# Supplement for: Variational Message Passing for Elaborate Response Regression Models

M. W. McLean[*] and M. P. Wand[†]

## S.1 Special Function Definitions and Results

Here we survey special functions that arise in the VMP updates for the elaborate distributions covered in this article.

### S.1.1 Modified Bessel Functions of the Second Kind

The *modified Bessel function of the second kind* of order $p \in \mathbb{R}$ is denoted by $K_p$. The argument of $K_p$ can be an arbitrary complex number. We restrict attention here to positive real arguments, which is sufficient for purposes of this article. Modified Bessel functions of the second kind have the following integral representation for positive arguments:

$$K_p(x) = \frac{\Gamma(|p| + \frac{1}{2})(2x)^{|p|}}{\sqrt{\pi}} \int_0^\infty \frac{\cos(t)}{(x^2 + t^2)^{|p|+1/2}}\, dt, \quad x > 0$$

(8.432(5) of Gradshteyn and Ryzhik, 1994). Note that

$$K_{-p}(x) = K_p(x) \quad \text{for all } p, x \in \mathbb{R}$$

(8.486(16) of Gradshteyn and Ryzhik, 1994). The following recursion formula also holds for all $p, x \in \mathbb{R}$:

$$xK_{p+1}(x) = 2pK_p(x) + xK_{p-1}(x) \tag{S.1.1}$$

(8.486(10) of Gradshteyn and Ryzhik, 1994).

Computation of $K_p(x)$ for $p \in \mathbb{R}$ and $x > 0$ is supported by various software packages such as R (R Core Team, 2017). The R command:

```
besselK(x,p)
```

returns $K_p(x)$, where `x` and `p` denote the respective values of $p$ and $x$.

It is common in variational inference algorithms to have updates involving the *ratios* of modified Bessel functions of the second kind such as

$$\frac{K_{p+1}(x)}{K_p(x)}, \quad x > 0. \tag{S.1.2}$$

[*]School of Mathematical and Physical Sciences, University of Technology Sydney, P.O. Box 123, Broadway 2007, Australia, mathew.w.mclean@gmail.com

[†]School of Mathematical and Physical Sciences, University of Technology Sydney, P.O. Box 123, Broadway 2007, Australia, matt.wand@uts.edu.au

Care needs to be taken with this computation since, for example, the numerator and denominator may be infinitesimal even though the ratio is not. Wand and Ormerod (2012) describe remedies to this problem involving *continued fraction* representation of ratios such as (S.1.2). From their Table 1 we have

$$\frac{K_{p+1}(x)}{K_p(x)} = \frac{2p+2x+1}{2x} + \cfrac{(p^2-\frac{1}{4})/x}{2(x+1) + \cfrac{p^2-3^2/4}{2(x+2)+\cfrac{p^2-5^2/4}{2(x+3)+\cfrac{p^2-7^2/4}{2(x+4)+\cdots}}}}.$$

As explained in Wand and Ormerod (2012), Lentz's Algorithm (e.g. Press et al., 1992) can be used to obtain continued fraction approximation. Other ratios can be handled using (S.1.1).

#### The Special Case of $p$ Being Half an Odd Integer

In $p = \frac{1}{2}(2k+1)$ for some $k \in \mathbb{Z}$ then $K_p$ admits explicit expressions. For example,

$$K_{1/2}(x) = \sqrt{\frac{\pi}{2\,x}}\; e^{-x}, \quad x > 0$$

combined with (S.1.1) can be used to obtain explicit forms for other modified Bessel functions of the second kind having order equal to half of an odd integer such as

$$K_{3/2}(x) = \frac{x+1}{x}\sqrt{\frac{\pi}{2\,x}}\; e^{-x}, \quad x > 0.$$

This leads to

$$\frac{K_{3/2}(x)}{K_{1/2}(x)} = 1 + \frac{1}{x}, \quad x > 0. \tag{S.1.3}$$

### S.1.2 Parabolic Cylinder Functions

The *parabolic cylinder function* of order $\nu \in \mathbb{R}$, is denoted by $D_\nu$. The parabolic cylinder functions of *negative order* can be expressed in terms of a simple integral as follows:

$$D_\nu(x) = \Gamma(-\nu)^{-1}\exp(-x^2/4)\int_0^\infty t^{-\nu-1}\exp(-xt - \tfrac{1}{2}t^2)\,dt, \quad \nu < 0,\ x \in \mathbb{R}.$$

Note that only such negative order members of the parabolic cylinder family arise in this article's VMP algorithms. A recursion formula for parabolic cylinder functions is

$$D_{\nu+1}(x) = x\,D_\nu(x) - \nu\,D_{\nu-1}(x). \tag{S.1.4}$$

(9.247(1) of Gradshteyn and Ryzhik, 1994).

The VMP updates in Algorithms 3–5 involve the follow ratio function:

$$\mathcal{R}_\nu(x) \equiv \frac{D_{-\nu-2}(x)}{D_{-\nu-1}(x)}, \quad \nu > -1, x \in \mathbb{R}, \tag{S.1.5}$$

which is studied in Neville et al. (2014). Care needs to be taken in the computation of $\mathcal{R}_\nu(x)$ to avoid overflow and underflow. For positive arguments of $\mathcal{R}_\nu$ we have the very simple continued fraction expression

$$\mathcal{R}_\nu(x) = \cfrac{1}{x + \cfrac{\nu+1}{x + \cfrac{\nu+2}{x + \cfrac{\nu+3}{x + \cdots}}}}, \quad x > 0. \tag{S.1.6}$$

As explained in Neville et al. (2014) and encapsulated in their Algorithm 4 (S.1.6) combined with Lentz's Algorithm leads to stable and efficient computation of $\mathcal{R}_\nu(x)$ for $x > 0$. However, as opposed to the situation in Neville et al. (2014), we also need $\mathcal{R}_\nu(x)$ for $x \leq 0$ and we are not aware of a continued fraction representation for the non-positive argument case. For general $x$ we have

$$\mathcal{R}_\nu(x) = \frac{\mathcal{J}^+(\nu+1, -x, \frac{1}{2})}{(\nu+1)\mathcal{J}^+(\nu, -x, \frac{1}{2})}.$$

where $\mathcal{J}^+(p,q,r) \equiv \int_0^\infty x^p \exp(qx - rx^2)\, dx$ is as defined in Wand et al. (2011). As described in Appendix B of Wand et al. (2011) it advisable to work with the representation

$$\mathcal{J}^+(p,q,r) = e^M \int_0^\infty \exp\{p\log(x) + qx - rx^2 - M\}\, dx,$$
$$\text{where } M \equiv \sup\{x > 0 : p\log(x) + qx - rx^2\},$$

and logarithms to avoid underflow and overflow. Lastly, note that (S.1.4) gives rise to expressions such as

$$\frac{D_{-\nu-3}(x)}{D_{-\nu-1}(x)} = \frac{1 - x\,\mathcal{R}_\nu(x)}{\nu+2}.$$

This affords efficient computation of quantities arising in Algorithms 3, 4 and 5. Relevant details are in Section S.2.3.

## S.1.3 Additional Integral-Defined Functions

For $p, q, r \geq 0$ and $s < -r$ define

$$\mathcal{D}(p,q,r,s) \equiv \int_0^\infty [x\log(x) - \log\{\Gamma(x)\}]^p\, x^q \exp\big(r[x\log(x) - \log\{\Gamma(x)\}] + sx\big)\, dx.$$

Similarly, $p, q \geq 0$, $r < 0$ and $|s| < -r$ we define

$$\mathcal{E}(p,q,r,s) \equiv \int_{-\infty}^{\infty} x^p (1+x^2)^q \exp\left(rx^2 + sx\sqrt{1+x^2}\right) dx.$$

Note that $\mathcal{E}$ is a special case of the $\mathcal{G}$ family of functions defined in Wand et al. (2011). Appendix B of Wand et al. (2011) describes a strategy for stable computation of functions such as $\mathcal{D}$ and $\mathcal{E}$. As explained there, working with logarithms is especially important to avoid underflow and overflow. The VMP algorithms in this article depend on ratios of $\mathcal{D}$ and $\mathcal{E}$ and logarithm arithmetic is recommended for computing such ratios.

## S.2 Additional Exponential Family Distributions

Section S.1 of the online supplement of Wand (2017) summarizes common exponential families. In particular, the sufficient statistics and natural parameters for each distribution are given. If $x$ is a univariate random variable having an exponential family distribution then the sufficient statistic is denoted by $\boldsymbol{T}(x)$. VMP updates reduce to expectations of natural statistics and Table S.1 of Wand (2017) lists expressions for $E\{\boldsymbol{T}(x)\}$ for each of the exponential family distributions covered there.

