Stat

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

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.196

Variational message passing for skew t regression

Luca Maestrini[a]* **and Matt P. Wand**[b]

Received 14 June 2018; Accepted 21 June 2018

We extend recent work concerning variational approximations via message passing to accommodate approximate fitting and inference for skew t regression models. Derivation of variational message passing is challenging owing to the presence of non-standard exponential families and numerical integration being needed. Nevertheless, the factor graph fragment approach means that algorithm updates only need to be derived once for a particular response model, which can be integrated in an arbitrarily complex model. Another advantage of our work is that all skew t parameters are inferred, rather than being held fixed. Furthermore, we show that posterior dependence arising in an auxiliary variable representation of a skew t model may lead to poor performances in terms of variational message passing approximation when using simple auxiliary variable representations of the likelihood fragment and convenient factorizations of the approximating densities. © 2018 John Wiley & Sons, Ltd.

Keywords: approximate Bayesian inference; auxiliary variables; factor graph; mean field variational Bayes; skew t; variational message passing

1 Introduction

Wand (2017) introduced the notion of factor graph fragments to design a general framework for variational Bayes approximate inference with a focus on common response distributions such as those in the Bernoulli, Poisson and normal families. Such an approach is extendible to models with more elaborate likelihood structures. Nolan & Wand (2017) provide accurate algebraic and numerical details for fitting logistic likelihood regression via variational message passing (VMP). McLean & Wand (2018) consider six other likelihood families: negative binomial, t, asymmetric Laplace, skew normal, finite normal mixture and support vector machine. We add to this recent body of work and derive VMP updates for approximate fitting and inference for skew t responses. The algorithm we propose allows inference on all the skew t parameters. The VMP framework is such that arbitrarily large skew t response semiparametric regression models can be handled using factor graph fragments. Furthermore, we investigate how various auxiliary random variable representations of the likelihood impact the variational approximating results.

Section 2 provides a brief description of VMP and its implementation via the notion of message passing on a factor graph. Section 3 is dedicated to the skew t likelihood fragment and includes an algorithm for approximate inference whose updates are derived under an auxiliary variable representation of the likelihood specification. Two different product density restrictions are considered. We prove that the more convenient one has a serious pitfall. Section 4 includes a numerical study to show the implications of different factorizations on VMP performances. We conclude with an illustration. Derivational details are given in the Supporting Information.

[a]**Department of Statistical Sciences, University of Padua, Via Cesare Battisti 241, 35121 Padua, PD, Italy**

[b]**School of Mathematical and Physical Sciences, University of Technology Sydney, PO Box 123, Broadway 2007, NSW, Australia**

*__Email: luca.maestrini@phd.unipd.it__

Stat

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.196

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

2 Variational message passing on factor graphs

Variational message passing is an approach to variational Bayes approximate inference that allows modularization through the notions of *factor graphs* and *message passing*. According to a factor graph message passing approach (e.g. Minka & Winn, 2008, Appendix A), calculations only need to be performed once for a particular fragment and can be integrated with other fragments to construct inference engines for arbitrarily large models.

Consider a Bayesian statistical model with observed data $\boldsymbol{D}$ and parameter vector $\boldsymbol{\theta}$. A *mean field variational approximation* $q^*(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta} \mid \boldsymbol{D})$ is the minimizer of the Kullback–Leibler divergence

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

subject to a 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 $\boldsymbol{\theta}$. It can be shown that the optimal q-density functions satisfy

$$q^*(\boldsymbol{\theta}_i) \propto \exp\left\{E_{q(\boldsymbol{\theta}\setminus\boldsymbol{\theta}_i)} \log p(\boldsymbol{\theta}_i \mid \boldsymbol{D}, \boldsymbol{\theta}\setminus\boldsymbol{\theta}_i)\right\}, \quad 1 \le i \le M, \tag{1}$$

where $\boldsymbol{\theta}\setminus\boldsymbol{\theta}_i$ denotes the entries of $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_i$ omitted. The previous expression gives rise to an iterative scheme for obtaining the parameters of the optimal density functions $q^*(\boldsymbol{\theta}_i)$, which is known as *mean field variational Bayes*. A listing of such an algorithm is provided, for instance, in Ormerod & Wand (2010).

Rather than using result (1), the VMP procedure is founded upon the notion of *messages* passed between any two neighbouring nodes, which is a particular function of the stochastic node that either sends or receives the message. We follow the approach of Minka (2005) among the several variants of VMP in the literature, which is described in Section 2.5 of Wand (2017). The approach is here briefly summarized.

Let N be 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

$$m_{\boldsymbol{\theta}_i \to f_j}(\boldsymbol{\theta}_i) \longleftarrow\!\propto \prod_{j' \ne j : i \in \text{neighbours}(j')} m_{f_{j'} \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \tag{2}$$

and the *factor to stochastic node* message updates have the form

$$m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \longleftarrow\!\propto \exp\left[E_{f_j \to \boldsymbol{\theta}_i}\left\{\log f_j\left(\boldsymbol{\theta}_{\text{neighbours}(j)}\right)\right\}\right], \tag{3}$$

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

$$\frac{\prod_{i' \in \text{neighbours}(j)\setminus\{i\}} m_{f_j \to \boldsymbol{\theta}_{i'}}(\boldsymbol{\theta}_{i'})\, m_{\boldsymbol{\theta}_{i'} \to f_j}(\boldsymbol{\theta}_{i'})}{\prod_{i' \in \text{neighbours}(j)\setminus\{i\}} \int m_{f_j \to \boldsymbol{\theta}_{i'}}(\boldsymbol{\theta}_{i'})\, m_{\boldsymbol{\theta}_{i'} \to f_j}(\boldsymbol{\theta}_{i'})\, \mathrm{d}\boldsymbol{\theta}_{i'}}. \tag{4}$$

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 neighbours$(j) \setminus \{i\} = \emptyset$, then the expectation in (3) can be dropped and the right-hand side of (3) is proportional to $f_j\left(\boldsymbol{\theta}_{\text{neighbours}(j)}\right)$. The optimal q-densities are then obtained from

$$q^*(\boldsymbol{\theta}_i) \propto \prod_{j : i \in \text{neighbours}(j)} m^*_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i), \tag{5}$$

where $m^*_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)$ are the optimal messages at convergence.

