Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.139

Accurate logistic variational message passing: algebraic and numerical details

Tui H. Nolan[a,b] and Matt P. Wand[a,b*]

Received 19 January 2017; Accepted 26 January 2017

We provide full algebraic and numerical details required for fitting accurate logistic likelihood regression-type models via variational message passing with factor graph fragments. Existing methodology of this type involves the Jaakkola–Jordan device, which is prone to poor accuracy. We examine two alternatives: the Saul–Jordan tilted bound device and conjugacy enforcement via multivariate normal prespecification of a key message. Both of these approaches appear in related literature. Our contributions facilitate immediate implementation within variational message passing schemes. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: approximate Bayesian inference; factor graph; generalized additive models; generalized linear mixed models; mean field variational Bayes

1 Introduction

Wand (2016) describes a general extendible framework for fast approximate inference in semiparametric regression and, conceptually, various other statistical models via mean field variational Bayes and its semiparametric extensions. The key components are message passing on factor graphs, known as *variational message passing* (VMP) in the mean field variational Bayes context, and factor graph fragments that allow compartmentalization of the updates. For the important special case of logistic likelihood semiparametric regression, Wand (2016) presents message updating formulae based on the Jaakkola–Jordan device (Jaakkola & Jordan, 2000). However, Knowles & Minka (2011) point out that the Jaakkola–Jordan message passing can lead to poor inferential accuracy and propose two alternatives for logistic VMP. There are

1. the tilted bound device developed by Saul & Jordan (1999) and
2. prespecification that the message from the logistic likelihood factor has a multivariate normal distribution as proposed by Knowles & Minka (2011) and further developed by Wand (2014).

The numerical studies given in Section 6 of Knowles & Minka (2011) show that these alternative approaches deliver more accurate Bayesian inference compared with that based on the Jaakkola–Jordan device.

In this article, we give full algebraic and numerical details for embedding these more accurate logistic likelihood fragment updates into the VMP framework given by Wand (2016). As mentioned by Knowles & Minka (2011), the Knowles–Minka–Wand approach has the drawback of intractable univariate integrals arising in the update expressions. Moreover, the integral families are such that standard numerical integration techniques can break down for certain

[a]School of Mathematical and Physical Sciences, University of Technology Sydney, PO Box 123, Broadway 2007, Australia
[b]Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers, Parkville 3052, Australia
*Email: matt.wand@uts.edu.au

parameter settings. The number of such integrals scales with sample size values, so in big applications, their evaluation can be a limiting factor. Using ideas from Monahan & Stefanski (1989), we formulate a robust and fast version of the Knowles–Minka–Wand logistic updates that is devoid of quadrature-related issues.

In Section 2, we give a brief description of VMP and its compartmentalization through factor graph fragments. Section 3 focuses on the logistic likelihood fragment and gives the algebraic and numerical details of the message parameter updates. Important practical issues concerning convergence stability and inferential accuracy are dealt with in Section 4 via a simulation study. It is demonstrated that the new approaches achieve greater levels of accuracy than the Jaakkola–Jordan device but, in some circumstances, suffer from lack of stability. Section 5 illustrates the utility of the more accurate logistic VMP schemes in the context of generalized additive model-based data analysis. Our conclusions are presented in Section 6.

1.1 Notation

Section 2.2 of Wand (2016) summarizes matrix notation useful for VMP, such as the vec and vec^{-1} functions. Such notation is also used here. An additional piece of notation, used in Algorithm 2, is that $\boldsymbol{A}/\boldsymbol{B}$ denotes the element-wise quotient of two equal-sized matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. The following scalar-argument functions are used throughout this article:

$$\text{expit}(x) \equiv \text{logit}^{-1}(x) = 1/\{1+\exp(-x)\}, \quad \phi(x) \equiv (2\pi)^{-1/2}\exp\left(-\frac{1}{2}x^2\right) \quad \text{and} \quad \Phi(x) \equiv \int_{-\infty}^{x}\phi(u)\,du.$$

2 Variational message passing and factor graph fragments

VMP is a general approach to obtaining mean field variational Bayes approximations to posterior density functions in a graphical model (e.g. Bishop, 2006). There are several variants of VMP in the literature, and here, we follow Minka (2005), which is summarized in Section 2.5 of Wand (2016). In Sections 3-5 of Wand (2016), a compartmental approach to VMP is described, with *factor graph fragments* acting as building blocks.

3 The logistic likelihood fragment

The logistic likelihood fragment is shown in Figure 1 and corresponds to the specification

$$y_i\,|\,\boldsymbol{\theta} \overset{\text{ind.}}{\sim} \text{Bernoulli}\big[\text{expit}\{(\boldsymbol{A\theta})_i\}\big], \quad 1 \le i \le n,$$

where $\boldsymbol{A}$ is a generic design matrix and $\boldsymbol{\theta}$ is a generic vector of coefficients. Section 5.1 of Wand (2016) describes Jaakkola–Jordan updates for the natural parameter of the message passed from $p(\boldsymbol{y}\,|\,\boldsymbol{\theta})$ to $\boldsymbol{\theta}$, which we denote by $m_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\to\boldsymbol{\theta}}(\boldsymbol{\theta})$. The Jaakkola–Jordan device entails replacement of $m_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\to\boldsymbol{\theta}}(\boldsymbol{\theta})$ by a multivariate normal message with natural parameter vector $\eta_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\to\boldsymbol{\theta}}$. Such an approximation ensures conjugacy with multivariate normal messages passed to $\boldsymbol{\theta}$ from other factors.

