Abstract:

A new variational approximation based on multivariate skew-normal densities is developed. This paper overcomes most of the computational difficulties associated with the proposed approximation and shows for several examples how the method can significantly improve upon accuracy of the Laplace approximation, integrated nested Laplace approximation and Gaussian variational approximations whilst being significantly faster than Markov chain Monte Carlo methods.

Key Words: Variational approximation; Skew-normal distribution; Laplace approximation; Integrated nested Laplace approximation; Machine learning; Markov chain Monte Carlo.

1 Introduction

In recent years there has been an increasing amount of research in approximate methods for Bayesian inference. The driving force behind this research is the fact that while Monte Carlo methods can be made to be arbitrarily accurate they can be unacceptably slow when applied to large or complex problems. Approximate methods for Bayesian inference include Laplace-like methods (Tierney & Kadane, 1986; Raudenbush, Yang & Yosef, 2000; Rue, Martino & Chopin, 2009), several variational approximation approaches (Bishop, 2006; Ormerod & Wand, 2010; Wand et al., 2011) and expectation propagation (Minka, 2001). These methods can be quite effective when the posterior density is log-concave, or at least unimodal.

The Laplace-like approximations are perhaps the most pervasive analytic approximations used in all of Statistics. Several variants exist in various contexts, including PQL (Breslow & Clayton, 1993), approximating ratios of integrals (Tierney & Kadane, 1986), higher order Laplace approximations (Raudenbush et al., 2000) and, more recently, the integrated nested Laplace approximation (INLA) of Rue et al. (2009). The Laplace approximation is elegant, easy to implement and highly efficient to compute. Unfortunately, if the posterior distribution is not similar in shape to a multivariate Gaussian the Laplace approximation may perform poorly in practise.

The INLA method of Rue et al. (2009) deserves special attention. The INLA method is a specialized method for approximating Gaussian latent effect models. The key idea of this method is to use Laplace-like methods to integrate out Gaussian latent effects while using other numerical methods to integrate out other parameters. A potential drawback of the method is that it focuses on the calculation of marginal posterior densities, rather than the full posterior. If the underlying posterior distribution of the latent effects is approximately Gaussian then
INLA can be quite effective. However, when the underlying posterior distribution of the latent effects is skewed INLA can perform poorly (for example see Fong, Rue & Wakefield, 2010).

Expectation propagation (EP) is another alternative to Monte Carlo methods and generally outperforms Laplace’s method in terms of accuracy (Kuss & Rasmussen, 2005). Some recent developments by Opper & Winther (2005), Paquet, Winter & Opper (2009) and Cseke & Heskes (2010) provide further improvements in terms of accuracy. However, EP can be difficult to apply to some models and can suffer from convergence issues. For these reasons we will not pursue EP in this paper.

Variational approximations take a variety of forms (for an introduction see Bishop, 2006, or Ormerod & Wand, 2010). This type of approximation, related to mean-field approximations, includes variational Bayes, although such approximations can also be applied in frequentist settings (see Hall, Ormerod & Wand, 2011a; Hall et al. 2011b; Ormerod & Wand, 2011). This type of approximation has been particularly successful when the model at hand is either sufficiently large and/or complex. Successful examples of these methods include document retrieval (e.g. Jordan, 2004) and functional magnetic resonance imaging (e.g. Flandin & Penny, 2007), cluster analysis for gene-expression data (Teschendorff et al., 2005) and finite mixture models (McGrory & Titterington, 2007).

In this paper we consider a particular type of parametric variational approximation, an alternative to variations Bayes. A particular parametric variational approximation that has recently received some attention is the Gaussian variational approximation (GVA) where the posterior, like Laplace’s method, is approximated by a multivariate Gaussian density (Opp & Archambeau, 2009; Ormerod & Wand, 2011). This type of approximation, as one might expect, has similar properties to the Laplace approximation but empirically tends to outperform it in terms of accuracy. However, like the Laplace approximation, the Gaussian variational approximation can perform poorly when the approximate posterior density is not similar in shape to the multivariate Gaussian density.

The lack of flexibility of the Laplace and GVA methods leads to their failure when, for instance, the underlying posterior density is skewed. In this paper we develop skew-normal variational approximations (SNVA) which seek to improve upon the Laplace and GVA methods by employing a multivariate skew-normal density in order to make up for this potential shortcoming. Although, many skew-normal distributions now exist (see Genton, 2004), we will use the original multivariate skew-normal distributions proposed by Azzalini & Dalla Valle (1996). This extension makes particular sense when attempting to approximate log-concave, or at least unimodal, densities.

In utilizing the increased flexibility of the multivariate skew-normal density, over multivariate Gaussian densities, the SNVA method can give substantial improvements over the Laplace’s or GVA methods. We illustrate and compare the SNVA method for examples including the normal sample, generalized linear models, generalized linear mixed models and an inhomogeneous noise model. In each of the examples the SNVA method gives better or at least comparable accuracy to INLA for almost all parameters where a direct comparison is possible. These examples also demonstrate two other advantages of SNVA over INLA, namely the fact that SNVA approximates the joint posterior distribution while INLA only approximates marginal posterior densities, and the fact that SNVA may be applied to models which INLA does not currently handle. The SNVA method is also reasonably fast.
The organization of the paper is as follows. Section 2 describes the basis for SNVA. Section 3 describes how to calculate the SNVA lower bound on the log-likelihood. Section 4 describes how we maximize this lower bound on the log-likelihood to improve the approximation. Section 5 demonstrates the advantages of the SNVA method for several Bayesian models and shows how for these examples the method requires at most univariate quadrature. Section 6 provides some conclusions and discussion.

2 Skew-Normal Variational Approximations

Bayesian inference is mainly concerned with the calculation of the posterior density

$$p(\theta | y) = \frac{p(y, \theta)}{p(y)}$$

(1)

where \( p(y) = \int_\Theta p(y, \theta) d\theta \), the vector \( y \) denotes the observed response values, and \( \theta \) denotes all unobserved or hidden variables, for example model parameters, latent and auxiliary variables or missing data, and \( \Theta \) is the \( m \)-dimensional parameter space of \( \theta \). In this paper we will assume, for simplicity, that \( \theta \) is continuous and that, unless otherwise specified, the domain of integration used in integrals is \( \Theta \). Unfortunately, the direct calculation of the quantity \( p(\theta | y) \) is only possible for special cases and approximation is required.

Most variational approximations are based on minimizing the Kullback-Leibler (KL) distance between a density \( q(\theta) \) and \( p(\theta | y) \) given by

$$\text{KL}(q(\theta), p(\theta | y)) = \int q(\theta) \log \left\{ \frac{q(\theta)}{p(\theta | y)} \right\} d\theta$$

(2)

noting that the Kullback-Leibler distance is positive and zero if and only if \( q(\theta) = p(\theta | y) \) almost everywhere (Kullback & Leibler, 1951).