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.196

2.1 Factor graph

Variational message passing arrives at variational Bayes approximation via message passing on an appropriate factor graph. A *factor graph* is a graphical representation of the argument groupings of a real-valued function. Consider, for example, the regression model

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2 \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}\right), \quad \boldsymbol{\beta} \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right), \quad \sigma^2 \mid a \sim \text{Inverse-}\chi^2\left(1, 1/a\right), \quad a \sim \text{Inverse-}\chi^2\left(1, 1/A^2\right), \tag{6}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of response data and $\boldsymbol{X}$ is an $n \times d$ design matrix. The $d \times 1$ vector $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, the $d \times d$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ and $A > 0$ are user-specified hyperparameters. A factor graph representation of this model based on the product density restriction $q\left(\boldsymbol{\beta}\right) q\left(\sigma^2\right) q\left(a\right)$ is that of Figure 1.

Each factor graph has a corresponding graphical representation based on nodes connected by edges. The word *node* is used for both a stochastic node $\boldsymbol{\theta}_i$, $1 \leq i \leq M$, and a factor f_j, $1 \leq j \leq N$. In detail, the shaded squares correspond to *factors*, which are single product components of the real-valued function. The unshaded circles are called *stochastic nodes* and refer to parameters expressing product dependencies in the approximating density. An *edge* connects a factor to the stochastic nodes included in that factor. Two nodes are neighbours of each other if they are joined by an edge. In Figure 1, stochastic nodes $\boldsymbol{\beta}$ and σ^2 are *neighbours* of the factor $p\left(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2\right)$, for instance. We denote by neighbours (j) the θ_i indices connected to the jth factor.

2.2 Notation

Before introducing VMP for skew t response models, we define some relevant notations.

For a vector $\boldsymbol{a}$ and scalar function s, we let $s\left(\boldsymbol{a}\right)$ 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 size. If $\boldsymbol{A}$ is a $d \times d$ matrix, then $\text{vec}\left(\boldsymbol{A}\right)$ 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 $\text{vec}^{-1}\left(\boldsymbol{a}\right)$ is the $d \times d$ matrix formed from listing the entries of $\boldsymbol{a}$ in a column-wise fashion in order from left to right. We denote by $\text{diag}\left(\boldsymbol{a}\right)$ the diagonal matrix containing the entries of $\boldsymbol{a}$ along the main diagonal. The vector $\boldsymbol{e}_i$ denotes a vector of zeroes with a 1 in the ith position. For a $d \times 1$ vector $\boldsymbol{v}_1$ and a $d^2 \times 1$ vector $\boldsymbol{v}_2$ such that $\text{vec}^{-1}\left(\boldsymbol{v}_2\right)$ is symmetric and given a $d \times d$ matrix $\boldsymbol{Q}$, a $d \times 1$ vector $\boldsymbol{r}$ and $s \in \mathbb{R}$, we define

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

The G_{VMP} function originates from the fact that $E_{\boldsymbol{\theta}}\left\{-\frac{1}{2}\left(\boldsymbol{\theta}^T\boldsymbol{Q}\boldsymbol{\theta} - 2\boldsymbol{r}^T\boldsymbol{\theta} + s\right)\right\} = G_{\text{VMP}}\left(\boldsymbol{\eta}; \boldsymbol{Q}, \boldsymbol{r}, s\right)$ when $\boldsymbol{\theta}$ is a $d \times 1$ multivariate normal random vector with natural parameter vector $\boldsymbol{\eta}$. Furthermore, we define

$$\boldsymbol{\eta}_{f \longleftrightarrow \boldsymbol{\theta}} \equiv \boldsymbol{\eta}_{f \to \boldsymbol{\theta}} + \boldsymbol{\eta}_{\boldsymbol{\theta} \to f}$$

for any natural parameter $\boldsymbol{\eta}$, factor f, and stochastic node $\boldsymbol{\theta}$. We introduce the notations $(E\boldsymbol{T})_2^{\text{ISRN}}$, $(E\boldsymbol{T})_3^{\text{ISRN}}$, $(E\boldsymbol{T})_2^{\text{SS}}$, $(E\boldsymbol{T})_3^{\text{SS}}$ and $(E\boldsymbol{T})_2^{\text{MR}}$ referring to the expected value of the sufficient statistic of particular exponential families that are

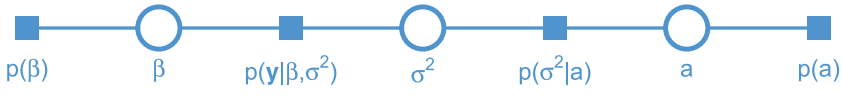

Figure 1. Factor graph for the regression model in (6) and restriction $q\left(\boldsymbol{\beta}\right) q\left(\sigma^2\right) q\left(a\right)$.

Stat

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.196

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

defined in the Supporting Information, Section S2, of McLean & Wand (2018): the inverse square root Nadarajah, sea sponge and moon rock distributions.

