Variational message passing for skew t regression

Luca Maestrini[1], Matt P. Wand[2]

[1] Department of Statistical Sciences, University of Padova, Italy
[2] School of Mathematical and Physical Sciences, University of Technology Sydney, Australia

E-mail for correspondence: luca.maestrini@phd.unipd.it

Abstract: We extend the class of variational message passing algorithms to approximate fitting and inference for skew t regression models, building on recent work concerning variational message passing on factor graph. A major advantage of a factor graph fragment approach is that calculations only need to be done once for the considered distribution family and can be easily adapted to accommodate more complex model structures. A simulation study shows how posterior dependence arising in auxiliary variable representation of a skew t response model may lead to poor performances in terms of variational message passing when using convenient factorizations of the approximating densities.

Keywords: Auxiliary variables; Factor graph fragment; Skew t; Variational approximation; Variational message passing.

1 Introduction

The notion of factor graph fragment introduced in Wand (2017) seminal paper represents a step forward towards the integration of variational approximations in mainstream statistics. This perspective provides a valid instrument to streamline algebra and computer coding when implementing variational message passing (VMP) for elaborate response regression models. A notable practical advantage is that algorithm updates only need to be derived once for a particular response model, which can be integrated in more complex structures such as those including, for instance, semiparametric components.
As shown in McLean and Wand (2018), modularity of variational message passing extends beyond the common distributions in the Bernoulli, Poisson

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

and Normal families. We widen this recent body of work to include skew t regression models. Attention is reserved to an auxiliary variable representation of the model. Auxiliary variables can reduce the algebraic complexity of algorithm derivations but at the same time affect variational message passing performances, according to the assumptions on the approximating density.

2 Variational message passing

Consider a Bayesian statistical model with observed data $\boldsymbol{D}$ and parameter vector $\boldsymbol{\theta}$. Variational approximations are conceived to treat models in which the posterior density function $p(\boldsymbol{\theta}|\boldsymbol{D})$ is analytically intractable.
A mean field variational approximation $q^*(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|\boldsymbol{D})$ is the minimizer of the Kullback-Leibler divergence

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

subject to 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

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

where $\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i$ 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. VMP arrives at the same approximation via message passing on an appropriate factor graph. Among the several variants of VMP in the literature we follow the approach of Minka (2005), as summarized in Section 2.5 of Wand (2017), to derive fragment updates that allow for skew t response models to be handled within the VMP framework.

3 The skew t likelihood fragment

The skew t (Azzalini and Capitanio, 2003) likelihood fragment corresponds to the likelihood specification

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

with $\boldsymbol{A}$ generic design matrix, $\boldsymbol{\theta}$ generic vector of coefficients and $\nu > 0$. If we introduce two auxiliary random variables a_{1i} and a_{2i}, $1 \leq i \leq n$, model (2) can be expressed as

$$\begin{aligned} &y_i|\boldsymbol{\theta}, \sigma^2, \lambda, a_{1i}, a_{2i} \overset{\text{ind.}}{\sim} N\left((\boldsymbol{A}\boldsymbol{\theta})_i + \frac{\sigma\lambda|a_{1i}|\sqrt{a_{2i}}}{\sqrt{1+\lambda^2}}, \frac{a_{2i}\sigma^2}{1+\lambda^2}\right), \\ &a_{1i} \overset{\text{ind.}}{\sim} N(0,1), \quad a_{2i} \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2(\nu,\nu). \end{aligned} \tag{3}$$

We then develop two VMP algorithms assuming that the optimal q-density admits the following alternative product restrictions:

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

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

The second partition avoids a more restrictive independence assumption on the auxiliary variables but requires an extensive use of non-standard conjugate exponential families and numerical integration when deriving a VMP algorithm. Figure 1 includes the factor graph representations of model (3) under both assumption A and B in support of VMP algorithm derivations. In detail, the shaded squares correspond to factors, which are single product components of model (3). The unshaded circles are called stochastic nodes and refer to parameters and variables expressing product dependencies in the approximating densities A and B. Edges connect factors to the stochastic nodes included in each factor.

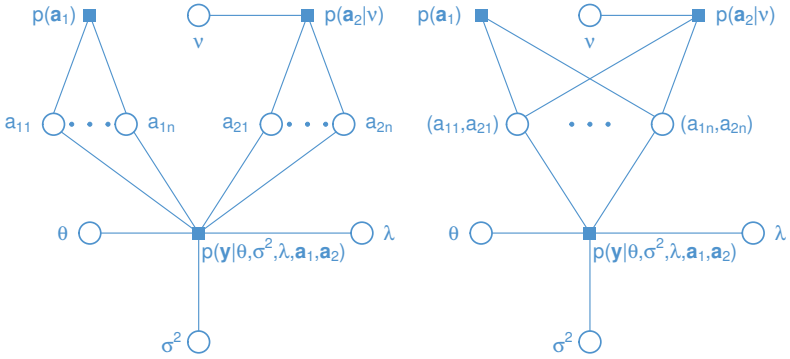

FIGURE 1. Factor graph for Skew t likelihood specification in (3) under assumption A (left panel) and B (right panel).

