Journal of Machine Learning Research 25 (2024) 1-25 Submitted 11/21; Revised 2/24; Published 3/24

Low-rank Variational Bayes correction to the Laplace method

Janet van Niekerk janet.vanNiekerk@kaust.edu.sa
Statistics Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia

Håvard Rue haavard.rue@kaust.edu.sa
Statistics Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia

Editor: Debdeep Pati

Abstract

Approximate inference methods like the Laplace method, Laplace approximations and variational methods, amongst others, are popular methods when exact inference is not feasible due to the complexity of the model or the abundance of data. In this paper we propose a hybrid approximate method called Low-Rank Variational Bayes correction (VBC), that uses the Laplace method and subsequently a Variational Bayes correction in a lower dimension, to the joint posterior mean. The cost is essentially that of the Laplace method which ensures scalability of the method, in both model complexity and data size. Models with fixed and unknown hyperparameters are considered, for simulated and real examples, for small and large data sets.

Keywords: Information processing rule, INLA, Laplace, LGM, Variational Bayes

1. Introduction

Bayesian methods involve a prior belief about a model and learning from the data to arrive at a new belief, which is termed the posterior belief. Mathematically, the posterior belief can be derived from the prior belief and the empirical evidence presented by the data using Bayes' rule. In this way Bayesian analysis is a natural statistical machine learning method (see Theodoridis (2015); Chen et al. (2016); Polson and Sokolov (2017); Rehman et al. (2019); Sambasivan et al. (2020); Vehtari et al. (2020); Moss et al. (2021); Richardson and Weiss (2021) amongst many others), and especially powerful for small data sets, missing data or complex models. Suppose we observe data $\boldsymbol{y}$ for which we formulate a data generating model, $\mathcal{F}$. Suppose we have unknown parameters $\boldsymbol{\psi}$ in $\mathcal{F}$ for which we can define priors. Then we want to find the posterior inference of $\boldsymbol{\psi}$, denoted as $\pi(\boldsymbol{\psi}|\boldsymbol{y})$. The question of how to calculate $\pi(\boldsymbol{\psi}|\boldsymbol{y})$ now arises.

From a computational viewpoint, various approaches have been proposed to perform Bayesian analysis, mainly exact (analytical or sampling-based) or approximate inferential approaches. Sampling-based methods like Markov Chain Monte Carlo (MCMC) sampling with its extensions (see Metropolis et al. (1953); Geman and Geman (1984); Casella and

George (1992); Andrieu et al. (2003), amongst others) gained popularity in the 1990's but suffers from slow speed and convergence issues exacerbated by large data and/or complex models. Hamiltonian Monte Carlo (HMC) methods (Betancourt and Girolami, 2015), as implemented in the STAN software, are showing promise for more efficient sampling-based inference. Partly motivated by the inefficency of sampling-based methods, approximate methods were developed to *approximate* the posterior density as a more efficient means of inference. Not all approximate methods are equally accurate or efficient. Some approximate methods are essentially sampling-based like the Monte Carlo Adjusted Langevin Algorithm (MALA) (Rossky et al., 1978; Roberts and Tweedie, 1996), pseudo- marginal MCMC (Andrieu and Roberts, 2009) and Approximate Bayesian Computation (ABC) (Beaumont et al., 2002; Tavaré et al., 1997), and thus still slow. Asymptotic approximate methods are not sampling-based and propose a specific form of the posterior like the Laplace method (Van der Vaart, 2000; Laplace, 1986; Tierney et al., 1989) and the Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009; Bakka et al., 2018; Van Niekerk et al., 2021) for Latent Gaussian models. Optimization-based approximate methods like Variational Bayes (VB) (Attias, 1999; Jordan et al., 1999; Blei et al., 2017; Hoffman et al., 2013), Expectation Propagation (EP) (Opper and Winther, 2000; Minka, 2001; Dehaene and Barthelmé, 2016) and discrete distributions approximations by Liu and Wang (2016); Nitanda and Suzuki (2017) are also popular.

The Laplace method using a second-order series expansion around the mode (Tierney et al., 1989; Laplace, 1986), is a common approach to calculate an approximate Gaussian density to an unknown function. Its popularity can be assigned to its simplicity and low computational cost. It is mostly used to approximate marginal likelihoods, or marginal posteriors in a Bayesian framework. The Hessian of the unknown function, evaluated at the mode provides a quantification of the uncertainty, and under some regularity assumptions is a consistent estimator (Newey and McFadden, 1994). From the Laplace method we can thus approximate $\pi(\boldsymbol{\psi}|\boldsymbol{y})$ as

$$\tilde{\pi}(\boldsymbol{\psi}|\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\psi}-\boldsymbol{\mu})^\top \boldsymbol{Q}(\boldsymbol{\psi}-\boldsymbol{\mu})\right),$$

such that $-\boldsymbol{Q}$ is the Hessian matrix of $\log \pi(\boldsymbol{\psi}|\boldsymbol{y})$ evaluated at $\boldsymbol{\mu}$, the mode of $\log \pi(\boldsymbol{\psi}|\boldsymbol{y})$.

When the function is not uni-modal or exhibit heavy-tail behavior, the Laplace method does not provide an accurate approximation to the function and other families besides the Gaussian could be considered. The Gaussian assumption is often too strict for marginal posteriors and more flexible families should be considered that would allow for some skewness or heavier tails. Variational Bayes (VB) methods are based on the optimization of a certain objective function (variational energy resulting in the evidence lower bound) for a specific family of distributions. As such, any family can be considered and the Gaussian assumption of the marginal posteriors is not needed.

Suppose we posit that the approximate posterior of $\boldsymbol{\psi}$ comes from family $\mathcal{G}$, with members g, then the VB approximation of $\pi(\boldsymbol{\psi}|\boldsymbol{y})$ is $\tilde{\pi}(\boldsymbol{\psi}|\boldsymbol{y}) = g(\boldsymbol{\psi})$, such that

$$\tilde{\pi}(\boldsymbol{\psi}|\boldsymbol{y}) = \arg\min_{g\in\mathcal{G}} \text{KLD}(g(\boldsymbol{\psi})||\pi(\boldsymbol{\psi}|\boldsymbol{y})), \tag{1}$$

where $\text{KLD}(g||h)$ is the Kullback-Leibler divergence from probability distribution g to probability distribution h. Now since $\pi(\boldsymbol{\psi}|\boldsymbol{y})$ is unknown, it is shown that the minimizer is also

the maximizer of the evidence lower bound (ELBO), such that

$$\tilde{\pi}(\boldsymbol{\psi}|\boldsymbol{y}) = \arg\max_{g\in\mathcal{G}} \mathrm{E}_g(\log g(\boldsymbol{\psi}) - \log \pi(\boldsymbol{\psi}, \boldsymbol{y})). \tag{2}$$

For a specific choice of $\mathcal{G}$, specialized optimization techniques can be developed and applied. Some works for $\mathcal{G}$ being the Gaussian family are the Gaussian flow or Gaussian particle flow techniques (Galy-Fajou et al., 2021), Stein Variational Gradient Descent (Zhuo et al., 2018; Korba et al., 2020; Lu et al., 2019), recursive VGI (Lambert et al., 2020) and exactly sparse VGA (Barfoot et al., 2020), amongst others. If the selected family includes the true posterior, then the variational Bayes approximation could recover the true posterior. Variational frameworks, however, are known to suffer from severe underestimation of the uncertainty of the point estimate, due to the non-availability of a consistent estimator of the variance of the variational estimate for an often simple form of $\mathcal{G}$ (see for example the Appendix of Rue et al. (2009) for more details). This underestimation will produce poor credible intervals and could result in incorrect decision-making. Furthermore, the parameters of the chosen family in (2) should all be estimated and the optimization problem should be solved in the dimension of the parameter space, which can be very large in for example spatial models. Even if scalable (in some sense) approaches are proposed to optimize (2), all unknown parameters will have to be solved for.

We consider the case where a Gaussian approximation is opted for (not necessarily for the marginals). Our assumption is that the unknown density is uni-modal, and thus the Hessian matrix provides a reasonable estimate of the curvature at the mode. In this paper we present a novel approach to approximate an unknown density function with a Gaussian density function, that provides reasonable first and second order properties. We achieve this by employing the Laplace method, and then we formulate a low-rank variational correction to the mode of this Gaussian approximation. The variational correction to the Laplace method's mode, is defined in dimension p, that is much smaller than the latent field dimension m. This is possible since we learn the graph of connectedness from the Laplace method, and we use that to propagate any change in the lower dimension to all elements in the higher dimensional latent field.

Although our proposal can be used in various ways, we show the impact it has in the Bayesian inference of latent Gaussian models by applying the proposal *not* to the latent marginal posteriors, but to the latent *conditional posteriors*, since the latent conditional posteriors are in fact more Gaussian-like as shown by Rue et al. (2009). This provides an accurate and very efficient approximate Bayesian inference tool for latent Gaussian models that include generalized additive mixed models, spatial models, temporal models, lifetime analysis models and many more.