3 The skew t likelihood fragment

The skew t likelihood fragment corresponds to the likelihood specification

$$y_i \,|\, \boldsymbol{\theta}, \sigma^2, \lambda, \nu \overset{\text{ind.}}{\sim} \text{Skew-}t\left((\boldsymbol{A}\boldsymbol{\theta})_i, \sigma^2, \lambda, \nu\right), \quad 1 \le i \le n, \tag{7}$$

where $\boldsymbol{A}$ is a generic design matrix, $\boldsymbol{\theta}$ is a generic vector of coefficients, $\sigma^2 > 0$, $\lambda \in \mathbb{R}$ and $\nu > 0$. Among the possible definitions of the skew t distribution, we consider the one described in Azzalini & Capitanio (2003). Their skew t distribution becomes a symmetric Student's t distribution when $\lambda = 0$, a conditional normal distribution as $\nu \to \infty$ and allows the inclusion of left-tailed or negative skewness when $\lambda < 0$ and right-tailed or positive skewness when $\lambda > 0$.

The response likelihood can be conveniently re-expressed in terms of auxiliary variables and more common distributions to aid the construction of a tractable VMP algorithm. The introduction of auxiliary variables has the practical advantage of reducing the complexity of message updates, which either can be expressed in a closed form or require only univariate numerical integration. If we introduce two auxiliary random variables a_{1i} and a_{2i}, $1 \le i \le n$, such that

$$a_{1i} \overset{\text{ind.}}{\sim} N(0,1) \quad \text{and} \quad a_{2i} \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2(\nu, \nu),$$

then, with standard distribution theoretical manipulations, the model in (7) can be alternatively written as

$$y_i \,|\, \boldsymbol{\theta}, \sigma^2, \lambda, a_{1i}, a_{2i} \overset{\text{ind.}}{\sim} N\left((\boldsymbol{A}\boldsymbol{\theta})_i + \frac{\sigma\lambda\,|a_{1i}|\,\sqrt{a_{2i}}}{\sqrt{1+\lambda^2}}, \frac{a_{2i}\sigma^2}{1+\lambda^2}\right), \quad a_{1i} \overset{\text{ind.}}{\sim} N(0,1), \quad a_{2i} \,|\, \nu \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2(\nu, \nu). \tag{8}$$

Considering this last model specification, we provide fragment updates that allow for the skew t distribution to be handled within the VMP framework. An assumption on the optimal q-density product restriction is required to produce the VMP solution in (5). An assumption producing one of the simplest VMP schemes is

$$q\left(\boldsymbol{\theta}, \sigma^2, \lambda, \nu, \boldsymbol{a}_1, \boldsymbol{a}_2\right) = q(\boldsymbol{\theta})\, q\left(\sigma^2\right) q(\lambda)\, q(\nu)\, q(\boldsymbol{a}_1)\, q(\boldsymbol{a}_2) = q(\boldsymbol{\theta})\, q\left(\sigma^2\right) q(\lambda)\, q(\nu) \prod_{i=1}^{n} q(a_{1i})\, q(a_{2i}). \tag{9}$$

Combining this product density restriction with the likelihood model in (8), we obtain the factor graph representation in Figure 2, left panel. The structure of messages from the likelihood factor to each node is obtained by manipulation of the log-likelihood factor as a function of the node of interest, according to the VMP Eqs. 2–4.

The messages passed from the likelihood factor to $\boldsymbol{\theta}$ take the form

$$m_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}(\boldsymbol{\theta}) = \exp\left\{\begin{bmatrix}\boldsymbol{\theta}\\ \text{vec}\left(\boldsymbol{\theta}\boldsymbol{\theta}^T\right)\end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}\right\},$$

which a has multivariate normal structure. Therefore, to ensure conjugacy, messages that $\boldsymbol{\theta}$ receives from factors outside of the skew t likelihood fragment, such as a prior on $\boldsymbol{\theta}$, have to be proportional to a multivariate normal density.

The messages passed from the likelihood factor to σ^2 have the form

$$m_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}\left(\sigma^2\right) = \exp\left\{\begin{bmatrix}\log\left(\sigma^2\right)\\ 1/\sigma\\ 1/\sigma^2\end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}\right\},$$

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

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.196

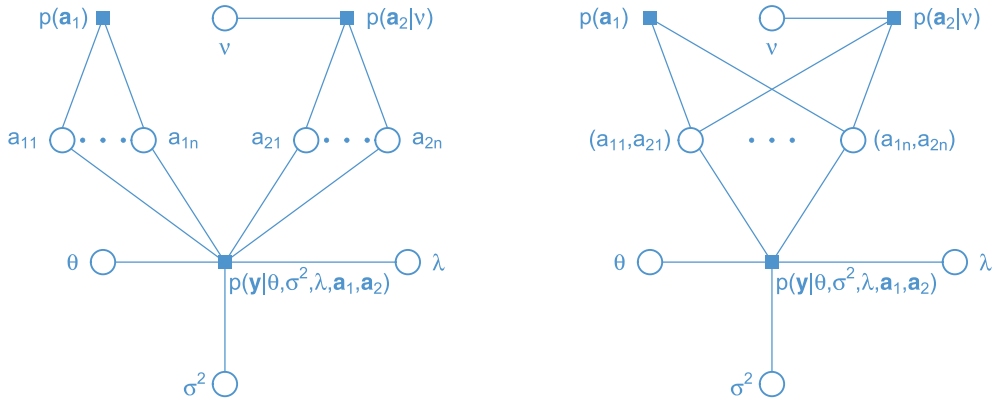

Figure 2. Factor graph for skew t likelihood specification in (8) under the assumption in (9) (left panel) and (11) (right panel) with independent $N(0,1)$ auxiliary variables $a_{11},\ldots,a_{1n}$ and independent Inverse-$\chi^2(\nu,\nu)$ auxiliary variables $a_{21},\ldots,a_{2n}$.

which is within the *inverse square root Nadarajah* family described in Section S.2.3 of McLean & Wand (2018). The imposition of conjugacy means that we assume that all messages that passed to σ^2 from factors outside of the skew normal likelihood fragment are also proportional to inverse square root Nadarajah density functions. For instance, an Inverse-χ^2 prior on σ^2 is suitable to ensure conjugacy.

