Variational Bayes for Elaborate Distributions

BY M.P. WAND[1], J.T. ORMEROD[2], S.A. PADOAN[3] & R. FRÜHWIRTH[4]

[1]*Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong 2522, AUSTRALIA*

[2]*School of Mathematics and Statistics, University of Sydney, Sydney 2006, AUSTRALIA*

[3]*Swiss Federal Institute of Technology, Lausanne, SWITZERLAND*

[4]*Institute of High Energy Physics, Austrian Academy of Sciences, Vienna, AUSTRIA*

16th September, 2010

SUMMARY

We develop strategies for variational Bayes approximate inference for models containing elaborate distributions. Such models suffer from the difficulty that the parameter updates do not admit closed form solutions. We circumvent this problem through a combination of (a) specially tailored auxiliary variables, (b) univariate quadrature schemes and (c) finite mixture approximations of troublesome density functions. An accuracy assessment is conducted and the new methodology is illustrated in an application.

Keywords: Asymmetric Laplace distribution; Auxiliary mixture sampling; Bayesian inference; Generalized Extreme Value distribution; Quadrature; Skew Normal Distribution.

1 Introduction

Variational Bayes refers to a general approach to approximate inference in hierarchical Bayesian models and offers a fast deterministic alternative to Markov chain Monte Carlo (MCMC). The approximations are driven by product assumptions on multi-parameter posterior distributions and lead to fast and, for some models, quite accurate Bayesian inference. The idea originated in the Physics literature (e.g. Parisi, 1988), where it is known as *mean field theory*. It was adopted by Computer Science for Bayesian inference in the late 1990s (e.g. Attias, 1999) and the term 'variational Bayes' was coined. During the 2000s it permeated into the statistical literature (e.g. Teschendorff *et al.*, 2005; McGrory & Titterington, 2007). Ormerod & Wand (2010) contains a summary of variational Bayes from a statistical standpoint.

A vital feature of variational Bayes, which allows it to be applied to a wide class of models, is the *localness property*. The localness property means that calculations concerning a particular parameter can be confined to 'nearby' parameters. It is best understood using graph theoretic representations of hierarchical Bayesian models, although we postpone the details on this to Section 3. Gibbs sampling also possesses the localness property and the software package BUGS (Lunn *et al.* 2000) relies on it to efficiently handle arbitrary complex models. Recently software packages that make use of the localness property of variational Bayes have emerged in an effort to streamline data analysis. The most prominent of these is Infer.NET (Minka, Winn, Guiver & Kannan, 2009) which is a suite of classes in .NET languages such as C++ and C♯.

Despite these developments, the vast majority of variational Bayes methodology and software is restricted to models where the random components have common distributions such as Normal, Gamma and Dirichlet, and the required calculations are analytic. This imposes quite stringent restrictions on the set of models that can be handled via variational Bayes. The current release of Infer.NET is subject to such restrictions.

In this article we explain how the class of distributions for variables in variational Bayes algorithms can be widened considerably. Specifically, we show how *elaborate* distributions such as t, Skew Normal, Asymmetric Laplace and Generalized Extreme Value can be handled within the variational Bayes framework. The incorporation of such distributions is achieved via a combination of

- specially tailored auxiliary variables,
- univariate quadrature schemes,
- finite mixture approximations to troublesome density functions.

Auxiliary variables have already enjoyed some use in variational Bayes contexts. Examples include Tipping & Lawrence (2003) for t-based robust curve fitting with fixed degrees of freedom, Archambeau & Bach (2008) and Armagan (2009) for Laplace and other exponential power distributions and Girolami & Rogers (2006) and Consonni & Marin (2007) for binary response regression. Quadrature and finite mixture approximations have received little, if any, attention in the variational Bayes literature.

We identify five distinct families of univariate integrals which arise in variational Bayes for the elaborate distributions treated here. The integrals within the families do not admit analytic solutions and quadrature is required. However, the integrands are well-behaved and we are able to tailor common quadrature schemes to achieve stable and accurate computation. It should also be noted that use of accurate quadrature schemes corresponds to exact variational Bayes updates as opposed to those based on Monte Carlo methods (e.g. Section 6.3 of Winn & Bishop, 2005).

A recent innovation in the MCMC literature is *auxiliary mixture sampling* (e.g. Frühwirth-Schnatter & Wagner, 2006; Frühwirth-Schnatter *et al.*, 2009). It involves approximation of particular density functions by finite, usually Normal, mixtures. The introduction of auxiliary indicator variables corresponding to components of the mixtures means that MCMC reduces to ordinary Gibbs sampling with closed form updates. The same idea is applicable to variational Bayes, and we use it for troublesome density functions such as those belonging to the Generalized Extreme Value family.

We confine much of our discussion to simple univariate models, since the forms of many of the updates for multi-parameter extensions are essentially the same. The localness property of variational Bayes means the these forms are unchanged when embedded into larger models.

A critical issue of variational Bayesian inference is accuracy compared with more exact approaches such as MCMC. We address this through a simulation study for a selection of elaborate distribution models. We find that the posterior densities of some parameters can be approximated very well. However the accuracy is only moderate to good for parameters which possess non-negligible posterior dependence with the introduced auxiliary variables. In particular, the spread of posterior densities are often under-approximated.

Section 2 contains all definitions and distributional results used in this article. Section 3 summarizes the variational Bayes and elaborates on the aforementioned localness property. In Section 4 we treat several location-scale models having elaborate distributional forms. Section 5 describes modifications when the alternative scale parameter priors are used. Multiparameter extensions are discussed in Section 6. In Section 7 we discuss extension to other elaborate distributions including discrete response models. The accuracy of variational Bayes for elaborate distribution models is assessed in Section 8. Section 9 applies some of the methodology developed in this paper to analysis of data from a respiratory health study. Discussion of the methodology and its performance is given in Section 9. Three appendices provide technical details.

2 Definitions and Distributional Results

Variational Bayes for elaborate distributions is very algebraic, and relies on several definitions and distributional results. We lay out each of them in this section. Each of the results can be obtained via standard distribution theoretic manipulations.

2.1 Non-analytic Integral Families

A feature of variational Bayes for elaborate distributions is that not all calculations can be done analytically. Some univariate quadrature is required. The following integral families comprise the full set of non-analytic integrals which arise in the models considered in this article:

$$\mathcal{F}(p,q,r,s,t) \equiv \int_s^t x^p \exp\left[q\{\tfrac{1}{2}x\log(x/2) - \log\Gamma(x/2)\} - \tfrac{1}{2}rx\right] dx,\ p \geq 0,\ q,r,s,t > 0;$$

$$\mathcal{G}(p,q,r,s,t) \equiv \int_{-\infty}^{\infty} x^p(1+x^2)^q \exp\left(-r\,x^2 + s\,x\sqrt{1+x^2} + tx\right)\,dx\ p,q \geq 0,\ r > 0;$$

$$\mathcal{H}(p,q,r) \equiv \int_0^{\infty} x^p \exp\{-q\,x^2 - \log(r + x^{-2})\}\,dx,\ p \geq 0,\ q,r > 0;$$

$$\mathcal{J}(p,q,r,s) \equiv \int_{-\infty}^{\infty} x^p \exp(qx - rx^2 - se^{-x})\,dx,\ p \geq 0,\ -\infty < q < \infty,\ r,s > 0$$

$$\text{and } \mathcal{J}^+(p,q,r) \equiv \int_0^{\infty} x^p \exp(qx - rx^2)\,dx,\ p \geq 0,\ -\infty < q < \infty,\ r > 0.$$

Since the integrals can take values that are arbitrarily large or small it is recommended that logarithmic storage and arithmetic be used. Appendix B discusses stable and efficient numerical computation of the members of each of these integral families.

2.2 Distributional Notation

The density function of a random vector $\boldsymbol{v}$ in a Bayesian model is denoted by $p(\boldsymbol{v})$. The conditional density of $\boldsymbol{v}$ given $\boldsymbol{w}$ is denoted by $p(\boldsymbol{v}|\boldsymbol{w})$. The covariance matrix of $\boldsymbol{v}$ is denoted by $\text{Cov}(\boldsymbol{v})$. If x_i has distribution D for each $1 \leq i \leq n$, and the x_i are independent, then we write $y_i \overset{\text{ind.}}{\sim} D$.

We use q to denote density functions that arise from variational Bayes approximation. For a generic random variable v and density function q we define:

$$\mu_{q(v)} \equiv E_q(v) \quad \text{and} \quad \sigma^2_{q(v)} \equiv \text{Var}_q(v).$$

For a generic random vector $\boldsymbol{v}$ and density function q we define:

$$\boldsymbol{\mu}_{q(\boldsymbol{v})} \equiv E_q(\boldsymbol{v}) \quad \text{and} \quad \Sigma_{q(\boldsymbol{v})} \equiv \text{Cov}_q(\boldsymbol{v}).$$

2.3 Distributional Definitions

We use the common notation, $N(\mu, \sigma^2)$, for the Normal distribution with mean μ and variance σ^2. The density and cumulative distribution functions of the $N(0,1)$ distribution are denoted by ϕ and Φ, respectively. Furthermore, we write $(\phi/\Phi)(x) \equiv \phi(x)/\Phi(x)$ for the ratio of these two functions.

The Inverse-Gaussian density function with mean $\mu > 0$ and precision $\gamma > 0$ is given by

$$p(x;\mu,\gamma) = \gamma^{1/2}(2\pi\,x^3)^{-1/2}\,\exp\left\{-\frac{\gamma(x-\mu)^2}{2\mu^2\,x}\right\},\quad x>0.$$

We write Inverse-Gaussian(μ,γ) for the corresponding family of distributions.

Table 1 provides the functional forms for the densities that are used for modelling observed data in Section 4. For simplicity, we given the density with location μ equal to zero and scale σ equal to one. The general location and scale density function involves the transition

$$f(x) \mapsto \frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$$

where $f(x)$ is as given in the second column of Table 1.

distribution	density in x ($\mu=0,\sigma=1$)	abbreviation
t	$\dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)(1+x^2/\nu)^{\frac{\nu+1}{2}}}$	$t(\mu,\sigma,\nu)\quad(\nu>0)$
Asymmetric Laplace	$\tau(1-\tau)\,e^{-\frac{1}{2}\lvert x\rvert+(\tau-\frac{1}{2})x}$,	Asymmetric-Laplace(μ,σ,τ) $(0<\tau<1)$
Skew Normal	$2\phi(x)\Phi(\lambda x)$	Skew-Normal(μ,σ,λ)
Finite Normal Mixture	$(2\pi)^{-1/2}\sum_{k=1}^K(w_k/s_k)$ $\times\phi((x-m_k)/s_k)$,	Normal-Mixture$(\mu,\sigma,\boldsymbol{w},\boldsymbol{m},\boldsymbol{s})$ $(\sum_{k=1}^K w_k=1,\ s_k>0)$
Generalized Extreme Value	$(1+\xi\,x)^{-1/\xi-1}$ $\times e^{-(1+\xi\,x)^{-1/\xi}},\ 1+\xi\,x>0$	GEV(μ,σ,ξ)

Table 1: *Density functions for modelling observed data. The functions ϕ and Φ are the density and cumulative distribution functions of the $N(0,1)$ distribution. The scale parameter is subject to the restriction $\sigma>0$ in all cases. The density function argument x and parameters range over $\mathbb{R}$ unless otherwise specified.*

In Table 2 we describe density families that are used for modelling scale parameters in the upcoming examples.

2.4 Distributional Results Involving Auxiliary Variables

In this section we give a collection of distributional results that link elaborate distributions to simpler ones. Each result is straightforward to derive. However, they play vital roles in variational Bayes for elaborate distributions.

Result 1. *Let x and a be random variables such that*

$$x|a \sim N\left(\mu, a\sigma^2\right)\quad\text{and}\quad a\sim\text{Inverse-Gamma}(\tfrac{\nu}{2},\tfrac{\nu}{2}).$$

Then $x\sim t(\mu,\sigma,\nu)$.

Result 2. *Let x and a be random variables such that*

$$x|a\sim N\left(\mu+\frac{(\frac{1}{2}-\tau)\sigma}{a\tau(1-\tau)},\,\frac{\sigma^2}{a\tau(1-\tau)}\right)\quad\textit{and}\quad a\sim\text{Inverse-Gamma}(1,\tfrac{1}{2}).$$

distribution	density in x	abbreviation
Inverse Gamma	$\frac{B^A}{\Gamma(A)}\,x^{-A-1}e^{-B/x}$	Inverse-Gamma(A,B) $\quad(A,B>0)$
Log Normal	$\frac{1}{Bx\sqrt{2\pi}}\,\exp[-\frac{1}{2B^2}\{\log(x)-A\}^2]$	Log-Normal(A,B) $\quad(B>0)$
Half Cauchy	$2A\{\pi(A^2+x^2)\}^{-1}$	Half-Cauchy(A) $\quad(A>0)$

Table 2: *Density functions used for modelling scale parameters. The density function argument x ranges over $x>0$.*

Then $x\sim$ Asymmetric-Laplace(μ,σ,τ).

Result 2 follows from Proposition 3.2.1 of Kotz, Kozubowski & Podgórski (2001).

Result 3. *Let x and a be random variables such that*

$$x|a\sim N\left(\mu+\frac{\sigma\lambda|a|}{\sqrt{1+\lambda^2}},\frac{\sigma^2}{1+\lambda^2}\right)\quad\textit{and}\quad a\sim N(0,1).$$

Then $x\sim$ Skew-Normal(μ,σ,λ).

Result 3 is an immediate consequence of Proposition 3 of Azzalini & Dalle Valle (1996). These authors trace the result back to Aigner, Lovell & Schmidt (1977).

Our last result involves the Multinomial$(1;\boldsymbol{\pi})$ distribution where $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_K)$ is such that $\sum_{k=1}^K\pi_k=1$. The corresponding probability mass function is $p(x_1,\ldots,x_K)=\prod_{k=1}^K\pi_k^{x_k},\ x_k=0,1$, for $1\le k\le K$.

Result 4. *Let x be a random variable and $\boldsymbol{a}$ be a $K\times 1$ random vector, having kth entry a_k, such that*

$$p(x|\boldsymbol{a})=\prod_{k=1}^K\left[(2\pi s_k^2)^{-1/2}\exp\{-\tfrac{1}{2}(x-m_k)^2/s_k^2\}\right]^{a_k},\quad -\infty<x<\infty,$$

$$\textit{and}\qquad \boldsymbol{a}\sim\text{Multinomial}(1;\boldsymbol{w}).$$

Then $x\sim$ Normal-Mixture$(0,1,\boldsymbol{w},\boldsymbol{m},\boldsymbol{s})$.

2.5 Expectation Results

The following expectation results are useful in some of the variational Bayes problems treated in Section 4. If $v\sim$ Inverse-Gamma(A,B) then

$$E(1/v)=A/B\quad\text{and}\quad E\{\log(v)\}=\log(B)-\text{digamma}(A).$$

If $v\sim$ Inverse-Gaussian(μ,γ) then

$$E(v)=\mu\quad\text{and}\quad E(1/v)=\frac{1}{\mu}+\frac{1}{\gamma}.\tag{1}$$

3 Variational Bayes

Variational Bayesian inference relies on product restrictions on posterior densities. For example, in a model with parameters μ and σ and observed data vector $\boldsymbol{x}$, the exact joint posterior density $p(\mu,\sigma|\boldsymbol{x})$ is replaced by the product density form

$$q(\mu)\,q(\sigma)\tag{2}$$

in the hope that the latter is more tractable. The Kullback-Leibler distance between $p(\mu,\sigma|\boldsymbol{x})$ and (2) is minimized by $q^*(\mu)$ and $q^*(\sigma)$ satisfying:

$$\begin{array}{rrcl} & q^*(\mu) & \propto & \exp\{E_{q(\sigma)}\log p(\mu|\sigma,\boldsymbol{x})\} \\ \text{and} & q^*(\sigma) & \propto & \exp\{E_{q(\mu)}\log p(\sigma|\mu,\boldsymbol{x})\}. \end{array} \tag{3}$$

The optimal parameters in these q-densities can be determined by an iterative scheme induced by (3). Each iteration is, under mild assumptions, guaranteed to lead to an increase in

$$\log \underline{p}(\boldsymbol{x};q) \equiv E_{q(\mu,\sigma)}\{\log p(\boldsymbol{x},\mu,\sigma) - \log q(\mu,\sigma)\}$$

(Luenberger & Ye; 2008, p. 253). Successive values of $\log \underline{p}(\boldsymbol{x};q)$ can be used to monitor convergence. At convergence $q^*(\mu)$, $q^*(\sigma)$ and $\log \underline{p}(\boldsymbol{x};\overline{q})$ are, respectively, the minimum Kullback-Leibler approximations to the posterior densities $p(\mu|\boldsymbol{x})$, $p(\sigma|\boldsymbol{x})$ and the marginal log-likelihood $\log p(\boldsymbol{x})$.

