Variational Message Passing for Elaborate Response Regression Models

M. W. McLean and M. P. Wand

Abstract. We build on recent work concerning message passing approaches to approximate fitting and inference for arbitrarily large regression models. The focus is on regression models where the response variable is modeled to have an elaborate distribution, which is loosely defined to mean a distribution that is more complicated than common distributions such as those in the Bernoulli, Poisson and Normal families. Examples of elaborate response families considered here are the Negative Binomial and t families. Variational message passing is more challenging due to some of the conjugate exponential families being non-standard and numerical integration being needed. Nevertheless, a factor graph fragment approach means the requisite calculations only need to be done once for a particular elaborate response distribution family. Computer code can be compartmentalized, including that involving numerical integration. A major finding of this work is that the modularity of variational message passing extends to elaborate response regression models.

MSC 2010 subject classifications: Primary 62F15, 62J05; secondary 62G08.

Keywords: Bayesian computing, factor graph, generalized additive models, generalized linear mixed models, mean field variational Bayes, support vector machine classification.

1 Introduction

We extend the variational message passing (VMP) body of work to accommodate elaborate response regression models. The notion of factor graph fragments, introduced in Wand (2017), is the vehicle for this extension. It affords a modular approach to mean field variational Bayes fitting and inference for large regression models. The factor graph fragment updates treated here only need to be derived and implemented once. Their addition to the variational message passing arsenal allows for fancier models, such as those having Negative Binomial and t responses, to be fitted.

VMP (Winn and Bishop, 2005; Minka, 2005; Minka and Winn, 2008) is a prescription for obtaining mean field variational Bayes approximations to posterior density functions that is amenable to modularization. The factor graph version of VMP (e.g Minka and Winn, 2008, Appendix A) is particularly attractive in this regard. Wand (2017) uses the notion of factor graph fragments to aid modularization for semiparametric regression models – a large class of regression-type models that includes, for example, generalized linear mixed models, generalized additive models and varying coefficient models

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

DOI: 0000

(e.g. Ruppert et al., 2003). However, the fragments in Wand (2017) only accommodate Gaussian, Bernoulli and Poisson response models. If, for example, a Negative Binomial response model is of interest then new fragment updates for this family are needed. Section 3.1 plugs this gap. Other elaborate response families are also treated in Section 3. Whilst we do not cover all possible families, our derivations for some elaborate families provide blueprints for future fragment derivations.

A major difference between simple response models and elaborate response models is that the latter involves non-standard exponential families. For the examples covered here four exponential families, beyond those covered in Wand (2017), emerge. Two of them seem to have little or no presence in the literature. The sufficient statistic expectations, which are needed for VMP updates, are not expressible in terms of common functions and require either evaluation of special functions, quadrature or continued fraction approximation.

The main contributions of this article may be summarized as follows:

1. If an analyst wants to build a mean field variational Bayes inference engine for arbitrarily large regression models then the message update formulae given in Section 3 allow for particular elaborate response families to be included;

2. The derivations in Section S.3 of the online supplement show how such update formulae can be obtained for the examples given in Section 3. They also serve as a template for handling other elaborate response likelihoods not covered here.

All of our new methodology is within the realm of deterministic variational approximate inference, with intractable integrals evaluated via quadrature. An alternative route is to use Monte Carlo methods to approximate such integrals, known as *stochastic* variational inference (e.g. Hoffman et al., 2013; Kucukelbir et al., 2017). See, for example, Titsias and Lázaro-Gredilla (2014) on the use of stochastic variational inference for non-conjugate circumstances similar to those arising in this article.

Some background on VMP is given in Section 2. Section 3 is the article's centerpiece and gives the fragment update for six elaborate response likelihoods. Illustration of their utility is then provided in Section 4. Closing remarks are given in Section 5. Derivational details are given in an online supplement.

2 Variational Message Passing and Factor Graph Fragments

Variational message passing (VMP) is an approach to obtaining mean field variational Bayes approximate posterior density functions in potentially large graphical models. It uses the concept of *message passing* on a *factor graph*.

Our starting point is a Bayesian statistical model with observed data $\boldsymbol{D}$ and parameter vector $\boldsymbol{\theta}$. The posterior density function $p(\boldsymbol{\theta}|\boldsymbol{D})$ is usually analytically intractable and a *mean field variational* approximation $q^*(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|\boldsymbol{D})$ is the minimizer of the Kullback-Leibler divergence

$$\int q(\boldsymbol{\theta}) \log \left\{ \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{D})} \right\} d\boldsymbol{\theta}$$

subject to the product density restriction $q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q(\boldsymbol{\theta}_i)$ where $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ is some partition of $\boldsymbol{\theta}$. The optimal q-density functions can be shown to satisfy

$$q^*(\boldsymbol{\theta}_i) \propto E_{q(\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i)}\{p(\boldsymbol{\theta}_i|\boldsymbol{D}, \boldsymbol{\theta}\backslash\boldsymbol{\theta}_i)\}, \quad 1 \leq i \leq M, \tag{2.1}$$

where $\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i$ denotes the entries of $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_i$ omitted. Expression (2.1) gives rise to an iterative scheme for determination of the optimal parameters of the $q^*(\boldsymbol{\theta}_i)$, which is known as *mean field variational Bayes*. A listing of such a scheme is provided by Algorithm 1 of Ormerod and Wand (2010).

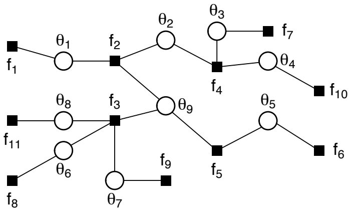

Figure 1: *Factor graph representation of the dependence of the stochastic nodes* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_9$ *on the factors* $f_1, \ldots, f_{11}$ *for the example given by (2.2).*

VMP arrives at the same approximation via message passing on an appropriate factor graph. Figure 1 is an example factor graph corresponding to an $M = 9$ example with

$$\begin{aligned} p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_9, \boldsymbol{D}) = f_1(\boldsymbol{\theta}_1) f_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_9) f_3(\boldsymbol{\theta}_6, \boldsymbol{\theta}_7, \boldsymbol{\theta}_8, \boldsymbol{\theta}_9) f_4(\boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4) f_5(\boldsymbol{\theta}_5, \boldsymbol{\theta}_9) \\ \times f_6(\boldsymbol{\theta}_5) f_7(\boldsymbol{\theta}_3) f_8(\boldsymbol{\theta}_6) f_9(\boldsymbol{\theta}_7) f_{10}(\boldsymbol{\theta}_4) f_{11}(\boldsymbol{\theta}_8). \end{aligned} \tag{2.2}$$

At least one of the f_j involves the data vector $\boldsymbol{D}$, but this dependence is suppressed. The unshaded circles are called *stochastic nodes* and the shaded rectangles are the *factors*. The word *node* is used for either a stochastic node or a factor and two nodes are neighbors of each other if they are joined by an edge. The edges join factors to stochastic nodes that are included in that factor. The θ_i indices connected to the jth factor are denoted by neighbors(j). For example, neighbors$(3) = \{6, 7, 8, 9\}$. Fuller details are in Sections 2.4 and 2.5 of Wand (2017).

A *message* passed between any two neighboring nodes is a particular function of the stochastic node that either sends or receives the message. Rather than using (2.1), the optimal q-densities are obtained from

$$q^*(\boldsymbol{\theta}_i) \propto \prod_{j:i\in\text{neighbors}(j)} m^*_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \tag{2.3}$$

where the $m^*_{f_j\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)$ are the optimal messages passed to $\boldsymbol{\theta}_i$ from each of the factors f_j in $p(\boldsymbol{\theta},\boldsymbol{D})$ that involve $\boldsymbol{\theta}_i$. For each j, this subset of $\{1,\ldots,M\}$ is denoted by neighbors(j) due to the definition of a factor graph, in which an edge is drawn between the θ_i and f_j nodes if and only if f_j depends on θ_i.

Letting N denote the number of factors, for each $1\le i\le M$ and $1\le j\le N$ the VMP stochastic node to factor message updates are

$$m_{\boldsymbol{\theta}_i\to\ f_j}(\boldsymbol{\theta}_i)\longleftarrow\propto\prod_{j'\neq j:\,i\in\text{neighbors}(j')} m_{f_{j'}\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \tag{2.4}$$

and the factor to stochastic node message updates are

$$m_{f_j\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)\longleftarrow\propto\exp\Big[E_{f_j\to\boldsymbol{\theta}_i}\Big\{\log f_j(\boldsymbol{\theta}_{\text{neighbors}(j)}\,)\Big\}\Big] \tag{2.5}$$

where $E_{f_j\to\boldsymbol{\theta}_i}$ denotes expectation with respect to the density function

$$\frac{\displaystyle\prod_{i'\in\text{neighbors}(j)\backslash\{i\}} m_{f_j\to\boldsymbol{\theta}_{i'}}(\boldsymbol{\theta}_{i'})\,m_{\boldsymbol{\theta}_{i'}\to f_j}(\boldsymbol{\theta}_{i'})}{\displaystyle\prod_{i'\in\text{neighbors}(j)\backslash\{i\}}\int m_{f_j\to\boldsymbol{\theta}_{i'}}(\boldsymbol{\theta}_{i'})\,m_{\boldsymbol{\theta}_{i'}\to f_j}(\boldsymbol{\theta}_{i'})\,d\boldsymbol{\theta}_{i'}}. \tag{2.6}$$

In (2.4) and (2.5) the $\longleftarrow\propto$ symbol means that the function of $\boldsymbol{\theta}_i$ on the left-hand side is updated according to the expression on the right-hand side but that multiplicative factors not depending on $\boldsymbol{\theta}_i$ can be ignored. If neighbors$(j)\backslash\{i\}=\emptyset$ then the expectation in (2.5) can be dropped and the right-hand side of (2.5) is proportional to $f_j(\boldsymbol{\theta}_{\text{neighbors}(j)}\,)$.

VMP fitting involves iteration of the updates (2.4) and (2.5)-(2.6) over each of the factors until the changes in all messages are negligible. When convergence is reached, the optimal q-densities of the model parameters are obtained from (2.3).

The algebra and coding for VMP can be compartmentalized using the notion of *factor graph fragments*, or *fragments* for short.

Definition: A *factor graph fragment*, or *fragment* for short, is a sub-graph of a factor graph consisting of a single factor and each of the stochastic nodes that are neighbors of the factor.

In the context of the current article, the fragment approach means that switching from a large regression-type model with a Gaussian likelihood to one with, say, a t likelihood can be achieved by replacing the Gaussian likelihood fragment by t likelihood fragments. The remainder of the model is unaffected in terms of the VMP updates.

Table 1 of Wand (2017) lists five fragments that are fundamental to semiparametric regression analysis via VMP. As explained there, a wide range of semiparametric regression models are accommodated by these five fragments but only for the Gaussian response case. In Section 5 of Wand (2017), additional fragments are introduced to handle logistic, probit and Poisson regression models. The next section adds to these response fragments.

3 Fragment Updates for Elaborate Response Likelihoods

We now provide fragment updates that allow for six more response distributions to be handled within the VMP framework. Most of them may be viewed as elaborations of the likelihoods covered by Wand (2017). For example, the Negative Binomial likelihood extends the Poisson likelihood for count response data and the t and Skew Normal likelihoods extend the Gaussian likelihood in different ways.

Each of the elaborate response likelihoods considered in this section are re-expressed in terms of auxiliary variables and more common distributions. This affords tractability, but comes at the cost of less accuracy compared with the case where auxiliary variables are not introduced. The auxiliary variables route is driven by the practical advantages of message updates being either closed form or requiring only univariate numerical integration. The alternative route, without auxiliary variables, is much more numerically challenging and often impractical.

Table 1 provides details on each of the distributions used in this article. It uses the following notation for the $N(0,1)$ density and cumulative distribution functions:

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

An additional functional notation is $\text{digamma}(x) \equiv \frac{d}{dx}\log\{\Gamma(x)\}$.

