Mean Field Variational Bayes for Continuous Sparse Signal Shrinkage: Pitfalls and Remedies

BY SARAH E. NEVILLE[1], J.T. ORMEROD[2] & M.P. WAND[3]

[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] *School of Mathematical Sciences, University of Technology Sydney, Broadway 2007, AUSTRALIA*

23rd May, 2014

SUMMARY

We investigate mean field variational approximate Bayesian inference for models that use continuous distributions, Horseshoe, Negative-Exponential-Gamma and Generalized Double Pareto, for sparse signal shrinkage. Our principal finding is that the most natural, and simplest, mean field variational Bayes algorithm can perform quite poorly due to posterior dependence among auxiliary variables. More sophisticated algorithms, based on special functions, are shown to be superior. Continued fraction approximations via Lentz's Algorithm are developed to make the algorithms practical.

Keywords: Approximate Bayesian inference; continued fraction; Generalized Double Pareto distribution; Horseshoe distribution; Lentz's Algorithm; Normal-Exponential-Gamma distribution; special function.

1 Introduction

We report on an investigation into the extension of mean field variational Bayes (MFVB) methodology to accommodate various *continuous sparse signal shrinkage* distributions. Our findings are two-pronged. Firstly, MFVB can possess pitfalls when applied naïvely – that is, when using natural auxiliary variable representations of continuous sparse signal distributions. Natural auxiliary variable representations are those that allow conjugate Gibbs sampling updates. The root cause is strong posterior dependence among auxiliary variables. Secondly, remedies are developed based on alternative auxiliary variable representations that remove the posterior dependence problem. These remedies involve a new MFVB tool: continued fraction approximations via Lentz's Algorithm.

Mean field approximation is a versatile and principled approach to approximate Bayesian inference in graphical models (e.g. Wainwright & Jordan, 2008). Over the past decade or so it has become increasingly popular as a fast alternative to Markov chain Monte Carlo (MCMC) for inference in hierarchical Bayesian models, where it has become known as *variational Bayes* or, more descriptively, *mean field variational Bayes* (MFVB). Much of the early MFVB literature treated models with standard distributions such as the Dirichlet, Gamma and Normal families (e.g. Attias, 1999; Teschendorff *et al.*, 2005; Flandin & Penny, 2007; McGrory & Titterington, 2007; Consonni & Marin, 2007). More recently efforts have targeted effective incorporation of more elaborate distributions into the MFVB framework such as the t, Laplace and Generalized Extreme Value distributions (e.g. Archambeau & Bach, 2008; Armagan, 2009; Wand, Ormerod, Padoan & Frürwirth, 2011). An earlier reference on MFVB for an elaborate distribution is Tipping & Lawrence (2003), who treated the t distribution with fixed degrees of freedom.

Continuous sparse signal shrinkage distributions are a topical class of elaborate distributions that, to date, have received little or no attention with regard to MFVB methodology. The primary motivation for their development is regression analysis for wide data ("$p \gg n$") settings, but they can also be contemplated for Bayesian approaches to wavelet nonparametric and semiparametric regression (e.g. Wand & Ormerod, 2011). Examples of continuous sparse signal distributions are:

- the Horseshoe distribution (Carvalho, Polson & Scott, 2010),
- the Normal-Exponential-Gamma and Normal-Gamma distributions (Griffin & Brown, 2011), and
- the Generalized Double Pareto distribution (Armagan, Dunson & Lee, 2013).

Several other examples are given in Polson & Scott (2010). In each of these references, the estimation properties of such distributions, when used as priors on coefficients in sparse signal regression models, are established. They represent purely continuous alternatives to so-called "slab-and-spike" priors such as Laplace-Zero mixtures (e.g. Johnstone & Silverman, 2004, 2005). The relative merits of the several options now available for coefficient priors in sparse signal models are not studied here. Rather, we devise MFVB algorithms and assess their quality once the shrinkage distribution has been chosen. Whilst we focus on the three classes of distributions listed above, we expect that the lessons apply generally to distributions of this type.

The current study is motivated by various versions of sparse signal regression, where it is desirable to force many of the estimated regression coefficients to be zero or effectively zero. In the past two decades, several Bayesian and non-Bayesian fitting methods for sparse regression have been developed. Given our focus on MFVB fitting, our discussion here is confined to Bayesian approaches to sparse regression. A generic form satisfied by many Bayesian sparse regression models is

$$g(E(\boldsymbol{y}|\,\beta_0, \boldsymbol{\beta})) = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}$$

where $\boldsymbol{y}$ is vector of response variables, $\mathbf{1}$ is a vector of ones and $\boldsymbol{X}$ is an $n \times p$ fixed design matrix and g is a link function. Situations for which a sparse estimate of $\boldsymbol{\beta}$ is desirable include:

- the columns of $\boldsymbol{X}$ correspond to wavelet basis functions of the observed values of a continuous predictor variable,
- $\boldsymbol{X}$ is very wide, in that $p \gg n$; i.e. there are many more candidate predictors than observations.

Sparseness may be achieved within the Bayesian paradigm via a prior specification of the form

$$\beta_j \overset{\text{ind.}}{\sim} p(\beta_j)$$

where p is a "slab-and-spike" density function:

$$p(x) = w p_{\text{cts.}}(x) + (1 - w)\delta_0(x), \quad 0 < w < 1,$$

with $p_{\text{cts.}}$ denoting a continuous density function. Johnstone & Silverman (2004, 2005), for example, make a compelling case for use of such priors. As mentioned earlier, several purely continuous alternatives to "slab-and-spike" density functions have been proposed in the last few years, and the versions studied here are defined in Appendix A. They do not lead to exactly sparse posterior distributions, but rather to "effectively" sparse results in that the posterior density function of β_j is very close to a point mass at zero. Despite this supposed drawback, purely continuous priors have been shown to have good properties in sparse signal contexts when $p < n$. For example, Carvalho, Polson & Scott (2010)

prove that the Horseshoe prior has tail robustness and super-efficient convergence properties. Markov chain Monte Carlo (MCMC) is the most common approach to Bayesian inference. However, $\boldsymbol{X}$ can be extremely wide in certain areas of application such as genomics. This can cause MCMC to be unacceptably slow and recent research has been concerned with fast MFVB alternatives in sparse signal regression (e.g. Logsdon, Hoffman & Mezey, 2010; Carbonetto & Stephens, 2011; Wand & Ormerod, 2011).

Section 2 is this article's centerpiece. We confine attention to simple univariate models involving continuous sparse signal shrinkage distributions. This means that the essence of MFVB for continuous sparse signal shrinkage can be delved into with minimal structure and notation. The locality property of MFVB (e.g. Wand *et al.*, 2011, Section 3) means that the lessons and methodology apply to other Bayesian models containing distributions of this type. We provide theory and numerical studies that point to serious pitfalls when MFVB is used naïvely, and then describe remedies. Some brief remarks on implications for sparse signal regression are made in Section 3. Appendix A contains necessary backgound material on special functions, distributional definitions and results. Some background on MFVB is also given. The derivations of the article's MFVB algorithms are given in Appendix B. Appendix C contains a proof of our main theoretical result, Theorem 1, concerning pitfalls of MFVB for continuous sparse signal shrinkage.

2 Univariate Scale Models

In a vein similar to Wand *et al.* (2011), we now concentrate on simple univariate scale models involving continuous sparse signal shrinkage distributions. These allow a deeper understanding of the issues, with a minimal amount of notational overhead. Unlike Wand *et al.* (2011), the location parameter is taken to be zero – which is in keeping with the use of such distributions in sparse signal regression. Definitions and results needed for this section are given in Appendix A.

2.1 Horseshoe Distribution

Consider the following Bayesian location-scale model for a univariate random sample from the Horseshoe distribution:

$$x_i|\,\sigma \overset{\text{ind.}}{\sim} \text{Horseshoe}(0,\sigma), \quad \sigma \sim \text{Half-Cauchy}(A), \tag{1}$$

where $A > 0$ is a hyperparameter.

Table 1 lists three new models that are equivalent to (1). The equivalences are due to Results 1a, 1b and 4 given in Appendix A. Model I introduces the single auxiliary variable a. Model II adds the $\boldsymbol{b} = (b_1, \ldots, b_n)$ vector of auxiliary variables. In Model III a third set of auxiliary variables, corresponding to the vector $\boldsymbol{c} = (c_1, \ldots, c_n)$, is added. Figure 1 shows the directed acyclic graphs corresponding to Models I, II and III.

An attraction of Model III is that each of the conditional distributions belong to the Normal and Gamma families. This translates to the full conditional distributions being standard distributions and Gibbs sampling being exact. Such is not the case for Models I and II.

Now consider MFVB approximation of the joint posterior density functions of σ and the auxiliary variables, according to the following product assumptions:

$$\begin{array}{lcll} p(\sigma, a|\boldsymbol{x}) & \approx & q(\sigma)\,q(a) & \text{for Model I,} \\ p(\sigma, a, \boldsymbol{b}|\boldsymbol{x}) & \approx & q(\sigma)\,q(a,\boldsymbol{b}) & \text{for Model II,} \\ p(\sigma, a, \boldsymbol{b}, \boldsymbol{c}|\boldsymbol{x}) & \approx & q(\sigma,\boldsymbol{c})\,q(a,\boldsymbol{b}) & \text{for Model III.} \end{array} \tag{2}$$

Model I	Model II	Model III
$x_i \mid \sigma \overset{\text{ind.}}{\sim} \text{Horseshoe}(0, \sigma)$	$x_i \mid \sigma, b_i \overset{\text{ind.}}{\sim} N(0, \sigma^2/b_i)$	$x_i \mid \sigma, b_i \overset{\text{ind.}}{\sim} N(0, \sigma^2/b_i)$
$\sigma^2 \mid a \sim \text{IG}(\frac{1}{2}, a^{-1})$	$\sigma^2 \mid a \sim \text{IG}(\frac{1}{2}, a^{-1})$	$\sigma^2 \mid a \sim \text{IG}(\frac{1}{2}, a^{-1})$
$a \sim \text{IG}(\frac{1}{2}, A^{-2})$	$a \sim \text{IG}(\frac{1}{2}, A^{-2})$	$a \sim \text{IG}(\frac{1}{2}, A^{-2})$
	$p(b_i) = \pi^{-1} b_i^{-1/2} (b_i + 1)^{-1},\ b_i > 0$	$b_i \mid c_i \overset{\text{ind.}}{\sim} \text{Gamma}(\frac{1}{2}, c_i)$
		$c_i \overset{\text{ind.}}{\sim} \text{Gamma}(\frac{1}{2}, 1)$

Table 1: *Three auxiliary variable models that each give rise to the Horseshoe model (1).*

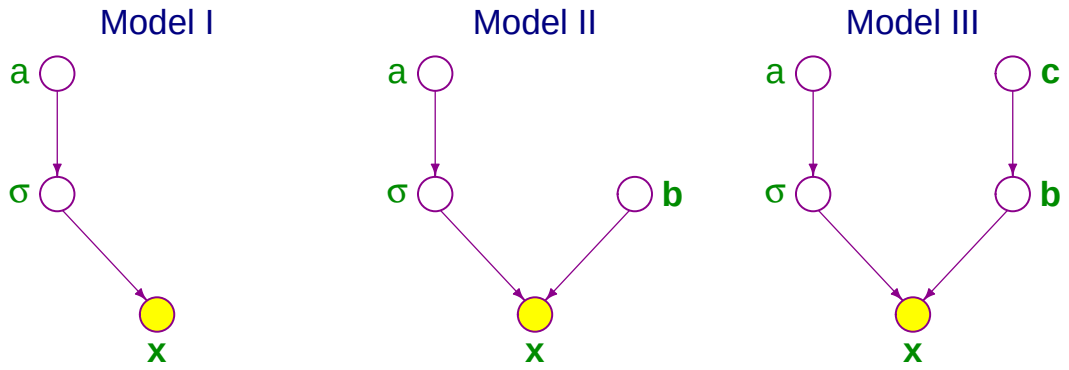

Figure 1: *Directed acyclic graphs corresponding to the three models listed in Table 1.*

Under Model I, the optimal q-densities satisfy

$$q^*(\sigma^2) \propto (\sigma^2)^{-\frac{1}{2}(n+3)} \exp\left\{ -\mu_{q(1/a)}/\sigma^2 + \sum_{i=1}^{n} \log p_{\text{HS}}(x_i/\sigma) \right\}.$$

The normalizing factor and moments of $q^*(\sigma^2)$ require numerical integration methods. The integrands involve n evaluations of the exponential integral function. This makes Model I quite challenging for MFVB and full assessment of the feasibility of such an approach is omitted here.

The appeal of Models II and III is the closed-form q-density for σ^2:

$$q^*(\sigma^2) \sim \text{IG}\left(\tfrac{1}{2}(n+1),\ \mu_{q(1/a)} + \tfrac{1}{2} \sum_{i=1}^{n} x_i^2\, \mu_{q(b_i)} \right). \tag{3}$$

We work with σ^2, rather than σ, due to $q^*(\sigma^2)$ being in a standard density function family.

The derivation of (3), as well as required optimal q-densities and relevant moments of the auxiliary variables, are given in Appendix B. These results give rise to Algorithm 1 for MFVB-approximate Bayesian inference for σ^2. We note that the Model III branch of Algorithm 1 is, to some degree, a special case of a procedure given in Section 4.1 of Armagan, Dunson & Clyde (2011).

Model II requires repeated evaluation of the function $\mathcal{Q}$, defined by (14) in Appendix A. Lentz's Algorithm (Lentz, 1976; Press *et al.*, 1992, pp. 169–171) is an effective method for continued fraction approximation of $\mathcal{Q}(x)$ to a prescribed accuracy. Algorithm 2 provides the details. Figure 2 shows that convergence is quite rapid for x bigger than about 1. For small $0 < x \leq 1$, we recommend direct computation of $\mathcal{Q}(x)$. In this case the

Initialize: $\mu_{q(1/\sigma^2)} > 0$.
If Model III, initialize: $\mu_{q(c_i)} > 0$, $1 \leq i \leq n$.
Cycle:

$$\mu_{q(1/a)} \leftarrow A^2/\{A^2\mu_{q(1/\sigma^2)} + 1\}.$$

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

$$G_i \leftarrow \tfrac{1}{2}\mu_{q(1/\sigma^2)}\, x_i^2$$

if Model II: $\mu_{q(b_i)} \leftarrow \{G_i \mathcal{Q}(G_i)\}^{-1} - 1$

if Model III: $\mu_{q(b_i)} \leftarrow 1/\left\{G_i + \mu_{q(c_i)}\right\}$; $\mu_{q(c_i)} \leftarrow 1/\left\{\mu_{q(b_i)} + 1\right\}$

$$\mu_{q(1/\sigma^2)} \leftarrow (n+1)\Big/\left\{2\mu_{q(1/a)} + \textstyle\sum_{i=1}^n x_i^2\, \mu_{q(b_i)}\right\}$$

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

Algorithm 1: *Mean field variational Bayes algorithm for determination of $q^*(\sigma^2)$ from data modelled according to (1). The schemes differ according to which auxiliary variable representations, Model II or Model III, from Table 1 is used.*

underflow threat, described in Appendix A.1.2, is absent since since $\exp(-x)$ is close to 1.