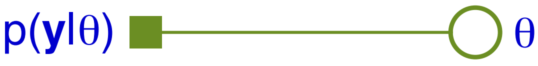

Figure 1. The logistic likelihood factor graph fragment.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.139

3.1 Saul–Jordan updates for the logistic likelihood fragment

Saul–Jordan updates for the logistic likelihood fragment are based on the following fact:

$$\text{if } x \sim N(\mu,\sigma^2) \text{ then, for any } \omega \in \mathbb{R},\ E\{\log(1+e^x)\} \le \frac{1}{2}\omega^2\sigma^2 + \log\left[1+\exp\left\{\mu+\frac{1}{2}(1-2\omega)\sigma^2\right\}\right] \tag{1}$$

(Saul & Jordan, 1999). The last expression in (1) has been labelled a *tilted bound* (e.g. Knowles & Minka, 2011) on $E\{\log(1+e^x)\}$ for $x \sim N(\mu,\sigma^2)$. It follows from the expression $E\{\log(1+e^x)\} = \omega\,\mu + E[\log\{e^{-\omega x}+e^{(1-\omega)x}\}]$, and application of Jensen's inequality to the second term.

To understand how (1) can aid logistic likelihood variational inference, consider the Bayesian logistic regression model

$$y_i\,|\,\boldsymbol{\beta} \overset{\text{ind.}}{\sim} \text{Bernoulli}\big[\text{expit}\{(\boldsymbol{X\beta})_i\}\big],\ 1 \le i \le n, \quad \boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \tag{2}$$

where $\boldsymbol{\beta}$ is a $d \times 1$ coefficient vector, $\boldsymbol{X}$ is a design matrix and $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ are hyperparameters. If the posterior density function $p(\boldsymbol{\beta}\,|\boldsymbol{y})$ is approximated by $q(\boldsymbol{\beta})$ where q is a $N(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})$ density function, then, courtesy of (1), the marginal log-likelihood can be bounded as follows:

$$\begin{aligned}
\log p(\boldsymbol{y}) \ge \log \underline{p}(\boldsymbol{y};q) \equiv{}& \frac{d}{2} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| + \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - E_q[\boldsymbol{1}^T\log\{1+\exp(\boldsymbol{X\beta})\}] \\
&- \frac{1}{2}\text{tr}\big[\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\{(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}-\boldsymbol{\mu}_{\boldsymbol{\beta}})(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^T + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\}\big] \\
\ge{}& \frac{d}{2} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| + \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \frac{1}{2}(\boldsymbol{\omega}^2)^T\text{diagonal}(\boldsymbol{X}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{X}^T) \\
&- \boldsymbol{1}^T\log\left[\boldsymbol{1}+\exp\left\{\boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \frac{1}{2}(\boldsymbol{1}-2\boldsymbol{\omega})\odot\text{diagonal}(\boldsymbol{X}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{X}^T)\right\}\right] \\
&- \frac{1}{2}\text{tr}\big[\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\{(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}-\boldsymbol{\mu}_{\boldsymbol{\beta}})(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^T + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\}\big]
\end{aligned} \tag{3}$$

where $\boldsymbol{\omega}$ is an $n \times 1$ vector of additional variational parameters. The Saul–Jordan updates are founded upon fixed-point iterative updates that aim to maximize the last lower bound in (3), although for a localized approximate marginal log-likelihood on a potentially much larger factor graph. The full details are given in Section S.2 of the supporting information and lead to Algorithm 1. The $\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\to\boldsymbol{\theta}}$ and $\boldsymbol{\eta}_{\boldsymbol{\theta}\to p(\boldsymbol{y}\,|\,\boldsymbol{\theta})}$ notations used in Algorithm 1 corresponds to the natural parameter vectors of the messages passed between $p(\boldsymbol{y}\,|\,\boldsymbol{\theta})$ and $\boldsymbol{\theta}$ as in Wand (2016). In addition,

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

Each of the messages is proportional to Multivariate Normal density functions. For example, the message from $p(\boldsymbol{y}\,|\,\boldsymbol{\theta})$ to $\boldsymbol{\theta}$ is

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

3.2 Knowles–Minka–Wand updates for the logistic likelihood fragment

Knowles–Minka–Wand updates for particular likelihood fragments are based on the general approach for handling non-conjugate messages, via prespecification of an exponential family form, given in Knowles & Minka (2011) and the explicit formulae for the multivariate normal case given in Wand (2014). The factor in the logistic likelihood fragment is