For a vector $\boldsymbol{a}$ and scalar function s we let $s(\boldsymbol{a})$ denote the vector containing the element-wise evaluations of s. Also, $\boldsymbol{A} \odot \boldsymbol{B}$ and $\boldsymbol{A}/\boldsymbol{B}$ respectively denote the element-wise product and element-wise quotient of vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ having the same sizes. If $\boldsymbol{A}$ is a $d \times d$ 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 column-wise fashion in order from left to right. The $d \times 1$ vector containing the diagonal entries of a $d \times d$ matrix $\boldsymbol{A}$ is denoted by $\text{diagonal}(\boldsymbol{A})$.

The $d \times 1$ vector $\boldsymbol{1}_d$ is such that all of its entries are equal to 1. The $d \times 1$ vector $\boldsymbol{e}_i$ is such that its ith entry is equal to 1 and all other entries are zero.

For a $d \times 1$ vector $\boldsymbol{v}_1$ and a $d^2 \times 1$ vector $\boldsymbol{v}_2$ such that $\text{vec}^{-1}(\boldsymbol{v}_2)$ is symmetric we define:

$$\begin{aligned} G_{\text{VMP}}\left(\begin{bmatrix}\boldsymbol{v}_1\\ \boldsymbol{v}_2\end{bmatrix}; \boldsymbol{Q}, \boldsymbol{r}, s\right) \equiv & -\tfrac{1}{8}\,\text{tr}\Big(\boldsymbol{Q}\{\text{vec}^{-1}(\boldsymbol{v}_2)\}^{-1}[\boldsymbol{v}_1\boldsymbol{v}_1^T\{\text{vec}^{-1}(\boldsymbol{v}_2)\}^{-1} - 2\boldsymbol{I}]\Big) \\ & -\tfrac{1}{2}\boldsymbol{r}^T\{\text{vec}^{-1}(\boldsymbol{v}_2)\}^{-1}\boldsymbol{v}_1 - \tfrac{1}{2}s. \end{aligned}$$

The secondary arguments of G_{VMP} are a $d \times d$ matrix $\boldsymbol{Q}$, a $d \times 1$ vector $\boldsymbol{r}$ and $s \in \mathbb{R}$. The genesis of the G_{VMP} function is the fact that

$$E_{\boldsymbol{\theta}}\{-\tfrac{1}{2}(\boldsymbol{\theta}^T\boldsymbol{Q}\boldsymbol{\theta} - 2\boldsymbol{r}^T\boldsymbol{\theta} + s)\} = G_{\text{VMP}}(\boldsymbol{\eta}; \boldsymbol{Q}, \boldsymbol{r}, s)$$

distribution	density/probability function in x	abbreviation
Multinomial	$\prod_{k=1}^{K} \pi_k^{x_k};\ x_k = 0,1,\ 1 \le k \le K;\ \sum_{k=1}^{K} \pi_k = 1$	Multinomial$(1, \boldsymbol{\pi})$
Poisson	$\lambda^x\, e^{-\lambda}/x!;\quad x = 0,1,\ldots;\ \lambda > 0$	Poisson(λ)
Negative Binomial	$\dfrac{\kappa^\kappa \Gamma(x+\kappa)\mu^x}{\Gamma(\kappa)(\kappa+\mu)\Gamma(x+1)};\ x = 0,1\ldots;\ \kappa,\mu > 0$	Negative-Binomial(μ,κ)
t	$\dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma(\nu/2)[1+\{(x-\mu)/\sigma\}^2/\nu]^{\frac{\nu+1}{2}}};\ \sigma,\nu > 0$	$t(\mu,\sigma,\nu)$
Asymmetric Laplace	$\frac{\tau(1-\tau)}{\sigma} \exp\left[-\frac{1}{2}\left\lvert\frac{x-\mu}{\sigma}\right\rvert + (\tau - \frac{1}{2})\left(\frac{x-\mu}{\sigma}\right)\right];$ $\sigma > 0, \quad 0 < \tau < 1$	Asymmetric-Laplace(μ,σ,τ)
Skew Normal	$\dfrac{2}{\sigma}\phi\left(\dfrac{x-\mu}{\sigma}\right)\Phi\left(\dfrac{\lambda(x-\mu)}{\sigma}\right);\ \sigma > 0$	Skew-Normal(μ,σ,λ)
Finite Normal Mixture	$\sum_{k=1}^{K} \dfrac{w_k}{\sigma\, s_k}\phi\left(\dfrac{(x-\mu)/\sigma - m_k}{s_k}\right);$ $w_k, s_k > 0,\ \sum_{k=1}^{K} w_k = 1$	Normal-Mixture$(\mu,\sigma,\boldsymbol{w},\boldsymbol{m},\boldsymbol{s})$
Gamma	$\dfrac{B^A\, x^{A-1} e^{-B\,x}}{\Gamma(A)};\quad x > 0,\ A,B > 0$	Gamma(A,B)
Inverse-χ^2	$\dfrac{(\lambda/2)^{\kappa/2}\, x^{-(\kappa/2)-1} e^{-(\lambda/2)/x}}{\Gamma(\kappa/2)};\quad x > 0;\ \kappa,\lambda > 0$	Inverse-$\chi^2(\kappa,\lambda)$

Table 1: *Distributions used in this article and their corresponding density/probability functions.*

when $\boldsymbol{\theta}$ is a $d \times 1$ Multivariate Normal random vector with natural parameter vector $\boldsymbol{\eta}$. A last piece of notation is

$$\boldsymbol{\eta}_{f\leftrightarrow\boldsymbol{\theta}} \equiv \boldsymbol{\eta}_{f\rightarrow\boldsymbol{\theta}} + \boldsymbol{\eta}_{\boldsymbol{\theta}\rightarrow f}$$

for any natural parameter $\boldsymbol{\eta}$, factor f and stochastic node $\boldsymbol{\theta}$.

3.1 Negative Binomial Likelihood

The Negative Binomial likelihood fragments are concerned with the likelihood specification

$$y_i|\boldsymbol{\theta},\kappa \overset{\text{ind.}}{\sim} \text{Negative-Binomial}[\exp\{(\boldsymbol{A\theta})_i\},\kappa], \quad 1 \le i \le n. \tag{3.1}$$

Introduce Gamma auxiliary random variables $a_i|\,\boldsymbol{\theta},\kappa \overset{\text{ind.}}{\sim} \text{Gamma}[\kappa,\kappa\,\exp\{-(\boldsymbol{A\theta})_i\}]$, $1 \le i \le n$. Then standard distribution theoretical manipulations lead to (3.1) being equivalent to

$$y_i|\,a_i \overset{\text{ind.}}{\sim} \text{Poisson}(a_i), \quad a_i|\,\boldsymbol{\theta},\kappa \overset{\text{ind.}}{\sim} \text{Gamma}[\kappa,\kappa\,\exp\{-(\boldsymbol{A\theta})_i\}].$$

The relevant factor graph fragments are shown in Figure 2 and corresponds to the mean field restriction

$$q(\boldsymbol{\theta},\kappa,\boldsymbol{a})=q(\boldsymbol{\theta})q(\kappa)\left\{\prod_{i=1}^{n}q(a_i)\right\}$$

that was used in Luts and Wand (2015).

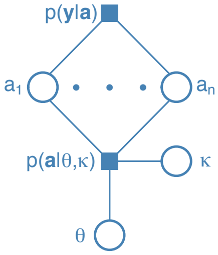

Figure 2: *Fragments for the Negative Binomial likelihood specification with independent Gamma auxiliary variables* $a_1,\ldots,a_n$.

First note that

$$m_{p(\boldsymbol{a}|\boldsymbol{\theta},\kappa)\,\to\,\boldsymbol{\theta}}(\boldsymbol{\theta})=\exp\Big[-E_{q(\kappa)}(\kappa)\big\{\mathbf{1}_n^T\boldsymbol{A}\boldsymbol{\theta}+E_{q(\boldsymbol{a})}(\boldsymbol{a})^T\exp(-\boldsymbol{A}\,\boldsymbol{\theta})\big\}\Big] \tag{3.2}$$

which is not conjugate with Multivariate Normal messages passed to $\boldsymbol{\theta}$ from other factors. Instead, we replace (3.2) with

$$\widetilde{m}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\to\,\boldsymbol{\theta}}(\boldsymbol{\theta})\equiv\exp\left\{\begin{bmatrix}\boldsymbol{\theta}\\ \mathrm{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\to\,\boldsymbol{\theta}}\right\} \tag{3.3}$$

to enforce conjugacy with Multivariate Normal messages. This is an instance of non-conjugate VMP (Knowles and Minka, 2011). We assume that each of the messages that $\boldsymbol{\theta}$ receives from factors outside of the Negative Binomial likelihood fragments are within the Multivariate Normal family. This leads to $q^*(\boldsymbol{\theta})$ having a Multivariate Normal distribution.

As explained in Section S.3.1 of the online supplement, the message from $p(\boldsymbol{a}|\boldsymbol{\theta},\kappa)$ to κ takes the form

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

which is proportional to the *Moon Rock* exponential family of density functions described in Section S.2.4 of the online supplement. We assume messages passed to κ from factors outside of the Negative Binomial likelihood fragments are also within the Moon Rock family or at least conjugate with the Moon Rock family. For example, if the only other factor passing messages to κ is its prior density function $p(\kappa)$ then we require that $p(\kappa)$ is a Moon Rock density function or conjugate with one. Note that, for example, Exponential density functions (Gamma$(1, B)$ density functions in the notation of Table 1) are conjugate with respect to the Moon Rock family but, strictly speaking, not within the Moon Rock family since $\alpha = 0$ in the notation of Section S.2.4. Hence, setting

$$p(\kappa) = B\exp(-B\,\kappa), \quad \kappa > 0,$$

for any $B > 0$ is permissible under the conjugacy constraint since it implies that

$$m_{p(\kappa)\,\rightarrow\,\kappa}(\kappa) = \exp\left\{\begin{bmatrix}\kappa\log(\kappa) - \log\{\Gamma(\kappa)\}\\ \kappa\end{bmatrix}^T\begin{bmatrix}0\\ -B\end{bmatrix}\right\}.$$

which is conjugate with respect to $m_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\kappa)\,\rightarrow\,\kappa}(\kappa)$.

Algorithm 1 lists the inputs, updates and outputs for the Negative Binomial likelihood fragments. The derivations are given in Section S.3.1 of the online supplement. The $(\boldsymbol{E}\boldsymbol{T})_2^{\text{MR}}$ notation, used in the first update, is explained in Section S.2.4 of the online supplement.

In Section 4.2 we provide illustration of Algorithm 1 in the context of additive model analysis.

3.2 *t* Likelihood

The t-distribution likelihood fragments arise from the likelihood specification

$$y_i|\,\boldsymbol{\theta},\sigma,\nu \overset{\text{ind.}}{\sim} t\Big((\boldsymbol{A\theta})_i,\sigma,\nu\Big), \quad 1 \le i \le n. \tag{3.4}$$

This likelihood is frequently used in regression applications as a robustness mechanism (e.g. Lange et al., 1989). If we introduce Inverse-χ^2 auxiliary random variables $a_i \overset{\text{ind.}}{\sim}$ Inverse-$\chi^2(\nu,\nu)$, $1 \le i \le n$, then (3.4) is equivalent to

$$y_i|\,\boldsymbol{\theta},\sigma^2,a_i \overset{\text{ind.}}{\sim} N\Big((\boldsymbol{A\theta})_i,a_i\sigma^2\Big), \quad a_i|\,\nu \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2(\nu,\nu). \tag{3.5}$$

It is common to use this representation of the t distribution for Bayesian computing. For example, the Markov chain Monte Carlo scheme of Verdinelli and Wasserman (1991) and the mean field variational Bayes scheme of Tipping and Lawrence (2003) each rely upon (3.5).

Figure 3 shows the factor graph fragments for the auxiliary variable representation (3.5) with q-density product restriction

$$q(\boldsymbol{\theta},\sigma^2,\nu,\boldsymbol{a}) = q(\boldsymbol{\theta})q(\sigma^2)q(\nu)\left\{\prod_{i=1}^n q(a_i)\right\}.$$

Data Inputs: $\boldsymbol{y}, \boldsymbol{A}$.

