Improved Laplace Approximation

for Marginal Likelihoods

Erlis Ruli*, Nicola Sartori and Laura Ventura

Department of Statistics, University of Padova, Italy

*ruli@stat.unipd.it, sartori@stat.unipd.it, ventura@stat.unipd.it

February 24, 2015

Abstract

Intractable multidimensional integrals arise very often both in Bayesian and frequentist applications. The Laplace formula is widely used to approximate such integrals. However, in large dimensions, when the shape of the integrand function is far from that of the Gaussian density and/or the sample size is small, the standard Laplace approximation can be inaccurate. We propose an improved Laplace approximation which increases asymptotically the accuracy of the standard Laplace formula by one order of magnitude, and which is also accurate in high-dimensions. Applications to Bayesian inference in nonlinear regression models and to frequentist inference in a generalized linear mixed model with crossed random effects demonstrate the superiority of the proposed method with respect to the standard Laplace formula. The accuracy
of the proposed method is comparable with that of other existing methods, which are computationally more demanding.

*Keywords*: Asymptotic expansions for integrals; Bayes Factor; Integrated likelihood; Nested optimisation; Numerical integration.

### 1 Introduction

Several Bayesian and frequentist applications involve the evaluation of finite integrals of the form

\[ I_n = \int_{\mathbb{R}^d} \exp\{-h_n(x)\} \, dx , \]

where \(-h_n(x)\) is a smooth and concave real function of the \(d\)-dimensional real vector \(x\) which may depend on a quantity \(n > 0\). For instance, in Bayesian analyses \(-h_n(x)\) may be the log-likelihood or the kernel of the log-posterior, or in Generalized Linear Mixed Models (GLMM) it may represent the sum of the log-likelihood and the log-density of the random effects. In both cases, \(n\) is related to the information in the sample, and is typically the sample size.

Integral (I) is often intractable but it can be approximated by numerical integration methods, such as Gaussian quadratures, Monte Carlo integration methods or by Laplace’s formula (see, e.g., Evans & Swartz, 2000). Numerical integration via quadrature rules is typically accurate but, due to the curse of dimensionality, it is feasible only for low-dimensional integrals, i.e. \(d < 10\) (Evans & Swartz, 1995). Monte Carlo methods are widely applicable to general functions \(h_n(x)\), especially for large \(d\), provided \(I_n\) is finite. However, Monte Carlo integration is typically simulation consistent, in the sense that it reaches the true value as the number of simulations goes to infinity. Alternatively, when \(-h_n(x)\) has a single mode,
which becomes increasingly dominant for large values of $n$, methods based on asymptotic expansions, such as Laplace’s formula, are computationally more convenient. Asymptotic expansions around the mode give analytical approximations of $I_n$ which do not require tuning, and which are consistent for $n \to \infty$. Hence, unlike Monte Carlo approximations, the accuracy of asymptotic methods is bounded by some function of the given $n$. Nevertheless, when $-h_n(x)$ is single peaked, even for moderate values of $n$ asymptotic approximations may outperform Monte Carlo methods for a given computational cost (see, e.g., Rue et al., 2009; Ruli et al., 2014). This makes asymptotic approximations particularly appealing in applications.

We focus on the Laplace method of integration for approximating marginal likelihoods of the form (1), both for Bayesian and frequentist inference. The Laplace method is widely used in the Bayesian literature for approximating posterior densities and posterior moments (see, e.g., Tierney & Kadane, 1986; Rue et al., 2009) or Bayes Factors (see, e.g., Kass & Raftery, 1995), and in a frequentist framework for integrating out random effects in GLMM (see, e.g., Breslow & Clayton, 1993) or to compute marginal likelihoods in group models (Barndorff-Nielsen & Cox, 1994, Sect. 2.8; Pace et al., 2006). In addition, Guihenneuc-Jouyaux & Rousseau (2005) use Laplace’s formula within a Markov chain Monte Carlo (MCMC) framework, in which the nuisance parameters are integrated out by the Laplace method and the parameters of interest are integrated by MCMC methods. Furthermore, Ruli et al. (2014) propose a simulation algorithm which draws posterior samples by inverting the approximate cumulative distribution function based on the Laplace approximation for marginal posterior densities. Lastly, Martino et al. (2011) and Rizopoulos et al. (2009) apply the Laplace method in the
context of survival analysis and joint modelling of survival and longitudinal data, respectively.

We propose an improved Laplace approximation for integrals of the form (1). This approximation achieves third-order accuracy, i.e. it has error of order $O(n^{-3/2})$, in a standard asymptotic setting, i.e. when the sample size $n$ diverges and $d$ is fixed. The main idea behind the proposed method is to approximate (1) indirectly, through an approximation of the density given by the normalised integrand function. Such an approximation involves $d$ scalar Laplace approximations for marginal densities (Tierney & Kadane, 1986), which can be easily renormalised numerically. The ratio between the integrand function and the approximated density, both evaluated in an arbitrary given point, provides the improved Laplace approximation of (1). This is similar in spirit to Chib (1995).

With respect to the standard Laplace approximation, the proposed method requires some additional optimisation tasks and $d$ scalar numerical integrations. However, the most demanding task is the localisation of the global mode, which is a requirement also for the standard Laplace method. More details on implementation issues are discussed in Section 4 and 5.

The rest of the article is structured as follows. Section 2 sets the background. Section 3 introduces the improved Laplace approximation. Section 4 illustrates the proposed method in three applications and Section 5 concludes the paper with some final remarks.