Parametric variational approximations attempt to minimize (2) subject to the restriction

$$q(\theta) = q(\theta; \xi)$$

where \( q(\theta; \xi) \) is some conveniently chosen parametric density and \( \xi \) are parameters of that density (called variational parameters). Using simple algebraic manipulations it is easy to show

$$\log p(y) = \int p(y, \theta) d\theta$$

(3)

$$= \int q(\theta; \xi) \log \left\{ \frac{p(y, \theta)}{q(\theta; \xi)} \right\} d\theta + \int q(\theta; \xi) \log \left\{ \frac{q(\theta; \xi)}{p(\theta | y)} \right\} d\theta$$

(4)

$$= \int q(\theta; \xi) \log \left\{ \frac{p(y, \theta)}{q(\theta; \xi)} \right\} d\theta + \text{KL}(q(\theta), p(\theta | y))$$

(5)

$$\geq \int q(\theta; \xi) \log p(y, \theta) d\theta - \int q(\theta; \xi) \log q(\theta; \xi) d\theta$$

(6)

$$= \int q(\theta; \xi) \log p(y, \theta) d\theta + \mathcal{E}_q(\xi)$$

(7)

$$\equiv \log p_q(y; \xi)$$

(8)
which follows from the fact that the KL-divergence, the second term on the right hand side of (4), is positive. We can also see from (4)–(8) that \( \log p(y) - \log p_q(y; \xi) = \text{KL}(q(\theta; \xi), p(\theta|y)) \) so that maximizing \( p_q(y; \xi) \) with respect to \( \xi \) is equivalent to minimizing \( \text{KL}(q(\theta; \xi), p(\theta|y)) \) with respect to \( \xi \). Furthermore, in order to calculate the lower bound for a particular \( q(\theta; \xi) \) we need to be able to calculate \( E_q(\xi) = - \int q(\theta; \xi) \log q(\theta; \xi) d\theta \), the (Shannon's) entropy of \( q(\theta; \xi) \).

As was motivated in the introduction we choose \( q \) to be the multivariate skew-normal density originally proposed by Azzalini & Dalla Valle (1996). We will say a \( m \)-dimensional random vector \( \theta \) has a multivariate skew-normal distribution with skewness vector \( \delta \) and covariance matrix \( \Lambda \), denoted \( \theta \sim SN(\mu, \Lambda, \delta) \), if its probability density function is given by

\[
q(\theta; \mu, \Lambda, \delta) = 2\phi(\theta - \mu)\Phi(d^T(\theta - \mu))
\]

(9)

where \( \theta \in \Theta = \mathbb{R}^m \), \( \phi(\theta - \mu) \) denotes the multivariate Gaussian density with mean vector \( \mu \) and covariance matrix \( \Lambda \), \( \Phi \) denotes the cumulative distribution function for the standard univariate Gaussian distribution with mean \( \mu \), \( \Lambda \) and \( d \) being the location, scale and skewness parameters respectively. Note that \( \mu \) and \( d \) are unrestricted in \( \mathbb{R}^m \) while \( \Lambda \) must be positive definite. It is easy to see that when \( d = 0 \) the multivariate Gaussian density is recovered.

We call the parametric variational approximations where \( q \) is given by (9) a skew-normal variational approximation (SNVA). For this \( q \) the variational parameters are \( \xi = (\mu, \Lambda, d) \). The Gaussian variational approximation (GVA) is the special case when \( d = 0 \) denoted \( p_G(y; \mu, \Lambda) \). Note that the SNVA is guaranteed to be a better approximation of the marginal likelihood than GVA in the sense that \( p_{SN} \) is a tighter lower bound on \( p(y) \) than \( p_G \) since

\[
p(y) \geq \sup_{\mu, \Lambda, d} \{ p_{SN}(y; \mu, \Lambda, d) \} \geq \sup_{\mu, \Lambda, d} \{ p_{SN}(y; \mu, \Lambda, d) \text{ such that } d = 0 \} = \sup_{\mu, \Lambda} \{ p_G(y; \mu, \Lambda) \}.
\]

We will now focus on the calculation of \( p_{SN} \).

3 Calculating the Variational Lower Bound

The calculation of \( p_{SN} \) requires applying various properties of the skew normal density (9). We will briefly review only the most pertinent of these to show some new results required by the skew-normal variational approximation. These results will also be useful for calculating the lower bound \( p_{SN} \) and its derivatives for the examples considered later on.

The moment generating function of \( q \) is given by \( M(t) = \mathbb{E}(\exp(t^T\theta)) = 2\exp(\mu^Tt + \frac{1}{2}t^T\Lambda t)\Phi(\delta^Tt) \) where \( \delta = \Lambda d / (1 + d^T\Lambda d)^{1/2} \) which we can use to show that \( \mathbb{E}(\theta) = \mu + \sqrt{2/\pi}\delta \) and \( \text{Cov}(\theta) = \Lambda - (2/\pi)\delta\delta^T \). Furthermore, we can use the moment generating function (González-Farías, Domínguez-Molina & Gupta, 2004) or alternatively (Azzalini & Capitanio, 1999) to show that the skew-normal distribution is closed under linear transformations, i.e. for any \( m \times r \) matrix \( \Lambda \) (of rank \( r \)) and \( r \times 1 \) vector \( b \) we have

\[
z = \Lambda \theta + b \sim SN(\mu_z, \Lambda_z, d_z)
\]

(10)

where \( \mu_z = \Lambda \mu + b \), \( \Lambda_z = \Lambda \Lambda \Lambda^T \) and \( d_z = \Lambda_z^{-1} \Lambda \Lambda d (1 + d^T\Lambda(I - \Lambda^T\Lambda_z^{-1}\Lambda \Lambda)d)^{-1/2} \). Using this expression it is easy to show that the "\( \delta \)" corresponding to this transformation is given by
\[ \delta_x = A \delta. \] From the closure of skew-normal densities under linear transformations it is easy to show that the skew-normal distribution is also closed under marginalization, i.e. by setting \( A \) to \( e_i^T \) in (10) where \( e_i \) is the zero vector except the value 1 in the \( i \)th element. Finally, using the closure under linear transformations property, we have

\[
\int_{\mathbb{R}^m} q(\theta; \mu, \Lambda, d) f(A \theta + b) d\theta = \mathbb{E} [f(A \theta + b)] = \mathbb{E} [f(z)] = \int_{\mathbb{R}^r} q(z; \mu_z, \Lambda_z, d_z) f(z) dz \quad (11)
\]

which, provided \( r < m \), reduces the dimensionality of the integral to be calculated. This fact is used to derive the following new result.

**Result 1 – Entropy Expression:** The entropy function of \( 2\phi_{\Lambda}(\theta - \mu)\Phi(d^T(\theta - \mu)) \) is given by

\[
\mathcal{E} (\mu, \Lambda, d) = - \int q(\theta; \mu, \Lambda, d) \log q(\theta; \mu, \Lambda, d) d\theta = \frac{1}{2} \log |2\pi \Lambda| - \log(2) - \Psi(d^T \Lambda d) \quad (12)
\]

where \( \Psi(\sigma^2) = \int_{\mathbb{R}} 2\phi_{\sigma}(z) \Phi(z) \log \Phi(z) dz. \)

**Proof:** See Appendix A.1.

The properties and the calculation of \( \Psi \) and its derivatives are discussed in Appendix A.1. To the best of the author’s knowledge the entropy expression for the skew-normal distribution has not been previously derived.