2. Proposal

Based on data $\boldsymbol{y}$ of size n, and unknown latent set $\boldsymbol{\psi} \in \mathbb{R}^m$, we formulate a data generating model, $\pi(\boldsymbol{y}|\boldsymbol{\psi})$ that depends on $\boldsymbol{\psi}$, such that the data is conditionally independent given $\boldsymbol{\psi}$. The goal is to infer $\boldsymbol{\psi}$ based on the data $\boldsymbol{y}$ and elective external information (prior information) $\pi(\boldsymbol{\psi})$. The joint density then is $\pi(\boldsymbol{\psi}, \boldsymbol{y})$. From this we can use Bayes' theorem

to formulate the posterior density of $\boldsymbol{\psi}$ as

$$\pi(\boldsymbol{\psi}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{\psi})\pi(\boldsymbol{\psi})}{\pi(\boldsymbol{y})}.$$

The Gaussian approximation of $\pi(\boldsymbol{\psi}|\boldsymbol{y})$ from the Laplace method is then derived from

$$\ln(\pi(\boldsymbol{\psi}|\boldsymbol{y})) = \ln(\pi(\boldsymbol{\psi}_0|\boldsymbol{y}) - \frac{1}{2}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)^\top \boldsymbol{H}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0) + \text{higher order terms},$$

where $\boldsymbol{\psi}_0$ is the mode of $\ln(\pi(\boldsymbol{\psi}|\boldsymbol{y}))$ and $\boldsymbol{H}$ is the negative Hessian matrix. Then

$$\tilde{\pi}(\boldsymbol{\psi}|\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)^\top \boldsymbol{H}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)\right), \tag{3}$$

so that $\boldsymbol{\psi}|\boldsymbol{y} \dot{\sim} N(\boldsymbol{\psi}_0, \boldsymbol{H}^{-1}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0})$ (approximately distributed as). To find the mode we solve for $\boldsymbol{\psi}_0$ in the system

$$\boldsymbol{H}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}\boldsymbol{\psi}_0 = \boldsymbol{\gamma}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} + \boldsymbol{H}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}\boldsymbol{\psi}_0, \tag{4}$$

where $\boldsymbol{\gamma}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}$ is the gradient of $\ln(\pi(\boldsymbol{\psi}|\boldsymbol{y}))$ evaluated at $\boldsymbol{\psi}=\boldsymbol{\psi}_0$. Now let $\boldsymbol{Q}_0 = \boldsymbol{H}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}$ and $\boldsymbol{b}_0 = \boldsymbol{\gamma}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} + \boldsymbol{H}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}\boldsymbol{\psi}_0$, then the system can be written as

$$\boldsymbol{Q}_0\boldsymbol{\psi}_0 = \boldsymbol{b}_0. \tag{5}$$

The precision matrix $\boldsymbol{Q}_0$, relates information about the conditional dependence amongst the elements in $\boldsymbol{\psi}$. Since the approximation in (3) is an approximation to the *joint* posterior, we still need to calculate the marginal posteriors. It is well-known that the marginal posteriors based on a joint Gaussian distribution can be computed as univariate Gaussian densities based on the elements of the joint mean and the diagonal elements of the inverse precision matrix, making the multivariate Gaussian assumption attractive.

We want to correct the mean of the Gaussian approximation to have a more accurate mean that is not necessarily the MAP (maximum a posteriori) estimator. As such we propose an updated mean,

$$\boldsymbol{\psi}_1 = \boldsymbol{\psi}_0 + \boldsymbol{\delta}, \tag{6}$$

where $\boldsymbol{\delta}$ can be viewed as corrections to the MAP estimator, such that the approximate posterior of $\boldsymbol{\psi}$ is then

$$\tilde{\pi}(\boldsymbol{\psi}|\boldsymbol{y}) = (2\pi)^{-m/2}|\boldsymbol{Q}_0|^{1/2}\exp\left(-\frac{1}{2}(\boldsymbol{\psi}-\boldsymbol{\psi}_1)^\top \boldsymbol{Q}_0(\boldsymbol{\psi}-\boldsymbol{\psi}_1)\right), \quad \boldsymbol{\psi}\in\mathbb{R}^m$$

where $\boldsymbol{\psi}_1 = \boldsymbol{\psi}_0 + \boldsymbol{\delta}$. Now the question arises: how can we estimate $\boldsymbol{\delta}$ in a fast and accurate way?

Since the dimension of $\boldsymbol{\psi}$ is m, we would need to find m values that produce a more accurate joint posterior mean. If the model is complex or contains many random effects, this dimension can be very large. For efficiency, we can use a variational framework since we only want to find a more accurate mean, while fixing the precision matrix based on the calculated Hessian. The ELBO for this problem is

$$\mathrm{E}_{\boldsymbol{\psi}\sim N(\boldsymbol{\psi}_0+\boldsymbol{\delta}, \boldsymbol{Q}_0^{-1})}(\log\phi(\boldsymbol{\psi}|\boldsymbol{\psi}_0+\boldsymbol{\delta}, \boldsymbol{Q}_0^{-1}) - \log\pi(\boldsymbol{\psi}, \boldsymbol{y})), \tag{7}$$

where $\phi(.)$ is the Gaussian density function. This optimization can be done in various ways using specialized techniques proposed in literature.

Rather than working with the ELBO, we revert back to the fundamental idea of Variational Bayes as introduced by Zellner (1988) (for more details see the Appendix) and more recently posed as an optimization view of Bayes' rule by Knoblauch et al. (2022). Based on the available information from the prior of the unknown parameters $\pi(\boldsymbol{\psi})$ and the conditional likelihood of the data $\pi(\boldsymbol{y}|\boldsymbol{\psi})$, we can derive two outputs: the marginal likelihood of the data $\pi(\boldsymbol{y})$ and the posterior of the unknown parameters $\pi(\boldsymbol{\psi}|\boldsymbol{y})$. If we want to use the input information optimally, then we find the approximate posterior $\tilde{\pi}(\boldsymbol{\psi}|\boldsymbol{y})$, such that

$$\tilde{\pi}(\boldsymbol{\psi}|\boldsymbol{y}) = \arg\min_{g\in\mathcal{G}} \left[E_{\boldsymbol{\psi}}[-\log \pi(\boldsymbol{y}|\boldsymbol{\psi})] + \mathrm{KLD}(g||\pi(\boldsymbol{\psi}))\right]. \tag{8}$$

In the work of Zellner (1988), it was shown that this variational framework produces the true posterior, from the appropriate family, as calculated from Bayes' theorem and thus implying that Bayes' theorem is an optimal rule for processing of information. Note that (8) does not contain the unknown true posterior $\pi(\boldsymbol{\psi}|\boldsymbol{y})$ as in (1), and can be directly optimized if the expected log-likelihood can be calculated in closed form. Thus to use (8) for the mean correction, we need to calculate

$$\tilde{\boldsymbol{\delta}} = \arg\min_{\boldsymbol{\delta}} \left[E_{\boldsymbol{\psi}\sim N(\boldsymbol{\psi}_0+\boldsymbol{\delta},\boldsymbol{Q}_0^{-1})}[-\log \pi(\boldsymbol{y}|\boldsymbol{\psi})] + \mathrm{KLD}\left(\phi(\boldsymbol{\psi}|\boldsymbol{\psi}_0+\boldsymbol{\delta},\boldsymbol{Q}_0^{-1})||\pi(\boldsymbol{\psi})\right)\right] \tag{9}$$

Whichever method is used to solve (9), the optimization is over an m-dimensional vector, thus the computational and memory cost will be based on m, which can be large.

2.1 Low-rank variational correction

Rather than an explicit correction to the MAP, we propose an implicit correction, by explicitly correcting the estimated gradient such that the improved posterior mean $\boldsymbol{\psi}_1$, satisfies the new system,

$$\boldsymbol{Q}_0\boldsymbol{\psi}_1 = \boldsymbol{b}_0 + \boldsymbol{\lambda} = \boldsymbol{b}_1. \tag{10}$$

Now, if $\boldsymbol{\lambda} \in \mathbb{R}^m$ then we would not gain any computational advantage over the proposal in (9), but because of the system, a change to any element in $\boldsymbol{b}_1$ will propagate changes to all the elements in $\boldsymbol{\psi}_1$. For a non-zero value of the j^{th} element of $\boldsymbol{\lambda}$ i.e. $\lambda_j \neq 0$, the change this value causes to the i^{th} element of $\boldsymbol{\psi}_1$, $\psi_{1,i}$ is

$$\frac{\partial \psi_{1,i}}{\partial \lambda_j}\lambda_j = \frac{\partial \psi_{1,i}}{\partial b_{1,j}}\frac{\partial b_{1,j}}{\partial \lambda_j}\lambda_j = Q_0^{ij}\lambda_j \tag{11}$$

where Q^{ij} denotes the element in the i^{th} row and j^{th} column of the inverse of $\boldsymbol{Q}$. Thus, in vector notation,

$$\frac{\partial \boldsymbol{\psi}_1}{\partial \lambda_j}\lambda_j = \boldsymbol{Q}_0^{\cdot j}\lambda_j \tag{12}$$

where $\boldsymbol{Q}^{\cdot j}$ denotes the j^{th} column of the inverse of $\boldsymbol{Q}$. This enables us to propose a low-rank Variational Bayes correction (VBC) since the dimension of $\boldsymbol{\lambda}$ is p, which can be much smaller than m and n, and p does not have to grow with m or n.

