Web-Supplement for:

Accurate Logistic Variational Message Passing: Algebraic and Numerical Details

BY TUI H. NOLAN AND MATT P. WAND

School of Mathematical and Physical Sciences, University of Technology Sydney, Broadway 2007, Australia

S.1 Proof of Theorem 1

First note that for any $a, b \in \mathbb{R}$,

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

From the first result in (S.1), it follows immediately that

$$\int_{-\infty}^{\infty} \text{expit}_k(\mu+\sigma\,x)\phi(x)\,dx = \sum_{i=1}^{k} p_{k,i}\Phi\left(\frac{\mu\,s_{k,i}}{\sqrt{1+\sigma^2\,s_{k,i}^2}}\right).$$

Hence, for all $\mu \in \mathbb{R}$ and $\sigma > 0$,

$$\begin{aligned} &\left|\mathcal{B}_0(\mu,\sigma^2) - \sum_{i=1}^{k} p_{k,i}\Phi\left(\frac{\mu\,s_{k,i}}{\sqrt{1+\sigma^2\,s_{k,i}^2}}\right)\right| \\ &\quad= \left|\int_{-\infty}^{\infty} \text{expit}(\mu+\sigma\,x)\phi(x)\,dx - \int_{-\infty}^{\infty} \text{expit}_k(\mu+\sigma\,x)\phi(x)\,dx\right| \\ &\quad\le \int_{-\infty}^{\infty} \left|\text{expit}(\mu+\sigma\,x) - \text{expit}_k(\mu+\sigma\,x)\right|\phi(x)\,dx \\ &\quad\le \sup_{u\in\mathbb{R}} \left|\text{expit}(u) - \text{expit}_k(u)\right| \int_{-\infty}^{\infty} \phi(x)\,dx = \Delta_k \end{aligned}$$

where the last equality follows from (7). Part (a) of Theorem 1 follows immediately.

The second result in (S.1) implies that

$$\int_{-\infty}^{\infty} x\,\text{expit}_k(\mu+\sigma\,x)\phi(x)\,dx = \sigma\sum_{i=1}^{k} \frac{p_{k,i}\,s_{k,i}}{\sqrt{1+\sigma^2\,s_{k,i}^2}}\,\phi\left(\frac{\mu\,s_{k,i}}{\sqrt{1+\sigma^2\,s_{k,i}^2}}\right)$$

For all $\mu \in \mathbb{R}$ and $\sigma > 0$ we then have

$$\begin{aligned}
&\left| \mathcal{B}_1(\mu, \sigma^2) - \sigma \sum_{i=1}^{k} \frac{p_{k,i}\, s_{k,i}}{\sqrt{1+\sigma^2\, s_{k,i}^2}} \phi\left(\frac{\mu\, s_{k,i}}{\sqrt{1+\sigma^2\, s_{k,i}^2}} \right) \right| \\
&\qquad = \left| \int_{-\infty}^{\infty} \text{expit}(\mu + \sigma\, x)\, x\phi(x)\, dx - \int_{-\infty}^{\infty} \text{expit}_k(\mu + \sigma\, x)\, x\phi(x)\, dx \right| \\
&\qquad \leq \int_{-\infty}^{\infty} \left| \text{expit}(\mu + \sigma\, x) - \text{expit}_k(\mu + \sigma\, x) \right| |x| \phi(x)\, dx \\
&\qquad \leq \sup_{u \in \mathbb{R}} \left| \text{expit}(u) - \text{expit}_k(u) \right| \int_{-\infty}^{\infty} |x| \phi(x)\, dx \\
&\qquad = 2\Delta_k \int_0^{\infty} x\, \phi(x)\, dx = \sqrt{2/\pi}\, \Delta_k.
\end{aligned}$$