The extension to general Bayesian models with arbitrary parameter vectors and latent variables is straightforward. Summaries may be found in, for example, Chapter 10 of Bishop (2006) and Ormerod & Wand (2010). As described in these references, directed acyclic graph (DAG) representations of Bayesian hierarchical models are very useful when deriving variational Bayes schemes for large models. We make use of DAG representations in the remainder of this section.

As mentioned in Section 1, an important feature of variational Bayes is the localness property. In the current paper, this implies that results established for the smaller models treated in Sections 4–6 also apply to much larger models. We now explain this property in graphical terms. Consider the generic hierarchical Bayesian model:

$$\begin{gathered} \boldsymbol{x}|\theta_1,\theta_2,\theta_3 \sim p(\boldsymbol{x}|\,\theta_1,\theta_2,\theta_3), \\ \theta_1|\,\theta_4 \sim p(\theta_1|\,\theta_4), \quad \theta_2|\,\theta_5,\theta_6 \sim p(\theta_2|\,\theta_5,\theta_6), \quad \theta_3|\,\theta_6 \sim p(\theta_3|\,\theta_6) \quad \text{independently,} \\ \theta_4 \sim p(\theta_4), \quad \theta_5 \sim p(\theta_5), \quad \theta_6 \sim p(\theta_6) \quad \text{independently.} \end{gathered} \tag{4}$$

The variational Bayes solutions satisfy

$$q^*(\theta_i) \propto \exp\{E_{q(\theta_{-i})}\log p(\theta_i|\boldsymbol{x},\theta_{-i})\}, \quad 1 \le i \le 6,$$

where θ_{-i} denotes the set $\{\theta_1,\ldots,\theta_6\}$ with θ_i excluded. However, from graphical model theory (Pearl, 1988), we have the result

$$p(\theta_i|\boldsymbol{x},\theta_{-i}) = p(\theta_i|\text{Markov blanket of }\theta_i)$$

where the Markov blanket of a node on a DAG is the set of parents, co-parents and children of that node. From this result we get the simplification

$$q^*(\theta_i) \propto \exp\{E_{q(\theta_{-i})}\log p(\theta_i|\text{Markov blanket of }\theta_i)\}, \quad 1 \le i \le 6. \tag{5}$$

The localness property of variational Bayes is encapsulated in Result (5). It affords considerable simplification for the model at hand, but also allows variational Bayes results for one model to be transferred to another. We now explain this graphically.

Figure 1 shows the Markov blankets for the each of $\theta_1,\ldots,\theta_6$. The θ_i are known as *hidden nodes* in graphical models parlance and the data vector $\boldsymbol{x}$ comprises the *evidence node*. The arrows convey conditional dependence among the random variables in the model. The Markov blanket for θ_1 is $\{\theta_2,\theta_3,\theta_4,\boldsymbol{x}\}$, which means that $q^*(\theta_1)$ depends on particular q-density moments of θ_2, θ_3 and θ_4, *but not on their distributions*. If, for example, $p(\theta_2|\theta_5)$ is changed from $\text{Inverse-Gamma}(0.07,\theta_5)$ to $\text{Log-Norma}(25,\theta_5)$ then this will not impact upon the form of $q^*(\theta_1)$. The variational Bayes solution for $q^*(\theta_4)$

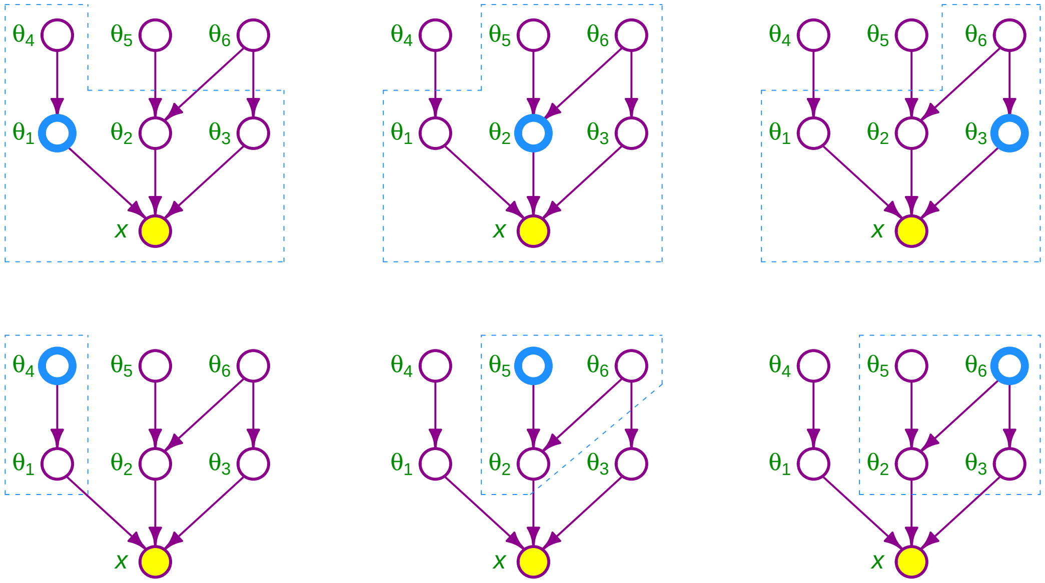

Figure 1: *Markov blankets for each of the six parameters (hidden nodes) in the example Bayesian hierarchical model (directed acyclic graph), given by (4). In each panel the Markov blanket is shown for the thick-circled blue node, using dashed lines. The shaded node $\boldsymbol{x}$ corresponds to the observed data (evidence node).*

provides a more dramatic illustration of the localness property, since the Markov blanket of θ_4 is simply $\{\theta_1\}$. This means that $q^*(\theta_4)$ is unaffected by the likelihood $p(\boldsymbol{x}|\theta_1, \theta_2, \theta_3)$. Therefore, results established for $q^*(\theta_4)$ for, say, $x_i|\theta_1, \theta_2, \theta_3 \overset{\text{ind.}}{\sim} t(\theta_1, \theta_2, \theta_3)$ also apply to $x_i|\theta_1, \theta_2, \theta_3 \overset{\text{ind.}}{\sim} \text{GEV}(\theta_1, \theta_2, \theta_3)$.

The upshot of the localness property of variational Bayes is that we can restrict attention to the simplest versions of models involving elaborate distributions with the knowledge that the forms that arise also apply to larger models. For this reason, Section 4 deals only with such simple models.

4 Univariate Location-Scale Models

Consider univariate Bayesian models of the form

$$x_1, \ldots, x_n|\mu, \sigma, \boldsymbol{\theta} \overset{\text{ind.}}{\sim} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}; \boldsymbol{\theta}\right) \tag{6}$$

where f is a fixed density function, $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ is the scale parameter and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a set of shape parameters. We call (6) a *univariate location-scale model*.

We will take the prior on μ to be Gaussian:

$$\mu \sim N(\mu_\mu, \sigma_\mu^2), \quad -\infty < \mu_\mu < \infty, \; \sigma_\mu^2 > 0$$

throughout this article. Gaussian priors for location parameters are generally adequate, and have a straightforward multi-parameter extension. Prior specification for scale parameters is somewhat more delicate (Gelman, 2006). In the current section we take the prior for σ to be of the form

$$p(\sigma) \propto \sigma^{-2A-1} e^{-B/\sigma^2}, \quad A, B > 0. \tag{7}$$

This is equivalent to the squared scale, σ^2, having an Inverse-Gamma prior. Due to conjugacy relationships between the Gaussian and Inverse-Gamma families, use of (7) results in variational Bayes algorithms with fewer intractable integrals. In Section 5 we treat alternative scale parameter priors. Let $p(\boldsymbol{\theta})$ denote the prior on $\boldsymbol{\theta}$. The form of $p(\boldsymbol{\theta})$ will change from one model to another.

The exact posterior density function for μ is

$$p(\mu|\boldsymbol{x}) = \frac{\exp\{-\frac{1}{2\sigma_\mu^2}(\mu-\mu_\mu)^2\}\int_{\boldsymbol{\Theta}}\int_0^\infty \sigma^{-n}\prod_{i=1}^n f\{(x_i-\mu)/\sigma;\boldsymbol{\theta}\}\,d\sigma\,d\boldsymbol{\theta}}{\int_{-\infty}^\infty \exp\{-\frac{1}{2\sigma_\mu^2}(\mu-\mu_\mu)^2\}\int_{\boldsymbol{\Theta}}\int_0^\infty \sigma^{-n}\prod_{i=1}^n f\{(x_i-\mu)/\sigma;\boldsymbol{\theta}\}\,d\sigma\,d\boldsymbol{\theta}\,d\mu}.$$

Similar expressions arise for $p(\sigma|\boldsymbol{x})$ and $p(\boldsymbol{\theta}|\boldsymbol{x})$. For elaborate f forms, the integrals in the normalizing factors are almost always intractable. For multi-parameter extensions we get stuck with multivariate integrals of arbitrary dimension.

The remainder of this section involves case-by-case treatment of the univariate location-scale models that arise when f is set to each of the densities in Table 1. These cases allow illustration of the difficulties that arise in variational Bayesian inference for elaborate distributions, and our strategy for overcoming them. Discussion concerning other f forms is given in Section 7.

4.1 t Model

A Bayesian t model for a univariate random sample is

$$\begin{gathered} x_i|\,\mu,\sigma \overset{\text{ind.}}{\sim} t(\mu,\sigma,\nu), \\ \mu\sim N(\mu_\mu,\sigma_\mu^2),\quad \sigma^2\sim\text{Inverse-Gamma}(A,B),\quad \nu\sim\text{Uniform}(\nu_{\text{min}},\nu_{\text{max}}) \end{gathered} \tag{8}$$

where μ_μ and $A, B, \nu_{\text{min}}, \nu_{\text{max}}, \sigma_\mu^2 > 0$ are hyperparameters. Model (8) and its multiparameter extensions (Section 6) possess attractive robustness properties (e.g. Lange, Little & Taylor, 1989). Section 9 contains a nonparametric regression example that uses the t distribution to achieve robustness.

Using Result 1 we can re-write (8) as

$$\begin{gathered} x_i|a_i,\mu,\sigma \overset{\text{ind.}}{\sim} N(\mu, a_i\,\sigma^2),\quad a_i|\nu \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{\nu}{2},\tfrac{\nu}{2}), \\ \mu\sim N(\mu_\mu,\sigma_\mu^2),\quad \sigma^2\sim\text{Inverse-Gamma}(A,B),\quad \nu\sim\text{Uniform}(\nu_{\text{min}},\nu_{\text{max}}). \end{gathered}$$

For variational Bayesian inference we impose the product restriction

$$q(\mu,\sigma,\nu,\boldsymbol{a}) = q(\mu,\nu)q(\sigma)q(\boldsymbol{a}).$$

This yields the following forms for the optimal densities:

$$\begin{aligned} &q^*(\mu)\sim N(\mu_{q(\mu)},\sigma^2_{q(\mu)}) \\ &q^*(\sigma^2)\sim\text{Inverse-Gamma}\left(A+\tfrac{n}{2}, B+\tfrac12\textstyle\sum_{i=1}^n \mu_{q(1/a_i)}\{(x_i-\mu_{q(\mu)})^2+\sigma^2_{q(\mu)}\}\right) \\ &q^*(a_i)\overset{\text{ind.}}{\sim}\text{Inverse-Gamma}\left(\tfrac{\mu_{q(\nu)}+1}{2},\tfrac12\left[\mu_{q(\nu)}+\mu_{q(1/\sigma^2)}\{(x_i-\mu_{q(\mu)})^2+\sigma^2_{q(\mu)}\}\right]\right) \\ &q^*(\nu) = \frac{\exp\left[n\left\{\frac{\nu}{2}\log(\nu/2)-\log\Gamma(\nu/2)\right\}-(\nu/2)C_1\right]}{\mathcal{F}(0,n,C_1,\nu_{\text{min}},\nu_{\text{max}})},\quad \nu_{\text{min}}<\nu<\nu_{\text{max}}. \end{aligned} \tag{9}$$

The last density uses the definition: $C_1 \equiv \sum_{i=1}^n\{\mu_{q(\log a_i)}+\mu_{q(1/a_i)}\}$. The parameters in (9) are determined from Algorithm 1.

```
Initialize: $\mu_{q(\mu)} \in \mathbb{R}$, $\sigma^2_{q(\mu)} > 0$, $\mu_{q(\nu)} \in [\nu_{\text{min}}, \nu_{\text{max}}]$ and $\mu_{q(1/\sigma^2)} > 0$.
Cycle:

    For $i = 1, \ldots, n$:

        $B_{q(a_i)} \leftarrow \frac{1}{2}\left[\mu_{q(\nu)} + \mu_{q(1/\sigma^2)}\{(x_i - \mu_{q(\mu)})^2 + \sigma^2_{q(\mu)}\}\right]$
        $\mu_{q(1/a_i)} \leftarrow \frac{1}{2}(\mu_{q(\nu)} + 1)/B_{q(a_i)}$
        $\mu_{q(\log a_i)} \leftarrow \log(B_{q(a_i)}) - \text{digamma}(\frac{1}{2}(\mu_{q(\nu)} + 1))$

    $\sigma^2_{q(\mu)} \leftarrow \left(\mu_{q(1/\sigma^2)} \sum_{i=1}^n \mu_{q(1/a_i)} + \frac{1}{\sigma^2_\mu}\right)^{-1}$

    $\mu_{q(\mu)} \leftarrow \sigma^2_{q(\mu)} \left(\mu_{q(1/\sigma^2)} \sum_{i=1}^n x_i \mu_{q(1/a_i)} + \frac{\mu_\mu}{\sigma^2_\mu}\right)$

    $C_1 \leftarrow \sum_{i=1}^n \{\mu_{q(\log a_i)} + \mu_{q(1/a_i)}\}$ ; $\mu_{q(\nu)} \leftarrow \frac{\mathcal{F}(1, n, C_1, \nu_{\text{min}}, \nu_{\text{max}})}{\mathcal{F}(0, n, C_1, \nu_{\text{min}}, \nu_{\text{max}})}$

    $B_{q(\sigma^2)} \leftarrow B + \frac{1}{2} \sum_{i=1}^n \mu_{q(1/a_i)}\{(x_i - \mu_{q(\mu)})^2 + \sigma^2_{q(\mu)}\}$ ; $\mu_{q(1/\sigma^2)} \leftarrow \frac{A + \frac{n}{2}}{B_{q(\sigma^2)}}$

until the increase in $\underline{p}(\boldsymbol{x}; q)$ is negligible.
```

Algorithm 1: *Iterative scheme for obtaining the parameters in the optimal densities $q^*(\boldsymbol{a})$, $q^*(\mu)$, $q^*(\nu)$ and $q^*(\sigma)$ for the t model.*

An explicit expression for $\log \underline{p}(\boldsymbol{x}; q)$ is:

$$\begin{aligned}
\log \underline{p}(\boldsymbol{x}; q) \quad = \quad & \tfrac{n+1}{2} + \tfrac{n}{2}\mu_{q(\nu)} - \tfrac{n}{2}\log(2\pi) + \tfrac{1}{2}\log(\sigma^2_{q(\mu)}/\sigma^2_\mu) - \frac{(\mu_{q(\mu)} - \mu_\mu)^2 + \sigma^2_{q(\mu)}}{2\sigma^2_\mu} \\
& + A\log(B) - \log\Gamma(A) - (A + \tfrac{n}{2})\log(B_{q(\sigma^2)}) + \log\Gamma(A + \tfrac{n}{2}) \\
& + \log\mathcal{F}(0, n, C_1, \nu_{\text{min}}, \nu_{\text{max}}) - \log(\nu_{\text{max}} - \nu_{\text{min}}) \\
& + n\log\Gamma(\tfrac{1}{2}(\mu_{q(\nu)} + 1)) - \tfrac{n}{2}(\mu_{q(\nu)} + 1)\,\text{digamma}\{\tfrac{1}{2}(\mu_{q(\nu)} + 1)\}
\end{aligned}$$

although it is only valid after each of the updates in Algorithm 1 have been performed.