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

(wileyonlinelibrary.com) DOI: 10.1002/sta4.139

Algorithm 1 The inputs, Saul–Jordan updates and outputs for the logistic likelihood fragment.

```
Data inputs: $\boldsymbol{y}, \boldsymbol{A}$

Parameter inputs: $\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\to\boldsymbol{\theta}},\ \boldsymbol{\eta}_{\boldsymbol{\theta}\to p(\boldsymbol{y}\,|\,\boldsymbol{\theta})},\ \boldsymbol{\omega}_1$

Updates:
    $\boldsymbol{\mu} \longleftarrow -\frac{1}{2}\boldsymbol{A}\Big\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\leftrightarrow\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\big(\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\leftrightarrow\boldsymbol{\theta}}\big)_1\ ;\ \boldsymbol{\sigma}^2 \longleftarrow -\frac{1}{2}\text{diagonal}\Big[\boldsymbol{A}\Big\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\leftrightarrow\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\boldsymbol{A}^T\Big]$
    $\boldsymbol{\omega}_0 \longleftarrow \boldsymbol{\mu} + \frac{1}{2}(1-2\boldsymbol{\omega}_1)\odot\boldsymbol{\sigma}^2\ ;\ \boldsymbol{\omega}_1 \longleftarrow \text{expit}(\boldsymbol{\omega}_0)\ ;\ \boldsymbol{\omega}_2 \longleftarrow 1/[\,2\{\boldsymbol{1}+\cosh(\boldsymbol{\omega}_0)\}]$
    $\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\to\boldsymbol{\theta}} \longleftarrow \begin{bmatrix}\boldsymbol{A}^T(\boldsymbol{y}-\boldsymbol{\omega}_1+\boldsymbol{\omega}_2\odot\boldsymbol{\mu})\\ -\frac{1}{2}\text{vec}\big(\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega}_2)\boldsymbol{A}\big)\end{bmatrix}$

Parameter outputs: $\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\to\boldsymbol{\theta}},\ \boldsymbol{\omega}_1$
```

which is a special case of the one-parameter exponential family likelihoods that take the general form

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \exp[\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{1}^T b(\boldsymbol{A}\boldsymbol{\theta}) + \boldsymbol{1}^T c(\boldsymbol{y})\}]$$

for functions b and c. The *Poisson* likelihood fragment is the other prominent member of this family, for which $b(x) = e^x$ and the Knowles–Minka–Wand updates for the Poisson likelihood fragment are summarized in Section 5.1 of Wand (2016) and derived in Section S.2.4. of that article's supporting information. The only important difference when switching to the logistic likelihood fragment is that now

$$b(x) \equiv \log(1+e^x). \tag{4}$$

In theory, the Knowles–Minka–Wand updates just involve replacement of $b(x) = e^x$ by (4). In practice, this change of b functions leads to intractable integrals and the need for robust numerical integration schemes. Introduce the notation

$$\mathcal{B}_r(\mu,\sigma^2) \equiv \int_{-\infty}^{\infty} x^r\,\text{expit}(\mu+\sigma\,x)\phi(x)\,dx, \quad r=0,1,\ \mu\in\mathbb{R},\ \sigma^2>0 \tag{5}$$

where expit and ϕ are defined in Section 1.1. Also, if $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are column vectors of the same dimension, then $\mathcal{B}_r(\boldsymbol{\mu},\boldsymbol{\sigma}^2)$ is evaluated in an element-wise fashion, as illustrated by

$$\mathcal{B}_r\left(\begin{bmatrix}9\\5\end{bmatrix},\begin{bmatrix}7\\2\end{bmatrix}\right) \equiv \begin{bmatrix}\mathcal{B}_r(9,7)\\ \mathcal{B}_r(5,2)\end{bmatrix}.$$

Then a reworking of the arguments given in Section S.2.4 of the supporting information of Wand (2016) leads to the following fragment updates:

$$\begin{aligned}
&\boldsymbol{\mu} \longleftarrow -\frac{1}{2}\boldsymbol{A}\Big\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\leftrightarrow\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\big(\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\leftrightarrow\boldsymbol{\theta}}\big)_1\\
&\boldsymbol{\sigma}^2 \longleftarrow -\frac{1}{2}\text{diagonal}\Big[\boldsymbol{A}\Big\{\text{vec}^{-1}\Big(\big(\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\leftrightarrow\boldsymbol{\theta}}\big)_2\Big)\Big\}^{-1}\boldsymbol{A}^T\Big]\\
&\boldsymbol{\omega}_1 \longleftarrow \mathcal{B}_0(\boldsymbol{\mu},\boldsymbol{\sigma}^2)\ ;\ \boldsymbol{\omega}_2 \longleftarrow \mathcal{B}_1(\boldsymbol{\mu},\boldsymbol{\sigma}^2)/\boldsymbol{\sigma}\ ;\ \boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\to\boldsymbol{\theta}} \longleftarrow \begin{bmatrix}\boldsymbol{A}^T(\boldsymbol{y}-\boldsymbol{\omega}_1+\boldsymbol{\omega}_2\odot\boldsymbol{\mu})\\ -\frac{1}{2}\text{vec}\big(\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega}_2)\boldsymbol{A}\big)\end{bmatrix}
\end{aligned}\ . \tag{6}$$

