{qx-re^x-se^x/(t+e^x)\}\,dx}{(t+e^x)^u},\\
&\qquad p\geq 0,\ q\in\mathbb{R}, r>0, s\geq 0, t>0, u>0.
\end{aligned}
\tag{S.1}
$$

An additional family of non-analytic functions that we need is:

$$
\mathcal{C}_b(p,q,r) \equiv \int_{-\infty}^{\infty} x^p\,\exp\{qx-rx^2-b(x)\}\,dx, \tag{S.2}
$$

where $q\in\mathbb{R}$, $r>0$ and $b:\mathbb{R}\to\mathbb{R}$ is any function for which $\mathcal{C}_b(p,q,r)$ exists.

To avoid underflow and overflow working with logarithms and suitably modified integrands is recommended. For example,

$$
\mathcal{C}(0,q,r) = e^M \int_{-\infty}^{\infty} \exp\{qx-rx^2-b(x)-M\}\,dx
$$

where $M\equiv\sup\{x\in\mathbb{R}: qx-rx^2-b(x)\}$ implies that

$$
\log\{\mathcal{C}(0,q,r)\} = M + \log\left[\int_{-\infty}^{\infty} \exp\{qx-rx^2-b(x)-M\}\,dx\right]
$$

Sine the last-written integrand has a maximum of 1, its values are not overly large or small.

S.1.2 Function Defined via Inversion

Theorem 1 of Kim & Wand (2016) asserts that the function $\log - \text{digamma}$ is a bijective mapping from $\mathbb{R}_+$ onto $\mathbb{R}_+$. Therefore its inverse

$$(\log - \text{digamma})^{-1} : \mathbb{R}_+ \to \mathbb{R}_+$$

is well-defined. Bounds given in Guo & Qi (2013) imply that

$$\frac{1}{2x} < (\log - \text{digamma})^{-1}(x) < \frac{1}{x} \quad \text{for all } x > 0.$$

Therefore,

$$(\log - \text{digamma})^{-1}(x) \approx \text{geometric mean of } \frac{1}{2x} \text{ and } \frac{1}{x} = \frac{1}{x\sqrt{2}}$$

and provides useful starting values for iterative inversion of $\log - \text{digamma}$.

In addition, care is required for evaluation of $(\log - \text{digamma})(x)$ for large x since direct computation round-off error can lead to an erroneous answer of zero. Software such as the function `logmdigamma()` in the R package statmod (Smyth, 2015).

S.1.3 Explicit Functions

The $N(0, 1)$ density function and cumulative distribution functions are denoted by

$$\phi(x) \equiv (2\pi)^{-1} e^{-x^2/2} \quad \text{and} \quad \Phi(x) \equiv \int_{-\infty}^{x} \phi(t)\, dt.$$

We also define

$$\zeta(x) \equiv \log\{2\Phi(x)\} \quad \text{so that} \quad \zeta'(x) \equiv \phi(x)/\Phi(x).$$

Stable computation of ζ' is available from, for example, the function `zeta()` in the R package sn (Azzalini, 2017).

The functions G^{N} and G^{IG1} defined in Kim & Wand (2016) are also needed here. The function G^{IG2} from Kim & Wand (2016) requires generalization to handle Half-t priors on standard deviation parameters with arbitrary degrees of freedom. The generalization is denoted by G^{IG3}. Each of G^{N}, G^{IG1} and G^{IG3} depend on the functions defined in Appendices S.1.1 and S.1.2 but otherwise ther are simple, albeit long-winded, functions with multiple vector arguments. Their definitions are given in Section A.4 of Kim & Wand (2016) and are repeated here for convenience. We also define the explicit functions H_{probit}, H_{logistic} and H_{Poisson}. First set:

$$\alpha\left(k, \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}\right) = \mathcal{A}\left(k, a_1, -a_2, \frac{-2c_2}{c_1}, \frac{c_3 - 2b_2}{c_1}, \frac{c_1 - 2b_1 - 2}{2}\right)$$

and

$$\beta\left(k, \ell, v, w, \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}\right) =$$
$$\mathcal{B}\left(k, \frac{\ell + c_1 - 1}{2} - a_1, \frac{c_1 c_3 - c_2^2}{2c_1} - a_2, -b_2\left(\frac{c_2}{c_1} + \frac{b_1}{2b_2}\right)^2, v, w\right).$$

Next, define

$$g(\ell, v, w, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = (\log - \text{digamma})^{-1}\left(\log\left\{\frac{\beta(0, \ell+1, v, w, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}{\beta(0, \ell-1, v, w, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}\right\}\right.$$
$$\left. - \frac{\beta(1, \ell-1, v, w, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}{\beta(0, \ell-1, v, w, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}\right).$$

Now we are ready to give the expressions for G^{N}, G^{IG1}, G^{IG2}, and G^{IG3}:

$$G^N(\boldsymbol{a},\boldsymbol{b};\boldsymbol{c}) = \left[\frac{\alpha(2,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})}{\alpha(0,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})} - \left\{\frac{\alpha(1,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})}{\alpha(0,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})}\right\}^2\right]^{-1} \left[\begin{array}{c} \alpha(1,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})/\alpha(0,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) \\ -1/2 \end{array}\right] - \boldsymbol{a},$$

$$G^{\text{IG1}}\left(\boldsymbol{a},\left[\begin{array}{c} b_1 \\ b_2 \end{array}\right];\left[\begin{array}{c} c_1 \\ c_2 \\ c_3 \end{array}\right]\right) = \left[\begin{array}{c} -1 - g(0,-2b_2/c_1,\frac{1}{2},\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) \\ \dfrac{-g(0,-2b_2/c_1,\frac{1}{2},\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})\,\beta(0,-1,-2b_2/c_1,\frac{1}{2},\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})}{\beta(0,1,-2b_2/c_1,\frac{1}{2},\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})} \end{array}\right] - \boldsymbol{a}$$

and

$$G^{\text{IG3}}\left(\boldsymbol{a},\left[\begin{array}{c} b_1 \\ b_2 \end{array}\right];k,\ell\right) =$$

$$\left[\begin{array}{c} -1 - g\left(k-2,-b_2/\ell,\ell-k/2-b_1,\boldsymbol{a},\left[\begin{array}{c} 0 \\ b_2 \end{array}\right],\left[\begin{array}{c} 2 \\ 0 \\ 0 \end{array}\right]\right) \\ \\ \dfrac{\left\{\begin{array}{c} -g\left(k-2,-b_2/\ell,\ell-k/2-b_1,\boldsymbol{a},\left[\begin{array}{c} 0 \\ b_2 \end{array}\right],\left[\begin{array}{c} 2 \\ 0 \\ 0 \end{array}\right]\right) \\ \times\beta\left(0,k-3,-b_2/\ell,\ell-k/2-b_1,\boldsymbol{a},\left[\begin{array}{c} 0 \\ b_2 \end{array}\right],\left[\begin{array}{c} 2 \\ 0 \\ 0 \end{array}\right]\right) \end{array}\right\}}{\beta\left(0,k-1,-b_2/\ell,\ell-k/2-b_1,\boldsymbol{a},\left[\begin{array}{c} 0 \\ b_2 \end{array}\right],\left[\begin{array}{c} 2 \\ 0 \\ 0 \end{array}\right]\right)} \end{array}\right] - \boldsymbol{a}.$$

This definition is a generalization of the function G^{IG2} given in Kim & Wand (2016, 2017) and is such that

$$G^{\text{IG3}}\left(\boldsymbol{a},\left[\begin{array}{c} b_1 \\ b_2 \end{array}\right];k,1\right) = G^{\text{IG2}}\left(\boldsymbol{a},\left[\begin{array}{c} b_1 \\ b_2 \end{array}\right];k\right).$$