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

$$m_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}(\lambda)=\exp\left\{\begin{bmatrix}\log\left(1+\lambda^2\right)\\ \lambda^2\\ \lambda\sqrt{1+\lambda^2}\end{bmatrix}^T\boldsymbol{\eta}_{p(\boldsymbol{y}\,|\,\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}\right\},$$

which is within the *Sea Sponge* family identified in McLean & Wand (2018), Section S.2.5. We assume that each of the messages that λ receives from factors outside of this fragment are conjugate to Sea Sponge density functions. If, for instance, the only factor that sends a message to λ is the prior density function $p(\lambda)$, then $m_{p(\lambda)\to\lambda}(\lambda)=p(\lambda)$ and, under conjugacy, $p(\lambda)$ must be of the form

$$p(\lambda)\propto\exp\left\{\begin{bmatrix}\log\left(1+\lambda^2\right)\\ \lambda^2\\ \lambda\sqrt{1+\lambda^2}\end{bmatrix}^T\boldsymbol{\eta}_\lambda\right\}\tag{10}$$

for some 3×1 vector $\boldsymbol{\eta}_\lambda$. A special case of (10) is priors of the form $\lambda\sim N\left(0,\sigma_\lambda^2\right)$, having $\boldsymbol{\eta}_\lambda=\left[0,-1/\left(2\sigma_\lambda^2\right),0\right]$.

As a function of ν, we have

$$\log p(\boldsymbol{a}_2\,|\,\nu)=\begin{bmatrix}(\nu/2)\log(\nu/2)-\log\{\Gamma(\nu/2)\}\\ (\nu/2)\end{bmatrix}^T\begin{bmatrix}n\\ -\mathbf{1}_n^T\{\log(\boldsymbol{a}_2)+\mathbf{1}_n/\boldsymbol{a}_2\}\end{bmatrix}+\text{const},$$

indicating that messages from $p(\boldsymbol{a}_2\,|\,\nu)$ to a_{2i}, $1\le i\le n$, are within the *Moon Rock* family defined in S.2.4 of McLean & Wand (2018). We assume that messages passed to ν from factors outside the skew t likelihood fragments are conjugate with the Moon Rock family. For example, if the only other factor passing messages to ν is its prior density function $p(\nu)$, then we require that $p(\nu)$ is a Moon Rock density function or conjugate with one, such as an exponential density function.

Stat

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.196

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

The structures of these messages serve as a base to build a VMP algorithm on assumption (9). A listing of the algorithm and the derivation of its updates are given in the Supporting Information. However, the implementation of such an algorithm in simulation studies reveals poor performances of VMP if roughly compared with the posterior densities of single parameters obtainable via Markov chain Monte Carlo (MCMC). The cause of this discrepancy is the strong posterior dependence between the two auxiliary variables a_{1i} and a_{2i} whereas the product density factorization in (9) ignores such a dependence. Figure 3 provides some insight into why VMP developed according to assumption (9) is prone to inaccuracy, showing pairwise scatterplots of series $|a_1|$ and $1/\sqrt{a_2}$ from an MCMC fitting output of a skew t random sample of size $n = 1000$ as in the figure description. MCMC draws are obtained through `rstan` with the `R` computing environment (R Core Team, 2018) interfacing via the `rstan` package (Stan Development Team, 2018). Expectations of these series involving the auxiliary random variables appear when deriving message updates. It is apparent that the posterior correlation between the auxiliary variables increases as the value of λ increases.

The following theoretical result confirms this posterior correlation problem.

Theorem 1

Consider random variables satisfying

$$x \,|\, a_1, a_2 \sim N\left(\mu_0 + \frac{\sigma_0\lambda_0 \,|a_1|\, \sqrt{a_2}}{\sqrt{1+\lambda_0^2}}, \frac{a_2\sigma_0^2}{1+\lambda_0^2}\right), \quad \textit{where} \quad a_1 \sim N(0,1) \quad \textit{and} \quad a_2 \sim \textit{Inverse-}\chi^2(\nu_0, \nu_0),$$

with $\mu_0, \lambda_0 \in \mathbb{R}$ *and* $\sigma_0, \nu_0 > 0$. *Then for any* $x_0 \in \mathbb{R}$ *and* μ_0, σ_0, ν_0

$$\lim_{|\lambda_0|\to\infty} \textit{Corr}\left(|a_1|, 1/\sqrt{a_2} \,|\, x = x_0\right) = 1.$$

A proof is given in the Supporting Information. As the q densities are assumed to approximate the posterior density structure, Theorem 1 suggests a modification on our previous assumption to a less simplistic product density restriction. At the cost of further algebra, we propose the replacement of the assumption in (9) with

$$q\left(\boldsymbol{\theta}, \sigma^2, \lambda, \nu, \boldsymbol{a}_1, \boldsymbol{a}_2\right) = q(\boldsymbol{\theta})\, q\left(\sigma^2\right) q(\lambda)\, q(\nu)\, q(\boldsymbol{a}_1, \boldsymbol{a}_2) = q(\boldsymbol{\theta})\, q\left(\sigma^2\right) q(\lambda)\, q(\nu) \prod_{i=1}^{n} q(a_{1i}, a_{2i}). \tag{11}$$