In practice, we require an effective strategy for computing the integrals given by (4) and (5). Knowles & Minka (2011) claim efficient computation of $\mathcal{B}_r(\mu,\sigma^2)$ using Gauss–Hermite or Clenshaw–Curtis quadrature, although details are not given. A similar advice is provided by Ormerod & Wand (2012) for integral families that include $\mathcal{B}_r(\mu,\sigma^2)$. Despite these claims, we found it quite challenging to devise quadrature schemes that provide robust, efficient and accurate evaluation of $\mathcal{B}_r(\mu,\sigma^2)$ for all $(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+$. Instead, we turned to the normal scale mixture uniform approximations of expit obtained by Monahan & Stefanski (1989), which, for any $k\in\mathbb{N}$, take the form

$$\sup_{x\in\mathbb{R}}\left|\text{expit}(x)-\text{expit}_k(x)\right| = \Delta_k \qquad (7)$$

where $\text{expit}_k(x)\equiv\sum_{i=1}^{k}p_{k,i}\Phi(s_{k,i}x)$ for constants $p_{k,i}$ and $s_{k,i}$. Monahan & Stefanski (1989) provide values of constants $p_{k,i}$ and $s_{k,i}$ and bounds Δ_k for $1\le k\le 8$. If expit is replaced by expit_k in the $\mathcal{B}_r(\mu,\sigma^2)$ expressions, then standard integral results given by Section S.1 in the supporting information lead to the approximations

$$\mathcal{B}_0(\mu,\sigma^2)\approx\sum_{i=1}^{k}p_{k,i}\Phi\left(\frac{\mu\,s_{k,i}}{\sqrt{1+\sigma^2 s_{k,i}^2}}\right)\quad\text{and}\quad\mathcal{B}_1(\mu,\sigma^2)\approx\sigma\sum_{i=1}^{k}\frac{p_{k,i}\,s_{k,i}}{\sqrt{1+\sigma^2 s_{k,i}^2}}\,\phi\left(\frac{\mu\,s_{k,i}}{\sqrt{1+\sigma^2 s_{k,i}^2}}\right).$$

Theorem 1 provides uniform bounds for the error in these approximations.

Theorem 1
Consider the integrals defined by (5). Then

(a) $\sup_{\mu\in\mathbb{R},\sigma>0}\left|\mathcal{B}_0(\mu,\sigma^2)-\sum_{i=1}^{k}p_{k,i}\Phi\left(\frac{\mu\,s_{k,i}}{\sqrt{1+\sigma^2 s_{k,i}^2}}\right)\right|\le\Delta_k$ *and*

(b) $\sup_{\mu\in\mathbb{R},\sigma>0}\left|\mathcal{B}_1(\mu,\sigma^2)-\sigma\sum_{i=1}^{k}\frac{p_{k,i}\,s_{k,i}}{\sqrt{1+\sigma^2 s_{k,i}^2}}\,\phi\left(\frac{\mu\,s_{k,i}}{\sqrt{1+\sigma^2 s_{k,i}^2}}\right)\right|\le\sqrt{2/\pi}\,\Delta_k.$

Table I lists $p_{8,i}$ and $s_{8,i}$ values, $1\le i\le 8$. From now on, we will use this $k=8$ version of the Monahan–Stefanski approximation, and let $\boldsymbol{p}$ and $\boldsymbol{s}$ be the vectors containing these constants. According to Theorem 1, the uniform bound on the error for the $\mathcal{B}_0(\mu,\sigma^2)$ approximation is $\Delta_8 = 2.9\times10^{-9}$, whilst that for the $\mathcal{B}_0(\mu,\sigma^2)$ approximation is

Table I. Entries of the 8×1 vectors $\boldsymbol{p}$ and $\boldsymbol{s}$ corresponding to the $k=8$ normal scale mixture uniform approximation of Monahan & Stefanski (1989).

Entries of $\boldsymbol{p}$	Entries of $\boldsymbol{s}$
0.00324 63432 72134	1.36534 08062 96348
0.05151 74770 33972	1.05952 39710 16916
0.19507 79126 73858	0.83079 13137 65644
0.31556 98236 32818	0.65073 21666 39391
0.27414 95761 58423	0.50813 54253 66489
0.13107 68806 95470	0.39631 33451 66341
0.02791 24187 27972	0.30890 42522 67995
0.00144 95678 05354	0.23821 26164 09306

$\sqrt{2/\pi} \times 2.9 \times 10^{-9} < 2.4 \times 10^{-9}$. Such bounds allow for the possibility of relative error checks on Monahan–Stefanski approximation of $\mathcal{B}_r(\mu, \sigma^2)$ when used in practice.

Even though (6) represents the pure form of the Knowles–Minka–Wand updates for the logistic likelihood fragment, with the method of numerical integration left to the user, Algorithm 2 provides a quick-to-compute and robust version that uses Monahan–Stefanski numerical integration. The updates for the vectors $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are given in matrix notation according to conventions of Section 2.2 of Wand (2016). Note that $\mathbf{1}_n$ is the $n \times 1$ vector of 1s where n is the number of rows in $\boldsymbol{y}$ and $\boldsymbol{A}$. Similarly, $\mathbf{1}_8$ is the 8×1 vectors of ones. Section 1.1 provides relevant matrix notation.