Figure 2 shows the results from application of Algorithm 1 to a simulated data set of size $n = 500$ from the $t(4, 0.5, 1.5)$ distribution. The algorithm was terminated when the relative increase in $\log \underline{p}(\boldsymbol{x}; q)$ was less than 10^{-6}. As shown in the first panel of Figure 2, this required about 75 iterations. The true parameter values are within the high probability regions of each approximate posterior density function, and this tended to occur for other realizations of the simulated data.

4.2 Asymmetric Laplace Model

The Asymmetric Laplace model for a univariate random sample is

$$\begin{gathered}
x_i |\, \mu, \sigma \overset{\text{ind.}}{\sim} \text{Asymmetric-Laplace}(\mu, \sigma, \tau), \\
\mu \sim N(\mu_\mu, \sigma^2_\mu), \quad \sigma^2 \sim \text{Inverse-Gamma}(A, B)
\end{gathered} \tag{10}$$

where $\mu_\mu \in \mathbb{R}$ and $A, B, \sigma^2_\mu > 0$ are hyperparameters.

We treat the case where the asymmetry parameter $0 < \tau < 1$ is fixed number to be specified by the user. Note that, μ equals the τ quantile of the distribution of the

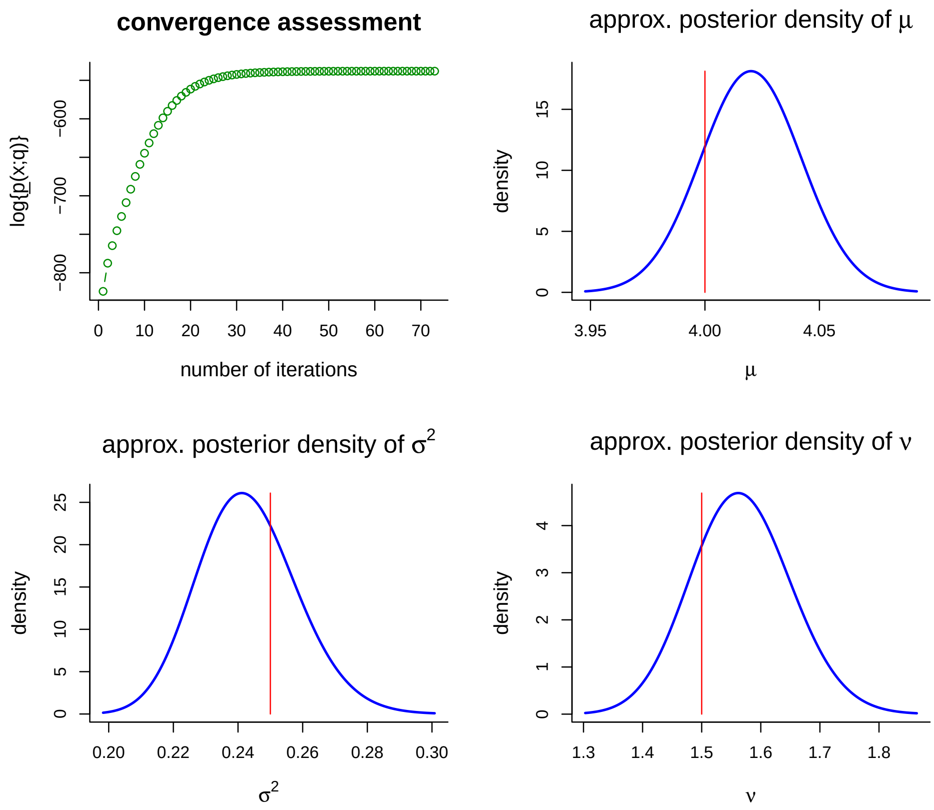

Figure 2: *Results of application of Algorithm 1 to a simulated random sample of size* $n = 500$ *from the* $t(4, 0.5, 1.5)$ *distribution. The upper-left panel shows successive values of* $\log \underline{p}(\boldsymbol{x}; q)$, *up until the meeting of a stringent convergence criterion. The other panels show the approximate posterior density functions for the three model parameters. The vertical lines correspond to the true values of the parameters from which the data were generated.*

x_is. Multiparameter extensions of (10), of the type described in Section 6, corresponds to Bayesian quantile regression (Yu & Moyeed, 2001). Laplacian variables also arise in Bayesian representations of the lasso (Park & Casella, 2009) and wavelet-based nonparametric regression.

Using Result 2 we can re-write model (10) as

$$x_i|a_i, \mu, \sigma \overset{\text{ind.}}{\sim} N\left(\mu + \frac{(\frac{1}{2} - \tau)\sigma}{a_i\tau(1-\tau)}, \frac{\sigma^2}{a_i\tau(1-\tau)}\right), \quad a_i \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(1, \tfrac{1}{2}),$$

$$\mu \sim N(\mu_\mu, \sigma_\mu^2), \quad \sigma^2 \sim \text{Inverse-Gamma}(A, B).$$

For variational Bayesian inference we impose the product restriction

$$q(\mu, \sigma, \boldsymbol{a}) = q(\mu)q(\sigma)q(\boldsymbol{a}). \tag{11}$$

The optimal densities take the forms:

$$\begin{aligned} q^*(\mu) &\sim N(\mu_{q(\mu)}, \sigma^2_{q(\mu)}), \\ q^*(\sigma) &\sim \frac{\sigma^{-(2A+n+1)}\exp(C_2/\sigma - C_3/\sigma^2)}{\mathcal{J}^+(2A+n-1, C_2, C_3)} \\ \text{and } q^*(a_i) &\overset{\text{ind.}}{\sim} \text{Inverse-Gaussian}(\mu_{q(a_i)}, \{4\tau(1-\tau)\}^{-1}). \end{aligned}$$

```
Initialize: $\mu_{q(\mu)} \in \mathbb{R}$ and $\sigma^2_{q(\mu)}, \mu_{q(1/\sigma)}, \mu_{q(1/\sigma^2)} > 0$.
Cycle:

    For $i = 1, \ldots, n$:

        $\mu_{q(a_i)} \leftarrow \Big[4\tau^2(1-\tau)^2\mu_{q(1/\sigma^2)}\{(x_i - \mu_{q(\mu)})^2 + \sigma^2_{q(\mu)}\}\Big]^{-1/2}$.
        $\mu_{q(1/a_i)} \leftarrow 1/\mu_{q(a_i)} + 4\tau(1-\tau)$

    $\sigma^2_{q(\mu)} \leftarrow \big\{\tau(1-\tau)\mu_{q(1/\sigma^2)}\sum_{i=1}^n \mu_{q(a_i)} + 1/\sigma^2_\mu\big\}^{-1}$

    $\mu_{q(\mu)} \leftarrow \sigma^2_{q(\mu)}\big\{\tau(1-\tau)\mu_{q(1/\sigma^2)}\sum_{i=1}^n x_i\mu_{q(a_i)} + n(\tau - \frac{1}{2})\mu_{q(1/\sigma)} + \mu_\mu/\sigma^2_\mu\big\}$

    $C_2 \leftarrow n(\overline{x} - \mu_{q(\mu)})(\frac{1}{2} - \tau)$

    $C_3 \leftarrow B + \frac{1}{2}\tau(1-\tau)\sum_{i=1}^n \mu_{q(a_i)}\{(x_i - \mu_{q(\mu)})^2 + \sigma^2_{q(\mu)}\}$

    $\mu_{q(1/\sigma^2)} \leftarrow \dfrac{\mathcal{J}^+(2A+n+1, C_2, C_3)}{\mathcal{J}^+(2A+n-1, C_2, C_3)}$ ; $\mu_{q(1/\sigma)} \leftarrow \dfrac{\mathcal{J}^+(2A+n, C_2, C_3)}{\mathcal{J}^+(2A+n-1, C_2, C_3)}$

until the increase in $\underline{p}(\boldsymbol{x}; q)$ is negligible.
```

Algorithm 2: *Iterative scheme for obtaining the parameters in the optimal densities $q^*(\boldsymbol{a})$, $q^*(\mu)$ and $q^*(\sigma)$ for the Asymmetric Laplace model.*

The parameters are determined from Algorithm 2.

An expression for $\log\,\underline{p}(\boldsymbol{x}; q)$, valid at the bottom of the loop in Algorithm 2, is:

$$\log\,\underline{p}(\boldsymbol{x}; q) = \tfrac{1}{2} + \log(2) + n\log\{\tau(1-\tau)\} - \frac{\sum_{i=1}^n\{1/\mu_{q(a_i)}\}}{8\tau(1-\tau)} + \tfrac{1}{2}\log(\sigma^2_{q(\mu)}/\sigma^2_\mu)$$
$$- \frac{(\mu_{q(\mu)} - \mu_\mu)^2 + \sigma^2_{q(\mu)}}{2\sigma^2_\mu} + A\log(B) - \log\Gamma(A) + \log\mathcal{J}^+(2A+n-1, C_2, C_3).$$

4.3 Skew Normal Model

A Bayesian Skew Normal model for a univariate random sample is

$$\begin{gathered} x_i|\,\mu,\sigma \overset{\text{ind.}}{\sim} \text{Skew-Normal}(\mu,\sigma,\lambda), \\ \mu \sim N(\mu_\mu, \sigma^2_\mu), \quad \sigma^2 \sim \text{Inverse-Gamma}(A,B), \quad \lambda \sim N(\mu_\lambda, \sigma^2_\lambda) \end{gathered} \tag{12}$$

where $\mu_\mu, \mu_\lambda \in \mathbb{R}$ and $A, B, \sigma^2_\mu, \sigma^2_\lambda > 0$ are hyperparameters. Model (12) is based on the version of the Skew Normal distribution used by Azzalini & Dalla Valle (1996).

Using Result 3 we can re-write model (12) as

$$x_i|a_i,\mu,\sigma,\lambda \overset{\text{ind.}}{\sim} N\left(\mu + \frac{\lambda|a_i|}{\sqrt{1+\lambda^2}}\,,\ \frac{\sigma^2}{1+\lambda^2}\right), \quad a_i \overset{\text{ind.}}{\sim} N(0,1),$$

$$\mu \sim N(\mu_\mu, \sigma^2_\mu), \quad \sigma^2 \sim \text{Inverse-Gamma}(A,B), \quad \lambda \sim N(\mu_\lambda, \sigma^2_\lambda).$$

For variational Bayesian inference we impose the product restriction

$$q(\mu,\sigma,\lambda,\boldsymbol{a}) = q(\mu)q(\sigma)q(\lambda)q(\boldsymbol{a}).$$

This leads to the following forms for the optimal densities:

$$\begin{aligned}
q^*(\mu) &\sim N(\mu_{q(\mu)}, \sigma^2_{q(\mu)})\\
q^*(\sigma) &= \frac{\sigma^{-(2A+n+1)}\exp(C_4/\sigma - C_5/\sigma^2)}{\mathcal{J}^+(2A+n-1, C_4, C_5)}\\
q^*(\lambda) &= \frac{(1+\lambda^2)^{n/2}\exp\left\{-C_6\,\lambda^2 + C_7\lambda\sqrt{1+\lambda^2} + (\mu_\lambda/\sigma^2_\lambda)\lambda\right\}}{\mathcal{G}(0, \frac{1}{2}n, C_6, C_7, (\mu_\lambda/\sigma^2_\lambda))}, \quad -\infty < \lambda < \infty\\
\text{and } q^*(a_i) &= \frac{\sqrt{1+\mu_{q(\lambda^2)}}\exp\left\{-\frac{1}{2}(1+\mu_{q(\lambda^2)})a_i^2 + C_{i8}|a_i|\right\}}{2(\Phi/\phi)(C_{i8}/\sqrt{1+\mu^2_{q(\lambda)}})}, \quad -\infty < a_i < \infty, \quad 1 \le i \le n.
\end{aligned}$$

The parameters are determined from Algorithm 3.

```
Initialize: $\mu_{q(\mu)} \in \mathbb{R}$ and $\sigma^2_{q(\mu)}, \mu_{q(1/\sigma)}, \mu_{q(1/\sigma^2)} > 0$.
Cycle:
    For $i = 1, \ldots, n$:
        $C_{i8} \leftarrow \mu_{q(1/\sigma)}\mu_{q(\lambda\sqrt{1+\lambda^2})}(x_i - \mu_{q(\mu)})$
        $\mu_{q(|a_i|)} \leftarrow \frac{C_{i8}}{1+\mu_{q(\lambda^2)}} + \frac{(\phi/\Phi)(C_{i8}/\sqrt{1+\mu_{q(\lambda^2)}})}{\sqrt{1+\mu_{q(\lambda^2)}}}$
        $\mu_{q(a_i^2)} \leftarrow \frac{1+\mu_{q(\lambda^2)}+C_{i8}^2}{(1+\mu_{q(\lambda^2)})^2} + \frac{C_{i8}(\phi/\Phi)(C_{i8}/\sqrt{1+\mu_{q(\lambda^2)}})}{(1+\mu_{q(\lambda^2)})\sqrt{1+\mu_{q(\lambda^2)}}}$
    $\sigma^2_{q(\mu)} \leftarrow \left\{\frac{1}{\sigma^2_\mu} + n\mu_{q(1/\sigma^2)}(1+\mu_{q(\lambda^2)})\right\}^{-1}$
    $\mu_{q(\mu)} \leftarrow \sigma^2_{q(\mu)}\left\{\frac{\mu_\mu}{\sigma^2_\mu} + n\mu_{q(1/\sigma^2)}(1+\mu_{q(\lambda^2)})\,\overline{x} - \mu_{q(1/\sigma)}\mu_{q(\lambda\sqrt{1+\lambda^2})}\sum_{i=1}^n \mu_{q(|a_i|)}\right\}$
    $C_4 \leftarrow \mu_{q(\lambda\sqrt{1+\lambda^2})}\sum_{i=1}^n \mu_{q(|a_i|)}(x_i - \mu_{q(\mu)})$
    $C_5 \leftarrow B + \frac{1}{2}(1+\mu_{q(\lambda^2)})\left\{\sum_{i=1}^n (x_i - \mu_{q(\mu)})^2 + n\sigma^2_{q(\mu)}\right\}$
    $\mu_{q(1/\sigma^2)} \leftarrow \frac{\mathcal{J}^+(2A+n+1,C_4,C_5)}{\mathcal{J}^+(2A+n-1,C_4,C_5)}$ ; $\mu_{q(1/\sigma)} \leftarrow \frac{\mathcal{J}^+(2A+n,C_4,C_5)}{\mathcal{J}^+(2A+n-1,C_4,C_5)}$
    $C_6 \leftarrow \mu_{q(1/\sigma^2)}\left\{\sum_{i=1}^n (x_i - \mu_{q(\mu)})^2 + n\sigma^2_{q(\mu)}\right\} + \sum_{i=1}^n \mu_{q(a_i^2)} + \frac{1}{\sigma^2_\lambda}$
    $C_7 \leftarrow \mu_{q(1/\sigma)}\sum_{i=1}^n \mu_{q(|a_i|)}(x_i - \mu_{q(\mu)})$.
    $\mu_{q(\lambda^2)} \leftarrow \frac{\mathcal{G}(2, \frac{1}{2}n, \frac{1}{2}C_6, C_7, (\mu_\lambda/\sigma^2_\lambda))}{\mathcal{G}(0, \frac{1}{2}n, \frac{1}{2}C_6, C_7, (\mu_\lambda/\sigma^2_\lambda))}$ ; $\mu_{q(\lambda\sqrt{1+\lambda^2})} \leftarrow \frac{\mathcal{G}(1, \frac{1}{2}(n+1), \frac{1}{2}C_6, C_7, (\mu_\lambda/\sigma^2_\lambda))}{\mathcal{G}(0, \frac{1}{2}n, \frac{1}{2}C_6, C_7, (\mu_\lambda/\sigma^2_\lambda))}$
until the increase in $\underline{p}(\boldsymbol{x}; q)$ is negligible.
```

Algorithm 3: *Iterative scheme for obtaining the parameters in the optimal densities $q^*_{\boldsymbol{a}}$, $q^*(\mu)$, $q^*(\sigma)$ and $q^*(\lambda)$ for the Skew Normal model.*

Note the simplified expression for use in Algorithm 3:

$$\begin{aligned}\log \underline{p}(\boldsymbol{x};q) \;=\;& \tfrac{1}{2} + n\log(2) - (n+\tfrac{1}{2})\log(2\pi) + A\log(B) - \log\Gamma(A)\\ & -\frac{(\mu_{q(\mu)}-\mu_\mu)^2+\sigma^2_{q(\mu)}}{2\sigma^2_\mu} + \tfrac{1}{2}\log(\sigma^2_{q(\mu)}/\sigma^2_\mu) - \tfrac{1}{2}\log(\sigma^2_\lambda) - \frac{\mu_\lambda^2}{2\sigma^2_\lambda}\\ & +\tfrac{1}{2}\mu_{q(\lambda^2)}\left[\mu_{q(1/\sigma^2)}\left\{\sum_{i=1}^n (x_i-\mu_{q(\mu)})^2 + n\sigma^2_{q(\mu)}\right\} + \sum_{i=1}^n \mu_{q(a_i^2)}\right]\\ & +\log\mathcal{G}(0,\tfrac{1}{2}n,\tfrac{1}{2}C_6,C_7,(\mu_\lambda/\sigma^2_\lambda)) + \log\mathcal{J}^+(2A+n-1,C_4,C_5)\\ & +\sum_{i=1}^n \Phi(C_{i8}/\sqrt{1+\mu_{q(\lambda^2)}}).\end{aligned}$$

4.4 Finite Normal Mixture Model

Consider the model

$$\begin{gathered} x_i|\,\mu,\sigma \overset{\text{ind.}}{\sim} \text{Normal-Mixture}(\mu,\sigma;\boldsymbol{w},\boldsymbol{m},\boldsymbol{s}),\\ \mu\sim N(\mu_\mu,\sigma^2_\mu),\quad \sigma^2\sim\text{Inverse-Gamma}(A,B)\end{gathered}\tag{13}$$

where μ_μ and $A, B, \sigma^2_\mu > 0$ are hyperparameters. Model (13) is not of great interest in its own right. However, as illustrated in Section 4.5, it becomes relevant when a troublesome response variable density function is replaced by an accurate normal mixture approximation.