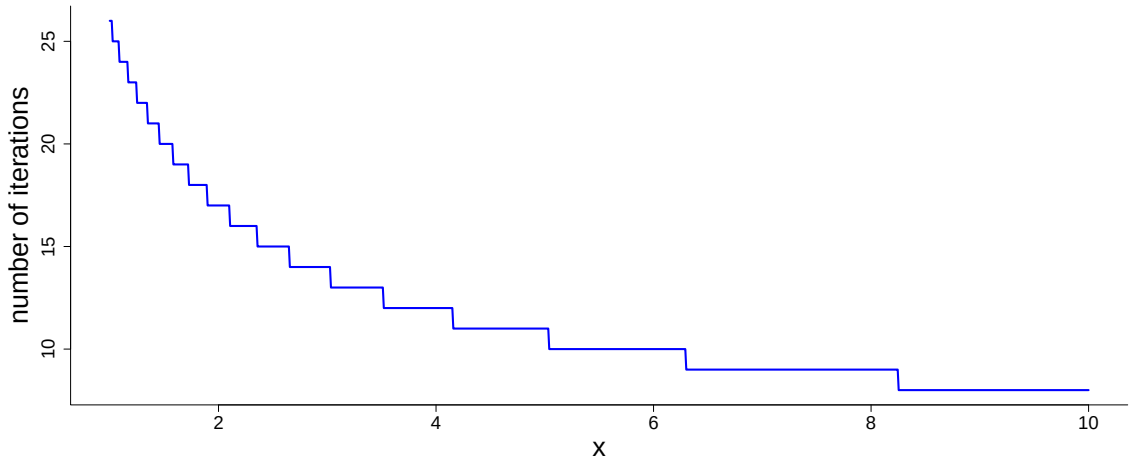

Figure 2: *The number of iterations required for Lentz's Algorithm to converge when used to approximate $\mathcal{Q}(x)$. The convergence criteria correspond to the default settings in Algorithm 2.*

It remains to discuss computation of the lower bound on the marginal log-likelihood, $\log\, \underline{p}(\boldsymbol{x}; q)$. The results in Appendix B lead to the explicit expressions:

$$\log \underline{p}(\boldsymbol{x}; q) = \begin{cases} \log \underline{p}(\boldsymbol{x}; q, \text{BASE}) - n\log(\pi) + \sum_{i=1}^n \left[\, G_i\, \mu_{q(b_i)} + \log\{\mathcal{Q}(G_i)\}\right] & \text{for Model II} \\ \log \underline{p}(\boldsymbol{x}; q, \text{BASE}) - n\log(\pi) + \sum_{i=1}^n [\mu_{q(b_i)}\{G_i + \mu_{q(c_i)}\} & \\ \quad - \log(G_i + \mu_{q(c_i)}) - \log(\mu_{q(b_i)} + 1)] & \text{for Model III} \end{cases}$$

where

$$\log \underline{p}(\boldsymbol{x}; q, \text{BASE}) \equiv \log\Gamma\left(\tfrac{n+1}{2}\right) - \tfrac{n}{2}\log(2\pi) - \log(\pi) - \log(A) - \log(\mu_{q(1/\sigma^2)} + A^{-2}) \\ -\tfrac{n+1}{2}\, \log\left(\mu_{q(1/a)} + \tfrac{1}{2}\, \textstyle\sum_{i=1}^n\, x_i^2\, \mu_{q(b_i)}\right) + \mu_{q(1/a)}\, \mu_{q(1/\sigma^2)}. \quad (4)$$

Inputs (with defaults): $x > 0$, $\varepsilon_1(10^{-30})$, $\varepsilon_2(10^{-7})$,

If $x > 1$ then (use Lentz's Algorithm)

$f_{\text{prev}} \leftarrow \varepsilon_1$; $C_{\text{prev}} \leftarrow \varepsilon_1$; $D_{\text{prev}} \leftarrow 0$; $\Delta = 2 + \varepsilon_2$; $j \leftarrow 1$

cycle while $|\Delta - 1| \geq \varepsilon_2$:

$j \leftarrow j + 1$; $D_{\text{curr}} \leftarrow x + 2j - 1 - (j-1)^2 D_{\text{prev}}$

$C_{\text{curr}} \leftarrow x + 2j - 1 - (j-1)^2 / C_{\text{prev}}$

$D_{\text{curr}} \leftarrow 1/D_{\text{curr}}$; $\Delta \leftarrow C_{\text{curr}} D_{\text{curr}}$; $f_{\text{curr}} \leftarrow f_{\text{prev}} \Delta$

$f_{\text{prev}} \leftarrow f_{\text{curr}}$; $C_{\text{prev}} \leftarrow C_{\text{curr}}$; $D_{\text{prev}} \leftarrow D_{\text{curr}}$

return $1/(x + 1 + f_{\text{curr}})$

Otherwise (use direct computation)

return $e^x E_1(x)$.

Algorithm 2: *Algorithm for stable and efficient computation of* $\mathcal{Q}(x)$.

2.1.1 Simplicity Comparison of Models II and III

Perusal of Algorithm 1 shows that Model III produces the simplest MFVB scheme since it involves only standard algebraic calculations such as taking square-roots.

Model II is obviously not as simple as Model III because of the requirement to compute $\mathcal{Q}$ for each $1 \leq i \leq n$ and for each coordinate ascent iteration. However, as indicated by Figure 2, $\mathcal{Q}$ evaluations are relatively cheap, and stable, for arguments exceeding 1. Special function software such as the R (R Development Core Team, 2014) function `expint_E1()` in the package **gsl** allows efficient and stable computation of $\mathcal{Q}$ for small x.

2.1.2 Simulation Comparison of Models II and III

Models II and III were compared via simulation. We generated 1000 data-sets according to

$$x_i \sim \text{Horseshoe}(0, 1), \quad 1 \leq i \leq n,$$

and sample sizes $n = 100$ and $n = 1000$. Hence σ^2 has a "true value" of 1. The accuracy of each MFVB approximation $q^*(\sigma^2)$ was assessed using

$$\text{accuracy} \equiv 1 - \tfrac{1}{2} \int_0^\infty \left| q^*(\sigma^2) - p_{\text{MCMC}}(\sigma^2 | \, \boldsymbol{x}) \right| d(\sigma^2)$$

where $p_{\text{MCMC}}(\sigma^2|\boldsymbol{x})$ is an accurate MCMC-based approximation to $p(\sigma^2|\boldsymbol{x})$, obtained using WinBUGS (Lunn *et al.* 2000) with R interfacing via the `BRugs` package (Ligges *et al.* 2011). MCMC samples of size 10000 were generated, with the first 5000 values discarded as burn-in and the remaining 5000 thinned by a factor of 5. Kernel density estimation, with direct plug-in bandwidth selection using the R package `KernSmooth` (Wand & Ripley, 2010), was used to obtain $p_{\text{MCMC}}(\sigma^2 | \, \boldsymbol{x})$ over a fine grid of σ^2 values. Note that $0 \leq \text{accuracy} \leq 1$, with an accuracy score of 1 implying perfect correspondence between the MFVB and MCMC approximations. We also kept track of coverage of the MFVB-approximate 95% credible intervals.

Table 2 summarizes the accuracy and coverage percentages for Model II and Model III MFVB. The Model III results are abysmal. In particular, none of the $n = 1000$ credible

	$n = 100$		$n = 1000$	
	accuracy	coverage	accuracy	coverage
Model II	54.3 (1.4)	55%	56.8 (0.9)	58%
Model III	6.3 (0.9)	4%	0.0 (0.0)	0%

Table 2: *Average (standard deviation) accuracy and percentage coverage of true σ^2 value by approximate 95% credible intervals based on MFVB for a simulation of size 1000 from (1).*

intervals include the true value of σ^2 and each of the $q^*(\sigma^2)$ densities has 0% accuracy. Model II has reasonably acceptable accuracy and interval coverage.
Figure 3 provides graphical comparison between $p_{\text{MCMC}}(\sigma^2|\boldsymbol{x})$ and $q^*(\sigma^2)$ for four replications. The Model III $q^*(\sigma^2)$ densities have their probability mass tightly centered on about 0.4, with negligible mass including the true value of 1. However, the Model II $q^*(\sigma^2)$ densities tend to be centered on 1 albeit with a lower amount of spread compared with $p_{\text{MCMC}}(\sigma^2|\boldsymbol{x})$.

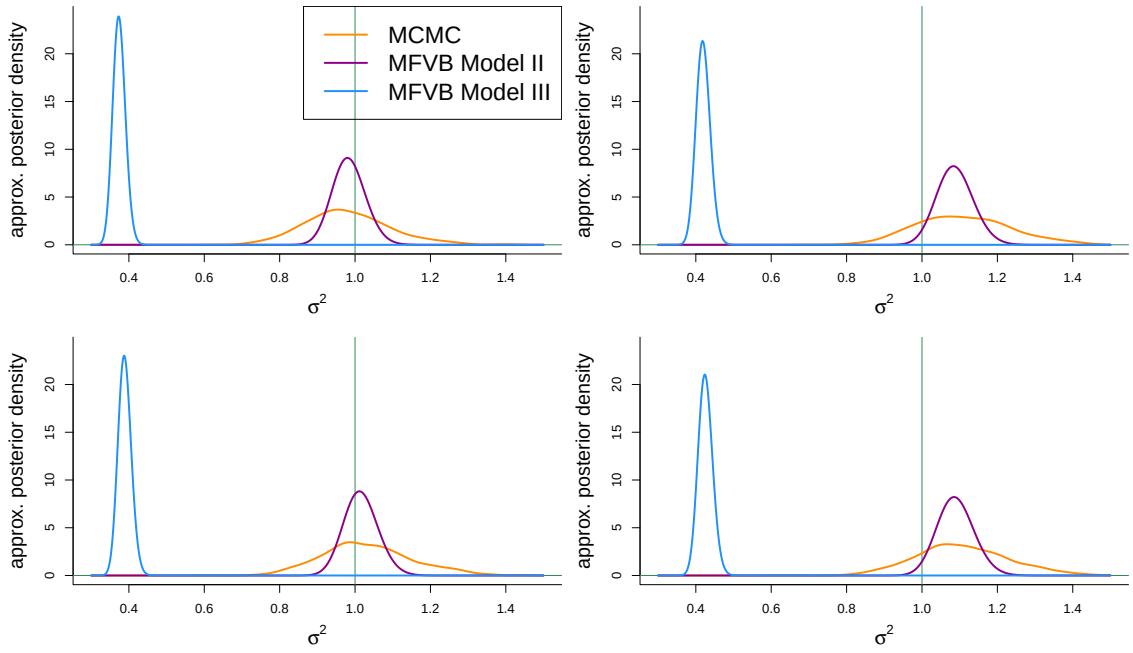

Figure 3: *Comparison of $p_{MCMC}(\sigma^2|\boldsymbol{x})$ and two $q^*(\sigma^2)$ densities, based Model II and Model III MFVB, for four replications from the simulation study corresponding to Table 2 with $n = 1000$.*

2.1.3 Theoretical Comparison of Models II and III

The simulation comparison of Section 2.1.2 shows that the most computationally convenient MFVB scheme, that based on Model III, has poor practical performance compared with that based on Model II. In this section we provide some theoretical explanations for these differences.

Our first theoretical observation is that the updates for $\mu_{q(b_i)}$ in Algorithm 1 may be written as:

$$\mu_{q(b_i)} \leftarrow \begin{cases} g^{\text{II}}(G_i) & \text{for Model II} \\ g^{\text{III}}(G_i) & \text{for Model III} \end{cases} \tag{5}$$

where

$$g^{\text{II}}(x) \equiv \{x\,\mathcal{Q}(x)\}^{-1} - 1 \quad \text{and} \quad g^{\text{III}}(x) \equiv \sqrt{\frac{1}{x} + \frac{1}{4}} - \frac{1}{2}. \tag{6}$$

The expression for $g^{\text{III}}(x)$ follows from algebraic reduction of the two equations in $\mu_{q(b_i)}$ and $\mu_{q(c_i)}$ in the Model III updates.

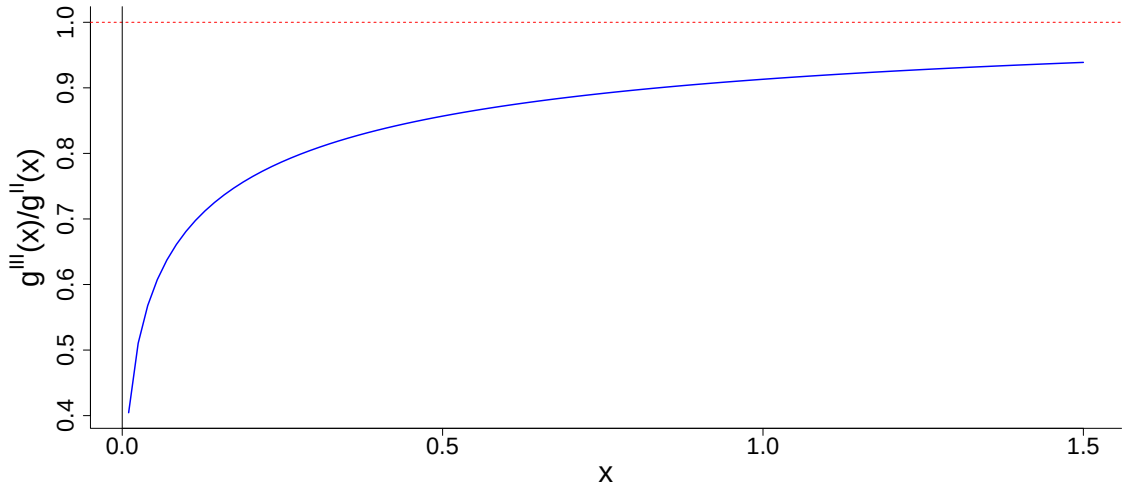

Figure 4: *Plot of $g^{III}(x)/g^{II}(x)$ for the functions g^{II} and g^{III} defined by (6).*

The interpretation of (5) is that use of the c_i auxiliary variables induces the approximation of g^{II} by the simpler g^{III}. But, as shown in Figure 4, the functions differ considerably for low positive arguments. This helps explain the poor performance of the Model III MFVB algorithm exhibited in Table 2 and Figure 3.

The root cause of this discrepancy is the strong posterior dependence between the b_i and c_i auxiliary variables, whereas MFVB assumes that there is no such dependence. This dependence can be described in simple terms by considering random variables x, b and c such that

$$x|\, b \sim N(0, 1/b), \quad b\,|\, c \sim \text{Gamma}(\tfrac{1}{2}, c), \quad c \sim \text{Gamma}(\tfrac{1}{2}, 1). \tag{7}$$

Figure 5 shows samples of $\{(\log(1/b), \log(c))|x = x_0\}$ for $x_0 = 1, 0.1, 0.01, 0.001$, along with corresponding sample correlation values. As x_0 approaches 0, the sample correlations are seen to get closer to 1.

The behavior exhibited in Figure 5 is described by Theorem 1, which uses $\text{Corr}(u, v|w)$ to denote the conditional correlation between two random variables u and v, given w:

Theorem 1. *Consider random variables x, b and c such that*

$$x|\, b \sim N(0, 1/b), \quad b\,|\, c \sim \text{Gamma}(\tfrac{1}{2}, c), \quad c \sim \text{Gamma}(\tfrac{1}{2}, 1).$$

Then

$$\lim_{x_0 \to 0} \text{Corr}(\log(1/b), \log(c)|\, x = x_0) = 1.$$

A proof of Theorem 1 is given in Appendix C. Theorem 1 verifies the behavior exhibited in Figure 5 and reinforces the inappropriateness of Model III for MFVB.