2 Background

In order to ease notation, in the following we drop $n$ form $I_n$, $h_n(x)$ and related quantities. Hence, let $h(x)$ be a smooth function of $x = (x_1, \ldots, x_d)$, at least twice differentiable and
with unique minimum at \( \hat{x} = (\hat{x}_1, \ldots, \hat{x}_d) \). In addition, we assume that \( \hat{x} \) is the only point for which
\[
\nabla h(x) = \left( \frac{\partial h(x)}{\partial x_1}, \ldots, \frac{\partial h(x)}{\partial x_d} \right) = 0,
\]
and that the Hessian matrix of \( h(x) \) at \( \hat{x} \), i.e.
\[
\hat{V} = V(\hat{x}) = \left. \frac{\partial^2 h(x)}{\partial x \partial x^T} \right|_{x=\hat{x}},
\]
is positive definite. Finally, let \( H(x) = \exp\{-h(x)\} \).

The standard Laplace approximation of (1) can be written as \( I = \hat{I}^L \{1 + O(n^{-1})\} \), with
\[
\hat{I}^L = (2\pi)^{d/2} |\hat{V}|^{-1/2} H(\hat{x});
\tag{2}
\]
see, e.g., Bleistein & Handelsman (1986, p. 335).

Consider a sample \( y = (y_1, \ldots, y_n) \) of size \( n \) from a model with probability mass or density function \( p(y; \theta), \theta \in \Theta \subseteq \mathbb{R}^d \). In Bayesian applications, \( -h(\cdot) \) can be the logarithm of the posterior distribution \( \pi(\theta | y) \propto L(\theta; y) \pi(\theta) \), where \( L(\theta; y) \) is the likelihood function for \( \theta \) and \( \pi(\theta) \) is the prior distribution. In this case, (1) gives the posterior normalizing constant that can be used, for instance, to compute Bayes Factors (BFs). In GLMM, \( -h(\cdot) \) is the logarithm of the likelihood, conditional on the random effects, times the density of the latter. In this case, the integration is performed with respect to the random effects and (1) gives the marginal likelihood of the fixed parameters given the observed data \( y \).

In the standard asymptotic setting with \( d \) fixed and \( n \to \infty \), the standard Laplace approximation (2) is second-order accurate. On the other hand, if also \( d \to \infty \), the asymptotic expansion requires more terms in order to achieve the same accuracy as in lower-dimensions. For instance, in the fortunate but unlikely case of (1) being factorisable into a product of
$d$ scalar identical integrals, the relative error would be of order $O(d/n)$ (Small, 2010, Sect. 6.9). However, in practice, $n$ is fixed and one may question if such accuracy is sufficient for practical purposes where $d$ may be large as well, and if it can be improved to some extent.

Most of the approaches for reducing the error of the standard Laplace approximation involve the inclusion of higher-order derivatives of $h(x)$ in the Taylor expansions. Lindley (1980) use this idea in a Bayesian context. Raudenbush et al. (2000) propose a higher-order Laplace approximation for GLMM, by considering derivatives of $h(\cdot)$ up to the the sixth order. Pace et al. (2006) use a similar approach for approximating marginal likelihoods in group models. However, when $d > 1$, the computation of higher-order derivatives can be demanding. Similar strategies are the Bayesian Bartlett correction proposed by DiCiccio et al. (1997), and the corrected Laplace approximation of Shun & McCullagh (1993). The latter solution is designed for situations, such as models with crossed random effects, in which the standard Laplace approximation may not be asymptotically valid.

Another improvement of the standard Laplace approximation is proposed by Nott et al. (2009), in which (1) is approximated by the product of scalar blocks, after a preliminary approximate orthogonalising variable transformation. Although the proposal of Nott et al. (2009) tends to work slightly better than the standard Laplace approximation, both methods have the same asymptotic error.

Finally, we notice that the proposed method differs from the integrated nested Laplace approximation (INLA) of Rue et al. (2009). Indeed, in INLA the approximation of the marginal likelihood requires multivariate integration whereas the improved Laplace breaks the multivariate integration problem in $d$ univariate integrations, by sequentially applying the renor-
malised Laplace approximation for scalar marginal densities.

3 The improved Laplace approximation

Let us consider the probability density function \( p(x) = \exp\{-h(x)\}/I \), where \( \exp\{-h(x)\} \) is the integrand function and \( I \) is the unknown quantity of interest.

By definition

\[
I = \frac{\exp\{-h(x)\}}{p(x)},
\]

and if \( p(x) \) is known, (3) readily provides \( I \) for every value of \( x \). The idea is to use a suitable estimate \( \hat{p}(x) \) of \( p(x) \) and to obtain from (3) an estimate \( \hat{I} \) of \( I \), given by \( \exp\{-h(x)\}/\hat{p}(x) \).

This was first explored by Chib (1995) for estimating Bayesian marginal likelihoods, with \( p(x) \) approximated by Gibbs sampling (see also Chib & Jeliazkov, 2001, for some extensions to general MCMC methods). In principle the method works for any \( x \). However, it is advisable to locate such point at a high density region (Chib, 1995). One possibility is to choose \( x = \hat{x} \), which in our case is also convenient from a computational point of view, as explained below.

We propose to approximate \( p(x) \) as follows. First, \( p(x) \) is split according to the law of total probability in \( d \) scalar densities. Each factor is approximated using the Laplace formula (2) for marginal densities proposed by Tierney & Kadane (1986) and renormalised numerically. The product of these approximations provides a third-order asymptotic estimate of \( p(x) \) which, when substituted in (3), provides the improved Laplace approximation of \( I \), denoted \( \hat{I}^{\text{IL}} \).

In the following, we give the details for the approximation of \( \hat{p}(\hat{x}) \). Let \( x_{1:q} = (x_1, \ldots, x_q) \) denote the first \( q \) components of \( x \), and similarly let \( x_{q+1:d} \) denote the last \( q + 1 \) components...
of $x$ ($q < d$). Moreover, let $\hat{x}_1$ be the constrained minimum of $h(x)$ with $x_1$ fixed and let $\hat{x}_{1:q};x_{q+1}$ be the constrained minimum with $x_{1:q}$ fixed at $\hat{x}_{1:q}$ and $x_{q+1}$ fixed. For the validity of the proposed method we require that $h(x)$ satisfies the assumptions in Section 2. In addition, we also assume that the Hessian of each constrained minimisation is positive definite. Essentially, these assumptions are similar to those required by the standard Laplace approximation for marginal posteriors (Tierney & Kadane, 1986).