Using the expression for the entropy of the skew-normal density (12) with (5) the skew-normal variational approximation the log of the marginal likelihood may be written as

\[
\log p_{\text{sn}}(y; \mu, \Lambda, d) = \frac{1}{2} \log |2\pi \Lambda| + \frac{m}{2} - \log(2) - \Psi(d^T \Lambda d) + f_{\text{sn}}(\mu, \Lambda, d)
\]

where \( f_{\text{sn}}(\mu, \Lambda, d) = \int 2\phi_{\Lambda}(\theta - \mu) \Phi(d^T(\theta - \mu)) \log p(y, \theta) d\theta. \) (13)

Many of the above expectations can be calculated using the moment generating function and linear transformations results above. Calculation of the expectations in two of the examples in Section 5 will be aided by the following result:

**Result 2:** Suppose that \( q(\theta; \mu, \Lambda, d) = 2\phi_{\Lambda}(\theta - \mu)\Phi(d^T(\theta - \mu)) \) then

\[
\exp(t^T \theta)q(\theta; \mu, \Lambda, d) = 2\exp(t^T \mu + t^T \Lambda t/2)\Phi(\delta^T t) \tilde{q}(\theta; \mu, \Lambda, d, t) \quad (14)
\]

where \( \tilde{q}(\theta; \mu, \Lambda, d, t) = \frac{\Phi(d^T(\theta - \mu))}{\Phi(\delta^T t)} \phi_{\Lambda}(\theta - \mu - \Lambda t). \) (15)

Note that \( \tilde{q}(\theta; \mu, \Lambda, d, t) \) is a special case of the multivariate skew-normal distribution first proposed by Arnold & Beaver (2002) and when \( t = 0 \) the skew-normal distribution of Azzalini & Dalla Valle (1996) is obtained. Let \( \Theta \) be a random variable whose density is (15). Then is can be shown that the moment generating function of \( \tilde{\Theta} \) is given by \( M(s) = 2\exp(s^T(\mu + \Lambda t) + s^T \Lambda s/2)\Phi(\delta^T (s + t))/\Phi(\delta^T t) \) and hence \( \mathbb{E}(\tilde{\Theta}) = \mu + \Lambda t + \zeta_1(\delta^T t) \delta \) and \( \text{Cov}(\tilde{\Theta}) = \Lambda + \zeta_2(\delta^T t) \delta \delta^T \) where \( \zeta_k(x) = d^k \log \Phi(x)/dx^k \). Note that special care needs to be taken to calculate \( \zeta_k(x) \) when \( x \) is large and negative for \( k \geq 1 \). A robust implementation of \( \zeta_k(x) \) is given in the \( \mathbb{R} \) package \( \text{sn} \) (Azzalini, 2010).
4 Maximising the Variational Lower Bound

We now address the problem of maximizing \( p_{SN}(y; \mu, \Lambda, d) \) with respect to the variational parameters \((\mu, \Lambda, d)\). Many efficient numerical optimization methods only require gradient information to perform this optimization (Nocedal & Wright, 1999; Luenberger & Ye, 2008). Quasi-Newton methods approximate the Hessian by using differences of successive iterations of the gradient vector. A particularly popular quasi-Newton method is the BFGS method, named after its inventors Broyden, Fletcher, Goldfarb and Shanno. This method is implemented in the \( \mathbb{R} \) function \code{optim}(). In order to use this method we will need the expressions for the derivatives of \( p_{SN}(y; \mu, \Lambda, d) \) with respect to the variational parameters \((\mu, \Lambda, d)\).

4.1 Derivative Expressions

There are a number of ways to parametrize \( \Lambda \). Suppose that, for the time being, that \( \operatorname{vech}(\Lambda) \) are the unique parameters of \( \Lambda \). Suppose that the \( 1 \times d \) derivative vector of a function \( f(x) \), denoted \( \nabla_x f(x) \), is the vector with \( i \)th entry \( \partial f(x) / \partial x_i \) and the corresponding Hessian matrix is denoted by \( \nabla^2_x f(x) = \nabla_x \{ \nabla_x f(x) \}^T \). Also, to simplify upcoming expressions let \( f(\theta) = \log p(y, \theta) \), \( g(\theta) = \nabla_{\theta} \log p(y, \theta) \) and \( H(\theta) = \nabla^2_{\theta} \log p(y, \theta) \). Then the expressions for the derivatives of \( p_{SN} \) with respect to \((\mu, \operatorname{vech}(\Lambda), d)\) are given in the following result.

Result 3 – Derivative Expressions: Assuming that \( \log p(y, \theta) \) is twice differentiable with respect to \( \theta \) the derivatives of \( \log p_{SN} \equiv \log p_{SN}(y; \mu, \Lambda, d) \) are given by

\[
\begin{align*}
D_{\mu} \log p_{SN} &= g_{SN} \\
D_{d} \log p_{SN} &= c_1 (\Lambda - \delta \delta^T) g_c - \Psi'(d^T \Lambda d) \Lambda d \\
\frac{d \log p_{SN}}{d \Lambda_{ij}} &= \frac{1}{2} \left[ \Lambda^{-1} + H_{SN} + c_1 (g_c d^T + d g_c^T) - c_2 dd^T \right]_{ij}
\end{align*}
\]

(16)

where \( c_1 = \sqrt{2/\pi}/\sqrt{1 + d^T \Lambda d}, c_2 = \Psi'(d^T \Lambda d) + c_1 d^T \Lambda g_c/(1 + d^T \Lambda d), \)

\[
\begin{align*}
g_{SN} &\equiv g_{SN}(\mu, \Lambda, d) = \int_{\mathbb{R}^m} 2 \phi_\Lambda(\theta - \mu) \Phi(d^T(\theta - \mu)) g(\theta) d\theta \\
H_{SN} &\equiv H_{SN}(\mu, \Lambda, d) = \int_{\mathbb{R}^m} 2 \phi_\Lambda(\theta - \mu) \Phi(d^T(\theta - \mu)) H(\theta) d\theta
\end{align*}
\]

and \( g_c \equiv g_{SN}(\mu, \Lambda - \delta \delta^T, 0) \).

Proof: See Appendix A.2.

The evaluation of \( D_{\mu} \log p_{SN}, D_{d} \log p_{SN} \) and \( D_{\operatorname{vech}(\Lambda)} \log p_{SN} \) would require different results, for example, if \( p(y, \theta) \) included Laplace distributed or other non-differentiable components.

A significant concern when we attempt to maximize \( p_{SN} \) is ensuring \( \Lambda \) remains positive definite. This constraint can be ensured by using the parameterization \( \Lambda(r) = R^T R \) where \( R \) is an upper triangular matrix with \( e^{r_{ii}} \) on the diagonal and \( r_{ij} \) in the upper right elements of the matrix. In this case the derivatives with respect to the \( r_{ij} \)s are given by

\[
\frac{\partial \log p_{SN}}{\partial r_{ij}} = \left[ R^{-T} + R \left( H_{SN} + c_1 (g_c d^T + d g_c^T) - c_2 dd^T \right) \right]_{ij} e^{r_{ij}I(i=j)}, \quad 1 \leq i \leq m, 1 \leq j \leq m
\]
where \( I(\cdot) \) is the indicator function taking the value 1 if the condition is true and 0 otherwise and the terms appearing are defined in the derivative expression results above.

### 4.2 Initial Values

Note that Result 3 coincides with the results obtained by Opper & Archambeau (2009) for GVA, i.e. when \( d = 0 \). We can also see, based on this observation, that if \((\mu_0^*, \Lambda_0^*)\) maximizes \( p_c \) then \((\mu_c^*, \Lambda_c^*, 0)\) is at least a local maximiser of \( p_{SN} \). Later we will see examples where \((\mu_c^*, \Lambda_c^*, 0)\) is neither the only maximiser nor the global maximiser of \( p_{SN} \).

We will use the Laplace approximation to provide ballpark starting values for SNVA. Assuming that the joint distribution \( p(y, \theta) \) is twice differentiable and \( \Theta = \mathbb{R}^m \) then Laplace’s method approximates the posterior density \( p(\theta|y) \) by \( q_l(\theta) = N(\mu, \Lambda) \) where \( \mu \) is the maximiser of \( p(y, \theta) \) and \( \Lambda = [-H(\mu)]^{-1} \). These values can be found quickly an easily by applying Newton’s method whose updates are given by \( \Lambda^{(t+1)} ← [-H(\mu^{(t)})] \) and \( \mu^{(t+1)} ← \mu^{(t)} + \Lambda^{(t)}g(\mu^{(t)}) \). Upon convergence \( \mu \) and \( \Lambda \) take the values \( \mu^{(t)} \) and \( \Lambda^{(t)} \) respectively. Given these values the Laplace approximation to the marginal likelihood is \( p_l(y) = |2\pi\Lambda|^{1/2}p(y, \mu) \).