In this section we add five more distributions to the list covered in Section S.1 of Wand (2017). One of them, the *Generalized Inverse Gaussian* distribution, is relatively well-known. Another is a distribution introduced and studied in Nadarajah (2008), which we simply call the *Nadarajah* distribution. The exponential family of distributions for which the reciprocal square root of its random variables have a Nadarajah distribution is an generalization of the Inverse Wishart conjugate family for squared scale parameters.

Lastly there are two exponential family distributions that arise in elaborate distribution VMP that, to the best of our knowledge, have not been identified in the statistical literature or given names. We have taken it upon ourselves to give them names in this article, since it aids readability as well as future applications and extensions of this work. Motivated by the fact that the distributions have new shapes, we have chosen the names *Moon Rock* distribution and *Sea Sponge* distribution.

### S.2.1 Generalized Inverse Gaussian

For any fixed $p \in \mathbb{R}$, the random variable $x$ has a Generalized Inverse Gaussian distribution with parameters $\alpha, \beta > 0$, written $x \sim \text{GIG}(\alpha, \beta; p)$, if the density function of $x$ is

$$p(x) = \frac{(\alpha/\beta)^{p/2}\, x^{p-1}}{2K_p(\sqrt{\alpha\beta})} \exp\left\{-\tfrac{1}{2}(\alpha\, x + \beta/x)\right\}, \quad x > 0,$$

where $K_p$ is the modified Bessel function of the second kind as described in Section S.1.1 of the online supplement. The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} x \\ 1/x \end{bmatrix} \quad \text{and} \quad h(x) = \tfrac{1}{2}\, x^{p-1} I(x > 0).$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} -a/2 \\ -b/2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -2\eta_1 \\ -2\eta_2 \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = \tfrac{1}{2}\, p \log(\eta_1/\eta_2) - \log K_p\big(2(\eta_1\eta_2)^{1/2}\big).$$

The expected value of the sufficient statistic is

$$E\{\boldsymbol{T}(x)\} = \begin{bmatrix} \dfrac{(\eta_2/\eta_1)^{1/2}\, K_{p+1}\big(2(\eta_1\eta_2)^{1/2}\big)}{K_p\big(2(\eta_1\eta_2)^{1/2}\big)} \\[2ex] \dfrac{(\eta_1/\eta_2)^{1/2}\, K_{p+1}\big(2(\eta_1\eta_2)^{1/2}\big)}{K_p\big(2(\eta_1\eta_2)^{1/2}\big)} + \dfrac{p}{\eta_2} \end{bmatrix}.$$

It follows from (S.1.3) that for the special case of $p = \frac{1}{2}$ we have

$$E\{\boldsymbol{T}(x)\} = \begin{bmatrix} \{\eta_1/(2\eta_2)\}^{1/2} - 1/(2\eta_2) \\[1ex] (\eta_1/\eta_2)^{1/2} \end{bmatrix}. \tag{S.2.1}$$

### S.2.2 Nadarajah Distribution

The random variable $x$ has the distribution introduced in Nadarajah (2008) with parameters $\alpha, \beta > 0$ and $\gamma \in \mathbb{R}$, written $x \sim \text{Nadarajah}(\alpha, \beta, \gamma)$, if the density function of $x$ is

$$p(x) = (2\beta)^{\alpha/2}/[\exp\{\gamma^2/(8\beta)\}\Gamma(\alpha)D_{-\alpha}(\gamma/\sqrt{2\beta})]\, x^{\alpha-1}\exp(-\beta\, x^2 - \gamma\, x), \quad x > 0.$$

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} \log(x) \\ x \\ x^2 \end{bmatrix} \quad \text{and} \quad h(x) = I(x > 0).$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} \alpha - 1 \\ -\gamma \\ -\beta \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \eta_1 + 1 \\ -\eta_3 \\ -\eta_2 \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = -\tfrac{1}{2}(\eta_1+1)\log(-2\eta_3) - \tfrac{1}{8}(\eta_2^2/\eta_3) + \log\{\Gamma(\eta_1+1)\} + \log\{D_{-\eta_1-1}(-\eta_2/\sqrt{-2\eta_3})\}$$

where the last term involves evaluation of the parabolic cylinder function of order $-\eta_1 - 1$. See Section S.1.2 of the online supplement for details on this family of functions. From (3) of Nadarajah (2008), the expectation of the sufficient statistic is

$$E\{\boldsymbol{T}(x)\} = \begin{bmatrix} \displaystyle\int_0^\infty \log(x)\, x^{\eta_1} \exp(\eta_2\, x + \eta_3\, x^2)\, dx \\ \dfrac{(\eta_1+1)\, D_{-\eta_1-2}(-\eta_2/\sqrt{-2\eta_3})}{\sqrt{-2\eta_3}\, D_{-\eta_1-1}(-\eta_2/\sqrt{-2\eta_3})} \\ \dfrac{(\eta_1+1)(\eta_1+2)\, D_{-\eta_1-3}(-\eta_2/\sqrt{-2\eta_3})}{(-2\eta_3)\, D_{-\eta_1-1}(-\eta_2/\sqrt{-2\eta_3})} \end{bmatrix}.$$

The integral in the first entry of $E\{\boldsymbol{T}(x)\}$ is expressible in terms of established special functions. However, this expectation is not needed for any of this article's algorithms.

### S.2.3 Inverse Square Root Nadarajah Distribution

A random variable $x$ has an Inverse Square Root Nadarajah distribution with parameters $\alpha, \beta > 0$ and $\gamma \in \mathbb{R}$, written $x \sim \text{Inverse-Square-Root-Nadarajah}(\alpha, \beta, \gamma)$, if and only if $1/\sqrt{x} \sim \text{Nadarajah}(\alpha, \beta, \gamma)$. Here we are using the same naming convention as used for the Log Normal distribution, where the transformation is the one applied to the new random variable to get to the established distribution.

The corresponding density function is

$$p(x) = (2\beta)^{\alpha/2}/[2\exp\{\gamma^2/(8\beta)\}\Gamma(\alpha)D_{-\alpha}(\gamma/\sqrt{2\beta})]\, x^{-(\alpha/2)-1}\exp(-\beta/x - \gamma/\sqrt{x}),\ x > 0.$$

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} \log(x) \\ 1/\sqrt{x} \\ 1/x \end{bmatrix} \quad \text{and} \quad h(x) = I(x > 0).$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} -(\alpha/2)-1 \\ -\gamma \\ -\beta \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} -2(\eta_1+1) \\ -\eta_3 \\ -\eta_2 \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = -\tfrac{1}{2}(\eta_1+1)\log(-2\eta_3) - \log(2) - \tfrac{1}{8}(\eta_2^2/\eta_3) + \log\{\Gamma(\eta_1+1)\} + \log\{D_{-\eta_1-1}(-\eta_2/\sqrt{-2\eta_3})\}.$$

The expectation of the sufficient statistic is

$$E\{\boldsymbol{T}(x)\} = \left[\begin{array}{c} \displaystyle\int_0^\infty \log(x)\, x^{\eta_1} \exp(\eta_2/\sqrt{x} + \eta_3/x)\, dx \\ \dfrac{-2(\eta_1+1)D_{2\eta_1+1}(-\eta_2/\sqrt{-2\eta_3})}{\sqrt{-2\eta_3}\, D_{2\eta_1+2}(-\eta_2/\sqrt{-2\eta_3})} \\ \dfrac{-(\eta_1+1)(2\eta_1+1)D_{2\eta_1}(-\eta_2/\sqrt{-2\eta_3})}{\eta_3\, D_{2\eta_1+2}(-\eta_2/\sqrt{-2\eta_3})} \end{array}\right]. \qquad \text{(S.2.2)}$$

Convenient notation based on (S.2.2), which we use in Algorithms 3–5, is

$$\begin{aligned} (E\boldsymbol{T})_2^{\text{ISRN}}(\boldsymbol{\eta}) &\equiv \frac{-2(\eta_1+1)D_{2\eta_1+1}(-\eta_2/\sqrt{-2\eta_3})}{\sqrt{-2\eta_3}\, D_{2\eta_1+2}(-\eta_2/\sqrt{-2\eta_3})} \\ \text{and} \quad (E\boldsymbol{T})_3^{\text{ISRN}}(\boldsymbol{\eta}) &\equiv \frac{-(\eta_1+1)(2\eta_1+1)D_{2\eta_1}(-\eta_2/\sqrt{-2\eta_3})}{\eta_3\, D_{2\eta_1+2}(-\eta_2/\sqrt{-2\eta_3})}. \end{aligned} \qquad \text{(S.2.3)}$$

See Section S.1.2 of the online supplement for advice concerning stable and efficient computation of $(E\boldsymbol{T})_2^{\text{ISRN}}(\boldsymbol{\eta})$ and $(E\boldsymbol{T})_3^{\text{ISRN}}(\boldsymbol{\eta})$.