The proposed improved Laplace approximation is obtained as follows. We write $p(x)$ as

$$p(x) = p(x_1) \times p(x_2|x_1) \times \cdots \times p(x_d|x_{d-1}, \ldots, x_1)$$

$$= \frac{\int \exp\{-h(x)\} dx_2 \cdots dx_d}{\int \exp\{-h(x)\} dx_1 \cdots dx_d} \times \cdots \times \frac{\exp\{-h(x)\}}{\int \exp\{-h(x)\} dx_d}.$$  

(4)

Then, we approximate separately each factor on the right-hand side of (4) by the Laplace approximation for marginal densities and we renormalise it numerically. In particular, using results from Tierney & Kadane (1986), for the marginal density $p(x_1)$ we have that

$$\hat{p}(x_1) = \frac{H(x_1, \hat{x}_1)}{H(\hat{x})} \left\{ \frac{|\hat{V}|}{2\pi |V_{2:d}(x_1, \hat{x}_1)|} \right\}^{1/2},$$

where $V_{2:d}(\cdot)$ is the block $(2:d, 2:d)$ of the Hessian matrix $V(\cdot)$. Upon numerical renormalisation, this marginal density is third-order accurate. Indeed, let

$$\hat{p}^*(x_1) = \frac{\hat{p}(x_1)}{\int \hat{p}(x_1) dx_1}.$$  

Then, we have $p(x_1) = \hat{p}^*(x_1)\{1 + O(n^{-3/2})\}$. From this result we readily have a third order-accurate approximation of $p(\hat{x}_1)$, given by $\hat{p}^*(\hat{x}_1)$.

Consider now the second ratio of integrals in (4), i.e. the conditional density $p(x_2|x_1)$. As for $p(x_1)$, we apply again the standard Laplace method to the numerator and denominator,
but with $x_1$ fixed at $\hat{x}_1$. This procedure yields the approximation

$$\hat{p}(x_2|\hat{x}_1) = \frac{H(\hat{x}_1, x_2, \hat{x}_1, x_2)}{H(\hat{x})} \left\{ \frac{|V_{2,2}|}{2\pi |V_{3,2}(\hat{x}_1, x_2, \hat{x}_2)|} \right\}^{1/2}.$$  

Since $\hat{p}(x_2|\hat{x}_1)$ is univariate, it can be easily normalised by scalar numerical integration to give

$$\hat{p}^*(x_2|\hat{x}_1) = \frac{\hat{p}(x_2|\hat{x}_1)}{\int \hat{p}(x_2|\hat{x}_1) \, dx_2},$$

which is such that $p(x_2|\hat{x}_1) = \hat{p}^*(x_2|\hat{x}_1)\{1 + O(n^{-3/2})\}$. Then, evaluated at $\hat{x}_2$, it gives a third-order approximation of $p(\hat{x}_2|\hat{x}_1)$.

More generally, the Laplace approximation of the conditional density $p(x_q|\hat{x}_{q-1}, \ldots, \hat{x}_1)$, with $q < d$, can be written as

$$\hat{p}(x_q|\hat{x}_{q-1}, \ldots, \hat{x}_1) = \frac{H(\hat{x}_{1:q-1}, x_q, \hat{x}_{1:q-1}, x_q)}{H(\hat{x})} \left\{ \frac{|V_{q,1}|}{2\pi |V_{q-1,2}(\hat{x}_{1:q-1}, x_q, \hat{x}_{1:q-1}, x_q)|} \right\}^{1/2},$$

and the renormalised version

$$\hat{p}^*(x_q|\hat{x}_{q-1}, \ldots, \hat{x}_1) = \frac{\hat{p}(x_q|\hat{x}_{q-1}, \ldots, \hat{x}_1)}{\int \hat{p}(x_q|\hat{x}_{q-1}, \ldots, \hat{x}_1) \, dx_q}$$

is third-order accurate.

Finally, notice that the last conditional density on the right hand side of (4), with $x_{p-1}, \ldots, x_1$ all fixed at their respective modal values, does not involve asymptotics nor optimisations, and only need renormalisation by numerical integration.

The product of these $d-1$ Laplace approximations evaluated at their modal values delivers

$$\hat{p}^*(\hat{x}) = \hat{p}^*(\hat{x}_1) \left\{ \prod_{q=2}^{d-1} \hat{p}^*(\hat{x}_q|\hat{x}_{q-1}, \ldots, \hat{x}_1) \right\} p(\hat{x}_d|\hat{x}_{d-1}, \ldots, \hat{x}_1)$$

$$= p(\hat{x}_1) \left\{ \prod_{q=2}^{d-1} p(\hat{x}_q|\hat{x}_{q-1}, \ldots, \hat{x}_1) \right\} p(\hat{x}_d|\hat{x}_{d-1}, \ldots, \hat{x}_1)\{1 + O(n^{-3/2})\}^{d-1}$$

$$= p(\hat{x})\{1 + O(n^{-3/2})\},$$

9
i.e. a third-order accurate approximation of \( p(x) \) at \( x = \hat{x} \). By replacing this approximate density in (3), we obtain

\[
I = \hat{I}^{iL} \{ 1 + O(n^{-3/2}) \}.
\]

Hence, in a standard asymptotic setting with \( d \) fixed and \( n \to \infty \), the proposed improved Laplace approximation is third-order accurate. On the other hand, if also \( d \to \infty \), it is easy to see that the error of the approximation becomes \( O(d/n^{3/2}) \).