S.2 Derivation of Algorithm 1

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

$$m_{p(\boldsymbol{y}|\boldsymbol{\theta}) \to \boldsymbol{\theta}}(\boldsymbol{\theta}) = \exp[\boldsymbol{y}^T \boldsymbol{A}\boldsymbol{\theta} - \mathbf{1}^T \log\{\mathbf{1} + \exp(\boldsymbol{A}\boldsymbol{\theta})\}]$$

is not conjugate with Multivariate Normal messages passed to $\boldsymbol{\theta}$ from other factors. A non-conjugate VMP remedy (Knowles & Minka, 2011) involves replacement of $m_{p(\boldsymbol{y}|\boldsymbol{\theta}) \to \boldsymbol{\theta}}(\boldsymbol{\theta})$ by

$$\widetilde{m}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \to \boldsymbol{\theta}}(\boldsymbol{\theta}) \equiv \exp\left\{ \begin{bmatrix} \boldsymbol{\theta} \\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T) \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \to \boldsymbol{\theta}} \right\} \tag{S.2}$$

to enforce conjugacy with Multivariate Normal messages. Under pre-specification (S.2) the current $q(\boldsymbol{\theta})$ density function satisfies

$$q(\boldsymbol{\theta}) \propto \widetilde{m}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \to \boldsymbol{\theta}}(\boldsymbol{\theta}) \times (\text{product of messages passed to } \boldsymbol{\theta} \text{ from its other neighbours})$$

From (10) of Wand (2017), we then get

$$q(\boldsymbol{\theta}) \propto \widetilde{m}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \to \boldsymbol{\theta}}(\boldsymbol{\theta}) m_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\boldsymbol{\theta})}(\boldsymbol{\theta}) = \exp\left\{ \begin{bmatrix} \boldsymbol{\theta} \\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T) \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} \right\}$$

where

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} \equiv \boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \to \boldsymbol{\theta}} + \boldsymbol{\eta}_{\boldsymbol{\theta} \to p(\boldsymbol{y}|\,\boldsymbol{\theta})}.$$

Let $\boldsymbol{\mu}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}}$ be the corresponding common parameters. The natural parameters and common parameters are the following functions of each other:

$$\begin{aligned}
&\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} = \begin{bmatrix} (\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}})_1 \\ (\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}})_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} \boldsymbol{\mu}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}}) \end{bmatrix} \\
&\text{and} \quad \begin{cases} \boldsymbol{\mu}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} &= -\frac{1}{2}\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}})_2\big)\}^{-1} (\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}})_1 \\ \boldsymbol{\Sigma}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} &= -\frac{1}{2}\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}})_2\big)\}^{-1}. \end{cases}
\end{aligned} \tag{S.3}$$

In the upcoming arguments we will use the shorthand:

$$\boldsymbol{\eta}_{q(\boldsymbol{\theta})} \equiv \boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}}, \quad \boldsymbol{\mu}_{q(\boldsymbol{\theta})} \equiv \boldsymbol{\mu}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}} \quad \text{and} \quad \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \equiv \boldsymbol{\Sigma}_{p(\boldsymbol{y}|\,\boldsymbol{\theta}) \leftrightarrow \boldsymbol{\theta}}.$$

Using the set-up of Section 2.2 of Rohde & Wand (2016), the $\boldsymbol{\theta}$-localized approximate marginal log-likelihood is

$$\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\eta}_{q(\boldsymbol{\theta})})^{[\boldsymbol{\theta}]} = \text{Entropy}\{q(\boldsymbol{\theta}; \boldsymbol{\eta}_{q(\boldsymbol{\theta})})\} + \text{NonEntropy}\{q(\boldsymbol{\theta}; \boldsymbol{\eta}_{q(\boldsymbol{\theta})})\}$$

where

$$\text{Entropy}\{q(\boldsymbol{\theta}); \boldsymbol{\eta}_{q(\boldsymbol{\theta})})\} = \tfrac{1}{2}\log\Big| - \tfrac{1}{2}\{\text{vec}^{-1}(\boldsymbol{\eta}_{q(\boldsymbol{\theta})})_2\}^{-1}\Big| + d\{1 + \log(2\pi)\}/2,$$

with d denoting the dimension of $\boldsymbol{\theta}$. Also,

$$\begin{aligned}\text{NonEntropy}\{q(\boldsymbol{\theta}; \boldsymbol{\eta}_{q(\boldsymbol{\theta})})\} \equiv\, & E_{q(\boldsymbol{\theta};\boldsymbol{\eta}_{q(\boldsymbol{\theta})})}\{\log\, p(\boldsymbol{y}|\boldsymbol{\theta})\} \\ & + E_{q(\boldsymbol{\theta};\boldsymbol{\eta}_{q(\boldsymbol{\theta})})}(\text{sum of other log-factors neighboring } \boldsymbol{\theta}) \\ = \,& \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} - E_{q(\boldsymbol{\theta};\boldsymbol{\eta}_{q(\boldsymbol{\theta})})}\left[\mathbf{1}^T\log\{\mathbf{1} + \exp(\boldsymbol{A}\boldsymbol{\theta})\} + \begin{bmatrix}\boldsymbol{\theta}\\ \text{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^T)\end{bmatrix}^T\boldsymbol{\eta}^\dagger\right]\end{aligned}$$