4 Stability and accuracy assessment

To test the stability and accuracy of the Saul–Jordan and Knowles–Minka–Wand updates, we ran a simulation study for fitting and inference in the simple linear Bayesian logistic regression model

$$y_i | \beta_0, \beta_1 \overset{\text{ind.}}{\sim} \text{Bernoulli}(\text{expit}(\beta_0 + \beta_1 x_i)), \quad 1 \le i \le 100, \quad \beta_0, \beta_1 \overset{\text{ind.}}{\sim} N(0, 10^{10}) \tag{8}$$

with the x_is drawn randomly from the uniform distribution on $(0, 1)$. The Jaakkola–Jordan updates were also considered in the study.

Preliminary runs indicated that the performance and relative performance of the three approaches were affected by the degree of posterior correlation between the intercept and slope parameters:

$$\text{corr}(\beta_0, \beta_1 \,|\, \boldsymbol{y}) \equiv \text{correlation between } \beta_0 \text{ and } \beta_1 \text{ given } \boldsymbol{y}.$$

Algorithm 2 The inputs, Knowles–Minka–Wand updates, with $k = 8$ Monahan–Stefanski numerical integration, and outputs of the logistic fragment.

Data inputs: $\boldsymbol{y}$, $\boldsymbol{A}$,

Constant inputs: $\boldsymbol{p}$, $\boldsymbol{s}$ as listed in Table I

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

Updates:

$$\boldsymbol{\mu} \longleftarrow -\frac{1}{2}\boldsymbol{A}\left\{\text{vec}^{-1}\left(\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta})\leftrightarrow\boldsymbol{\theta}}\right)_2\right)\right\}^{-1}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta})\leftrightarrow\boldsymbol{\theta}}\right)_1 \;;\; \boldsymbol{\sigma}^2 \longleftarrow -\frac{1}{2}\text{diagonal}\left[\boldsymbol{A}\left\{\text{vec}^{-1}\left(\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta})\leftrightarrow\boldsymbol{\theta}}\right)_2\right)\right\}^{-1}\boldsymbol{A}^T\right]$$

$$\boldsymbol{\Omega} \longleftarrow \sqrt{\mathbf{1}_n\mathbf{1}_8^T + (\sigma^2)(\boldsymbol{s}^2)^T} \;\;;\;\; \boldsymbol{\omega}_3 \longleftarrow \Phi\left((\boldsymbol{\mu}\boldsymbol{s}^T)/\boldsymbol{\Omega}\right)\boldsymbol{p} \;\;;\;\; \boldsymbol{\omega}_4 \longleftarrow \left\{\phi\left((\boldsymbol{\mu}\boldsymbol{s}^T)/\boldsymbol{\Omega}\right)/\boldsymbol{\Omega}\right\}(\boldsymbol{p} \odot \boldsymbol{s})$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta})\to\boldsymbol{\theta}} \longleftarrow \begin{bmatrix} \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{\omega}_3 + \boldsymbol{\omega}_4 \odot \boldsymbol{\mu}) \\ -\frac{1}{2}\text{vec}(\boldsymbol{A}^T\text{diag}(\boldsymbol{\omega}_4)\boldsymbol{A}) \end{bmatrix}.$$

Parameter outputs: $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\theta})\to\boldsymbol{\theta}}$

(wileyonlinelibrary.com) DOI: 10.1002/sta4.139

Table II. True values of β_0 and β_1 used in the Bayesian logistic regression simulation study and corresponding approximate value of corr$(\beta_0, \beta_1|\boldsymbol{y})$.

Setting	True β_0	True β_1	Approximate corr$(\beta_0, \beta_1\,|\,\boldsymbol{y})$
1	0.5	3.18	–0.8
2	–2.2	3.8	–0.9
3	–7.5	9.36	–0.98
4	16.1	–19.05	–0.995
5	–24.0	28.03	–0.9975

Using an approximation to corr$(\beta_0, \beta_1\,|\,\boldsymbol{y})$ described in Section S.3 of the supporting information, we settled upon the true parameter values given in Table II.

For each setting in Table II, we generated 100 samples from (8) and then applied VMP with each of three types of updates for the logistic likelihood fragment. We used 25 iterations of the Jaakkola–Jordan approach to obtain starting values for the Saul–Jordan and Knowles–Minka–Wand approaches. We also obtained 1,000,0000 Markov chain Monte Carlo draws from the posterior distribution of (β_0, β_1) using the R package rstan (Stan Development Team, 2016). The warmup value was set to 1,000.

Convergence of the VMP iterations was assessed using relative absolute change in the approximate marginal log-likelihood corresponding to each approach, given in Section S.4 of the supporting information. Specifically, if $\log\{\underline{p}(y;q)\}_{\text{prev}}$ and $\log\{\underline{p}(y;q)\}_{\text{curr}}$, respectively, denote the approximate marginal log-likelihood for the previous and current iterations, then the condition

$$\left|\frac{\log\{\underline{p}(y;q)\}_{\text{curr}}}{\log\{\underline{p}(y;q)\}_{\text{prev}}} - 1\right| < 10^{-10}$$

was used to determine whether or not convergence had been achieved. For settings 1 and 2, convergence was usually achieved in 10–20 iterations for all three approaches. For setting 3, the VMP algorithms based on the Saul–Jordan and Knowles–Minka–Wand approaches failed to converge after 1,000 iterations for about 5–10% of the replications. In settings 4 and 5, there were much higher percentages of non-convergence with the approximate marginal log-likelihood often oscillating between two convergents rather than a single one. More seriously, for some of the replications, the q-density parameters arising from the VMP iterations diverged to values very far from the true simulation values. For setting 5, this happened most of the time for the Saul–Jordan updates, and occasionally for the Knowles–Minka–Wand updates. The Jaakkola–Jordan updates appear to be immune to such a breakdown. Figure 2 uses traffic light colour-coded bar charts to summarize the convergence/divergence behaviour of the VMP algorithms.