Put

$$H_b\left(\left[\begin{array}{c} a_1 \\ a_2 \end{array}\right];y\right) = \left[\begin{array}{c} \dfrac{\mathcal{C}_b(1,a_1+y,-a_2)/\mathcal{C}_b(0,a_1+y,-a_2)}{\dfrac{\mathcal{C}_b(2,a_1+y,-a_2)}{\mathcal{C}_b(0,a_1+y,-a_2)} - \dfrac{\mathcal{C}_b(1,a_1+y,-a_2)^2}{\mathcal{C}_b(0,a_1+y,-a_2)^2}} \\ \\ \dfrac{-1/2}{\dfrac{\mathcal{C}_b(2,a_1+y,-a_2)}{\mathcal{C}_b(0,a_1+y,-a_2)} - \dfrac{\mathcal{C}_b(1,a_1+y,-a_2)^2}{\mathcal{C}_b(0,a_1+y,-a_2)^2}} \end{array}\right] - \left[\begin{array}{c} a_1 \\ a_2 \end{array}\right]$$

for any $a_1 \in \mathbb{R}$, $a_2 < 0$ and $y \in \mathbb{R}$ and then let

$$\begin{aligned} H_{\text{logistic}}\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix};y\right) &\equiv H_b\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix};y\right) \quad \text{for} \quad b(x) = \log(1+e^x) \\ \text{and} \quad H_{\text{Poisson}}\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix};y\right) &\equiv H_b\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix};y\right) \quad \text{for} \quad b(x) = e^x. \end{aligned} \tag{S.3}$$

Lastly,

$$H_{\text{probit}}\left(\left[\begin{array}{c} a_1 \\ a_2 \end{array}\right];y\right) \equiv \frac{1}{1-2a_2-\zeta'(r)\{r+\zeta'(r)\}}\begin{bmatrix} a_1(1-2a_2) \\ +(2y-1)\zeta'(r)\sqrt{2a_2(2a_2-1)} \\ \\ a_2(1-2a_2) \end{bmatrix} - \begin{bmatrix} a_1 \\ \\ \\ a_2 \end{bmatrix}$$

where $r \equiv (2y-1)a_1/\sqrt{2a_2(2a_2-1)}$.

S.2 Derivations

We now provide derivations of each of algorithms given in Section 3. Definitions used in the derivations are:

$$A_{\mathrm{N}}(\boldsymbol{\eta}) \equiv -\tfrac{1}{4}(\eta_1^2/\eta_2) - \tfrac{1}{2}\log(-2\eta_2) \quad \text{and} \quad A_{\mathrm{I}\chi^2}(\boldsymbol{\eta}) \equiv (\eta_1+1)\log(-\eta_2) + \log\Gamma(-\eta_1-1)$$

for the log-partition functions of the Normal and Inverse Chi-Squared families respectively. In a similar vein, $\text{proj}_{\mathrm{N}}[\,\cdot\,]$ denotes Kullback-Leibler projection onto the (possibly Multivariate) Normal family of density functions and $\text{proj}_{\mathrm{I}\chi^2}[\,\cdot\,]$ denotes Kullback-Leibler projection onto the Inverse Chi-Squared family of density functions. We use Z to denote the normalizing factor of the function inside a Kullback-Leibler projection operator.

S.2.1 Derivation of Algorithm 1

Plugging into (9) we get

$$m_{p(\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\text{proj}_{\mathrm{N}}\Big[m_{\theta\,\to\,p(\boldsymbol{\theta})}(\boldsymbol{\theta})\,p(\boldsymbol{\theta})/Z\Big]}{m_{\theta\,\to\,p(\boldsymbol{\theta})}(\boldsymbol{\theta})} = \frac{m_{\theta\,\to\,p(\boldsymbol{\theta})}(\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{m_{\theta\,\to\,p(\boldsymbol{\theta})}(\boldsymbol{\theta})} = p(\theta).$$

The second equality follows from the fact that $m_{\theta\,\to\,p(\boldsymbol{\theta})}(\boldsymbol{\theta})$ is a Multivariate Normal density function, which is a consequence of (11). Since

$$p(\boldsymbol{\theta}) \propto \exp\big\{-\tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}})\big\} \propto \exp\left\{\begin{bmatrix}\boldsymbol{\theta}\\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}}\\ -\tfrac{1}{2}\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})\end{bmatrix}\right\}$$

we have

$$m_{p(\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}}(\boldsymbol{\theta}) = \exp\left\{\begin{bmatrix}\boldsymbol{\theta}\\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T\boldsymbol{\eta}_{p(\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}}\right\}$$

where

$$\boldsymbol{\eta}_{p(\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}} = \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}}\\ -\tfrac{1}{2}\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})\end{bmatrix}.$$

S.2.2 Derivation of Algorithm 2

Using arguments similar to those given in Section S.2.1 for the Gaussian prior fragment lead to

$$\begin{aligned}
m_{p(\boldsymbol{\Theta})\,\to\,\boldsymbol{\Theta}}(\boldsymbol{\Theta}) \quad &\propto \quad p(\boldsymbol{\Theta}) \propto |\boldsymbol{\Theta}|^{-(\kappa_{\boldsymbol{\Theta}}+d^{\Theta}+1)/2}\exp\{-\tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}\boldsymbol{\Theta}^{-1})\}\\
&= \quad \exp\left\{\begin{bmatrix}\log|\boldsymbol{\Theta}|\\ \text{vec}(\boldsymbol{\Theta}^{-1})\end{bmatrix}^T\begin{bmatrix}-\tfrac{1}{2}(\kappa_{\boldsymbol{\Theta}}+d^{\Theta}+1)\\ -\tfrac{1}{2}\text{vec}(\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}^{-1})\end{bmatrix}\right\}.
\end{aligned}$$

Hence

$$m_{p(\boldsymbol{\Theta})\,\to\,\boldsymbol{\Theta}}(\boldsymbol{\Theta}) = \exp\left\{\begin{bmatrix}\log|\boldsymbol{\Theta}|\\ \text{vec}(\boldsymbol{\Theta}^{-1})\end{bmatrix}^T\boldsymbol{\eta}_{p(\boldsymbol{\Theta})\,\to\,\boldsymbol{\Theta}}\right\}$$

where

$$\boldsymbol{\eta}_{p(\boldsymbol{\Theta})\,\to\,\boldsymbol{\Theta}} = \begin{bmatrix}-\tfrac{1}{2}(\kappa_{\boldsymbol{\Theta}}+d^{\Theta}+1)\\ -\tfrac{1}{2}\text{vec}(\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}^{-1})\end{bmatrix}.$$

S.2.3 Derivation of Algorithm 3

The message from $p(\sigma^2|a)$ to σ^2 is

$$m_{p(\sigma^2|a)\,\to\,\sigma^2}(\sigma^2) \longleftarrow \frac{\text{proj}_{\text{I}\chi^2}\left[m_{\sigma^2\,\to\,p(\sigma^2|a)}(\sigma^2)\int_0^\infty p(\sigma^2|a)\,m_{a\,\to\,p(\sigma^2|a)}(a)\,da/Z\right]}{m_{\sigma^2\,\to\,p(\sigma^2|a)}(\sigma^2)}. \quad \text{(S.4)}$$