**Example 1: Fixed \( d \) and \( n \to \infty \).** Let \( y = (y_1, \ldots, y_n) \) be a random sample of size \( n \) from a Gompertz random variable with density function \( p(y; \theta) = \exp\{\alpha + \beta + e^\beta y + e^\alpha - \alpha e^\beta y\} \) and parameter \( \theta = (\alpha, \beta) \in \mathbb{R}^2 \). For each sample size \( n = \psi + \sqrt{\psi} \), for \( \psi = 3, 8, 13, \ldots \) (rounded at the nearest integer), we simulate 20 dataset from the Gomperz distribution with \( \alpha = 2 \) and \( \beta = 3 \) and compute the normalizing constant of \( L(\theta; y)\pi(\theta) \) with adaptive quadrature, with the standard Laplace approximation and with the improved Laplace approximation. The prior \( \pi(\theta) \) is chosen as the product of two Gaussian densities with zero means and variances equal to 100.

The aim is to compare the behaviour the standard \((\hat{I}^L)\) and the improved \((\hat{I}^{iL})\) Laplace approximations with the target value \((I)\) computed by adaptive quadrature methods, as \( n \) diverges. In particular, following Davison et al. (2006), let \( c_1 > 0 , c_2 > 0 \), \((b_1, b_2) \in \mathbb{R}^2 \) and suppose that

\[
I = \hat{I}^{iL}(1 + b_1 n^{-c_1}) + o(n^{-c_1}),
\]

and that

\[
I = \hat{I}^L(1 + b_2 n^{-c_2}) + o(n^{-c_2}),
\]
Figure 1: Asymptotic error of the improved and standard Laplace methods. Left panel: log-log plot of \( \hat{I}_{iL} / I \) (\( \bullet \)) and \( \hat{I}_L / I \) (\( \circ \)) against \( n \). Right panel: log-log plot of the relative error \( |\hat{I}_{iL} / I - 1| \) (\( \bullet \)) and \( |\hat{I}_L / I - 1| \) (\( \circ \)) against \( n \); the solid and dashed lines are the corresponding least-squares regression lines. The empirical slopes are -1.48 with 0.99 confidence interval (-1.52, -1.43) for the solid line, and -0.98 with 0.99 confidence interval (-1.01,-0.95) for the dashed line.

for \( n \to \infty \). Then, \( \lim_{n \to \infty} \{ \hat{I}_{iL} / I \} = 1 \) and \( \lim_{n \to \infty} \{ \hat{I}_L / I \} = 1 \), and if the improved Laplace approximation is more accurate than the standard Laplace, then the first limit should approach 1 faster.

The left panel of Figure 1 reports the log-log plot of \( \hat{I}_{iL} / I \) and \( \hat{I}_L / I \), both averaged over the repetitions at each value of \( n \), against the sample size \( n \). This plot highlights that the improved Laplace approximation is more accurate than the standard Laplace method, and that the convergence at 1 of \( \hat{I}_{iL} / I \) is almost immediate.
The log-log plot of the relative error against $n$ on the right panel of Figure 1 shows that the improved Laplace method achieves third-order accuracy whereas the standard Laplace formula is second-order accurate, e.g. $c_1 = 1.5$ and $c_2 = 1$.

**Example 2: Fixed $n$ and $d \to \infty$.** Consider the multivariate $t$-skew-$t$ distribution proposed by Jones (2002), which is given by the multivariate $t$-Student density centred at 0 and with identity scale matrix, where the marginal density of the first component is replaced with the univariate $t$-skew-$t$ density (see, Jones, 2002, Eq. (6)). As in Jones (2002, p. 95), we consider the parametrisation with $a$, $c$ and $\nu$, where $a > 0$ and $c > 0$ determine the distribution of the skewed marginal and $\nu$, the degrees of freedom (df), controls the tail behaviour of the distribution. The case with $a = c = \nu/2$ leads the the ordinary multivariate $t$-Student distribution with identity scale matrix and $\nu$ df.

We approximate the normalizing constant of the multivariate $t$-skew-$t$ density, which is obviously 1, with the improved and with the standard Laplace approximations, in two scenarios: one with $a = c = 1.5$ and one with $a = 12$ and $c = 0.5$. For each scenario we consider multivariate $t$-skew-$t$ densities with dimensions $3, 8, \ldots, 98$. The results are shown in Figure 2. First of all, note that the standard Laplace approximation rapidly deteriorates in higher dimensions, in both scenarios. Second, the standard Laplace approximation is worse for lower degrees of freedom. This is obvious since the Laplace approximation works better with quadratic functions, which in this case are obtained in the limit as $\nu \to \infty$ and $a = c = \nu/2$. Finally, also with higher skewness (large $a$ and small $c$ or vice versa) again the quality of the standard Laplace approximation deteriorates. On the contrary, the improved Laplace approximation is not affected by any of these conditions, and gives almost exact
Figure 2: Standard (red colored symbols) and improved Laplace (black) approximations of the normalizing constant of the multivariate \(t\)/skew-\(t\) density against the dimension, with various parameter values and df: 3 (○), 5 (∆), 10 (+) and 20 (×).

answers in both scenarios. A similar example was considered also by Nott et al. (2009), in order to test the accuracy of their modified Laplace approximation. However, their method is substantially less accurate, only slightly improving the poor quality of the standard Laplace approximation. For instance, the normalising constant of the 10-variate \(t\)/skew-\(t\) density with \(a = 4, c = 1\) and \(\nu = 3\), is 0.013 with the standard Laplace method, 0.02 with the modified Laplace approximation of Nott et al. (2009) and 0.9981 with the proposed improved Laplace approximation.
4 Applications

In this section we illustrate the accuracy of the proposed method by three applications. The first two applications focus on Bayesian inference in nonlinear regression models, where the aim is to choose between the normal and the \( t \)-Student error distributions. The third application is a logistic regression model with crossed random effects applied to the well known Salamander data (see McClullagh & Nelder, 1989, p. 440). The Maximum Likelihood Estimate (MLE), obtained by maximising the approximate marginal log-likelihood computed by the improved Laplace method, is compared with the analogous obtained with other competing methods.