We also recorded accuracy scores defined by

$$\text{accuracy}(\beta_j) \equiv 1 - \frac{1}{2}\int_{-\infty}^{\infty} |q^*(\beta_j) - p(\beta_j\,|\,\boldsymbol{y})|\,d\beta_j, \quad j = 0, 1.$$

Because $p(\beta_j\,|\,\boldsymbol{y})$ does not have a closed form, we approximated it using kernel density estimation with a default direct plug-in bandwidth selector in the R package KernSmooth (Wand & Ripley, 2015). The kernel density estimates should be of high quality given the sample sizes of 1,000,000 and the simple shapes of the estimands. Figure 3 provides boxplot summaries of the accuracy scores. The replications for which the VMP algorithm wildly diverged are omitted in these accuracy score summaries. Therefore, Figures 2 and 3 should be viewed together when evaluating the relative performance of the three approaches in terms of stability and accuracy.

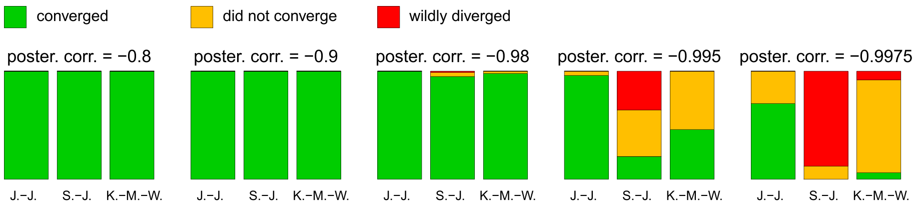

Figure 2. Proportions of replications of the Bayesian simple logistic regression simulation study for which (a) the approximate marginal log-likelihood converged in less than 1,000 iterations ("converged"), (b) the approximate marginal log-likelihood failed to converge after 1,000 iterations but still returning reasonable estimates ("did not converge") and (c) diverged without returning reasonable estimates ("wildly diverged").

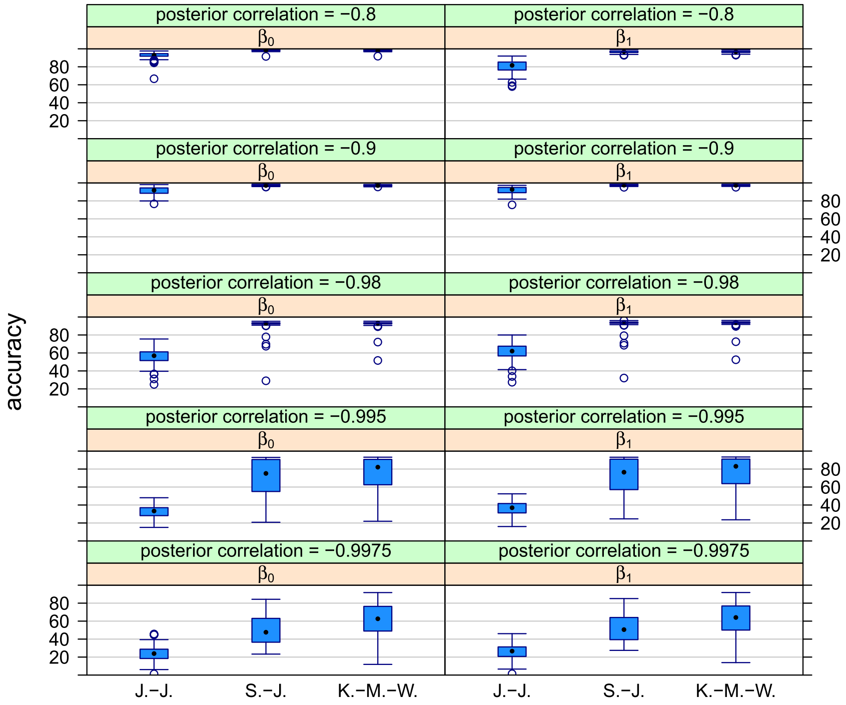

Figure 3. Side-by-side boxplots of the accuracy scores of $q^*(\beta_0)$ and $q^*(\beta_1)$ for the Bayesian simple logistic regression simulation study, with variational message passing based on each of the Jaakkola–Jordan, Saul–Jordan and Knowles–Minka–Wand updates.