Suppose we have a set of indices $i \in I \subset \{1, 2, ..., m\}$ for which we want to correct $b_{0,i}$, then we extract the relevant columns of $\boldsymbol{Q}_0^{-1}$ and denote it by $\boldsymbol{Q}_I^{-1}$. The improved mean is thus

$$\boldsymbol{\psi}_1 = \boldsymbol{\psi}_0 + \boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}.$$

Now we can optimize (9), but for $\boldsymbol{\lambda}$ in dimension p instead of $\boldsymbol{\delta}$ in dimension m as follows

$$\tilde{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \left[E_{\boldsymbol{\psi}\sim N(\boldsymbol{\psi}_0 + \boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}, \boldsymbol{Q}_0^{-1})}[-\log \pi(\boldsymbol{y}|\boldsymbol{\psi})] + \text{KLD}\left(\phi(\boldsymbol{\psi}|\boldsymbol{\psi}_0 + \boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}, \boldsymbol{Q}_0^{-1}) || \pi(\boldsymbol{\psi})\right)\right]. \tag{13}$$

This proposal allows us to correct an m-dimensional MAP estimator with a rank p update with $p \ll m$, resulting in a computational cost of about $O(mp^2)$, since we do not need to calculate the entire inverse of $\boldsymbol{Q}_0$ but only the selected elements based on I. Moreover, from Zellner (1988), this optimization is optimally information efficient and converges to the true posterior when the true family is selected. We illustrate this convergence using simulated and real examples, and we compare the posterior from a Gaussian approximation with the VB correction, to the posterior from MCMC samples in Section 3.

Our proposal to approximate the joint posterior can be summarized as follows:

1. Calculate the gradient $\boldsymbol{\gamma}$, and the negative Hessian matrix $\boldsymbol{H}$, of $\log \pi(\boldsymbol{\psi}|\boldsymbol{y})$.

2. Find the MAP estimator by solving for $\boldsymbol{\psi}_0$ such that

   $$\boldsymbol{H}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}\boldsymbol{\psi}_0 = \boldsymbol{\gamma}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} + \boldsymbol{H}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}\boldsymbol{\psi}_0,$$

   and define $\boldsymbol{Q}_0 = \boldsymbol{H}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}$ and $\boldsymbol{b}_0 = \boldsymbol{\gamma}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} + \boldsymbol{H}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}\boldsymbol{\psi}_0$.

3. Decide on the set of indices for correction, I, construct the $p \times m$ matrix $\boldsymbol{Q}_I^{-1}$ from the columns of the inverse of $\boldsymbol{Q}_0$, $\boldsymbol{Q}_0^{-1}$, and solve for $\boldsymbol{\lambda}$ such that

   $$\tilde{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \left[E_{\boldsymbol{\psi}\sim N(\boldsymbol{\psi}_0 + \boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}, \boldsymbol{Q}_0^{-1})}[-\log \pi(\boldsymbol{y}|\boldsymbol{\psi})] + \text{KLD}\left(\phi(\boldsymbol{\psi}|\boldsymbol{\psi}_0 + \boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}, \boldsymbol{Q}_0^{-1}) || \pi(\boldsymbol{\psi})\right)\right].$$

4. The approximate posterior of $\boldsymbol{\psi}$ is Gaussian with mean $\boldsymbol{\psi}_1 = \boldsymbol{\psi}_0 + \boldsymbol{Q}_I^{-1}\tilde{\boldsymbol{\lambda}}$ and precision matrix $\boldsymbol{Q}_0$.

Now we consider the choice of the index set I. Since a change in any one element of $\boldsymbol{b}_1$ is propagated to the posterior mean of the entire latent field, similar choices of the index set I, will result in a similar improved joint posterior mean, since the proposal is based on an improved joint Gaussian approximation for the entire field. We are thus not solving for element-wise corrections, and from the work of Zellner (1988) we are assured of a joint improvement. From our experience, we want to mainly correct the Gaussian approximation for those elements in $\boldsymbol{\psi}$ that are most influential and connected to many datapoints. Hence, we explicitly correct the posterior means of the fixed effects, and those random effects that are connected to many datapoints (short length random effects). We return to this in Sections 3, 4.3 and 5.

Even though the proposal looks basic and simple, various computational details are intricate and complicated to ensure a low computational cost while maintaining accuracy. Some of these details are presented in the next section.

2.2 Computational aspects

In this section we focus on computational aspects regarding the proposed variation Bayes correction to the Laplace method. The gradient and Hessian matrix, can be calculated numerically and we can use various gradient descent or Newton-Raphson type algorithms. In our approach we use the smart gradient proposed by Fattah et al. (2022). The efficient calculation of the expected log-likelihood in (13) requires some attention and we present our approach in this section.

2.2.1 Smart gradient

Numerical gradients are important in various optimization techniques (as in our proposal) such as stochastic gradient descent, trust region and Newton-type methods, to name a few. The smart gradient approach can be used to calculate the gradient (and Hessian) numerically, more accurately by using previous descent directions and a transformed coordinate basis. Instead of using the canonical basis at each step, a new orthonormal basis is constructed based on the previous direction using the Modified Gram-Schmidt orthogonalization (see for example Picheny et al. (2013)). This transformed basis results in more accurate numeric gradients, which could lead to finding optimums more accurately and more efficiently. For more details see Fattah et al. (2022).

2.2.2 Expected log-likelihood

For some likelihoods (as in Section 3), the expectation can be calculated analytically, but for others we have to numerically approximate this expectation. Note that as previously stated, the data is assumed conditionally independent given the latent set $\boldsymbol{\psi}$, and hence the log-likelihood can be constructed by a simple sum of the log-likelihoods from each datapoint. The expected log-likelihood of each datapoint can then be approximated using Gauss-Hermite quadrature, since the integral is with respect to a Gaussian kernel. The expected log-likelihood with respect to the approximate posterior of $\boldsymbol{\psi}$ is,

$$E_{\boldsymbol{\psi}\sim N(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda},\boldsymbol{Q}_0^{-1})}\left[-\log\pi(\boldsymbol{y}|\boldsymbol{\psi})\right]=\int_{\mathbb{R}^m}-\sum_{i=1}^{n}\log\pi(y_i|\boldsymbol{\psi})\phi(\boldsymbol{\psi}|\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda},\boldsymbol{Q}_0^{-1})d\boldsymbol{\psi}. \quad (14)$$

Now if we consider a generalized linear model, with the design matrix $\boldsymbol{A}$ that links the data to the parameter $\boldsymbol{\psi}$, then the linear predictors can be calculated as

$$\boldsymbol{\eta}=\boldsymbol{A}\boldsymbol{\psi}, \quad (15)$$

such that the posterior mean for the i^{th} linear predictor for y_i, η_i is $\boldsymbol{A}_{i\boldsymbol{\cdot}}(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda})$, where $\boldsymbol{A}_{i\boldsymbol{\cdot}}$ is the i^{th} row of $\boldsymbol{A}$. Then,

$$E_{\boldsymbol{\psi}\sim N(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda},\boldsymbol{Q}_0^{-1})}[-\log\pi(\boldsymbol{y}|\boldsymbol{\psi})]=E_{\boldsymbol{\eta}\sim N(\boldsymbol{A}(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}),\boldsymbol{A}\boldsymbol{Q}_0^{-1}\boldsymbol{A}^\top)}[-\log\pi(\boldsymbol{y}|\boldsymbol{\eta})], \quad (16)$$

since $\pi(\boldsymbol{y}|\boldsymbol{\psi})$ only depends on ψ through η. Since the data are conditionally independent,

$$\begin{aligned} E_{\boldsymbol{\eta}\sim N(\boldsymbol{A}(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}),\boldsymbol{A}\boldsymbol{Q}_0^{-1}\boldsymbol{A}^\top)}\left[-\log\pi(\boldsymbol{y}|\boldsymbol{\eta})\right] &= E_{\boldsymbol{\eta}\sim N(\boldsymbol{A}(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}),\boldsymbol{A}\boldsymbol{Q}_0^{-1}\boldsymbol{A}^\top)}\left[-\sum_{i=1}^n \log\pi(y_i|\boldsymbol{\eta})\right] \\ &= E_{\boldsymbol{\eta}\sim N(\boldsymbol{A}(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}),\boldsymbol{A}\boldsymbol{Q}_0^{-1}\boldsymbol{A}^\top)}\left[-\sum_{i=1}^n \log\pi(y_i|\eta_i)\right] \\ &= -\sum_{i=1}^n E_{\eta_i}[\log\pi(y_i|\eta_i)]. \end{aligned}$$