All examples can be reproduced by using the R (R Core Team, 2014) package \texttt{iLaplace}, available online as Supplementary Material. The core functions needed to run the examples are implemented in C++ and are called from R via the \texttt{Rcpp} package. For the minimisations we use the function \texttt{nlminb} to which we supply the analytical gradient and Hessian. The nested optimisations are started at the global minimum and the numerical integrations are performed with the function \texttt{integrate}.

4.1 Nonlinear regression with BOD2 data

Consider the nonlinear regression model

\[
y_i = \beta_1 \exp\{1 - \exp(-x_i/\beta_2)\} + \sigma \varepsilon_i,
\]

where \( y_i \) is the response variable, \( x_i \) is a covariate, \((\beta_1, \beta_2)\) are unknown regression parameters and the \( \varepsilon_i \) are independent error terms, \( i = 1, \ldots, n \).
We focus on two possible distributions for the error terms: the normal and the $t$-Student with unknown df $\nu$, distributions. The aim is to choose among them using the BF. For the normal model we need to integrate out the parameter $(\beta, \sigma)$, whereas for the $t$-Student model the parameter to be integrated out is $(\beta, \sigma, \nu)$.

As an example we consider the BOD2 dataset (Bates & Watts, 1988, p. 305), which concern a study on biochemical oxygen concentration ($y$) as function of time ($x$). We assume the parameter a priori independent with a bivariate $t$-Student distribution centred at zero, with scale matrix $10I_2$ and 5 df for $\beta$. For the scale parameter of both models, following the recommendations of Gelman (2006), we assume a half-Cauchy prior with scale equal to 10. Finally, the Jeffreys rule prior proposed by Fonseca et al. (2008) is taken for $\nu$. To avoid numerical issues, $\sigma$ and $\nu$ are considered in logarithmic scale.

We compare the improved Laplace approximation with the standard Laplace and the Bartlett-corrected Laplace approximations. Comparisons are performed also with the importance sampling (IS) approximation and with the method proposed by Chib & Jeliazkov (2001). The Bartlett-corrected Laplace approximation is performed with a $10^7$ MCMC sample, which after a burn-in period and suitable thinning resulted in $10^6$ final draws. The same posterior sample is used also for computing the marginal likelihood with Chib & Jeliazkov’s method. For IS we consider $10^6$ draws from the multivariate $t$-Student distribution with 3 df, centred at the posterior mode, with scale matrix equal to $c > 0$ times the inverse of the posterior Hessian at the modal value. Finally, we also compare the results with those obtained using the adaptive numerical integration method implemented in the R package cubature.

From the results, shown in Table I, we notice that the standard Laplace approximation
Table 1: BOD2 data. Comparison of the improved Laplace method with adaptive numerical integration over hypercubes, the standard Laplace, the Bartlett-corrected Laplace, importance sampling (IS) and Chib & Jeliazkov’s approximations.

and its Bartlett-corrected version are quite inaccurate, since both lead to substantial false evidence for the normal model against the t-Student model (see Kass & Raftery, 1995 for the interpretation of the BF). More accurate approximations, such as adaptive integration, IS and Chib & Jeliazkov’s method do not confirm such evidence. The result of the improved Laplace method is in reasonable accordance with IS, Chib & Jeliazkov’s approximation and adaptive numerical integration.

The reason of the inaccuracy of the standard Laplace approximation in the case of the t-Student model is most likely due to the non normality of the marginal posterior of $(\log \sigma, \log \nu)$. Indeed, a look at the bivariate kernel density estimate of this marginal bivariate posterior (not reported here) reveals that it is banana-shaped, and therefore it is quite far from being elliptical. Nevertheless, such a shape is well accommodated by the improved Laplace approximation.
4.2 Nonlinear regression with Lubricant data

As a second example we consider the Lubricant data (Bates & Watts, 1988, p. 275), which concern the kinematic viscosity of a lubricant \( y \) as a function of temperature \( x_1 \) and pressure \( x_2 \). For the \( i \)th observation, the model can be written as

\[
y_i = \frac{\beta_1}{\beta_2 + x_{i1}} + \beta_3 x_{i2} + \beta_4 x_{i2}^2 + \beta_5 x_{i2}^3 + (\beta_6 + \beta_7 x_{i2}) x_{i2} \exp \left( -\frac{x_{i1}}{\beta_8 + \beta_9 x_{i2}^2} \right) + \sigma \varepsilon_i,
\]

\( i = 1, \ldots, n \). As in the previous example, the aim is to choose between the normal and the \( t \)-Student with \( \nu \) unknown df, error distributions using the BF. For the normal model, the parameter to be integrated out is \( (\beta, \sigma) \), whereas for the \( t \)-Student model the parameter is \( (\beta, \sigma, \nu) \), with \( \beta = (\beta_1, \ldots, \beta_9) \).

We assume the parameter a priori independent, and a 9-variate \( t \)-Student distribution with location \( 0_9 \), scale matrix \( 50I_9 \) and 5 df for \( \beta \). Moreover, for \( \sigma \) and \( \nu \) we take a half-Cauchy distribution with scale 10 and the Jeffreys product rule prior (Fonseca et al., 2008), respectively. To avoid numerical issues, \( \sigma \) and \( \nu \) are considered in logarithmic scale.