2.2 Normal-Exponential-Gamma Distribution

As in Section 2.1, we consider the model for a univariate random sample, this time arising from the Normal-Exponential-Gamma distribution:

$$x_i|\, \sigma \overset{\text{ind.}}{\sim} \text{NEG}(0, \sigma, \lambda), \quad \sigma \sim \text{Half-Cauchy}(A). \tag{8}$$

Table 3 lists three models that are equivalent to (8). The directed acyclic graphs in Figure 1 also convey the conditional dependence structure of Models I, II and III in Table 3.

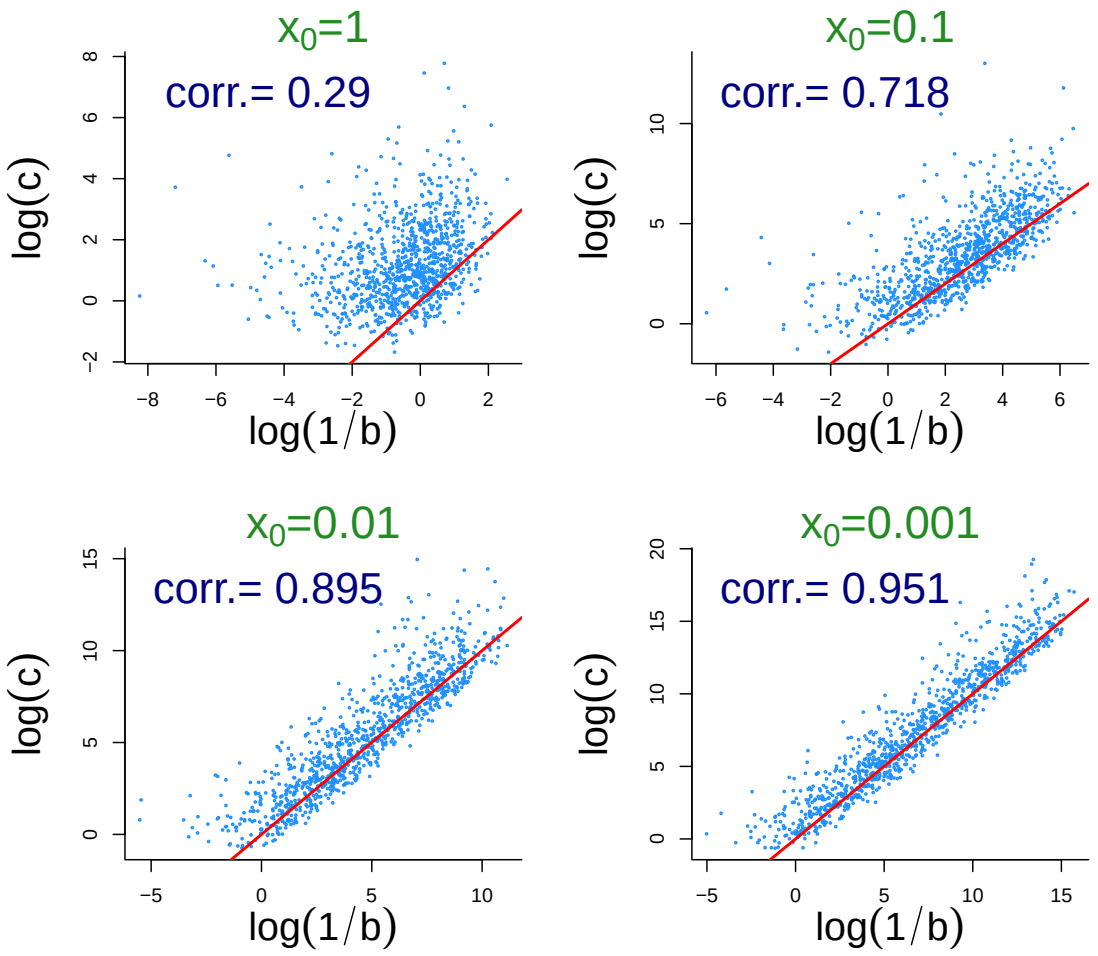

Figure 5: *MCMC samples of size 1000 from the distribution* $\{(log(1/b), log(c))|x = x_0\}$ *for* $x_0 = 1, 0.1, 0.01, 0.001$. *The corresponding sample correlations are also shown.*

Model I	Model II	Model III
$x_i \mid \sigma \overset{\text{ind.}}{\sim} \text{NEG}(0, \sigma, \lambda)$	$x_i \mid \sigma, b_i \overset{\text{ind.}}{\sim} N(0, \sigma^2/b_i)$	$x_i \mid \sigma, b_i \overset{\text{ind.}}{\sim} N(0, \sigma^2/b_i)$
$\sigma^2 \mid a \sim \text{IG}(\frac{1}{2}, a^{-1})$	$\sigma^2 \mid a \sim \text{IG}(\frac{1}{2}, a^{-1})$	$\sigma^2 \mid a \sim \text{IG}(\frac{1}{2}, a^{-1})$
$a \sim \text{IG}(\frac{1}{2}, A^{-2})$	$a \sim \text{IG}(\frac{1}{2}, A^{-2})$	$a \sim \text{IG}(\frac{1}{2}, A^{-2})$
	$p(b_i) = \lambda\, b_i^{\lambda-1}(1+b_i)^{-\lambda-1},\ b_i > 0$	$b_i \mid c_i \overset{\text{ind.}}{\sim} \text{IG}(1, c_i)$
		$c_i \overset{\text{ind.}}{\sim} \text{Gamma}(\lambda, 1)$

Table 3: *Three auxiliary variable models that each give rise to the Negative-Exponential-Gamma model (8).*

Again we consider MFVB approximation of the joint posterior density function of σ^2, according to product restrictions (2). We eliminate Model I immediately since, as for the Horseshoe distribution, the MFVB equations are very computationally challenging.

Models II and III representations lead to $q^*(\sigma^2)$ having an Inverse-Gamma distribution of the form (3). The q-density can be determined from Algorithm 3. Note that, as for Algorithm 1, the Model III branch of Algorithm 1 is, to some degree, a special case of a procedure given in Section 4.1 of Armagan, Dunson & Clyde (2011).

Initialize: $\mu_{q(1/\sigma^2)} > 0$.
If Model III, initialize: $\mu_{q(c_i)} > 0$, $1 \le i \le n$.
Cycle:

$$\mu_{q(1/a)} \leftarrow A^2/\{A^2\mu_{q(1/\sigma^2)} + 1\}.$$

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

$$G_i \leftarrow \tfrac{1}{2}\mu_{q(1/\sigma^2)}\, x_i^2$$

if Model II: $\mu_{q(b_i)} \leftarrow (2\lambda+1)\,\mathcal{R}_{2\lambda}\left(\sqrt{2\,G_i}\right) \Big/ \sqrt{2G_i}$

if Model III: $\mu_{q(b_i)} \leftarrow \sqrt{\mu_{q(c_i)}/G_i}$; $\mu_{q(1/b_i)} \leftarrow 1/\mu_{q(b_i)} + 1/\{2\mu_{q(c_i)}\}$

$$\mu_{q(c_i)} \leftarrow (\lambda+1)/\{\mu_{q(1/b_i)} + 1\}$$

$$\mu_{q(1/\sigma^2)} \leftarrow (n+1) \Big/ \left\{2\mu_{q(1/a)} + \textstyle\sum_{i=1}^n x_i^2\, \mu_{q(b_i)}\right\}$$

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

Algorithm 3: *Mean field variational Bayes algorithm for determination of $q^*(\sigma^2)$ from data modelled according to (8). The schemes differ according to which auxiliary variable representations, Model II or Model III, from Table 3 is used.*

Derivation of the updates in Algorithm 3 is given in Appendix B. The lower bounds on the marginal log-likelihood can be shown to have explicit expressions

$$\log \underline{p}(\boldsymbol{x}; q) = \begin{cases} \log \underline{p}(\boldsymbol{x}; q, \text{BASE}) + n\log(\lambda) + n(\lambda+\tfrac{1}{2})\,\log(2) + n\log\{\Gamma(\lambda+\tfrac{1}{2})\} \\ \quad + \sum_{i=1}^n \left[\{\mu_{q(b_i)} + \tfrac{1}{2}\}G_i + \log\{D_{-2\lambda-1}(\sqrt{2G_i})\}\right] & \text{for Model II} \\ \log \underline{p}(\boldsymbol{x}; q, \text{BASE}) + n\log(\lambda) + \tfrac{n}{2}\{1 + \log(\pi)\} \\ \quad - \sum_{i=1}^n \left[\tfrac{1}{2}\log\{\mu_{q(c_i)}\} + (\lambda+1)\log\{\mu_{q(1/b_i)} + 1\}\right] & \text{for Model III} \end{cases}$$

where $\log \underline{p}(\boldsymbol{x}; q, \text{BASE})$ is given by (4).

In the case of Model II, Algorithm 3, with $\nu = 2\lambda$, should be accompanied by Algorithm 4 to handle the $\mathcal{R}_{2\lambda}$ evaluations. Note that $\mathcal{R}_\nu$ is defined in Appendix A.

Inputs (with defaults): $x \ge 0$, $\lambda > 0$, $\varepsilon_1(10^{-30})$, $\varepsilon_2(10^{-7})$,

If $(\nu > 20)$ or $(x > 0.2)$ then (use Lentz's Algorithm)

$$f_{\text{prev}} \leftarrow \varepsilon_1 \;;\; C_{\text{prev}} \leftarrow \varepsilon_2 \;;\; D_{\text{prev}} \leftarrow 0 \;\;;\;\; \Delta = 2 + \varepsilon_2 \;\;;\; j \leftarrow 1$$

cycle while $|\Delta - 1| \ge \varepsilon_2$:

$$j \leftarrow j+1 \;\;;\; D_{\text{curr}} \leftarrow x + (\nu+j)D_{\text{prev}} \;\;;\; C_{\text{curr}} \leftarrow x + (\nu+j)/C_{\text{prev}}$$
$$D_{\text{curr}} \leftarrow 1/D_{\text{curr}} \;;\; \Delta \leftarrow C_{\text{curr}}\, D_{\text{curr}} \;;\; f_{\text{curr}} \leftarrow f_{\text{prev}}\, \Delta$$
$$f_{\text{prev}} \leftarrow f_{\text{curr}} \;;\; C_{\text{prev}} \leftarrow C_{\text{curr}} \;;\; D_{\text{prev}} \leftarrow D_{\text{curr}}$$

return $1/(x + f_{\text{curr}})$

Otherwise (use direct computation)

return $D_{-\nu-2}(x)/D_{-\nu-1}(x)$.

Algorithm 4: *Algorithm for stable and efficient computation of $\mathcal{R}_\nu(x)$.*

2.2.1 Simplicity Comparison of Models II and III

The comments that we made in Section 2.1.1 for the Horseshoe distribution also apply to MFVB for models containing Negative-Exponential-Gamma distributions.

Model II requires repeated evaluation of $\mathcal{R}_{2\lambda}$, defined in Appendix A, via Algorithm 4. Note the algorithm uses direct evaluation the ratio $D_{-2\lambda-2}(x)/D_{-2\lambda-1}(x)$ only for $x \leq 0.2$ and $\lambda < 40$. Otherwise Lentz's Algorithm is used. Careful checking of Lentz's Algorithm, via plots similar to Figure 2, found convergence to be quite rapid with these cut-offs. In the case of R implementation, direct evaluation for low x and λ can be handled using the function `whittakerW()` in the package `fAsianOptions`, as explained in Appendix A.

2.2.2 Simulation Comparison of Models II and III

Models II and III were compared via a simulation study analogous to that described in Section (2.1.2). We generated 500 data-sets for sizes $n = 100$ and $n = 1000$ according to

$$x_i \sim \text{NEG}(0, 1, \lambda), \quad 1 \leq i \leq n,$$

and with

$$\lambda \in \{0.1, 0.2, 0.4, 0.8, 1.6\}.$$

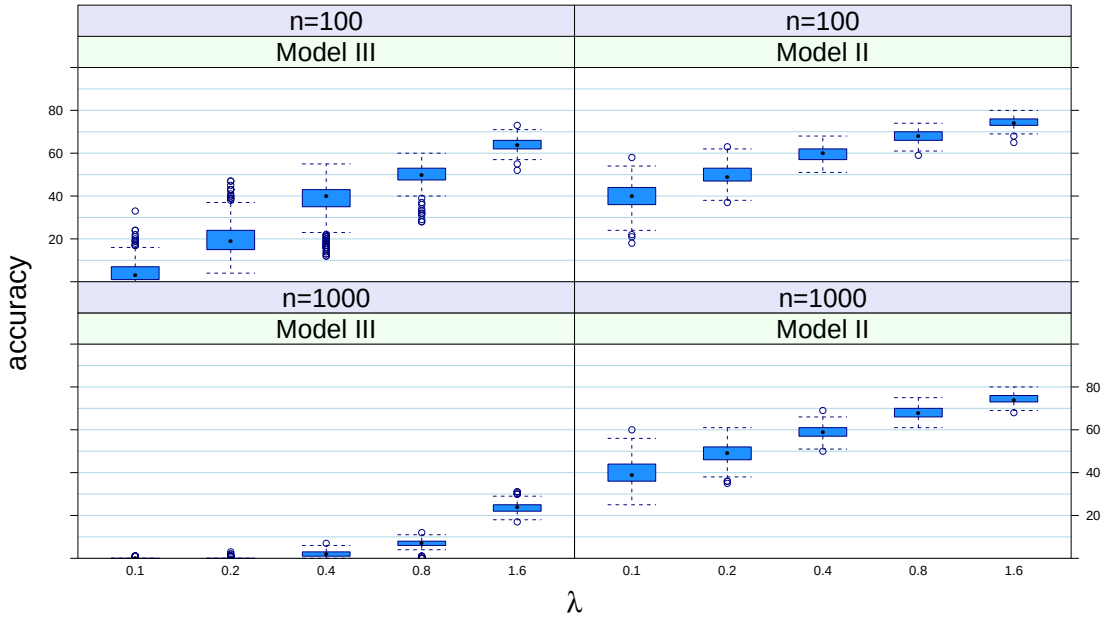

Figure 6: *Side-by-side boxplots of accuracy values for the simulation study described in the text.*

Figure 7 compares the various approximations to $p(\sigma^2|\boldsymbol{x})$ for four replications from the simulation study that produced Figure 6 for $\lambda = 0.1$. The $q^*(\sigma^2)$ density based on Model III is seen to suffer from a pronounced locational shift to the right of $p_{\text{MCMC}}(\sigma^2|\boldsymbol{x})$. On other hand, the Model II $q^*(\sigma^2)$ tends to have its central location matching that of $p_{\text{MCMC}}(\sigma^2|\boldsymbol{x})$, although its spread is considerably lower.

Finally, we compared $q^*(\sigma^2)$ for Models II and III in terms of 95% credible interval coverage of the known true value of σ^2. Table 4 shows the resulting coverage percentages.

Table 4 reveals that Model III can lead to very poor approximate inference when λ is low.