Using Result 4 we can rewrite model (13) as

$$p(\boldsymbol{x}|\,\mu,\sigma,\boldsymbol{a}_i) = \textstyle\prod_{i=1}^n\prod_{k=1}^K \left[\sigma^{-1}(2\pi s_k^2)^{-1/2}\exp\left\{-\tfrac{1}{2}(\tfrac{x_i-\mu}{\sigma}-m_k)^2/s_k^2\right\}\right]^{a_{ik}},$$

$$\boldsymbol{a}_i \overset{\text{ind.}}{\sim} \text{Multinomial}(1;\boldsymbol{w})\quad \mu\sim N(\mu_\mu,\sigma^2_\mu),\quad \sigma^2\sim\text{Inverse-Gamma}(A,B)$$

and a_{ik} denotes the kth entry of $\boldsymbol{a}_i$. The auxiliary random vectors $\boldsymbol{a}_i$, $1\le i\le n$, facilitate more tractable variational Bayes calculations as is apparent from the derivations given in Section A.4 of Appendix A. Under the product restriction

$$q(\mu,\sigma,\boldsymbol{a}) = q(\mu)q(\sigma)q(\boldsymbol{a})$$

the optimal densities take the form:

$$\begin{aligned} q^*(\mu) &\sim N(\mu_{q(\mu)},\sigma^2_{q(\mu)}),\\ q^*(\sigma) &= \frac{\sigma^{-2A-n-1}\exp(C_9/\sigma - C_{10}/\sigma^2)}{\mathcal{J}^+(2A+n-1,C_9,C_{10})},\quad \sigma>0,\\ \text{and } q^*(\boldsymbol{a}_i) &\overset{\text{ind.}}{\sim} \text{Multinomial}(1;\boldsymbol{\mu}_{q(\boldsymbol{a}_i)}).\end{aligned}$$

The parameters are determined from Algorithm 4.

An explicit expression for $\log \underline{p}(\boldsymbol{x};q)$ is:

$$\begin{aligned}\log \underline{p}(\boldsymbol{x};q) \;=\;& \tfrac{1}{2} - \tfrac{n}{2}\log(2\pi) + \log(2) + A\log(B) - \log\Gamma(A) + \log\mathcal{J}^+(2A+n-1,C_9,C_{10},0)\\ & +\sum_{k=1}^K \mu_{q(a_{\bullet k})}\{\log(w_k) - \tfrac{1}{2}\log(s_k^2) - \tfrac{1}{2}(m_k^2/s_k^2)\}\\ & +\tfrac{1}{2}\log\left(\frac{\sigma^2_{q(\mu)}}{\sigma^2_\mu}\right) - \frac{(\mu_{q(\mu)}-\mu_\mu)^2+\sigma^2_{q(\mu)}}{2\sigma^2_\mu} - \sum_{i=1}^n\sum_{k=1}^K \mu_{q(a_{ik})}\log(\mu_{q(a_{ik})}).\end{aligned}$$

```
Initialize: $\mu_{q(\mu)} \in \mathbb{R}$ and $\sigma^2_{q(\mu)}, \mu_{q(1/\sigma)}, \mu_{q(1/\sigma^2)} > 0$
Cycle:
    For $i = 1, \ldots, n$, $k = 1, \ldots, K$:
        $\nu_{ik} \leftarrow \log(w_k) - \tfrac{1}{2}\,\log(s_k^2) - \frac{1}{2s_k^2}\Big[\mu_{q(1/\sigma^2)}\{(x_i - \mu_{q(\mu)})^2 + \sigma^2_{q(\mu)}\}$
              $-2\mu_{q(1/\sigma)}\, m_k(x_i - \mu_{q(\mu)}) + m_k^2\Big]$
    For $i = 1, \ldots, n$, $k = 1, \ldots, K$:  $\mu_{q(a_{ik})} \leftarrow \exp(\nu_{ik}) \Big/ \sum_{k=1}^K \exp(\nu_{ik})$
    For $k = 1, \ldots, K$:  $\mu_{q(a_{\bullet k})} \leftarrow \sum_{i=1}^n \mu_{q(a_{ik})}$
    $\sigma^2_{q(\mu)} \leftarrow \Big(1/\sigma^2_\mu + \mu_{q(1/\sigma^2)} \sum_{k=1}^K \frac{\mu_{q(a_{\bullet k})}}{s_k^2}\Big)^{-1}$
    $\mu_{q(\mu)} \leftarrow \sigma^2_{q(\mu)} \Big\{\mu_{q(1/\sigma^2)} \sum_{i=1}^n \sum_{k=1}^K \frac{\mu_{q(a_{ik})} x_i}{s_k^2} - \mu_{q(1/\sigma)} \sum_{k=1}^K \frac{\mu_{q(a_{\bullet k})}\, m_k}{s_k^2} + \frac{\mu_\mu}{\sigma^2_\mu}\Big\}$
    $C_9 \leftarrow \sum_{i=1}^n \sum_{k=1}^K \frac{\mu_{q(a_{ik})} m_k (x_i - \mu_{q(\mu)})}{s_k^2}$
    $C_{10} \leftarrow B + \tfrac{1}{2} \sum_{i=1}^n \sum_{k=1}^K \frac{\mu_{q(a_{ik})}\{(x_i - \mu_{q(\mu)})^2 + \sigma^2_{q(\mu)}\}}{s_k^2}$
    $\mu_{q(1/\sigma^2)} \leftarrow \frac{\mathcal{J}^+(2A + n + 1, C_9, C_{10})}{\mathcal{J}^+(2A + n - 1, C_9, C_{10})}$ ; $\mu_{q(1/\sigma)} \leftarrow \frac{\mathcal{J}^+(2A + n, C_9, C_{10})}{\mathcal{J}^+(2A + n - 1, C_9, C_{10})}$
until the increase in $\underline{p}(\boldsymbol{x}; q)$ is negligible.
```

Algorithm 4: *Iterative scheme for obtaining the parameters in the optimal densities $q^*(\boldsymbol{a})$, $q^*(\mu)$ and $q^*(\sigma)$ for the Finite Normal Mixture model.*

4.5 Generalized Extreme Value Model

Now consider the case where f is the standard Generalized Extreme Value density function with shape parameter $-\infty < \xi < \infty, \xi \neq 0$:

$$f(x;\xi) = (1+\xi\,x)^{-1/\xi-1}e^{-(1+\xi\,x)^{-1/\xi}},\ 1+\xi\,x > 0.$$

Letting $\xi \to 0$ results in the standard Gumbel density

$$f(x;0) = \exp(-x - e^{-x}), \quad -\infty < x < \infty.$$

Direct variational Bayes is problematic for the location-scale model (6) when f is $\text{GEV}(0,1,\xi)$, since the likelihood induced by $f(;\xi)$ has complicated dependence on the parameters. A reasonable way out is to work with normal mixture approximations to the $f(\cdot;\xi)$:

$$f(x;\xi) \approx \sum_{k=1}^{K} \frac{w_{k,\xi}}{s_{k,\xi}}\,\phi\left(\frac{x-m_{k,\xi}}{s_{k,\xi}}\right). \tag{14}$$

Approximations for $f(x;0)$ have been employed successfully by Frühwirth-Schnatter & Wagner (2006) for Markov chain Monte Carlo-based inference. A number of extensions of this work now exist, including Frühwirth-Schnatter *et al.* (2009) and Nakajima *et al.* (2009). In Appendix C we describe normal mixture approximation for other members of the $\text{GEV}(0,1,\xi)$ family of density functions.

Let Ξ be a finite parameter space for the ξ parameter and consider the univariate GEV location-scale model:

$$\begin{gathered} x_i|\,\mu,\sigma \overset{\text{ind.}}{\sim} \text{GEV}(\mu,\sigma,\xi), \\ \mu \sim N(\mu_\mu,\sigma_\mu^2), \quad \sigma^2 \sim \text{Inverse-Gamma}(A,B), \quad \xi \sim p(\xi) \end{gathered} \tag{15}$$

where $\mu_\mu \in \mathbb{R}$ and $A, B, \sigma_\mu^2 > 0$ are hyperparameters and $p(\xi)$ is a fixed probability mass function over $\xi \in \Xi$.

For any fixed $\xi \in \Xi$, suppose we have a normal mixture approximation to $f(\cdot;\xi)$. Then we can use Algorithm 4 to obtain variational approximations, with the restriction $q(\mu,\sigma) = q(\mu)q(\sigma)$, to the *conditional* posterior densities $p(\mu|\boldsymbol{x},\xi)$ and $p(\sigma|\boldsymbol{x},\xi)$. Let these approximations be denoted by $q^*(\mu|\xi)$ and $q^*(\sigma|\xi)$, respectively. From the relationships

$$p(\mu|\boldsymbol{x}) = \sum_{\xi\in\Xi} p(\xi|\boldsymbol{x})p(\mu|\boldsymbol{x},\xi) \quad \text{and} \quad p(\sigma|\boldsymbol{x}) = \sum_{\xi\in\Xi} p(\xi|\boldsymbol{x})p(\sigma|\boldsymbol{x},\xi)$$

it follows that

$$p(\mu|\boldsymbol{x}) \approx \sum_{\xi\in\Xi} p(\xi|\boldsymbol{x})q^*(\mu|\xi) \quad \text{and} \quad p(\sigma|\boldsymbol{x}) \approx \sum_{\xi\in\Xi} p(\xi|\boldsymbol{x})q^*(\sigma|\boldsymbol{x},\xi). \tag{16}$$

The posterior probability mass function for ξ can be approximated by noting the relationship

$$p(\xi|\boldsymbol{x}) = \frac{p(\xi)p(\boldsymbol{x}|\xi)}{\sum_{\xi'\in\Xi} p(\xi')p(\boldsymbol{x}|\xi)}$$

and plugging in the marginal likelihood approximations $\underline{p}(\boldsymbol{x}|\xi)$, obtained from Algorithm 4 for each fixed $\xi \in \Xi$. This leads to

$$p(\xi|\boldsymbol{x}) \approx q^*(\xi) \equiv \frac{p(\xi)\underline{p}(\boldsymbol{x}|\xi)}{\sum_{\xi'\in\Xi} p(\xi')\underline{p}(\boldsymbol{x}|\xi)}.$$

In view of (16), appropriate variational Bayes approximations to $p(\mu|\boldsymbol{x})$ and $p(\sigma|\boldsymbol{x})$ are then

$$q^*(\mu) \equiv \sum_{\xi\in\Xi} q^*(\xi)q^*(\mu|\xi) \quad \text{and} \quad q^*(\sigma) \equiv \sum_{\xi\in\Xi} q^*(\xi)q^*(\sigma|\boldsymbol{x},\xi).$$

Finally, note that the overall marginal log-likelihood is approximated by

$$\underline{p}(\boldsymbol{y};q) \equiv \sum_{\xi\in\Xi} q^*(\xi)\underline{p}(\boldsymbol{x}|\xi).$$

Algorithm 5 summarizes this finite normal mixture approach to variational Bayesian inference for (μ,σ,ξ) in (15). The algorithm assumes that finite normal mixture approximations of the form (14) have been obtained for each $\xi\in\Xi$. Such calculations only need to be done once and can be stored in a look-up table. As described in Appendix C, we have done them for $\xi\in\{-1,-0.995,\ldots,0.995,1\}$ with $K=24$.

For each $\xi\in\Xi$:

1. Retrieve the normal mixture approximation vectors: $(w_{k,\xi},m_{k,\xi},s_{k,\xi})$, $1\le k\le K$, for approximation of the GEV$(0,1,\xi)$ density function.
2. Apply Algorithm 4 with (w_k,m_k,s_k) set to $(w_{k,\xi},m_{k,\xi},s_{k,\xi})$, $1\le k\le K$.
3. Store the parameters needed to define $q^*(\mu|\xi)$ and $q^*(\sigma|\xi)$.
4. Store the converged marginal likelihood lower bound $\underline{p}(\boldsymbol{x}|\xi)$.

Form the approximations to the posteriors $p(\xi|\boldsymbol{x})$, $p(\mu|\boldsymbol{x})$ and $p(\sigma|\boldsymbol{x})$ as follows:

$$q^*(\xi)=\frac{p(\xi)\underline{p}(\boldsymbol{x}|\xi)}{\sum_{\xi'\in\Xi}p(\xi')\underline{p}(\boldsymbol{x}|\xi)},\quad q^*(\mu)=\sum_{\xi\in\Xi}q^*(\xi)q^*(\mu|\xi),\quad q^*(\sigma)=\sum_{\xi\in\Xi}q^*(\xi)q^*(\sigma|\xi).$$

Algorithm 5: *Scheme for approximation of the posteriors $p(\xi|\boldsymbol{x})$, $p(\mu|\boldsymbol{x})$ and $p(\sigma|\boldsymbol{x})$ for the Generalized Extreme Value model.*

4.6 General Univariate Location-Scale Models

As demonstrated in the previous section for the GEV univariate location-scale model, the auxiliary normal mixture approach offers itself as a viable 'last resort' for troublesome density functions. Provided f in (6) is reasonably smooth, one can approximate it arbitrarily well by a finite normal mixture and then use Algorithm 4. If additional parameters are present, such as the GEV shape parameter ξ, then there is the option of imposing a discrete finite prior and using the approach exemplified by Algorithm 5.

Hence, the auxiliary mixture approach can be used for variational Bayesian inference for general univariate location-scale models.

5 Alternative Scale Parameter Priors

The Inverse Gamma distribution is the conjugate family for variance parameters in Normal mean-scale models. Since, after the introduction of auxiliary variables, many of the models in Section 4 involve Normal distributions the conjugacy property helps reduce the number of non-analytic forms. However, alternative scale parameter priors are often desirable. Gelman (2006) argues that Half t densities are better for achieving non-informativity of scale parameters, and pays particular attention to Half Cauchy scale priors. The Bayesian variable selection models of Cottet, Kohn & Nott (2008) use Log Normal priors for scale parameters. In this section we briefly describe the handling of these alternative scale parameter priors in variational Bayesian inference.

5.1 Half-Cauchy Prior

Consider the following alternative to (8):

$$x_i|\,\mu,\sigma \overset{\text{ind.}}{\sim} t(\mu,\sigma,\nu),$$

$$\mu \sim N(\mu_\mu,\sigma_\mu^2),\quad \sigma \sim \text{Half-Cauchy}(A),\quad \nu \sim \text{Uniform}(\nu_{\text{min}},\nu_{\text{max}})$$

where $\mu_\mu \in \mathbb{R}$ and $A,\nu_{\text{min}},\nu_{\text{max}},\sigma_\mu^2 > 0$ are hyperparameters. The only difference is $\sigma^2 \sim$ Inverse-Gamma(A,B) is replaced with $\sigma \sim$ Half-Cauchy(A). As before, we introduce auxiliary variables of the form $a_i|\nu \overset{\text{ind.}}{\sim}$ Inverse-Gamma$(\frac{\nu}{2},\frac{\nu}{2})$, which allows us to re-write the data model as

$$x_i|a_i,\mu,\sigma \overset{\text{ind.}}{\sim} N(\mu,a_i\,\sigma^2).$$

The optimal q densities are the same as (9), but with

$$q^*(\sigma) = \frac{\exp(-C_{11}/\sigma^2)}{\mathcal{H}(n-2,C_{11},A^2)\sigma^n(A^2+\sigma^2)},\quad \sigma > 0,$$

where $\mathcal{H}(p,q,r)$ is as defined in Section 2.1.