Parameter Inputs: $\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\kappa)\,\rightarrow\,\boldsymbol{\theta}}, \boldsymbol{\eta}_{\boldsymbol{\theta}\,\rightarrow\,p(\boldsymbol{a}|\,\boldsymbol{\theta},\kappa)}, \boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\kappa)\,\rightarrow\,\kappa}, \boldsymbol{\eta}_{\kappa\,\rightarrow\,p(\boldsymbol{a}|\,\boldsymbol{\theta},\kappa)}$.

Updates:

$$\mu_{q(\kappa)} \longleftarrow (\boldsymbol{ET})_2^{\mathrm{MR}}\big(\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\kappa)\,\leftrightarrow\,\kappa}\big)$$

$$\boldsymbol{\omega}_1 \longleftarrow -\tfrac{1}{2}\boldsymbol{A}\Big\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\big(\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_1$$

$$\boldsymbol{\omega}_2 \longleftarrow \exp\Bigg(-\boldsymbol{\omega}_1 - \tfrac{1}{4}\mathrm{diagonal}\Big[\boldsymbol{A}\Big\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\boldsymbol{A}^T\Big]\Bigg)$$

$$\boldsymbol{\omega}_3 \longleftarrow \Big\{\boldsymbol{\omega}_2 \odot \big(\boldsymbol{y} + \mu_{q(\kappa)}\boldsymbol{1}_n\big)\Big\}\Big/\big(\boldsymbol{1}_n + \mu_{q(\kappa)}\,\boldsymbol{\omega}_2\big)$$

$$\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\kappa)\,\rightarrow\,\boldsymbol{\theta}} \longleftarrow \mu_{q(\kappa)}\begin{bmatrix} \boldsymbol{A}^T\Big\{\boldsymbol{\omega}_3 \odot (\boldsymbol{1}_n + \boldsymbol{\omega}_1) - \boldsymbol{1}_n\Big\} \\ -\tfrac{1}{2}\mathrm{vec}\big(\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_3)\boldsymbol{A}\big) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\kappa)\,\rightarrow\,\kappa} \longleftarrow \begin{bmatrix} n \\ \boldsymbol{1}_n^T\Big\{\mathrm{digamma}\big(\mu_{q(\kappa)}\boldsymbol{1}_n + \boldsymbol{y}\big) - \boldsymbol{\omega}_1 \\ -\log\big(\boldsymbol{1}_n + \mu_{q(\kappa)}\boldsymbol{\omega}_2\big) - \boldsymbol{\omega}_3\Big\} \end{bmatrix}$$

Parameter Outputs: $\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\kappa)\,\rightarrow\,\boldsymbol{\theta}}, \boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\kappa)\,\rightarrow\,\kappa}$.

Algorithm 1: *The inputs, updates and outputs of the Negative Binomial likelihood fragment.*

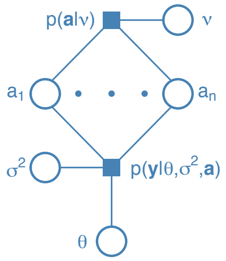

Figure 3: *Fragments for the t likelihood specification with the shape parameter prior with independent* Inverse-$\chi^2(\nu,\nu)$ *auxiliary variables* $a_1,\ldots,a_n$.

The message from $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a})$ to $\boldsymbol{\theta}$ is proportional to a Multivariate Normal density function, while that from $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a})$ to σ^2 is within the Inverse-χ^2 family.

The message from $p(\boldsymbol{a}|\nu)$ to ν has the form

$$m_{p(\boldsymbol{a}|\nu)\,\rightarrow\,\nu}(\nu)=\exp\left\{\left[\begin{array}{c}(\nu/2)\log(\nu/2)-\log\{\Gamma(\nu/2)\}\\ \nu/2\end{array}\right]^T\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\nu)\,\rightarrow\,\nu}\right\}$$

with details given in Section S.3.2 of the online supplement. Note that $m_{p(\boldsymbol{a}|\nu)\,\rightarrow\,\nu}(\nu)$ is proportional to a factor of 2 rescaling of the Moon Rock exponential family of density functions introduced in Section S.2.4 of the online supplement. The conjugacy constraint dictates that

$$m_{\nu\,\rightarrow\,p(\boldsymbol{a}|\nu)}(\nu)=\exp\left\{\left[\begin{array}{c}(\nu/2)\log(\nu/2)-\log\{\Gamma(\nu/2)\}\\ \nu/2\end{array}\right]^T\boldsymbol{\eta}_{\nu\,\rightarrow\,p(\boldsymbol{a}|\,\nu)}\right\}$$

which occurs if all message passed to ν from factors outside of the t likelihood fragments are also within the same rescaled Moon Rock family, or at least conjugate with respect to it. The $(E\boldsymbol{T})_2^{\mathrm{MR}}$ notation is defined in Section S.2.4 of the online supplement.

Algorithm 2 provides the inputs, updates and outputs for the t likelihood fragments. The derivations are given in Section S.3.2 of the online supplement.

3.3 Asymmetric Laplace Likelihood

Now consider the Asymmetric Laplace likelihood specification

$$y_i|\,\boldsymbol{\theta},\sigma^2\overset{\text{ind.}}{\sim}\text{Asymmetric-Laplace}\big((\boldsymbol{A\theta})_i,\sigma,\tau\big),\quad 1\le i\le n,\tag{3.6}$$

where $0<\tau<1$ is a fixed constant. As explained in, for example, Yu and Moyeed (2001), the likelihood specification (3.6) corresponds to τth-quantile regression. Yang et al. (2016) discuss valid posterior inference for Bayesian quantile regression.

If we introduce auxiliary random variables $a_i\overset{\text{ind.}}{\sim}\text{Inverse-}\chi^2(2,1)$, $1\le i\le n$, then Proposition 3.2.1 of Kotz et al. (2001) implies that (3.6) is equivalent to

$$y_i|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a}\overset{\text{ind.}}{\sim}N\left((\boldsymbol{A\theta})_i+\frac{(\frac{1}{2}-\tau)\sigma}{a_i\tau(1-\tau)},\frac{\sigma^2}{a_i\tau(1-\tau)}\right),\quad a_i\overset{\text{ind.}}{\sim}\text{Inverse-}\chi^2(2,1).\tag{3.7}$$

We assume that the optimal q-density admits the product restriction

$$q(\boldsymbol{\theta},\sigma^2,\boldsymbol{a})=q(\boldsymbol{\theta})q(\sigma^2)q(\boldsymbol{a})=q(\boldsymbol{\theta})q(\sigma^2)\prod_{i=1}^n q(a_i).$$

The corresponding factor graph fragments are shown in Figure 4.

As shown in Section S.3.3 of the online supplement,

$$m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\boldsymbol{\theta}}(\boldsymbol{\theta})=\exp\left\{\left[\begin{array}{c}\boldsymbol{\theta}\\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{array}\right]^T\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\boldsymbol{\theta}}\right\}$$

Data Inputs: $\boldsymbol{y}, \boldsymbol{A}$.

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

$\boldsymbol{\eta}_{\sigma^2\,\to\,p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})}, \boldsymbol{\eta}_{p(\boldsymbol{a}|\,\nu)\,\to\,\nu}, \boldsymbol{\eta}_{\nu\,\to\,p(\boldsymbol{a}|\,\nu)}.$

Updates:

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

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

$$\boldsymbol{\omega}_4 \longleftarrow \Big[\ G_{\mathrm{VMP}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}}; \boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{A}, \boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{y}, y_i^2\Big)\ \Big]_{1\le\imath\le n}$$

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

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\to\,\boldsymbol{\theta}} \longleftarrow \mu_{q(1/\sigma^2)}\begin{bmatrix} \boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_5)\boldsymbol{y} \\ -\frac{1}{2}\mathrm{vec}\big(\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_5)\boldsymbol{A}\big) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\to\,\sigma^2} \longleftarrow \begin{bmatrix} -\frac{1}{2}\,n \\ G_{\mathrm{VMP}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}}; \boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_5)\boldsymbol{A}, \\ \boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_5)\boldsymbol{y}, \boldsymbol{y}^T\mathrm{diag}(\boldsymbol{\omega}_5)\boldsymbol{y}\Big) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\nu)\,\to\,\nu} \longleftarrow \begin{bmatrix} n \\ \\ n\,\mathrm{digamma}\left(\frac{\mu_{q(\nu)}+1}{2}\right) - \mathbf{1}_n^T\Big\{\log\Big(\frac{1}{2}\,\mu_{q(\nu)}\mathbf{1}_n - \mu_{q(1/\sigma^2)}\boldsymbol{\omega}_4\Big) \\ +\boldsymbol{\omega}_5\Big\} \end{bmatrix}$$

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

Algorithm 2: *The inputs, updates and outputs of the t likelihood fragment.*

which is conjugate with Multivariate Normal messages passed to $\boldsymbol{\theta}$ from factors outside of the Asymmetric Laplace likelihood fragments.

However, the message from the likelihood factor to σ^2 takes the form

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

which is not within a standard exponential family. However, $m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\to\,\sigma^2}(\sigma^2)$ is

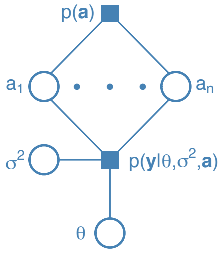

Figure 4: *Fragments for the Asymmetric Laplace likelihood specification with independent* Inverse-$\chi^2(2,1)$ *auxiliary variables* $a_1,\ldots,a_n$.

proportional to the family of density functions of random variables such that their reciprocal square roots are distributed according to members of a family proposed in Nadarajah (2008). Sections S.2.2 and S.2.3 of the online supplement contain the relevant details. We will assume that messages passed to σ^2 from factors outside of the Asymmetric Laplace likelihood fragments are within the *Inverse Square Root Nadarajah* family (Section S.2.3 of the online supplement). Note that messages proportional to Inverse Chi-Squared density functions are conjugate with this family.

Algorithm 3 provides the inputs, updates and outputs for the Asymmetric Laplace likelihood fragments with derivations deferred to Section S.3.3 of the online supplement. Note that the second update of Algorithm 3 involves the function $\mathcal{R}_\nu$, which is defined in Section S.1.2 of the online supplement. Efficient and stable computation of $\mathcal{R}_\nu$ is discussed there.

In Section 4.1 we show that Algorithm 3 facilitates quantile nonparametric regression embellishment of ordinary nonparametric regression.

Laplace Likelihood Special Case

The case of $\tau=\frac{1}{2}$ corresponds to the special case of the Laplace likelihood, and (3.6) reduces to *median* regression. In this special case, the second entry of $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\to\,\sigma^2}$ is zero and messages passed to σ^2 are proportional to Inverse Chi-Squared density functions. In addition, the $\mu_{q(1/\sigma)}$ update in Algorithm 3 is not needed and that for $\mu_{q(1/\sigma^2)}$ reduces to

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

where $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\sigma^2}$ is an Inverse Chi-Squared natural parameter vector.

3.4 Skew Normal Likelihood

In this section, we consider fragments involving the Skew Normal likelihood:

$$y_i|\,\boldsymbol{\theta},\sigma^2,\lambda\sim\text{Skew-Normal}\{(\boldsymbol{A\theta})_i,\sigma,\lambda\},\quad 1\le i\le n. \tag{3.8}$$

Data Inputs: $\boldsymbol{y}, \boldsymbol{A}, \tau$.