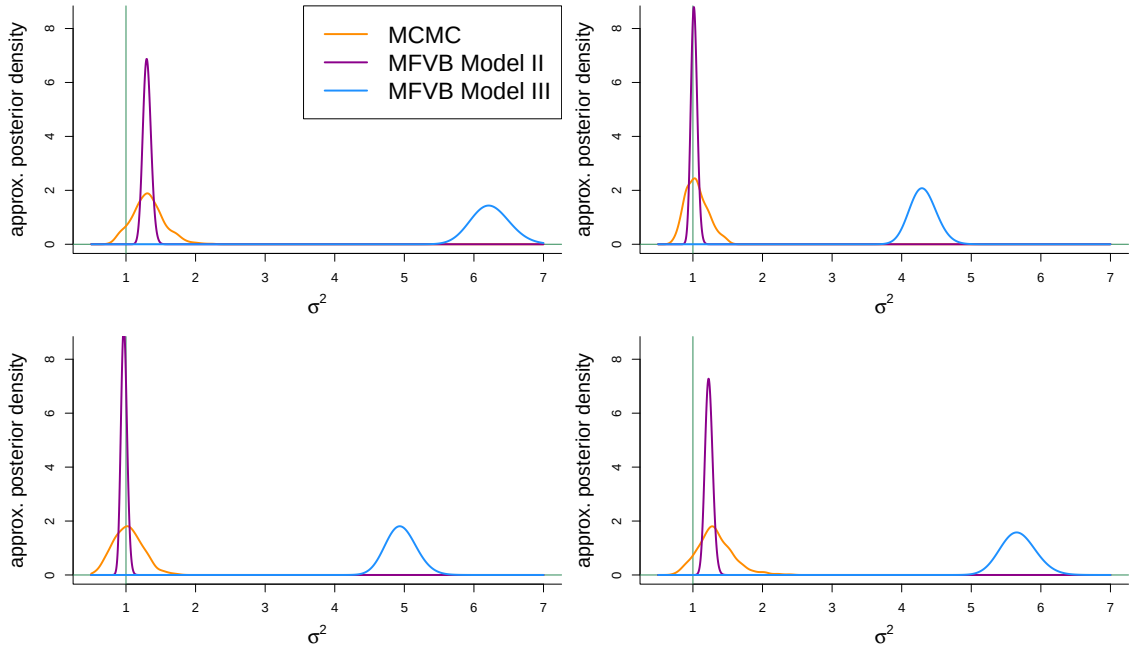

Figure 7: *Comparison of $p_{MCMC}(\sigma^2|\boldsymbol{x})$ and two $q^*(\sigma^2)$ densities, based on Model II and Model III MFVB, for four replications from the simulation study described in the text, with $\lambda = 0.1$ and $n = 1000$.*

value of λ	0.1	0.2	0.4	0.8	1.6
Model II	31	47	56	64	74
Model III	0	0	1	6	21

Table 4: *Percentage coverage of true σ^2 value by approximate 95% credible intervals based on MFVB approximate posterior density functions with $n = 1000$.*

2.2.3 Theoretical Comparison of Models II and III

Note that the update for $\mu_{q(1/b_i)}$ in Algorithm 3 may be written as:

$$\mu_{q(1/b_i)} \leftarrow \begin{cases} g_\lambda^{\text{II}}(G_i) & \text{for Model II} \\ g_\lambda^{\text{III}}(G_i) & \text{for Model III} \end{cases} \tag{9}$$

where

$$g_\lambda^{\text{II}}(x) \equiv \frac{(2\lambda+1)\mathcal{R}_{2\lambda}(\sqrt{2x})}{\sqrt{2x}} \quad \text{and} \quad g_\lambda^{\text{III}}(x) \equiv \sqrt{\frac{2\lambda+1}{2x} + \frac{1}{4}} - \frac{1}{2}.$$

The expression for $g_\lambda^{\text{III}}(x)$ follows from algebraic reduction of the three equations in $\mu_{q(b_i)}$, $\mu_{q(c_i)}$ and $\mu_{q(1/b_i)}$ corresponding to the Model III updates. The g_λ^{III} expression is quite similar to that for g^{III} in Section 2.1.3 for the Horseshoe theoretical comparison and its interpretation as an approximation to g_λ^{II} is analogous to the one described there. In Figure 8 these ratios of the two functions are compared across different values of λ. It is apparent that the gap between g_λ^{II} and g_λ^{III} widens as λ becomes small. This transfers to worse comparative performance of Method III for lower values of λ.

Figure 9 shows MCMC-based samples from the posterior distribution of $(\log(b_i), \log(c_i))$ for simulated data generated according to (8) with $n = 5$ and the same values of λ as Figure 8. Note that the posterior correlation is quite strong for $\lambda = 1.6$, and increases to be near perfect correlation as λ decreases. Such behavior is directly at odds with the $q(\boldsymbol{b}, \boldsymbol{c}) = q(\boldsymbol{b})\, q(\boldsymbol{c})$ product restriction on which Model III MFVB is based.

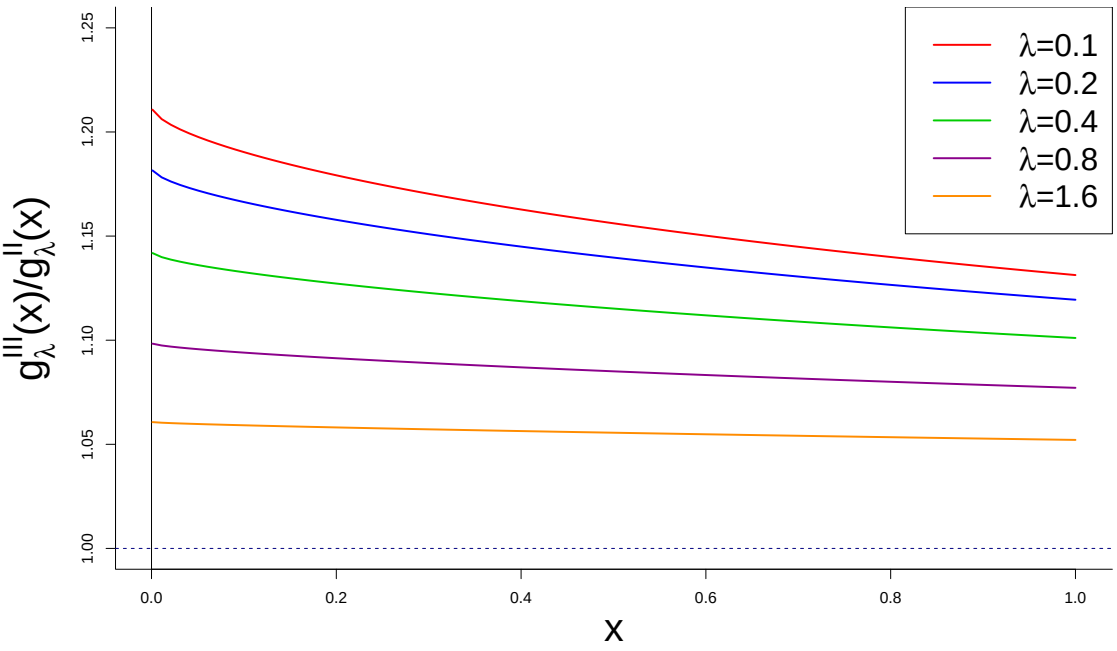

Figure 8: *Plots of* $g_\lambda^{III}(x)/g_\lambda^{II}(x)$ *for* $\lambda \in \{0.1, 0.2, 0.4, 0.8, 1.6\}$.

It is conjectured that analogues of Theorem 1 also hold for Model III in the NEG case. Numerical corroboration of such conjectures are provided in Neville (2013).

2.3 Generalized Double Pareto Distribution

The univariate location-scale model that we consider is:

$$x_i|\,\sigma \overset{\text{ind.}}{\sim} \text{GDP}(0, \sigma, \lambda), \quad \sigma \sim \text{Half-Cauchy}(A). \tag{10}$$

Table 5 lists three alternative representations of this model. The directed acyclic graph structure depicted in Figure 1 applies to these models as well.

Model I	Model II	Model III
$x_i\|\,\sigma \overset{\text{ind.}}{\sim} \text{GDP}(0, \sigma, \lambda)$	$x_i\|\,\sigma, b_i \overset{\text{ind.}}{\sim} N(0, \sigma^2/b_i)$	$x_i\|\,\sigma, b_i \overset{\text{ind.}}{\sim} N(0, \sigma^2/b_i)$
$\sigma^2\|\,a \sim \text{IG}(\frac{1}{2}, a^{-1})$	$\sigma^2\|\,a \sim \text{IG}(\frac{1}{2}, a^{-1})$	$\sigma^2\|\,a \sim \text{IG}(\frac{1}{2}, a^{-1})$
$a \sim \text{IG}(\frac{1}{2}, A^{-2})$	$a \sim \text{IG}(\frac{1}{2}, A^{-2})$	$a \sim \text{IG}(\frac{1}{2}, A^{-2})$
	$p(b_i) = \frac{1}{2}(\lambda+1)\lambda^{\lambda+1}\, b_i^{(\lambda-2)/2}$	$b_i\|c_i \overset{\text{ind.}}{\sim} \text{IG}(1, \frac{1}{2}c_i^2)$
	$\times\, e^{\lambda^2 b_i/4} D_{-\lambda-2}(\lambda\sqrt{b_i}),\ b_i > 0$	$c_i \overset{\text{ind.}}{\sim} \text{Gamma}(\lambda, \lambda)$

Table 5: *Three auxiliary variable models that each give rise to the Generalized Double Pareto model (10).*

Algorithm 5 sets out the MFVB algorithms corresponding to Models II and III. Justification is given in Appendix B. Note that Algorithm 5 uses the result

$$D_{-\lambda-4}(x)/D_{-\lambda-2}(x) = \{1 - x\,\mathcal{R}_{\lambda+1}(x)\}/(\lambda+3), \quad \lambda > 0,$$

which follows from the recurrence formula for parabolic cylinder functions (Gradshteyn & Ryzhik, 1994).

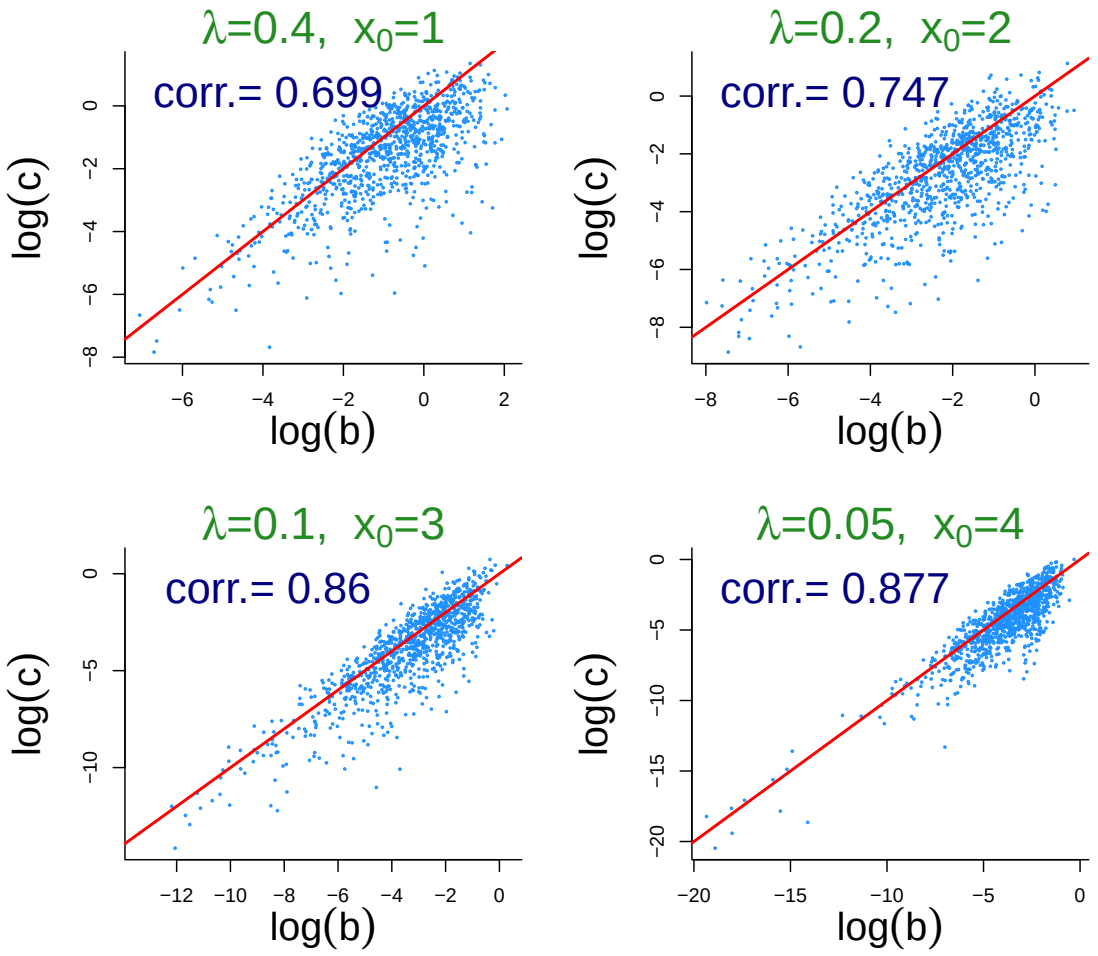

Figure 9: *MCMC samples from the posterior distributions of* $(\log(b_i), \log(c_i))$ *for simulated data generated according to (8) with varying values of* λ *and* x_0*. The sample correlations are also shown.*

$$\log \underline{p}(\boldsymbol{x}; q) = \begin{cases} \log \underline{p}(\boldsymbol{x}; q, \text{BASE}) + \frac{n}{2}\log(\pi) + n(\lambda+1)\log(\lambda) - \frac{n}{2}(3\lambda+4)\,\log(2) \\ + \sum_{i=1}^{n}\left[\mu_{q(b_i)}G_i - \frac{1}{2}(\lambda+1)\log(G_i)\right. \\ \left. + \log\{{}_2F_1(\frac{1}{2}\lambda+1; \frac{1}{2}\lambda+\frac{1}{2}; \lambda+2, 1-\lambda^2/(2G_i))\}\right] \qquad \text{for Model II} \\ \log \underline{p}(\boldsymbol{x}; q, \text{BASE}) + n[\lambda\log(\lambda) + \log\{\lambda(\lambda+1)\} - \log(2) + \frac{1}{2}\log(\pi) + \frac{1}{2}] \\ - \sum_{i=1}^{n}\left[\frac{1}{2}(\lambda+1)\log\{\mu_{q(1/b_i)}\} - \frac{1}{4}\lambda^2/\mu_{q(1/b_i)}\right. \\ \left. \qquad - \log\, D_{-\lambda-2}(\lambda/\sqrt{\mu_{q(1/b_i)}})\right] \qquad \text{for Model III} \end{cases}$$

where $\log \underline{p}(\boldsymbol{x}; q, \text{BASE})$ is given by (4).

2.3.1 Comparison of Models II and III

MFVB for GDP has the unexpected feature of being *simpler* for Model II than it is for Model III, since the latter involves special functions whereas the former does not. This represents a reversal of relative complexities compared with the Horseshoe and NEG cases. There is no compelling reason for introduction of the c_i auxiliary variables and Model III does not seem worthy of further consideration.

Fuller details, and associated numerical work, are provided in Neville (2013).