Assumption (13) implies that

$$m_{a\,\to\,p(\sigma^2|a)}(a) = \exp\left\{\begin{bmatrix}\log(a)\\ 1/a\end{bmatrix}^T \boldsymbol{\eta}_{a\,\to\,p(\sigma^2|a)}\right\} = a^{\eta_1^\clubsuit}\exp(\eta_2^\clubsuit/a)$$

and

$$m_{\sigma^2\,\to\,p(\sigma^2|a)}(\sigma^2) = \exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma^2\end{bmatrix}^T \boldsymbol{\eta}_{\sigma^2\,\to\,p(\sigma^2|a)}\right\} = (\sigma^2)^{\eta_1^\spadesuit}\exp(\eta_2^\spadesuit/\sigma^2)$$

where

$$\boldsymbol{\eta}^\clubsuit \equiv \boldsymbol{\eta}_{a\,\to\,p(\sigma^2|a)} \quad \text{and} \quad \boldsymbol{\eta}^\spadesuit \equiv \boldsymbol{\eta}_{\sigma^2\,\to\,p(\sigma^2|a)}.$$

As a function of σ^2, the integral in (S.4) is

$$\begin{aligned}
&\int_0^\infty \frac{(\nu/a)^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-(\nu/2)-1}\exp\{-\nu/(\sigma^2 a)\}\,a^{\eta_1^\clubsuit}\exp(\eta_2^\clubsuit/a)\,da\\
&\qquad\propto (\sigma^2)^{-(\nu/2)-1}\int_0^\infty a^{-\{-\eta_1^\clubsuit+(\nu/2)-1\}-1}\exp[-\{(\nu/\sigma^2)-\eta_2^\clubsuit\}/a]\,da\\
&\qquad\propto (\sigma^2)^{-(\nu/2)-1}\{(\nu/\sigma^2)-\eta_2^\clubsuit\}^{\eta_1^\clubsuit-(\nu/2)+1}.
\end{aligned}$$

Therefore, the density function inside the $\text{proj}_{\text{I}\chi^2}[\,\cdot\,]$ is proportional to

$$(\sigma^2)^{\eta_1^\spadesuit-(\nu/2)-1}\{(\nu/\sigma^2)-\eta_2^\clubsuit\}^{\eta_1^\clubsuit-(\nu/2)+1}\exp(\eta_2^\spadesuit/\sigma^2)$$

and the numerator of (S.4) is proportional to

$$\exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma^2\end{bmatrix}^T \boldsymbol{\eta}_{\text{numer}}\right\}$$

where

$$\boldsymbol{\eta}_{\text{numer}} \equiv (\nabla A_{\text{I}\chi^2})^{-1}\left(\begin{bmatrix}\dfrac{\int_0^\infty \log(\sigma^2)(\sigma^2)^{\eta_1^\spadesuit-(\nu/2)-1}\{(\nu/\sigma^2)-\eta_2^\clubsuit\}^{\eta_1^\clubsuit-(\nu/2)+1}\exp(\eta_2^\spadesuit/\sigma^2)\,d\sigma^2}{\int_0^\infty (\sigma^2)^{\eta_1^\spadesuit-(\nu/2)-1}\{(\nu/\sigma^2)-\eta_2^\clubsuit\}^{\eta_1^\clubsuit-(\nu/2)+1}\exp(\eta_2^\spadesuit/\sigma^2)\,d\sigma^2}\\[2ex] \dfrac{\int_0^\infty (1/\sigma^2)(\sigma^2)^{\eta_1^\spadesuit-(\nu/2)-1}\{(\nu/\sigma^2)-\eta_2^\clubsuit\}^{\eta_1^\clubsuit-(\nu/2)+1}\exp(\eta_2^\spadesuit/\sigma^2)\,d\sigma^2}{\int_0^\infty (\sigma^2)^{\eta_1^\spadesuit-(\nu/2)-1}\{(\nu/\sigma^2)-\eta_2^\clubsuit\}^{\eta_1^\clubsuit-(\nu/2)+1}\exp(\eta_2^\spadesuit/\sigma^2)\,d\sigma^2}\end{bmatrix}\right).$$

Steps analogous to those given in Appendix A.5.4 of Kim & Wand (2016) can then be used to derive the $\boldsymbol{\eta}_{p(\sigma^2|a)\,\to\,\sigma^2}$ update. However, note that Algorithm 3 supports general Half-t(A,ν) prior distributions and Kim & Wand (2016) only deal with the $\nu = 1$ (Half-Cauchy) special case. Hence, the function G^{IG2} in Kim & Wand (2016) has to be generalized to the G^{IG3} function.

Also, note that the $\boldsymbol{\eta}_{p(\sigma^2|a)\,\to\,\sigma^2}$ update has $\xleftarrow{\varepsilon}$ rather than $\longleftarrow$ to allow for the damping adjustment defined by (10). The same adjustment applies to the remainder of the derivations given in Section S.2.

The message from $p(\sigma^2|a)$ to a has a similar form and arguments analogous to those just given for the message from $p(\sigma^2|a)$ to σ^2 lead to the update for $\boldsymbol{\eta}_{p(\sigma^2|a)\,\to\,a}$ given in Algorithm 3.

S.2.4 Derivation of Algorithm 4

Algorithm 4 is the $d^{\boldsymbol{\theta}} = 1$ special case of Algorithm 5 and therefore its derivation is covered by the general $d^{\boldsymbol{\theta}}$ case given next in Section S.2.5.

S.2.5 Derivation of Algorithm 5

Algorithm 5 depends on Result 2 below, which extends Theorem 1 of Kim & Wand (2017) to multivariate linear combination derived variables.

Let
$$p_{N_d}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$$
denote the density function of a d-variate $\mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ random vector. Then we have:

Result 2. *For all $d\times d'$ matrices $\boldsymbol{L}$ such $\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}$ is positive definite and $d'\times 1$ vectors $\boldsymbol{v}$:*

$$\int_{\mathbb{R}^d} p_{N_d}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\,\delta\big(\boldsymbol{v}-\boldsymbol{L}^T\boldsymbol{x}\big)\,d\boldsymbol{x} = p_{N_{d'}}(\boldsymbol{v};\boldsymbol{L}^T\boldsymbol{\mu},\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}). \tag{S.5}$$

Derivation of Result 2.

Via the change of variable $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu})$, the left-hand side of (S.5) is

$$\int_{\mathbb{R}^d} p_{N_d}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})\,\delta\big(\boldsymbol{v}-\boldsymbol{L}^T\boldsymbol{\mu}-\boldsymbol{L}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{z}\big)\,d\boldsymbol{z} = \int_{\mathbb{R}^d} p_{N_d}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})\,\delta\big(\boldsymbol{v}^\dagger-(\boldsymbol{L}^\dagger)^T\boldsymbol{z}\big)\,d\boldsymbol{z}$$

where

$$\boldsymbol{v}^\dagger \equiv \boldsymbol{v}-\boldsymbol{L}^T\boldsymbol{\mu} \quad \text{and} \quad \boldsymbol{L}^\dagger \equiv \boldsymbol{\Sigma}^{1/2}\boldsymbol{L}.$$