It is clear that Knowles–Minka–Wand updates, with $k = 8$ Monahan–Stefanski numerical integration, provide more accurate inference than the Jaakkola–Jordan updates. Their accuracy advantage over the Saul–Jordan updates is less pronounced, but they are much more stable. When both accuracy and stability are taken into account, the Knowles–Minka–Wand approach is the best of the three approaches considered here. In addition, unless the absolute posterior correlation between β_0 and β_1 is very high, Figure 3 indicates that the Knowles–Minka–Wand approach usually leads to very accurate Bayesian inference with accuracy scores close to 100%.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.139

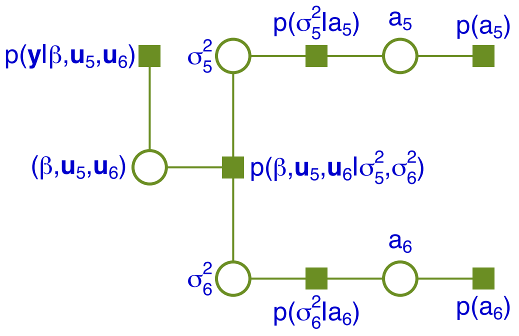
$p(\mathbf{y}|\beta,\mathbf{u}_5,\mathbf{u}_6)$
σ_5^2
$p(\sigma_5^2|a_5)$
a_5
$p(a_5)$
$(\beta,\mathbf{u}_5,\mathbf{u}_6)$
$p(\beta,\mathbf{u}_5,\mathbf{u}_6|\sigma_5^2,\sigma_6^2)$
σ_6^2
a_6
$p(\sigma_6^2|a_6)$
$p(a_6)$

Figure 4. Factor graph for the logistic additive model (9).

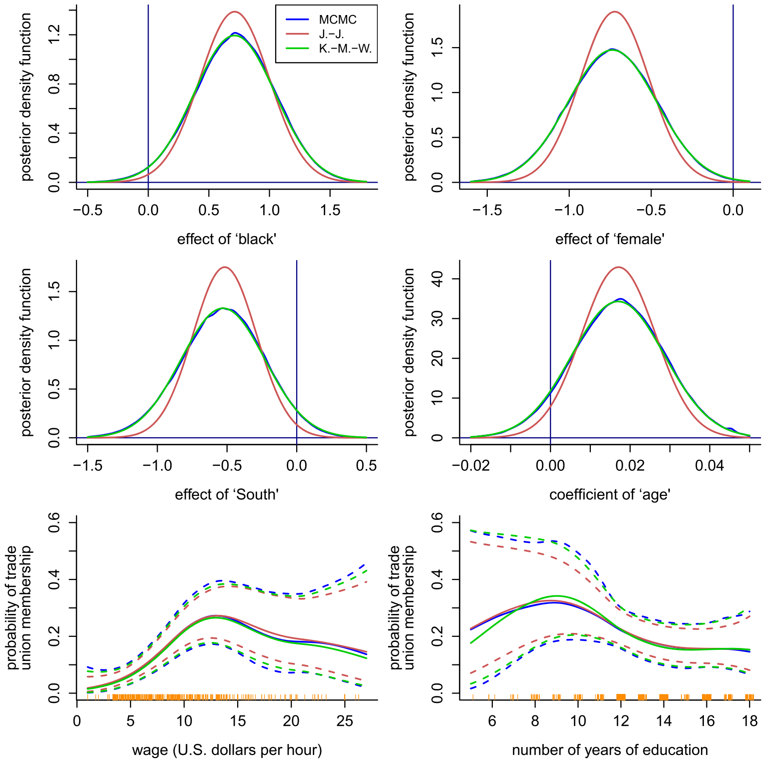

Figure 5. Comparison of approximate Bayesian inference methods for the trade union logistic additive model example. The methods are 1,000,000 Markov chain Monte Carlo draws, variational message passing (VMP) with the Jaakkola–Jordan logistic updates and VMP with the Knowles–Minka–Wand updates. VMP with the Saul–Jordan updates was not stable in this example and broke down after 19 iterations after initialization with 25 Jaakkola–Jordan logistic updates. Upper four panels: approximate posterior density density functions of the coefficients of the linear components. Lower two panels: estimated non-linear functions of wage and number of years of education with all other variables set to their average value. The solid curves correspond to posterior medians, and the dashed curves correspond to pointwise 95% credible intervals.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.139

5 Logistic additive model illustration

As explained by Wand (2016), the fragment updates given in Algorithms 1 and 2 can be used to fit arbitrarily large semiparametric regression models. We provide illustrations here for a logistic additive model analysis of data on trade union membership (e.g. Berndt, 1991). The model is

$$\begin{gathered}y_i\,|\,\boldsymbol{\beta},\boldsymbol{u}_5,\boldsymbol{u}_6 \overset{\text{ind.}}{\sim} \text{Bernoulli}\big(\text{expit}(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i}+\beta_4 x_{4i}+f_5(x_{5i})+f_6(x_{6i}))\big),\\ \boldsymbol{\beta} \equiv [\beta_0,\ldots,\beta_6]^T \sim N(\boldsymbol{\mu_\beta},\boldsymbol{\Sigma_\beta}) \quad f_j(x)=\beta_j x+\sum\nolimits_{k=1}^{K_j} u_{jk}\,z_{jk}(x),\ j=5,6,\\ \boldsymbol{u}_j \equiv [u_{j1},\ldots,u_{jK_j}]^T\,|\sigma_j^2 \sim N(\boldsymbol{0},\sigma_j^2\boldsymbol{I}),\quad \sigma_j^2|a_j \sim \text{Inverse}-\chi^2(1,1/a_j),\quad a_j \sim \text{Inverse}-\chi^2(1,1/A_{uj}^2),\ j=5,6\end{gathered} \tag{9}$$

where $\{z_{jk} : 1 \le k \le K_j\}$ are suitable spline bases (e.g. Wand & Ormerod, 2008).