Initialize: $\mu_{q(1/\sigma^2)} > 0$.
If Model III, initialize: $\mu_{q(c_i)} > 0$, $1 \leq i \leq n$.
Cycle:

$$\mu_{q(1/a)} \leftarrow A^2/\{A^2\mu_{q(1/\sigma^2)} + 1\}.$$

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

$$G_i \leftarrow \tfrac{1}{2}\mu_{q(1/\sigma^2)}\, x_i^2$$

if Model II: $\mu_{q(b_i)} \leftarrow \dfrac{\lambda + 1}{\sqrt{2G_i}(\lambda + \sqrt{2G_i})}$

if Model III: $\mu_{q(b_i)} \leftarrow \sqrt{\mu_{q(c_i^2)}/(2G_i)}$; $\mu_{q(1/b_i)} \leftarrow 1/\mu_{q(b_i)} + 1/\mu_{q(c_i^2)}$

$$\mu_{q(c_i^2)} \leftarrow (\lambda + 2)[1 - \{\lambda/\sqrt{\mu_{q(b_i)}}\}\, \mathcal{R}_{\lambda+1}(\lambda/\sqrt{\mu_{q(b_i)}})]/\mu_{q(b_i)}$$

$$\mu_{q(1/\sigma^2)} \leftarrow (n+1) \Big/ \left\{2\mu_{q(1/a)} + \textstyle\sum_{i=1}^n x_i^2\, \mu_{q(b_i)}\right\}$$

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

Algorithm 5: *Mean field variational Bayes algorithm for determination of $q^*(\sigma^2)$ from data modelled according to (10). The schemes differ according to which auxiliary variable representations, Model II or Model III, from Table 5 is used.*

2.4 Conclusion

The numerical and theoretical results presented in Sections 2.1, 2.2 and 2.3 all point to the same conclusion: the two-level auxiliary variable representations of continuous sparse signal shrinkage distributions, that involve simple distributions and give rise to simple MCMC algorithms, lead to serious pitfalls when used in MFVB algorithms. On the other hand, one-level auxiliary variable representations are reasonably well-behaved and, hence, provide remedies to these pitfalls.

3 Implications for Sparse Signal Regression

The findings laid out in Section 2 have immediate implications for MFVB fitting and inference in sparse signal regression with continuous sparse signal shrinkage priors. Because of the locality property of MFVB, the pitfalls of high posterior dependence among auxiliary variables can impact the quality of inference for parameters close to those auxiliary variables on the regression model's directed acyclic graph.

We confine discussion here to the Horseshoe prior. Similar comments apply to the Normal-Exponential-Gamma, Generalized Double Pareto and other similar priors. Consider the sparse signal regression model

$$\begin{gathered}
\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \sigma_\varepsilon \sim N(\mathbf{1}\,\beta_0 + \boldsymbol{X}\,\boldsymbol{\beta}, \sigma_\varepsilon^2\,\boldsymbol{I}), \\
\beta_0 \sim N(0, \sigma_{\beta_0}^2), \quad \beta_j|\,\sigma_\beta \overset{\text{ind.}}{\sim} \text{Horseshoe}(0, \sigma_{\boldsymbol{\beta}}), \quad 1 \leq j \leq p, \\
\sigma_\varepsilon \sim \text{Half-Cauchy}(0, A_\varepsilon), \quad \sigma_{\boldsymbol{\beta}} \sim \text{Half-Cauchy}(0, A_{\boldsymbol{\beta}})
\end{gathered} \tag{11}$$

where $\boldsymbol{X}$ is $n \times p$ and $\sigma_{\beta_0}^2, A_\varepsilon, A_{\boldsymbol{\beta}} > 0$ are hyperparameters. Analogously to the univariate location-scale models, Model (11) has auxiliary variable representations based on Results 1a and 1b being applied to the Horseshoe distribution. We will continue to use

the Model II and Model III labeling. For example, Model III involves replacement of $\beta_j \overset{\text{ind.}}{\sim} \text{Horseshoe}(0, \sigma_{\boldsymbol{\beta}})$ by

$$\beta_j | \, \sigma_{\boldsymbol{\beta}}, b_i \overset{\text{ind.}}{\sim} N(0, \sigma^2_{\boldsymbol{\beta}}/b_i), \quad b_j | \, c_j \overset{\text{ind.}}{\sim} \text{Gamma}(\tfrac{1}{2}, c_i) \quad \text{and} \quad c_j \overset{\text{ind.}}{\sim} \text{Gamma}(\tfrac{1}{2}, 1).$$

We also continue to use Result 4 for auxiliary variable representation of Half Cauchy distributions. Figure 10 shows the corresponding directed acyclic graphs.

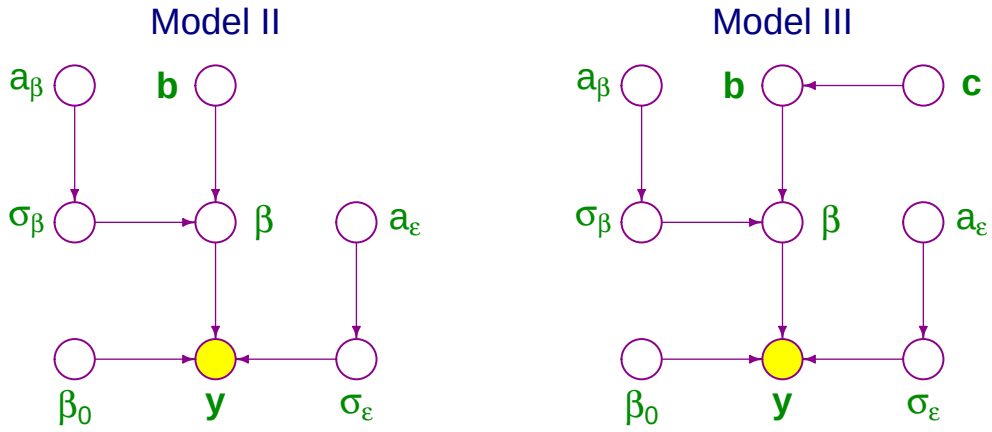

Figure 10: *Directed acyclic graphs corresponding to Model II and Model III representations of (11).*

The pitfalls caused by the high correlation between the b_j and c_j described in Section 2 still apply to the Model III version of (11), and this can impact the inference for nodes close to $\boldsymbol{b}$ and $\boldsymbol{c}$ on the graph — namely $\boldsymbol{\beta}$ and $\sigma_{\boldsymbol{\beta}}$.

We ran a small simulation study involving wavelet regression to see if, and to what degree, Model II and Model III differ in terms of quality of the regression fit. We generated 1000 samples according to

$$y_i = f_{\text{WO}}(x_i) + \varepsilon_i, \quad 1 \le i \le n,$$

where $x_i \sim$ Uniform(0,1) and $\varepsilon_i \sim N(0,1)$ and $n \in \{1000, 5000, 10000\}$. Here f_{WO} is the jagged/jumpy regression function used throughout Wand & Ormerod (2011) and defined by

$$\begin{aligned} f_{\text{WO}}(x) \quad \equiv \quad 18\Big[& \sqrt{x(1-x)}\sin(1.6\pi/(x+0.2)) + 0.4\, I(x > 0.13) \\ & -0.7\, I(0.32 < x < 0.38) + 0.43\{(1 - |(x - 0.65)/0.03|)_+\}^4 \\ & +0.42\{(1 - |(x - 0.91)/0.015|)_+\}^4\Big], \qquad 0 < x < 1, \end{aligned}$$

where $I(\mathcal{P}) = 1$ if $\mathcal{P}$ is true and zero otherwise. Estimation of f_{WO} involved MFVB fitting of (11) with $\boldsymbol{X}$ containing 255 Daubechies 5 wavelet basis functions applied to the x_is using the construction described in Section 3.1 of Wand & Ormerod (2011) with $L = 8$ levels. The quality of the resulting estimator, $\widehat{f}_{\text{WO}}$, was measured using the average squared error:

$$n^{-1} \sum_{i=1}^{n} \{\widehat{f}_{\text{WO}}(x_i) - f_{\text{WO}}(x_i)\}^2.$$

Figure 11 provides a visual summary of the simulation results, with the ratios of the average squared error values plotted as boxplots for each sample size. Model II is the clear winner, with a superior average squared error performance across all 1000 replications. The advantage of Model II is seen to be greater for lower sample sizes.

We are planning to conduct a large-scale simulation to assess the fuller implications of Model II versus Model III for sparse signal regression, particularly when p is very high, but this is yet to be carried out.

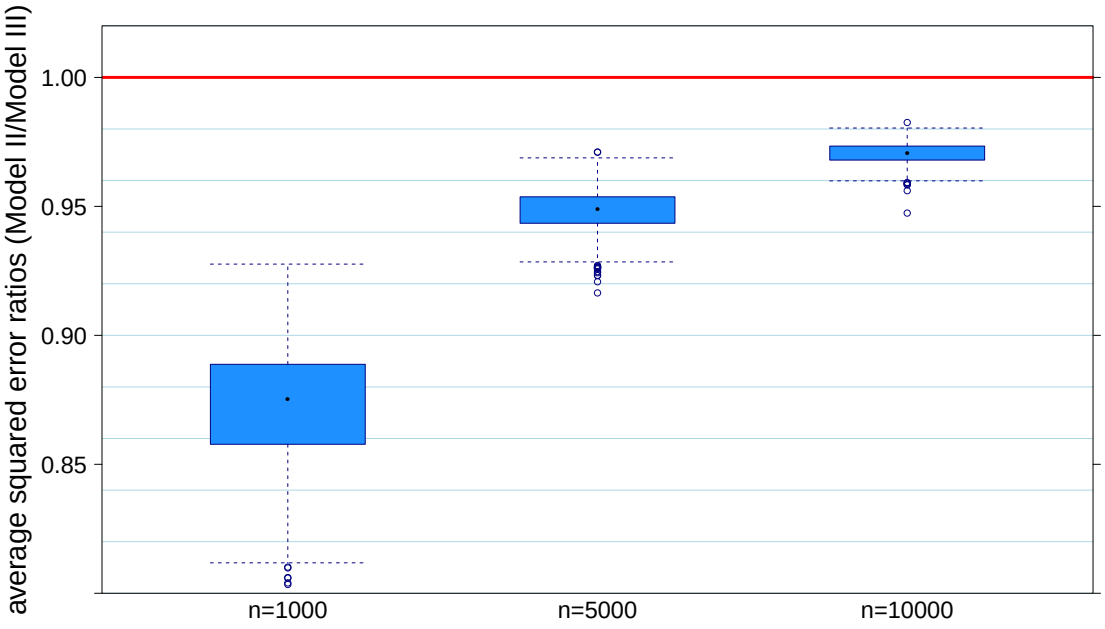

Figure 11: *Boxplots of ratios of average squared errors for the wavelet regression simulation study described in the text. The Model II errors are divided by the Model III errors.*

Appendix A: Background

In this appendix we assemble all special function and distributional definitions and results required for the study given in Section 2. We also provide a brief description of mean field variational Bayes.

A.1 Special function results

Continuous sparse signal shrinkage distributions depend on certain special functions. Additionally, the superior MFVB algorithms that we develop for models containing such distributions involve ratios of such functions. The necessary background material is laid out here.

A.1.1 Special function definitions

We now define all special functions used in later sections. We follow the conventions and notation of Gradshteyn & Ryzhik (1994). Their evaluation in the R computing environment (R Development Core Team, 2014), which is now ubiquitous in mainstream Statistics, is also dealt with.

The *exponential integral function of order 1*, E_1, is defined by

$$E_1(x) \equiv \int_x^\infty \frac{e^{-t}}{t}\, dt, \quad x \in \mathbb{R},\ x \neq 0.$$

Evaluation of E_1 is supported by the function `expint_E1()` in the R package **gsl** (Hankin, 2007), which uses the GNU Scientific Library (Galassi *et al.*, 2009).

The *parabolic cylinder function* of order $\nu \in \mathbb{R}$, is denoted by D_ν. The parabolic cylinder functions of *negative order* admit the integral expression

$$D_\nu(x) = \Gamma(-\nu)^{-1} \exp(-x^2/4) \int_0^\infty t^{-\nu-1} \exp(-xt - \tfrac{1}{2}t^2)\, dt, \quad \nu < 0,\ x \in \mathbb{R}.$$

Note that only such negative order members of the parabolic cylinder family arise in the present article. Consequently, we have the relationship

$$\int_0^\infty x^p \exp(qx - rx^2)\, dx = (2r)^{-(p+1)/2}\Gamma(p+1)\exp\{q^2/(8r)\}D_{-p-1}(-q/\sqrt{2r}), \quad p > -1, q \in \mathbb{R}, r > 0. \tag{12}$$

Note that

$$D_\nu(x) = 2^{\nu/2+1/4}\, W_{\nu/2+1/4,\,-1/4}(\tfrac{1}{2}\,x^2)/\sqrt{x}, \quad x > 0, \tag{13}$$

where $W_{k,m}$ is a *confluent hypergeometric function* as defined in Whittaker & Watson (1990). Due to (13) and $W_{k,m}$ being supported by the R function `whittakerW()` within the package **fAsianOptions** (Wuertz *et al.*, 2009), evaluation of $D_\nu(x)$, for $\nu \in \mathbb{R}$, $x > 0$, can be achieved via:

```
library(fAsianOptions)
2^(nu/2+1/4)*Re(whittakerW(x^2/2,nu/2+1/4,-1/4))/sqrt(x)
```

where `nu` and `x` denote the respective values of ν and x.

Gauss's hypergeometric function of order (α, β, γ) has an infinite series definition (Gradshteyn & Ryzhik, 1994), but has the integral representation

$${}_2F_1(\alpha, \beta; \gamma; x) = \frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)} \int_0^1 (1-tx)^{-\alpha}\, t^{\beta-1}(1-t)^{\gamma-\beta-1}\, dt \quad \text{for} \quad \gamma > \beta > 0.$$

Section 9.130 of Gradshteyn & Ryzhik (1994) gives conditions under which ${}_2F_1(\alpha, \beta; \gamma; x)$ converges. Evaluation of ${}_2F_1(\alpha, \beta; \gamma; \cdot)$ is supported by the function `hyperg_2F1` in the R package **gsl**.

A.1.2 Additional function definitions and continued fraction representations

The following new function definitions permit convenient listing and analysis of our MFVB algorithms in Sections 2.1–2.3:

$$\begin{aligned} \mathcal{Q}(x) &\equiv e^x\, E_1(x), \quad x > 0, \\ \text{and} \quad \mathcal{R}_\nu(x) &\equiv \frac{D_{-\nu-2}(x)}{D_{-\nu-1}(x)}, \quad \nu > 0,\ x > 0. \end{aligned} \tag{14}$$

Whilst both of these functions are simple forms involving special functions, care needs to be taken with their computation, as we now explain.

First note that $\mathcal{Q}$ can be written as

$$\mathcal{Q}(x) = \frac{E_1(x)}{\exp(-x)}, \quad x > 0. \tag{15}$$

As is well-known, the denominator on the right-hand side of (15) is strictly positive, and rapidly approaches zero as $x \to \infty$. Unfortunately, the numerator has the same properties, and accurate evaluation of the ratio is impeded by underflow for large x. A remedy would be to work with $\log\{E_1(x)\}$, but we know of no established software for accurate computation of this function for large positive x. Fortunately, $\mathcal{Q}(x)$ admits the simple continued fraction expansion:

$$\mathcal{Q}(x) = \cfrac{1}{x + 1 - \cfrac{1^2}{x + 3 - \cfrac{2^2}{x + 5 - \cfrac{3^2}{x + 7 - \cdots}}}} \tag{16}$$

(Equation (14.1.23) of Cuyt *et al.*, 2008).