The univariate expectations are calculate using Gauss-Hermite quadrature with m_g weights $\boldsymbol{w}$, and roots $\boldsymbol{x}^w$, such that,

$$E_{\boldsymbol{\psi}\sim N(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda},\boldsymbol{Q}_0^{-1})}[-\log\pi(\boldsymbol{y}|\boldsymbol{\psi})] \approx \frac{-1}{\sqrt{\pi}}\sum_{r=1}^{m_g}\left[w_r\sum_{i=1}^n \log\pi\left(y_i|\eta_i(x_r^w)\right)\right]. \tag{17}$$

To optimize (13) numerically, we expand (17) around $\boldsymbol{\lambda}=\boldsymbol{0}$ using a second order Taylor series expansion such that,

$$E_{\boldsymbol{\psi}\sim N(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda},\boldsymbol{Q}_0^{-1})}[-\log\pi(\boldsymbol{y}|\boldsymbol{\psi})] \approx \text{constant} + \boldsymbol{B}^\top\boldsymbol{A}\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda} + \frac{1}{2}(\boldsymbol{A}\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda})^\top\text{diag}(\boldsymbol{C})\boldsymbol{A}\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}, \tag{18}$$

where the i^{th} entries of $\boldsymbol{B}$ and $\boldsymbol{C}$, respectively, are,

$$B_i = \sum_{r=1}^{m_g}\frac{w_r x_r^w}{S_i}\log\pi\left(y_i|\eta_i = x_r^w S_i + \boldsymbol{A}_i.\boldsymbol{\psi}_0\right)$$

and

$$C_i = \sum_{r=1}^{m_g}\frac{w_r[(x_r^w)^2-1]}{S_i^2}\log\pi\left(y_i|\eta_i = x_r^w S_i + \boldsymbol{A}_i.\boldsymbol{\psi}_0\right),$$

with $S_i = \sqrt{\left(\boldsymbol{A}^\top\boldsymbol{Q}_0^{-1}\boldsymbol{A}\right)_{ii}}$. Thus, for a generalized linear model the expected log-likelihood can be calculated in closed form.

3. Illustrative example - low-count Poisson regression model

Here we provide the details for a generalized linear model for count data, where we use a Poisson response model.

Suppose we have data $\boldsymbol{y}$, of size n with covariates $\boldsymbol{X}$ and random effect covariates $\boldsymbol{u}$, then

$$\begin{aligned} Y_i|\beta_0,\boldsymbol{\beta},\boldsymbol{f} &\sim \text{Poisson}(\exp(\eta_i)) \\ \eta_i &= \beta_0 + \boldsymbol{X}_i\boldsymbol{\beta} + \sum_{k=1}^K f^k(\boldsymbol{u}_k). \end{aligned}$$

3.1 Expected log-likelihood

For the Poisson likelihood, we can obtain a closed-form expression of the expected log-likelihood and no numerical integration is required. Note that

$$\begin{aligned} E_{\boldsymbol{\psi}|\boldsymbol{\theta}\sim N(\boldsymbol{\psi}_1,\boldsymbol{Q}_0^{-1})}\left[-\log\pi(\boldsymbol{\psi}|\boldsymbol{y})\right] &= \int_{\mathbb{R}^m} -\log\pi(\boldsymbol{\psi}|\boldsymbol{y})\phi(\boldsymbol{\psi}|\boldsymbol{\psi}_1,\boldsymbol{Q}_0^{-1})d\boldsymbol{\psi} \\ &= \int_{\mathbb{R}^m}\sum_{i=1}^{n}\left(\exp(\boldsymbol{A}_{i\cdot}\boldsymbol{\psi}) - \boldsymbol{A}_{i\cdot}\boldsymbol{\psi}y_i + \log(y_i!)\right)\phi(\boldsymbol{\psi}|\boldsymbol{\psi}_1,\boldsymbol{Q}_0^{-1})d\boldsymbol{\psi} \\ &= \sum_{i=1}^{n}\left(\exp\left(\boldsymbol{A}_{i\cdot}\boldsymbol{\psi}_1 + \frac{(\boldsymbol{A}^\top\boldsymbol{Q}_0^{-1}\boldsymbol{A})_{ii}}{2}\right) - y_i(\boldsymbol{A}_{i\cdot}\boldsymbol{\psi}_1) + \log(y_i!)\right). \end{aligned}$$

Now, from (13), we find $\boldsymbol{\lambda}$, where

$$\begin{aligned} \tilde{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \quad & \left[\sum_{i=1}^{n}\left(\exp\left(\boldsymbol{A}_{i\cdot}(\boldsymbol{\psi}_0 + \boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}) + \frac{(\boldsymbol{A}^\top\boldsymbol{Q}_0^{-1}\boldsymbol{A})_{ii}}{2}\right) - y_i(\boldsymbol{A}_{i\cdot}(\boldsymbol{\psi}_0 + \boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}))\right)\right. \\ & \left.+\mathrm{KLD}\left(\phi(\boldsymbol{\psi}|\boldsymbol{\psi}_0 + \boldsymbol{Q}_I^{-1}\boldsymbol{\lambda}, \boldsymbol{Q}_0^{-1})||\pi(\boldsymbol{\psi})\right)\right]. \end{aligned}$$

3.2 Simulation results

In this section we present an example of the proposed method. We focus on count data with low counts since this is usually a challenging case because the likelihood is maximized at $-\infty$, and the second-order expansion of the log-likelihood is less accurate. We use MCMC, a Gibbs sampler (using the *runjags* library) and HMC (using Stan) with a burn-in of 10^2 and a sample of size 10^5, as the gold standard, and compare VBC to the Laplace method in terms of computational efficiency and accuracy.

We consider the following over-dispersed count model defined as follows for a data set of size n:

$$y_i \sim \text{Poisson}(\exp(\eta_i)), \quad \eta_i = \beta_0 + \beta_1 x_i + u_i, \tag{19}$$

with a sum-to-zero constraint on $\boldsymbol{u}$, to ensure identifiability of β_0. We use $\beta_0 = -1, \beta_1 = -0.5$ and a continuous covariate x, simulated as $x \sim N(0,1)$. The overdispersion is simulated as $u_i \sim N(0, 0.25)$. We design the study with the intent of having mostly low counts. We want to perform full Bayesian inference for the latent field $\boldsymbol{\psi} = \{\beta_0, \beta_1, \boldsymbol{u}\}$, and the linear predictors $\boldsymbol{\eta} = \{\eta_1, \eta_2, ..., \eta_n\}$. We assume the following illustrative priors,

$$\beta_0 \sim t(5), \quad \beta_1 \sim U(-3,3) \quad \text{and} \quad \boldsymbol{u} \sim N(\boldsymbol{0}, 0.25\boldsymbol{I})$$

i.e. β_0 follows a Student's t prior with 5 degrees of freedom, β_1 follows a uniform distribution in $(-3,3)$ and the random effects are independent and identically distributed with a fixed marginal precision of 4. The vector of estimable parameters is thus $\boldsymbol{\psi} = \{\beta_0, \beta_1, u_1, u_2, ..., u_n\}$ of dimension $n+2$.

To illustrate the effect of the low-rank correction we apply the VBC only to β_0, β_1 and then propagate the induced corrections to the n random intercepts, $\boldsymbol{u}$ and the n linear predictors, $\boldsymbol{\eta}$. We thus perform a two-dimensional optimization instead of an $(n+2)$-dimensional optimization, as would be necessary with other variational Bayes approaches.

We simulate two samples from the proposed model, one of size $n = 20$ and another of size $n = 100$, and the data are presented in Figure 1 (left). The posterior means for the Laplace method, MCMC, HMC and the VBC methods are presented in Table 1 for the latent field and selected linear predictors. We can clearly see the improved accuracy in the mean of the VBC to the Laplace method when compared with the MCMC and HMC output, from Table 1 and Figure 1 (center and right), especially for the smaller data set. In the case of a larger data set, we note that the Gaussian approximation performs well and the VBC applies only a slight correction. With the VBC we can achieve similar posterior inference for the latent field and linear predictors, to the MCMC and HMC approaches, more efficiently. Note that for a small data set the computational time is small for all the methods, as expected due to the small dimension of the latent field $\boldsymbol{\psi} = \{\beta_0, \beta_1, u_1, u_2, ..., u_{20}\}$ and linear predictors $\boldsymbol{\eta} = \{\eta_1, \eta_2, ..., \eta_{20}\}$, although for a larger data set like $n = 100$ the excessive computational time for MCMC and HMC is clear, even with this simple model, because the parameter space is of dimension 102. With the VBC, the correction space is of dimension 2 and thus the cost of VBC compared to any other inferential framework based on the entire parameter space of dimension 102, will be much less. The time comparison can be misleading since neither the VBC, MCMC or HMC code has been optimized for this specific model and priors. Nonetheless, the VBC scales well with increasing data size as shown in this example, since the size of the correction space for the optimization stays 2, while the parameter space size grows from 22 to 102. This fictitious example illustrates the stability and scalability of the VBC, for a fixed hyperparameter, which is an unrealistic scenario. In the next Section we provide a hybrid method using the VBC and the integrated nested Laplace approximation (INLA) methodology to address Bayesian inference for more realistic models, including those with hyperparameters, in order to perform Bayesian inference for the latent field and hyperparameters simultaneously.