### 5 Illustrations

In this section we demonstrate the value of the SNVA method. We will compare various approximations with an accurate MCMC approximation which we will use as a “gold standard”. To this end, for each of models fitted in this paper, we will generate a long run of MCMC samples using the R package BRugs (Ligges et al., 2009). Unless otherwise specified in the following examples we will generate \( 5 \times 10^4 \) burn-in samples followed by a further \( 5 \times 10^6 \) samples using a thinning factor of 5 so that a total of \( 10^6 \) samples are available for inference. Furthermore, we will use the over-relaxed form of MCMC as described by Neal (1998) to improve mixing. In BRugs this involves setting the option overRelax=TRUE when calling the modelUpdate() function. Kernel density estimates \( q_{MCMC}(\theta_i) \) of the \( p(\theta_i|y) \)s were then constructed using the \( 10^6 \) MCMC samples. These We should expect these Kernel density estimates to be nearly perfect posterior density approximations.

Let the vector \( \theta_i \) denote a generic model parameter of length at most two. There are various ways to measure the accuracy between an approximate posterior density \( q(\theta_i) \) and a “gold standard” \( q_{GS}(\theta_i) \). We will compare various posterior density estimates and use the \( L_1 \)-norm or Integrated Absolute Error (IAE) defined by

\[
\text{IAE}(q) = \int \left| q(\theta_i) - q_{GS}(\theta_i) \right| d\theta_i
\]

where for the purposes of this paper we will use the MCMC approximation described above as the “gold standard”, i.e. \( q_{GS}(\theta_i) = q_{MCMC}(\theta_i) \). This error measure has the advantages of being invariant to monotone transformations of the parameter \( \theta \) and is a scale independent number between 0 and 2 (e.g. Devroye & Györfi, 1985). This motivates the following measure of accuracy

\[
\text{Accuracy}(q(\theta_i)) = 1 - \frac{1}{2} \text{IAE}(q(\theta_i))
\]
noting that \(0 \leq \text{Accuracy}(q(\theta_i)) \leq 1\) and will be expressed as a percentage. We will restrict \(\theta_i\) to be at most two-dimensional so that \(\text{IAE}(\theta_i)\) can be approximated numerically using one or two-dimensional quadrature.

We could also use some alternative measures of accuracy. Other measures such as the KL-divergence between \(q(\theta|y)\) and \(p(\theta|y)\) is computationally both more demanding and delicate due to the fact that this requires the calculation of the marginal likelihood. We might also be tempted to use the KL-divergence between marginal posterior densities. However, as noted by Hall (1987), this measure is dominated by the tail behavior of \(q(\theta)\) and \(q_{\text{MCMC}}(\theta)\).

Using the above measure of accuracy we will compare the Laplace, INLA, GVA and SNVA methods for the normal sample model (Example 1), generalized linear models (Example 2), generalized linear mixed models (Example 3) and linear regression with inhomogeneous noise (Example 4). In Ormerod & Wand (2010) the variational Bayes (VB) approach was described for Example 1 and so for this example we will also compare VB. Note that for the INLA method we use the \(R\) package \(\text{INLA}\) (available from \url{http://www.r-inla.org/}) and use all default options except that the option \(\text{control.inla=list(strategy="FIT_SCGAUSSIAN")}\) was used as the option which is expected to give the best accuracy.

### 5.1 Example 1 – Normal Sample

A good place to start is to consider one of the simplest Bayesian models for which integration is required, the estimation of a normal sample. It is by no means a difficult problem, since MCMC takes negligible time to fit, however this simple example serves to illustrate the differences in various approximation methods.

Consider the model

\[
y_i|\mu, \sigma^2 \sim N(\mu, \sigma^2), \quad 1 \leq i \leq n,
\]

with priors \(\mu \sim N(0, \sigma^2)\) and \(\sigma^2 \sim \text{IG}(a, b)\).

The marginal likelihood may be written as

\[
p(y) = \int_{0}^{\infty} \int_{-\infty}^{\infty} p(y, \mu, \sigma^2) d\mu d\sigma^2
\]

where \(p(y, \mu, \sigma^2) = \exp\{(-(A+n/2+1) \log(\sigma^2) - \sigma^{-2}(B+\|y-\mu 1\|_2^2/2) - \mu^2/(2\sigma^2\mu) + \kappa}\}

where the sum of the constant terms, \(\kappa\), is given by \(\kappa = -\log(2\pi) - \log(\sigma^2)/2 - A\log(B) - \log \Gamma(A)\). In order to transform the domain of the integral to \(\mathbb{R}^2\), so that we can use SNVA, we use the transformation \(\sigma^2 = e^\gamma\). Based on Theorem 2.1.5 of Casella & Berger (2002) the prior for \(\gamma\) becomes \(p(\gamma) = A^B \exp\{(-A\gamma - B e^{-\gamma})/\Gamma(\gamma)\}\). Let \(\theta = (\mu, \gamma)\) so that the joint likelihood may be written as

\[
p(y, \theta) = \exp\{-(A+n/2) \gamma - e^{-\gamma}(B+\|y-\mu 1\|_2^2/2) - \mu^2/(2\sigma^2\mu) + \kappa\}.
\]

Next, consider the partitions

\[
\mu = \begin{bmatrix} \mu_\mu \\ \mu_\gamma \end{bmatrix} , \quad \delta = \begin{bmatrix} \delta_\mu \\ \delta_\gamma \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix} \Lambda_{\mu\mu} & \Lambda_{\mu\gamma} \\ \Lambda_{\gamma\mu} & \Lambda_{\gamma\gamma} \end{bmatrix}
\]

of \(\mu\), \(\delta\) and \(\Lambda\) respectively, corresponding to the elements of \(\theta\). Note that we will be using similar partitions of \(\mu\), \(\delta\) and \(\Lambda\) in later problems.
Then, using the results in Section 3, we can show that expressions for $f_{SN}$, $g_{SN}$ and $H_{SN}$ are

$$f_{SN} = - (A + \frac{n}{2}) \left( \mu_\gamma + \sqrt{\frac{2}{\pi}} \delta_\gamma \right) - M \left( B + \frac{1}{2} T \right) - \frac{1}{2\sigma_\mu} \left( \mu_\mu^2 + 2\sqrt{\frac{2}{\pi}} \mu_\mu \delta_\mu + \Lambda_{\mu\mu} \right)$$

$$g_{SN} = \left[ M1_n^T (y - \tilde{\mu}1_n) - (\mu_\mu + \sqrt{\frac{2}{\pi}} \delta_\mu)/\sigma_\mu \right] M \left( B + \frac{1}{2} T \right) - \left( A + \frac{n}{2} \right)$$

and $H_{SN} = - \left[ \frac{Mn + \sigma_\mu^2}{M1_n^T (y - \tilde{\mu}1_n)} \right] M \left( B + \frac{1}{2} T \right)$

respectively where $M = 2e^{-\mu_\gamma + \Lambda_{\gamma\gamma}/2}\Phi(-\delta_\gamma)$, $T = \|y - \tilde{\mu}1_n\|^2 + n(\Lambda_{\mu\mu} + \zeta_2(\delta_\gamma)\delta_\mu^2)$ and $\bar{\mu} = \mu_\mu - \Lambda_{\gamma\mu} + \zeta_1(\delta_\gamma)\delta_\mu$.