The optimal parameters can be obtained using an iterative algorithm similar to Algorithm 1. The only change is that

$$B_{q(\sigma^2)} \leftarrow B + \tfrac{1}{2}\sum_{i=1}^n \mu_{q(1/a_i)}\{(x_i-\mu_{q(\mu)})^2+\sigma^2_{q(\mu)}\}\ \ ;\ \ \mu_{q(1/\sigma^2)} \leftarrow \frac{A+\frac{n}{2}}{B_{q(\sigma^2)}}$$

is replaced with

$$C_{11} \leftarrow \tfrac{1}{2}\sum_{i=1}^n \mu_{q(1/a_i)}\left\{(x_i-\mu_{q(\mu)})^2+\sigma_{q(\mu)}\right\}\quad ;\quad \mu_{q(1/\sigma^2)} \leftarrow \frac{\mathcal{H}(n,C_{11},A^2)}{\mathcal{H}(n-2,C_{11},A^2)}.$$

The expression for $\log\,\underline{p}(\boldsymbol{x};q)$ becomes:

$$\begin{aligned}\log\,\underline{p}(\boldsymbol{x};q) \quad = \quad & \tfrac{n+1}{2} + \tfrac{n}{2}\mu_{q(\nu)} - \tfrac{n}{2}\log(2\pi) + \tfrac{1}{2}\log(\sigma^2_{q(\mu)}/\sigma^2_\mu) - \frac{(\mu_{q(\mu)}-\mu_\mu)^2+\sigma^2_{q(\mu)}}{2\sigma^2_\mu}\\ & +\log(2A/\pi) + \log\mathcal{H}(n-2,C_{11},A^2) + \log\mathcal{F}(0,n,C_1,\nu_{\text{min}},\nu_{\text{max}})\\ & -\log(\nu_{\text{max}}-\nu_{\text{min}}) + n\log\Gamma(\tfrac{1}{2}(\mu_{q(\nu)}+1))\\ & -\tfrac{n}{2}(\mu_{q(\nu)}+1)\,\text{digamma}\{\tfrac{1}{2}(\mu_{q(\nu)}+1)\}.\end{aligned}$$

5.2 Log Normal Prior

Next, consider the following alternative to (8):

$$x_i|\,\mu,\sigma \overset{\text{ind.}}{\sim} t(\mu,\sigma,\nu),$$

$$\mu \sim N(\mu_\mu,\sigma_\mu^2),\quad \sigma \sim \text{Log-Normal}(A,B),\quad \nu \sim \text{Uniform}(\nu_{\text{min}},\nu_{\text{max}})$$

where $\mu_\mu \in \mathbb{R}$ and $A,B,\nu_{\text{min}},\nu_{\text{max}},\sigma_\mu^2 > 0$ are hyperparameters. Once again, we introduce auxiliary variables $a_i|\nu \overset{\text{ind.}}{\sim}$ Inverse-Gamma$(\frac{\nu}{2},\frac{\nu}{2})$, and work with

$$x_i|a_i,\mu,\sigma \overset{\text{ind.}}{\sim} N(\mu,a_i\,\sigma^2).$$

The optimal q densities are the same as (9), but with

$$q^*(\sigma) = \frac{2\,\sigma^{(A/B^2)-n-1}\exp\{-C_6/\sigma^2-(\log\sigma)^2/(2B^2)\}}{\mathcal{J}(0,\frac{A}{2B^2}-\frac{n}{2},\frac{1}{8B^2},C_6)},\quad \sigma>0.$$

The optimal parameters can be obtained using an iterative algorithm similar to Algorithm 1. The only change is that

$$B_{q(\sigma^2)} \leftarrow B+\tfrac{1}{2}\sum_{i=1}^{n}\mu_{q(1/a_i)}\{(x_i-\mu_{q(\mu)})^2+\sigma^2_{q(\mu)}\}\ \ ;\ \ \mu_{q(1/\sigma^2)} \leftarrow \frac{A+\frac{n}{2}}{B_{q(\sigma^2)}}$$

is replaced with

$$C_{11} \leftarrow \tfrac{1}{2}\sum_{i=1}^{n}\mu_{q(1/a_i)}\left\{(x_i-\mu_{q(\mu)})^2+\sigma_{q(\mu)}\right\}\quad;\quad \mu_{q(1/\sigma^2)} \leftarrow \frac{\mathcal{J}(0,\frac{A}{2B^2}-1-\frac{n}{2},\frac{1}{8B^2},C_6)}{\mathcal{J}(0,\frac{A}{2B^2}-\frac{n}{2},\frac{1}{8B^2},C_6)}.$$

The expression for $\log\,\underline{p}(\boldsymbol{x};q)$ becomes:

$$\begin{aligned}\log\,\underline{p}(\boldsymbol{x};q) \quad = \quad & \tfrac{n+1}{2}+\tfrac{n}{2}\mu_{q(\nu)}-\tfrac{n}{2}\log(2\pi)+\tfrac{1}{2}\log(\sigma^2_{q(\mu)}/\sigma^2_{\mu})-\frac{(\mu_{q(\mu)}-\mu_{\mu})^2+\sigma^2_{q(\mu)}}{2\sigma^2_{\mu}}\\ & -\tfrac{1}{2}\log(2\pi)-\tfrac{1}{2}(A^2/B^2)-\log(B)+\log\mathcal{J}(0,\tfrac{A}{2B^2}-\tfrac{n}{2},\tfrac{1}{8B^2},C_6)\\ & +\log\mathcal{F}(0,n,C_1,\nu_{\text{min}},\nu_{\text{max}})-\log(\nu_{\text{max}}-\nu_{\text{min}})\\ & +n\log\Gamma(\tfrac{1}{2}(\mu_{q(\nu)}+1))-\tfrac{n}{2}(\mu_{q(\nu)}+1)\,\text{digamma}\{\tfrac{1}{2}(\mu_{q(\nu)}+1)\}.\end{aligned}$$

6 Multiparameter Extensions

Up until we have restricted attention to univariate models. This has the advantage that the various issues that arise with elaborate distributions in variational Bayes can be addressed with minimal notational effort. The localness property of variational Bayes means that the non-analytic forms that were identified in Sections 4 and 5 still apply for larger models. In this section we examine the most common multiparameter extension: from univariate models to regression models. For shape parameters such as ν, the t distribution degrees of freedom, this extension has no impact on the updates. The scale parameter updates are only mildly impacted. The location parameter μ is replaced by a vector of regression coefficients $\boldsymbol{\beta}$. Algebraically, this involves replacement of

$$\mathbf{1}\mu \quad \text{by} \quad \boldsymbol{X\beta}$$

in the model specification. The updates for $\boldsymbol{\beta}$ then involve matrix algebraic counterparts of $\mu_{q(\mu)}$ and $\sigma^2_{q(\mu)}$. We we will provide details on this extension for the t-distribution model with Inverse Gamma priors. Extensions for other models are similar.

A Bayesian t regression model (e.g. Lange, Little & Taylor, 1989) is

$$\begin{gathered} y_i|\,\boldsymbol{\beta},\sigma \overset{\text{ind.}}{\sim} t((\boldsymbol{X\beta})_i,\sigma,\nu),\\ \boldsymbol{\beta}\sim N(\boldsymbol{\mu_\beta},\boldsymbol{\Sigma_\beta}),\quad \sigma^2\sim\text{Inverse-Gamma}(A,B),\quad \nu\sim\text{Uniform}(\nu_{\text{min}},\nu_{\text{max}})\end{gathered}\tag{17}$$

where $\boldsymbol{\mu_\beta}$ and $\boldsymbol{\Sigma_\beta}$ hyperparameters for $\boldsymbol{\beta}$. We can re-write (17) as

$$y_i|a_i,\mu,\sigma \overset{\text{ind.}}{\sim} N((\boldsymbol{X\beta})_i,a_i\,\sigma^2),\quad a_i|\nu \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(\tfrac{\nu}{2},\tfrac{\nu}{2})$$

$$\boldsymbol{\beta}\sim N(\boldsymbol{\mu_\beta},\boldsymbol{\Sigma_\beta}),\quad \sigma^2\sim\text{Inverse-Gamma}(A,B),\quad \nu\sim\text{Uniform}(\nu_{\text{min}},\nu_{\text{max}}).$$

For variational Bayesian inference we impose the product restriction

$$q(\boldsymbol{\beta},\sigma,\nu,\boldsymbol{a}) = q(\boldsymbol{\beta},\nu)q(\sigma)q(\boldsymbol{a}).$$

This yields the following forms for the optimal densities:

$$\begin{aligned}
&q^*(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu}_{q(\boldsymbol{\beta})},\boldsymbol{\Sigma}_{q(\boldsymbol{\mu})}),\\
&q^*(\sigma^2) \sim \text{Inverse-Gamma}\Big(A+\tfrac{n}{2},\\
&\qquad\qquad B+\tfrac{1}{2}\,\mu_{q(1/\sigma^2)}\Big[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})^T\boldsymbol{C}_{12}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})+\text{tr}\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{X}^T\boldsymbol{C}_{12}\boldsymbol{X}\}\Big]\Big),\\
&q^*(a_i) \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\tfrac{\mu_{q(\nu)}+1}{2},\tfrac{1}{2}\Big[\mu_{q(\nu)}+\mu_{q(1/\sigma^2)}\{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})_i^2+(\boldsymbol{X}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{X}^T)_{ii}\}\Big]\right)\\
&\text{and } q^*(\nu) = \frac{\exp\Big[n\left\{\frac{\nu}{2}\log(\nu/2)-\log\Gamma(\nu/2)\right\}-(\nu/2)C_1\Big]}{\mathcal{F}(0,n,C_1,\nu_{\text{min}},\nu_{\text{max}})},\quad \nu_{\text{min}}<\nu<\nu_{\text{max}}.
\end{aligned}\tag{18}$$

The last density uses the same definition for C_1 that was in the univariate case: $C_1 \equiv \sum_{i=1}^n\{\mu_{q(\log a_i)}+\mu_{q(1/a_i)}\}$. The parameters in (18) are determined from Algorithm 6.

```
Initialize: $\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \in \mathbb{R}^{k+1}$, $\mu_{q(\nu)} \in [\nu_{\text{min}},\nu_{\text{max}}]$ and $\mu_{q(1/\sigma^2)} > 0$.
Cycle:
    For $i = 1,\ldots,n$:
        $B_{q(a_i)} \leftarrow \frac{1}{2}\Big[\mu_{q(\nu)}+\mu_{q(1/\sigma^2)}\{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})_i^2+(\boldsymbol{X}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{X}^T)_{ii}\}\Big]$
        $\mu_{q(1/a_i)} \leftarrow \frac{1}{2}(\mu_{q(\nu)}+1)/B_{q(a_i)}$
        $\mu_{q(\log a_i)} \leftarrow \log(B_{q(a_i)}) - \text{digamma}(\frac{1}{2}(\mu_{q(\nu)}+1))$
    $\boldsymbol{C}_{12} \leftarrow \text{diag}_{1\le i\le n}\{\mu_{q(1/a_i)}\}$ ; $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \leftarrow \Big\{\mu_{q(1/\sigma^2)}\boldsymbol{X}^T\boldsymbol{C}_{12}\boldsymbol{X}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\Big\}^{-1}$
    $\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\Big\{\mu_{q(1/\sigma^2)}\boldsymbol{X}^T\boldsymbol{C}_{12}\boldsymbol{y}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\Big\}$
    $C_1 \leftarrow \sum_{i=1}^n\{\mu_{q(\log a_i)}+\mu_{q(1/a_i)}\}$ ; $\mu_{q(\nu)} \leftarrow \dfrac{\mathcal{F}(1,n,C_1,\nu_{\text{min}},\nu_{\text{max}})}{\mathcal{F}(0,n,C_1,\nu_{\text{min}},\nu_{\text{max}})}$
    $B_{q(\sigma^2)} \leftarrow B+\frac{1}{2}\,\mu_{q(1/\sigma^2)}\Big[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})^T\boldsymbol{C}_{12}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})+\text{tr}\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\boldsymbol{X}^T\boldsymbol{C}_{12}\boldsymbol{X}\}\Big]$
    $\mu_{q(1/\sigma^2)} \leftarrow (A+\frac{n}{2})/B_{q(\sigma^2)}$
until the increase in $\underline{p}(\boldsymbol{x};q)$ is negligible.
```

Algorithm 6: *Iterative scheme for obtaining the parameters in the optimal densities $q^*(\boldsymbol{a})$, $q^*(\boldsymbol{\beta})$, $q^*(\nu)$, $q^*(\sigma)$ for the t regression model.*

The lower bound on the marginal log-likelihood admits the expression:

$$\begin{aligned}
\log\,\underline{p}(\boldsymbol{x};q) \;=\;& \tfrac{n+k+1}{2}+\tfrac{n}{2}\mu_{q(\nu)}-\tfrac{n}{2}\log(2\pi)\\
&+\tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}|-\tfrac{1}{2}\Big\{(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}-\boldsymbol{\mu}_{\boldsymbol{\beta}})+\text{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\Big\}\\
&+A\log(B)-\log\Gamma(A)-(A+\tfrac{n}{2})\log(B_{q(\sigma^2)})+\log\Gamma(A+\tfrac{n}{2})\\
&+\log\mathcal{F}(0,n,C_1,\nu_{\text{min}},\nu_{\text{max}})-\log(\nu_{\text{max}}-\nu_{\text{min}})\\
&+n\log\Gamma(\tfrac{1}{2}(\mu_{q(\nu)}+1))-\tfrac{n}{2}(\mu_{q(\nu)}+1)\,\text{digamma}\{\tfrac{1}{2}(\mu_{q(\nu)}+1)\}.
\end{aligned}$$

7 Other Elaborate Response Models

Many other elaborate continuous response distributions could be entertained. Examples include Skew t (e.g. Azzalini & Capitanio, 2003), Generalized Inverse Gaussian and Generalized Hyperbolic distributions. There are also numerous elaborate distributions appropriate for discrete responses, such as Negative Binomial and Beta Binomial distributions. In the multiparameter case, corresponding to Bayesian generalized additive models, the link function also impacts tractability of variational Bayes schemes (e.g. Girolami & Rogers, 2006).

Clearly we cannot cover all possible elaborate response distributions. However, we note that the strategies used in Sections 4 to 6 involving judicious use of auxiliary variables, quadrature and finite mixture approximations apply generally. For example, equation (25) of Azzalini & Capitanio (2003) suggests a useful auxiliary variable representation for Skew t response models. As mentioned in Section 4.6, finite mixture approximation to the response density is always available as a last resort.

8 Accuracy Assessment

We conducted a simulation study to assess the accuracy of the univaraite location-scale variational Bayes algorithms described in Sections 4 and 5. One hundred random samples of size $n = 500$ were drawn from the t distribution, Asymmetric Laplace, Skew Normal and Generalized Extreme Value distributions. Without loss of generality we set the location and scale parameters to be $\mu = 0$ and $\sigma = 1$. The shape parameters were:

$$\begin{aligned}
& \nu = 1.5 \quad \text{for the } t\text{-distribution models,} \\
& \tau = 0.75 \quad \text{for the Asymmetric Laplace distribution model,} \\
& \lambda = 5 \quad \text{for the skew-Normal distribution model} \\
\text{and} \quad & \xi = 0.5 \quad \text{for the Generalized Extreme Value distribution model.}
\end{aligned}$$

The hyperparameters for μ were fixed at $\mu_\mu = 0$ and $\sigma^2_\mu = 10^8$. For Inverse Gamma priors on the squared scale we used $A = B = 0.01$. For the Half Cauchy prior on the scale we used $A = 25$ and for the Log Normal prior on the scale we used $A = 100$ and $B = 10$. Shape parameter hyperparameters were $\nu_{\text{min}} = 0.01$, $\nu_{\text{max}} = 100$, $\mu_\lambda = 0$ and $\sigma^2_\lambda = 10^8$. Finally, $p(\xi)$ was a uniform discrete distribution on $\Xi = \{0, 0.01, \ldots, 0.99, 1\}$.