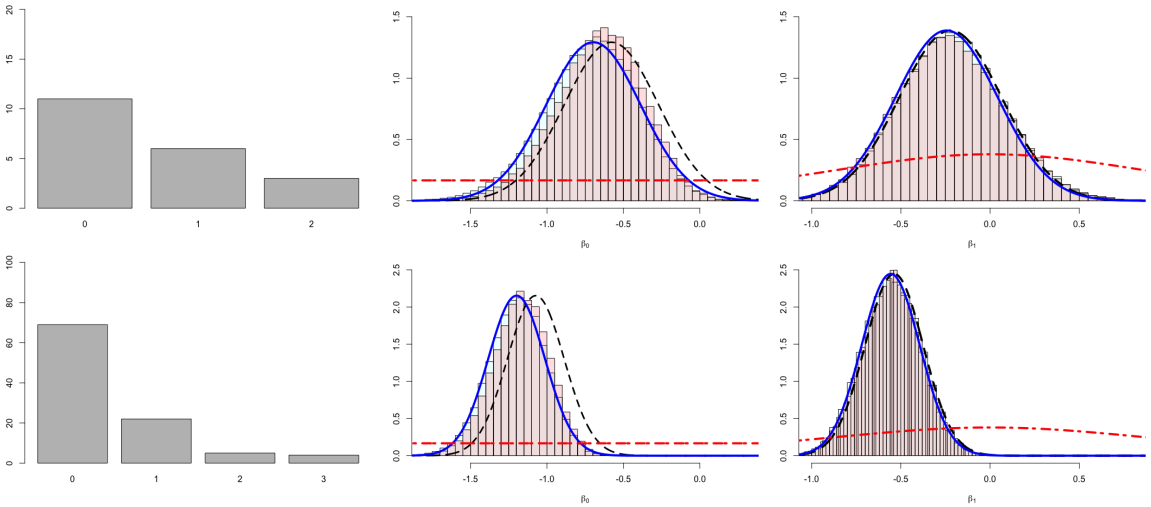

Figure 1: Poisson counts simulated from (19) (left) and the marginal posterior of β_0 (center) and β_1 (right) from MCMC (blue histogram), HMC (red histogram), the Laplace method (dashed line) and VBC (solid line) based on the prior (broken line) for $n = 20$ (top) and $n = 100$ (bottom)

	n=20				n=100			
	LM	VBC	MCMC	HMC	LM	VBC	MCMC	HMC
β_0	-0.579	-0.706	-0.715	-0.657	-1.073	-1.199	-1.196	-1.180
β_1	-0.222	-0.224	-0.218	-0.226	-0.538	-0.567	-0.552	-0.552
u_1	-0.158	-0.148	-0.152	-0.159	0.177	0.174	0.175	0.174
u_8	0.098	0.098	0.099	0.099	-0.046	-0.052	-0.049	-0.044
u_{15}	0.122	0.115	0.118	0.121	-0.074	-0.079	-0.077	-0.073
Time(s)	2.21	5.78	22.537	12.438	9.48	17.36	384.12	169.57

Table 1: Posterior means from the Laplace method, VBC, MCMC and HMC

4. Application to latent Gaussian models

The proposal in Section 2 can be used to accurately and efficiently calculate the joint posterior for the latent field, only. Many problems, however, include hyperparameters and as such we can embed our proposal into another framework to perform full Bayesian inference for the latent field *and* the hyperparameter set. We use the INLA methodology as proposed by Rue et al. (2009) and propose an INLA-VBC methodology. Various strategies, with varying accuracy and efficiency, are available in the INLA framework. The most accurate strategy is the Laplace strategy which involves nested use of the Laplace method, while the least accurate strategy is the Gaussian strategy where the Laplace method is used only once. Naturally, the Gaussian strategy is most efficient while the Laplace strategy is least efficient. Details of these two strategies are presented in the next section. We aim to achieve accuracy similar to that of the Laplace strategy, with a similar cost than that of the Gaussian strategy. We focus our attention to latent Gaussian models (LGMs) for which INLA is developed and show how our proposal can be used.

A latent Gaussian model appears naturally in statistical modeling since it is a hierarchical Bayesian model with a Gaussian Markov random field as the latent prior. We define an LGM based on data $\boldsymbol{y}$ of size n, a latent field $\boldsymbol{\psi}$ of size m and n linear predictors

$$\boldsymbol{\eta} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{k=1}^{K} f^k(\boldsymbol{u}_k), \tag{20}$$

such that $\{\boldsymbol{f}\}$ are unknown functions (random effects) of $\boldsymbol{u}$, and $\boldsymbol{\beta}$ contains the coefficients for the linear effects of $\boldsymbol{X}$ on $\boldsymbol{\eta}$. The latent field is defined as $\boldsymbol{\psi} = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{f}\}$. Often, hyperparameters either from the likelihood or the prior of the latent field, form part of the LGM and we denote these by $\boldsymbol{\theta}$ and assume a moderate dimension q (usually $q < 30$). In an LGM the latent field is assumed to follow a Gaussian prior with a sparse precision matrix. The sparseness is often satisfied as most generalized additive mixed models exhibit a sparse precision matrix by construction. Thus an LGM can be summarized as follows:

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\psi}, \boldsymbol{\theta}_1 &\sim \prod_{i=1}^{n} \pi(y_i|\boldsymbol{\psi}, \boldsymbol{\theta}_1) \\ \boldsymbol{\psi}|\boldsymbol{\theta}_2 &\sim N(\mathbf{0}, \boldsymbol{Q}_\pi^{-1}(\boldsymbol{\theta}_2)) \\ \boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\} &\sim \pi(\boldsymbol{\theta}). \end{aligned} \tag{21}$$

The aim is thus to estimate the latent posteriors $\pi(\psi_j|\boldsymbol{y}), j = 1, 2, ..., m$, and the hyperparameter posteriors $\pi(\theta_k|\boldsymbol{y}), k = 1, 2, ..., q$.

A specialized methodology called the integrated nested Laplace approximation (INLA) was introduced by Rue et al. (2009), to accurately and efficiently approximate the marginal posteriors of $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ for an LGM. This methodology is based on a series of Gaussian approximations to *conditional* posteriors, using the Laplace method. There is no parametric assumption on the form of the marginal posteriors.

4.1 INLA

The INLA methodology from Rue et al. (2009) can be summarized as follows,

$$\begin{aligned}
\pi(\boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{y}) &= \pi(\boldsymbol{\theta})\pi(\boldsymbol{\psi}|\boldsymbol{\theta}_2) \prod_{i=1}^{n} \pi(y_i|\boldsymbol{\psi}, \boldsymbol{\theta}_1) \\
\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) &\propto \left. \frac{\pi(\boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{y})}{\pi_{\text{LM}}(\boldsymbol{\psi}|\boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0(\boldsymbol{\theta})} \\
\tilde{\pi}(\theta_j|\boldsymbol{y}) &= \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-j} \\
\tilde{\pi}(\psi_j|\boldsymbol{y}) &= \int \tilde{\pi}(\psi_j|\boldsymbol{\theta}, \boldsymbol{y}) \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}, \qquad (22)
\end{aligned}$$

where $\pi_{\text{LM}}(\boldsymbol{\psi}|\boldsymbol{\theta}, \boldsymbol{y})$ is the approximation based on the Laplace method at the mode $\boldsymbol{\psi}_0(\boldsymbol{\theta})$ with precision matrix $\boldsymbol{Q}_0(\boldsymbol{\theta})$ from (5), and $\tilde{f}(.)$ denotes an approximation to $f(.)$. Note that the Laplace method is used for the approximation based on a fixed $\boldsymbol{\theta}$. For convenience we will use $\boldsymbol{\psi}_0$ and $\boldsymbol{Q}_0$, to denote $\boldsymbol{\psi}_0(\boldsymbol{\theta})$ and $\boldsymbol{Q}_0(\boldsymbol{\theta})$, respectively.