To test each approximation we simulated data from (17) with sample size $n = 6$ and true parameters $\nu = 6$ and $\sigma^2 = 225$. We used the hyperparameters $\sigma_\mu^2 = 10^4$ and $A = B = 0.01$. For the purposes of comparison we obtained highly accurate approximations to $p(\mu|y)$ and $p(\sigma^2|y)$ via one-dimensional numerical quadrature instead of using a large number of MCMC samples. Figure 1 illustrates the quality of the Laplace, VB, GVA and SNVA approximations.

![Figure 1: Posterior approximations based on the Laplace, Variational Bayes (VB), GVA and SNVA approximations. The ‘true’ posterior density, based on accurate numerical integration, is in black while the approximations are in grey.](image)

In Figure 1 we see that the Laplace approximation is a Gaussian approximation around the mode of the posterior. The variational Bayes approximation does quite a good job for this example, which is to be expected since, asymptotically the posterior distributions of $\mu$ and $\gamma$ are independent. The GVA approximation gives a slight improvement over the Laplace approximation and SNVA is the most accurate approximation for this case.

Figure 1 also illustrates that given different initial values the SNVA approximation can give different answers. If supplied with the GVA approximation as initial values then SNVA will give identical results to GVA. However, depending on the initial values the SNVA approximation may compensate for one or the other of the tails of the posterior density of $\mu$.

### 5.2 Example 2 - Generalized Linear Models

Generalized linear models are one of the most commonly used models in Statistics. Fitting these models by maximum likelihood is trivial. However, in a Bayesian setting, it can be extremely difficult to find adequate analytic approximations to the posterior distribution for these
models. Furthermore, Bayesian generalized linear models closely share many of the computational difficulties with generalized linear mixed models, an active area of research, and are thus worthy of consideration.

Consider the model \( y|\beta \sim \text{Poisson}(\expit(X\beta)) \) or \( y|\beta \sim \text{Binomial}(\expit(X\beta)) \) where \( \expit(x) = \frac{e^x}{1 + e^x} \). We employ the prior \( \beta \sim N(0, \Sigma) \) so that the joint likelihood may be written as 
\[
p(y|\beta) \propto \exp \left\{ \mathbf{y}^T \mathbf{X} \beta - \frac{1}{2} \mathbf{X} \beta \Sigma^{-1} \beta \right\}
\]
where \( b(x) = \expit(x) \) for the Poisson model and \( b(x) = \log(1 + e^x) \) for the logistic model. The functions \( f_{SN}, g_{SN} \) and \( H_{SN} \) are given by

\[
 f_{SN} = \mathbf{y}^T \mathbf{X} (\mu + \sqrt{\frac{2}{\pi}} \delta) - 1^T B^{(0)}(\mathbf{X} \mu, \text{dg}(\mathbf{X} \Lambda \mathbf{X}^T), \mathbf{X} \delta) - \frac{1}{2} \mu^T \Sigma^{-1} \mu - \sqrt{\frac{2}{\pi}} \mu^T \Sigma^{-1} \delta - \frac{1}{2} \text{tr} [\Sigma^{-1} \Lambda]
\]
\[
 g_{SN} = \mathbf{X}^T (\mathbf{y} - B^{(1)}(\mathbf{X} \mu, \text{dg}(\mathbf{X} \Lambda \mathbf{X}^T), \mathbf{X} \delta)) - \Sigma^{-1} (\mu + \sqrt{\frac{2}{\pi}} \delta)
\]
\[
 H_{SN} = -\mathbf{X}^T \text{diag}(B^{(2)}(\mathbf{X} \mu, \text{dg}(\mathbf{X} \Lambda \mathbf{X}^T), \mathbf{X} \delta)) \mathbf{X} - \Sigma^{-1}
\]

where \( B^{(r)}(\mu, \sigma^2, \delta) = \int_{-\infty}^{\infty} 2 \phi_\sigma(x - \mu) \Phi(d(x - \mu)) b(x) dx \) with \( d = (\delta/\sigma)/\sqrt{1 - \delta^2/\sigma^2} \) and \( \text{dg}(\mathbf{A}) \) is the vector obtained from taking the diagonal elements of a square matrix \( \mathbf{A} \). For the Poisson model the function \( B^{(r)}(\mu, \sigma^2, \delta) \) is available analytically and is given by \( B^{(r)}(\mu, \sigma^2, \delta) = 2 \exp \left\{ \mu + \frac{1}{2} \sigma^2 \right\} \Phi(\delta) \) whereas for the logistic model univariate numerical quadrature is required. The adaptive Gauss-Hermite quadrature method described by Liu & Pierce (1994) or Pinheiro & Bates (1995) is a particularly suitable for this type of integral.

### 5.2.1 O-ring Data

Analysis of the failure temperature of o-rings has been argued to be the probable cause of the 1986 Challenger disaster and mainstay of virtually every Statistics course involving logistic regression. The dataset is available from many sources including Congdon (2006). We fit the logistic regression model described in Section 5.2 where \( y_i \) is the indicator of o-ring failure and \( x_i \) is the temperature in Fahrenheit for \( 1 \leq i \leq 23 \). We standardized the \( x_i \)s.

The approximate joint posterior densities based on the Laplace, GVA and SNVA approximations are illustrated in Figure 2. Based on Figure 2 the SNVA method is clearly the most accurate method at estimating the joint posterior density.

Note that we have not made comparisons here with the INLA method which only approximates marginal posterior densities, unlike the other methods which approximates the joint posterior density. The joint posterior density may be of interest in itself, for example when approximating credible intervals for the mean. Using (10) we can find the distribution of \( \mathbf{X} \beta \) to obtain credible intervals for the mean. Figure (2) illustrates the posterior mean and 95% credible intervals of the mean based on the Laplace, GVA and SNVA approximations from which it is clear that SNVA provides the best approximations.

### 5.2.2 Breastfeed Data

Consider the data from a study conducted in a UK hospital on the decision of pregnant women to breastfeed their babies or not, see Moustaki et al. (1998). For the study, 135 expectant mothers were asked how they intended to feed their baby. The responses were coded as \( y = 1 \) for including breastfeeding and mixed breast/bottle feeding, and \( y = 0 \) for exclusive bottle-feeding.
Figure 2: [Top panels] Illustrations of the Laplace, GVA and SNVA approximations (gray) of the joint posterior density for logistic regression model applied to the o-ring dataset along with the “exact” posterior based on bivariate numerical integration (black). [Bottom panels] Illustrations of the Laplace, GVA and SNVA approximations (gray) of the posterior mean and 95% credible intervals of the mean for logistic regression model applied to the o-ring dataset along with these “exact” quantities based on MCMC simulations (black).

Available covariates included the advancement of the pregnancy (end pregnancy = 1 or beginning pregnancy = 0), how the mothers were fed as babies (howfed), how the mother’s friend fed their babies (howfedfr), whether they had a partner (partner), the mother’s age (age), the age at which they left full-time education (educat), their ethnic group (ethnic) and if they have ever smoked (smokebf) or if they had stopped smoking (smokenow). The data has also been analysed by and is available from the website of the book by Heritier et al. (2009).

The Bayesian logistic regression model described in Section 5.2 using all of the available covariates where the variables age and educat were standardized and used the prior covariance \( \Sigma = 10^4 I \). Results from each of the methods considered are summarized in Figure 3.