## S.2.4 Moon Rock Distribution

The random variable $x$ has a Moon Rock distribution with parameters $\alpha > 0$ and $\beta > \alpha$, written $x \sim \text{Moon-Rock}(\alpha, \beta)$, if the density function of $x$ is

$$p(x) = \left[\int_0^\infty \{t^t/\Gamma(t)\}^\alpha \exp(-\beta\, t)\, dt\right]^{-1} \{x^x/\Gamma(x)\}^\alpha \exp(-\beta\, x), \quad x > 0.$$

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \left[\begin{array}{c} x\log(x) - \log\{\Gamma(x)\} \\ x \end{array}\right] \quad \text{and} \quad h(x) = I(x > 0).$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \left[\begin{array}{c} \eta_1 \\ \eta_2 \end{array}\right] = \left[\begin{array}{c} \alpha \\ -\beta \end{array}\right] \quad \text{and} \quad \left[\begin{array}{c} \alpha \\ \beta \end{array}\right] = \left[\begin{array}{c} \eta_1 \\ -\eta_2 \end{array}\right]$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = \log\left\{\int_0^\infty \{t^t/\Gamma(t)\}^{\eta_1} \exp(\eta_2\, t)\, dt\right\}. \qquad \text{(S.2.4)}$$

The expectation of the sufficient statistic is

$$E\{\boldsymbol{T}(x)\} = \exp\{-A(\boldsymbol{\eta})\} \left[\begin{array}{c} \displaystyle\int_0^\infty [x\log(x) - \log\{\Gamma(x)\}]\, \{x^x/\Gamma(x)\}^{\eta_1} e^{\eta_2\, x}\, dx \\ \displaystyle\int_0^\infty x\, \{x^x/\Gamma(x)\}^{\eta_1} \exp(\eta_2\, x)\, dx \end{array}\right]. \qquad \text{(S.2.5)}$$

It seems that the integrals appearing in this section are not expressible in terms of established special functions. The function $\mathcal{D}$ defined in (S.1.3) is tailor-made to summarize such integrals succinctly. Expressions (S.2.4) and (S.2.5) can be re-written:

$$A(\boldsymbol{\eta}) = \log\{\mathcal{D}(0,0,\eta_1,\eta_2)\} \quad \text{and} \quad E\{\boldsymbol{T}(x)\} = \begin{bmatrix} \mathcal{D}(1,0,\eta_1,\eta_2) \\ \mathcal{D}(0,1,\eta_1,\eta_2) \end{bmatrix} \Big/ \mathcal{D}(0,0,\eta_1,\eta_2).$$

Working with logarithms is strongly recommended to avoid underflow and overflow.

A convenient notation based on (S.2.5), which we use in Algorithms 1 and 2, is

$$\begin{aligned} (E\boldsymbol{T})_2^{\text{MR}}(\boldsymbol{\eta}) &\equiv \mathcal{D}(0,1,\eta_1,\eta_2)/\mathcal{D}(0,0,\eta_1,\eta_2) \\ &= \exp\big[\log\{\mathcal{D}(0,1,\eta_1,\eta_2)\} - \log\{\mathcal{D}(0,0,\eta_1,\eta_2)\}\big]. \end{aligned} \tag{S.2.6}$$

### S.2.5 Sea Sponge Distribution

The random variable $x$ has a Sea Sponge distribution with parameters $\alpha > 0$, $\beta > 0$ and $|\gamma| < \beta$, written $x \sim \text{Sea-Sponge}(\alpha,\beta,\gamma)$, if the density function of $x$ is

$$p(x) = \left\{\int_{-\infty}^{\infty} (1+t^2)^{\alpha} \exp\left(-\beta\, t^2 + \gamma\, t\sqrt{1+t^2}\right) dt\right\}^{-1} (1+x^2)^{\alpha} \times \exp\left(-\beta\, x^2 + \gamma\, x\sqrt{1+x^2}\right).$$

The sufficient statistic and base measure are

$$\boldsymbol{T}(x) = \begin{bmatrix} \log(1+x^2) \\ x^2 \\ x\sqrt{1+x^2} \end{bmatrix} \quad \text{and} \quad h(x) = 1.$$

The natural parameter vector and its inverse mapping are

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} \alpha \\ -\beta \\ \gamma \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \eta_1 \\ -\eta_2 \\ \eta_3 \end{bmatrix}$$

and the log-partition function is

$$A(\boldsymbol{\eta}) = \log\left\{\int_{-\infty}^{\infty} (1+t^2)^{\eta_1} \exp\left(\eta_2\, t^2 + \eta_3\, t\sqrt{1+t^2}\right) dt\right\}.$$

The expectation of the sufficient statistic is

$$E\{\boldsymbol{T}(x)\} = e^{-A(\boldsymbol{\eta})} \begin{bmatrix} \int_{-\infty}^{\infty} \log(1+x^2)\,(1+x^2)^{\eta_1} \exp\left(\eta_2\, x^2 + \eta_3\, x\sqrt{1+x^2}\right) dx \\ \int_{-\infty}^{\infty} x^2\,(1+x^2)^{\eta_1} \exp\left(\eta_2\, x^2 + \eta_3\, x\sqrt{1+x^2}\right) dx \\ \int_{-\infty}^{\infty} x\sqrt{1+x^2}\,(1+x^2)^{\eta_1} \exp\left(\eta_2\, x^2 + \eta_3\, x\sqrt{1+x^2}\right) dx \end{bmatrix}. \tag{S.2.7}$$

The log-partition function and expected sufficient statistic can be written as

$$A(\boldsymbol{\eta}) = \log\{\mathcal{E}(0, \eta_1, \eta_2, \eta_3)\}$$

where the function $\mathcal{E}$ is defined in Section S.1.3 of the online supplement. Notation analogous to that given for the Inverse Square Root Nadarajah and Moon Rock distributions, based on (S.2.7), is:

$$\begin{aligned}(E\boldsymbol{T})^{\text{ss}}_2(\boldsymbol{\eta}) &\equiv \mathcal{E}(2, \eta_1, \eta_2, \eta_3)/\mathcal{E}(0, \eta_1, \eta_2, \eta_3)\\ &= \exp\big[\log\{\mathcal{E}(2, \eta_1, \eta_2, \eta_3)\} - \log\{\mathcal{E}(0, \eta_1, \eta_2, \eta_3)\}\big]\end{aligned} \tag{S.2.8}$$

and

$$\begin{aligned}(E\boldsymbol{T})^{\text{ss}}_3(\boldsymbol{\eta}) &\equiv \mathcal{E}(1, \eta_1 + \tfrac{1}{2}, \eta_2, \eta_3)/\mathcal{E}(0, \eta_1, \eta_2, \eta_3)\\ &= \exp\big[\log\{\mathcal{E}(1, \eta_1 + \tfrac{1}{2}, \eta_2, \eta_3)\} - \log\{\mathcal{E}(0, \eta_1, \eta_2, \eta_3)\}\big]\end{aligned} \tag{S.2.9}$$

and appears in Algorithm 4.

# S.3 Derivations

Each of the fragment updates in Algorithms 1–6 involve repeated application of the VMP equations (4)–(6) and the occasional non-conjugate VMP (Knowles and Minka, 2011) modification. We now provide full derivational details.

Throughout these derivations we use 'const' to denote terms that do not depend on the variable of interest.