The accuracy of variational Bayes approximate posterior density functions was measured via L_1 distance. Let θ be a generic parameter in any one of the models considered in Section 4 or 5. Then the L_1 error, or *integrated absolute error (IAE)* of q^*, given by

$$\text{IAE}(q^*) = \int_{-\infty}^{\infty} \big|\, q^*(\theta) - p(\theta|\boldsymbol{x}) \,\big|\, d\theta.$$

Note that L_1 error is a scale-independent number between 0 and 2 and is invariant to monotone transformations on the parameter θ (Devroye & Györfi, 1985). The latter property implies, for example, that the IAEs for $q^*(\sigma)$ and $q^*(\sigma^2)$ are identical. The *accuracy* of q^*

$$\text{accuracy}(q^*) = 1 - \{\text{IAE}(q^*)/ \sup_{q \text{ a density}} \text{IAE}(q)\} = 1 - \tfrac{1}{2}\text{IAE}(q^*). \tag{19}$$

Since $0 \leq \text{accuracy}(q^*) \leq 1$ we express this measure as a percentage in our accuracy assessments. Exact computation of $p(\theta|\boldsymbol{x})$ is difficult so we worked with MCMC samples obtained using BRugs (Ligges *et al.* 2010) with a burnin of size 10000. A thinning factor of 5 was applied to post-burnin samples of size 50000. This resulted in MCMC samples of size 10000 for density estimation. Density estimates were obtained using the binned kernel density estimate bkde() function in the R package KernSmooth (Wand & Ripley, 2009). The bandwidth was chosen using a direct plug-in rule, corresponding to the default version of the dpik() function in KernSmooth.

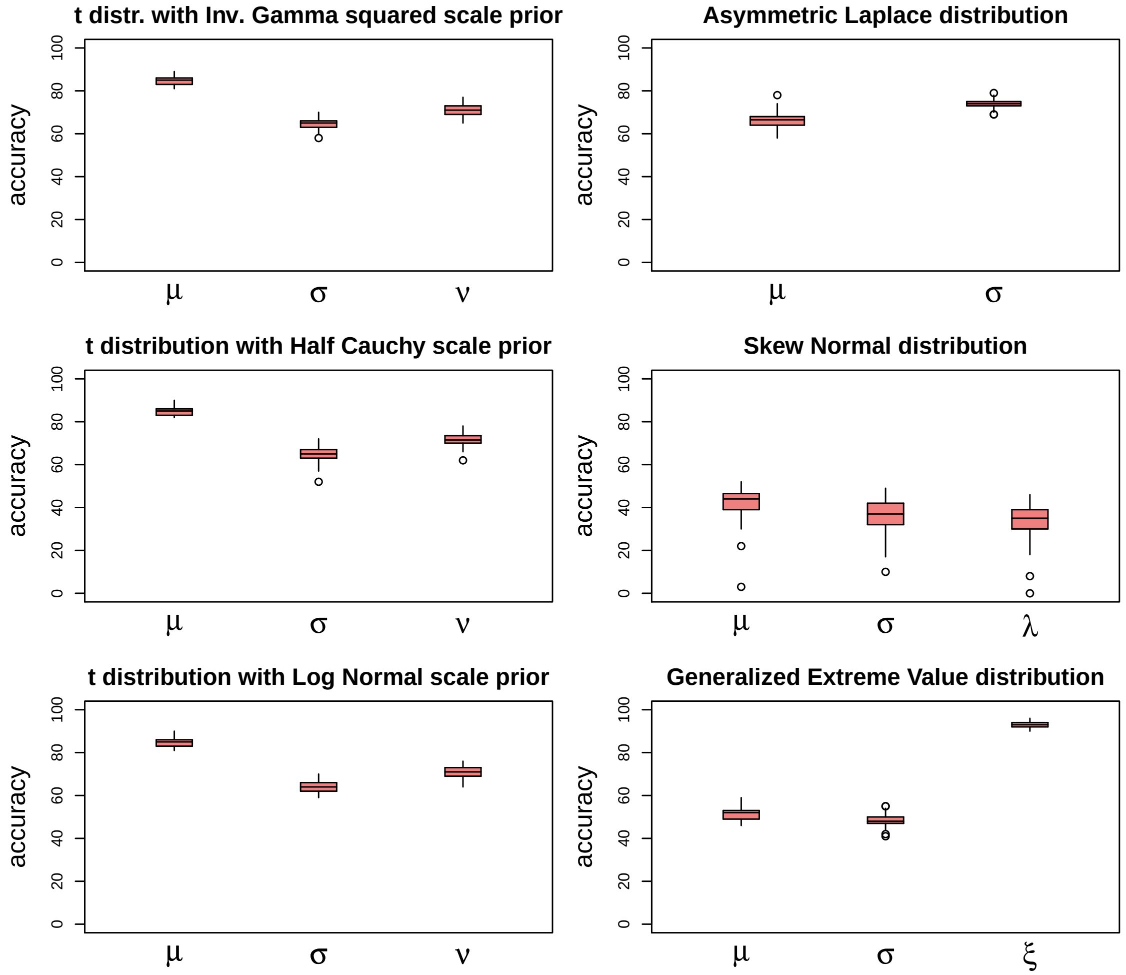

Figure 3: *Boxplots of accuracy measurements for the simulation study described in the text.*

Figure 3 summarizes the accuracy measures obtained from 100 replications of each of six models. The left-hand panels show the accuracy of variational Bayes for the three t models. The results are similar, regardless of form of the scale parameter prior. There is also very little between sample variability in the accuracy measures and, hence, we will simply quote average accuracy here. The location parameter μ has its posterior approximated with about 84% accuracy. For the degrees of freedom parameter ν the accuracy drops to about 71%, while it is only about 65% for the scale parameter σ. The results for the Asymmetric Laplace show an approximate reversal with the scale parameter having 74% accuracy, but the location parameter posterior at 66% accuracy. The accuracy values for the Skew Normal model are between 37% and 42% for the three model parameters μ, σ and λ, indicating that this distribution is particularly challenging for variational Bayes. For the Generalized Extreme Value model the location and scale have accuracy each around 50%. But the accuracy for the shape parameter ξ is excellent at 93%.

The nature of the inaccuracies is shown in Figure 4, in which the approximate densities are shown for the first replication of the simulation study. Since there is very little variability in the accuracies, these plots show typical performance. There is a pronounced tendency for the variational Bayes densities to be too narrow.

Figure 5 provides some insight into why variational Bayes is prone to inaccuracy for the models in Sections 4–6. It shows pairwise scatterplots of the MCMC output when

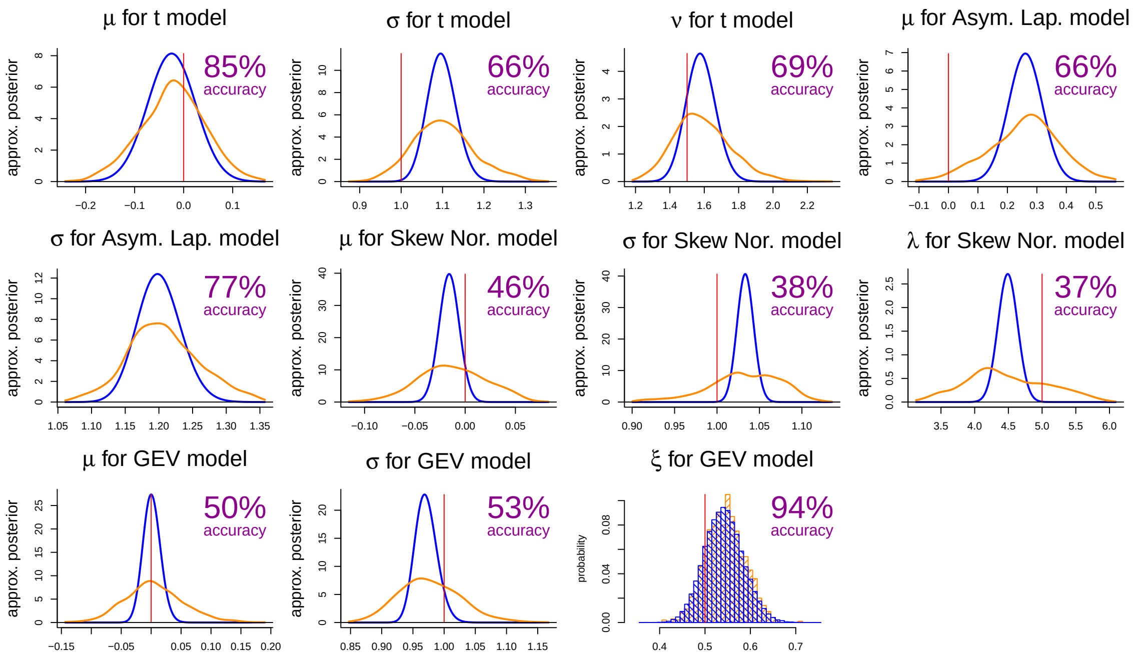

Figure 4: *Comparison of variational Bayes and 'exact' (MCMC-based) posterior density functions and probability mass function for several parameters from the simulation study. In each case, the approximate posterior densities are obtained from the first replication of the simulation study. For the density function comparisons, variational Bayes approximations are shown as blue curves and the 'exact' densities are shown as orange curves. Analogous colour-coding applies to the probability mass functions.*

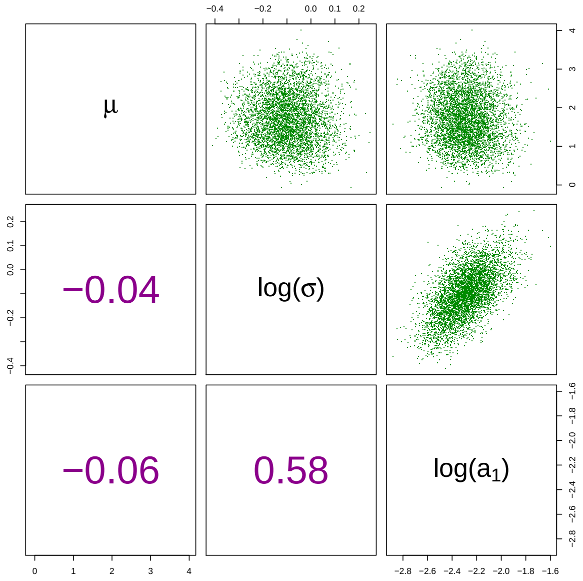

Figure 5: *Pairwise scatterplots and sample correlations of MCMC output for* μ, $\log(\sigma)$ *and* $\log(a_1)$ *when fitting a univariate asymmetric Laplace model to a sample of size* $n = 100$ *with shape parameter* $\tau = 0.95$. *The MCMC sample size is 5000.*

fitting the asymmetric Laplace model to a simulated random sample of size 100. The shape parameter was set at $\tau = 0.95$. It is apparent from Figure 5 that the posterior correlation between σ and a_1 is quite strong. The variational Bayes approximation with product restriction (11) ignores this dependence and, consequently, its accuracy suffers.

9 Application

Variational Bayes for elaborate distributions has enormous potential for use in applications. The localness property means that the results for the simple models in Sections 4–6 can be used in larger models tailored to the data at hand. In this section we provide a brief illustration: robust nonparametric regression based on the t-distribution for data from a respiratory health study. The data, shown in Figure 6, correspond to measurements on one subject during two separate respiratory experiments. The data are from a study conducted by Professor Russ Hauser at Harvard School of Public Health, Boston, USA. In each panel, the (x_i, y_i) predictor/response pairs are:

$$\begin{aligned} x_i &= \text{time in seconds since exposure to air containing particulate matter} \\ y_i &= \log(\text{adjusted time of exhalation}). \end{aligned}$$

The adjusted time of exhalation is obtained by subtracting the average time of exhalation at baseline, prior to exposure to filtered air. Interest centres upon the mean response as a function of the time, so an appropriate model is

$$y_i = f(x_i) + \varepsilon_i$$

However, the y_is contain outlying values due to an occasional cough or sporadic breath and it is appropriate to model the errors as $\varepsilon_i \overset{\text{ind.}}{\sim} t(0, \sigma_\varepsilon, \nu)$. A penalized spline model for f is

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

where $\{z_1(x), \ldots, z_K(x)\}$ is an appropriate set of spline basis functions (e.g. Wand & Ormerod, 2008). Staudenmayer, Lake & Wand (2009) considered a non-Bayesian version of this model and described fitting via an Expectation-Maximization (EM) algorithm. Here we consider the Bayesian hierarchical model

$$\begin{gathered} y_i|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon \overset{\text{ind.}}{\sim} t((\boldsymbol{X\beta} + \boldsymbol{Zu})_i, \sigma_\varepsilon, \nu), \quad \boldsymbol{u}|\sigma_u \overset{\text{ind.}}{\sim} N(\boldsymbol{0}, \sigma_u^2 \boldsymbol{I}) \\ \boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{I}), \quad \sigma_\varepsilon \sim \text{Half-Cauchy}(A_\varepsilon), \qquad (20) \\ \sigma_u \sim \text{Half-Cauchy}(A_u), \quad \nu \sim \text{Uniform}(\nu_{\text{min}}, \nu_{\text{max}}). \end{gathered}$$

where

$$\boldsymbol{X} = [1\ x_i]_{1 \le i \le n} \quad \text{and} \quad \boldsymbol{Z} = [z_1(x_i)\ \cdots\ z_K(x_i)]_{1 \le i \le n}.$$

We used the following product density assumption in our variational Bayes approximation:

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \nu, \sigma_u, \sigma_\varepsilon) = q(\boldsymbol{\beta}, \boldsymbol{u}, \nu) q(\sigma_u, \sigma_\varepsilon).$$

Up until now, variational Bayes fitting of (20) has been challenging due to the elaborate form of the response and the non-conjugate prior distributions on the standard deviation parameters. However, simple extension of the methodology in Sections 5.1 and 6 permits its fitting. In particular, all calculations are either analytic or involve members of the $\mathcal{F}(p, q, r, s, t)$ and $\mathcal{H}(p, q, r)$ integral families.

The hyperparameters are set at $\sigma_{\boldsymbol{\beta}}^2 = 10^8$, $A_u = A_\varepsilon = 25$, $\nu_{\text{min}} = 0.01$ and $\nu_{\text{max}} = 100$ with standardized versions of the (x_i, y_i) data used in the fitting. This imposes non-informativeness for all parameters (Gelman, 2006). The results were then transformed back to the original units.

Inspection of Figure 6 shows that the variational Bayes fits and pointwise 95% credible sets are quite close to those obtained using MCMC. This high accuracy is aligned with

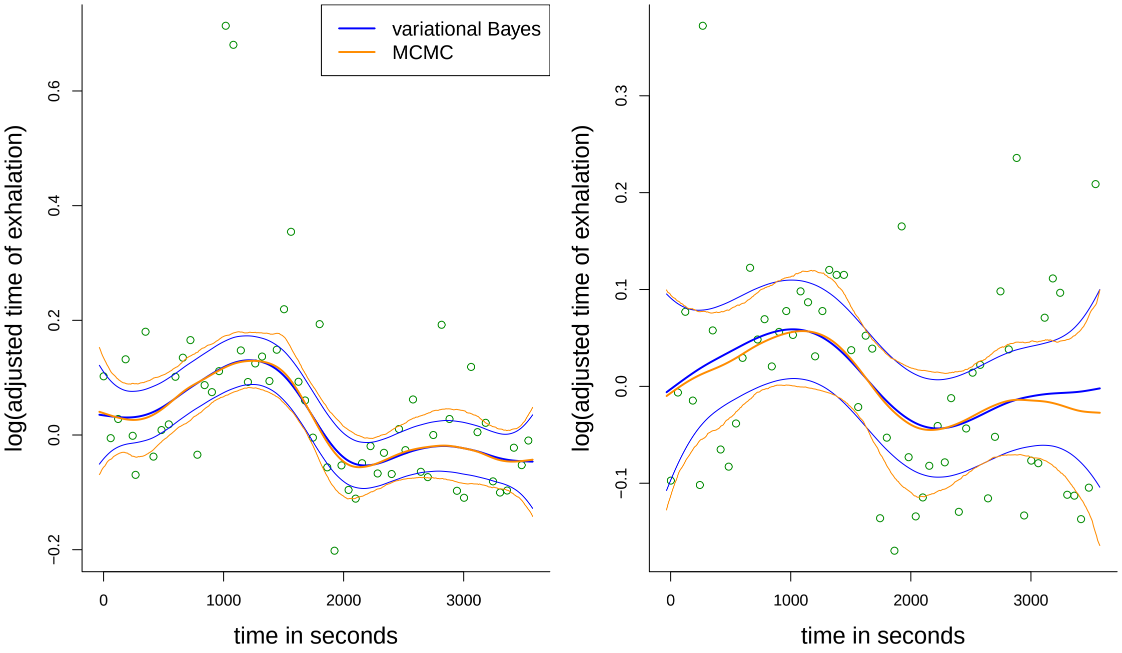

Figure 6: *Comparison of variational Bayes and MCMC fits for robust t-based nonparametric regression, corresponding to model 20, for data from the respiratory experiment described in the text.*

that exhibited by $q^*(\mu)$ for the univariate t-distribution model (upper left-hand panel of Figure 3). Staudenmayer *et al.* (2009) admit that EM-based fitting of these data requires several hours of computing time. The MCMC fits shown in Figure 6 took 3 minutes on the first author's laptop computer (Mac OS X; 2.33 GHz processor, 3 GBytes of random access memory). A simplistic R implementation of the variational Bayes approximation took about 15 seconds.