The approximate conditional posterior of ψ_j, $\tilde{\pi}(\psi_j|\boldsymbol{\theta}, \boldsymbol{y})$, can be calculated in one of two ways, extracted from $\pi_{\text{LM}}(\boldsymbol{\psi}|\boldsymbol{\theta}, \boldsymbol{y})$ (Gaussian strategy) as

$$\tilde{\pi}(\psi_j|\boldsymbol{\theta}, \boldsymbol{y}) \approx \phi(\psi_j|\psi_{0j}, Q_0^{jj}), \qquad (23)$$

or subsequent Gaussian approximations (Laplace strategy) as follows,

$$\tilde{\pi}(\psi_j|\boldsymbol{\theta}, \boldsymbol{y}) \propto \left. \frac{\tilde{\pi}(\boldsymbol{\psi}, \boldsymbol{\theta}|\boldsymbol{y})}{\pi_{\text{LM}}(\boldsymbol{\psi}_{-j}|\psi_j, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{\psi}_{-j}=\boldsymbol{\mu}_{-j}(\boldsymbol{\theta})}, \qquad (24)$$

where $\boldsymbol{\mu}_{-j}(\boldsymbol{\theta})$ is the mode from the Gaussian approximation to $\pi(\boldsymbol{\psi}_{-j}|\psi_j, \boldsymbol{\theta}, \boldsymbol{y})$ based on the Laplace method, and $\boldsymbol{\psi}_{-j}$ is $\boldsymbol{\psi}$ without the j^{th} element. As the dimension of the latent field grow it is clear that the Laplace strategy will become costly due to the multiple Gaussian approximations. It was shown by Rue et al. (2009) and multiple works there after that the posteriors from INLA using the Laplace strategy is accurate when compared with those obtained from MCMC sampling, while being much more time and memory efficient than MCMC sampling, even for a large hyperparameter set due to parallel integration strategies (Gaedke-Merzhäuser et al., 2023). The Gaussian strategy is more efficient, but the resulting marginal posteriors are not accurate enough for some cases. Our proposal in Section 2 thus fits naturally into this framework, where we can find a more accurate Gaussian approximation based on the Laplace method and the VBC, by correcting the mode of (23).

4.2 INLA-VBC

The INLA methodology provides a deterministic framework to approximate the posteriors of the hyperparameters as well as the latent field elements. We apply the proposal from Section 2 to the INLA methodology with the hope of achieving more efficient yet accurate approximations of the latent posteriors. Conditional on the hyperparameters, $\boldsymbol{\theta}$, define the corrected posterior mean of the joint conditional as $\boldsymbol{\psi}_1 = \boldsymbol{\psi}_0 + \boldsymbol{\delta}$, where we calculate $\boldsymbol{\delta}$ implicitly from the correction to $\boldsymbol{b}_0$ such that $\boldsymbol{b}_1 = \boldsymbol{b}_0 + \boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is non-zero, only for those elements in I, the set of p indices to which we formulate the explicit correction. Note that the latent prior $\pi(\boldsymbol{\psi}|\boldsymbol{\theta}_2)$ is Gaussian by construction of the LGM as in (21), so the KLD term simplifies to the KLD between two multivariate Gaussian densities. Then from (13) and (18), we solve for $\boldsymbol{\lambda}$ (conditionally on $\boldsymbol{\theta}$) as

$$\begin{aligned}\tilde{\boldsymbol{\lambda}} &= \arg\min_{\boldsymbol{\lambda}} \left[E_{\boldsymbol{\psi}|\boldsymbol{\theta}\sim N(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda},\boldsymbol{Q}_0^{-1})}[-\log\pi(\boldsymbol{y}|\boldsymbol{\psi})] + \mathrm{KLD}\left(\phi(\boldsymbol{\psi}|\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda},\boldsymbol{Q}_0^{-1})||\phi(\boldsymbol{\psi}|\boldsymbol{0},\boldsymbol{Q}_\pi)\right)\right] \\ &= \arg\min_{\boldsymbol{\lambda}} \left[E_{\boldsymbol{\psi}|\boldsymbol{\theta}\sim N(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda},\boldsymbol{Q}_0^{-1})}[-\log\pi(\boldsymbol{y}|\boldsymbol{\psi})] + \frac{1}{2}(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda})^\top \boldsymbol{Q}_\pi(\boldsymbol{\psi}_0+\boldsymbol{Q}_I^{-1}\boldsymbol{\lambda})\right], \qquad (25)\end{aligned}$$

where $\boldsymbol{Q}_I^{-1}$ is constructed from specific columns of $\boldsymbol{Q}_0^{-1}$. Thus the improved Gaussian approximation to $\pi(\boldsymbol{\psi}|\boldsymbol{\theta},\boldsymbol{y})$ is

$$\boldsymbol{\psi}|\boldsymbol{\theta},\boldsymbol{y} \sim N(\boldsymbol{\psi}_1, \boldsymbol{Q}_0^{-1}). \qquad (26)$$

Now we can use this improved Gaussian approximation to the *conditional* joint posterior, to extract the conditional posteriors for the latent field elements as

$$\tilde{\pi}(\psi_j|\boldsymbol{\theta},\boldsymbol{y}) \approx \phi(\psi_j|\psi_{1j}, Q_0^{jj}), \qquad (27)$$

instead of the more cumbersome series of Gaussian approximations as in (24). Finally, using the INLA methodology (22), the marginal posteriors of the latent field elements can be calculated as

$$\tilde{\pi}(\psi_j|\boldsymbol{y}) = \sum_{k=1}^{K} \tilde{\pi}(\psi_j|\boldsymbol{\theta},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{y})\Delta_k,$$

where $\{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,...,\boldsymbol{\theta}_K\}$ is a set of values calculated from the joint posterior $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ using a central composite design (CCD) (Box and Wilson, 1951), and Δ_k is the step size. Thus, the proposed INLA-VBC methodology can be summarized as

$$\begin{aligned}\pi(\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{y}) &= \pi(\boldsymbol{\theta})\pi(\boldsymbol{\psi}|\boldsymbol{\theta}_2)\prod_{i=1}^{n}\pi(y_i|\boldsymbol{\psi},\boldsymbol{\theta}_1) \\ \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) &\propto \left.\frac{\pi(\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{y})}{\pi_{\mathrm{LM}}(\boldsymbol{\psi}|\boldsymbol{\theta},\boldsymbol{y})}\right|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} \\ \tilde{\pi}(\theta_j|\boldsymbol{y}) &= \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j} \\ \tilde{\pi}(\psi_j|\boldsymbol{y}) &= \int \tilde{\pi}_{\mathrm{VBC}}(\psi_j|\boldsymbol{\theta},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \qquad (28)\end{aligned}$$

where $\tilde{\pi}_{\mathrm{VBC}}(\psi_j|\boldsymbol{\theta},\boldsymbol{y})$ is the VB corrected Gaussian approximation from (27). Next we show how accurate and efficient this proposal is for approximate Bayesian inference of latent Gaussian models, based on a simulated sample from an overdispersed Poisson model.

4.3 Simulation results

We use an overdispersed Poisson regression model with Gaussian priors for the latent field such that

$$y_i \sim \text{Poisson}(\exp(\eta_i)), \quad \eta_i = \beta_0 + \beta_1 x_i + u_i, \tag{29}$$

for $i = 1, 2, ..., n$, where $u_i|\tau \sim N(0, \tau^{-1})$, $\log \tau \sim \text{loggamma}(1, 5 \times 10^{-5})$, $\beta_0 \sim N(0, 1)$ and $\beta_1 \sim N(0, 1)$. The data is simulated based on $\beta_0 = -1, \beta_1 = -0.5, \tau = 1$ and a continuous covariate x, simulated as $x \sim N(0, 1)$. We want to perform Bayesian inference for the latent field $\boldsymbol{\psi} = \{\beta_0, \beta_1, u_1, u_2, ..., u_n\}$, the linear predictors $\{\eta_1, \eta_2, ..., \eta_n\}$ and the set of hyperparameters $\boldsymbol{\theta} = \{\tau\}$.

We simulate a sample of $n = 1000$ counts and the data is presented in Figure 2 (left). In this case we apply the VBC only to the fixed effects β_0 and β_1, and the associated changes are then propagated to the posterior means of $\boldsymbol{u}$ and the linear predictors $\boldsymbol{\eta}$, we thus have a two-dimensional optimization instead of a 1002-dimensional optimization as with other variational Bayes approaches, conditional on the hyperparameter τ.

The posterior means for the Gaussian strategy (GA), Laplace strategy (INLA), MCMC and the INLA-VBC methods are presented in Table 2. We can clearly see the improved accuracy of the INLA-VBC to the Gaussian strategy when compared with the MCMC output, from Table 2 and Figure 2 (center and right), without much additional computational cost based on the time. With the INLA-VBC we can achieve similar results to that of the MCMC approach, more efficiently, while inferring the hyperparameters as well. For the MCMC we used a Gibbs sampler with a burn-in of 10^3 and a sample of size 10^5.