Analogous underflow problems afflict direct computation of $\mathcal{R}_\nu(x)$, and for its stable computation we call upon:

$$\mathcal{R}_\nu(x) = \cfrac{1}{x + \cfrac{\nu+2}{x + \cfrac{\nu+3}{x + \cfrac{\nu+4}{x + \cdots}}}} \tag{17}$$

(Equation (16.5.7) of Cuyt *et al.*, 2008).

Algorithms 2 and 4 achieve practical computation of $\mathcal{Q}$ and $\mathcal{R}_\nu$ based on these continued fraction representations.

A succinct summary of continued fraction enhancement of Bayesian computing is given in Wand & Ormerod (2012).

A.2 Distributional definitions and results

Mean field variational Bayes for models containing continuous sparse shrinkage distributions depend on certain special functions and distributional results, which we give here.

A.2.1 Continuous sparse signal shrinkage density functions

The standard Horseshoe density function is

$$p_{\text{HS}}(x) = (2\pi^3)^{-1/2} \exp(x^2/2) E_1(x^2/2). \tag{18}$$

If the random variable x has density function $\sigma^{-1} p_{\text{HS}}((x-\mu)/\sigma)$ then we write

$$x \sim \text{Horseshoe}(\mu, \sigma).$$

The standard Normal-Exponential-Gamma density function, with shape parameter $\lambda > 0$, is

$$p_{\text{NEG}}(x;\lambda) = \pi^{-1/2} \lambda\, 2^\lambda \Gamma(\lambda + \tfrac{1}{2})\, \exp(x^2/4)\, D_{-2\lambda-1}(|\,x|). \tag{19}$$

If the random variable x has density function $\sigma^{-1} p_{\text{NEG}}((x-\mu)/\sigma;\lambda)$ then we write

$$x \sim \text{NEG}(\mu, \sigma, \lambda).$$

The standard Generalized Double Pareto density function is

$$p_{\text{GDP}}(x;\lambda) = \frac{1}{2(1+|x|/\lambda)^{\lambda+1}}. \tag{20}$$

If the random variable x has density function $\sigma^{-1} p_{\text{GDP}}((x-\mu)/\sigma)$ then we write

$$x \sim \text{GDP}(\mu, \sigma, \lambda).$$

Figure 12 depicts standard ($\mu = 0$, $\sigma = 1$) Horseshoe, Normal-Exponential-Gamma and Generalized Double Pareto density functions with varying values of corresponding shape parameters. For the Normal-Exponential-Gamma and Generalized Double Pareto distributions a decrease of the shape parameter λ results in a density function having higher kurtosis.

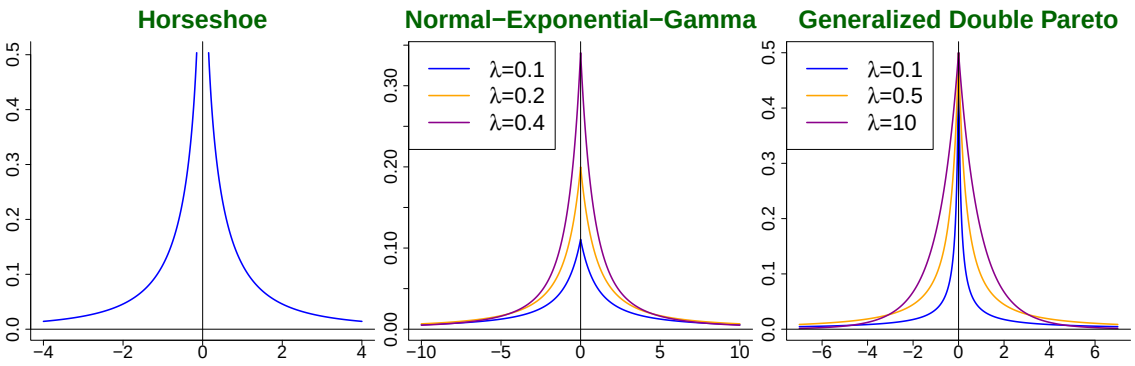

Figure 12: *Left panel: the standard Horseshoe density function. Middle panel: three standard Normal-Exponential-Gamma density functions with varying shape parameter λ. Right panel: three standard Double Generalized Pareto density functions with varying shape parameter λ.*

A.2.2 Related distributional results

The notation $v \sim \text{Gamma}(A, B)$ means that v has a Gamma distribution with *shape* parameter $A > 0$ and *rate* parameter $B > 0$. The corresponding density function is

$$p(v) = B^A \Gamma(A)^{-1} v^{A-1} \exp(-Bv), \quad v > 0.$$

The notation $v \sim \text{IG}(A, B)$ means that v has an Inverse-Gamma distribution with *shape* parameter $A > 0$ and *rate* parameter $B > 0$. The corresponding density function is

$$p(v) = B^A \Gamma(A)^{-1} v^{-A-1} \exp(-B/v), \quad v > 0.$$

Note that $v \sim \text{IG}(A, B)$ if and only if $1/v \sim \text{Gamma}(A, B)$.

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

$$x|\, b \sim N(\mu, \sigma^2/b), \quad b\,|\, c \sim \text{Gamma}(\tfrac{1}{2}, c) \quad \textit{and} \quad c \sim \text{Gamma}(\tfrac{1}{2}, 1).$$

Then $x \sim$ Horseshoe(μ, σ).

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

$$x|\, b \sim N(\mu, \sigma^2/b) \quad \textit{and} \quad p(b) = \pi^{-1} b^{-1/2} (b+1)^{-1}, \quad b > 0.$$

Then $x \sim$ Horseshoe(μ, σ).

Results 1a and 1b follow from results on the Horseshoe distribution given in Carvalho, Polson & Scott (2010) as well as Proposition 1 of Armagan, Dunson & Clyde (2011) or Result 5 of Wand *et al.* (2011).

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

$$x|\, b \sim N(\mu, \sigma^2/b), \quad b\,|\, c \sim \text{IG}(1, c) \quad \textit{and} \quad c \sim \text{Gamma}(\lambda, 1).$$

Then $x \sim$ NEG(μ, σ, λ).

Result 2b. *Let x and b such that*

$$x|\, b \sim N(\mu, \sigma^2/b) \quad \textit{and} \quad p(b) = \lambda\, b^{\lambda-1}\, (b+1)^{-\lambda-1}, \quad b > 0.$$

Then $x \sim$ NEG(μ, σ, λ).

Results 2a and 2b are related to results on the Normal-Exponential-Gamma distribution given in Griffin & Brown (2011).

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

$$x|\, b \sim N(\mu, \sigma^2/b), \quad b\,|\, c \sim \text{IG}(1, \tfrac{1}{2}\, c^2) \quad \textit{and} \quad c \sim \text{Gamma}(\lambda, \lambda).$$

Then $x \sim GDP(\mu, \sigma, \lambda)$.

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

$$x|\, b \sim N(\mu, \sigma^2/b) \quad \textit{and} \quad p(b) = \tfrac{1}{2}(\lambda+1)\lambda^{\lambda+1}\, b^{(\lambda-2)/2}\, e^{\lambda^2\, b/4} D_{-\lambda-2}(\lambda\sqrt{b}), \quad b > 0.$$

Then $x \sim GDP(\mu, \sigma, \lambda)$.

Results 3a and 3b are related to results on the Generalized Double Pareto distribution given in Armagan, Dunson & Lee (2013).

A.2.3 Half-Cauchy distribution

The notation $v \sim \text{Half-Cauchy}(A)$ means that v has a Half Cauchy distribution with scale parameter $A > 0$. The corresponding density function is

$$p(x) = \frac{2A}{\pi(x^2 + A^2)}, \quad x > 0.$$

We use the Half-Cauchy family to impose non-informative priors on scale parameters.

The following result, a special case of Proposition 1 in Armagan, Dunson & Clyde (2011) and Result 5 in Wand *et al.* (2011), is useful for mean field variational Bayes calculations for models containing Half Cauchy random variables:

Result 4. *Let x and a be random variables such that*

$$x|\, a \sim \text{IG}(\tfrac{1}{2}, a^{-1}) \quad \textit{and} \quad a \sim \text{IG}(\tfrac{1}{2}, A^{-2}).$$

Then $\sqrt{x} \sim$ Half-Cauchy(A).

A.3 Mean field variational Bayes

Consider a Bayesian model (graphical model) with observed data vector $\boldsymbol{x}$ (evidence node), parameter θ and auxiliary variable vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Concrete examples of such a model are given in Sections 2.1–2.3. Typically, the joint posterior density function $p(\theta, \boldsymbol{a}, \boldsymbol{b}|\boldsymbol{x})$ is intractable. A mean field variational approach postulates an approximation such as

$$p(\theta, \boldsymbol{a}, \boldsymbol{b}|\, \boldsymbol{x}) \approx q(\theta)\, q(\boldsymbol{a}, \boldsymbol{b}) \tag{21}$$

and chooses densities $q(\theta)$ and $q(\boldsymbol{a}, \boldsymbol{b})$ to minimize the following Kullback-Liebler distance between the two joint density functions:

$$\int q(\theta)\, q(\boldsymbol{a}, \boldsymbol{b})\, \log\left\{\frac{q(\theta)\, q(\boldsymbol{a}, \boldsymbol{b})}{p(\theta, \boldsymbol{a}, \boldsymbol{b}|\boldsymbol{x})}\right\} d\theta\, d\boldsymbol{a}\, d\boldsymbol{b}.$$

The solutions can be shown to satisfy

$$\begin{array}{rrcl} & q^*(\theta) & \propto & \exp\{E_{q(\boldsymbol{a},\boldsymbol{b})} p(\theta|\, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b})\} \\ \text{and} & q^*(\boldsymbol{a}, \boldsymbol{b}) & \propto & \exp\{E_{q(\theta)} p(\boldsymbol{a}, \boldsymbol{b}|\, \boldsymbol{x}, \theta)\}. \end{array} \tag{22}$$

These conditions gives rise to an iterative coordinate ascent algorithm which is guaranteed to converge under mild conditions. Convergence can be monitored using the following lower bound on the marginal log-likelihood:

$$\log \underline{p}(\boldsymbol{x}; q) \equiv E_{q(\theta,\boldsymbol{a},\boldsymbol{b})}[\log p(\boldsymbol{x}, \theta, \boldsymbol{a}, \boldsymbol{b}) - \log\{q(\theta)\, q(\boldsymbol{a}, \boldsymbol{b})\}] \leq \log p(\boldsymbol{x}),$$

since each iteration leads to an improvement in the bound. Several illustrative examples are given in Section 2.2 of Ormerod & Wand (2010). The number of iterations required for convergence with strict tolerances is typically in the tens or hundreds.

Note that it is possible that the optimal density $q^*(\boldsymbol{a}, \boldsymbol{b})$ factorizes as $q^*(\boldsymbol{a})\, q^*(\boldsymbol{b})$ even though this restriction is not imposed by (21). This is known as *induced* factorization in the MFVB literature (e.g. Section 10.2.5 of Bishop, 2006).

Ease of implementation and speed depends on the ease with which the expectations in (22) can be evaluated. For simple models involving common distributions, the expectations often admit explicit forms – in which case computation can be quite rapid. Models involving more complicated distributions may be such that quadrature or Monte Carlo is required, which tends to compromise speed.

Description of the MFVB algorithms in the upcoming sections benefit from notation such as

$$\mu_{q(v)} \equiv \int_{-\infty}^{\infty} v\, q(v)\, dv,\ \mu_{q(v^2)} \equiv \int_{-\infty}^{\infty} v^2\, q(v)\, dv. \quad \text{and} \quad \sigma^2_{q(v)} \equiv \int_{-\infty}^{\infty} \{v - \mu_{q(v)}\}^2\, q(v)\, dv.$$

Appendix B: Mean Field Variational Bayes Derivations

Algorithms 1, 3 and 5 depend on the following derivations of the optimal density functions and relevant moments under product restrictions (2). Throughout the derivations, the symbol 'rest' denotes all other random variables in the Bayesian model at hand. Constants with respect to the function argument are denoted by 'const'.

The notation $v \sim \text{Inverse-Gaussian}(\mu, \gamma)$ means that v has an Inverse-Gaussian distribution with mean μ and rate parameter γ. The corresponding density function is

$$p(v) = \sqrt{\frac{\gamma}{2\pi\, v^3}} \exp\left\{-\frac{\gamma(v-\mu)^2}{2\mu^2 v}\right\}, \quad v > 0,$$

and is such that $E(v) = \mu$ and $E(1/v) = 1/\mu + 1/\gamma$.

B.1 Horseshoe Models

The full conditional of a satisfies

$$\log p(a|\text{rest}) = -2\log(a) - (\sigma^{-2} + A^{-2})/a + \text{const.}$$

For Models II and III, the full conditional of σ^2 satisfies

$$\log p(\sigma^2|\text{rest}) = -\tfrac{1}{2}\,(n+3)\log(\sigma^2) - \left(\tfrac{1}{2}\sum_{i=1}^{n} b_i\, x_i^2 + a^{-1}\right)\Big/\sigma^2 + \text{const.}$$

For Model II, the full conditionals of the b_is satisfy

$$\log p(b_i|\text{rest}) = -\log(b_i+1) - \frac{b_i\, x_i^2}{2\sigma^2} + \text{const.}$$

For Model III, the full conditionals of the b_is and c_is satisfy

$$\log p(b_i|\text{rest}) = -\left(\frac{x_i^2}{2\sigma^2} + c_i\right) b_i + \text{const.}$$

and

$$\log p(c_i|\text{rest}) = -\,(b_i+1)\, c_i + \text{const.}$$

B.1.1 Expressions for $q^*(a)$ and $\mu_{q(1/a)}$

$$q^*(a) \sim \text{IG}\left(1, \mu_{q(1/\sigma^2)} + A^{-2}\right)$$

and

$$\mu_{q(1/a)} = 1 \Big/ \left\{\mu_{q(1/\sigma^2)} + A^{-2}\right\}.$$

Derivations:

$$E_q\{\log p(a|\text{rest})\} = E_q\left\{-2\log(a) - \left(\sigma^{-2} + A^{-2}\right)\Big/ a\right\} + \text{const}$$

and so

$$q^*(a) \propto a^{-2} \exp\left\{-\left(\mu_{q(1/\sigma^2)} + A^{-2}\right)\Big/ a\right\}.$$

Standard manipulations involving the Inverse Gamma family of density functions lead to the stated results.

B.1.2 Expressions for $q^*(\sigma^2)$ and $\mu_{q(1/\sigma^2)}$ for Models II and III

$$q^*(\sigma^2) \sim \text{IG}\left(\tfrac{1}{2}(n+1), \tfrac{1}{2}\sum_{i=1}^{n} x_i^2\, \mu_{q(b_i)} + \mu_{q(1/a)}\right)$$

and

$$\mu_{q(1/\sigma^2)} = \tfrac{1}{2}(n+1) \Bigg/ \left\{\tfrac{1}{2}\sum_{i=1}^{n} x_i^2\, \mu_{q(b_i)} + \mu_{q(1/a)}\right\}.$$

Derivations:

$$E_q\{\log p(\sigma^2|\text{rest})\} = E_q\left[-\tfrac{1}{2}(n+3)\log(\sigma^2) - \left(\tfrac{1}{2}\sum_{i=1}^{n} x_i^2 b_i + a^{-1}\right)\Big/\sigma^2\right] + \text{const}$$

and so

$$q^*(\sigma^2) \propto (\sigma^2)^{-\frac{1}{2}(n+3)} \exp\left[-\left\{\tfrac{1}{2}\sum_{i=1}^{n} x_i^2\, \mu_{q(b_i)} + \mu_{q(1/a)}\right\}\Big/\sigma^2\right].$$

Standard manipulations involving the Inverse Gamma family of density functions lead to the stated results.