4 Simulation study and application

We conduct a simulation study to assess the performances of VMP generating one hundred datasets of size $n = 500$ according to model (3), using a Uniform $(0, 1)$ predictor. We set the vector of location parameters $\boldsymbol{\theta} \equiv \boldsymbol{\beta} = (\beta_0, \beta_1)$ to be $\beta_0 = 1$ and $\beta_1 = 2$. The scale parameter is $\sigma = 1$ and the shape parameters are $\lambda = 5$ and $\nu = 1.5$. The hyperparameters for $\boldsymbol{\beta}$ are fixed to $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = 10^{10}\boldsymbol{I}$ over a prior $N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right)$. We use an Inverse-$\chi^2(0.01, 0.01)$ prior on the squared scale. The prior for the parameter of symmetry λ is assumed to be $N\left(0, 10^{10}\right)$ and that for the

degrees of freedom ν to be $\Gamma(1, 0.01)$. For each single parameter, we measure the accuracy of VMP approximations $q^*(\theta_i)$ from (1) under both the assumptions A and B as

$$\text{accuracy}(q^*) = 1 - \frac{1}{2}\int_{-\infty}^{\infty} |q^*(\theta_i) - p(\theta_i|\mathcal{D})|\, d\theta_i,$$

so that $0 \leq \text{accuracy}(q^*) \leq 1$. The L_1 error appearing in this last expression is a scale independent number that is invariant to monotone transformation on the parameter of interest. This implies, for instance, that the accuracy values for $q^*(\sigma)$ and $q^*(\sigma^2)$ coincide. Exact computation of $p(\theta_i|\mathcal{D})$ is replaced by MCMC samples obtained using `rstan` (Stan Development Team, 2018). MCMC samples of size 10000 were generated setting a burn-in of 5000 values and thinning the remaining 5000 by a factor of 5.
Table 1 summarizes mean and standard deviation of accuracy percentages from the simulation study, indicating a notable improvement when adopting the less restrictive assumption B.

TABLE 1. Average (standard deviation) accuracy from the simulation study.

	Accuracy			
Parameter	Assump. A		Assump. B	
β_0	0.0	(0.0)	37.7	(6.2)
β_1	51.2	(17.7)	57.1	(4.5)
σ^2	18.0	(16.9)	11.0	(3.5)
λ	0.0	(0.0)	10.0	(3.4)
ν	0.0	(0.0)	56.6	(6.1)

We also consider the dataset examined in Azzalini and Capitanio (2003) with the linear model

$$y_i = \beta_0 + \beta_1 \text{CRSP}_i + \varepsilon_i, \quad \varepsilon_i \sim \text{Skew-t}\left(0, \sigma^2, \lambda, \nu\right), \quad 1 \leq i \leq n$$

where the variables y_i and CRSP_i denote the excess rate of the Martin Marietta company and the index of return excess for the whole New York Stock Exchange respectively. Data over a period of $n = 60$ consecutive months from January 1982 to December 1986 are available. We estimate the posterior density functions via VMP approximation under assumption B and MCMC with the same settings of the simulation study.
The plots in Figure 2 show that VMP curves are roughly located around the modes of MCMC densities. Note also that the variance of VMP approximating densities appears lower than that of MCMC posterior densities, coherently with the theoretical results in Wang and Blei (2017). Better VMP performances could be achieved by proposing more generic assumptions on the approximating density function. However, our choice of the

product density restrictions is a compromise among algebraic complexity, feasibility and quality of the approximation.

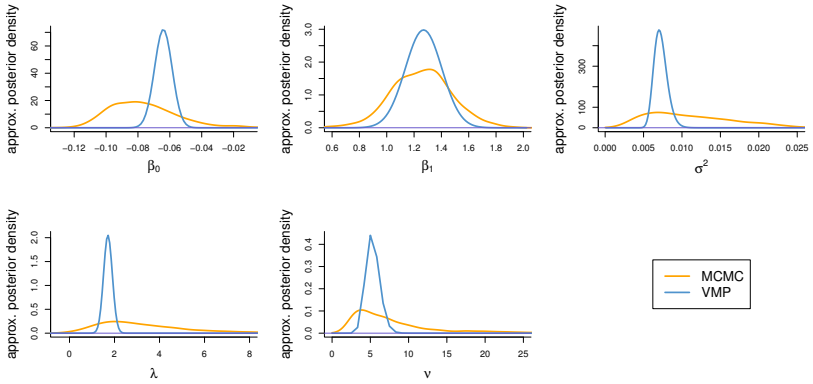

FIGURE 2. Martin Marietta data: posterior density plots via MCMC and VMP.

Acknowledgments: 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. Special Thanks to Alessandra Salvan and Nicola Sartori for their assistance with this research.

References

Azzalini, A. and 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*, **65**, 367 – 389.

McLean, M.W. and Wand, M.P. (2018). Variational Message Passing for Elaborate Response Regression Models. *Bayesian Analysis*, in press.

Minka, T. (2005). Divergence measures and message passing. *Microsoft Research Technical Report Series*, **173**, 1 – 17.

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

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

Wang, Y. and Blei, D. M. (2017). Frequentist Consistency of Variational Bayes. *arXiv:1705.03439*.