	GA	INLA	INLA-VBC	MCMC
β_0	-0.972	-0.664	-0.972	-0.934
β_1	-0.484	-0.532	-0.531	-0.529
τ	1.056	1.056	1.056	1.037
Time(s)	5.067	18.299	5.718	207.445

Table 2: Posterior means from the Gaussian strategy (GA), Laplace strategy (INLA), INLA-VBC and MCMC

5. Real data examples

We consider two real data examples of different sized data sets. The first example includes a stochastic spline model while the second example is a time to event model based on a continuously-indexed spatial field. Both these models are latent Gaussian models and both involve hyperparameters. We thus use the proposed INLA-VBC methodology from Section 4.2 for full Bayesian inference of these models. Due to the complexity of these models we only compare the results with that of MCMC for the small scale example. Instead, we compare the results based on the Gaussian strategy and the Laplace strategy within the INLA framework of Rue et al. (2009), with those of the INLA-VBC proposal. These examples illustrate the gain in accuracy, without an increased computational cost when using INLA-VBC.

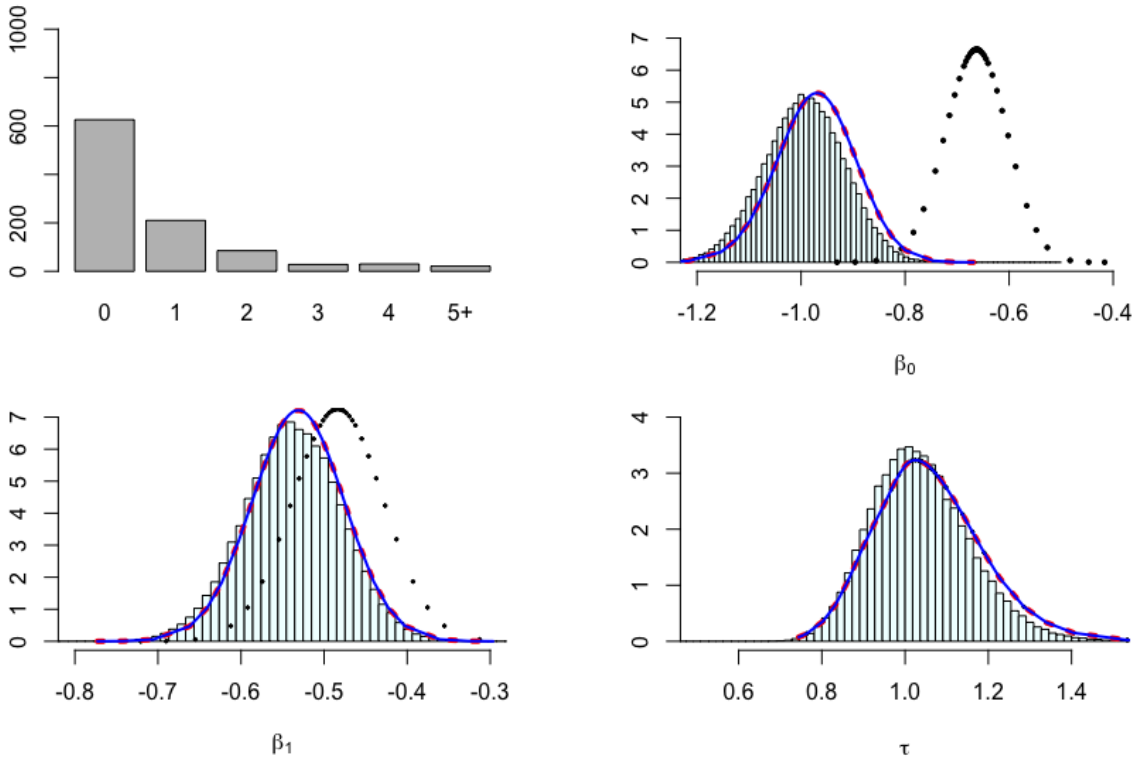

Figure 2: Poisson counts simulated from (19) (top left) and the marginal posterior of β_0 (top right), β_1 (bottom left) and τ (bottom right) from the Gaussian strategy (points), Laplace strategy (dashed line), INLA-VBC (solid line) and MCMC

5.1 Cyclic second order random walk - small scale

The Tokyo data set (Rue and Held, 2005) in the R-INLA library contains information on the number of times the daily rainfall measurements in Tokyo was more than 1mm on a specific day t for two consecutive years. In order to model the annual rainfall pattern, a stochastic spline model with fixed precision is used to smooth the data. In this example we use a cyclic random walk order two model defined as follows:

$$\begin{aligned} y_i|\boldsymbol{\psi} &\sim Bin\left(n_i, p_i = \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}\right) \\ (\alpha_{i+1} - 2\alpha_i + \alpha_{i-1})|\tau &\overset{\text{iid}}{\sim} N(0, \tau^{-1}), \end{aligned}$$

where $i = 1, 2, ..., 366$, $\boldsymbol{\alpha}$ is a stochastic spline second order random walk model on a circle (Rue and Held, 2005), and $n_{60} = 1$ else $n_i = 2$. The latent field is $\boldsymbol{\psi} = \{\boldsymbol{\alpha}\}$ and we fix the hyperparameter $\tau = 1$. Here we apply the correction to $\boldsymbol{\alpha}$, so that $I = \{1, 2, ..., n\}$. In Figure 3 we present the posterior mean of the spline, estimated with each of the methods and also the posterior marginal for one specific element, to illustrate the uncertainty in the different posteriors. We can see clearly that the approximate posterior mean of the spline with INLA-VBC is significantly improved from the Laplace method, while it is very

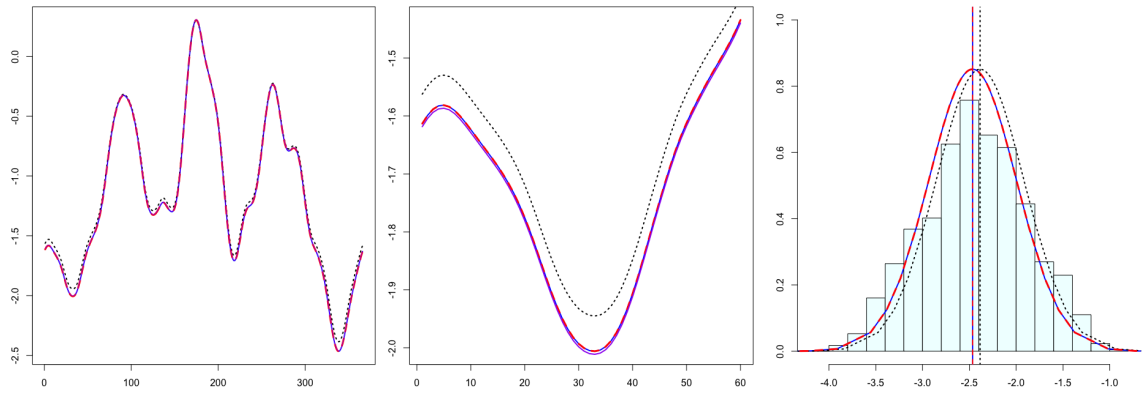

Figure 3: Posterior mean of $\boldsymbol{\alpha}$ (left) (zoomed for the first two months (center)) and the marginal posterior of α_{339} (right) from the Gaussian strategy (points), INLA-VBC (solid blue line), INLA (broken line) and MCMC (solid purple line and histogram)

close to the posterior mean from INLA and MCMC. All methods estimate similar posterior uncertainty as shown in Figure 3.

As a measure of error consolidation, we note that the mean of the absolute errors produced between the Gaussian strategy and INLA is 0.0358 while for INLA-VBC it is 0.0009, underpinning the findings as illustrated in Figure 1. The time for all methods were less than 6 seconds, as expected due to the small dimension of the data and latent field, although the time for MCMC was 87.63 seconds. In this real example, the INLA-VBC does not offer much computational gain over the Laplace strategy, although we have a larger difference in computational cost for an increase in data size or model complexity, as shown in the next example.

5.2 Leukemia data set - large scale

Consider the R data set `Leuk` that features the survival times of 1043 patients with acute myeloid leukemia (AML) in Northwest England between 1982 to 1998, for more details see Henderson et al. (2002). Exact residential locations and districts of the patients are known and indicated by the dots in Figure 4. The aim is to model the survival time based on various covariates $\boldsymbol{X}$ and space $\boldsymbol{s}$, with a Cox proportional hazards model,

$$h(t, \boldsymbol{s}) = h_0(t) \exp(\boldsymbol{\beta X} + \boldsymbol{u}(\boldsymbol{s})),$$

where the baseline hazard function $h_0(t)$ is modeled with a stochastic spline as in Section 5.1, with hyperparameter τ. The baseline hazard is modeled using 100 time intervals and we use the data augmented Poisson regression technique as proposed by Holford (1980); Laird and Olivier (1981). This then implies an augmented data set of size 11738.

As fixed effects we use scaled age (Age), scaled white blood cell count at diagnosis (WBC) and the scaled Townsend score (TPI) as prognostic factors. Then to account for spatial variation we use a Gaussian effect $\boldsymbol{u}$ with a Matérn covariance structure with hyperparameters, marginal variance σ_u^2 and nominal range $r = 2/\kappa$ (Lindgren et al., 2011).