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

Updates:

$$\mu_{q(1/\sigma)} \longleftarrow (E\boldsymbol{T})_2^{\mathrm{ISRN}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\sigma^2}\Big)$$

$$\mu_{q(1/\sigma^2)} \longleftarrow (E\boldsymbol{T})_3^{\mathrm{ISRN}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\sigma^2}\Big)$$

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

$$\boldsymbol{\omega}_8 \longleftarrow \Big\{-8\,\tau^2(1-\tau)^2\,\mu_{q(1/\sigma^2)}\,\boldsymbol{\omega}_7\Big\}^{-1/2}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\boldsymbol{\theta}} \longleftarrow \tau(1-\tau)\mu_{q(1/\sigma^2)}\begin{bmatrix} \boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_8)\boldsymbol{y} \\ -\tfrac{1}{2}\mathrm{vec}\Big(\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_8)\boldsymbol{A}\Big) \end{bmatrix}$$

$$+\,(\tau-\tfrac{1}{2})\mu_{q(1/\sigma)}\begin{bmatrix} \boldsymbol{A}^T\boldsymbol{1}_n \\ \boldsymbol{0} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\sigma^2} \longleftarrow \begin{bmatrix} -n/2 \\ (\tfrac{1}{2}-\tau)\Big[\boldsymbol{y}+\tfrac{1}{2}\boldsymbol{A}\Big\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1} \\ \times\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_1\Big]^T\boldsymbol{1}_n \\ \tau(1-\tau)G_{\mathrm{VMP}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}}; \boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_8)\boldsymbol{A}, \\ \boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_8)\boldsymbol{y}, \boldsymbol{y}^T\mathrm{diag}(\boldsymbol{\omega}_8)\boldsymbol{y}\Big) \end{bmatrix}$$

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

Algorithm 3: *The inputs, updates and outputs of the Asymmetric Laplace likelihood fragments.*

Regression-type models having Skew Normal responses may be found in, for example, Frühwirth-Schnatter and Pyne (2010) and Lachos et al. (2010).

If we introduce auxiliary random variables $a_i \overset{\text{ind.}}{\sim} N(0,1),\ 1 \le i \le n$, then Proposition 3 of Azzalini and Dalla Valle (1996) implies that (3.8) is equivalent to

$$y_i|\,\boldsymbol{\theta},\sigma^2,\lambda,a_i \overset{\text{ind.}}{\sim} N\left((\boldsymbol{A\theta})_i+\frac{\sigma\lambda|a_i|}{\sqrt{1+\lambda^2}},\frac{\sigma^2}{1+\lambda^2}\right),\quad a_i \overset{\text{ind.}}{\sim} N(0,1). \tag{3.9}$$

We assume the optimal q-density admits the product restriction

$$q(\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})=q(\boldsymbol{\theta})q(\sigma^2)q(\lambda)q(\boldsymbol{a})=q(\boldsymbol{\theta})q(\sigma^2)q(\lambda)\prod_{i=1}^n q(a_i).$$

The corresponding factor graph fragments are shown in Figure 5.

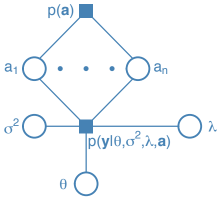

Figure 5: *Fragments for the Skew Normal likelihood specification with independent* $N(0,1)$ *auxiliary variables* $a_1,\ldots,a_n$ *.*

The messages passed from the likelihood factor to $\boldsymbol{\theta}$ and σ^2 take the forms

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

and

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

As for the Asymmetric Laplace likelihood fragments, the latter is within the Inverse Square Root Nadarajah family. The imposition of conjugacy means that we assume that all messages passed to σ^2 from factors outside of the Skew Normal likelihood fragments are also proportional to Inverse Square Root Nadarajah density functions.

The message from the likelihood factor to λ has the exponential family form

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

We have not been able to find any mention of this family in the literature. In Section S.2.5 of the online supplement we dub it the *Sea Sponge* family. We assume that each of the messages that λ receives from factors outside of the Skew Normal likelihood fragments are also proportional to Sea Sponge density functions. As an example, suppose that the only other factor that sends a message to λ is the prior density function $p(\lambda)$. Then, $m_{p(\lambda)\,\rightarrow\,\lambda}(\lambda) = p(\lambda)$ and, under conjugacy, $p(\lambda)$ must be of the form

$$p(\lambda) \propto \exp\left\{ \begin{bmatrix} \log(1+\lambda^2) \\ \lambda^2 \\ \lambda\sqrt{1+\lambda^2} \end{bmatrix}^T \boldsymbol{\eta}_\lambda \right\} \tag{3.10}$$

for some 3×1 vector $\boldsymbol{\eta}_\lambda$. Priors of the form $\lambda \sim N(0,\sigma_\lambda^2)$ are allowable under conjugacy constraints since these are a special case of (3.10) with $\boldsymbol{\eta}_\lambda = [0 \quad -1/(2\sigma_\lambda^2) \quad 0]^T$.

The message natural parameter updates depend on the first derivative of

$$\zeta(x) \equiv \log\{2\Phi(x)\} \quad \text{which leads to} \quad \zeta'(x) \equiv \frac{\phi(x)}{\Phi(x)}.$$

Software such as the function `zeta()` within the package `sn` (Azzalini, 2017) of the `R` computing environment (`R` Core Team, 2017) supports stable computation of ζ'.

Algorithm 4 provides the inputs, updates and outputs for the Skew Normal likelihood fragments. The $(\boldsymbol{ET})_2^{\text{SS}}$ and $(\boldsymbol{ET})_3^{\text{SS}}$ notation is explained in Section S.2.5 of the online supplement.

Justification for Algorithm 4 is given in Section S.3.4 of the online supplement.

3.5 Finite Normal Mixture Likelihood

The Finite Normal Mixture likelihood fragments involve the likelihood

$$y_i|\,\boldsymbol{\theta},\sigma^2 \overset{\text{ind.}}{\sim} \text{Normal-Mixture}\Big((\boldsymbol{A\theta})_i,\sigma,\boldsymbol{w},\boldsymbol{m},\boldsymbol{s}\Big), \quad 1\le i\le n, \tag{3.11}$$

where $\boldsymbol{w},\boldsymbol{m}$ and $\boldsymbol{s}$ are each constant $K\times 1$ vectors. Finite Normal Mixture approximation of difficult response density functions can be a "last resort" for development of tractable Bayesian inference algorithms. See, for example, Frühwirth-Schnatter and Wagner (2006) and Frühwirth-Schnatter et al. (2009). In the variational inference context, Wand et al. (2011) showed how Finite Normal Mixture approximation benefits variational inference for Generalized Extreme Value response models.

If we introduce auxiliary random variables $\boldsymbol{a}_i \equiv (a_{i1},\ldots,a_{iK})^T$ such that

$$\boldsymbol{a}_i \overset{\text{ind.}}{\sim} \text{Multinomial}(1,\boldsymbol{w}), \quad 1\le i\le n,$$

then we can re-express (3.11) as

$$\begin{aligned} &p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n) = \\ &\qquad \prod_{i=1}^n\prod_{k=1}^K\left[\sigma^{-1}(2\pi s_k^2)^{-1/2}\exp\left\{-\frac{1}{2s_k^2}\left(\frac{(\boldsymbol{y}-\boldsymbol{A\theta})_i}{\sigma}-m_k\right)^2\right\}\right]^{a_{ik}}, \\ &\qquad \boldsymbol{a}_i \overset{\text{ind.}}{\sim} \text{Multinomial}(1,\boldsymbol{w}). \end{aligned} \tag{3.12}$$

Data Inputs: $\boldsymbol{y}, \boldsymbol{A}$.

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

Updates:

$$\mu_{q(1/\sigma)} \longleftarrow (E\boldsymbol{T})_2^{\mathrm{ISRN}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\leftrightarrow\sigma^2}\Big)$$

$$\mu_{q(1/\sigma^2)} \longleftarrow (E\boldsymbol{T})_3^{\mathrm{ISRN}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\leftrightarrow\sigma^2}\Big)$$

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

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

$$\boldsymbol{\omega}_{10} \longleftarrow \boldsymbol{y} + \tfrac{1}{2}\boldsymbol{A}\Big\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\leftrightarrow\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\leftrightarrow\boldsymbol{\theta}}\big)_1$$

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

$$\boldsymbol{\omega}_{12} \longleftarrow \frac{\mu_{q(1/\sigma)}\,\mu_{q(\lambda\sqrt{1+\lambda^2})}\,\boldsymbol{\omega}_{10}}{\sqrt{1+\mu_{q(\lambda^2)}}} \quad ; \quad \boldsymbol{\omega}_{13} \longleftarrow \frac{\boldsymbol{\omega}_{12}+\zeta'(\boldsymbol{\omega}_{12})}{\sqrt{1+\mu_{q(\lambda^2)}}}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\to\boldsymbol{\theta}} \longleftarrow \{1+\mu_{q(\lambda^2)}\}\mu_{q(1/\sigma^2)}\begin{bmatrix} \boldsymbol{A}^T\boldsymbol{y} \\ -\frac{1}{2}\mathrm{vec}\big(\boldsymbol{A}^T\boldsymbol{A}\big) \end{bmatrix}$$
$$-\mu_{q(\lambda\sqrt{\lambda^2+1})}\mu_{q(1/\sigma)}\begin{bmatrix} \boldsymbol{A}^T\boldsymbol{\omega}_{13} \\ \boldsymbol{0} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\to\sigma^2} \longleftarrow \begin{bmatrix} -n/2 \\ \mu_{q(\lambda\sqrt{1+\lambda^2})}\,\boldsymbol{\omega}_{10}^T\boldsymbol{\omega}_{13} \\ \big\{1+\mu_{q(\lambda^2)}\big\}\omega_{11} \end{bmatrix}$$

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

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

Algorithm 4: *The inputs, updates and outputs of the Skew Normal likelihood fragments.*

Even though the $\boldsymbol{a}_i$ are vectors, we will use the abbreviation $\boldsymbol{a} \equiv \boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ from now onwards. The q-density product form we consider is

$$q(\boldsymbol{\theta},\sigma^2,\boldsymbol{a})=q(\boldsymbol{\theta})q(\sigma^2)q(\boldsymbol{a})=q(\boldsymbol{\theta})q(\sigma^2)\prod_{i=1}^{n}q(\boldsymbol{a}_i).$$

The factor graph fragments for the Finite Normal Mixture likelihood are shown in Figure 6.

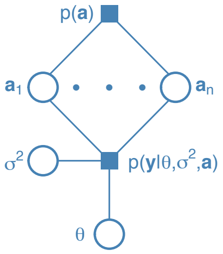

Figure 6: *Fragments for the Finite Normal Mixture likelihood specification with independent* Multinomial$(1,\boldsymbol{w})$ *auxiliary variables* $\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n$.

As in Sections 3.3 and 3.4, the conjugate distribution for σ^2 is the Inverse Square Root Nadarajah distribution (Section S.2.3 of the online supplement).

The inputs, updates and outputs for the Finite Normal Mixture likelihood fragments are listed in Algorithm 5, and justifications are in Section S.3.5 of the online supplement.