We compare the improved Laplace approximation with the standard Laplace and the Bartlett-corrected Laplace approximations. Comparisons are performed also with the IS approximation and the method of Chib & Jeliazkov (2001). We do not pursue comparisons with adaptive numerical integration as the dimension of the integral is too high for the results to be trustworthy within a reasonable amount of time. The Bartlett-corrected Laplace approximation is performed with a \( 10^7 \) MCMC sample, which after a burn-in period and suitable thinning resulted in \( 10^6 \) final draws. The same posterior sample is used also for computing the marginal likelihood with Chib & Jeliazkov’s method in one block sampling. For IS we consider \( 10^6 \) draws from the multivariate \( t \)-Student with 3 df, centred at the
posterior mode, with scale matrix equal to $c > 0$ times the inverse of the posterior Hessian at the modal value.

| Model             | Laplace | Improved | Bartlett-corrected | Chib & Jeliazkov | IS |
|-------------------|---------|----------|--------------------|------------------|----|
| Normal            | 12.043  | 12.503   | 12.570             | 12.327           | 12.500 |
| $t$-Student       | 8.399   | 9.886    | 9.478              | 9.544            | 9.927 |

$\log_{10} BF$ | 1.583   | 1.137    | 1.343              | 1.209            | 1.117 |

(Normal vs $t$)

Table 2: Lubricant viscosity data. Comparison of the improved Laplace method with the standard Laplace, the Bartlett-corrected Laplace, the importance sampling (IS) and the Chib & Jeliazkov’s approximations.

The various approximations of the posterior normalizing constant (in logarithmic scale) and of the related $\log_{10} BF$ are illustrated in Table 2. All the approximation methods offer some evidence against the normal model in favour of the more robust $t$-Student model. However, compared with IS and Chib & Jeliazkov’s method, the standard Laplace and the related Bartlett-corrected approximations appear to be slightly higher, whereas the improved Laplace gives very similar results.

Also here the inaccuracy of the standard Laplace approximation is most presumably due to the non-elliptically shaped marginal posterior of $(\log \sigma, \log \nu)$.  

18
4.3 Salamander mating data

We consider the problem of approximating the marginal likelihood for the fixed parameters in a model with crossed random effects (Shun & McCullagh, 1995; Shun, 1997). The marginal likelihood for the fixed parameters \((\theta, \theta_u)\) in mixed-effects models can be generally written as

\[
L(\theta, \theta_u; y) = \int L(\theta, \theta_u, u; y) f(u; \theta_u) \, du,
\]

where \(f(u; \theta_u)\) is the density of the random effects. This function can be written also as

\[
L(\theta, \theta_u; y) = \int \exp\{\ell(\theta, \theta_u, u; y) + \log f(u; \theta_u)\} \, du
\]

\[
= \int e^{-h(u; \theta, \theta_u, y)} \, du
\]

\[
= \frac{\exp\{-h(\hat{u}; \theta, \theta_u, y)\}}{f(\hat{u}; \theta, \theta_u, y)},
\]

where \(f(u; \theta, \theta_u, y)\) is such that \(\int f(u; \theta, \theta_u, y) \, du = 1\). Hence, in this case, given \((\theta, \theta_u)\) we aim at integrating out \(u\) from the product of the density of the data and the density of \(u\).

As an example, consider the Salamander mating data first published in McClullagh & Nelder (1989, p. 439). These data have been analysed also by Karim & Zeger (1992), Shun (1997), Booth & Hobert (1999), Bellio & Varin (2005), Sung & Geyer (2007), among others, and consist of three separate experiments, each performed according to the design given in McClullagh & Nelder (1989, Table 14.3). Each experiment involved matings among salamanders in two closed groups. Both groups contained five species R females, five species W females, five species R males and five species W males. Within each group, only 60 of the possible 100 heterosexual crosses were observed owing to time constraints. Thus, each experiment resulted in 120 binary observations indicating which matings were successful and which were not.
As in McClullagh & Nelder (1989, p. 441), the data are modelled as if different sets of 20 male and 20 female salamanders were used in each experiment. Let $y_{ij}$ be the indicator of a successful mating between female $i$ and male $j$, for $i, j = 1, \ldots, 60$, where only 360 of the $(i,j)$ pairs are relevant. Let $u_i^f$ denote the random effect that the $i$th female salamander has across matings in which she is involved, and define $u_j^m$ similarly for the $j$th male. The data $y_{ij}$ are assumed conditionally independent with

$$y_{ij} | u_i^f, u_j^m \sim \text{Bernoulli}(\pi_{ij}),$$

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = x_{ij}^T \beta + u_i^f + u_j^m,$$

where $x_{ij}$ is a 4-dimensional row vector of zeros and ones indicating the type of cross and $\beta$ is the vector of regression parameters. Moreover, $u_i^f$ and $u_j^m$ are assumed to be independent normals with zero mean and variance $\sigma_f^2$ and $\sigma_m^2$, respectively.

As a first example we consider the estimation of the fixed effect and $(\sigma_f^2, \sigma_m^2)$ for each separate experiment – following the same model structure as in Shun (1997) – by maximising the approximate marginal likelihood. The aim is to compare the MLE based on the marginal likelihood approximated by the improved Laplace method with those based on the modified Laplace approximation proposed by Shun & McCullagh (1995) and Shun (1997). It is well known that in models with crossed random effects the standard Laplace approximation is not asymptotically valid (Shun & McCullagh, 1995), and it may give poor results.