## S.3.1 Negative Binomial Likelihood Fragment Updates

From (5) and (6), the messages from $p(\boldsymbol{y}|\boldsymbol{a})$ to each of the $a_i$ are

$$m_{p(\boldsymbol{y}|\,\boldsymbol{a})\,\to\,a_i}(a_i) = \exp\left\{\begin{bmatrix}\log(a_i)\\ a_i\end{bmatrix}^T\begin{bmatrix}y_i\\ -1\end{bmatrix}\right\},\quad 1 \le i \le n.$$

Similarly, the messages from $p(\boldsymbol{a}|\boldsymbol{\theta}, \kappa)$ to each of the $a_i$ are, for $1 \le i \le n$,

$$m_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\to\,a_i}(a_i) = \exp\left\{\begin{bmatrix}\log(a_i)\\ a_i\end{bmatrix}^T\begin{bmatrix}\mu_{q(\kappa)} - 1\\ -\mu_{q(\kappa)}\,E_{q(\boldsymbol{\theta})}[\exp\{-(\boldsymbol{A\theta})_i\}]\end{bmatrix}\right\}$$

where $\mu_{q(\kappa)}$ is the mean of the density function formed by normalizing the message product:

$$m_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\to\,\kappa}(\kappa)\,m_{\kappa\,\to\,p(\boldsymbol{a}\,|\,\boldsymbol{\theta},\,\kappa)}(\kappa)$$

and $E_{q(\boldsymbol{\theta})}$ denotes expectation with respect to the normalization of

$$m_{p(\boldsymbol{a}|\boldsymbol{\theta},\,\kappa)\,\to\,\boldsymbol{\theta}}(\boldsymbol{\theta})\,m_{\boldsymbol{\theta}\,\to\,p(\boldsymbol{a}|\boldsymbol{\theta},\,\kappa)}(\boldsymbol{\theta}).$$

Since, from (4), $m_{a_i\,\rightarrow\,p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)}(a_i) \longleftarrow m_{p(\boldsymbol{y}|\,\boldsymbol{a})\,\rightarrow\,a_i}(a_i)$ we then have

$$
\begin{aligned}
q^*(a_i) &\propto m_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\rightarrow\,a_i}(a_i)\, m_{a_i\,\rightarrow\,p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)}(a_i)\\
&= \exp\left\{ \begin{bmatrix} \log(a_i) \\ a_i \end{bmatrix}^T \begin{bmatrix} y_i + \mu_{q(\kappa)} - 1 \\ -1 - \mu_{q(\kappa)}\, E_{q(\boldsymbol{\theta})}[\exp\{-(\boldsymbol{A\theta})_i\}] \end{bmatrix} \right\}, \quad 1 \le i \le n.
\end{aligned}
$$

This is proportional to a Gamma density function with mean

$$
E_{q(a_i)}(a_i) \equiv \frac{y_i + \mu_{q(\kappa)}}{1 + \mu_{q(\kappa)}\, E_{q(\boldsymbol{\theta})}[\exp\{-(\boldsymbol{A\theta})_i\}]} \tag{S.3.1}
$$

where $E_{q(a_i)}$ denotes expectation with respect to $q^*(a_i)$. The corresponding logarithmic expectations are

$$
E_{q(a_i)}(\log a_i) \equiv \text{digamma}(y_i + \mu_{q(\kappa)}) - \log\Big(1 + \mu_{q(\kappa)}\, E_{q(\boldsymbol{\theta})}[\exp\{-(\boldsymbol{A\theta})_i\}]\Big). \tag{S.3.2}
$$

For the message passed from $p(\boldsymbol{a}|\boldsymbol{\theta},\kappa)$ to $\kappa$ note that

$$
\begin{aligned}
\log\, p(\boldsymbol{a}|\boldsymbol{\theta},\kappa) \;=\; & \begin{bmatrix} \kappa\log(\kappa) - \log\{\Gamma(\kappa)\} \\ \kappa \end{bmatrix}^T \begin{bmatrix} n \\ -\boldsymbol{1}_n^T\boldsymbol{A\theta} + \boldsymbol{1}_n^T\log(\boldsymbol{a}) - \boldsymbol{a}^T\exp(-\boldsymbol{A\theta}) \end{bmatrix}\\
&+\text{const}.
\end{aligned}
$$

Hence

$$
m_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\rightarrow\,\kappa}(\kappa) = \exp\left\{ \begin{bmatrix} \kappa\log(\kappa) - \log\{\Gamma(\kappa)\} \\ \kappa \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\rightarrow\,\kappa} \right\} \tag{S.3.3}
$$

where

$$
\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\rightarrow\,\kappa} \longleftarrow \begin{bmatrix} n \\ -\boldsymbol{1}_n^T\boldsymbol{A}\,E_{q(\boldsymbol{\theta})}(\boldsymbol{\theta}) + \boldsymbol{1}_n^T E_{q(\boldsymbol{a})}\{\log(\boldsymbol{a})\} - E_{q(\boldsymbol{a})}(\boldsymbol{a})^T E_{q(\boldsymbol{\theta})}\{\exp(-\boldsymbol{A\theta})\} \end{bmatrix}.
$$

Note that (S.3.3) is proportional to a Moon Rock density function (defined in Section S.2.4 of the online supplement) and, under the constraint of conjugacy,

$$
m_{\kappa\,\rightarrow\,p(\boldsymbol{a}\,|\,\boldsymbol{\theta},\,\kappa)}(\kappa) = \exp\left\{ \begin{bmatrix} \kappa\log(\kappa) - \log\{\Gamma(\kappa)\} \\ \kappa \end{bmatrix}^T \boldsymbol{\eta}_{\kappa\,\rightarrow\,p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)} \right\}
$$

is also proportional to a Moon Rock density function and

$$
q^*(\kappa) \propto \exp\left\{ \begin{bmatrix} \kappa\log(\kappa) - \log\{\Gamma(\kappa)\} \\ \kappa \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\leftrightarrow\,\kappa} \right\}.
$$

Hence

$$\mu_{q(\kappa)} = \int_0^\infty \kappa\, q^*(\kappa)\, d\kappa \longleftarrow (E\boldsymbol{T})_2^{\text{MR}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\leftrightarrow\,\kappa}\Big)$$

where $(E\boldsymbol{T})_2^{\text{MR}}$ is given by (S.2.6).

The message passed from $p(\boldsymbol{a}|\boldsymbol{\theta},\kappa)$ to $\boldsymbol{\theta}$ is

$$m_{p(\boldsymbol{a}|\boldsymbol{\theta},\,\kappa)\,\rightarrow\,\boldsymbol{\theta}}(\boldsymbol{\theta}) = \exp\left[-\mu_{q(\kappa)}\{\boldsymbol{1}_n^T\boldsymbol{A\theta} + E_{q(\boldsymbol{a})}(\boldsymbol{a})^T\exp(-\boldsymbol{A\theta})\}\right]$$

which is not conjugate with Multivariate Normal messages passed to $\boldsymbol{\theta}$ from other factors. A non-conjugate variational message passing remedy (Knowles and Minka, 2011) is to replace $m_{p(\boldsymbol{a}|\boldsymbol{\theta},\,\kappa)\,\rightarrow\,\boldsymbol{\theta}}(\boldsymbol{\theta})$ with

$$\widetilde{m}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\rightarrow\,\boldsymbol{\theta}}(\boldsymbol{\theta}) \equiv \exp\left\{\begin{bmatrix}\boldsymbol{\theta}\\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\rightarrow\,\boldsymbol{\theta}}\right\}.$$

Working with $\widetilde{m}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\rightarrow\,\boldsymbol{\theta}}(\boldsymbol{\theta})$ instead of $m_{p(\boldsymbol{a}|\boldsymbol{\theta},\,\kappa)\,\rightarrow\,\boldsymbol{\theta}}(\boldsymbol{\theta})$ implies that $E_{q(\boldsymbol{\theta})}$ involves expectation with respect to a Multivariate Normal random vector and we get

$$E_{q(\boldsymbol{\theta})}\{\exp(-\boldsymbol{A\theta})\} \longleftarrow \boldsymbol{\omega}_2$$

where

$$\boldsymbol{\omega}_2 \equiv \exp\left(\boldsymbol{\omega}_1 - \tfrac{1}{4}\text{diagonal}\Big[\boldsymbol{A}\Big\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\boldsymbol{A}^T\Big]\right)$$

and

$$\boldsymbol{\omega}_1 \equiv \tfrac{1}{2}\boldsymbol{A}\Big\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_1.$$

Arguments similar to those given in Section S.2.3 of Wand (2017) lead to the updates in Algorithm 1.

### S.3.2 *t* Likelihood Fragment Updates

The log-likelihood in (11) is

$$\log\, p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a}) = -\tfrac{n}{2}\log(\sigma^2) - \tfrac{1}{2}\sum_{i=1}^n \log(a_i) - \tfrac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{A\theta})^T\text{diag}(\boldsymbol{a})^{-1}(\boldsymbol{y}-\boldsymbol{A\theta}) + \text{const}.$$

Arguments analogous to those used in Section 4.1.5 of Wand (2017) for the Gaussian likelihood fragment lead to

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\,\sigma^2,\,\boldsymbol{a})\,\rightarrow\,\boldsymbol{\theta}} \longleftarrow \mu_{q(1/\sigma^2)}\begin{bmatrix}\boldsymbol{A}^T\text{diag}\{E_{q(\boldsymbol{a})}(1/\boldsymbol{a})\}\boldsymbol{y}\\ -\tfrac{1}{2}\text{vec}\big(\boldsymbol{A}^T\text{diag}\{E_{q(\boldsymbol{a})}(1/\boldsymbol{a})\}\boldsymbol{A}\big)\end{bmatrix}$$

and

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\to\,\sigma^2} \longleftarrow \begin{bmatrix} -n/2 \\ G_{\text{VMP}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}};\boldsymbol{A}^T\text{diag}\{E_{q(\boldsymbol{a})}(1/\boldsymbol{a})\}\boldsymbol{A}, \\ \boldsymbol{A}^T\text{diag}\{E_{q(\boldsymbol{a})}(1/\boldsymbol{a})\}\boldsymbol{y},\boldsymbol{y}^T\text{diag}\{E_{q(\boldsymbol{a})}(1/\boldsymbol{a})\}\boldsymbol{y}\Big) \end{bmatrix}$$

where

$$\mu_{q(1/\sigma^2)} = \Big\{\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\sigma^2}\big)_1+1\Big\}\Big/\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\sigma^2}\big)_2.$$