3.6 Support Vector Machine Pseudo-likelihood

Luts and Ormerod (2014) derived mean field variational Bayes algorithms for support vector machine classification using the auxiliary variable representation of the hinge loss psuedo-likelihood of Polson and Scott (2011). The approach is founded upon the following result:

$$\int_0^\infty (2\pi\,a)^{-1/2}\exp\left\{-\frac{(1+a-x)^2}{2a}\right\}da=\exp\{-2(1-x)_+\} \tag{3.13}$$

where $u_+\equiv\max(0,u)$ for any $u\in\mathbb{R}$. Letting $I(\mathcal{P})$ be the indicator of whether the proposition $\mathcal{P}$ is true, note that (3.13) can be re-expressed as follows:

$$\begin{gathered}\text{if } p(x|a) \text{ is the } N(a+1,a) \text{ density function in } x \text{ and } \check{p}(a)=I(a>0) \text{ then}\\ \check{p}(x)\equiv\int_{-\infty}^{\infty}p(x|a)\check{p}(a)\,da=\exp\{-2(1-x)_+\}.\end{gathered} \tag{3.14}$$

In (3.14) the pseudo-density function $\check{p}(x)$ is represented as a mixture of a particular Normal density function and the auxiliary variable pseudo-density function $\check{p}(a)$. As we will see, such a representation is amenable to the VMP updating equations with pseudo-density functions treated as ordinary density functions. As explained in Polson and Scott

```
**Data Inputs:** $\boldsymbol{y}, \boldsymbol{A}, K, \boldsymbol{w}, \boldsymbol{m}, \boldsymbol{s}$.
**Parameter Inputs:** $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\boldsymbol{\theta}}, \boldsymbol{\eta}_{\boldsymbol{\theta}\,\rightarrow\,p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})}, \boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\sigma^2}$,
                      $\boldsymbol{\eta}_{\sigma^2\,\rightarrow\,p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})}$.
**Updates:**
    $\mu_{q(1/\sigma)} \longleftarrow (\boldsymbol{ET})_2^{\mathrm{ISRN}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\sigma^2}\Big)$; $\mu_{q(1/\sigma^2)} \longleftarrow (\boldsymbol{ET})_3^{\mathrm{ISRN}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\sigma^2}\Big)$
    $\boldsymbol{\omega}_{15} \longleftarrow \boldsymbol{y} + \frac{1}{2}\boldsymbol{A}\Big\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_1$
    $\boldsymbol{\omega}_{16} \longleftarrow \Big[\;G_{\mathrm{VMP}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}}; \boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{A}, \boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{y}, y_i^2\Big)\;\Big]_{1\le\imath\le n}$
    $\boldsymbol{\Omega}_{17} \longleftarrow \mu_{q(1/\sigma)}\boldsymbol{\omega}_{15}(\boldsymbol{m}/\boldsymbol{s}^2)\Big)^T + \mu_{q(1/\sigma^2)}\boldsymbol{\omega}_{16}\big(\boldsymbol{1}_K/\boldsymbol{s}^2\big)^T + \boldsymbol{1}_n\Big\{\log(\boldsymbol{w}/\boldsymbol{s}) - (\boldsymbol{m}^2)/(2\boldsymbol{s}^2)\Big\}^T$
    $\boldsymbol{\Omega}_{18} \longleftarrow \exp(\boldsymbol{\Omega}_{17})/\{\exp(\boldsymbol{\Omega}_{17})\boldsymbol{1}_K\boldsymbol{1}_K^T\}$ ; $\boldsymbol{\omega}_{19} \longleftarrow \boldsymbol{\Omega}_{18}(\boldsymbol{1}_K/\boldsymbol{s}^2)$
    $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\boldsymbol{\theta}} \longleftarrow \mu_{q(1/\sigma^2)}\begin{bmatrix}\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_{19})\boldsymbol{y}\\ -\frac{1}{2}\mathrm{vec}\Big(\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_{19})\boldsymbol{A}\Big)\end{bmatrix} - \mu_{q(1/\sigma)}\begin{bmatrix}\boldsymbol{A}^T\boldsymbol{\Omega}_{18}(\boldsymbol{m}/\boldsymbol{s}^2)\\ \boldsymbol{0}\end{bmatrix}$
    $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\sigma^2} \longleftarrow \begin{bmatrix}-n/2\\ \boldsymbol{\omega}_{15}^T\boldsymbol{\Omega}_{18}(\boldsymbol{m}/\boldsymbol{s}^2)\\ G_{\mathrm{VMP}}\Big(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}}; \boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_{19})\boldsymbol{A},\\ \boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_{19})\boldsymbol{y}, \boldsymbol{y}^T\mathrm{diag}(\boldsymbol{\omega}_{19})\boldsymbol{y}\Big)\end{bmatrix}$
**Parameter Outputs:** $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\boldsymbol{\theta}}, \boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\sigma^2}$.
```

Algorithm 5: *The inputs, updates and outputs of the Finite Normal Mixture likelihood fragments.*

(2011), the hinge loss pseudo-density function could be replaced by an ordinary density function via normalization. However, the pseudo-density function version leads to the traditional support vector machine classifier.

The Support Vector Machine pseudo-likelihood fragments are concerned with the pseudo-likelihood specification

$$\check{p}(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n}\exp[-2\{1-(2y_i-1)(\boldsymbol{A\theta})_i\}_+] \tag{3.15}$$

where the $y_i \in \{0,1\}$ are indicators of class membership in a two-class classification setting. If we now introduce an auxiliary variable vector $\boldsymbol{a} = (a_1,\ldots,a_n)$ with entries

a_i, $1 \le i \le n$, with each independently having the pseudo-density function $\check{p}(a_i) = I(a_i > 0)$ then, using (3.14), (3.15) is equivalent to

$$\begin{aligned} \check{p}(\boldsymbol{y}|\boldsymbol{a},\boldsymbol{\theta}) &= \prod_{i=1}^{n} (2\pi\, a_i)^{-1/2} \exp\left[-\frac{\{1 + a_i - (2y_i - 1)(\boldsymbol{A\theta})_i\}^2}{2a_i} \right], \\ \check{p}(\boldsymbol{a}) &= \prod_{i=1}^{n} I(a_i > 0). \end{aligned} \tag{3.16}$$

The corresponding factor graph fragments are shown in Figure 7.

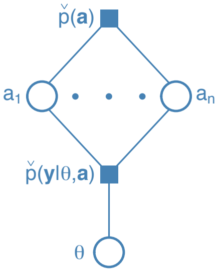

Figure 7: *Fragments for the Support Vector Machine pseudo-likelihood specification with independent auxiliary variables* $\boldsymbol{a} = (a_1, \ldots, a_n)$ *having psuedo-density function* $\check{p}(\boldsymbol{a}) = \prod_{i=1}^{n} I(a_i > 0)$.

Under the assumption that all messages passed to $\boldsymbol{\theta}$ are Multivariate Normal, Algorithm 6 provides updates for the natural parameter vector passed from $\check{p}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{a})$ to $\boldsymbol{\theta}$. An attractive feature of the Support Vector Machine pseudo-likelihood fragment updates is that each of them are simple closed form operations.

4 Illustrations

We now provide some illustrations of how the fragment updates of Section 3 can be used to move from one variational inference analysis to another, without having to start from scratch.

4.1 Ordinary to Quantile Nonparametric Regression

First consider ordinary nonparametric regression via the Bayesian mixed model-based penalized spline model used in Section 3.2.1 of Wand (2017). We quickly recap the details here. The data are the predictor/response pairs (x_i, y_i), $1 \le i \le n$, and the nonparametric regression model is:

$$y_i|f, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N\big(f(x_i), \sigma_\varepsilon^2\big),$$

```
Data Inputs: $\boldsymbol{y}, \boldsymbol{A}$.

Parameter Inputs: $\boldsymbol{\eta}_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\to\,\boldsymbol{\theta}}, \boldsymbol{\eta}_{\boldsymbol{\theta}\,\to\,\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})}$

Updates:
    $\boldsymbol{\omega}_{20} \longleftarrow -\tfrac{1}{2}\boldsymbol{A}\Big\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\big(\boldsymbol{\eta}_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_1\Big]$
    $\boldsymbol{\omega}_{21} \longleftarrow -\tfrac{1}{2}\mathrm{diagonal}\Big[\boldsymbol{A}\Big\{\mathrm{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\leftrightarrow\,\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\boldsymbol{A}^T\Big]$
    $\boldsymbol{\omega}_{22} \longleftarrow \big[\{(2\boldsymbol{y}-\boldsymbol{1}_n)\odot\boldsymbol{\omega}_{20}-\boldsymbol{1}_n\}^2+\boldsymbol{\omega}_{21}\big]^{-1/2}$
    $\boldsymbol{\eta}_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\to\,\boldsymbol{\theta}} \longleftarrow \begin{bmatrix} \boldsymbol{A}^T\{(\boldsymbol{1}_n+\boldsymbol{\omega}_{22})\odot(2\boldsymbol{y}-1)\} \\ -\tfrac{1}{2}\mathrm{vec}\big(\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{\omega}_{22})\boldsymbol{A}\big) \end{bmatrix}$

Parameter Outputs: $\boldsymbol{\eta}_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\to\,\boldsymbol{\theta}}$
```

Algorithm 6: *The inputs, updates and outputs of the Support Vector Machine pseudo-likelihood fragments.*

where the model for the mean function f takes the form

$$f(x) = \beta_0 + \beta_1\,x + \sum_{k=1}^{K} u_k\,z_k(x) \quad \text{with} \quad u_k|\sigma_u^2 \overset{\text{ind.}}{\sim} N(0,\sigma_u^2) \tag{4.1}$$

and $\{z_k : 1 \le k \le K\}$ is a suitable spline basis. The full model used in Wand (2017) is

$$\begin{gathered}\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2 \sim N(\boldsymbol{X\beta}+\boldsymbol{Zu},\sigma_\varepsilon^2\,\boldsymbol{I}), \quad \left[\begin{array}{c}\boldsymbol{\beta}\\ \boldsymbol{u}\end{array}\right]\Bigg|\,\sigma_u^2 \sim N\left(\left[\begin{array}{c}\boldsymbol{\mu_\beta}\\ \boldsymbol{0}\end{array}\right],\left[\begin{array}{cc}\boldsymbol{\Sigma_\beta} & \boldsymbol{0}\\ \boldsymbol{0} & \sigma_u^2\,\boldsymbol{I}\end{array}\right]\right),\\ \sigma_u^2|a_u \sim \text{Inverse-}\chi^2(1,1/a_u), \quad a_u \sim \text{Inverse-}\chi^2(1,1/A_u^2),\\ \sigma_\varepsilon^2|\,a_\varepsilon \sim \text{Inverse-}\chi^2(1,1/a_\varepsilon), \quad a_\varepsilon \sim \text{Inverse-}\chi^2(1,1/A_\varepsilon^2)\end{gathered} \tag{4.2}$$

where

$$\boldsymbol{X} \equiv \left[\begin{array}{cc} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{array}\right] \quad \text{and} \quad \boldsymbol{Z} \equiv \left[\begin{array}{ccc} z_1(x_1) & \cdots & z_K(x_1) \\ \vdots & \ddots & \vdots \\ z_1(x_n) & \cdots & z_K(x_n) \end{array}\right].$$

The 2×1 vector $\boldsymbol{\mu_\beta}$, 2×2 symmetric positive definite matrix $\boldsymbol{\Sigma_\beta}$ and the positive numbers A_u and A_ε are user-specified hyperparameters. Note that

$$\sigma_u^2|a_u \sim \text{Inverse-}\chi^2(1,1/a_u), \quad a_u \sim \text{Inverse-}\chi^2(1,1/A_u^2)$$

is equivalent to σ_u having a Half Cauchy prior with scale parameter A_u, but this auxiliary variable representation has advantages for VMP fitting. The final choice is the form of the z_k and the value of K. In the upcoming example we used canonical cubic O'Sullivan splines (Wand and Ormerod, 2008) with $K = 27$.

The joint posterior density function is approximated according to the following product restriction

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_u^2, a_u, \sigma_\varepsilon^2, a_\varepsilon | \boldsymbol{y}) \approx q(\boldsymbol{\beta}, \boldsymbol{u}) q(\sigma_u^2) q(a_u)\, q(\sigma_\varepsilon^2) q(a_\varepsilon). \tag{4.3}$$

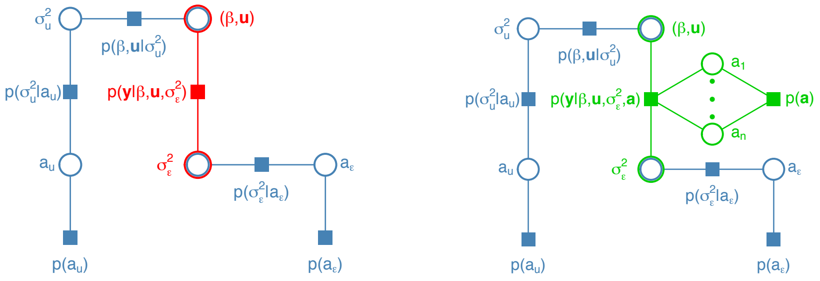

Figure 8: *Left panel: Factor graph for the ordinary nonparametric regression model. The Gaussian likelihood fragment is shown in red. Right panel: Factor graph for the quantile nonparametric regression model. The Asymmetric Laplace likelihood fragments are shown in green.*

VMP fitting of (4.2) can be accomplished by using the natural parameter updates for each of the fragments described in Section 4.1 of Wand (2017). The relevant factor graph is in the left panel of Figure 8 with the Gaussian likelihood fragment shown in red. We applied the VMP fitting procedure to data on 4,847 Zambian children from a 1992 demographic and health survey. These data are part of the data frame `Zambia` in the R package `INLA` (Rue et al., 2016). The predictor and response data are

$$\begin{aligned} x_i &= \text{age of the } i\text{th child in months} \\ \text{and } y_i &= \text{undernutrition score of the } i\text{th child}, \quad 1 \le i \le 4,847. \end{aligned} \tag{4.4}$$

All data were standardized and the hyperparameters we set at $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{0}$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = 10^{10}\boldsymbol{I}$ and $A_u = A_\varepsilon = 10^5$. The fits were back-transformed to the original units for plotting. The estimated nonparametric regression function and corresponding pointwise 95% credible set are shown in the left panel of Figure 9. The estimate shows mean undernutrition falling during the infancy period of the children before levelling off at about 2 years of age.