From Figure 3 we see that the marginal posterior densities approximated by SNVA are nearly indistinguishable to the gold standard MCMC approximation with \( 10^6 \) posterior samples for most coefficients and quite better than its nearest analytic approximation method INLA. Furthermore, it is more accurate than an MCMC approximation based on only \( 10^3 \) samples (taking about a minute on the author’s laptop) and is of comparable accuracy with the MCMC approximation based on \( 10^4 \) samples (taking about a 10 minutes on the author’s laptop). In comparison SNVA took around 5 seconds to fit this model. Thus the SNVA method has a considerable advantage over MCMC and INLA for this example.
5.2.3 Simulated Data

We first compare the Laplace approximation, GVA, INLA and SNVA approximations for simulated data. For large $n$ the posterior distribution becomes increasingly Gaussian. Thus, for sufficiently large $n$ each of these methods will perform similarly well.

We have chosen the following simulation settings so that $n$ is small relative to $m$ so as to induce the more difficult case where the posterior distribution can become quite skewed: $m \in (4, 8, 16, 20, 30, 40), \beta = (2, -2, \ldots, 2, -2)/m, x_{ij} \sim N(0, 1), 1 \leq i \leq m, 1 \leq i \leq n$ with the prior covariance $\Sigma = 10^4I$. We set $n = 3m/2$ for the Poisson case and $n = 2m$ for the logistic case. Also, in order to avoid pathological cases, we discarded simulations where the R function glm failed to converge or the number of zeros was greater than $m - 1$ for the Poisson case or where either the number of successes or failures was less than 2. Lastly, for the purposes of simulation, we used results based on $10^4$ MCMC samples (instead of $10^6$) as the gold standard to reduce the total time of running these simulations. We compared the accuracies for 50 simulated data with the above settings and the results are summarized in Figure 4.

From Figure 4 we can see that SNVA performs better than the other methods for all settings for both the Poisson case and the logistic case. However, for the Poisson case, the relative advantage of SNVA over GVA and INLA seems to diminish for larger $m$. This is perhaps because, for larger $m$, the sample size $n$ is relatively large and the posterior density is reasonably Gaussian. Unfortunately this increased accuracy comes at a price. For the logistic case, although such a comparison is not completely fair, our implementation of SNVA was roughly 10 times slower than INLA.

Figure 3: The left panel illustrates the posterior approximations based on MCMC ($10^6$ samples), SNVA and INLA approximations for the logistic regression model for the breastfeed dataset. The right panel is a boxplot of the accuracies for the Laplace, GVA, SNVA, INLA and MCMC methods with $10^3$ samples (MC3) and $10^4$ samples (MC4).
5.3 Example 3 - Generalized Linear Mixed Models

We have thus far only considered fairly simple models. We now consider Generalized Linear Mixed Models (GLMM) which are significantly more difficult to fit using analytic approximations. For simplicity we will consider a simple subclass of GLMMs described by the model

\[ y|\beta, u \sim \text{Bernoulli}(\expit(X\beta + Zu)), \quad \text{or} \quad y|\beta, u \sim \text{Poisson}(\exp(X\beta + Zu)) \]  
(18)

using the same definitions of \(X, Z, \beta\) and \(u\) to those in Example 3. More general classes of GLMMs are described by Zhao et al. (2006) and Rue et al. (2009) but for simplicity will not be considered here. Using the same definitions, priors and transformations to those used in Example 3 the joint likelihood which may be written as

\[ p(y, \theta) \propto \exp \left\{ y^T C \nu - 1^T b(C \nu) - \frac{\| \theta / 2\|^2}{2 \sigma^2} - (A + \frac{m}{2}) \gamma - e^{-\gamma} \left( B + \frac{\| u \|^2}{2} \right) \right\} \]

where \( \theta = (\beta, u, \gamma) \).

The functions \(f_{SN}, g_{SN}\) and \(H_{SN}\) are given by

\[
f_{SN} = y^T C(\mu_\nu + \sqrt{\frac{2}{\pi}} \delta_\nu) - 1^T B(0)(C\mu_\nu, dg(C\Lambda_{\nu\nu}C^T), C\delta_\nu) - \frac{1}{2\sigma_\beta} \left[ \| \mu_\beta \|^2 + 2 \frac{\sqrt{2/\pi}}{\mu_\beta \delta_\beta + \text{tr} (\Lambda_{\beta\beta})} - (A + \frac{m}{2}) \left( \mu_\gamma + \sqrt{\frac{2}{\pi}} \delta_\gamma \right) - M \left( B + \frac{1}{2} T \right) \right],
\]

\[
g_{SN} = \begin{bmatrix} X^T (y - B(1)(C\mu_\nu, dg(C\Lambda_{\nu\nu}C^T), C\delta_\nu)) - (\mu_\beta + \sqrt{\frac{2}{\pi}} \delta_\beta) / \sigma_\beta \\ X^T (y - B(1)(C\mu_\nu, dg(C\Lambda_{\nu\nu}C^T), C\delta_\nu)) - (\mu_\beta + \sqrt{\frac{2}{\pi}} \delta_\beta) / \sigma_\beta \\ M \left( B + \frac{1}{2} T \right) - (A + \frac{m}{2}) \end{bmatrix}
\]

and

\[
H_{SN} = \begin{bmatrix} -X^T W X - \sigma_\beta^{-2} I & -X^T W Z & 0 \\ -Z^T W X & -X^T W X - M \tilde{u} & M \tilde{u} \\ 0 & M \tilde{u}^T & -M \left( B + \frac{1}{2} T \right) \end{bmatrix}
\]

where \( W = \text{diag}(B(2)(C\mu_\nu, dg(C\Lambda_{\nu\nu}C^T), C\delta_\nu)) \), \( M = 2e^{-\mu_\gamma + \Lambda_{\gamma\gamma} / 2 \Phi(-\delta_\gamma)}, T = \| \tilde{u} \|^2 + \text{tr} (\Lambda_{uu} + \zeta_2 (-\delta_\gamma) \delta_\delta u^T) \) and \( \tilde{u} = \mu_u - \Lambda_{\gamma u} + \zeta_1 (-\delta_\gamma) \delta u \).

As an example the Epilepsy dataset represents data collected from a clinical trial of 59 epileptics (Thall & Vail, 1990). This dataset has been described and analysed in several other places.
The reader may refer to Thall & Vail (1990), Breslow & Clayton (1993), Rue et al. (2009) or Ormerod & Wand (2011) for a full description. We will consider the Poisson random intercept model

\[ y_{ij}|\beta, u_i \sim \text{Poisson}\{\exp\{\beta_0 + u_i + \beta_{\text{visit}} \text{visit}_j + \beta_{\text{base}} \log(\text{base}_i/4) \\
+ \beta_{\text{trt}} \text{trt}_i + \beta_{\text{age}} \log(\text{age}_i)\}\} \]

where \( y_{ij} \) is a response of seizure counts, \( \text{base}_i, \text{trt}_i \) and \( \text{age}_i \) are covariates for \( 1 \leq i \leq 59, 1 \leq j \leq 4 \). Using an appropriate assignment of \( y, X, Z, \beta \) and \( u \) the model may be described by (18). Figure 5 summarizes the results obtained by using the Laplace, GVA, SNVA and INLA methods.

![Coefficient Accuracies](image)

**Figure 5:** Boxplot accuracies for the Laplace, GVA, SNVA and INLA methods and illustration of the approximations of the marginal posterior densities for \( \gamma \) for the Epilepsy example. The numbers in the legend correspond to the accuracies of each method.

From Figure 5 we see that the GVA, SNVA and INLA methods give much better results than the Laplace method. The GVA and SNVA methods are quite accurate, although the results of INLA are slightly better than GVA and SNVA. Although such a comparison, for various reasons, is not exactly fair the GVA, SNVA and INLA methods took around 2.5, 3.6 and 1.5 seconds respectively while MCMC took around 10 minutes to run.