Next note that

$$\begin{aligned}\int_{\mathbb{R}^d} p_{N_d}(\boldsymbol{x};\boldsymbol{0},\boldsymbol{I})\,\delta\big(\boldsymbol{v}^\dagger-(\boldsymbol{L}^\dagger)^T\boldsymbol{z}\big)\,d\boldsymbol{z} &= \int_{\mathbb{R}^d}\lim_{\varepsilon\to 0} p_{N_d}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})\,p_{N_{d'}}\big(\boldsymbol{v}^\dagger-(\boldsymbol{L}^\dagger)^T\boldsymbol{z};\boldsymbol{0},\varepsilon\,\boldsymbol{I}\big)\,d\boldsymbol{z}\\ &= (2\pi)^{-d/2}\int_{\mathbb{R}^d}\lim_{\varepsilon\to 0}\big[(2\pi\varepsilon)^{-d'/2}\exp\{r(\boldsymbol{z},\boldsymbol{v}^\dagger,\boldsymbol{L}^\dagger,\varepsilon)\}\big]\,d\boldsymbol{z}\end{aligned}$$

where

$$\begin{aligned}r(\boldsymbol{z},\boldsymbol{v}^\dagger,\boldsymbol{L}^\dagger,\varepsilon) &= -\tfrac{1}{2}\boldsymbol{z}^T\boldsymbol{z}-\tfrac{1}{2\varepsilon}\{\boldsymbol{v}^\dagger-(\boldsymbol{L}^\dagger)^T\boldsymbol{z}\}^T\{\boldsymbol{v}^\dagger-(\boldsymbol{L}^\dagger)^T\boldsymbol{z}\}\\ &= -\tfrac{1}{2}(\boldsymbol{z}-\boldsymbol{m})^T\{\boldsymbol{I}+\tfrac{1}{\varepsilon}\boldsymbol{L}^\dagger(\boldsymbol{L}^\dagger)^T\}(\boldsymbol{z}-\boldsymbol{m})-\tfrac{1}{2\varepsilon}(\boldsymbol{v}^\dagger)^T\boldsymbol{v}^\dagger\\ &\quad +\tfrac{1}{2\varepsilon^2}(\boldsymbol{L}^\dagger\boldsymbol{v}^\dagger)^T\{\boldsymbol{I}+\tfrac{1}{\varepsilon}\boldsymbol{L}^\dagger(\boldsymbol{L}^\dagger)^T\}^{-1}\boldsymbol{L}^\dagger\boldsymbol{v}^\dagger\end{aligned}$$

with $\boldsymbol{m} = \boldsymbol{m}(\boldsymbol{v}^\dagger,\boldsymbol{L}^\dagger,\varepsilon) \equiv \frac{1}{\varepsilon}(\boldsymbol{I}+\frac{1}{\varepsilon}\boldsymbol{L}\boldsymbol{L}^T)^{-1}(\boldsymbol{L}^\dagger)^T\boldsymbol{v}^\dagger$. We then have

$$\begin{aligned}\textstyle\int_{\mathbb{R}^d} p_{N_d}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})\,\delta\big(\boldsymbol{v}^\dagger-(\boldsymbol{L}^\dagger)^T\boldsymbol{z}\big)\,d\boldsymbol{z} = \lim_{\varepsilon\to 0}\Big[&(2\pi\varepsilon)^{-d'/2}|\boldsymbol{I}+\tfrac{1}{\varepsilon}\boldsymbol{L}^\dagger(\boldsymbol{L}^\dagger)^T|^{-1/2}\\ &\times\exp\big\{\tfrac{1}{2\varepsilon^2}(\boldsymbol{L}^\dagger\boldsymbol{v}^\dagger)^T\{\boldsymbol{I}+\tfrac{1}{\varepsilon}\boldsymbol{L}^\dagger(\boldsymbol{L}^\dagger)^T\}^{-1}\boldsymbol{L}^\dagger\boldsymbol{v}^\dagger-\tfrac{1}{2\varepsilon}(\boldsymbol{v}^\dagger)^T\boldsymbol{v}^\dagger\big\}\Big]\end{aligned} \tag{S.6}$$

where we have used the fact

$$\int_{\mathbb{R}^d}(2\pi)^{-d/2}|\boldsymbol{I}+\tfrac{1}{\varepsilon}\boldsymbol{L}^\dagger(\boldsymbol{L}^\dagger)^T|^{1/2}\exp\Big\{-\tfrac{1}{2}(\boldsymbol{z}-\boldsymbol{m})^T\big\{\boldsymbol{I}+\tfrac{1}{\varepsilon}\boldsymbol{L}^\dagger(\boldsymbol{L}^\dagger)^T\big\}(\boldsymbol{z}-\boldsymbol{m})\Big\}\,d\boldsymbol{z} = 1$$

and assumed that the integrand possesses properties that are sufficient to justify interchanging the limit as $\varepsilon\to 0$ and the integral over $\mathbb{R}^d$. Using Theorem 18.1.1 of Harville (2008), the determinant in (S.6) is

$$|\boldsymbol{I}+\tfrac{1}{\varepsilon}\boldsymbol{L}^\dagger(\boldsymbol{L}^\dagger)^T| = \varepsilon^{-d'}|\varepsilon\,\boldsymbol{I}+(\boldsymbol{L}^\dagger)^T\boldsymbol{L}^\dagger| = \varepsilon^{-d'}|\varepsilon\,\boldsymbol{I}+\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}|. \tag{S.7}$$

Next, application of Theorem 18.2.8 of Harville (2008) gives

$$\{\boldsymbol{I}+\tfrac{1}{\varepsilon}\boldsymbol{L}^\dagger(\boldsymbol{L}^\dagger)^T\}^{-1}=\boldsymbol{I}-\boldsymbol{L}^\dagger\big\{\varepsilon\boldsymbol{I}+(\boldsymbol{L}^\dagger)^T\boldsymbol{L}^\dagger\big\}^{-1}(\boldsymbol{L}^\dagger)^T \tag{S.8}$$

which implies that

$$\begin{aligned}(\boldsymbol{L}^\dagger\boldsymbol{v}^\dagger)^T\{\boldsymbol{I}+\tfrac{1}{\varepsilon}\boldsymbol{L}^\dagger(\boldsymbol{L}^\dagger)^T\}^{-1}\boldsymbol{L}^\dagger\boldsymbol{v}^\dagger&=(\boldsymbol{v}^\dagger)^T(\boldsymbol{L}^\dagger)^T\boldsymbol{L}^\dagger\boldsymbol{v}^\dagger-(\boldsymbol{L}^\dagger\boldsymbol{v}^\dagger)^T\boldsymbol{L}^\dagger\big\{\varepsilon\boldsymbol{I}+(\boldsymbol{L}^\dagger)^T\boldsymbol{L}^\dagger\big\}^{-1}(\boldsymbol{L}^\dagger)^T\boldsymbol{L}^\dagger\boldsymbol{v}^\dagger\\&=(\boldsymbol{v}^\dagger)^T\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}\,\boldsymbol{v}^\dagger-(\boldsymbol{v}^\dagger)^T\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}\big(\varepsilon\boldsymbol{I}+\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}\big)^{-1}\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}\boldsymbol{v}^\dagger.\end{aligned}$$

The exponent in the expression on the right-hand side of (S.6) is then

$$\begin{aligned}&\tfrac{1}{2\varepsilon^2}\Big\{(\boldsymbol{v}^\dagger)^T\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}\,\boldsymbol{v}^\dagger-(\boldsymbol{v}^\dagger)^T\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}\big(\varepsilon\boldsymbol{I}+\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}\big)^{-1}\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}\boldsymbol{v}^\dagger-\varepsilon(\boldsymbol{v}^\dagger)^T\boldsymbol{v}^\dagger\Big\}\\&\quad=-\tfrac{1}{2}(\boldsymbol{v}^\dagger)^T\big(\varepsilon\boldsymbol{I}+\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}\big)^{-1}\boldsymbol{v}^\dagger=-\tfrac{1}{2}(\boldsymbol{v}-\boldsymbol{L}^T\boldsymbol{\mu})^T\big(\varepsilon\boldsymbol{I}+\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}\big)^{-1}(\boldsymbol{v}-\boldsymbol{L}^T\boldsymbol{\mu}).\end{aligned}$$

Substitution of this result and (S.7) into (S.6) then gives

$$\begin{aligned}&\int_{\mathbb{R}^d}p_{N_d}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})\,\delta\big(\boldsymbol{v}^\dagger-(\boldsymbol{L}^\dagger)^T\boldsymbol{z}\big)\,d\boldsymbol{z}\\&\quad=\lim_{\varepsilon\to0}\big[(2\pi)^{-d'/2}|\varepsilon\boldsymbol{I}+\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L}|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{v}-\boldsymbol{L}^T\boldsymbol{\mu})^{-1}(\varepsilon\boldsymbol{I}+\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L})^{-1}(\boldsymbol{v}-\boldsymbol{L}^T\boldsymbol{\mu})\}\big].\\&\quad=p_{N_{d'}}(\boldsymbol{v};\boldsymbol{L}^T\boldsymbol{\mu},\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{L})\end{aligned}$$

and the result follows.