Here $E_{q(\boldsymbol{a})}$ denotes expectation with respect to $q^*(\boldsymbol{a}) \equiv \prod_{i=1}^n q^*(a_i)$ and $q^*(a_i)$ is proportional to

$$\begin{aligned} & m_{p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\to\,a_i}(a_i)\, m_{p(\boldsymbol{a}|\nu)\,\to\,a_i}(a_i) \\ & = \exp\left\{\begin{bmatrix}\log(a_i)\\ 1/a_i\end{bmatrix}^T \begin{bmatrix} -\tfrac{1}{2}\mu_{q(\nu)}-\tfrac{3}{2} \\ \mu_{q(1/\sigma^2)}G_{\text{VMP}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}};\boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{A},\boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{y},y_i^2\Big) \\ -\tfrac{1}{2}\mu_{q(\nu)} \end{bmatrix}\right\} \end{aligned}$$

and $\mu_{q(\nu)} \equiv \int_0^\infty \nu\, q^*(\nu)\, d\nu$. Since $q^*(a_i)$ is an Inverse-$\chi^2$ density function the $i$th entry of $E_{q(\boldsymbol{a})}(1/\boldsymbol{a})$ is

$$E_{q(a_i)}(1/a_i) = \frac{-\tfrac{1}{2}\mu_{q(\nu)}-\tfrac{3}{2}+1}{\mu_{q(1/\sigma^2)}G_{\text{VMP}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}};\boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{A},\boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{y},y_i^2\Big)-\tfrac{1}{2}\mu_{q(\nu)}}.$$

Immediately it follows that

$$\boldsymbol{\omega}_5 \equiv E_{q(\boldsymbol{a})}(1/\boldsymbol{a}) = \frac{(\mu_{q(\nu)}+1)\mathbf{1}_n}{\mu_{q(\nu)}\mathbf{1}_n - 2\mu_{q(1/\sigma^2)}\boldsymbol{\omega}_4}$$

where $\boldsymbol{\omega}_4$ has $i$th entry equal to

$$(\boldsymbol{\omega}_4)_i \equiv G_{\text{VMP}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}};\boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{A},\boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{y},y_i^2\Big).$$

Also,

$$E_{q(a_i)}\{\log(a_i)\} = \log\Big(\tfrac{1}{2}\mu_{q(\nu)} - \mu_{q(1/\sigma^2)}(\boldsymbol{\omega}_4)_i\Big) - \text{digamma}\left(\frac{\mu_{q(\nu)}+1}{2}\right).$$

As a function of $\nu$ we have

$$\log\, p(\boldsymbol{a}|\nu) = n\{(\nu/2)\log(\nu/2) - \log\Gamma(\nu/2)\} - (\nu/2)\mathbf{1}_n^T\{\log(\boldsymbol{a}) + (1/\boldsymbol{a})\} + \text{const}.$$

Hence

$$m_{p(\boldsymbol{a}|\nu)\,\to\,\nu}(\nu) = \exp\left\{\begin{bmatrix}(\nu/2)\log(\nu/2)-\log\{\Gamma(\nu/2)\}\\ \nu/2\end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{a}|\,\nu)\,\to\,\nu}\right\} \qquad \text{(S.3.4)}$$

where this message's natural parameter is

$$\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\nu)\,\to\,\nu} \longleftarrow \begin{bmatrix} n \\ -\mathbf{1}_n^T\, E_{q(\boldsymbol{a})}\{\log(\boldsymbol{a}) + (1/\boldsymbol{a})\} \end{bmatrix}$$

which involves $\boldsymbol{\omega}_4$ and $\boldsymbol{\omega}_5$ given above. Note that (S.3.4) is proportional to a factor of 2 rescaling of a Moon Rock density function. Under conjugacy the message $m_{\nu\,\to\,p(\boldsymbol{a}|\nu)}(\nu)$ is in this same exponential family and, hence

$$q^*(\nu) \propto \exp\left\{ \begin{bmatrix} (\nu/2)\log(\nu/2) - \log\{\Gamma(\nu/2)\} \\ \nu/2 \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{a}|\,\nu)\,\leftrightarrow\,\nu} \right\}.$$

It follows that $\mu_{q(\nu)}$ is updated according to

$$\mu_{q(\nu)} \longleftarrow 2(E\boldsymbol{T})_2^{\text{MR}}\big(\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\nu)\,\leftrightarrow\,\nu}\big).$$

### S.3.3 Asymmetric Laplace Fragment Updates

From (13), the logarithm of the likelihood factor is

$$\log\, p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a}) = -\tfrac{n}{2}\log(\sigma^2) + \tfrac{1}{2}\sum_{i=1}^n \log(a_i)$$
$$-\tfrac{\tau(1-\tau)}{2\sigma^2}\left\{\boldsymbol{y} - \boldsymbol{A\theta} - \frac{(\frac{1}{2}-\tau)\sigma\mathbf{1}_n}{\tau(1-\tau)\boldsymbol{a}}\right\}^T \text{diag}(\boldsymbol{a})\left\{\boldsymbol{y} - \boldsymbol{A\theta} - \frac{(\frac{1}{2}-\tau)\sigma\mathbf{1}_n}{\tau(1-\tau)\boldsymbol{a}}\right\} + \text{const.}$$

Steps analogous to those given in Section 4.1.5 of Wand (2017) for the Gaussian likelihood fragment lead to the message from $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a})$ to $\boldsymbol{\theta}$ being Multivariate Normal with natural parameter update

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\to\,\boldsymbol{\theta}} \longleftarrow \tau(1-\tau)\mu_{q(1/\sigma^2)} \begin{bmatrix} \boldsymbol{A}^T\text{diag}\{E_{q(\boldsymbol{a})}(\boldsymbol{a})\}\boldsymbol{y} \\ -\tfrac{1}{2}\text{vec}\big(\boldsymbol{A}^T\text{diag}\{E_{q(\boldsymbol{a})}(\boldsymbol{a})\}\boldsymbol{A}\big) \end{bmatrix}$$
$$+(\tau - \tfrac{1}{2})\mu_{q(1/\sigma)} \begin{bmatrix} \boldsymbol{A}^T\mathbf{1}_n \\ \mathbf{0} \end{bmatrix}$$

where $\mu_{q(1/\sigma^k)} \equiv \int_0^\infty (1/\sigma^k)\, q^*(\sigma^2)\, d\sigma^2$ for $k = 1, 2$.

Noting that, as a function of $\sigma^2$,

$$\log\, p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a}) = -\tfrac{n}{2}\log(\sigma^2) + (\tfrac{1}{2}-\tau)\{\boldsymbol{y} - A\boldsymbol{\theta})\}^T\mathbf{1}_n(1/\sigma)$$
$$-\tfrac{1}{2}\tau(1-\tau)(\boldsymbol{y} - \boldsymbol{A\theta})^T\text{diag}(\boldsymbol{a})(\boldsymbol{y} - \boldsymbol{A\theta})(1/\sigma^2) + \text{const}$$

the message from $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a})$ to $\sigma^2$ is

$$m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\,\sigma^2,\,\boldsymbol{a})\,\to\,\sigma^2}(\sigma^2)=\exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma\\ 1/\sigma^2\end{bmatrix}^T\begin{bmatrix}-n/2\\ (\frac{1}{2}-\tau)\big\{\boldsymbol{y}-\boldsymbol{A}E_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})\big\}^T\boldsymbol{1}_n\\ -\frac{1}{2}\tau(1-\tau)\mathrm{tr}\big[E_{q(\boldsymbol{\theta})}\{(\boldsymbol{y}-\boldsymbol{A\theta})(\boldsymbol{y}-\boldsymbol{A\theta})^T\}\\ \times\,\mathrm{diag}\{E_{q(\boldsymbol{a})}(\boldsymbol{a})\}\big]\end{bmatrix}\right\} \tag{S.3.5}$$

where $E_{q(\boldsymbol{\theta})}$ denotes expectation with respect to the normalization of

$$m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\,\sigma^2,\,\boldsymbol{a})\,\to\,\boldsymbol{\theta}}(\boldsymbol{\theta})\,m_{\boldsymbol{\theta}\,\to\,p(\boldsymbol{y}|\,\boldsymbol{\theta},\,\sigma^2,\,\boldsymbol{a})}(\boldsymbol{\theta})$$

and $E_{q(\boldsymbol{a})}$ is defined similarly for $\boldsymbol{a}$. Conjugacy with (S.3.5) requires that the message from $\sigma^2$ to $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a})$ is also of the form

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

and this is the case provided that messages passed to $\sigma^2$ from other factors outside of the Asymmetric Laplace likelihood fragments are within or conjugate to the Inverse Square Root Nadarajah family. The optimal $q$-density for $\sigma^2$ is such that

$$q^*(\sigma^2)\propto\exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma\\ 1/\sigma^2\end{bmatrix}^T\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\,\sigma^2,\,\boldsymbol{a})\,\leftrightarrow\,\sigma^2}\right\}$$

and the $\mu_{q(1/\sigma)}$ and $\mu_{q(1/\sigma^2)}$ updates follow from (S.2.2).