### 5.4 Linear Regression with Inhomogeneous Noise

The INLA method is a specialized method for Gaussian latent effect models and cannot be used to handle every complication. Examples of complications which INLA cannot currently handle include missing data and inhomogeneous noise. Consider the model \( y|\beta, \alpha \sim N(X\beta, \text{diag}(\exp(Z\alpha))) \) where \( X \) and \( Z \) are design matrices and the vectors \( \beta \) and \( \alpha \) are their corresponding coefficients. We will employ the priors \( \beta \sim N(0, \sigma^2_\beta I) \) and \( \alpha \sim N(0, \sigma^2_\alpha I) \). The joint log-likelihood may be written as:

\[
p(y, \theta) \propto \exp\left\{ -\frac{1}{2} y^T Z \alpha - \frac{1}{2} (y - X\beta)^T \text{diag}(\exp(-Z\alpha))(y - X\beta) - \frac{\|\beta\|^2}{2\sigma^2_\beta} - \frac{\|\alpha\|^2}{2\sigma^2_\alpha} \right\}.
\]
The functions $f_{SN}$, $g_{SN}$ and $H_{SN}$ are given by

$$
f_{SN} = -\frac{1}{2} 1^T Z \left( \mu_\alpha + \sqrt{\frac{2}{\pi}} \delta_\alpha \right) - \frac{1}{2} 1^T M (y - \tilde{\mu})^2
$$

$$
- \frac{1}{2\sigma_\beta^2} \left[ ||\mu_\beta||^2 \sqrt{\frac{8}{\pi} \mu_\beta^T \delta_\beta} + \text{tr}(\Lambda_\beta) \right] - \frac{1}{2\sigma_\beta^2} \left[ ||\mu_\alpha||^2 + \sqrt{\frac{8}{\pi} \mu_\alpha^T \delta_\alpha} + \text{tr}(\Lambda_\alpha) \right],
$$

$$
g_{SN} = \begin{bmatrix}
X^T M (y - \tilde{\mu}) - (\mu_\beta + \sqrt{\frac{2}{\pi}} \delta_\beta) / \sigma_\beta^2 \\
\frac{1}{2} Z^T (M (y - \tilde{\mu})^2 - 1) - (\mu_\alpha + \sqrt{\frac{2}{\pi}} \delta_\alpha) / \sigma_\alpha^2
\end{bmatrix}
$$

and

$$
H_{SN} = \begin{bmatrix}
-X^T M X - \sigma_\beta^{-2} I \\
-Z^T M \text{diag}((y - \tilde{\mu})) X \\
-Z^T M \text{diag}((y - \tilde{\mu})) Z - \frac{1}{2} Z^T M \text{diag}((y - \tilde{\mu})^2) Z - \sigma_\alpha^{-2} I
\end{bmatrix}
$$

where $M$ is an $n \times n$ diagonal matrix and $\tilde{\mu}$ is a vector of length $n$ whose elements are given by $M_{ii} = 2e^{-z_i^T \mu_\alpha + z_i^T \Lambda_\alpha z_i} \phi(-z_i^T \delta_\alpha)$ and $\tilde{\mu}_i = x_i^T \mu_\beta - x_i^T \Lambda_\beta z_i + \zeta_1 (-z_i^T \delta_\alpha) x_i^T \delta_\beta$ respectively for $1 \leq i \leq n$ with $x_i$ and $z_i$ being the $i$th rows of $X$ and $Z$ respectively.

Consider data from a study investigating the metabolic effect of cross-country skiing (Zuliani et al., 1983). The response variable $y$ are measurements of the enzyme creatine phosphokinase (CPK) which is contained within muscle cells and is necessary for the storage and release of energy. It can be released into the blood in response to vigorous exercise from leaky muscle cells and occurs often even in healthy athletes. Subjects were participants in a 24 hour cross-country relay where age, weight and CPK concentration 12 hours into the relay were recorded. This data is available from the website http://www.statsci.org/data/general/bloodcpk.html.

We will use the design matrices with $x_i = (1, \text{weight}_i)$ and $z_i = (1, \text{age}_i)$. The results based on the Laplace, GVA and SNVA methods are illustrated in Figure 6. Noting that the mean accuracies for Laplace, GVA and SNVA are 80%, 91% and 97% respectively the SNVA method is clearly the most accurate.

![Figure 6: Illustration of the approximations of the marginal posterior distributions based on the the Laplace, GVA and SNVA methods for the CPK data.](image-url)
6 Conclusion and Discussion

In this paper we have shown for a number of simple models of interest that a multivariate skew-normal distribution can be used to efficiently fit posterior densities with high accuracy via parametric variational approximations. On the examples considered the SNVA method is shown to be either the most accurate (almost) analytic method or has comparable accuracy to INLA for most parameters. For one particular example SNVA had comparable accuracy to MCMC approximations using $10^5$ samples. Furthermore, the computational effort of the method is comparable to INLA.

In particular, based on the generalized linear models examples considered in this paper, SNVA appears to perform on par, if not better, than INLA, in terms of accuracy. It should be noted that for generalized linear models the Laplace approximation has error order $O(n^{-1})$ and the INLA method for this problem is simply an implementation of the Laplace approximation applied to ratios proposed by Tierney & Kadane (1986) which has error order $O(n^{-2})$. While the theory for GVA has not been thoroughly established it appears to have similar asymptotic properties to Laplace’s method, at least empirically. Lastly, the best SNVA approximation should be no worse than GVA.

The implementation which we pursued here was driven on the basis of practicality. There are several computational issues which could be addressed have potential to speed up SNVA. For example, the numerical quadrature routine used to calculate $B^r$ could be further tuned for better performance, alternative optimization routines could be explored, alternative initial values could be used and the method could be modified to take advantage sparseness which arises for various models.

Lastly, the ideas presented in this paper could foreseeably be extended to more general classes of skewed distributions, potentially increasing the accuracy of the method presented here.

Appendix A: Skew-Normal Distribution Results

A.1 Entropy Expression

Proof – Entropy Expression: It is straightforward to show, using the properties of the skew-normal distribution (outlined in Section 3), that

$$E(\mu, \Lambda, d) = \frac{1}{2} \log |2\pi\Lambda| + \frac{m}{2} - \log(2) - \int_{\mathbb{R}^m} \phi_\Lambda(\theta - \mu) 2\Phi(d^T(\theta - \mu)) \log \Phi(d^T(\theta - \mu)) d\theta.$$ 

By the affine transformation result (10) from $z = d^T(\theta - \mu)$ has density $q(z) = 2\phi_{\sqrt{d^T\Lambda d}}(z)\Phi(z)$. Using this fact, and in light of (11), the result follows directly.

Note that the function $\Psi(\sigma^2)$ is strictly increasing with $\Psi(0) = -\log(2)$. In practise we can calculate $\Psi(\sigma^2)$ and its derivatives using one-dimensional integration. Let $\Psi^{(r)}(\sigma^2) = \int_\mathbb{R} \phi(z)\psi^{(r)}(\sigma z) dz$ where $\psi(z) = \Phi(z)\log\Phi(z)$. Then the derivatives with respect to $\sigma^2$ are aided by the result $\partial\Psi^{(r)}(\sigma^2)/\partial\sigma^2 = \Psi^{(r+2)}(\sigma^2)/2$. Also note that $\lim_{\sigma^2 \to 0} \Psi^{(r)}(\sigma^2) = \psi^{(r)}(0)$.  