This gives rise to the factor graph representation in Figure 2, right panel, and Algorithm 1, whose output at convergence provides the optimal approximating densities according to (5), without alteration of previous message

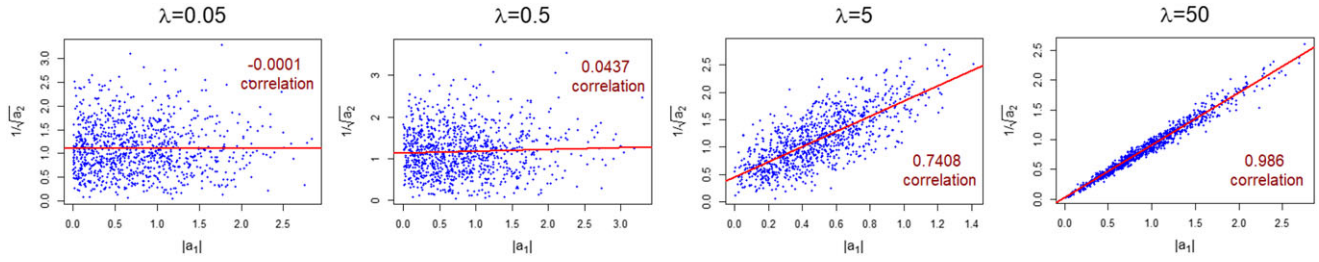

Figure 3. Markov chain Monte Carlo samples ($n = 1000$) drawn via `rstan` from the distribution $\left\{|a_1|, 1/\sqrt{a_2} \,|\, \text{rest}\right\}$ for a skew t random sample with $\boldsymbol{\theta} = \mu = 0$, $\sigma = 1$, $\nu = 1.5$ and $\lambda = (0.05, 0.5, 5, 50)$, using the hyperparameters specified in Section 4. Sample correlations are also shown.

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.196

Algorithm 1 The inputs, updates and outputs of the skew t likelihood fragment assuming $q\left(\boldsymbol{\theta},\sigma^2,\lambda,\nu,\boldsymbol{a}_1,\boldsymbol{a}_2\right) = q\left(\boldsymbol{\theta}\right)q\left(\sigma^2\right)q\left(\lambda\right)q\left(\nu\right)\prod_{i=1}^{n}q\left(a_{1i},a_{2i}\right)$.

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

Parameter Inputs: $\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}$, $\boldsymbol{\eta}_{\boldsymbol{\theta}\to p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}$, $\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}$, $\boldsymbol{\eta}_{\sigma^2\to p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}$, $\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}$, $\boldsymbol{\eta}_{\lambda\to p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)}$, $\boldsymbol{\eta}_{p(\boldsymbol{a}_2\mid\nu)\to\nu}$, $\boldsymbol{\eta}_{\nu\to p(\boldsymbol{a}_2\mid\nu)}$.

Updates:

$$\mu_{q(1/\sigma)} \longleftarrow (E\boldsymbol{T})_2^{\mathrm{ISRN}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\sigma^2}\right)$$

$$\mu_{q(1/\sigma^2)} \longleftarrow (E\boldsymbol{T})_3^{\mathrm{ISRN}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\sigma^2}\right)$$

$$\mu_{q(\lambda^2)} \longleftarrow (E\boldsymbol{T})_2^{\mathrm{SS}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\lambda}\right)$$

$$\mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)} \longleftarrow (E\boldsymbol{T})_3^{\mathrm{SS}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\lambda}\right)$$

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

$$\boldsymbol{\tau}_1 \longleftarrow \boldsymbol{y} + \frac{1}{2}\boldsymbol{A}\left\{\mathrm{vec}^{-1}\left(\left(\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}}\right)_2\right)\right\}^{-1}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}}\right)_1$$

$$\boldsymbol{\tau}_2 \longleftarrow \left[G_{\mathrm{VMP}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}};\boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{A},\boldsymbol{A}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{y},y_i^2\right)\right]_{1\le i\le n}$$

$$\boldsymbol{\eta}_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)} \longleftarrow \begin{bmatrix} -\frac{1}{2}\left(1+\mu_{q(\lambda^2)}\right) \\ \mu_{q\left(\lambda\sqrt{1+\lambda}\right)}\mu_{q(1/\sigma)}\boldsymbol{\tau}_1 \\ \left(1+\mu_{q(\lambda^2)}\right)\mu_{q(1/\sigma^2)}\boldsymbol{\tau}_2 - \frac{1}{2}\mu_{q(\nu)} \\ -\frac{1}{2}\left(3+\mu_{q(\nu)}\right) \end{bmatrix}$$

$$E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{a}_1^2\right) \longleftarrow (E\boldsymbol{T})_1^{\mathrm{MW}}\left(\boldsymbol{\eta}_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\right)$$

$$E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\left|\boldsymbol{a}_1\right|/\sqrt{\boldsymbol{a}_2}\right) \longleftarrow (E\boldsymbol{T})_2^{\mathrm{MW}}\left(\boldsymbol{\eta}_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\right)$$

$$E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right) \longleftarrow (E\boldsymbol{T})_3^{\mathrm{MW}}\left(\boldsymbol{\eta}_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\right)$$

$$E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left\{\log\left(\boldsymbol{a}_2\right)\right\} \longleftarrow (E\boldsymbol{T})_4^{\mathrm{MW}}\left(\boldsymbol{\eta}_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\right)$$