where $\boldsymbol{\eta}^\dagger$ is the sum of the natural parameters of the messages passed to $\boldsymbol{\theta}$ other than the message from $p(\boldsymbol{y}|\boldsymbol{\theta})$. Ideally we would maximize $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\eta}_{q(\boldsymbol{\theta})})^{[\boldsymbol{\theta}]}$ over $\boldsymbol{\eta}_{q(\boldsymbol{\theta})}$ but are thwarted by the intractability of the q-density expectation of $\log\{\mathbf{1} + \exp(\boldsymbol{A}\boldsymbol{\theta})\}$. To get around this we apply (1) to obtain

$$\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\eta}_{q(\boldsymbol{\theta})})^{[\boldsymbol{\theta}]} \geq \text{Entropy}\{q(\boldsymbol{\theta}; \boldsymbol{\eta}_{q(\boldsymbol{\theta})})\} + \underline{\text{NonEntropy}}\{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}, \boldsymbol{\omega}_1\}$$

where

$$\begin{aligned}\underline{\text{NonEntropy}}\{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}, \boldsymbol{\omega}_1\} \equiv\, & \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} - \tfrac{1}{2}(\boldsymbol{\omega}_1^2)^T\text{diagonal}(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T) \\ & -\mathbf{1}^T\log[\mathbf{1} + \exp\{\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \tfrac{1}{2}(\mathbf{1} - 2\boldsymbol{\omega}_1)\odot\text{diagonal}(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T)\}] \\ & + \left\{\begin{bmatrix}\boldsymbol{\mu}_{q(\boldsymbol{\theta})}\\ \text{vec}\Big(\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} + \boldsymbol{\mu}_{q(\boldsymbol{\theta})}\boldsymbol{\mu}_{q(\boldsymbol{\theta})}^T\Big)\end{bmatrix}\right\}^T\boldsymbol{\eta}^\dagger\end{aligned}$$

and $\boldsymbol{\omega}_1$ is an $n \times 1$ vector of variational parameters. We now seek to maximize

$$\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\eta}_{q(\boldsymbol{\theta})}, \boldsymbol{\omega}_1)^{[\boldsymbol{\theta}]} \equiv \text{Entropy}\{q(\boldsymbol{\theta}; \boldsymbol{\eta}_{q(\boldsymbol{\theta})})\} + \underline{\text{NonEntropy}}\{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}, \boldsymbol{\omega}_1\}.$$

Using arguments analogous to those given in Section 4 of Rohde & Wand (2016), the function $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\eta}_{q(\boldsymbol{\theta})}, \boldsymbol{\omega}_1)^{[\boldsymbol{\theta}]}$ has a stationary point in the $[\boldsymbol{\eta}_{q(\boldsymbol{\theta})}^T\ \boldsymbol{\omega}_1^T]^T$ space if and only if

$$\boldsymbol{\eta}_{q(\boldsymbol{\theta})} = \Big\{\mathsf{H}_{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}}A(\boldsymbol{\eta}_{q(\boldsymbol{\theta})})\Big\}^{-1}\mathsf{D}_{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}}\underline{\text{NonEntropy}}\{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}, \boldsymbol{\omega}_1\}^T \text{ and} \tag{S.4}$$

$$\mathbf{0} = \mathsf{D}_{\boldsymbol{\omega}_1}\underline{\text{NonEntropy}}\{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}, \boldsymbol{\omega}_1\}^T \tag{S.5}$$

with A denoting the Multivariate Normal log-partition function and $\mathsf{D}_{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}}$ and $\mathsf{H}_{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}}$ respectively denoting the derivative vector and Hessian matrix with respect to $\boldsymbol{\eta}_{q(\boldsymbol{\theta})}$ as defined in Rohde & Wand (2016). Standard vector calculus steps (e.g. Wand, 2002) lead to

$$\begin{aligned}\mathsf{D}_{\boldsymbol{\omega}_1}\underline{\text{NonEntropy}}\{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}, \boldsymbol{\omega}_1\}^T = \big[&\text{expit}\{\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \tfrac{1}{2}(\mathbf{1} - 2\boldsymbol{\omega}_1)\odot\text{diagonal}(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T)\} - \boldsymbol{\omega}_1\big] \\ &\odot\text{diagonal}(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T).\end{aligned}$$