∎

From (9) we have

$$\boldsymbol{\eta}_{\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})\to\boldsymbol{\alpha}}(\boldsymbol{\alpha})\longleftarrow\frac{\text{proj}_{\text{N}}\Big[m_{\boldsymbol{\alpha}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}(\boldsymbol{\alpha})\displaystyle\int_{\mathbb{R}^{d^{\boldsymbol{\theta}}}}\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})m_{\boldsymbol{\theta}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}(\boldsymbol{\theta})\,d\boldsymbol{\theta}/Z\Big]}{m_{\boldsymbol{\alpha}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}(\boldsymbol{\alpha})} \tag{S.9}$$

where $d^{\boldsymbol{\theta}}$ is the dimension of $\boldsymbol{\theta}$. It follows from assumption (15) and (8) that

$$m_{\boldsymbol{\alpha}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}(\boldsymbol{\alpha})=\exp\left\{\begin{bmatrix}\boldsymbol{\alpha}\\\text{vec}(\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\end{bmatrix}^T\boldsymbol{\eta}_{\boldsymbol{\alpha}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}\right\} \tag{S.10}$$

and

$$\begin{aligned}m_{\boldsymbol{\theta}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}(\boldsymbol{\theta})&=\exp\left\{\begin{bmatrix}\boldsymbol{\theta}\\\text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T\boldsymbol{\eta}_{\boldsymbol{\theta}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}\right\}\\&\propto(2\pi)^{-d^{\boldsymbol{\theta}}/2}|\boldsymbol{\Sigma}_{\odot}|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\odot})^T\boldsymbol{\Sigma}_{\odot}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\odot})\}\end{aligned}$$

where

$$\boldsymbol{\mu}_{\odot}\equiv-\tfrac{1}{2}\Big\{\text{vec}^{-1}\big(\boldsymbol{\eta}_{\boldsymbol{\theta}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}\big)_2\Big\}^{-1}\Big(\boldsymbol{\eta}_{\boldsymbol{\theta}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}\Big)_1$$

and

$$\boldsymbol{\Sigma}_{\odot}\equiv-\tfrac{1}{2}\Big\{\text{vec}^{-1}\big(\boldsymbol{\eta}_{\boldsymbol{\theta}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}\big)_2\Big\}^{-1}$$

are the common parameters matching $\boldsymbol{\eta}_{\boldsymbol{\theta}\to\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})}$. From Result 2 in Appendix S.2.5,

$$\begin{aligned}&\int_{\mathbb{R}^{d^{\boldsymbol{\theta}}}}\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})(2\pi)^{-d^{\boldsymbol{\theta}}/2}|\boldsymbol{\Sigma}_{\odot}|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\odot})^T\boldsymbol{\Sigma}_{\odot}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\odot})\}\,d\boldsymbol{\theta}\\&\quad=|2\pi\,\boldsymbol{A}^T\boldsymbol{\Sigma}_{\odot}\boldsymbol{A}|^{-1/2}\exp\Big\{-\tfrac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\mu}_{\odot})^T(\boldsymbol{A}^T\boldsymbol{\Sigma}_{\odot}\boldsymbol{A})^{-1}(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\mu}_{\odot})\Big\}.\end{aligned}$$

Substitution into (S.9) then leads to

$$
\begin{aligned}
&\boldsymbol{\eta}_{\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})\,\to\,\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \\
&\frac{\mathrm{proj}_{\mathrm{N}}\left[\exp\left\{\begin{bmatrix}\boldsymbol{\alpha}\\ \mathrm{vec}(\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\end{bmatrix}\boldsymbol{\eta}_{\boldsymbol{\alpha}\,\to\,\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})}\right\}\exp\left\{\begin{bmatrix}\boldsymbol{\alpha}\\ \mathrm{vec}(\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\end{bmatrix}^T\begin{bmatrix}(\boldsymbol{A}^T\boldsymbol{\Sigma}_{\odot}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{\mu}_{\odot}\\ -\frac{1}{2}\mathrm{vec}\big((\boldsymbol{A}^T\boldsymbol{\Sigma}_{\odot}\boldsymbol{A})^{-1}\big)\end{bmatrix}\right\}\Big/ Z\right]}{\exp\left\{\begin{bmatrix}\boldsymbol{\alpha}\\ \mathrm{vec}(\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\end{bmatrix}\boldsymbol{\eta}_{\boldsymbol{\alpha}\,\to\,\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})}\right\}} \\
&= \exp\left\{\begin{bmatrix}\boldsymbol{\alpha}\\ \mathrm{vec}(\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\end{bmatrix}^T\begin{bmatrix}(\boldsymbol{A}^T\boldsymbol{\Sigma}_{\odot}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{\mu}_{\odot}\\ -\frac{1}{2}\mathrm{vec}\big((\boldsymbol{A}^T\boldsymbol{\Sigma}_{\odot}\boldsymbol{A})^{-1}\big)\end{bmatrix}\right\}.
\end{aligned}
$$

Therefore, setting $\boldsymbol{\Omega} \longleftarrow -2\boldsymbol{\Sigma}_{\odot}\boldsymbol{A}$, we get

$$
\boldsymbol{\eta}_{\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})\,\to\,\boldsymbol{\alpha}} \longleftarrow \begin{bmatrix}(\boldsymbol{\Omega}^T\boldsymbol{A})^{-1}\boldsymbol{\Omega}^T\Big(\boldsymbol{\eta}_{\boldsymbol{\theta}\,\to\,\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})}\Big)_1\\ \mathrm{vec}\big((\boldsymbol{\Omega}^T\boldsymbol{A})^{-1}\big)\end{bmatrix}
$$

which are the first two updates of Algorithm 4.