$$\boldsymbol{\tau}_3 \longleftarrow G_{\mathrm{VMP}}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\leftrightarrow\boldsymbol{\theta}};\boldsymbol{A}^T\mathrm{diag}\left\{E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{A},\boldsymbol{A}^T\mathrm{diag}\left\{E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y},\right.$$
$$\left.\boldsymbol{y}^T\mathrm{diag}\left\{E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y}\right)$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}} \longleftarrow \left(1+\mu_{q(\lambda^2)}\right)\mu_{q(1/\sigma^2)}\begin{bmatrix}\boldsymbol{A}^T\mathrm{diag}\left\{E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{y} \\ -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{A}^T\mathrm{diag}\left\{E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{1}_n/\boldsymbol{a}_2\right)\right\}\boldsymbol{A}\right)\end{bmatrix}$$
$$-\mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\mu_{q(1/\sigma)}\begin{bmatrix}\boldsymbol{A}^T E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\left|\boldsymbol{a}_1\right|/\sqrt{\boldsymbol{a}_2}\right) \\ 0\end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2} \longleftarrow \begin{bmatrix} -n/2 \\ \mu_{q\left(\lambda\sqrt{1+\lambda^2}\right)}\boldsymbol{\tau}_1^T E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\left|\boldsymbol{a}_1\right|/\sqrt{\boldsymbol{a}_2}\right) \\ \left(1+\mu_{q(\lambda^2)}\right)\boldsymbol{\tau}_3 \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda} \longleftarrow \begin{bmatrix} n/2 \\ \mu_{q(1/\sigma^2)}\boldsymbol{\tau}_3 - \frac{1}{2}\boldsymbol{1}_n^T E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\boldsymbol{a}_1^2\right) \\ \mu_{q(1/\sigma)}\boldsymbol{\tau}_1^T E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left(\left|\boldsymbol{a}_1\right|/\sqrt{\boldsymbol{a}_2}\right) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{a}_2\mid\nu)\to\nu} \longleftarrow \begin{bmatrix} n \\ -\boldsymbol{1}_n^T E_{q(\boldsymbol{a}_1,\boldsymbol{a}_2)}\left\{\log\left(\boldsymbol{a}_2\right)+\boldsymbol{1}_n/\boldsymbol{a}_2\right\}\end{bmatrix}.$$

Parameter Outputs: $\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\boldsymbol{\theta}}$, $\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\sigma^2}$, $\boldsymbol{\eta}_{p(\boldsymbol{y}\mid\boldsymbol{\theta},\sigma^2,\lambda,\boldsymbol{a}_1,\boldsymbol{a}_2)\to\lambda}$, $\boldsymbol{\eta}_{p(\boldsymbol{a}_2\mid\nu)\to\nu}$.

Stat

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.196

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

structures. Further details about derivations and notation $(E\boldsymbol{T})_j^{\text{MW}}$, $j = 1, \ldots, 4$, are displayed in the Supporting Information. As explained in Section S.3 of the Supporting Information, under assumption (11), the moments with respect to $q^*(a_{1i}, a_{2i})$ are expressible in a closed form. However, further numerical integration may be required when the arguments of the Gaussian hypergeometric functions appearing in moment expressions are close to 1.

4 Simulation study

We propose a simulation study to compare the performances of the VMP algorithm designed around the assumption in (9) and Algorithm 1, which is based on the assumption in (11). We generated 100 datasets of size $n = 500$ setting two regression parameters to be $\theta_0 = 1$ and $\theta_1 = 2$, scale parameter $\sigma = 1$ and shape parameters $\lambda = 5$ and $\nu = 1.5$. The hyperparameters for $\boldsymbol{\theta}$ were fixed to $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = 10^{10}\boldsymbol{I}$ over a prior $N(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$. We used an Inverse-$\chi^2(A, B)$ prior on the squared scale with $A = B = 0.01$. The prior for the parameter of symmetry λ was assumed to be $N(\mu_\lambda, \sigma_\lambda^2)$ with $\mu_\lambda = 0$ and $\sigma_\lambda^2 = 10^{10}$ and that for the degrees of freedom ν to be a Gamma(α_ν, β_ν) with $\alpha_\nu = 1$ and $\beta_\nu = 0.01$.

Let ξ be a generic parameter. The accuracy of each VMP approximation $q^*(\xi)$ as from (5) can be assessed using the L_1 error, or *integrated absolute error (IAE)* of q^*, given by

$$\text{IAE}(q^*) = \int_{-\infty}^{\infty} |q^*(\xi) - p(\xi \mid \boldsymbol{D})| \, \mathrm{d}\xi.$$

As pointed out in Wand et al. (2011), the L_1 error is a scale independent number that is invariant to a monotone transformation on the parameter ξ. This implies, for instance, that the IAE values for $q^*(\sigma)$ and $q^*(\sigma^2)$ coincide. Note that the L_1 error is a number between 0 and 2. To express this measure as a percentage, we can then define the accuracy as

$$\text{accuracy}(q^*) = 1 - \left\{ \text{IAE}(q^*) \Big/ \sup_{q \text{ a density}} \text{IAE}(q) \right\} = 1 - \frac{1}{2}\text{IAE}(q^*),$$

so that $0 \leq \text{accuracy}(q^*) \leq 1$, with 1 reflecting perfect correspondence between VMP approximations and posterior densities. The computation of $p(\xi \mid \boldsymbol{D})$ is complex, so we worked with MCMC samples obtained using `rstan`. MCMC samples of size 10,000 were generated setting a burn-in of 5000 values and thinning the remaining 5000 by a factor of 5. Table I includes the accuracy values from the simulation study. As expected, the algorithm based on assumption (11) is seen to provide more accurate inference. However, accuracy and percentage of coverage of σ^2 and λ are particularly low for both the algorithms. This might suggest the application of less generic product density restriction

Table I. Average (standard deviation) accuracy and percentage coverage from the simulation study. "VMP 1" and "VMP 2" refer to the variational message passing algorithms constructed according to assumptions in (9) and (11), respectively.