The messages from $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a})$ to $a_i$, $1\le i\le n$, are

$$m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\,\sigma^2,\,\boldsymbol{a})\,\to\,a_i}(a_i)=a_i^{1/2}\exp\left\{\begin{bmatrix}a_i\\ 1/a_i\end{bmatrix}^T\begin{bmatrix}-\frac{1}{2}\tau(1-\tau)\mu_{q(1/\sigma^2)}E_{q(\boldsymbol{\theta})}\{(\boldsymbol{y}-\boldsymbol{A\theta})_i^2\}\\ \frac{1}{2}-\frac{1}{8\tau(1-\tau)}\end{bmatrix}\right\}I(a_i>0)$$

whilst those from $p(\boldsymbol{a})$ to $a_i$, $1\le i\le n$, are

$$m_{p(\boldsymbol{a})\,\to\,a_i}(a_i)=a_i^{-2}\exp\{-1/(2a_i)\}I(a_i>0).$$

This leads to the $q^*(a_i)$ being Inverse-Gaussian density functions and

$$E_{q(\boldsymbol{a})}(\boldsymbol{a})=\{-8\tau^2(1-\tau)^2\mu_{q(1/\sigma^2)}\boldsymbol{\omega}_7\}^{-1/2}\equiv\boldsymbol{\omega}_8$$

where

$$\boldsymbol{\omega}_7\equiv\Big[\;G_{\text{VMP}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\,\sigma^2,\,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}};\boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{A},\boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{y},y_i^2\Big)\;\Big]_{1\le i\le n}.$$

The updates in Algorithm 3 quickly follow.

### S.3.4 Skew Normal Fragment Updates

It follows from (15) that the logarithm of the likelihood factor is

$$\log\, p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}) = -\tfrac{n}{2}\log(\sigma^2)+\tfrac{n}{2}\log(1+\lambda^2)-\frac{1+\lambda^2}{2\sigma^2}\left\|\boldsymbol{y}-\boldsymbol{A\theta}-\frac{\lambda\sigma|\boldsymbol{a}|}{\sqrt{1+\lambda^2}}\right\|^2+\text{const},$$

where, here and elsewhere, $\|\boldsymbol{v}\|\equiv\sqrt{\boldsymbol{v}^T\boldsymbol{v}}$ for any vector $\boldsymbol{v}$. Using steps similar to those given in Section 4.1.5 of Wand (2017) for the Gaussian likelihood fragment, the message from $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})$ to $\boldsymbol{\theta}$ is proportional to a Multivariate Normal density function with natural parameter update

$$\begin{aligned}\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\to\boldsymbol{\theta}} \longleftarrow\ & \{1+\mu_{q(\lambda^2)}\}\mu_{q(1/\sigma^2)}\begin{bmatrix}\boldsymbol{A}^T\boldsymbol{y}\\ -\frac{1}{2}\text{vec}\big(\boldsymbol{A}^T\boldsymbol{A}\big)\end{bmatrix}\\ & -\mu_{q(\lambda\sqrt{\lambda^2+1})}\mu_{q(1/\sigma)}\begin{bmatrix}\boldsymbol{A}^TE_{q(\boldsymbol{a})}|\boldsymbol{a}|\\ \boldsymbol{0}\end{bmatrix}\end{aligned}$$

where $\mu_{q(1/\sigma^k)}\equiv\int_0^\infty(1/\sigma^k)\,q^*(\sigma^2)\,d\sigma^2$ for $k=1,2,$

$$\mu_{q(\lambda^2)}\equiv\int_{-\infty}^{\infty}\lambda^2\,q^*(\lambda)\,d\lambda\quad\text{and}\quad\mu_{q(\lambda\sqrt{\lambda^2+1})}\equiv\int_{-\infty}^{\infty}\lambda\sqrt{\lambda^2+1}\,q^*(\lambda)\,d\lambda.$$

The message from $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})$ to $\sigma^2$ is

$$m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\to\sigma^2}(\sigma^2) = \exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma\\ 1/\sigma^2\end{bmatrix}^T\begin{bmatrix}-n/2\\ \mu_{q(\lambda\sqrt{1+\lambda^2})}\big\{\boldsymbol{y}-\boldsymbol{A}\,E_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})\big\}^TE_{q(\boldsymbol{a})}|\boldsymbol{a}|\\ -\frac{1}{2}\big(1+\mu_{q(\lambda^2)}\big)E_{q(\boldsymbol{\theta})}\{\|\boldsymbol{y}-\boldsymbol{A\theta}\|^2\}\end{bmatrix}\right\} \tag{S.3.6}$$

which is in the Inverse Square Root Nadarajah Family (Section S.2.3 of the online supplement). The treatment of $m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\to\sigma^2}(\sigma^2)$ is analogous to that for the messages from the likelihood factor to $\sigma^2$ for the Asymmetric Laplace fragments.

The message from $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})$ to $\lambda$ is

$$m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\to\lambda}(\lambda) = \exp\left\{\begin{bmatrix}\log(1+\lambda^2)\\ \lambda^2\\ \lambda\sqrt{1+\lambda^2}\end{bmatrix}^T\begin{bmatrix}n/2\\ -\frac{1}{2}[\mu_{q(1/\sigma^2)}\,E_{q(\boldsymbol{\theta})}\{\|\boldsymbol{y}-\boldsymbol{A\theta}\|^2\}+E_{q(\boldsymbol{a})}\|\boldsymbol{a}\|^2]\\ \mu_{q(1/\sigma)}\big\{\boldsymbol{y}-\boldsymbol{A}\,E_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})\big\}^TE_{q(\boldsymbol{a})}|\boldsymbol{a}|\end{bmatrix}\right\} \tag{S.3.7}$$

which is proportional to density functions within the Sea Sponge exponential family defined in Section S.2.5 of the online supplement. Under the conjugacy restriction $q^*(\lambda)$

is a Sea Sponge density function with natural parameter $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\leftrightarrow\lambda}$. Using definitions (S.2.8) and (S.2.9) we then get

$$\mu_{q(\lambda^2)} \longleftarrow (E\boldsymbol{T})^{\rm SS}_2\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\leftrightarrow\lambda}\big)$$

and

$$\mu_{q(\lambda\sqrt{1+\lambda^2})} \longleftarrow (E\boldsymbol{T})^{\rm SS}_3\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\leftrightarrow\lambda}\big).$$

The messages from $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})$ to the $a_i$, $1\le i\le n$, are

$$m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\,\rightarrow\,a_i}(a_i)=\exp\left\{\left[\begin{array}{c}|a_i|\\[1ex] a_i^2\end{array}\right]^T\left[\begin{array}{c}\mu_{q(1/\sigma)}\mu_{q(\lambda\sqrt{1+\lambda^2})}\{\boldsymbol{y}-\boldsymbol{A}\,E_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})\}_i\\[1ex] -\frac{1}{2}\mu_{q(\lambda^2)}\end{array}\right]\right\}$$

whilst those from $p(\boldsymbol{a})$ to $a_i$ are

$$m_{p(\boldsymbol{a})\,\rightarrow\,a_i}(a_i)=\exp(-\tfrac{1}{2}\,a_i^2).$$

Hence

$$q^*(a_i)\propto\exp\left\{\left[\begin{array}{c}|a_i|\\[1ex] a_i^2\end{array}\right]^T\left[\begin{array}{c}\mu_{q(1/\sigma)}\mu_{q(\lambda\sqrt{1+\lambda^2})}\{\boldsymbol{y}-\boldsymbol{A}\,E_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})\}_i\\[1ex] -\frac{1}{2}\{\mu_{q(\lambda^2)}+1\}\end{array}\right]\right\}$$

and standard manipulations involving the Standard Normal distribution density and cumulative distribution functions lead to

$$E_{q(\boldsymbol{a})}|\boldsymbol{a}|=\boldsymbol{\omega}_{13}\qquad\text{and}\qquad E_{q(\boldsymbol{a})}\|\boldsymbol{a}\|^2=\frac{n+\boldsymbol{1}_n^T[\boldsymbol{\omega}_{12}\odot\{\boldsymbol{\omega}_{12}+\zeta'(\boldsymbol{\omega}_{12})\}]}{\mu_{q(\lambda^2)}+1}$$

where $\boldsymbol{\omega}_{12}$ and $\boldsymbol{\omega}_{13}$ are defined by the relevant updates in Algorithm 4.