B.1.3 Expressions for $q^*(b_i)$ and $\mu_{q(b_i)}$ for Model II

$$q^*(b_i) = \frac{1}{(b_i+1)\,\exp\{G_i\,(b_i+1)\}E_1(G_i)}, \quad b_i > 0$$

and

$$\mu_{q(b_i)} = \frac{1}{G_i\,\exp(G_i)\,E_1(G_i)} - 1$$

where

$$G_i \equiv \tfrac{1}{2}\,\mu_{q(1/\sigma^2)}\,x_i^2. \tag{23}$$

Derivations:

$$E_q\{\log p(b_i|\text{rest})\} = E_q\left\{-\log(b_i+1) - \frac{x_i^2\, b_i}{2\sigma^2}\right\} + \text{const}$$

and so

$$\begin{aligned} q^*(b_i) &\propto (b_i+1)^{-1} \exp\left[-\tfrac{1}{2}\, \mu_{q(1/\sigma^2)} \{ (x_i - \mu_{q(\mu)})^2 + \sigma^2_{q(\mu)} \}\, b_i \right], \quad b_i > 0 \\ &= (b_i+1)^{-1} \exp(-G_i\, b_i), \quad b_i > 0. \end{aligned}$$

where G_i is given by (23). The normalizing factor is

$$\int_0^\infty (b_i+1)^{-1} \exp(-G_i\, b_i)\, db_i = \exp(G_i)\, E_1(G_i),$$

which follows from 3.352(4) of Gradshteyn & Ryzhik (1994).

The numerator of $\mu_{q(b_i)}$ is

$$\int_0^\infty b_i (b_i+1)^{-1} \exp(-G_i, b_i)\, db_i = G_i^{-1} - \exp(G_i) E_1\left(G_i\right),$$

by application of 3.353(5) of Gradshteyn & Ryzhik (1994). The stated results then follow immediately.

B.1.4 Expressions for $q^*(b_i)$ and $\mu_{q(b_i)}$ for Model III

$$q^*(b_i) \sim \text{Gamma}(1, G_i + \mu_{q(c_i)}), \quad b_i > 0,$$

and

$$\mu_{q(b_i)} = 1 \Big/ \left\{ G_i + \mu_{q(c_i)} \right\}$$

where G_i is given by (23).

Derivations:

$$E_q\{\log p(b_i|\text{rest})\} = E_q \left[-\left\{ \frac{x_i^2}{2\sigma^2} + c_i \right\} b_i \right] + \text{const}$$

and so

$$q^*(b_i) \propto \exp[-\{G_i + \mu_{q(c_i)}\}\, b_i], \quad b_i > 0.$$

Standard manipulations involving the Gamma family of density functions lead to the stated results.

B.1.5 Expressions for $q^*(c_i)$ and $\mu_{q(1/c_i)}$ for Model III

$$q^*(c_i) \sim \text{Gamma}(1, \mu_{q(b_i)} + 1)$$

and

$$\mu_{q(c_i)} = 1 \Big/ \left\{ \mu_{q(b_i)} + 1 \right\}.$$

Derivations:

$$E_q\{\log p(c_i|\text{rest})\} = E_q \left\{ -\left(1 + b_i\right) c_i \right\} + \text{const}.$$

Hence

$$q^*(c_i) \propto \exp[-\{\mu_{q(b_i)} + 1\}\, c_i], \quad c_i > 0.$$

The stated results follow from properties of the Inverse Gamma family of distributions.

B.2 Normal-Exponential-Gamma Models

The calculations for $q^*(\sigma^2)$ and $q(a)$ are identical to those for the Horseshoe models.

For Model II, the full conditionals of the b_is satisfy

$$\log p(b_i|\text{rest}) = (\lambda - \tfrac{1}{2})\log(b_i) - (\lambda+1)\log(b_i+1) - \frac{b_i\, x_i^2}{2\sigma^2} + \text{const.}$$

For Model III, the full conditionals of the b_is and c_is satisfy

$$\log p(b_i|\text{rest}) = -\tfrac{3}{2}\log(b_i) - \frac{b_i\, x_i^2}{2\sigma^2} - \frac{c_i}{b_i} + \text{const}$$

and

$$\log p(c_i|\text{rest}) = \lambda\, \log(c_i) - \left(\frac{1}{b_i} + 1\right) c_i + \text{const.}$$

B.2.1 Expressions for $q^*(b_i)$ and $\mu_{q(b_i)}$ for Model II

$$q^*(b_i) = \frac{b_i^{\lambda-1/2}(b_i+1)^{-\lambda-1}\, \exp(-G_i\, b_i)}{2^{\lambda+\frac{1}{2}}\Gamma(\lambda+\frac{1}{2})\exp(G_i/2)\, D_{-2\lambda-1}\left(\sqrt{2G_i}\right)}, \quad b_i > 0$$

and

$$\mu_{q(b_i)} = \frac{(2\lambda+1)\, D_{-2\lambda-2}(\sqrt{2G_i}\,)}{\sqrt{2G_i}\, D_{-2\lambda-1}\left(\sqrt{2G_i}\right)}$$

where G_i is given by (23).

Derivations:

$$E_q\{\log p(b_i|\text{rest})\} = E_q\left\{(\lambda-\tfrac{1}{2})\log(b_i) - (\lambda+1)\log(b_i+1) - \frac{b_i\, x_i^2}{2\sigma^2}\right\} + \text{const}$$

and so

$$q^*(b_i) \propto b_i^{\lambda-1/2}(b_i+1)^{-\lambda-1}\, \exp(-G_i\, b_i), \quad b_i > 0.$$

The normalizing factor is

$$\int_0^\infty b_i^{(\lambda+\frac{1}{2})-1}(b_i+1)^{-(\lambda+\frac{1}{2})-\frac{1}{2}}\, \exp(-G_i\, b_i)\, db_i$$
$$= 2^{\lambda+\frac{1}{2}}\Gamma(\lambda+\tfrac{1}{2})\exp(G_i/2)\, D_{-2\lambda-1}\left(\sqrt{2G_i}\right)$$

which follows from 3.383(7) of Gradshteyn & Ryzhik (1994).

The numerator of $\mu_{q(b_i)}$ is

$$\int_0^\infty b_i^{(\lambda+\frac{3}{2})-1}(b_i+1)^{-(\lambda+\frac{3}{2})+\frac{1}{2}}\, \exp(-G_i\, b_i)\, db_i$$
$$= 2^{\lambda+\frac{3}{2}}\Gamma(\lambda+\tfrac{3}{2})\exp(G_i/2)\, D_{-2\lambda-2}\left(\sqrt{2G_i}\right)\Big/\sqrt{2G_i}$$

with the last line being an application of 3.383(6) of Gradshteyn & Ryzhik (1994). The stated result then follows from the fact that

$$2^{\lambda+\frac{3}{2}}\Gamma(\lambda+\tfrac{3}{2})/\{2^{\lambda+\frac{1}{2}}\Gamma(\lambda+\tfrac{1}{2})\} = 2\lambda+1.$$

B.2.2 Expressions for $q^*(b_i)$, $\mu_{q(1/b_i)}$ and $\mu_{q(b_i)}$ for Model III

$$q^*(b_i) \sim \text{Inverse-Gaussian}\left(\sqrt{\frac{\mu_{q(c_i)}}{G_i}}\,,\, 2\mu_{q(c_i)}\right),$$

$$\mu_{q(b_i)} = \sqrt{\frac{\mu_{q(c_i)}}{G_i}} \quad \text{and} \quad \mu_{q(1/b_i)} = \frac{1}{\mu_{q(b_i)}} + \frac{1}{2\mu_{q(c_i)}}$$

where G_i is given by (23).

Derivations:

$$E_q\{\log p(b_i|\text{rest})\} = E_q\left\{-\tfrac{3}{2}\,\log(b_i) - \frac{x_i^2\,b_i}{2\sigma^2} - \frac{c_i}{b_i}\right\} + \text{const}$$

and so

$$q^*(b_i) \propto b_i^{-3/2} \exp\left\{-\,G_i\,b_i - \frac{\mu_{q(c_i)}}{b_i}\right\}, \quad b_i > 0.$$

Standard manipulations involving the Inverse Gaussian family of density functions lead to the stated results.

B.2.3 Expressions for $q^*(c_i)$ and $\mu_{q(c_i)}$ for Model III

$$q^*(c_i) \sim \text{Gamma}(\lambda + 1, \mu_{q(1/b_i)} + 1)$$

and

$$\mu_{q(c_i)} = \frac{\lambda + 1}{\mu_{q(1/b_i)} + 1}.$$

Derivations:

$$E_q\{\log p(c_i|\text{rest})\} = E_q\left\{\lambda\,\log(c_i) - \left(\frac{1}{b_i} + 1\right)\,c_i\right\} + \text{const.}$$

Hence

$$q^*(c_i) \propto c_i^{(\lambda+1)-1}\,\exp[-\{\mu_{q(1/b_i)} + 1\}\,c_i], \quad c_i > 0,$$

which is proportional to the Gamma$(\lambda+1, \mu_{q(1/b_i)}+1)$ density function. The stated results follow from properties of the Gamma family of distributions.

B.3 Generalized Double Pareto Models

The calculations for $q^*(\sigma^2)$ and $q(a)$ are identical to those for the Horseshoe models.

For Model II, the full conditionals of the b_is satisfy

$$\log p(b_i|\text{rest}) = \tfrac{1}{2}\,(\lambda - 1)\,\log(b_i) + \left(\frac{\lambda^2}{4} - \frac{x_i^2}{2\sigma^2}\right)\,b_i + \log\,D_{-\lambda-2}\left(\lambda\,\sqrt{b_i}\right) + \text{const.}$$

For Model III, the full conditionals of the b_is and c_is satisfy

$$\log p(b_i|\text{rest}) = -\tfrac{3}{2}\,\log(b_i) - \frac{x_i^2\,b_i}{2\sigma^2} - \frac{c_i^2}{2b_i}$$

and

$$\log p(c_i|\text{rest}) = (\lambda + 1)\log(c_i) - \lambda\,c_i - \frac{c_i^2}{2b_i} + \text{const.}$$

B.3.1 Expressions for $q^*(b_i)$ and $\mu_{q(b_i)}$ for Model II

$$q^*(b_i) = \frac{2^{(3\lambda+2)/2}(\lambda+1)G_i^{(\lambda+1)/2}\, b_i^{(\lambda-1)/2} \exp\{(\lambda^2/4 - G_i)b_i\} D_{-\lambda-2}(\lambda\sqrt{b_i})}{\sqrt{\pi}\, {}_2F_1\left(\frac{\lambda+2}{2}, \frac{\lambda+1}{2}; \lambda+2; 1-\lambda^2(2G_i)\right)}, \quad b_i > 0$$

and

$$\mu_{q(b_i)} = \frac{\lambda+1}{\sqrt{2G_i}(\lambda+\sqrt{2G_i})}.$$

Derivations:

$$E_q\{\log p(b_i|\text{rest})\} = E_q\left\{\tfrac{1}{2}(\lambda-1)\,\log(b_i) + \left(\frac{\lambda^2}{4} - \frac{x_i^2}{2\sigma^2}\right) b_i + \log\, D_{-\lambda-2}\left(\lambda\sqrt{b_i}\right)\right\} + \text{const}$$

and so

$$q^*(b_i) \propto b_i^{(\lambda-1)/2} \exp\left\{\left(\tfrac{1}{4}\lambda^2 - G_i\right)\, b_i\right\}\, D_{-\lambda-2}(\lambda\sqrt{b_i}), \quad b_i > 0.$$

The expression for the normalizing factor follows from 7.725(6) of Gradshteyn & Ryzhik (1994), existence of the hypergeometric function for all $G_i > 0$ depends on Stieltjes integral transform theory described in Sections 5.2 and 15.2 of Cuyt *et al.* (2008).

Therefore

$$\begin{aligned}
\mu_{q(b_i)} &= \frac{\int_0^\infty b_i^{(\lambda+1)/2} \exp\left\{\left(\tfrac{1}{4}\lambda^2 - G_i\right)\, b_i\right\}\, D_{-\lambda-2}(\lambda\sqrt{b_i})\, db_i}{\int_0^\infty b_i^{(\lambda-1)/2} \exp\left\{\left(\tfrac{1}{4}\lambda^2 - G_i\right)\, b_i\right\}\, D_{-\lambda-2}(\lambda\sqrt{b_i})\, db_i} \\
&= \frac{\int_0^\infty e^{-zt} t^{-1+\beta_n/2}\, D_{-\nu}(2\sqrt{kt})\, dt}{\int_0^\infty e^{-zt} t^{-1+\beta_d/2}\, D_{-\nu}(2\sqrt{kt})\, dt}
\end{aligned}$$

where

$$z = G_i - \tfrac{1}{4}\lambda^2, \quad \beta_n = \lambda+3, \quad \beta_d = \lambda+1, \quad \nu = \lambda+2 \quad \text{and} \quad k = \tfrac{1}{4}\lambda^2.$$

Application of 7.725(6) of Gradshteyn & Ryzhik (1994) to the numerator and denominator results in the expression

$$\mu_{q(b_i)} = \frac{(\lambda+1)\, {}_2F_1\left(\frac{1}{2}\lambda+1, \frac{1}{2}\lambda+\frac{3}{2}; \lambda+3\,; 1-\lambda^2/(2G_i)\right)}{4\,G_i\, {}_2F_1\left(\frac{1}{2}\lambda+1, \frac{1}{2}\lambda+\frac{1}{2}; \lambda+2\,; 1-\lambda^2/(2G_i)\right)}. \tag{24}$$

Results 15.1.1 and 15.1.13 of Abramowitz & Stegun (1972) are, respectively,

$${}_2F_1(a,b;c;x) = {}_2F_1(b,a;c;x) \quad \text{and} \quad {}_2F_1\left(a, a+\tfrac{1}{2}, 2a+1, x\right) = 2^{2a}\left(1+\sqrt{1-x}\right)^{-2a}.$$

These imply that

$$\begin{aligned}
{}_2F_1\left(\tfrac{1}{2}\lambda+1, \tfrac{1}{2}\lambda+\tfrac{3}{2}; \lambda+3; x\right) &= 2^{\lambda+2}\left(1+\sqrt{1-x}\right)^{-(\lambda+2)} \quad \text{and} \\
{}_2F_1\left(\tfrac{1}{2}\lambda+1, \tfrac{1}{2}\lambda+\tfrac{1}{2}; \lambda+2; x\right) &= {}_2F_1\left(\tfrac{1}{2}\lambda+\tfrac{1}{2}, \tfrac{1}{2}\lambda+1; \lambda+2; x\right) = 2^{\lambda+1}\left(1+\sqrt{1-x}\right)^{-(\lambda+1)}.
\end{aligned}$$

The stated result for $\mu_{q(b_i)}$ follows immediately.