	Accuracy			
Parameter	VMP 1		VMP 2	
β_0	0.0	(0.0)	37.7	(6.2)
β_1	51.2	(17.7)	57.1	(4.5)
σ^2	0.0	(0.0)	11.0	(3.5)
λ	0.0	(0.0)	10.0	(3.4)
ν	0.0	(0.0)	56.6	(6.1)

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.196

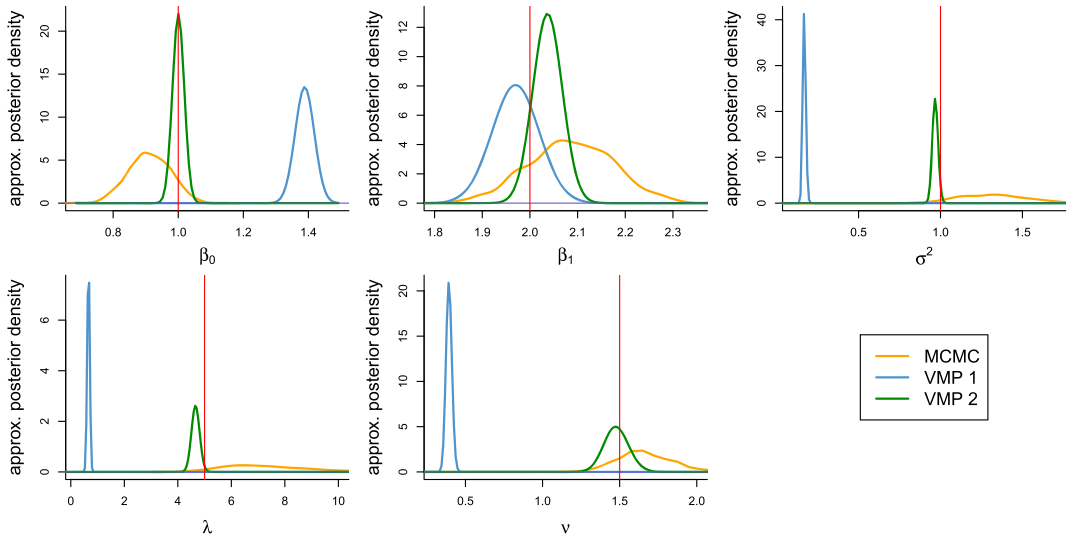

Figure 4. VMP-approximate and MCMC posterior density functions from a single dataset of the simulation study. "VMP 1" and "VMP 2" respectively refer to the VMP algorithm constructed around the assumption in (9) and Algorithm 1. VMP, variational message passing; MCMC, Markov chain Monte Carlo.

to take into account other possible posterior dependence among variables. Nonetheless, this choice would imply more involved message update derivations and further numerical integration. Figure 4 permits visualization of these results with the plot of approximate and MCMC posterior densities from a single simulation. The density curves produced by Algorithm 1 are sensibly closer to the modes of MCMC posterior densities than are those from the other VMP algorithm. Note also the lower variance of variational approximating densities, which corresponds with the theoretical results in Wang & Blei (2018) concerning variance underestimation of variational Bayes.

5 Application

Here, we provide an application that illustrates how the derivations in Section 3 can be integrated to perform variational inference on extensions of the skew t likelihood fragment without deriving a VMP scheme from scratch. We consider the dataset `Workinghours` from the `R` package `Ecdat` (Croissant, 2016), which contains a cross-section study of 3382 observations. The response variable is *income* divided by a factor of 10 (the other household income in thousands of dollars) versus the variable *age* (age of the wife). The pairs of predictors and responses (x_i, y_i), $1 \leq i \leq n$, are analysed via non-parametric regression and the following penalized spline model in a Bayesian mixed model form:

$$y_i \,|\, f, \sigma_\varepsilon^2, \lambda, \nu \overset{\text{ind.}}{\sim} \text{Skew-}t\left(f\left(x_i\right), \sigma_\varepsilon^2, \lambda, \nu\right),$$

with the function f structured as $f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k z_k(x)$ with $u_k \,|\, \sigma_u^2 \overset{\text{ind.}}{\sim} N\left(0, \sigma_u^2\right)$ and $\{z_k : 1 \leq k \leq K\}$ suitable spline basis. The full model with auxiliary variable representation is

$$\begin{aligned}
&y_i \,|\, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2, \lambda, a_{1i}, a_{2i} \overset{\text{ind.}}{\sim} N\left(\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}\right)_i + \frac{\sigma_\varepsilon \lambda \left|a_{1i}\right| \sqrt{a_{2i}}}{\sqrt{1+\lambda^2}}, \frac{a_{2i}\sigma_\varepsilon^2}{1+\lambda^2}\right), \quad a_{1i} \overset{\text{ind.}}{\sim} N(0,1), \quad a_{2i} \,|\, \nu \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2(\nu, \nu), \\
&\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \Bigg| \, \sigma_u^2 \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_u^2 \boldsymbol{I} \end{bmatrix}\right), \quad \sigma_u^2 \sim \text{Inverse-}\chi^2\left(A_{\sigma_u^2}, B_{\sigma_u^2}\right), \quad \sigma_\varepsilon^2 \sim \text{Inverse-}\chi^2\left(A_{\sigma_\varepsilon^2}, B_{\sigma_\varepsilon^2}\right), \\
&\lambda \sim N\left(\mu_\lambda, \sigma_\lambda^2\right), \quad \nu \sim \text{Gamma}\left(\alpha_\nu, \beta_\nu\right),
\end{aligned} \tag{12}$$