We are now in a position to simplify the messages (S.3.6) and (S.3.7). The natural parameter update for the first of these messages is

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\,\rightarrow\,\sigma^2}\longleftarrow\left[\begin{array}{c}-n/2\\[2ex] \mu_{q(\lambda\sqrt{1+\lambda^2})}\,\boldsymbol{\omega}_{10}^T\boldsymbol{\omega}_{13}\\[1ex] \big(1+\mu_{q(\lambda^2)}\big)\omega_{11}\end{array}\right]$$

where

$$\omega_{11}\equiv G_{\rm VMP}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}};\boldsymbol{A}^T\boldsymbol{A},\boldsymbol{A}^T\boldsymbol{y},\boldsymbol{y}^T\boldsymbol{y}\Big).$$

That for the second is

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\,\rightarrow\,\lambda}\longleftarrow\left[\begin{array}{c}n/2\\[2ex] \mu_{q(1/\sigma^2)}\omega_{11}-\dfrac{n+\boldsymbol{1}_n^T[\boldsymbol{\omega}_{12}\odot\{\boldsymbol{\omega}_{12}+\zeta'(\boldsymbol{\omega}_{12})\}]}{2\{\mu_{q(\lambda^2)}+1\}}\\[1ex] \mu_{q(1/\sigma)}\,\boldsymbol{\omega}_{10}^T\boldsymbol{\omega}_{13}\end{array}\right].$$

### S.3.5 Finite Normal Mixture Fragment Updates

From (18), the logarithm of the likelihood factor is

$$\log\, p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a}) = -\frac{n}{2}\log(\sigma^2) + \sum_{i=1}^{n}\sum_{k=1}^{K} a_{ik}\left[-\tfrac{1}{2}\log(s_k^2) - \frac{1}{2s_k^2}\left(\frac{(\boldsymbol{y}-\boldsymbol{A\theta})_i}{\sigma} - m_k\right)^2\right] +\text{const.}$$

As with each of the previous derivations in this section, the message from the likelihood factor to $\boldsymbol{\theta}$ is proportional to a Multivariate Normal density function and takes the form

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\to\,\boldsymbol{\theta}} \longleftarrow \mu_{q(1/\sigma^2)}\begin{bmatrix}\boldsymbol{A}^T\text{diag}\{E_{q(\boldsymbol{a})}(\boldsymbol{\mathscr{A}})(\boldsymbol{1}_K/\boldsymbol{s}^2)\}\boldsymbol{y}\\ -\tfrac{1}{2}\text{vec}\big(\boldsymbol{A}^T\text{diag}\{E_{q(\boldsymbol{a})}(\boldsymbol{\mathscr{A}})(\boldsymbol{1}_K/\boldsymbol{s}^2)\}\boldsymbol{A}\big)\end{bmatrix}$$

$$-\mu_{q(1/\sigma)}\begin{bmatrix}\boldsymbol{A}^T\,E_{q(\boldsymbol{a})}(\boldsymbol{\mathscr{A}})(\boldsymbol{m}/\boldsymbol{s}^2)\\ \boldsymbol{0}\end{bmatrix}$$

where $\mu_{q(1/\sigma^k)} \equiv \int_0^\infty (1/\sigma^k)\, q^*(\sigma^2)\, d\sigma^2$ for $k=1,2$ and

$$\boldsymbol{\mathscr{A}} \equiv \begin{bmatrix}\boldsymbol{a}_1^T\\ \vdots\\ \boldsymbol{a}_n^T\end{bmatrix}.$$

The messages from $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a})$ to the $\boldsymbol{a}_i$, $1\le i\le n$, are

$$m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\to\,\boldsymbol{a}_i}(\boldsymbol{a}_i) = \exp\Bigg\{\boldsymbol{a}_i^T\Bigg(-\tfrac{1}{2}\log(\boldsymbol{s}^2) - \frac{\mu_{q(1/\sigma^2)}E_{q(\boldsymbol{\theta})}\{(\boldsymbol{y}-\boldsymbol{A\theta})_i\}^2\boldsymbol{1}_K - 2\mu_{q(1/\sigma)}\{\boldsymbol{y}-\boldsymbol{A}\,E_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})\}_i\boldsymbol{m} - \boldsymbol{m}^2}{2\boldsymbol{s}^2}\Bigg)\Bigg\}$$

whereas the messages from $p(\boldsymbol{a})$ to the $\boldsymbol{a}_i$ are

$$m_{p(\boldsymbol{a})\,\to\,\boldsymbol{a}_i}(\boldsymbol{a}_i) = \exp\{\boldsymbol{a}_i^T\log(\boldsymbol{w})\}.$$

For each $i$, both messages passed to $\boldsymbol{a}_i$ from its neighboring factors are proportional to Multinomial probability mass functions. Hence $q^*(\boldsymbol{a}_i)$ is a Multinomial probability mass function and standard calculations lead to $E_{q(\boldsymbol{a})}(\boldsymbol{\mathscr{A}}) = \boldsymbol{\Omega}_{18}$ where $\boldsymbol{\Omega}_{18}$ is defined by the updates given in Algorithm 5.

The message from $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a})$ to $\sigma^2$ is

$$m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\to\,\sigma^2}(\sigma^2) = \exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma\\ 1/\sigma^2\end{bmatrix}^T\begin{bmatrix}-n/2\\ \{\boldsymbol{y}-\boldsymbol{A}\,E_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})\}^T E_{q(\boldsymbol{a})}(\boldsymbol{\mathscr{A}})(\boldsymbol{m}/\boldsymbol{s}^2)\\ -\tfrac{1}{2}\,E_{q(\boldsymbol{\theta},\boldsymbol{a})}\Big[\big(\boldsymbol{y}-\boldsymbol{A\theta}\big)^T\text{diag}\{\boldsymbol{\mathscr{A}}(1/\boldsymbol{s}^2)\}\big(\boldsymbol{y}-\boldsymbol{A\theta}\big)\Big]\end{bmatrix}\right\}.$$

This means that the message passed from $m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\sigma^2}(\sigma^2)$ to $\sigma^2$ is proportional to an Inverse Square Root Nadarajah density function. Noting that

$$\{\boldsymbol{y}-\boldsymbol{A}\,E_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})\}^T E_{q(\boldsymbol{a})}(\boldsymbol{\mathscr{A}})(\boldsymbol{m}/\boldsymbol{s}^2)=\boldsymbol{\omega}_{15}^T\,\boldsymbol{\Omega}_{18}\,(\boldsymbol{m}/\boldsymbol{s}^2)$$

where $\boldsymbol{\omega}_{15}$ is defined by the update in Algorithm 5 and

$$\begin{aligned}
&-\tfrac{1}{2}\,E_{q(\boldsymbol{\theta},\boldsymbol{a})}\Big[\big(\boldsymbol{y}-\boldsymbol{A\theta}\big)^T\mathrm{diag}\{\boldsymbol{\mathscr{A}}(1/\boldsymbol{s}^2)\}\big(\boldsymbol{y}-\boldsymbol{A\theta}\big)\Big]\\
&\qquad=-\tfrac{1}{2}\,E_{q(\boldsymbol{\theta},\boldsymbol{a})}\Big[\boldsymbol{\theta}^T\boldsymbol{A}^T\mathrm{diag}\{\boldsymbol{\mathscr{A}}(\boldsymbol{1}_K/\boldsymbol{s}^2)\}\boldsymbol{A\theta}-2\boldsymbol{\theta}^T\boldsymbol{A}^T\mathrm{diag}\{\boldsymbol{\mathscr{A}}(\boldsymbol{1}_K/\boldsymbol{s}^2)\}\boldsymbol{y}\\
&\qquad\qquad+\boldsymbol{y}^T\mathrm{diag}\{\boldsymbol{\mathscr{A}}(\boldsymbol{1}_K/\boldsymbol{s}^2)\}\boldsymbol{y}\Big]\\
&\qquad=-\tfrac{1}{2}\,E_{q(\boldsymbol{\theta})}\Big[\boldsymbol{\theta}^T\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_{19})\boldsymbol{A\theta}-2\boldsymbol{\theta}^T\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_{19})\boldsymbol{y}+\boldsymbol{y}^T\mathrm{diag}(\boldsymbol{\omega}_{19})\boldsymbol{y}\Big]\\
&\qquad=G_{\mathrm{VMP}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}};\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_{19})\boldsymbol{A},\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_{19})\boldsymbol{y},\boldsymbol{y}^T\mathrm{diag}(\boldsymbol{\omega}_{19})\boldsymbol{y}\Big),
\end{aligned}$$

with $\boldsymbol{\omega}_{19}\equiv E_{q(\boldsymbol{a})}(\boldsymbol{\mathscr{A}})(\boldsymbol{1}_K/\boldsymbol{s}^2)=\boldsymbol{\Omega}_{18}(\boldsymbol{1}_K/\boldsymbol{s}^2)$, the update for $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\sigma^2}$ in Algorithm 5 follows.