Now suppose that $100\tau\%$ quantile nonparametric regression for the same data is of interest. This involves replacement of

$$\boldsymbol{y} |\, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2 \sim N(\boldsymbol{X\beta} + \boldsymbol{Zu}, \sigma_\varepsilon^2\, \boldsymbol{I})$$

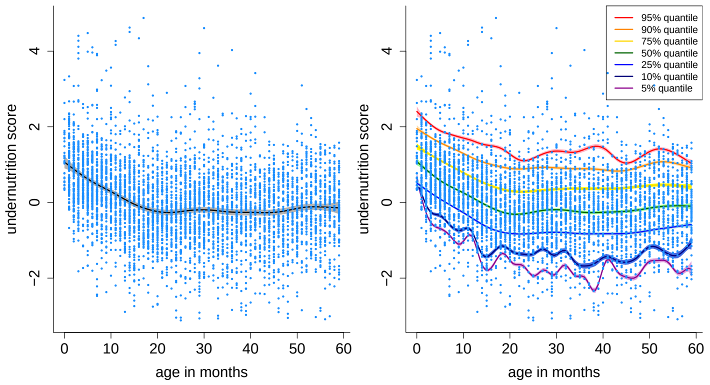

Figure 9: *Left panel: VMP nonparametric regression fit to the variables on Zambian children given by (4.4). The curve is the approximate posterior mean and the shaded region corresponds to pointwise approximate 95% credible sets. The estimates are based on VMP applied to model (4.2) according to product restriction (4.3). The relevant factor graph is shown in the left panel of Figure 8. Right panel: VMP quantile nonparametric regression fits to the same data. The curves and shaded regions have the same definitions as for the left panel.*

by

$$y_i|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\boldsymbol{a} \overset{\text{ind.}}{\sim} N\left((\boldsymbol{X\beta}+\boldsymbol{Zu})_i + \frac{(\frac{1}{2}-\tau)\sigma}{a_i\tau(1-\tau)}, \frac{\sigma^2}{a_i\tau(1-\tau)}\right), \quad a_i \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2(2,1)$$

in model (4.2). In terms of factor graphs it involves replacement of the Gaussian likelihood fragment by the Asymmetric Laplace likelihood fragments of Figure 4. The new fragments are shown in green in the right panel of Figure 8. The VMP updates corresponding to messages away from the likelihood are identical for both models. Algorithm 3 is used for the quantile nonparametric regression fitting and inference.

As a check, the same models were fit to the data using Markov chain Monte Carlo. The nonparametric regression and quantile regression curves are very close to their VMP counterparts. However, the 95% credible set bands are narrower for VMP. This is a consequence of the loss of inferential accuracy incurred by variational approximations involving auxiliary variables (see e.g. Wand et al., 2011).

4.2 Poisson to Negative Binomial Additive Model Analysis

Our second illustration involves additive model analysis when the response variable is a count. First we carried out a Poisson additive model analysis similar to those described in Section 12.3 of Ruppert et al. (2003). The data involve daily ragweed pollen counts in Kalamazoo, U.S.A., during the 1991–1994 ragweed seasons. The model is of the form

$$y_i \overset{\text{ind.}}{\sim} \text{Poisson}\big\{\exp\big(\beta_0 + \beta_1\, x_{1i} + \beta_2\, x_{2i} + \beta_3\, x_{3i} + f_{z_i}(x_{4i})\big)\big\}, \quad 1 \le i \le n \qquad (4.5)$$

where $n = 334$ is the total number of days when ragweed pollen was in season during 1991-1994. The variables appearing in (4.5) are ragweed pollen count on the ith day (y_i), temperature residual on the ith day (x_{1i}), indicator of significant rain on the ith day (x_{2i}), wind speed in knots on the ith day (x_{3i}), day number of ragweed pollen season for the current year on which y_i was recorded (x_{4i}) and a categorical variable for the year in which y_i was recorded (one of 1991, 1992, 1993 or 1994) (z_i). Here temperature residuals are the residuals from fitting penalized splines, each having 5 effective degrees of freedom, to temperature (in degrees Fahrenheit) versus day number for each annual ragweed pollen season. Note that (4.5) is not an additive model in the usual sense since $f_{z_i}(x_{4i})$ represents an interaction between year and day in ragweed pollen season. Mixed model-based penalized splines analogous to (4.1) are used for modelling the f_z, $z \in \{1992, 1992, 1993, 1994\}$. Let $\sigma^2_{u\ell}$, $1 \le \ell \le 4$, denote the variance parameters used to penalize each of the four penalized splines. The full model is

$$\begin{gathered} \boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u} \sim \text{Poisson}\big\{\exp(\boldsymbol{X\beta} + \boldsymbol{Zu})\big\}, \\ \left[\begin{array}{c}\boldsymbol{\beta}\\ \boldsymbol{u}\end{array}\right]\Bigg|\,\sigma^2_{u1},\sigma^2_{u2},\sigma^2_{u3},\sigma^2_{u4} \sim N\left(\left[\begin{array}{c}\boldsymbol{\mu_\beta}\\ \\ \boldsymbol{0}\end{array}\right], \left[\begin{array}{cc}\boldsymbol{\Sigma_\beta} & \boldsymbol{0}\\ \boldsymbol{0} & \underset{1\le\ell\le4}{\text{blockdiag}}(\sigma^2_{u\ell}\boldsymbol{I})\end{array}\right]\right), \qquad (4.6) \\ \sigma^2_{u\ell}|a_{u\ell} \sim \text{Inverse-}\chi^2(1, 1/a_{u\ell}), \quad a_{u\ell} \sim \text{Inverse-}\chi^2(1, 1/A^2_{u\ell}), \quad 1 \le \ell \le 4. \end{gathered}$$

Here

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} \cdots x_{41} & I(z_1{=}1992) & x_{41}I(z_1{=}1992) & \cdots & I(z_1{=}1994) & x_{41}I(z_1{=}1994) \\ \vdots & \vdots \quad \vdots \quad \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} \cdots x_{4n} & I(z_n{=}1992) & x_{4n}I(z_n{=}1992) & \cdots & I(z_n{=}1994) & x_{4n}I(z_n{=}1994) \end{bmatrix}$$

and $\boldsymbol{Z} = [\boldsymbol{Z}_{1991}\, \boldsymbol{Z}_{1992}\, \boldsymbol{Z}_{1993}\, \boldsymbol{Z}_{1994}]$ where $\boldsymbol{Z}_{1991}$ is an $n \times K$ matrix with (i, k) entry equal to $I(z_i = 1991)z_k(x_{4i})$ and $\boldsymbol{Z}_{1992}, \ldots, \boldsymbol{Z}_{1994}$ are defined analogously. The $\boldsymbol{\beta}$ and $\boldsymbol{u}$ vectors contain the coefficients to match the columns of $\boldsymbol{X}$ and $\boldsymbol{Z}$ respectively.

Despite the simplicity of Poisson response regression models, it is often the case that the Poisson likelihood is inadequate for modeling count response data that typically arises in practice. The crux of this inadequacy is the Poisson distribution restriction of the variance equalling the mean. It is common for the variability of count responses to be much higher than that imposed by the Poisson likelihood. If such overdispersion is ignored then standard errors are underestimated and valid statistical inference is compromised. The Negative Binomial family is an extension of the Poisson family that allows for the variance to exceed the mean. The move from this Poisson additive model to a Negative Binomial additive model involves replacement of

$$\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u} \sim \text{Poisson}\big\{\exp(\boldsymbol{X\beta} + \boldsymbol{Zu})\big\}$$

by

$$y_i|\,a_i \overset{\text{ind.}}{\sim} \text{Poisson}(a_i), \quad a_i|\,\boldsymbol{\beta},\boldsymbol{u},\kappa \overset{\text{ind.}}{\sim} \text{Gamma}[\kappa, \kappa\, \exp\{-(\boldsymbol{X\beta}+\boldsymbol{Zu})_i\}],$$

which corresponds to the likelihood specification

$$y_i|\,\boldsymbol{\beta},\boldsymbol{u},\kappa \overset{\text{ind.}}{\sim} \text{Negative-Binomial}[\exp\{-(\boldsymbol{X\beta}+\boldsymbol{Zu})_i\},\kappa].$$

Figure 10 shows the old and the new factor graphs according to this replacement. Almost all of the fragments in these factor graphs are covered by Wand (2017) and Algorithm 1. The exception is the fragment containing the factor $p(\kappa)$, which corresponds to placing a prior distribution on κ. In the ragweed data analysis we used the prior $p(\kappa) = 0.01\exp(-0.01\kappa),\ \kappa > 0$, which implies that the message sent from $p(\kappa)$ to κ is

$$m_{p(\kappa)\,\rightarrow\,\kappa}(\kappa) = \exp\left\{ \begin{bmatrix} \kappa\log(\kappa) - \log\{\Gamma(\kappa)\} \\ \kappa \end{bmatrix}^T \begin{bmatrix} 0 \\ -0.01 \end{bmatrix} \right\}.$$

This prior and message simply correspond to the Exponential distribution with rate parameter 0.01. We use the Moon Rock-type representation since it is conjugate with messages passed from $p(\boldsymbol{a}|\boldsymbol{\beta},\boldsymbol{u},\kappa)$ to κ.

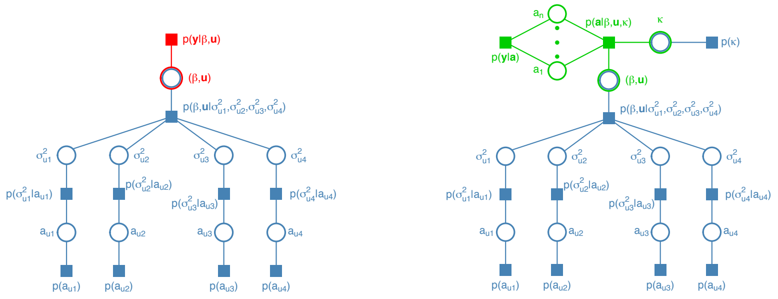

Figure 10: *Left panel: Factor graph for the Poisson additive regression model. The Poisson likelihood fragment is shown in red. Right panel: Factor graph for the Negative Binomial additive model. The Negative Binomial likelihood fragments are shown in green.*

Figure 11 provides some visual summaries of the model fits. The first row shows posterior density functions for the coefficients of the predictors that enter the models linearly, and the Negative Binomial shape parameter. The posterior density functions for the Poisson model are considerably narrower than those for the Negative Binomial model, which is indicative of overdispersion being ignored in the former model. In the same vein, the posterior density function of κ has most of its support between 1.4 and 2.2. Such low κ values indicate superiority of the Negative Binomial model since the Poisson model corresponds to the $\kappa \rightarrow \infty$ limiting case.