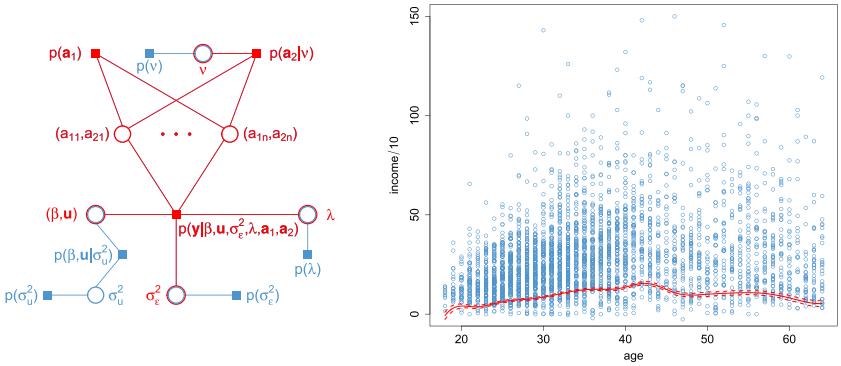

Figure 5. Study of data from `Workinghours` dataset. Left panel: factor graph corresponding to the model in (12) under the product density restriction in (13). Right panel: approximate posterior mean and pointwise 95% credible sets obtained via variational message passing, integrating Algorithm 1; 20 observations whose "income/10" value exceeds 150 have been excluded from the plot.

where

$$\boldsymbol{X} \equiv \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{Z} \equiv \begin{bmatrix} z_1(x_1) & \ldots & z_K(x_1) \\ \vdots & \ddots & \vdots \\ z_1(x_n) & \ldots & z_K(x_n) \end{bmatrix}.$$

The 2×1 vector $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, 2×2 symmetric positive definite matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$; positive numbers $A_{\sigma_u^2}$, $B_{\sigma_u^2}$, $A_{\sigma_\varepsilon^2}$, $B_{\sigma_\varepsilon^2}$, σ_λ^2, α_ν and β_ν; and number μ_λ are user-specified hyperparameters. We adopt canonical cubic O'Sullivan splines (Wand & Ormerod, 2008) with $K = 14$, relying on a common rule of thumb in the penalized spline literature for which the number of interior knots can be chosen as $\min(n_U/4, 35)$, where n_U is the number of unique x_is (e.g. Ruppert et al., 2003).

Assuming that the joint posterior density approximation admits the product density approximation

$$p\left(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2, \sigma_u^2, \lambda, \nu | \boldsymbol{y}\right) \approx q\left(\boldsymbol{\beta}, \boldsymbol{u}\right) q\left(\sigma_\varepsilon^2\right) q\left(\sigma_u^2\right) q\left(\lambda\right) q\left(\nu\right) \prod_{i=1}^{n} q\left(a_{1i}, a_{2i}\right), \tag{13}$$

Algorithm 1 can be integrated with updates involving the blue nodes in the factor graph in Figure 5 to fit the regression model (12) via VMP. The estimated non-parametric regression function and corresponding pointwise 95% credible set are shown in the right panel of Figure 5. The results show higher mean income of the other household for average-age wives, which tends to decrease more remarkably around the age of 43 and 57.

6 Conclusions

Variational message passing offers a flexible framework to standardize variational Bayes algorithm derivations and numerical integration steps. Motivated by the desire to have fast approximate inference methods for additional notable likelihood models, we have developed a VMP algorithm for fitting and inference for skew t regression models. As indicated by the simulation study in Section 4, the performance of variational Bayes is not always satisfactory, especially in presence of a high number of parameters and strong correlation among model parameters. The VMP algorithm we propose is designed around a choice of the mean field restriction, which is a compromise among algebraic complexity, feasibility and quality of the approximation. As revealed by the application on a real dataset, VMP allows one to integrate several fragments and compose algorithms for more complex models without significant additional effort.

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

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.196

Acknowledgements

The work of Luca Maestrini was carried out during a visiting period at the School of Mathematical and Physical Sciences, University of Technology Sydney, Australia. We thank Alessandra Salvan and Nicola Sartori for their assistance with this research.

References

Azzalini, A & Capitanio, A (2003), 'Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**, 367–389.

Croissant, Y (2016), Ecdat: Data sets for econometrics. R package version 0.3.1. https://CRAN.R-project.org/package=Ecdat.

McLean, MW & Wand, MP (2018), 'Variational message passing for elaborate response regression models', *Bayesian Analysis*, **13**, 1–28.

Minka, T (2005), 'Divergence measures and message passing', *Microsoft Research Technical Report Series*, **173**, 1–17.

Minka, T & Winn, J (2008), 'Gates: A graphical notation for mixture models', *Microsoft Research Technical Report Series*, **185**, 1–16.

Nolan, TH & Wand, MP (2017), 'Accurate logistic variational message passing: Algebraic and numerical details', *Stat*, **6**, 102–112.

Ormerod, JT & Wand, MP (2010), 'Explaining variational approximations', *The American Statistician*, **64**, 140–153.

R Core Team (2018), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ruppert, D, Wand, MP & Carroll, RJ (2003), *Semiparametric Regression*, *Cambridge University Press*, New York.

Stan Development Team (2018), rstan: the R interface to Stan. R package version 2.17.3. http://mc-stan.org/.

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

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

Wand, MP, Ormerod, JT, Padoan, SA & Frühwirth, RF (2011), 'Mean field variational Bayes for elaborate distributions', *Bayesian Analysis*, **6**, 847–900.

Wang, Y & Blei, DM (2018), 'Frequentist consistency of variational Bayes', *arXiv preprint*. arXiv:1705.03439v2.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article