Substitution of this result into (S.5) and a reworking of the arguments that lead to Result 2 of Rohde & Wand (2016), but applied to $\underline{\text{NonEntropy}}\{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}, \boldsymbol{\omega}_1\}$, rather than $\text{NonEntropy}\{q(\boldsymbol{\theta}; \boldsymbol{\eta}_{q(\boldsymbol{\theta})})\}$, lead to the fixed-point updates:

$$\begin{cases}\boldsymbol{\omega}_1 \;\longleftarrow\; \text{expit}\{\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \tfrac{1}{2}(\mathbf{1} - 2\boldsymbol{\omega}_1)\odot\text{diagonal}(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T)\} \\ \boldsymbol{v}_{q(\boldsymbol{\theta})} \;\longleftarrow\; \mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\theta})}}\underline{\text{NonEntropy}}\{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}, \boldsymbol{\omega}_1\}^T \\ \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \;\longleftarrow\; -\Big[\mathsf{H}_{\boldsymbol{\mu}_{q(\boldsymbol{\theta})}}\underline{\text{NonEntropy}}\{\boldsymbol{\eta}_{q(\boldsymbol{\theta})}, \boldsymbol{\omega}_1\}\Big]^{-1} \\ \boldsymbol{\mu}_{q(\boldsymbol{\theta})} \;\longleftarrow\; \boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{v}_{q(\boldsymbol{\theta})}\end{cases} \tag{S.6}$$

Further vector calculus leads to the explicit forms:

$$\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\theta})}}\underline{\text{NonEntropy}}\{\boldsymbol{\eta}_{q(\boldsymbol{\theta})},\boldsymbol{\omega}_1\}^T = \boldsymbol{A}^T\Big(\boldsymbol{y} - \text{expit}\big[\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \tfrac{1}{2}(\mathbf{1}-2\boldsymbol{\omega}_1)\odot\text{diagonal}(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T)\big]\Big)$$
$$+(\boldsymbol{\eta}^\dagger)_1 + 2\,\text{vec}^{-1}\big((\boldsymbol{\eta}^\dagger)_2\big)\boldsymbol{\mu}_{q(\boldsymbol{\theta})}$$

and

$$\mathsf{H}_{\boldsymbol{\mu}_{q(\boldsymbol{\theta})}}\underline{\text{NonEntropy}}\{\boldsymbol{\eta}_{q(\boldsymbol{\theta})},\boldsymbol{\omega}_1\}^T =$$
$$-\boldsymbol{A}^T\text{diag}\left(\frac{\frac{1}{2}\mathbf{1}}{\mathbf{1}+\cosh\left[\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \frac{1}{2}(\mathbf{1}-2\boldsymbol{\omega}_1)\odot\text{diagonal}(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T)\right]}\right)\boldsymbol{A} + 2\,\text{vec}^{-1}\big((\boldsymbol{\eta}^\dagger)_2\big).$$

Introduction of $\boldsymbol{\omega}_0 \equiv \text{logit}(\boldsymbol{\omega}_1)$ and substitution into (S.6) then gives

$$\begin{cases} \boldsymbol{\omega}_0 \;\longleftarrow\; \boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \frac{1}{2}(\mathbf{1}-2\boldsymbol{\omega}_1)\odot\text{diagonal}(\boldsymbol{A}\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\boldsymbol{A}^T) \\ \boldsymbol{\omega}_1 \;\longleftarrow\; \text{expit}(\boldsymbol{\omega}_0)\;\;;\;\; \boldsymbol{\omega}_2 \longleftarrow \mathbf{1}/[2\{\mathbf{1}+\cosh(\boldsymbol{\omega}_0)\}] \\ \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} \;\longleftarrow\; \Big\{\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega}_2)\boldsymbol{A} - 2\text{vec}((\boldsymbol{\eta}^\dagger)_2)\Big\}^{-1} \\ \boldsymbol{\mu}_{q(\boldsymbol{\theta})} \;\longleftarrow\; \boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}\Big\{\boldsymbol{A}^T(\boldsymbol{y}-\boldsymbol{\omega}_1) + (\boldsymbol{\eta}^\dagger)_1 + 2\,\text{vec}^{-1}\big((\boldsymbol{\eta}^\dagger)_2\big)\boldsymbol{\mu}_{q(\boldsymbol{\theta})}\Big\}. \end{cases} \tag{S.7}$$