Let $\beta = (\beta_0, \beta_{WS_f}, \beta_{WS_m}, \beta_{WS_f \times WS_m})$ be the fixed effects, where $\beta_0$ is a constant, $\beta_{WS_f}$ is the effect of the dummy variable WS which takes one if the observation is from a species W female and zero otherwise, and so on. The marginal likelihood is given by

$$L(\beta, \sigma_f^2, \sigma_m^2) = \int \int L(\beta, \sigma_f^2, \sigma_m^2; u_f, u_m) f(u_f; \sigma_f^2) f(u_m; \sigma_m^2) du_f du_m.$$
which involves a 40 dimensional integral that cannot be reduced to a product of lower dimensional integrals, even if the random effects have independent normal distributions. The approximate MLEs for the three separate experiments (reported in Table 3) are compared with those of Shun (1997, Tab. 2 and Tab. 3). In this example the improved Laplace approximation is run in parallel over 11 threads. The maximisations of the approximate marginal likelihood computed with the improved Laplace is initialised at \((\beta, \log \sigma_f^2, \log \sigma_m^2) = (0, 0, 0, 0, 0, 0)\), and in the three experiments took 5, 3 and 4 minutes, respectively in a CentOS workstation with 64Gb Ram. The convergence is reached after 17, 13 and 14 iterations, respectively.

From Table 3 we notice that the approximate MLEs obtained with the improved Laplace method are very close to those based on the modified Laplace approximation of Shun (1997). However, the improved Laplace method compared to Shun (1997) is easier to compute since it does not require log-posterior derivatives beyond the second-order.

Consider now the joint analysis of the Salamander data, by independently combining the three experiments’ data. In this case the marginal likelihood entails the computation of three 40-dimensional integrals. To compare our method with other results available in the literature we consider a slightly modified version of the fixed effects. In particular, here \(\beta\) is equal to \((\beta_{R/R}, \beta_{R/W}, \beta_{W/R}, \beta_{W/W})\), where \(\beta_{R/R}\) denotes the effect of the cross between a species R female and a species R male, and so on. The minimisation of the negative marginal likelihood approximated by the improved Laplace method, started at the same point as before, took 20 minutes and it converged after 23 iterations.

The approximate MLE obtained from the improved Laplace approximation, the Monte Carlo Expectation-Maximisation (MC-EM) algorithm of Booth & Hobert (1999), the quasi-
### Table 3: Salamander mating data. Comparison of the improved Laplace method with the standard Laplace and the modified Laplace approximation of Shun (1997).

| Methods | Approximate MLE |
|---------|-----------------|
|         | $\beta_0$  | $\beta_{WS_f}$ | $\beta_{WS_m}$ | $\beta_{WS_f \times WS_m}$ | $\sigma_f^2$ | $\sigma_m^2$ |
| Laplace: | | | | | | |
| Exper. 1 | 1.34 | -2.94 | -0.42 | 3.18 | 1.59 | 0.073 |
| Exper. 2 | 0.57 | -2.46 | -0.77 | 3.71 | 1.82 | 0.92 |
| Exper. 3 | 1.02 | -3.23 | -0.82 | 3.82 | 0.35 | 1.85 |
| Laplace of Shun (1997)$^a$: | | | | | | |
| Exper. 1 | 1.37 | -3.02 | -0.44 | 3.27 | 1.72 | 0.185 |
| Exper. 2 | 0.58 | -2.54 | -0.77 | 3.79 | 2.10 | 1.10 |
| Exper. 3 | 1.04 | -3.31 | -0.83 | 3.90 | 0.46 | 2.07 |
| Improved Laplace: | | | | | | |
| Exper. 1 | 1.37 | -3.02 | -0.44 | 3.27 | 1.74 | 0.19 |
| Exper. 2 | 0.56 | -2.52 | -0.75 | 3.76 | 2.13 | 1.14 |
| Exper. 3 | 1.03 | -3.30 | -0.82 | 3.90 | 0.49 | 2.12 |

$^a$Corrected values from Shun (1997)
| Methods                | $\beta_0$ | $\beta_{WS_f}$ | $\beta_{WS_m}$ | $\beta_{WS_f \times WS_m}$ | $\sigma_f^2$ | $\sigma_m^2$ |
|------------------------|-----------|----------------|----------------|----------------------------|---------------|---------------|
| Laplace                | 1.01      | 0.31           | -1.90          | 0.99                       | 1.17          | 1.04          |
| Improved Laplace       | 1.02      | 0.32           | -1.95          | 1.00                       | 1.39          | 1.25          |
| MC-EM (Booth & Hobert, 1999) | 1.03      | 0.32           | -1.95          | 0.99                       | 1.40          | 1.25          |
| Gibbs (Karim & Zeger, 1992) | 1.03      | 0.34           | -1.98          | 1.07                       | 1.50          | 1.36          |
| PQL$^a$                | 0.87      | 0.28           | -1.69          | 0.95                       | 1.35          | 0.93          |

$^a$From Booth & Hobert (1999)

Table 4: Salamander mating data analysed jointly. Comparisons of the improved Laplace approximation with the Monte Carlo Expectation-Maximisation (MC-EM) of Booth & Hobert (1999), the Gibbs sampling of Karim & Zeger (1992), the quasi-likelihood (PQL) approach of Breslow & Clayton (1993) and the standard Laplace approximation.
likelihood approach of Breslow & Clayton (1993), the standard Laplace approximation and the posterior mean taken with the Gibbs sampling proposed by Karim & Zeger (1992) are illustrated in Table 4. We notice that the standard Laplace approximation underestimates the variance parameters (see also Shun, 1997). The estimate of $\beta$ and that of the variance parameters based on the proposed improved Laplace approximation are quite similar to those of the MC-EM procedure of Booth & Hobert (1999) (see also Sung & Geyer, 2007). Hence, the improved Laplace method gives approximate MLE that are quite similar to MC-EM, but compared to MC-EM the proposed method requires less intervention from the practitioner.

5 Final remarks

Laplace’s method for integrals is still widely used in both Bayesian and frequentist applications. However, when the sample size is small or when the dimension of the integral is high, the second-order Laplace approximation may deliver inaccurate results or it may even be asymptotically not valid as with the Salamander data. In this paper we proposed an improved Laplace approximation, and showed its superiority with respect to the standard Laplace method. Moreover, the three applications considered here indicated that our method performs comparably well with other existing gold standard methods, which are computationally more demanding.