The lower four panels of Figure 11 show the estimates of $f_{1991}, \ldots, f_{1994}$ for the Poisson and Negative Binomial models. The solid curves correspond to the posterior mean for each day in season value, while the dashed curves are pointwise 95% credible

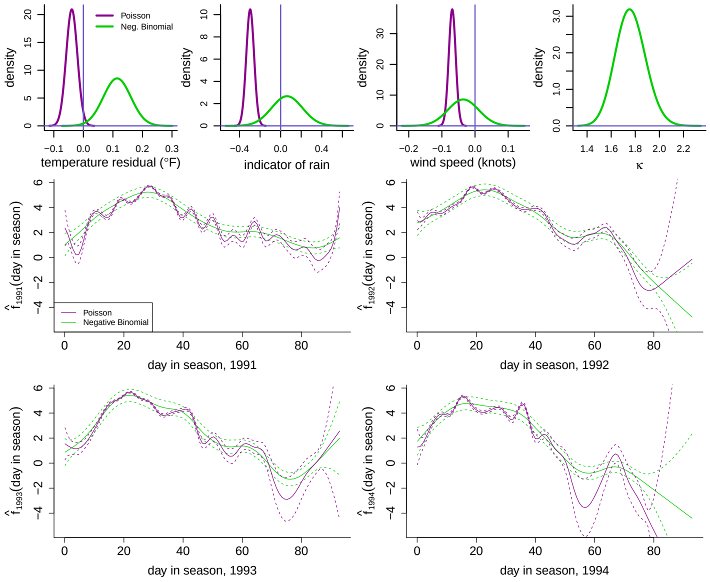

Figure 11: *First three panels: VMP-approximate posterior density functions of the coefficients of temperature residual, indicator of rain and wind speed for both the Poisson additive model (4.6) and the Negative Binomial additive model for the ragweed data example. Fourth panel: VMP-approximate posterior density function of the κ parameter for the Negative Binomial additive model. Lower four panels: VMP-based estimates of $f_{1992}, \ldots, f_{1994}$ according to each model. The solid curves are posterior means and the dashed curves are pointwise 95% credible intervals based on VMP approximate inference.*

sets according to the VMP approximation. The estimates are similar for each model, but the credible set bands are narrower for the Poisson model, in keeping with their ignorance of overdispersion.

Computing times for the Poisson and Negative Binomial additive models were also compared. All computing was performed using version 3.4.1 of the R language (R Core Team, 2017) on a desktop personal computer with 8 gigabytes of random access memory and a 3.2 gigahertz processor. Firstly, we determined that 250 iterations were sufficient for convergence of VMP for each model. The elapsed times were 5.5 seconds for the Poisson model and 6.9 seconds for the Negative Binomial model.

We also compared the VMP-approximate posterior density functions and additive model components with those obtained using Markov chain Monte Carlo. Excellent agreement was observed in almost all cases. An exception concerned the posterior density function for κ, with VMP under-approximating the posterior standard deviation. This phenomenon was also observed in Luts and Wand (2015).

5 Closing Remarks

As exemplified in Section 4, the algorithms presented in Section 3 concerning fragments updates for elaborate likelihoods greatly enhances the utility of VMP for semiparametric regression analyses. In addition to the primitives for VMP-based semiparametric regression laid down in Wand (2017) we have identified a small set of new primitives, corresponding to sufficient statistic expectations of the Inverse Square Root Nadarajah, Moon Rock and Sea Sponge distributions. Once their computation is established in a suite of computer programmes, a much richer class of models can be handled via the VMP paradigm.

References

Azzalini, A. (2017). *The R package sn: The skew-normal and related distributions, such as the skew-t (version 1.5).*
URL http://azzalini.stat.unipd.it/SN 15

Azzalini, A. and Dalla Valle, A. (1996). "The multivariate skew-normal distribution." *Biometrika*, 83: 715–726. 14

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., and Rue, H. (2009). "Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data." *Statistics and Computing*, 19: 479–492. 15

Frühwirth-Schnatter, S. and Pyne, S. (2010). "Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions." *Biostatistics*, 11: 317–336. 13

Frühwirth-Schnatter, S. and Wagner, H. (2006). "Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling." *Biometrika*, 93: 827–841. 15

Gradshteyn, I. S. and Ryzhik, I. M. (1994). *Tables of Integrals, Series, and Products, Fifth Edition*. San Diego, California: Academic Press. 1, 3

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. W. (2013). "Stochastic variational inference." *Journal of Machine Learning Research*, 14: 1303–1347. 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. 7, 9, 11

Kotz, S., Kozubowski, T. J., and Podgórski, K. (2001). *The Laplace Distribution and Generalizations*. Boston: Birkhäuser. 10

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). "Automatic differentiation variational inference." *Journal of Machine Learning Research*, 18: 1–45. 2

Lachos, V. H., Ghosh, P., and Arellano-Valle, R. B. (2010). "Likelihood based inference for skew-normal independent linear mixed models." *Statistica Sinica*, 303–322. 13

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). “Robust statistical modeling using the t distribution.” *Journal of the American Statistical Association*, 84: 881–896. 8

Luts, J. and Ormerod, J. T. (2014). “Mean field variational Bayesian inference for support vector machine classification.” *Computational Statistics & Data Analysis*, 73: 163–176. 17

Luts, J. and Wand, M. P. (2015). “Variational inference for count response semiparametric regression.” *Bayesian Analysis*, 10: 991–1023. 7, 25

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

Minka, T. and Winn, J. (2008). “Gates: A graphical notation for mixture models.” *Microsoft Research Technical Report Series*, MSR-TR-2008-185: 1–16. 1

Nadarajah, S. (2008). “A new model for symmetric and skewed data.” *Probability in the Engineering and Informational Sciences*, 22: 261–271. 12, 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. 3

Ormerod, J. T. and Wand, M. P. (2010). “Explaining variational approximations.” *The American Statistician*, 64: 140–153. 3

Polson, N. G. and Scott, S. L. (2011). “Data augmentation for support vector machines.” *Bayesian Analysis*, 6: 1–23. 17

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. 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/ 15, 25, 1

Rue, H., Martino, S., Lindgren, F., Simpson, D., Riebler, A., and Krainski, E. (2016). *The R package ’INLA’: Functions which allow to perform full Bayesian analysis of latent Gaussian models using integrated nested Laplace approximation (version 0.0)*.
URL http://www.r-inla.org 21

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press. 2, 22

Tipping, M. E. and Lawrence, N. D. (2003). “A variational approach to robust Bayesian interpolation.” In *Institute of Electrical and Electronics Engineers Workshop of Neural Networds for Signal Processing*, 229–238. 8

Titsias, M. K. and Lázaro-Gredilla, M. (2014). “Doubly stochastic variational Bayes for non-conjugate inference.” *Proceedings of Machine Learning Research*, 32: 1971–1979. 2

Verdinelli, I. and Wasserman, L. (1991). "Bayesian analysis of outlier problems using the Gibbs sampler." *Statistics and Computing*, 1: 105–117. 8

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. 1, 2, 3, 4, 5, 19, 20, 21, 24, 26, 11, 13, 15

Wand, M. P. and Ormerod, J. T. (2008). "On semiparametric regression with O'Sullivan penalized splines." *Australian & New Zealand Journal of Statistics*, 50: 179–198. 21

— (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. 15, 22, 3, 4

Winn, J. and Bishop, C. M. (2005). "Variational message passing." *Journal of Machine Learning Research*, 6: 661–694. 1

Yang, Y., Wang, H. J., and He, X. (2016). "Posterior inference in Bayesian quantile regression with Asymmetric Laplace likelihood." *International Statistical Review*, 84: 327–344. 10

Yu, K. and Moyeed, R. A. (2001). "Bayesian quantile regression." *Statistics and Probability Letters*, 54: 437–447. 10

Acknowledgments

This research was partially supported by the Australian Research Council Discovery Project DP140100441. We are grateful to Peter Alspach, Chris Jones, Luca Maestrini, John Maindonald, Saralees Nadarajah and Thomas Yee for their assistance with this research.

Supplement for:

Variational Message Passing for Elaborate Response Regression Models

M. W. McLean and M. P. Wand

School of Mathematical and Physical Sciences,
University of Technology Sydney,
P.O. Box 123, Broadway 2007, Australia

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 $p, x \in \mathbb{R}$:

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

(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

$$\frac{K_{p+1}(x)}{K_p(x)}, \quad x > 0. \tag{S.2}$$

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.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).

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) 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

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

S.1.2 Parabolic Cylinder Functions

The *parabolic cylinder function* of order $\nu \in \mathbb{R}$, is denoted by D_ν. 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

$$D_{\nu+1}(x) = x\,D_\nu(x) - \nu\,D_{\nu-1}(x). \tag{S.4}$$

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

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

$$\mathcal{R}_\nu(x) \equiv \frac{D_{-\nu-2}(x)}{D_{-\nu-1}(x)}, \quad \nu > -1, x \in \mathbb{R}, \tag{S.5}$$

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

$$\mathcal{R}_\nu(x) = \cfrac{1}{x + \cfrac{\nu+1}{x + \cfrac{\nu+2}{x + \cfrac{\nu+3}{x+\cdots}}}}, \quad x > 0. \tag{S.6}$$

As explained in Neville et al. (2014) and encapsulated in their Algorithm 4 (S.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 $x \le 0$ 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.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 $p, q, r \ge 0$ 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.3) that for the special case of $p = \frac{1}{2}$ we have

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

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

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

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

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

The expectation of the sufficient statistic is

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

Convenient notation based on (S.2), which we use in Algorithms 3–5 , is

$$\begin{aligned} (E\boldsymbol{T})_2^{\text{ISRN}}(\boldsymbol{\eta}) &\equiv \frac{-2(\eta_1+1)D_{2\eta_1+1}(-\eta_2/\sqrt{-2\eta_3})}{\sqrt{-2\eta_3}\, D_{2\eta_1+2}(-\eta_2/\sqrt{-2\eta_3})} \\ \text{and} \quad (E\boldsymbol{T})_3^{\text{ISRN}}(\boldsymbol{\eta}) &\equiv \frac{-(\eta_1+1)(2\eta_1+1)D_{2\eta_1}(-\eta_2/\sqrt{-2\eta_3})}{\eta_3\, D_{2\eta_1+2}(-\eta_2/\sqrt{-2\eta_3})}. \end{aligned} \tag{S.3}$$

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

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

The natural parameter vector and its inverse mapping are

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

and the log-partition function is

$$A(\boldsymbol{\eta}) = \log\left\{\int_0^\infty \{t^t/\Gamma(t)\}^{\eta_1} \exp(\eta_2\, t)\, dt\right\}. \tag{S.4}$$

The expectation of the sufficient statistic is

$$E\{\boldsymbol{T}(x)\} = \exp\{-A(\boldsymbol{\eta})\} \begin{bmatrix} \displaystyle\int_0^\infty [x\log(x) - \log\{\Gamma(x)\}]\, \{x^x/\Gamma(x)\}^{\eta_1} e^{\eta_2\, x}\, dx \\ \displaystyle\int_0^\infty x\, \{x^x/\Gamma(x)\}^{\eta_1} \exp(\eta_2\, x)\, dx \end{bmatrix}. \tag{S.5}$$

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.4) and (S.5) can be re-written:

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

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

A convenient notation based on (S.5), which we use in Algorithms 1 and 2, is

$$\begin{aligned} (E\boldsymbol{T})_2^{\text{MR}}(\boldsymbol{\eta}) &\equiv \mathcal{D}(0,1,\eta_1,\eta_2)/\mathcal{D}(0,0,\eta_1,\eta_2) \\ &= \exp\big[\log\{\mathcal{D}(0,1,\eta_1,\eta_2)\} - \log\{\mathcal{D}(0,0,\eta_1,\eta_2)\}\big]. \end{aligned} \tag{S.6}$$

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

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

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

$$E\{\boldsymbol{T}(x)\} = e^{-A(\boldsymbol{\eta})} \begin{bmatrix} \displaystyle\int_{-\infty}^{\infty} \log(1+x^2)\,(1+x^2)^{\eta_1} \exp\left(\eta_2\, x^2 + \eta_3\, x\sqrt{1+x^2}\right) dx \\ \displaystyle\int_{-\infty}^{\infty} x^2\,(1+x^2)^{\eta_1} \exp\left(\eta_2\, x^2 + \eta_3\, x\sqrt{1+x^2}\right) dx \\ \displaystyle\int_{-\infty}^{\infty} x\sqrt{1+x^2}\,(1+x^2)^{\eta_1} \exp\left(\eta_2\, x^2 + \eta_3\, x\sqrt{1+x^2}\right) dx \end{bmatrix}. \tag{S.7}$$

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.7), is:

$$\begin{aligned}(E\boldsymbol{T})^{\text{ss}}_2(\boldsymbol{\eta}) &\equiv \mathcal{E}(2, \eta_1, \eta_2, \eta_3)/\mathcal{E}(0, \eta_1, \eta_2, \eta_3)\\ &= \exp\left[\log\{\mathcal{E}(2, \eta_1, \eta_2, \eta_3)\} - \log\{\mathcal{E}(0, \eta_1, \eta_2, \eta_3)\}\right]\end{aligned} \tag{S.8}$$

and

$$\begin{aligned}(E\boldsymbol{T})^{\text{ss}}_3(\boldsymbol{\eta}) &\equiv \mathcal{E}(1, \eta_1 + \tfrac{1}{2}, \eta_2, \eta_3)/\mathcal{E}(0, \eta_1, \eta_2, \eta_3)\\ &= \exp\left[\log\{\mathcal{E}(1, \eta_1 + \tfrac{1}{2}, \eta_2, \eta_3)\} - \log\{\mathcal{E}(0, \eta_1, \eta_2, \eta_3)\}\right]\end{aligned} \tag{S.9}$$

and appears in Algorithm 4.