The remainder of the derivation of Algorithm 1 involves expressing (S.7) in terms of the input and output natural parameter vectors

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}} \quad\text{and}\quad \boldsymbol{\eta}_{\boldsymbol{\theta}\,\to\,p(\boldsymbol{y}|\,\boldsymbol{\theta})}.$$

Using (S.3), the $\boldsymbol{\omega}_0$ update is equivalent to

$$\begin{cases} \boldsymbol{\mu} \;\longleftarrow\; -\frac{1}{2}\boldsymbol{A}\Big\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}})_2\big)\Big\}^{-1}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}})_1 \\ \boldsymbol{\sigma}^2 \;\longleftarrow\; -\frac{1}{2}\text{diagonal}\Big[\boldsymbol{A}\Big\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}})_2\big)\Big\}^{-1}\boldsymbol{A}^T\Big] \\ \boldsymbol{\omega}_0 \;\longleftarrow\; \boldsymbol{\mu} + \frac{1}{2}(\mathbf{1}-2\boldsymbol{\omega}_1)\odot\boldsymbol{\sigma}^2. \end{cases}$$

The $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$ update can be written as

$$-\tfrac{1}{2}\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\leftrightarrow\,\boldsymbol{\theta}})_2\big)\}^{-1} \longleftarrow \Big\{\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega}_2)\boldsymbol{A} - 2\,\text{vec}^{-1}\big((\boldsymbol{\eta}^\dagger)_2\big)\Big\}^{-1}$$

which is equivalent to

$$(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}})_2 + (\boldsymbol{\eta}^\dagger)_2 \longleftarrow\; -\tfrac{1}{2}\text{vec}(\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega}_2)\boldsymbol{A}) + (\boldsymbol{\eta}^\dagger)_2$$

which, in turn, is equivalent to the second component of $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}}$ being updated according to

$$(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}})_2 \longleftarrow\; -\tfrac{1}{2}\text{vec}(\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega}_2)\boldsymbol{A}). \tag{S.8}$$

For the update of the first component of $\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})\,\to\,\boldsymbol{\theta}}$ we note that the last update of (S.7) is equivalent to

$$\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\theta})}\mu_{q(\boldsymbol{\theta})} \longleftarrow \boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\theta})}\boldsymbol{\mu}_{q(\boldsymbol{\theta})} + \boldsymbol{A}^T(\boldsymbol{y}-\boldsymbol{\omega}_1) + (\boldsymbol{\eta}^\dagger)_1 + 2\,\text{vec}^{-1}\big((\boldsymbol{\eta}^\dagger)_2\big)\boldsymbol{\mu}_{q(\boldsymbol{\theta})} \tag{S.9}$$

where, on the right-hand side,

$$\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\theta})} = \boldsymbol{A}^T\text{diag}(\boldsymbol{\omega}_2)\boldsymbol{A} - 2\,\text{vec}^{-1}\big((\boldsymbol{\eta}^\dagger)_2\big) \tag{S.10}$$

according to its updated value and

$$\mu_{q(\boldsymbol{\theta})} = -\tfrac{1}{2}\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_2\big)\}^{-1}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_1 \tag{S.11}$$

is the terms of the natural parameter from the previous iteration before (S.8) has taken place. Substitution of (S.10) and (S.11) into (S.9) we get

$$\begin{aligned}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \rightarrow \boldsymbol{\theta})_1 + (\boldsymbol{\eta}^{\dagger})_1 \longleftarrow \\ \Big\{ -\tfrac{1}{2}\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega}_2)\boldsymbol{A} + \text{vec}^{-1}\big((\boldsymbol{\eta}^{\dagger})_2\big)\Big\}\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_2\big)\}^{-1}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_1 \\ +\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{\omega}_1) + (\boldsymbol{\eta}^{\dagger})_1 - \text{vec}^{-1}\big((\boldsymbol{\eta}^{\dagger})_2\big)\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_2\big)\}^{-1}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_1\end{aligned}$$