16
A.2 Derivative Expressions

In all of the results below we let $\Lambda \Lambda^T$ be the Cholesky factorization of $\Lambda$. Result 4 is used a few times to simplify some expressions later on.

**Result 4:** $2\phi_\Lambda(\theta)\phi(d^T\theta) = \sqrt{2/\pi}\phi_{\Lambda-\delta\delta^T}(\theta)/\sqrt{1 + d^T\Lambda d}$.

**Proof:** Firstly,

$$2\phi_\Lambda(\theta)\phi(d^T\theta) = \sqrt{2/\pi}\frac{1}{|\Lambda|^{1/2}} \exp\left\{-\frac{1}{2}\theta^T(\Lambda^{-1} + dd^T)\theta\right\} = \sqrt{2/\pi}\frac{|2\pi(\Lambda-\delta\delta^T)|^{1/2}}{|\Lambda|^{1/2}} \phi_{\Lambda-\delta\delta^T}(\theta)$$

since, via the Sherman-Morrison-Woodbury formula, we have $(\Lambda^{-1} + dd^T)^{-1} = \Lambda - \Lambda d(1 + d^T\Lambda d)^{-1}d^T\Lambda = \Lambda - \delta\delta^T$ and using standard algebraic manipulations it can be shown that $|2\pi(\Lambda-\delta\delta^T)|^{1/2}/|\Lambda|^{1/2} = (1 + d^T\Lambda d)^{-1/2}$.

**Result 5:** For any differentiable function $f(\theta)$

$$\int_{\mathbb{R}^m} f(\theta + \mu)\theta 2\phi_\Lambda(\theta)\Phi(d^T\theta)d\theta = \Lambda \left[ \int_{\mathbb{R}^m} [D_\theta f(\theta)] 2\phi_\Lambda(\theta)\Phi(d^T(\theta - \mu))d\theta \right] + \frac{\sqrt{2/\pi}}{\sqrt{1 + d^T\Lambda d}} \int_{\mathbb{R}^m} f(\theta)\phi_{\Lambda-\delta\delta^T}(\theta - \mu)d\theta$$

**Proof:** First note that

$$D_\theta 2\phi_\Lambda(\theta)\Phi(d^T\theta) = -\Lambda^{-1}2\phi_\Lambda(\theta)\Phi(d^T\theta) + \frac{\sqrt{2}}{\pi} \phi_{\Lambda-\delta\delta^T}(\theta)d$$

which follows from Result 4. Hence,

$$\int_{\mathbb{R}^m} f(\theta + \mu)\theta 2\phi_\Lambda(\theta)\Phi(d^T\theta)d\theta$$

$$= \Lambda \int_{\mathbb{R}^m} f(\theta + \mu) \left[ -D_\theta 2\phi_\Lambda(\theta)\Phi(d^T\theta) \right] d\theta + \frac{\sqrt{2}}{\pi} \left[ \int_{\mathbb{R}^m} f(\theta)\phi_{\Lambda-\delta\delta^T}(\theta - \mu)d\theta \right] \frac{d}{\sqrt{1 + d^T\Lambda d}}$$

Result 5 then follows by applying integration by parts and a change of variables.

In particular from Result 5, setting $d = 0$, we obtain

$$\int_{\mathbb{R}^m} f(\theta + \mu)\theta 2\phi_\Lambda(\theta)d\theta = \Lambda \int_{\mathbb{R}^m} [D_\theta f(\theta)] \phi_\Lambda(\theta - \mu)d\theta.$$  \hspace{1cm} (19)

**Proof of Derivative Expressions:** We will first consider derivatives of $f_{\Delta}(\mu, \Lambda, d)$ with respect to $\xi = (\mu, \Lambda, d)$. We will assume that for all $\xi_i$ that we can take derivatives with respect to $\xi_i$ inside the integral, i.e.

$$\frac{\partial}{\partial \xi_i} \int_{\mathbb{R}^m} q(\theta; (\mu, \Lambda, d))\log p(y, \theta)d\theta = \int_{\mathbb{R}^m} \frac{\partial}{\partial \xi_i} q(\theta; (\mu, \Lambda, d))\log p(y, \theta)d\theta$$
and that the first and second derivatives of $\log p(y, \theta)$ exist. Let $\log p(y, \theta) = f(\theta)$. Then the derivatives with respect to $\mu$ are

$$D_\mu f_{SN}(\mu, \Lambda, d) = D_\mu \int 2\phi_\Lambda(\theta)\Phi(d^T\theta)f(\mu + \theta)d\theta = \int 2\phi_\Lambda(\theta)\Phi(d^T\theta)D_\mu f(\mu + \theta)d\theta$$

$$= \int 2\phi_\Lambda(\theta)\Phi(d^T\theta)D_\mu \log p(y, \mu + \theta)d\theta = \int 2\phi_\Lambda(\theta - \mu)\Phi(d^T(\theta - \mu))\log p(y, \theta)d\theta = g_{SN}(\mu, \Lambda, d).$$

Next,

$$D_d f_{SN}(\mu, \Lambda, d) = D_d \int 2\phi_\Lambda(\theta)\Phi(d^T\theta)f(\mu + \theta)d\theta = \int 2\phi_\Lambda(\theta)\Phi(d^T\theta)[D_d \log p(y, \mu + \theta)d\theta$$

$$= \int 2\phi_\Lambda(\theta)\Phi(d^T\theta)f(\mu + \theta)d\theta = \sqrt{\frac{2}{\pi}} (\Lambda - \delta\delta^T)g_c(\mu, \Lambda - \delta\delta^T) \frac{\sqrt{1 + d^T\Lambda d}}{\sqrt{1 + \delta\delta^T}}.$$  

The last equality follows from Result 4, (19) and a change of variables. Finally,

$$\frac{\partial}{\partial \Lambda_{ij}} f_{SN}(\mu, \Lambda, d) = \int \frac{\partial}{\partial \Lambda_{ij}} f(\mu + A\theta)2\phi(\theta)\Phi(d^T A\theta)d\theta$$

$$= \left[ \int f(\mu + A\theta)d^T \left( \frac{\partial A}{\partial \Lambda_{ij}} \right) \theta \right] 2\phi(\theta)\phi(d^T\Lambda\theta)d\theta + \int \left[ g(\mu + A\theta)^T \left( \frac{\partial A}{\partial \Lambda_{ij}} \right) \theta \right] 2\phi(\theta)\Phi(d^T\Lambda\theta)d\theta$$

$$= \frac{1}{2} d^T E_{ij} \Lambda^{-1} \left[ \int f(\mu + \theta)\theta \sqrt{\frac{2}{\pi}} \frac{\phi_\Lambda - \delta\delta^T(\theta)}{\sqrt{1 + d^T\Lambda d}} d\theta + \frac{1}{2} \int g(\mu + \theta)^T E_{ij} \Lambda^{-1} \theta \right] 2\phi_\Lambda(\theta)\Phi(d^T\theta)d\theta$$

$$= \frac{1}{2} \text{tr} \left[ \left( H_{SN} + \sqrt{\frac{2}{\pi}} \Lambda^{-1}(\Lambda - \delta\delta^T)g_c\delta^T\Lambda^{-1} + \sqrt{\frac{2}{\pi}} \Lambda^{-1}\delta g_c^T \right) E_{ij} \right]$$

$$= \frac{1}{2} \text{tr} \left[ \left( H_{SN} + \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{\Lambda + d^T\Lambda d}} \left( g_c d^T + d g_c^T \right) - \frac{d^T A g_c}{1 + d^T\Lambda d} d^T \right) E_{ij} \right]$$

where $E_{ij}$ is the zero matrix with 1 in the $(i,j)$th entry.

**References**