An advantage of the proposed method is that the scalar integrals can be easily run in parallel. Moreover, it can approximate accurately also high-dimensional integrals, provided $h(\cdot)$ satisfies the required assumptions. The main computational burden of our method is given by the constrained optimisations and Hessian determinants at the constrained optimum.
values. Both tasks can be greatly enhanced by considering analytical first and second-order derivatives of \(-h(\cdot)\). This is the strategy used throughout the examples. An option to analytical differentiation is the automatic differentiation, which provides on-line function differentiation during its evaluation (see, e.g., Fournier et al., 2012).

The improved Laplace approximation requires the integrand to be concave and unimodal but not necessarily symmetric or with normal tails. From our experience, the standard Laplace approximation tends to work poorly when many variables of the integrand lay on the positive subset of \(\mathbb{R}\). Indeed, despite applying logarithmic reparametrisations, such variables may lead to an asymmetric or heavy-tailed integrand.

While in Bayesian applications the \(-h(\cdot)\) may not always be regular as required by the improved Laplace method, in GLMM the integrand often results in a unimodal function. In many instances, with independent random effects, the application of the Laplace approximation or quadrature rules with few quadrature points are accurate enough. Indeed standard GLMM can now be fitted quite accurately by available \texttt{R} packages such as \texttt{lme4}. However, in models with complicated, dependent and/or crossed random effects, Laplace method may perform poorly, and quadrature methods may require a large number of quadrature points, hence leading to a higher computational overhead. Our method seems more suited for these contexts, as was also shown by the example with the Salamander data.

**Supplementary Material**

The Supplementary Material contains necessary \texttt{R} codes and the \texttt{iLaplace} package, which can be used for reproducing the plots/analyses considered in the paper. In particular:
• to reproduce the plots in the Gompertz example, use the R code in the file `gompertz.R`;

• to reproduce the plots in the t/skew-t example use the R code in the file `mvtstud.R`;

• to reproduce the analysis with the BOD2 dataset use the R code in the file `bod2.R`;

• to reproduce the analysis with the Lubricant dataset use the R code in the file `lubricant.R`;

• to reproduce the analysis with the Salamander data use the R code in the file `salamander.R`.

Notice that the R code and the `iLaplace` package may require the installation of additional packages available on CRAN.

References

Barndorff-Nielsen, O. E. & Cox, D. R. (1994). Inference and Asymptotics. Boca Ranton, Florida: Chapman & Hall/CRC.

Bates, D. M. & Watts, D. G. (1988). Nonlinear Regression Analysis and Its Applications. New York: Wiley Online Library.

Bellio, R. & Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. Statistical Modelling 5, 217–227.

Bleistein, N. & Handelsman, R. (1986). Asymptotic Expansions of Integrals. New York: Dover.
Booth, J. G. & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B* 61, 265–285.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88, 9–25.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90, 1313–1321.

Chib, S. & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association* 96, 270–281.

Davison, A. C., Fraser, D. A. S. & Reid, N. (2006). Improved likelihood inference for discrete data. *Journal of the Royal Statistical Society: Series B* 68, 495–508.

DiCiccio, T. J., Kass, R. E., Raftery, A. & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association* 92, 903–915.

Evans, M. & Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science* 10, 254–272.

Evans, M. & Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press.

Fonseca, T. C., Ferreira, M. A. R. & Migon, H. S. (2008). Objective Bayesian analysis for the student-$t$ regression model. *Biometrika* 95, 325–333.
Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A. & Sibert, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software* **27**, 233–249.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian analysis* **1**, 515–534.

Guihenneuc-Jouyaux, C. & Rousseau, J. (2005). Laplace expansions in Markov chain Monte Carlo algorithms. *Journal of Computational and Graphical Statistics* **14**, 75–94.

Jones, M. C. (2002). Marginal replacement in multivariate densities, with application to skewing spherically symmetric distributions. *Journal of Multivariate Analysis* **81**, 85–99.

Karim, M. R. & Zeger, S. L. (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics* **48**, 631.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Lindley, D. V. (1980). Approximate Bayesian methods. *Trabajos de Estadistica Y de Investigacion Operativa* **31**, 223–245.

Martino, S., Akerkar, R. & Rue, H. (2011). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics* **38**, 514–528.

McClullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall, 2nd ed.
Nott, D. J., Fielding, M. & Leonte, D. (2009). On a generalization of the Laplace approximation. Statistics & Probability Letters 79, 1397–1403.

Pace, L., Salvan, A. & Ventura, L. (2006). Likelihood-based discrimination between separate scale and regression models. Journal of Statistical Planning and Inference 136, 3539 – 3553.

R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Raudenbush, S. W., Yang, M.-L. & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. Journal of Computational and Graphical Statistics 9, 141–157.

Rizopoulos, D., Verbeke, G. & Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. Journal of the Royal Statistical Society: Series B 71, 637–654.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B 71, 319–392.

Ruli, E., Sartori, N., Ventura, L. et al. (2014). Marginal posterior simulation via higher-order tail area approximations. Bayesian Analysis 9, 129–146.

Shun, Z. (1997). Another look at the salamander mating data: A modified Laplace approximation approach. Journal of the American Statistical Association 92, 341–349.
SHUN, Z. & MCCULLAGH, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B* **57**, 749–760.

SMALL, C. G. (2010). *Expansions and Asymptotics for Statistics*. Boca Ranton, Florida: Chapman & Hall/CRC.

SUNG, Y. J. & GEYER, C. J. (2007). Monte Carlo likelihood inference for missing data models. *The Annals of Statistics* **35**, 990–1011.

TIERNEY, L. & KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.