which is equivalent to

$$\begin{aligned}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \rightarrow \boldsymbol{\theta})_1 &\longleftarrow -\tfrac{1}{2}\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega}_2)\boldsymbol{A}\{\text{vec}^{-1}\big((\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_2\big)\}^{-1}(\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \leftrightarrow \boldsymbol{\theta})_1 + \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{\omega}_1) \\ &= \boldsymbol{A}^T(\boldsymbol{\omega}_2 \odot \boldsymbol{\mu}) + \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{\omega}_1) = \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{\omega}_1 + \boldsymbol{\omega}_2 \odot \boldsymbol{\mu}).\end{aligned}$$

Combining this update with that given in (S.8) we get the following update for the full natural parameter vector:

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\,\boldsymbol{\theta})} \rightarrow \boldsymbol{\theta} \longleftarrow \begin{bmatrix} \boldsymbol{A}^T\big(\boldsymbol{y} - \boldsymbol{\omega}_1 + \boldsymbol{\omega}_2 \odot \boldsymbol{\mu}\big) \\ -\tfrac{1}{2}\text{vec}\big(\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega}_2)\boldsymbol{A}\big) \end{bmatrix}$$

which matches Algorithm 1.

S.3 Approximation of corr$(\beta_0, \beta_1|\boldsymbol{y})$

Consider the Bayesian linear regression model (8). Then, given the approximate noninformativity of prior distribution of $\boldsymbol{\beta} = [\beta_0\ \beta_1]$, the posterior covariance matrix of $\boldsymbol{\beta}$, Cov$(\boldsymbol{\beta}|\boldsymbol{y})$ is such that

$$\begin{aligned}\text{Cov}(\boldsymbol{\beta}|\boldsymbol{y}) &\approx \text{the inverse Fisher information matrix of } \boldsymbol{\beta} = [\boldsymbol{X}^T\text{diag}\{b''(\boldsymbol{X}\boldsymbol{\beta})\}\boldsymbol{X}]^{-1} \\ &= \left(\boldsymbol{X}^T\text{diag}\left[\frac{\boldsymbol{1}}{2\{\boldsymbol{1} + \cosh(\boldsymbol{X}\boldsymbol{\beta})\}}\right]\boldsymbol{X}\right)^{-1}\end{aligned}$$

where

$$\boldsymbol{X} \equiv \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

Straightforward algebra then leads to the approximate posterior correlation between β_0 and β_1 being

$$\text{corr}(\beta_0, \beta_1|\boldsymbol{y}) \approx \frac{-\frac{1}{n}\sum_{i=1}^n [x_i/\{1 + \cosh(\beta_0 + \beta_1 x_i)]}{\sqrt{\frac{1}{n}\sum_{i=1}^n [1/\{1 + \cosh(\beta_0 + \beta_1 x_i)\}]\ \frac{1}{n}\sum_{i=1}^n [x_i^2/\{1 + \cosh(\beta_0 + \beta_1 x_i)\}]}}.$$

However, the x_is are uniformly distributed on $(0, 1)$ so replacement of sample means by population means leads to the final approximation

$$\text{corr}(\beta_0, \beta_1|\boldsymbol{y}) \approx \frac{-\int_0^1 [x/\{1 + \cosh(\beta_0 + \beta_1 x)\}]\,dx}{\sqrt{\int_0^1 [1/\{1 + \cosh(\beta_0 + \beta_1 x)\}]\,dx\ \int_0^1 [x^2/\{1 + \cosh(\beta_0 + \beta_1 x)\}]\,dx}}.$$