The arguments used to obtain the $\mu_{q(1/\sigma^k)}$ updates are similar to those given in Section S.3.3 for the Asymmetric Laplace fragment. Algorithm 5 ensues.

### S.3.6 Support Vector Machine Fragment Updates

According to (5) and (6), the messages from $\check{p}(\boldsymbol{a})$ to each of the $a_i$ are

$$m_{\check{p}(\boldsymbol{a})\,\rightarrow\,a_i}(a_i)=\check{p}(a_i)=I(a_i>0),\quad 1\le i\le n. \tag{S.3.8}$$

Similarly, the messages from $\check{p}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{a})$ to each of the $a_i$ are, for $1\le i\le n$,

$$\begin{aligned}
&m_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\rightarrow\,a_i}(a_i)\\
&\qquad=a_i^{-1/2}\exp\left\{\begin{bmatrix}a_i\\1/a_i\end{bmatrix}^T\begin{bmatrix}-\tfrac{1}{2}\\-\tfrac{1}{2}\,E_{q(\boldsymbol{\theta})}\big[\{(2y_i-1)(\boldsymbol{A\theta})_i-1\}^2\big]\end{bmatrix}\right\}
\end{aligned}\tag{S.3.9}$$

where $E_{q(\boldsymbol{\theta})}$ denotes expectation with respect to the normalized

$$m_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\rightarrow\,\boldsymbol{\theta}}(\boldsymbol{\theta})\,m_{\boldsymbol{\theta}\,\rightarrow\,\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})}(\boldsymbol{\theta}). \tag{S.3.10}$$

For the message from $\check{p}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{a})$ to $\boldsymbol{\theta}$ we first note that, as a function of $\boldsymbol{\theta}$,

$$\begin{aligned}
\log\,\check{p}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{a})\;&=\;-\tfrac{1}{2}\sum_{i=1}^n\left[\frac{\{1+a_i-(2y_i-1)(\boldsymbol{A\theta})_i\}^2}{a_i}\right]+\mathrm{const}\\
&=\begin{bmatrix}\boldsymbol{\theta}\\\mathrm{vec}(\boldsymbol{\theta\theta}^T)\end{bmatrix}^T\begin{bmatrix}\boldsymbol{A}^T\{(\boldsymbol{1}_n+\boldsymbol{1}_n/\boldsymbol{a})\odot(2\boldsymbol{y}-\boldsymbol{1}_n)\}\\-\tfrac{1}{2}\mathrm{vec}\big\{\boldsymbol{A}^T\mathrm{diag}\,(\boldsymbol{1}_n/\boldsymbol{a})\,\boldsymbol{A}\big\}\end{bmatrix}+\mathrm{const}
\end{aligned}$$

where 'const' denotes terms not depending on $\boldsymbol{\theta}$. Therefore

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

where

$$\boldsymbol{\eta}_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\rightarrow\,\boldsymbol{\theta}} \longleftarrow \begin{bmatrix}\boldsymbol{A}^T[\{\boldsymbol{1}_n + E_{q(\boldsymbol{a})}(\boldsymbol{1}_n/\boldsymbol{a})\}\odot(2\boldsymbol{y}-\boldsymbol{1}_n)]\\ -\frac{1}{2}\text{vec}\big[\boldsymbol{A}^T\text{diag}\{E_{q(\boldsymbol{a})}(\boldsymbol{1}_n/\boldsymbol{a})\}\boldsymbol{A}\big]\end{bmatrix} \tag{S.3.11}$$

and

$$E_{q(\boldsymbol{a})}(\boldsymbol{1}_n/\boldsymbol{a}) \equiv [E_{q(a_1)}(1/a_1),\ldots,E_{q(a_n)}(1/a_n)]^T$$

with $E_{q(a_i)}$ denoting expectation with respect to the normalized

$$q^*(a_i) \propto m_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\rightarrow\,a_i}(a_i)\, m_{\check{p}(\boldsymbol{a})\,\rightarrow\,a_i}(a_i), \quad 1 \le i \le n. \tag{S.3.12}$$

On combining (S.3.8), (S.3.9) and (S.3.12) it is apparent that $E_{q(a_i)}$ denotes expectation with respect to a Generalized Inverse Gaussian distribution with $p = \frac{1}{2}$ and natural parameter vector

$$\begin{bmatrix}-\frac{1}{2}\\ -\frac{1}{2}E_{q(\boldsymbol{\theta})}\big[\{(2y_i-1)(\boldsymbol{A}\boldsymbol{\theta})_i - 1\}^2\big]\end{bmatrix}.$$

Then, from (S.2.1) in Section S.2.1

$$\begin{aligned}E_{q(a_i)}(1/a_i) &= \big(E_{q(\boldsymbol{\theta})}\big[\{(2y_i-1)(\boldsymbol{A}\boldsymbol{\theta})_i-1\}^2\big]\big)^{-1/2}\\ &= \Big(\big[(2y_i-1)E_{q(\boldsymbol{\theta})}\big\{(\boldsymbol{A}\boldsymbol{\theta})_i\big\}-1\big]^2 + \text{Var}_{q(\boldsymbol{\theta})}\big\{(\boldsymbol{A}\boldsymbol{\theta})_i\big\}\Big)^{-1/2}\\ &= \Big[\big\{(2y_i-1)(\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})})_i-1\big\}^2 + (\boldsymbol{A}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T)_{ii}\Big]^{-1/2}\end{aligned}$$

where $\boldsymbol{\mu}_{q(\boldsymbol{\theta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$ are the common parameters of the Multivariate Normal that arises from normalization of (S.3.10). Now set the updates

$$\begin{aligned}&\boldsymbol{\omega}_{20} \longleftarrow \boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})}, \quad \boldsymbol{\omega}_{21} \longleftarrow \text{diagonal}(\boldsymbol{A}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T),\\ &\text{and} \quad \boldsymbol{\omega}_{22} \longleftarrow [\{(2\boldsymbol{y}-\boldsymbol{1}_n)\odot\boldsymbol{\omega}_{20}-\boldsymbol{1}_n\}^2+\boldsymbol{\omega}_{21}]^{-1/2}.\end{aligned} \tag{S.3.13}$$

Then the updates in Algorithm 6 follow from updates (S.3.11) and (S.3.13) with $\boldsymbol{\mu}_{q(\boldsymbol{\theta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$ replaced by their natural parameter counterparts according to (S.4) in the online supplement of Wand (2017).

## References

Gradshteyn, I. S. and Ryzhik, I. M. (1994). *Tables of Integrals, Series, and Products, Fifth Edition*. San Diego, California: Academic Press. MR1398882. 1, 2

Knowles, D. A. and Minka, T. (2011). "Non-conjugate variational message passing for multinomial and binary regression." In *Advances in Neural Information Processing Systems*, 1701–1709. 9, 11

Nadarajah, S. (2008). "A new model for symmetric and skewed data." *Probability in the Engineering and Informational Sciences*, 22: 261–271. MR2399727. doi: https://doi.org/10.1017/S0269964808000156. 4, 5, 6

Neville, S. E., Ormerod, J. T., and Wand, M. P. (2014). "Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies." *Electronic Journal of Statistics*, 8: 1113–1151. MR3263115. doi: https://doi.org/10.1214/14-EJS910. 3

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes: The Art of Scientific Computing, Second Edition*. New York: Cambridge University Press. MR1414682. 2

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ 1

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. MR3646558. doi: https://doi.org/10.1080/01621459.2016.1197833. 4, 11, 13, 15, 19

Wand, M. P. and Ormerod, J. T. (2012). "Continued fraction enhancement of Bayesian computing." *Stat*, 1: 31–41. 2

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Frühwirth, R. F. (2011). "Mean field variational Bayes for elaborate distributions." *Bayesian Analysis*, 6: 847–900. MR2869967. doi: https://doi.org/10.1214/11-BA631. 3, 4