For the third update, note that (9) gives

$$
\begin{aligned}
&m_{\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}}(\boldsymbol{\theta}) \\
&\longleftarrow \frac{\mathrm{proj}_{\mathrm{N}}\Big[m_{\boldsymbol{\theta}\,\to\,\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})}(\boldsymbol{\theta})\int_{\mathbb{R}^{d\boldsymbol{\alpha}}}\delta(\boldsymbol{\alpha}-\boldsymbol{A}^T\boldsymbol{\theta})m_{\boldsymbol{\alpha}\,\to\,\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})}(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}/Z\Big]}{m_{\boldsymbol{\theta}\,\to\,\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})}(\boldsymbol{\theta})} \\
&= \frac{\mathrm{proj}_{\mathrm{N}}\Big[m_{\boldsymbol{\theta}\,\to\,\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})}(\boldsymbol{\theta})\,m_{\boldsymbol{\alpha}\,\to\,\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})}(\boldsymbol{A}^T\boldsymbol{\theta})/Z\Big]}{m_{\boldsymbol{\theta}\,\to\,\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})}(\boldsymbol{\theta})} \\
&= \frac{\mathrm{proj}_{\mathrm{N}}\left[\exp\left\{\begin{bmatrix}\boldsymbol{\theta}\\ \mathrm{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T\boldsymbol{\eta}_{\boldsymbol{\theta}\,\to\,\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})}\right\}\exp\left\{\begin{bmatrix}\boldsymbol{A}^T\boldsymbol{\theta}\\ \mathrm{vec}\big((\boldsymbol{A}^T\boldsymbol{\theta})(\boldsymbol{A}^T\boldsymbol{\theta})^T\big)\end{bmatrix}^T\boldsymbol{\eta}_{\boldsymbol{\alpha}\,\to\,\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})}\right\}\Big/ Z\right]}{\exp\left\{\begin{bmatrix}\boldsymbol{\theta}\\ \mathrm{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T\boldsymbol{\eta}_{\boldsymbol{\theta}\,\to\,\delta(\boldsymbol{\alpha}\,-\,\boldsymbol{A}^T\boldsymbol{\theta})}\right\}} \\
&= \exp\left\{\begin{bmatrix}\boldsymbol{\theta}\\ \mathrm{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{A}\Big(\boldsymbol{\eta}_{\alpha\,\to\,\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})}\Big)_1\\ (\boldsymbol{A}\otimes\boldsymbol{A})\Big(\boldsymbol{\eta}_{\alpha\,\to\,\delta(\alpha\,-\,\boldsymbol{a}^T\boldsymbol{\theta})}\Big)_2\end{bmatrix}\right\}.
\end{aligned}
$$

The last step uses the identity $\mathrm{vec}(\boldsymbol{BCD}) = (\boldsymbol{D}^T\otimes\boldsymbol{B})\mathrm{vec}(\boldsymbol{C})$ for any compatible matrices $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$. The third update follows immediately.

S.2.6 Derivation of Algorithm 6

The message from $p(y|\alpha,\sigma^2)$ to α is, from (9),

$$
m_{p(y|\alpha,\sigma^2)\,\to\,\alpha}(\alpha) = \frac{\mathrm{proj}_{\mathrm{N}}\big[m_{\alpha\,\to\,p(y|\alpha,\sigma^2)}(\alpha)\,\int_0^\infty p(y|\alpha,\sigma^2)\,m_{\sigma^2\,\to\,p(y|\alpha,\sigma^2)}(\sigma^2)\,d\sigma^2\big]}{m_{\alpha\,\to\,p(y|\alpha,\sigma^2)}(\alpha)}. \tag{S.11}
$$

It follows from (8) and (18) that

$$
m_{\sigma^2\,\to\,p(y|\alpha,\sigma^2)}(\sigma^2) = \exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma^2\end{bmatrix}^T\boldsymbol{\eta}_{\sigma^2\,\to\,p(y|\alpha,\sigma^2)}\right\} = (\sigma^2)^{\eta_1^{\boxplus}}\exp(\eta_2^{\boxplus}/\sigma^2).
$$

where $\eta^{\boxplus} \equiv \boldsymbol{\eta}_{\sigma^2 \,\to\, p(y|\alpha,\,\sigma^2)}$. Hence

$$\int_0^\infty p(y|\alpha,\sigma^2)\, m_{\sigma^2 \,\to\, p(y|\alpha,\,\sigma^2)}(\sigma^2)\, d\sigma^2 \;=\; (2\pi)^{-1/2} \int_0^\infty (\sigma^2)^{\eta_1^{\boxplus}-1/2} \exp[\{\eta_2^{\boxplus} - \tfrac{1}{2}(y-\alpha)^2\}/\sigma^2)]\, d\sigma^2$$

$$\propto \left\{\tfrac{1}{2}(y-\alpha)^2 - \eta_2^{\boxplus}\right\}^{\eta_1^{\boxplus}+\frac{1}{2}}.$$

Therefore, letting $\eta^{\text{✡}} \equiv \boldsymbol{\eta}_{\alpha \,\to\, p(y|\alpha,\,\sigma^2)}$, the numerator of the right-hand side of (S.11) is

$$\begin{aligned}
&\mathrm{proj}_{\mathrm{N}}\left[\frac{\exp(\alpha\eta_1^{\text{✡}} + \alpha^2\eta_1^{\text{✡}})}{\left\{(y-\alpha)^2 - 2\eta_2^{\boxplus}\right\}^{-\eta_1^{\boxplus}-\frac{1}{2}}}\right] \\
&= \exp\left\{\begin{bmatrix}\alpha \\ \alpha^2\end{bmatrix}^T (\nabla A_{\mathrm{N}})^{-1}\left(\begin{bmatrix}\dfrac{\int_{-\infty}^{\infty}\alpha\exp(\alpha\eta_1^{\text{✡}}+\alpha^2\eta_2^{\text{✡}})\Big/\left\{(y-\alpha)^2-2\eta_2^{\boxplus}\right\}^{-\eta_1^{\boxplus}-\frac{1}{2}}\,d\alpha}{\int_{-\infty}^{\infty}\exp(\alpha\eta_1^{\text{✡}}+\alpha^2\eta_2^{\text{✡}})\Big/\left\{(y-\alpha)^2-2\eta_2^{\boxplus}\right\}^{-\eta_1^{\boxplus}-\frac{1}{2}}\,d\alpha} \\[2ex] \dfrac{\int_{-\infty}^{\infty}\alpha^2\exp(\alpha\eta_2^{\text{✡}}+\alpha^2\eta_2^{\text{✡}})\Big/\left\{(y-\alpha)^2-2\eta_2^{\boxplus}\right\}^{-\eta_1^{\boxplus}-\frac{1}{2}}\,d\alpha}{\int_{-\infty}^{\infty}\exp(\alpha\eta_1^{\text{✡}}+\alpha^2\eta_2^{\text{✡}})\Big/\left\{(y-\alpha)^2-2\eta_2^{\boxplus}\right\}^{-\eta_1^{\boxplus}-\frac{1}{2}}\,d\alpha}\end{bmatrix}\right)\right\} \\
&= \exp\left\{\begin{bmatrix}\alpha \\ \alpha^2\end{bmatrix}^T (\nabla A_N)^{-1}\left(\begin{bmatrix}\dfrac{\mathcal{A}(1,\eta_1^{\text{✡}},-\eta_1^{\text{✡}},-2y,y^2-2\eta_2^{\boxplus},-\eta_1^{\boxplus}-\frac{1}{2})}{\mathcal{A}(0,\eta_1^{\text{✡}},-\eta_2^{\text{✡}},-2y,y^2-2\eta_2^{\boxplus},-\eta_1^{\boxplus}-\frac{1}{2})} \\ \dfrac{\mathcal{A}(2,\eta_1^{\text{✡}},-\eta_1^{\text{✡}},-2y,y^2-2\eta_2^{\boxplus},-\eta_1^{\boxplus}-\frac{1}{2})}{\mathcal{A}(0,\eta_1^{\text{✡}},-\eta_2^{\text{✡}},-2y,y^2-2\eta_2^{\boxplus},-\eta_1^{\boxplus}-\frac{1}{2})}\end{bmatrix}\right)\right\}.
\end{aligned}$$

Hence

$$\boldsymbol{\eta}_{p(y|\alpha,\,\sigma^2)\,\to\,\alpha} \longleftarrow (\nabla A_N)^{-1}\left(\begin{bmatrix}\dfrac{\mathcal{A}(1,\eta_1^{\text{✡}},-\eta_1^{\text{✡}},-2y,y^2-2\eta_2^{\boxplus},-\eta_1^{\boxplus}-\frac{1}{2})}{\mathcal{A}(0,\eta_1^{\text{✡}},-\eta_2^{\text{✡}},-2y,y^2-2\eta_2^{\boxplus},-\eta_1^{\boxplus}-\frac{1}{2})} \\ \dfrac{\mathcal{A}(2,\eta_1^{\text{✡}},-\eta_1^{\text{✡}},-2y,y^2-2\eta_2^{\boxplus},-\eta_1^{\boxplus}-\frac{1}{2})}{\mathcal{A}(0,\eta_1^{\text{✡}},-\eta_2^{\text{✡}},-2y,y^2-2\eta_2^{\boxplus},-\eta_1^{\boxplus}-\frac{1}{2})}\end{bmatrix}\right) - \boldsymbol{\eta}_{\alpha \,\to\, p(y|\alpha,\,\sigma^2)}$$

and the first update in Algorithm 6 follows from the definition of G^{N} given in Section S.1.3.