Closing Discussion

Variational Bayes provides an alternative to MCMC when speed is at a premium. In this article we have significantly enriched the class of models which can be handled via the variational Bayes paradigm. The numerical studies in Section 8 show that, as with simpler distributions, variational Bayes for elaborate distributions entails a loss in accuracy for the convenient product restrictions used in our illustrations. Yet to be explored are less stringent product restrictions for elaborate distribution models of the type considered in Sections 4–6. These promise higher accuracy, but at the expense of higher computational overhead.

Appendix A: Derivations

The derivations in make use of the following convenient shorthand. By 'rest' we mean all other random variables in the Baysian model at hand. Additive constants with respect to the function argument are denoted by 'const.'. The sample mean of $x_1, \ldots, x_n$ is denoted by $\overline{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$.

A.1. t Model

Each of the full conditional density functions satisfy:

$$\begin{aligned}
\log\, p(\mu|\text{rest}) &= -\tfrac{1}{2}\left[\left\{\frac{\sum_{i=1}^n(1/a_i)}{\sigma^2}+\frac{1}{\sigma_\mu^2}\right\}\mu^2-2\left\{\frac{\sum_{i=1}^n x_i/a_i}{\sigma^2}+\frac{\mu_\mu}{\sigma_\mu^2}\right\}\mu\right]+\text{const.},\\
\log\, p(\sigma^2|\text{rest}) &= -(A+\tfrac{1}{2}\,n)\log(\sigma^2)-\left\{B+\tfrac{1}{2}\sum_{i=1}^n\frac{(x_i-\mu)^2}{a_i}\right\}\Big/\sigma^2+\text{const.},\\
\log p(\nu|\text{rest}) &= n\{\tfrac{\nu}{2}\log(\nu/2)-\log\Gamma(\nu/2)\}\\
&\qquad -(\nu/2)\sum_{i=1}^n\{\log(a_i)+(1/a_i)\}+\text{const.},\quad \nu_{\text{min}}<\nu<\nu_{\text{max}},\\
\text{and }\ \log p(\boldsymbol{a}|\text{rest}) &= \sum_{i=1}^n\left[-\tfrac{1}{2}(\nu+1)\log(a_i)-\tfrac{1}{2}(1/a_i)\left\{\nu+\frac{(x_i-\mu)^2}{\sigma^2}\right\}\right]+\text{const.}
\end{aligned}$$

Then

$$\begin{aligned}
q^*(\mu) &\propto \exp\left\{E_{q(\sigma^2,\boldsymbol{a})}\log\, p(\mu|\text{rest})\right\}\\
&= \exp\left(-\tfrac{1}{2}\left[\left\{\mu_{q(1/\sigma^2)}\sum_{i=1}^n\mu_{q(1/a_i)}+\frac{1}{\sigma_\mu^2}\right\}\mu^2-2\left\{\mu_{q(1/\sigma^2)}\sum_{i=1}^n x_i\mu_{q(1/a_i)}+\frac{\mu_\mu}{\sigma_\mu^2}\right\}\mu\right]\right).
\end{aligned}$$

Standard manipulations lead to $q^*(\mu)$ being the $N(\mu_{q(\mu)},\sigma^2_{q(\mu)})$ density function, where

$$\sigma^2_{q(\mu)}=\left(\mu_{q(1/\sigma^2)}\sum_{i=1}^n\mu_{q(1/a_i)}+\frac{1}{\sigma_\mu^2}\right)^{-1}\quad\text{and}\quad \mu_{q(\mu)}=\sigma^2_{q(\mu)}\left(\mu_{q(1/\sigma^2)}\sum_{i=1}^n x_i\mu_{q(1/a_i)}+\frac{\mu_\mu}{\sigma_\mu^2}\right).$$

The derivations for $q^*(\sigma^2)$, $q^*(\nu)$ and $q^*(\boldsymbol{a})$ involve similar and standard manipulations.

A.2. Asymmetric Laplace Model

The full conditionals satisfy:

$$\begin{aligned}
\log p(\mu|\text{rest}) &= -\tfrac{1}{2}\left\{\frac{1}{\sigma_\mu^2}+\frac{\tau(1-\tau)\sum_{i=1}^n a_i}{\sigma^2}\right\}\mu^2+\left\{\frac{\mu_\mu}{\sigma_\mu^2}+\frac{\tau(1-\tau\sum_{i=1}^n a_ix_i)}{\sigma^2}+\frac{n(\tau-\frac{1}{2})}{\sigma}\right\}\mu\\
&\qquad +\text{const.},\\
\log p(\sigma|\text{rest}) &= -(2A+n+1)\log(\sigma)-\frac{1}{\sigma^2}\left(B+\tfrac{1}{2}\tau(1-\tau)\sum_{i=1}^n a_i(x_i-\mu)^2\right)\\
&\qquad +\frac{1}{\sigma}\,n(\overline{x}-\mu)(\tfrac{1}{2}-\tau)+\text{const.}\\
\text{and }\ \log p(\boldsymbol{a}|\text{rest}) &= \sum_{i=1}^n\left[-\tfrac{3}{2}\log(a_i)-\tfrac{1}{2}\left\{a_i\frac{(x_i-\mu)^2\tau(1-\tau)}{\sigma^2}+\frac{1}{a_i4\tau(1-\tau)}\right\}\right]+\text{const.}.
\end{aligned}$$

The derivation of $q^*(\mu)$ is similar to that given in Section A.2 for the t model. The optimal q-density for σ satisfies

$$q^*(\sigma)\propto\exp[E_{q(\mu,\boldsymbol{a})}\{p(\sigma|\text{rest})\}]=\exp(C_2/\sigma-C_3/\sigma^2),\quad \sigma>0,$$

where

$$C_2\equiv n(\overline{x}-\mu_{q(\mu)})(\tfrac{1}{2}-\tau)\quad\text{and}\quad C_3\equiv B+\tfrac{1}{2}\tau(1-\tau)\sum_{i=1}^n\mu_{q(a_i)}\{(x_i-\mu_{q(\mu)})^2+\sigma^2_{q(\mu)}\}.$$

Noting that

$$\int_0^\infty\sigma^{-(2A+n+1+k)}\exp(C_2/\sigma-C_3/\sigma^2)=\mathcal{J}^+(2A+n-k-1,C_2,C_3)$$

for each of $k \in \{-2,-1,0\}$ we get the normalizing factor for $q^*(\sigma)$ being $\mathcal{J}^+(2A+n-1,C_2,C_3)$ and the expressions for $\mu_{q(1/\sigma^2)}$ and $\mu_{q(1/\sigma)}$ appearing in Algorithm 2. Lastly,

$$q^*(\boldsymbol{a}) \propto \prod_{i=1}^n a_i^{-3/2} \exp\left[-\tfrac{1}{2}\left\{a_i\mu_{q(1/\sigma^2)}(x_i-\mu)^2\tau(1-\tau)+\frac{1}{a_i 4\tau(1-\tau)}\right\}\right], \quad a_i>0.$$

After a little algebra, it becomes clear that $q^*(\boldsymbol{a})$ is a product of Inverse Gaussian densities $q^*(a_i)$ with means

$$\mu_{q(a_i)} = \left[4\tau^2(1-\tau)^2\mu_{q(1/\sigma^2)}\{(x_i-\mu_{q(\mu)})^2+\sigma^2_{q(\mu)}\}\right]^{-1/2}, \quad 1\le i\le n,$$

and common precision parameter $\gamma_{q(a_i)} = \{4\tau(1-\tau)\}^{-1}$. The expression for $\mu_{q(1/a_i)}$ in Algorithm 2 follows from the expectation results (1).

A.3. Skew Normal Model

The full conditionals satisfy

$$\begin{aligned}
\log p(\mu|\text{rest}) &= -\tfrac{1}{2}\left\{\frac{1}{\sigma_\mu^2}+\frac{n(1+\lambda^2)}{\sigma^2}\right\}\mu^2+\left\{\frac{\mu_\mu}{\sigma_\mu^2}+\frac{n(1+\lambda^2)\overline{x}}{\sigma^2}-\frac{\lambda\sqrt{1+\lambda^2}\sum_{i=1}^n|a_i|}{\sigma}\right\}\mu+\text{const.},\\
\log p(\sigma|\text{rest}) &= -(2A+n+1)\log(\sigma)-\frac{1}{\sigma^2}\left(B+\tfrac{1}{2}(1+\lambda^2)\sum_{i=1}^n(x_i-\mu)^2\right)\\
&\qquad +\frac{1}{\sigma}\lambda\sqrt{1+\lambda^2}\sum_{i=1}^n|a_i|(x_i-\mu)+\text{const.},\\
\log p(\lambda|\text{rest}) &= \tfrac{n}{2}\log(1+\lambda^2)-\tfrac{1}{2}\left[\frac{1}{\sigma^2}\sum_{i=1}^n(x_i-\mu)^2+\sum_{i=1}^n a_i^2+\frac{1}{\sigma_\lambda^2}\right]\lambda^2\\
&\qquad +\frac{\lambda\sqrt{1+\lambda^2}}{\sigma}\sum_{i=1}^n|a_i|(x_i-\mu)+\frac{\mu_\lambda\lambda}{\sigma_\lambda^2}+\text{const.}\\
\text{and }\log p(\boldsymbol{a}|\text{rest}) &= \sum_{i=1}^n\left\{-\frac{(1+\lambda^2)a_i^2}{2}+\frac{\lambda\sqrt{1+\lambda^2}(x_i-\mu)|a_i|}{\sigma}\right\}+\text{const.}.
\end{aligned}$$

The derivation for $q^*(\mu)$ is similar to that given for each of the previous models. The derivation for $q^*(\sigma)$ is similar to that given for the Asymmetric Laplace model.

To obtain $q^*(\lambda)$ note that

$$\begin{aligned}
E_q\{\log p(\lambda|\text{rest})\} &= \tfrac{n}{2}\log(1+\lambda^2)-\tfrac{1}{2}\left[\mu_{q(1/\sigma^2)}\left\{\sum_{i=1}^n(x_i-\mu_{q(\mu)})^2+n\sigma^2_{q(\mu)}\right\}+\sum_{i=1}^n\mu_{q(a_i^2)}+\frac{1}{\sigma_\lambda^2}\right]\lambda^2\\
&\qquad +\left[\mu_{q(1/\sigma)}\sum_{i=1}^n\mu_{q(|a_i|)}(x_i-\mu_{q(\mu)})\right]\lambda\sqrt{1+\lambda^2}+\frac{\mu_\lambda\lambda}{\sigma_\lambda^2}+\text{const.}.
\end{aligned}$$

Hence

$$q^*(\lambda)\propto(1+\lambda^2)^{n/2}\exp\left\{-C_6\,\lambda^2+C_7\lambda\sqrt{1+\lambda^2}+(\mu_\lambda/\sigma_\lambda^2)\lambda\right\}, \quad -\infty<\lambda<\infty,$$

where

$$C_6\equiv\mu_{q(1/\sigma^2)}\left\{\sum_{i=1}^n(x_i-\mu_{q(\mu)})^2+n\sigma^2_{q(\mu)}\right\}+\sum_{i=1}^n\mu_{q(a_i^2)}+\frac{1}{\sigma_\lambda^2}\text{ and }C_7\equiv\mu_{q(1/\sigma)}\sum_{i=1}^n\mu_{q(|a_i|)}(x_i-\mu_{q(\mu)}).$$

The normalizing factor is

$$\int_{-\infty}^{\infty}(1{+}\lambda^2)^{n/2}\exp\left\{-C_6\,\lambda^2+C_7\lambda\sqrt{1+\lambda^2}+(\mu_\lambda/\sigma^2_\lambda)\lambda\right\}\,d\lambda=\mathcal{G}(0,\tfrac{1}{2}n,\tfrac{1}{2}C_6,C_7,(\mu_\lambda/\sigma^2_\lambda)).$$

The expressions for $\mu_{q(\lambda^2)}$ and $\mu_{q(\lambda\sqrt{1+\lambda^2})}$ involve similar manipulations. Finally,

$$E_{q(\mu,\sigma,\lambda)}\{\log p(\boldsymbol{a}|\text{rest})\} \;=\; \sum_{i=1}^{n}\left\{-\tfrac{1}{2}(1+\mu_{q(\lambda^2)})a_i^2+\mu_{q(1/\sigma)}\mu_{q(\lambda\sqrt{1+\lambda^2})}(x_i-\mu_{q(\mu)})|a_i|\right\}+\text{const.}.$$

Hence,

$$q^*(\boldsymbol{a})\propto\prod_{i=1}^{n}\exp\left\{-\tfrac{1}{2}(1+\mu_{q(\lambda^2)})a_i^2+C_{i8}|a_i|\right\},\quad -\infty<a_i<\infty,\quad 1\le i\le n,$$

where

$$C_{i8}\equiv\mu_{q(1/\sigma)}\mu_{q(\lambda\sqrt{1+\lambda^2})}(x_i-\mu_{q(\mu)}).$$

The normalizing factors and moment expressions follow from involve standard manipulations involving the normal density and cumulative distribution functions.

A.4. Finite Normal Mixture Model

The full conditionals satisfy:

$$\begin{aligned}\log\,p(\mu|\text{rest}) \;&=\; -\tfrac{1}{2}\left[\left\{\frac{1}{\sigma^2}\sum_{k=1}^{K}\frac{a_{\bullet k}}{s_k^2}+\frac{1}{\sigma^2_\mu}\right\}\mu^2-2\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{a_{ik}x_i}{s_k^2}-\frac{1}{\sigma}\sum_{k=1}^{K}\frac{a_{\bullet k}\,m_k}{s_k^2}+\frac{\mu_\mu}{\sigma^2_\mu}\right\}\mu\right],\\ &\quad +\text{const.}\\ \log\,p(\sigma|\text{rest}) \;&=\; -\,(2A+n+1)\log(\sigma)-\frac{1}{\sigma^2}\left\{B+\tfrac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{a_{ik}(x_i-\mu)^2}{s_k^2}\right\}\\ &\quad +\frac{1}{\sigma}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{a_{ik}m_k(x_i-\mu)}{s_k^2}+\text{const.}\\ \text{and }\;\log\,p(\boldsymbol{a}|\text{rest}) \;&=\; \sum_{i=1}^{n}\sum_{k=1}^{K}a_{ik}\left\{\log(w_k)-\tfrac{1}{2}\log(s_k^2)-\frac{(x_i-\mu-\sigma\,m_k)^2}{2\sigma^2 s_k^2}\right\}+\text{const.}.\end{aligned}$$

where $a_{\bullet k}\equiv\sum_{i=1}^{n}a_{ik}$. The derivation for $q^*(\mu)$ is similar to that given for each of the previous models. The derivation for $q^*(\sigma)$ is similar to that given for the Asymmetric Laplace and Skew Normal models. To obtain $q^*(\boldsymbol{a})$, first note that

$$E_{q(\mu,\sigma)}\{\log\,p(\boldsymbol{a}|\text{rest})\}=\sum_{i=1}^{n}\sum_{k=1}^{K}a_{ik}\nu_{ik}+\text{const.}$$

where ν_{ik} is given in Algorithm 4. It follows that $q^*(\boldsymbol{a})\propto\prod_{i=1}^{n}\prod_{k=1}^{K}\{\exp(\nu_{ik})\}^{a_{ik}}$. The requirement that $\sum_{k=1}^{K}\mu_{q(a_{ik})}=1$ for all $1\le i\le n$ then leads to

$$q^*(\boldsymbol{a})=\prod_{i=1}^{n}\prod_{k=1}^{K}\{\mu_{q(a_{ik})}\}^{a_{ik}}\quad\text{where}\quad \mu_{q(a_{ik})}=\exp(\nu_{ik})\Big/\sum_{k=1}^{K}\exp(\nu_{ik}).$$

Appendix B: Numerical Integration

Many of the non-analytic integrals that arise in variational Bayes for elaborate distributions are of the form

$$\mathcal{I}(\boldsymbol{\theta}) = \int_a^b \exp\{h(x;\boldsymbol{\theta})\}\,dx$$

where $h''(x;\boldsymbol{\theta}) < 0$ for all $a < x < b$ and $\boldsymbol{\theta}$. In other words, the integrand is *log-concave* over the domain for all values of its parameters which, as explained in Appendix B.1., aids numerical integration strategies. This is the case for the integral families $\mathcal{H}(p,q,r)$, $\mathcal{J}(p,q,r,s)$ and $\mathcal{J}^+(p,q,r)$ defined in Section 2.1. The family $\mathcal{F}(p,q,r,s,t)$ does not have this property, so needs to be treated more carefully. We will give the details for log-concave integrands.