B.3.2 Expressions for $q^*(b_i)$, $\mu_{q(1/b_i)}$ and $\mu_{q(b_i)}$ for Model III

$$q^*(b_i) \sim \text{Inverse-Gaussian}\left(\sqrt{\frac{\mu_{q(c_i^2)}}{2G_i}}\,,\, \mu_{q(c_i^2)}\right),$$

$$\mu_{q(b_i)} = \sqrt{\frac{\mu_{q(c_i^2)}}{2G_i}} \quad \text{and} \quad \mu_{q(1/b_i)} = \frac{1}{\mu_{q(b_i)}} + \frac{1}{\mu_{q(c_i^2)}}$$

where G_i is defined by (23).

Derivations:

$$E_q\{\log p(b_i|\text{rest})\} = E_q\left\{-\tfrac{3}{2}\,\log(b_i) - \frac{x_i^2\,b_i}{2\sigma^2} - \frac{c_i^2}{2\,b_i}\right\} + \text{const}$$

and so

$$q^*(b_i) \propto b_i^{-3/2}\exp\left\{-G_i\,b_i - \mu_{q(c_i^2)}/(2\,b_i)\right\}, \quad b_i > 0.$$

Standard manipulations involving the Inverse Gaussian family of density functions lead to the stated results.

B.3.3 Expressions for $q^*(c_i)$ and $\mu_{q(c_i^2)}$ for Model III

$$q^*(c_i) = \frac{\mu_{q(1/b_i)}^{(\lambda+2)/2}\,c_i^{\lambda+1}\,\exp\{-\lambda\,c_i - \frac{1}{2}\,\mu_{q(1/b_i)}\,c_i^2\}}{\Gamma(\lambda+2)\,\exp\{\lambda^2/(4\,\mu_{q(1/b_i)})\}\,D_{-\lambda-2}(\,\lambda/\sqrt{\mu_{q(1/b_i)}}\,)}, \quad c_i > 0,$$

and

$$\mu_{q(c_i^2)} = \frac{(\lambda+2)(\lambda+3)\,D_{-\lambda-4}(\lambda/\sqrt{\mu_{q(1/b_i)}}\,)}{\mu_{q(1/b_i)}\,D_{-\lambda-2}(\lambda/\sqrt{\mu_{q(1/b_i)}}\,)}.$$

Derivations:

$$E_q\{\log p(c_i|\text{rest})\} = E_q\left\{(\lambda+1)\log(c_i) - \lambda\,c_i - \frac{c_i^2}{2\,b_i}\right\} + \text{const.}$$

Hence

$$q^*(c_i) \propto c_i^{\lambda+1}\,\exp\{-\lambda\,c_i - \tfrac{1}{2}\,\mu_{q(1/b_i)}\,c_i^2\}, \quad c_i > 0.$$

From (12), the normalizing factor is

$$\int_0^\infty c_i^{\lambda+1}\,\exp\{-\lambda\,c_i - \tfrac{1}{2}\,\mu_{q(1/b_i)}\,c_i^2\}\,dc_i = \mu_{q(1/b_i)}^{-(\lambda+2)/2}\,\Gamma(\lambda+2)\,\exp\left(\frac{\lambda^2}{4\,\mu_{q(1/b_i)}}\right)\,D_{-\lambda-2}\left(\lambda/\sqrt{\mu_{q(1/b_i)}}\right)$$

and the expression for $q^*(c_i)$ follows. Another application of (12) results in

$$\int_0^\infty c_i^{\lambda+3}\,\exp\{-\lambda\,c_i - \tfrac{1}{2}\,\mu_{q(1/b_i)}\,c_i^2\}\,dc_i = \mu_{q(1/b_i)}^{-(\lambda+4)/2}\,\Gamma(\lambda+4)\,\exp\left(\frac{\lambda^2}{4\,\mu_{q(1/b_i)}}\right)\,D_{-\lambda-4}\left(\lambda/\sqrt{\mu_{q(1/b_i)}}\right),$$

which immediately leads to the stated result for $\mu_{q(c_i^2)}$.

Appendix C: Proof of Theorem 1

First note that

$$\begin{aligned} E\{\log(1/b)\log(c)|x\} &= \int_0^\infty\int_0^\infty \log(b)\log(c)\,p(b,c|x)\,db\,dc\Big/p_{\text{HS}}(x) \\ &= \int_0^\infty \log(1/b)p(x|\,b)\left\{\int_0^\infty \log(c)\,p(b|c)\,p(c)\right\}\,db\Big/p_{\text{HS}}(x) \end{aligned}$$

where $p_{\rm HS}$ is given by (18). The inner integral is

$$\int_0^\infty \log(c)\exp\{-c(b+1)\}\,dc = \pi^{-1}\,b^{-1/2}\,(b+1)^{-1}\{\psi(1)-\log(b+1)\}$$

where $\psi(x) \equiv \frac{d}{dx}\log\Gamma(x)$ is the digamma function. Substitution of the expressions for $p_{\rm HS}(x)$ and $p(x|b)$ then leads to the univariate integral expression

$$E\{\log(1/b)\log(c)|x=\pm\sqrt{2\kappa}\} = \frac{\widetilde{\widetilde{B}}_2(\kappa)-\psi(1)B_1(\kappa)}{B_0(\kappa)}$$

where

$$\widetilde{\widetilde{B}}_2(\kappa) \equiv \int_0^\infty \frac{\exp\{-\kappa\,(b+1)\}\,\log(b)\,\log(b+1)}{b+1}\,db.$$

and

$$B_j(\kappa) \equiv \int_0^\infty \frac{\exp\{-\kappa\,(b+1)\}\,\{\log(b)\}^j}{b+1}\,db, \quad j=0,1.$$

After obtaining similar expressions for

$$E\{\{\log(1/b)\}^j|x=\pm\sqrt{2\kappa}\} \quad \text{and} \quad E\{\{\log(c)\}^j|x=\pm\sqrt{2\kappa}\}, \quad j=1,2,$$

in terms of $B_j(\kappa)$ and

$$\widetilde{B}_j(\kappa) \equiv \int_0^\infty \frac{\exp\{-\kappa\,(b+1)\}\,\{\log(b+1)\}^j}{b+1}\,db, \tag{25}$$

straightforward algebraic manipulations then lead to

$$\begin{aligned}
&\text{Corr}\{\log(1/b),\log(c)|x=\pm\sqrt{2\kappa}\} \\
&= \frac{B_0(\kappa)\widetilde{\widetilde{B}}_2(\kappa)-B_1(\kappa)\widetilde{B}_1(\kappa)}{\sqrt{\{B_0(\kappa)B_2(\kappa)-B_1(\kappa)^2\}\{B_0(\kappa)\widetilde{B}_2(\kappa)-\widetilde{B}_1(\kappa)^2+\frac{1}{6}\pi^2B_0(\kappa)^2\}}} \\
&= \Big[1+\{B_0(\kappa)\,\widetilde{\widetilde{\Delta}}_2(\kappa)/\widetilde{D}(\kappa)\}+\{\widetilde{B}_1(\kappa)\,\widetilde{\Delta}_1(\kappa)/\widetilde{D}(\kappa)\}\Big]\Bigg(\Big[1-\{B_0(\kappa)\,\widetilde{\Delta}_2(\kappa)/\widetilde{D}(\kappa)\} \\
&\quad +2\{\widetilde{B}_1(\kappa)\,\widetilde{\Delta}_1(\kappa)/\widetilde{D}(\kappa)\}-\{\widetilde{\Delta}_1(\kappa)^2/\widetilde{D}(\kappa)\}\Big]\Big[1+\tfrac{1}{6}\{\pi^2B_0(\kappa)^2/\widetilde{D}(\kappa)\}\Big]\Bigg)^{-1/2}
\end{aligned}$$

where

$$\widetilde{D}(\kappa) \equiv B_0(\kappa)\widetilde{B}_2(\kappa)-\widetilde{B}_1(\kappa)^2, \quad \widetilde{\Delta}_1(\kappa) \equiv \widetilde{B}_1(\kappa)-B_1(\kappa),$$
$$\widetilde{\Delta}_2(\kappa) \equiv \widetilde{B}_2(\kappa)-B_2(\kappa) \quad \text{and} \quad \widetilde{\widetilde{\Delta}}_2(\kappa) \equiv \widetilde{\widetilde{B}}_2(\kappa)-\widetilde{B}_2(\kappa).$$

We then note that

$$\begin{aligned}
\widetilde{\Delta}_1(0) &= \int_0^\infty \frac{\log(b+1)-\log(b)}{b+1}\,db = \tfrac{1}{6}\pi^2, \\
\widetilde{\Delta}_2(0) &= \int_0^\infty \frac{\{\log(b+1)\}^2-\{\log(b)\}^2}{b+1}\,db = 0 \\
\text{and} \quad \widetilde{\widetilde{\Delta}}_2(0) &= \int_0^\infty \frac{\log(b+1)\{\log(b+1)-\log(b)\}}{b+1}\,db = -\zeta(3)
\end{aligned}$$

where $\zeta(x) \equiv \sum_{j=1}^\infty j^{-x}$ denotes the Riemann zeta function. From this we have, for example, that

$$\lim_{\kappa\to 0}\{\widetilde{\Delta}_1(\kappa)^2/\widetilde{D}(\kappa)\} = \tfrac{1}{36}\,\pi^4\lim_{\kappa\to 0}\{1/\widetilde{D}(\kappa)\}.$$

Next note that, via the substitution $v = \log(b+1)$ into (25), we have for $j = 0, 1, 2$:

$$\widetilde{B}_j(\kappa) = \int_0^\infty \exp(-\kappa\, e^v)\, v^j\, dv.$$

From this we have

$$\begin{aligned}
2\widetilde{D}(\kappa) &= \left\{\int_0^\infty \exp(-\kappa\, e^v)\, dv\right\}\left\{\int_0^\infty \exp(-\kappa\, e^w)\, w^2\, dw\right\} \\
&\quad + \left\{\int_0^\infty \exp(-\kappa\, e^v)\, v^2\, dv\right\}\left\{\int_0^\infty \exp(-\kappa\, e^w)\, dw\right\} \\
&\quad -2\left\{\int_0^\infty \exp(-\kappa\, e^v)\, v\, dv\right\}\left\{\int_0^\infty \exp(-\kappa\, e^w)\, w\, dw\right\} \\
&= \int_0^\infty \int_0^\infty \exp\{-\kappa\,(e^v + e^w)\}(v-w)^2\, dv\, dw \\
&> \exp\left(-2\kappa\, e^M\right) \int_0^M \int_0^M (v-w)^2\, dv\, dw = \exp\left(-2\kappa\, e^M\right)\, M^4/6
\end{aligned}$$

for any $M > 0$. Therefore,

$$\lim_{\kappa\to 0}\{1/\widetilde{D}(\kappa)\} \le (12/M^4) \lim_{\kappa\to 0} \exp\left(-2\kappa\, e^M\right) = 12/M^4.$$

Since M is arbitrary we must have $\lim_{\kappa\to 0}\{1/\widetilde{D}(\kappa)\} = 0$. Hence $\{\widetilde{\Delta}_1(\kappa)^2/D(\kappa)\}$ vanishes as $\kappa \to 0$. Similar arguments can be used to show

$$\lim_{\kappa\to 0}\{B_0(\kappa)\, \widetilde{\widetilde{\Delta}}_2(\kappa)/\widetilde{D}(\kappa)\} = \lim_{\kappa\to 0}\{\widetilde{B}_1(\kappa)\, \widetilde{\Delta}_1(\kappa)/\widetilde{D}(\kappa)\} = \lim_{\kappa\to 0}\{B_0(\kappa)^2/\widetilde{D}(\kappa)\} = 0$$

and Theorem 1 immediately follows.

Acknowledgments

We are grateful to David Balding, David Nott and Tung Pham for their inputs. This research was partially supported by Australian Research Council Discovery Project DP110100061 and an Australian Postgraduate Award.

References

Abramowitz, M. & Stegun, I.A. eds. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover Publications.

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.

Armagan, A., Dunson, D.B. & Clyde, M. (2011). Generalized beta mixtures of Gaussians. In J. Shawe-Taylor, R.S. Zamel, P. Bartlett, F. Pereira and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 523-531.

Armagan, A., Dunson, D.B. & Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, **23**, 119–143.

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.

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

Carbonetto, P. & Stephens, M. (2011). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, **6**, Number 4, 1–42.

Carvalho, C.M., Polson, N.G. & Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.

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

Cuyt, A., Petersen, V.B., Verdonk, B., Waadeland, H. & Jones, W.B. (2008). *Handbook of Continued Fractions for Special Functions*. New York: Springer.

Flandin, G. & Penny, W.D. (2007). Bayesian fMRI data analysis with sparse spatial basis function priors. *NeuroImage*, **34**, 1108–1125.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., Rossi, F. (2009). *GNU Scientific Library Reference Manual*, 3rd Edition, Version 1.12, Bristol UK: Network Theory.

Gradshteyn, I.S. & Ryzhik, I.M. (1994). *Tables of Integrals, Series, and Products*, 5th Edition, San Diego, California: Academic Press.

Griffin, J.E. & Brown, P.J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian and New Zealand Journal of Statistics*, **53**, 423–442.

Hankin, R.K.S. (2007). gsl 1.9. Wrapper for the Gnu Scientific Library. R package. `http://cran.r-project.org`

Johnstone, I.M. & Silverman, B.W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, **32**, 1594–1649.

Johnstone, I.M. & Silverman, B.W. (2005). Bayes selection of wavelet thresholds. *The Annals of Statistics*, **33**, 1700–1752.

Lentz, W.J. (1976). Generating Bessel functions in Mie scattering calculations using continued fractions. *Applied Optics*, **3**, 668–671.

Ligges, U., Thomas, A., Spiegelhalter, D., Best, N., Lunn, D., Rice, K. & Sturtz, S. (2011). BRugs 0.5: OpenBUGS and its R/S-PLUS interface BRugs. R package. `http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/2.13`

Logsdon, B.A., Hoffman, G.E. & Mezey, J.G. (2010). A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, **11**:58, 1–13.

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.

Neville, S.E. (2013). *Elaborate Distribution Semiparametric Regression via Mean Field Variational Bayes*. PhD Thesis, University of Wollongong.

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

Polson, N.G. & Scott, J.G. (2010). Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Bayesian Statistics 9*, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West (Eds.). Oxford: Oxford University Press.

Press, W., Teukolosky, S., Vetterling, W. & Flannery, B. (1992). *Numerical Recipes: the Art of Scientific Computing*, 2nd Edition. New York: Cambridge University Press.

R Development Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, `http://www.R-project.org`

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.

Wainwright, M.J. & Jordan, M.I. (2008). Graphical models, exponential families, and variational inference. *Foundation and Trends in Machine Learning*, **1**, 1–305.

Wand, M.P. & Ormerod, J.T. (2011). Penalized wavelets: embedding wavelets into semiparametric regression. *Electronic Journal of Statistics*, **5**, 1654–1717.

Wand, M.P. and Ormerod, J.T. (2012). Continued fraction enhancement of Bayesian computing. *Stat*, **1**, 31–41.

Wand, M.P., Ormerod, J.T., Padoan, S.A. and Frühwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, **6**, Number 4, 847–900.

Wand, M.P. & Ripley, B.D. (2010). 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`

Whittaker, E. T. & Watson, G. N. (1990). *A Course in Modern Analysis*, 4th ed. Cambridge UK: Cambridge University Press.

Wuertz, D. & others. (2009). fAsianOptions 2100.76. Exponential Brownian motion and Asian option evaluation. R package. `http://cran.r-project.org`