S.4 Approximate Marginal Log-Likelihood Expressions

The simulation study described in Section 4 concerning variational inference for the Bayesian logistic regression model (2) used approximate marginal log-likelihood expressions, appropriate for the particular approach, as a means to assess convergence. Each of the expressions are given in this section. The last one uses the definition

$$B(\mu, \sigma^2) \equiv \int_{-\infty}^{\infty} b(\mu + \sigma\, x)\phi(x)\, dx$$

where, as defined in Section 3.2,

$$b(x) \equiv \log(1 + e^x).$$

As with the $\mathcal{B}_r$ notation given there, evaluations of $B(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ when $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are equal-sized column vectors are defined in an element-wise fashion, as illustrated by

$$B\left(\left[\begin{array}{c} 19 \\ 11 \end{array}\right], \left[\begin{array}{c} 36 \\ 28 \end{array}\right]\right) \equiv \left[\begin{array}{c} B(19, 36) \\ B(11, 28) \end{array}\right].$$

S.4.1 Jaakkola-Jordan Updates

$$\begin{aligned} \log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\xi}) \;\; = \;\; & \tfrac{1}{2} \log |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta};\boldsymbol{\xi})}| + \tfrac{1}{2} \boldsymbol{\mu}_{q(\boldsymbol{\beta};\boldsymbol{\xi})}^T \boldsymbol{\Sigma}_{q(\boldsymbol{\beta};\boldsymbol{\xi})}^{-1} \boldsymbol{\mu}_{q(\boldsymbol{\beta};\boldsymbol{\xi})} - \tfrac{1}{2} \boldsymbol{\mu}_{\boldsymbol{\beta}}^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} \\ & + \sum_{i=1}^{n} \{\xi/2 - \log(1 + e^{\xi_i}) + (\xi/4) \tanh(\xi_i/2)\} - \frac{1}{2} \log |\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| \end{aligned}$$

where

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta};\boldsymbol{\xi})} \equiv \left[\boldsymbol{X}^T \text{diag} \left\{ \frac{\tanh(\boldsymbol{\xi}/2)}{2\boldsymbol{\xi}} \right\} \boldsymbol{X} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \right]^{-1} \quad \text{and} \quad \boldsymbol{\mu}_{q(\boldsymbol{\beta};\boldsymbol{\xi})} \equiv \boldsymbol{\Sigma}_{q(\boldsymbol{\beta};\boldsymbol{\xi})} \{ \boldsymbol{X}^T (\boldsymbol{y} - \tfrac{1}{2}\mathbf{1}) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} \}.$$

and $\boldsymbol{\xi}$ is the current value of the variational parameter vector that arises in the Jaakkola-Jordan device. See, for example, Section 5.1 of Wand (2017).

S.4.2 Saul-Jordan Updates

$$\begin{aligned} \log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\omega}) \;\; = \;\; & \tfrac{1}{2} \log |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| - \tfrac{1}{2} \text{tr} \big[\big\{ \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \big\{ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} + (\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}})(\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^T \big\} \big] \\ & + \boldsymbol{y}^T \boldsymbol{X} \boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \tfrac{1}{2} (\boldsymbol{\omega}^2)^T \text{diagonal}(\boldsymbol{X} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \boldsymbol{X}^T) \\ & - \mathbf{1}^T \log[\mathbf{1} + \exp\{ \boldsymbol{X} \boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \tfrac{1}{2} (\mathbf{1} - 2\boldsymbol{\omega}) \odot \text{diagonal}(\boldsymbol{X} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \boldsymbol{X}^T) \}] \\ & + \frac{d}{2} - \frac{1}{2} \log |\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| \end{aligned}$$

where $\boldsymbol{\omega}$ is the current value of the variational parameter vector that arises in the Saul-Jordan device.

S.4.3 Knowles-Minka-Wand Updates

$$\begin{aligned} \log \underline{p}(\boldsymbol{y}; q) \;\; = \;\; & \tfrac{1}{2} \log |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| - \tfrac{1}{2} \text{tr} \big[\big\{ \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \big\{ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} + (\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}})(\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^T \big\} \big] \\ & + \boldsymbol{y}^T \boldsymbol{X} \boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \mathbf{1}^T B\big(\boldsymbol{X} \boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \text{diagonal}(\boldsymbol{X} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \boldsymbol{X}^T) \big) \\ & + \frac{d}{2} - \frac{1}{2} \log |\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| \end{aligned}$$

Additional References

Rohde, D. and Wand, M.P. (2016). Semiparametric mean field variational Bayes: General principles and numerical issues. *Journal of Machine Learning Research*, **17(172)**, 1–47.

Wand, M.P. (2002). Vector differential calculus in statistics. *The American Statistician*, **56**, 55–62.