The message $p(y|\alpha,\sigma^2)$ to σ^2 is

$$m_{p(by|\alpha,\,\sigma^2)\,\to\,\sigma^2}(\sigma^2) = \frac{\mathrm{proj}_{\mathrm{I}\chi^2}\left[m_{\sigma^2 \,\to\, p(y|\alpha,\,\sigma^2)}(\sigma^2)\,\int_{-\infty}^{\infty} p(y|\alpha,\sigma^2)\, m_{\alpha \,\to\, p(y|\alpha,\,\sigma^2)}(\alpha)\, d\alpha\right]}{m_{\sigma^2 \,\to\, p(y|\alpha,\,\sigma^2)}(\sigma^2)}. \tag{S.12}$$

It follows from (8) and (18) that

$$m_{\alpha \,\to\, p(y|\alpha,\,\sigma^2)}(\alpha) = \exp\left\{\begin{bmatrix}\alpha \\ \alpha^2\end{bmatrix}^T \boldsymbol{\eta}_{\alpha \,\to\, p(y|\alpha,\,\sigma^2)}\right\} = \exp\left(\alpha\,\eta_1^{\oplus} + \alpha^2\eta_2^{\oplus}\right)$$

where $\boldsymbol{\eta}^{\oplus} \equiv \boldsymbol{\eta}_{\alpha \,\to\, p(y|\alpha,\,\sigma^2)}$. The integral in (S.12) is

$$\int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\left\{\frac{-(y-\alpha)^2}{2\sigma^2}\right\} \exp\left(\alpha\,\eta_1^{\oplus} + \alpha^2\eta_2^{\oplus}\right) d\alpha$$

$$\propto \int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\left\{\frac{-(y-\alpha)^2}{2\sigma^2}\right\} \{2\pi(\sigma^{\oplus})^2\}^{-1/2} \exp\left\{\frac{-(y-\mu^{\oplus})^2}{2(\sigma^{\oplus})^2}\right\} d\alpha$$

where

$$\begin{bmatrix}\mu^{\oplus} \\ (\sigma^{\oplus})^2\end{bmatrix} \equiv \begin{bmatrix}-\eta_1^{\oplus}/(2\eta_2^{\oplus}) \\ -1/(2\eta_2^{\oplus})\end{bmatrix}$$

is the common parameter vector corresponding to $\boldsymbol{\eta}^{\oplus}$. Using (A.2) of Wand and Jones (1993), the last-written integral is

$$\{2\pi(\sigma^2+(\sigma^{\oplus})^2\}^{-1/2}\exp\left\{\frac{-(y-\mu^{\oplus})^2}{2\{\sigma^2+(\sigma^{\oplus})^2\}}\right\}.$$

Therefore, the function inside the $\text{proj}_{\text{I}\chi^2}[\,\cdot\,]$ in (S.12) is

$$(\sigma^2)^{\eta_1^{\boxplus}}\exp(\eta_2^{\boxplus}/\sigma^2)[2\pi\{\sigma^2-1/(2\eta_2^{\oplus})\}]^{-1/2}\exp\left[\frac{-\{y+\eta_1^{\oplus}/(2\eta_2^{\oplus})\}^2}{2\{\sigma^2-1/(2\eta_2^{\oplus})\}}\right].$$

Plugging into (S.12) we then have

$$m_{p(by|\alpha,\sigma^2)\,\rightarrow\,\sigma^2}(\sigma^2)=\exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma^2\end{bmatrix}^T\boldsymbol{\eta}_{p(y|\alpha,\sigma^2)\,\rightarrow\,\sigma^2}\right\}$$

where

$$\boldsymbol{\eta}_{p(y|\alpha,\sigma^2)\,\rightarrow\,\sigma^2}\longleftarrow(\nabla A_{\text{I}\chi^2})^{-1}\left(\begin{bmatrix}\dfrac{\int_0^{\infty}\log(\sigma^2)(\sigma^2)^{\eta_1^{\boxplus}}\{\sigma^2-1/(2\eta_2^{\oplus})\}^{-1/2}\exp\left[\frac{\eta_2^{\boxplus}}{\sigma^2}-\frac{\{y+\eta_1^{\oplus}/(2\eta_2^{\oplus})\}^2}{2\{\sigma^2-1/(2\eta_2^{\oplus})\}}\right]d\sigma^2}{\int_0^{\infty}(\sigma^2)^{\eta_1^{\boxplus}}\{\sigma^2-1/(2\eta_2^{\oplus})\}^{-1/2}\exp\left[\frac{\eta_2^{\boxplus}}{\sigma^2}-\frac{\{y+\eta_1^{\oplus}/(2\eta_2^{\oplus})\}^2}{2\{\sigma^2-1/(2\eta_2^{\oplus})\}}\right]d\sigma^2}\\ \dfrac{\int_0^{\infty}(1/\sigma^2)(\sigma^2)^{\eta_1^{\boxplus}}\{\sigma^2-1/(2\eta_2^{\oplus})\}^{-1/2}\exp\left[\frac{\eta_2^{\boxplus}}{\sigma^2}-\frac{\{y+\eta_1^{\oplus}/(2\eta_2^{\oplus})\}^2}{2\{\sigma^2-1/(2\eta_2^{\oplus})\}}\right]d\sigma^2}{\int_0^{\infty}(\sigma^2)^{\eta_1^{\boxplus}}\{\sigma^2-1/(2\eta_2^{\oplus})\}^{-1/2}\exp\left[\frac{\eta_2^{\boxplus}}{\sigma^2}-\frac{\{y+\eta_1^{\oplus}/(2\eta_2^{\oplus})\}^2}{2\{\sigma^2-1/(2\eta_2^{\oplus})\}}\right]d\sigma^2}\end{bmatrix}\right)-\boldsymbol{\eta}^{\boxplus}.$$

The change of variable $\sigma^2=e^{-x}$ and some simple algebra then leads to

$$\boldsymbol{\eta}_{p(y|\alpha,\sigma^2)\,\rightarrow\,\sigma^2}\longleftarrow(\nabla A_{\text{I}\chi^2})^{-1}\left(\begin{bmatrix}\dfrac{-\mathcal{B}(1,-\eta_1^{\boxplus},-\eta_2^{\boxplus},-2\eta_2^{\oplus}\{y+\eta_1^{\oplus}/(2\eta_2^{\oplus})\}^2,-2\eta_2^{\oplus},\frac{1}{2})}{\mathcal{B}(0,-\eta_1^{\boxplus},-\eta_2^{\boxplus},-2\eta_2^{\oplus}\{y+\eta_1^{\oplus}/(2\eta_2^{\oplus})\}^2,-2\eta_2^{\oplus},\frac{1}{2})}\\ \dfrac{\mathcal{B}(0,1-\eta_1^{\boxplus},-\eta_2^{\boxplus},-2\eta_2^{\oplus}\{y+\eta_1^{\oplus}/(2\eta_2^{\oplus})\}^2,-2\eta_2^{\oplus},\frac{1}{2})}{\mathcal{B}(0,-\eta_1^{\boxplus},-\eta_2^{\boxplus},-2\eta_2^{\oplus}\{y+\eta_1^{\oplus}/(2\eta_2^{\oplus})\}^2,-2\eta_2^{\oplus},\frac{1}{2})}\end{bmatrix}\right)-\boldsymbol{\eta}_{\sigma^2\,\rightarrow\,p(y|\alpha,\sigma^2)}.$$

The second update in Algorithm 6 follows from the definition of G^{IG1} given in Section S.1.3.