S.3 Derivations

Each of the fragment updates in Algorithms 1–6 involve repeated application of the VMP equations (2.4)–(2.6) 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 (2.5) and (2.6), the messages from $p(\boldsymbol{y}|\boldsymbol{a})$ to each of the a_i are

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

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

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

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

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

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

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

Since, from (2.4), $m_{a_i\,\rightarrow\,p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)}(a_i) \longleftarrow m_{p(\boldsymbol{y}|\,\boldsymbol{a})\,\rightarrow\,a_i}(a_i)$ we then have

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

This is proportional to a Gamma density function with mean

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

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

$$E_{q(a_i)}(\log a_i) \equiv \text{digamma}(y_i+\mu_{q(\kappa)}) - \log\Big(1+\mu_{q(\kappa)}\,E_{q(\boldsymbol{\theta})}[\exp\{-(\boldsymbol{A\theta})_i\}]\Big). \tag{S.2}$$

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

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

Hence

$$m_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\rightarrow\,\kappa}(\kappa) = \exp\left\{\left[\begin{array}{c}\kappa\log(\kappa)-\log\{\Gamma(\kappa)\}\\ \kappa\end{array}\right]^T\boldsymbol{\eta}_{p(\boldsymbol{a}|\,\boldsymbol{\theta},\,\kappa)\,\rightarrow\,\kappa}\right\} \tag{S.3}$$

where

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

Note that (S.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,

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

is also proportional to a Moon Rock density function and

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

Hence

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

where $(\boldsymbol{ET})_2^{\mathrm{MR}}$ is given by (S.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

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

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

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

and

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

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 (3.5) is

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

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

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

and

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

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

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

and $\mu_{q(\nu)} \equiv \int_0^\infty \nu\, q^*(\nu)\, d\nu$. Since $q^*(a_i)$ is an Inverse-χ^2 density function the ith 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

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

where $\boldsymbol{\omega}_4$ has ith 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,

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

As a function of ν we have

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

Hence

$$m_{p(\boldsymbol{a}|\nu)\,\rightarrow\,\nu}(\nu) = \exp\left\{\begin{bmatrix}(\nu/2)\log(\nu/2) - \log\{\Gamma(\nu/2)\} \\ \nu/2\end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{a}|\,\nu)\,\rightarrow\,\nu}\right\} \qquad \text{(S.4)}$$

where this message's natural parameter is

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

which involves $\boldsymbol{\omega}_4$ and $\boldsymbol{\omega}_5$ given above. Note that (S.4) is proportional to a factor of 2 rescaling of a Moon Rock density function. Under conjugacy the message $m_{\nu\,\rightarrow\,p(\boldsymbol{a}|\nu)}(\nu)$ 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 (3.7), 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

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

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 σ^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 σ^2 is

$$
\begin{array}{l}
m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\sigma^2}(\sigma^2) = \\
\qquad \exp\left\{\left[\begin{array}{c}\log(\sigma^2)\\ 1/\sigma \\ 1/\sigma^2\end{array}\right]^T\left[\begin{array}{c}-n/2\\ (\frac{1}{2}-\tau)\big\{\boldsymbol{y}-\boldsymbol{A}E_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})\big\}^T\boldsymbol{1}_n\\ -\frac{1}{2}\tau(1-\tau)\mathrm{tr}\big[E_{q(\boldsymbol{\theta})}\{(\boldsymbol{y}-\boldsymbol{A\theta})(\boldsymbol{y}-\boldsymbol{A\theta})^T\}\\ \times\,\mathrm{diag}\{E_{q(\boldsymbol{a})}(\boldsymbol{a})\}\big]\end{array}\right]\right\}
\end{array}
\tag{S.5}
$$

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

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

and $E_{q(\boldsymbol{a})}$ is defined similarly for $\boldsymbol{a}$. Conjugacy with (S.5) requires that the message from σ^2 to $p(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2,\boldsymbol{a})$ is also of the form

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

and this is the case provided that messages passed to σ^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 σ^2 is such that

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

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

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

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

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

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

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

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

The updates in Algorithm 3 quickly follow.

S.3.4 Skew Normal Fragment Updates

It follows from (3.9) that the logarithm of the likelihood factor is

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

where, here and elsewhere, $\Vert\boldsymbol{v}\Vert\equiv\sqrt{\boldsymbol{v}^T\boldsymbol{v}}$ 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

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

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 σ^2 is

$$m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\,\to\,\sigma^2}(\sigma^2)=\exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma\\ 1/\sigma^2\end{bmatrix}^T\begin{bmatrix}-n/2\\ \mu_{q(\lambda\sqrt{1+\lambda^2})}\big\{\boldsymbol{y}-\boldsymbol{A}\,E_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})\big\}^TE_{q(\boldsymbol{a})}|\boldsymbol{a}|\\ -\tfrac{1}{2}\big(1+\mu_{q(\lambda^2)}\big)E_{q(\boldsymbol{\theta})}\{\Vert\boldsymbol{y}-\boldsymbol{A\theta}\Vert^2\}\end{bmatrix}\right\} \tag{S.6}$$

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 σ^2 for the Asymmetric Laplace fragments.

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

$$m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a})\,\to\,\lambda}(\lambda)=\exp\left\{\begin{bmatrix}\log(1+\lambda^2)\\ \lambda^2\\ \lambda\sqrt{1+\lambda^2}\end{bmatrix}^T\begin{bmatrix}n/2\\ -\tfrac{1}{2}[\mu_{q(1/\sigma^2)}\,E_{q(\boldsymbol{\theta})}\{\Vert\boldsymbol{y}-\boldsymbol{A\theta}\Vert^2\}+E_{q(\boldsymbol{a})}\Vert\boldsymbol{a}\Vert^2]\\ \mu_{q(1/\sigma)}\big\{\boldsymbol{y}-\boldsymbol{A}\,E_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})\big\}^TE_{q(\boldsymbol{a})}|\boldsymbol{a}|\end{bmatrix}\right\} \tag{S.7}$$

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.8) and (S.9) we then get

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

and

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

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

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

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

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

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.6) and (S.7). The natural parameter update for the first of these messages is

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

where

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

That for the second is

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

S.3.5 Finite Normal Mixture Fragment Updates

From (3.12), 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

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\,\sigma^2,\,\boldsymbol{a})\,\rightarrow\,\boldsymbol{\theta}} \longleftarrow \mu_{q(1/\sigma^2)}\begin{bmatrix} \boldsymbol{A}^T\text{diag}\{E_{q(\boldsymbol{a})}(\boldsymbol{\mathscr{A}})(\boldsymbol{1}_K/\boldsymbol{s}^2)\}\boldsymbol{y} \\ -\tfrac{1}{2}\text{vec}\big(\boldsymbol{A}^T\text{diag}\{E_{q(\boldsymbol{a})}(\boldsymbol{\mathscr{A}})(\boldsymbol{1}_K/\boldsymbol{s}^2)\}\boldsymbol{A}\big)\end{bmatrix}$$
$$-\mu_{q(1/\sigma)}\begin{bmatrix}\boldsymbol{A}^T\,E_{q(\boldsymbol{a})}(\boldsymbol{\mathscr{A}})(\boldsymbol{m}/\boldsymbol{s}^2) \\ \boldsymbol{0}\end{bmatrix}$$

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

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

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

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

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 σ^2 is

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

This means that the message passed from $m_{p(\boldsymbol{y}|\,\boldsymbol{\theta},\sigma^2,\boldsymbol{a})\,\rightarrow\,\sigma^2}(\sigma^2)$ to σ^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 (2.5) and (2.6), the messages from $\check{p}(\boldsymbol{a})$ to each of the a_i are

$$m_{\check{p}(\boldsymbol{a})\,\rightarrow\,a_i}(a_i)=\check{p}(a_i)=I(a_i>0),\quad 1\le i\le n. \tag{S.8}$$

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

$$\begin{aligned}
&m_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\rightarrow\,a_i}(a_i)\\
&\quad=a_i^{-1/2}\exp\left\{\left[\begin{array}{c}a_i\\1/a_i\end{array}\right]^T\left[\begin{array}{c}-\frac{1}{2}\\-\frac{1}{2}\,E_{q(\boldsymbol{\theta})}\left[\{(2y_i-1)(\boldsymbol{A\theta})_i-1\}^2\right]\end{array}\right]\right\}
\end{aligned}\tag{S.9}$$

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

$$m_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\rightarrow\,\boldsymbol{\theta}}(\boldsymbol{\theta})\,m_{\boldsymbol{\theta}\,\rightarrow\,\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})}(\boldsymbol{\theta}). \tag{S.10}$$

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

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

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

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

where

$$\boldsymbol{\eta}_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\rightarrow\,\boldsymbol{\theta}} \longleftarrow \begin{bmatrix}\boldsymbol{A}^T[\{\boldsymbol{1}_n + E_{q(\boldsymbol{a})}(\boldsymbol{1}_n/\boldsymbol{a})\}\odot(2\boldsymbol{y}-\boldsymbol{1}_n)]\\ -\tfrac{1}{2}\mathrm{vec}\big[\boldsymbol{A}^T\mathrm{diag}\{E_{q(\boldsymbol{a})}(\boldsymbol{1}_n/\boldsymbol{a})\}\boldsymbol{A}\big]\end{bmatrix} \tag{S.11}$$

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

$$q^*(a_i) \propto m_{\check{p}(\boldsymbol{y}|\,\boldsymbol{\theta},\boldsymbol{a})\,\rightarrow\,a_i}(a_i)\, m_{\check{p}(\boldsymbol{a})\,\rightarrow\,a_i}(a_i), \quad 1 \le i \le n. \tag{S.12}$$

On combining (S.8), (S.9) and (S.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

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

Then, from (S.1) in Section S.2.1

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

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.10). Now set the updates

$$\begin{aligned}&\boldsymbol{\omega}_{20} \longleftarrow \boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})}, \quad \boldsymbol{\omega}_{21} \longleftarrow \mathrm{diagonal}(\boldsymbol{A}\,\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T),\\ &\text{and} \quad \boldsymbol{\omega}_{22} \longleftarrow [\{(2\boldsymbol{y}-\boldsymbol{1}_n)\odot\boldsymbol{\omega}_{20} - \boldsymbol{1}_n\}^2 + \boldsymbol{\omega}_{21}]^{-1/2}.\end{aligned} \tag{S.13}$$

Then the updates in Algorithm 6 follow from updates (S.11) and (S.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).