Figure 4 displays the factor graph corresponding to (9) with the joint posterior density approximation admitting the product density approximation $p(\boldsymbol{\beta},\boldsymbol{u}_5,\boldsymbol{u}_6,\sigma_5^2,\sigma_6^2,a_5,a_6\,|\,\boldsymbol{y}) \approx q(\boldsymbol{\beta},\boldsymbol{u}_5,\boldsymbol{u}_6)q(\sigma_5^2)q(\sigma_6^2)q(a_5)q(a_6)$.

Figure 5 shows the results of (9) to the trade union membership data. The response is an indicator of whether or not a USA-based worker is in a trade union. The predictors that enter the model linearly are indicator of having black ethnicity, indicator of being female, indicator of being in the south region of the USA and age in years. The predictors that enter the model nonlinearly are wage in US dollars per hour and number of years of education. VMP based on the Saul–Jordan approach diverged away from a reasonable fit despite being warmed up with Jaakkola–Jordan iterations. The fits for the Jaakkola–Jordan and Knowles–Minka–Wand approaches are shown as well as those based on 1,000,000 Markov chain Monte Carlo draws. VMP based on Knowles–Minka–Wand updates is seen to provide the more accurate inference in this example.

6 Conclusion

When both numerical stability and inferential accuracy are taken into account, it is apparent from the preceding two sections that the Knowles–Minka–Wand logistic likelihood updates are superior to those based on the Jaakkola–Jordan and Saul–Jordan devices in approximate inference via VMP. Moreover, adoption of the Monahan–Stefanski numerical integration methodology means that there is no computational overhead due to using the Knowles–Minka–Wand approach. The Saul–Jordan approach is inferior in terms of accuracy and is susceptible to algorithmic breakdown, and we do not recommend it for use in VMP inference procedures.

As revealed by the simulation study of Section 4, the Jaakkola–Jordan updates have excellent stability properties. Therefore, it is recommended that the Jaakkola–Jordan updates be used in tandem with the Knowles–Minka–Wand updates—with the former providing starting values for the latter and also guarding against occasional divergence of the Knowles–Minka–Wand updates. As indicated by our numerical studies, the Knowles–Minka–Wand updates will often lead to accurate fast approximate Bayesian inference in logistic likelihood semiparametric regression models.

Acknowledgements

This research was partially supported by the Australian Research Council Discovery Project DP140100441.

References

Berndt, ER (1991), *The Practice of Econometrics*, *Addison-Wesley*, New York.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.139

The ISI's Journal for the Rapid Dissemination of Statistics Research

Bishop, CM (2006), *Pattern Recognition and Machine Learning*, *Springer*, New York.

Jaakkola, T & Jordan, MI (2000), 'Bayesian parameter estimation via variational methods', *Statistics and Computing*, **10**, 25–37.

Knowles, DA & Minka, TP (2011), 'Non-conjugate message passing for multinomial and binary regression', *Advances in Neural Information Processing Systems 24* in Shawe-Taylor, J, Zamel, RS, Bartlett, P, Pereira, F & Weinberger, KQ (eds), *MIT Press*, Cambridge, Massachusetts, 1701–1709.

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

Monahan, JF & Stefanski, LA (1989), 'Normal scale mixture approximations to $F^*(z)$ and computation of the logistic-normal integral', *Handbook of the Logistic Distribution* in Balakrishnan, N (ed), *Marcel Dekker*, New York, 529–540.

Ormerod, JT & Wand, MP (2012), 'Gaussian variational approximate inference for generalized linear mixed models', *Journal of Computational and Graphical Statistics*, **21**, 2–17.

Saul, LK & Jordan, MI (1999), 'A mean field learning algorithm for unsupervised neural networks', *Learning in Graphical Models* in Jordan, MI (ed), *MIT Press*, Cambridge, Massachusetts, 105–162.

Stan Development Team (2016), rstan: R interface to Stan. R package version 2.12. http://mc-stan.org.

Wand, MP (2014), 'Fully simplified multivariate normal updates in non-conjugate variational message passing', *Journal of Machine Learning Research*, **15**, 1351–1369.

Wand, MP (2016), 'Fast approximate inference for arbitrarily large semiparametric regression models via message passing (with discussion)', *Journal of the American Statistical Association*, **112**. in press. arXiv:1602.07412.

Wand, MP & Ormerod, JT (2008), 'On semiparametric regression with O'Sullivan penalized splines', *Australian and New Zealand Journal of Statistics*, **50**, 179–198.

Wand, MP & Ripley, BD (2015), KernSmooth: functions for kernel smoothing supporting the book 'Kernel Smoothing' by Wand and Jones (1995). R package version 2.23. http://www.r-project.org.

Supporting information

Additional supporting information may be found online in the supporting information tab for this article.