S.2.7 Derivation of Algorithm 7

The only factor to stochastic node message for the logistic fragment is, from (9):

$$m_{p(y|\alpha)\,\rightarrow\,\alpha}(\alpha)=\frac{\text{proj}[m_{\alpha\,\rightarrow\,p(y|\alpha)}(\alpha)\,p(y|\alpha)/Z]}{m_{\alpha\,\rightarrow\,p(y|\alpha)}(\alpha)}$$

with projection onto an appropriate exponential family. Assumption (19) and conjugacy considerations implies projection onto the Univariate Normal family. Setting

$$\boldsymbol{\eta}^{\#}\equiv\boldsymbol{\eta}_{\alpha\,\rightarrow\,p(y|\alpha)},$$

we then have

$$m_{p(y|\alpha)\,\rightarrow\,\alpha}(\alpha)=\frac{\text{proj}_{\text{N}}[\exp(\eta_1^{\#}\alpha+\eta_2^{\#}\alpha^2)\,\exp\{y\alpha-\log(1+e^{\alpha})\}]}{\exp(\eta_1^{\#}\alpha+\eta_2^{\#}\alpha^2)}.$$

The numerator is

$$
\begin{aligned}
&\text{proj}_{\text{N}}\big[\exp\{(\eta_1^\# + y)\alpha + \eta_2^\#\alpha^2 - \log(1+e^\alpha)\}\big]\\
&\qquad \propto \exp\left\{\begin{bmatrix}\alpha\\ \alpha^2\end{bmatrix}^T (\nabla A_N)^{-1}\left(\begin{bmatrix}\frac{\int_{-\infty}^{\infty}\alpha\exp\{(\eta_1^\#+y)\alpha+\eta_2^\#\alpha^2-\log(1+e^\alpha)\}\,d\alpha}{\int_{-\infty}^{\infty}\exp\{(\eta_1^\#+y)\alpha+\eta_2^\#\alpha^2-\log(1+e^\alpha)\}\,d\alpha}\\ \frac{\int_{-\infty}^{\infty}\alpha^2\exp\{(\eta_1^\#+y)\alpha+\eta_2^\#\alpha^2-\log(1+e^\alpha)\}\,d\alpha}{\int_{-\infty}^{\infty}\exp\{(\eta_1^\#+y)\alpha+\eta_2^\#\alpha^2-\log(1+e^\alpha)\}\,d\alpha}\end{bmatrix}\right)\right\}\\
&\qquad \propto \exp\left\{\begin{bmatrix}\alpha\\ \alpha^2\end{bmatrix}^T (\nabla A_N)^{-1}\left(\begin{bmatrix}\frac{\mathcal{C}(1,\eta_1^\#+y,\eta_1^\#)}{\mathcal{C}(0,\eta_1^\#+y,\eta_1^\#)}\\ \frac{\mathcal{C}(2,\eta_1^\#+y,\eta_1^\#)}{\mathcal{C}(0,\eta_1^\#+y,\eta_1^\#)}\end{bmatrix}\right)\right\}.
\end{aligned}
$$

Hence

$$
\boldsymbol{\eta}_{p(y|\alpha)\,\to\,\alpha} \longleftarrow (\nabla A_N)^{-1}\left(\begin{bmatrix}\frac{\mathcal{C}(1,\eta_1^\#+y,\eta_1^\#)}{\mathcal{C}(0,\eta_1^\#+y,\eta_1^\#)}\\ \frac{\mathcal{C}(2,\eta_1^\#+y,\eta_1^\#)}{\mathcal{C}(0,\eta_1^\#+y,\eta_1^\#)}\end{bmatrix}\right) - \boldsymbol{\eta}_{\alpha\,\to\,p(y|\alpha)}
$$

and the update in Algorithm 7 follows from the definition of H_{logistic} given in Section S.1.3.

S.2.8 Derivation of Algorithm 8

Via arguments analogous to those given in Section S.2.7, the factor to stochastic node natural parameter update is

$$
\boldsymbol{\eta}_{p(y|\alpha)\,\to\,\alpha} \longleftarrow (\nabla A_N)^{-1}\left(\begin{bmatrix}\frac{\int_{-\infty}^{\infty}\alpha\Phi((2y-1)\alpha)\exp\{\eta_1^\#\alpha+\eta_2^\#\alpha^2\}\,d\alpha}{\int_{-\infty}^{\infty}\Phi((2y-1)\alpha)\exp\{\eta_1^\#\alpha+\eta_2^\#\alpha^2\}\,d\alpha}\\ \frac{\int_{-\infty}^{\infty}\alpha^2\Phi((2y-1)\alpha)\exp\{\eta_1^\#\alpha+\eta_2^\#\alpha^2\}\,d\alpha}{\int_{-\infty}^{\infty}\Phi((2y-1)\alpha)\exp\{\eta_1^\#\alpha+\eta_2^\#\alpha^2\}\,d\alpha}\end{bmatrix}\right) - \boldsymbol{\eta}_{\alpha\,\to\,p(y|\alpha)}
$$

where, as before, $\boldsymbol{\eta}^\# \equiv \boldsymbol{\eta}_{\alpha\,\to\,p(y|\alpha)}$. The integral results

$$
\begin{aligned}
\int_{-\infty}^{\infty}\Phi(a+bx)\phi(x)\,dx &= \Phi\left(\frac{a}{\sqrt{b^2+1}}\right),\\
\int_{-\infty}^{\infty}x\,\Phi(a+bx)\phi(x)\,dx &= \frac{b}{\sqrt{b^2+1}}\,\phi\left(\frac{a}{\sqrt{b^2+1}}\right)\\
\text{and}\quad \int_{-\infty}^{\infty}x^2\,\Phi(a+bx)\phi(x)\,dx &= \Phi\left(\frac{a}{\sqrt{b^2+1}}\right) - \frac{ab^2}{\sqrt{(b^2+1)^3}}\,\phi\left(\frac{a}{\sqrt{b^2+1}}\right),
\end{aligned}
\tag{S.13}
$$

and standard algebraic manipulations lead to

$$
\boldsymbol{\eta}_{p(y|\alpha)\,\to\,\alpha} \longleftarrow H_{\text{probit}}\left(\boldsymbol{\eta}_{\alpha\,\to\,p(y|\alpha)}\right)
$$

where H_{probit} is defined in Section S.1.3.

S.2.9 Derivation of Algorithm 9

The Poisson likelihood fragment derivation is very similar to that given in Section S.2.7. The only change is that

$$
\exp\{y\alpha - \log(1+e^\alpha)\} \quad \text{is replaced by} \quad \exp(y\alpha - e^\alpha)
$$

which leads to the H_{Poisson} function appearing in the factor to stochastic node update rather than the H_{logistic} function. Section S.1.3 provides the definition of H_{Poisson}.