The value of $\mathcal{I}(\boldsymbol{\theta})$ can be arbitrarily small or large for various values of $\boldsymbol{\theta}$. Hence, it is prudent to work with $\log\{\mathcal{I}(\boldsymbol{\theta})\}$ instead.

B.1. Transforming the integrand to a 'nice' scale

In this section, we suppress the dependence of $\mathcal{I}$ and h on the parameters $\boldsymbol{\theta}$. We transform the integrand to a nice scale by borrowing from the ideas of Laplace approximation and Gauss-Hermite quadrature (e.g. Liu & Pierce, 1994). The log-concavity property means that the equation

$$h'(x) = 0$$

has a unique solution. Using the ideas of Laplace approximation, we use the substitution

$$u = \frac{x - \mu_0}{\sigma_0\sqrt{2}}$$

where

$$\mu_0 \equiv \text{the solution to } h'(x) = 0 \quad \text{and} \quad \sigma_0 \equiv 1/\sqrt{-h''(\mu_0)}.$$

On substitution into the $\log\mathcal{I}$ expression we get

$$\log\mathcal{I} = h(\mu_0) + \log(\sigma_0\sqrt{2}) + \log(\mathcal{I}_0)$$

where

$$\mathcal{I}_0 \equiv \int_{(a-\mu_0)/(\sigma_0\sqrt{2})}^{(b-\mu_0)/(\sigma_0\sqrt{2})} \exp\{h(\mu_0 + u\,\sigma_0\sqrt{2}) - h(\mu_0)\}\,du.$$

B.2. Quadrature for $\mathcal{I}_0$

We have now reduced the problem to one involving numerical integration for $\mathcal{I}_0$. The integrand for $\mathcal{I}_0$ has the properties of being unimodal, bounded above by unity and has support 'similar' to the standard normal density. For the families $\mathcal{G}(p,q,r,s,t)$, $\mathcal{H}(p,q,r)$, $\mathcal{J}(p,q,r,s)$ and $\mathcal{J}^+(p,q,r)$ the integrands have exponentially decaying tails. Therefore, even simple quadrature such as the trapezoidal or Simpson's rules can be very accurate provided we (a) determine the effective support of the integrand; and (b) use a high number of quadrature points. For (a) a reasonable way to do this is to sequentially enlarge the support (L, U) until

$$\max\{\exp\{h(\mu_0 + L\sigma_0\sqrt{2}) - h(\mu_0)\}, \exp\{h(\mu_0 + U\sigma_0\sqrt{2}) - h(\mu_0)\}\} < \varepsilon$$

for some 'small' ε such as 10^{-15}. For (b) we use doubling of the number of quadrature points until the relative error is below some nominal threshold such as 10^{-5}.

Appendix C: Finite Normal Mixture Approximation

In this appendix we describe our strategy for finite normal mixture approximation of Generalized Extreme Value density functions, as required for Algorithm 5.

Recall that the GEV$(0,1,\xi)$ family of density functions is given by

$$f(x;\xi) = \begin{cases} (1+\xi\,x)^{-1/\xi-1}e^{-(1+\xi\,x)^{-1/\xi}}, & 1+\xi\,x>0, \xi\neq 0 \\ \exp(-x-e^{-x}), & \xi=0. \end{cases}$$

The support is $[-1/\xi,\infty)$ for $\xi>0$, $\mathbb{R}$ for $\xi=0$ and $(-\infty,-1/\xi]$ for $\xi<0$. For $\xi=-1$ the density function has a jump discontinuity at $x=1$ and for $\xi<-1$ it has a pole at $x=-1/\xi$. In the present article we have restricted attention to $-1\leq\xi\leq 1$. In applications, this sub-family is usually found to be adequate for modelling sample extrema.

Let

$$f^{\text{NM}}(x;\boldsymbol{w}_\xi,\boldsymbol{m}_\xi,\boldsymbol{s}_\xi)\equiv\sum_{k=1}^{K}\frac{w_{k,\xi}}{s_{k,\xi}}\,\phi\left(\frac{x-m_{k,\xi}}{s_{k,\xi}}\right)$$

be a normal mixture approximation to $f(x;\xi)$. The notation is the same as that used in Table 1, with the addition of a ξ subscript. After fixing K, we considered choice of $(\boldsymbol{w}_\xi,\boldsymbol{m}_\xi,\boldsymbol{s}_\xi)$ by minimizing both L_1 distance:

$$\text{IAE}(\boldsymbol{w}_\xi,\boldsymbol{m}_\xi,\boldsymbol{s}_\xi;\xi)\equiv\int_{-\infty}^{\infty}|f^{\text{NM}}(x;\boldsymbol{w}_\xi,\boldsymbol{m}_\xi,\boldsymbol{s}_\xi)-f(x;\xi)|\,dx$$

and χ^2 distance

$$\chi^2(\boldsymbol{w}_\xi,\boldsymbol{m}_\xi,\boldsymbol{s}_\xi;\xi)\equiv\int_S\{f^{\text{NM}}(x;\boldsymbol{w}_\xi,\boldsymbol{m}_\xi,\boldsymbol{s}_\xi)-f(x;\xi)\}^2/f^{\text{NM}}(x;\boldsymbol{w}_\xi,\boldsymbol{m}_\xi,\boldsymbol{s}_\xi)\,dx.$$

where S is the effective support of $f^{\text{NM}}(\cdot;\boldsymbol{w}_\xi,\boldsymbol{m}_\xi,\boldsymbol{s}_\xi)$. The Nelder-Mead simplex algorithm (Nelder & Mead, 1965) was used for optimization via the MATLAB function `fminsearch`. The entries of $\boldsymbol{w}_\xi$ were constrained to be at least 10^{-6}. The component means and variances were constrained to compact intervals. The algorithm was terminated after convergence to a local minimum.

The integrals were approximated using the trapezoidal rule on an adaptive grid. For $\xi<0$ we used 540 grid points to the left of the mode at $x=\{(1+\xi)^{-\xi}-1\}/\xi$ and 540 grid points between the mode at the upper end of the support at $x=-1/\xi$. An additional 120 grid points were used in the interval $[-1/\xi,-1/\xi+4]$ since $f^{\text{NM}}(\cdot;\boldsymbol{w}_\xi,\boldsymbol{m}_\xi,\boldsymbol{s}_\xi)$ may have a small amount of probability mass above $-1/\xi$. For $\xi\geq 0$ the grid point strategy required more care due to the heavy right-hand tail. An equi-spaced grid between the 10^{-8} and $1-10^{-6}$ quantiles of $f^{\text{NM}}(\cdot;\boldsymbol{w}_\xi,\boldsymbol{m}_\xi,\boldsymbol{s}_\xi)$, but right-truncated at 100000, was used. The grid sizes increased linearly from 1100 for $\xi=0$ to 7500 for $\xi=0.3$ and was fixed at 7500 for $0.3<\xi\leq 1$. Approximating mixtures were determined for $\xi\in[-1,1]$ over an equally-spaced grid of size 401.

Figure 7 shows some indications of the accuracy of $K=24$ mixture normal mixture approximations to the $f(\cdot;\xi)$ density functions. The top panel shows the accuracy of the L_1-based approximation. Since the L_1 distance between two density functions is a scale-independent number between 0 and 2, the vertical axis is immediately meangingful. The fact that the L_1 distance is uniformly below 0.01 implies that the accuracy measure defined by (19) always exceeds 99.5%. The second panel shows accuracy of χ^2-based approximation. The bottom panel compares the two types of approximation in terms of Kullback-Leibler distance and shows that the χ^2-based approximation is almost uniformly better. Further error analyses reveal that chi-squared distance leads to better accuracy in the tails. This is particularly important for $\xi>0$ since the upper tail of $f(\cdot;\xi)$ is heavy compared with those of normal densities. Hence, we recommend the

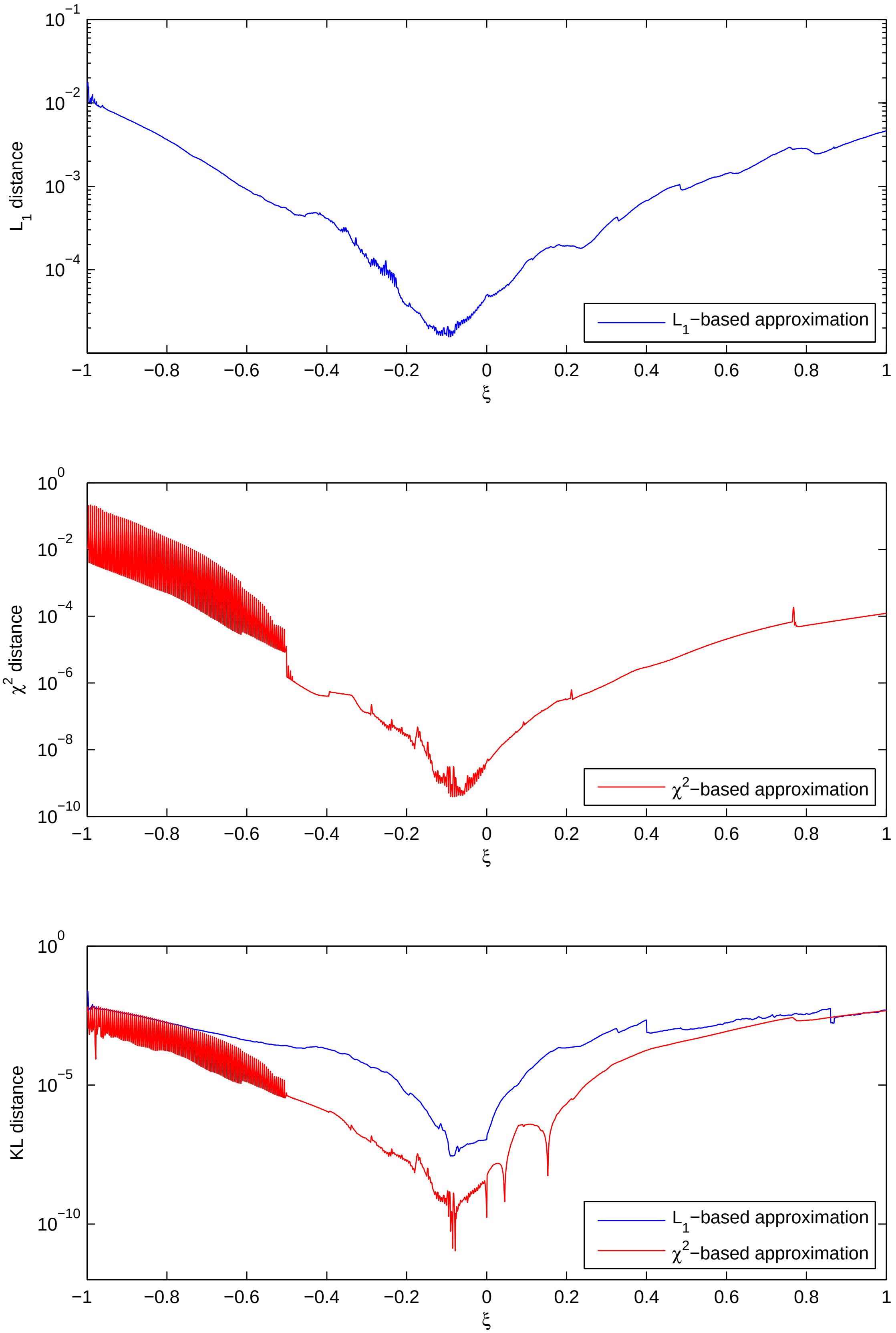

Figure 7: *Accuracy of both L_1-based and χ^2-based approximation to GEV$(0, 1, \xi)$ density functions using $K = 24$ normal mixtures. The top panel plots L_1 distance versus $-1 \le \xi \le 1$ for L_1-based approximation. The second panel shows an analogous plot for χ^2 distance. The bottom panel plots Kullback-Leibler distance versus $-1 \le \xi \le 1$ for both types of approximation.*

chi-squared based normal mixture approximations and these are used in Section 4.5 and the remainder of this appendix.

A text file containing the fitted normal parameters over the fine grid of ξ values is available as web-supplement to this article.

Acknowledgements

The first author is grateful to Professor A.C. Davison and the Department of Mathematics at the Swiss Federal Institute of Technology, Lausanne, Switzerland, for their hospitality during the early stages of this research. We thank Sarah Neville for her assistance with some of the derivations. This research was partially supported by Australian Research Council Discovery Project DP0877055.

References

Aigner, D.J., Lovell, C.A.K. & Schmidt, P. (1977). Formulation estimation of stochastic frontier production function model. *Journal of Econometrics*, **12**, 21–37.

Archambeau, C. & Bach, F. (2008). Sparse probabilistic projections. *21st Annual Conference on Neural Information Processing Systems, Vancouver, Canada, December 8–1, 2008*.

Armagan, A. (2009). Variational bridge regression. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, **5**, 17–24.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 21–30.

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*, **65**, 367–389.

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

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer.

Consonni, G. & Marin, J.-M. (2007). Mean-field variational approximate Bayesian inference for latent variable models. *Computational Statistics and Data Analysis*, **52**, 790–798.

Cottet, R., Kohn, R.J. & Nott, D.J. (2008). Variable selection and model averaging in semiparametric overdispersed generalized linear models. *Journal of the American Statistical Association*, **103**, 661–671.

Devroye, L & Györfi, L. (1985). *Density Estimation: The L_1 View*. New York: Wiley.

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

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.

Girolami, M. & Rogers, S. (2006). Variational Bayesian multinomial probit regression. *Neural Computation*, **18**, 1790–1817.

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

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

Ligges, U., Thomas, A., Spiegelhalter, D., Best, N., Lunn, D., Rice, K. & Sturtz, S. (2010). BRugs 0.5: OpenBUGS and its R/S-PLUS interface BRugs. R package. `http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/2.10`

Liu, Q. & Pierce, D.A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, **81**, 624–629.

Luenberger, D.G & Ye, Y. (2008). *Linear and Nonlinear Programming, Third Edition*. New York: Springer.

Lunn, D.J., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.

McGrory, C.A. & Titterington, D.M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis*, **51**, 5352–5367.

Minka, T., Winn, J., Guiver, G. & Kannan, A. (2009). Infer.Net 2.3. Microsoft Research Cambridge, Cambridge, UK.

Nakajima, J., Kunihama, T., Omori, Y. & Frühwirth-Schnatter, S. (2009). Generalized extreme value distribution with time-dependence using the ARMA model in state space form. Unpublished manuscript.

Nelder, J.A. & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308–313.

Ormerod, J.T. & Wand, M.P. (2010). Explaining variational approximations. *The American Statistician*, **64**, 140–153.

Parisi, G. (1988). *Statistical Field Theory*. Redwood City, California: Addison-Wesley.

Park, T. & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–686.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, California: Morgan Kaufmann.

Staudenmayer, J., Lake, E.E. & Wand, M.P. (2009). Robustness for general design mixed models using the t-distribution. *Statistical Modelling*, **9**, 235–255.

Teschendorff, A.E., Wang, Y., Barbosa-Morais, N.L., Brenton, J.D. & Caldas C. (2005). A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics*, **21**, 3025–3033.

Tipping, M.E. & Lawrence, N.D. (2003). A variational approach to robust Bayesian interpolation. *IEEE Workshop on Neural Networks for Signal Processing*, 229–238.

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

Wand, M.P. & Ripley, B.D. (2009). KernSmooth 2.23. Functions for kernel smoothing corresponding to the book: Wand, M.P. & Jones, M.C. (1995) "Kernel Smoothing". R package. `http://cran.r-project.org`

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

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