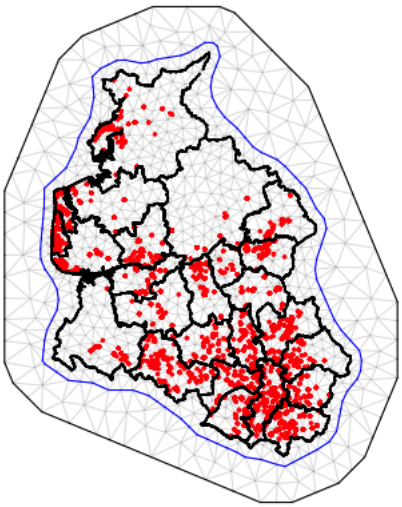

Figure 4: Exact residential locations of patients with AML (dots) and the triangulated mesh for the finite element method estimation of the SPDE of the spatial field

The model for the linear predictor is

$$\eta_i(s) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{WBC}_i + \beta_3 \text{TPI}_i + u(s). \tag{30}$$

The model for $\boldsymbol{u}$ is continuously indexed in space and we use the finite element method and the mesh presented in Figure 4 to estimate this model (for more details regarding the SPDE approach to continuous spatial modeling see Lindgren et al. (2011); Lindgren and Rue (2015); Krainski et al. (2018)). The mesh contains 2032 triangles, and through the mapping of the data to the mesh, we get an augmented latent field of size $m = 39158$. The hyperparameters to be estimated is $\boldsymbol{\theta} = \{\tau, \sigma_u^2, r\}$. Here we apply the correction to the fixed effects only, hence $p = 4$, while the other $m = 39154$ corrections implicitly follow. We present the fixed effects posterior means for this example in Table 3, as well as the computational time (which includes the time to estimate $\boldsymbol{\theta}$). It is clear that we can achieve the same accuracy as the Laplace strategy (INLA), at a fraction of the computational cost with the INLA-VBC.

The marginal posteriors of $\beta_0, \beta_1, \beta_2$ and β_3 are presented in Figure 5 and the accuracy of the correction is clear. Note that for β_1, β_2 and β_3, the posterior means from the Gaussian strategy are already very close to those from INLA. In this case we see that the correction from INLA-VBC is stable by estimating only a slight correction. The posterior mean and 95% credible interval of the baseline hazard $h_0(t)$ is presented in Figure 5 and we see that even though we only explicitly correct the four fixed effects, the posterior mean of the baseline hazard is also corrected. Additionally, the posterior mean of the Gaussian field, $\boldsymbol{u}(\boldsymbol{s})$ is presented in Figure 6 for INLA-VBC (left), as well as the posterior standard deviation (center). Based on the posterior mean of $\boldsymbol{u}(\boldsymbol{s})$ we can clearly identify areas which have increased risk (red) of death due to AML and also areas where the risk is lower (blue). These areas can be used to inform public health interventions to be targeted towards those

	GA	INLA	INLA-VBC
β_0	-2.023	-2.189	-2.189
β_1	0.596	0.597	0.597
β_2	0.242	0.241	0.241
β_3	0.108	0.108	0.108
τ	0.340	0.340	0.340
σ_u	0.223	0.223	0.223
r	0.202	0.202	0.202
Time(s)	25.9	1276	26.3

Table 3: Posterior means from the Gaussian strategy, INLA and INLA-VBC - all fixed effects are significant (see Figure 5)

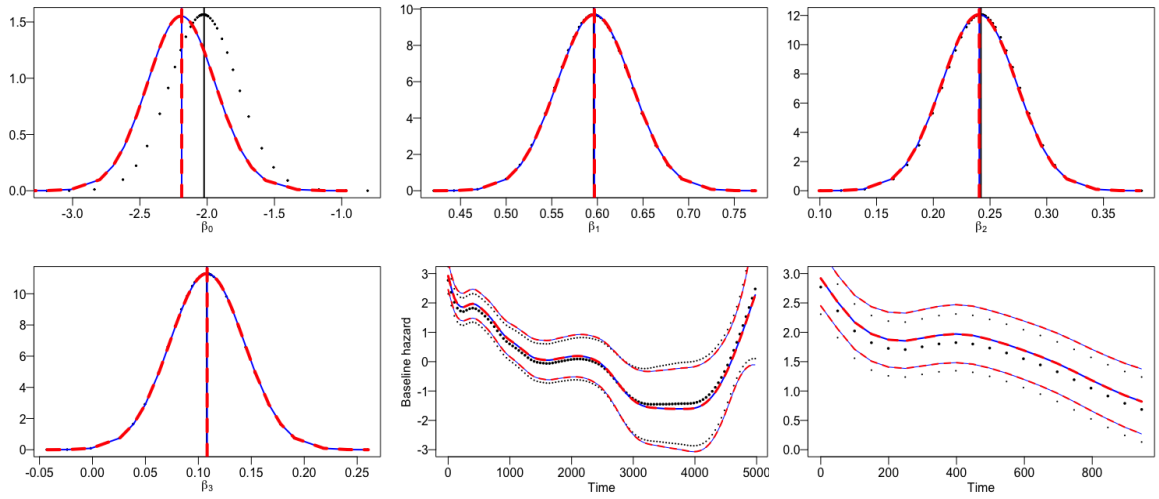

Figure 5: Marginal posteriors from the Gaussian strategy (points), INLA-VBC (solid line) and INLA (broken line) for the fixed effects and posterior mean and 95% credible interval for the baseline hazard $h_0(t)$

areas in need. Since we propagated the explicit correction of the fixed effects to the spatial field as well, we see that corrections were made to the Gaussian field based on the INLA-VBC strategy for most locations (see Figure 6 (right)).

This real example illustrates the potential and need of our proposal, to perform more accurate approximations to the posterior mean (and thus point estimates) for *all* model components (in this example we have made $m = 39158$ improvements to the joint posterior mean) by calculating an optimization in a very small dimension (in this example we solved (25) with $p = 4$).

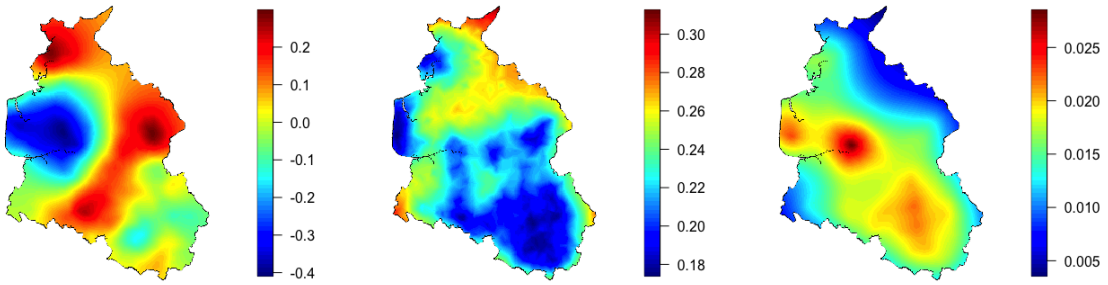

Figure 6: Posterior mean (left) and posterior standard deviation (center) of $\boldsymbol{u}(\boldsymbol{s})$ from INLA-VBC with the absolute difference between the posterior means of $\boldsymbol{u}(\boldsymbol{s})$ from the Gaussian strategy and INLA-VBC (right)

6. Discussion and future directions

In this paper we proposed a method to correct the posterior mean from a Laplace method using a low-rank correction that propagates to a higher dimensional latent parameter space. We use a variational framework to calculate this lower-dimensional correction. This proposal is useful for problems where a Gaussian approximation from the Laplace method is used to approximate an unknown function, or as an intermediate step in a specific algorithm. We show that the VBC works well compared to MCMC for unimodal unknown functions, in terms of location and uncertainty estimation. Moreover, we apply the VBC to the INLA methodology and construct INLA-VBC that performs full Bayesian inference for latent Gaussian models in an accurate and efficient manner. INLA-VBC achieves similar accuracy in the mean than that of more costly procedures (like MCMC and the Laplace strategy of INLA), without much additional computational cost to that of the Laplace method. INLA-VBC is implemented in the *R-INLA* library and available for use with the *inla* function, more details are available at `www.r-inla.org`.

VBC is not merely a different technique to perform optimization for a Variational Bayes problem, but rather poses a new variational optimization that can be defined on a much smaller dimension than the dimension of the parameter space, while providing results for the entire parameter space. As such, VBC is not to be pinned against other VB computational approaches like inducing point methods, stochastic variational inference, minibatching, boosting approaches, normalizing flow etc, but rather proposes a new framework within which these techniques can be applied.

VBC can also be used to do a VB (Bayesian) correction to the maximum likelihood estimator (MLE) from a generalized linear model. This results in an approximate Bayesian inference framework, starting from a Frequentist basis. The MLE, $\boldsymbol{\mu}$, of $\boldsymbol{\psi}$ is calculated as

$$\boldsymbol{\mu} = \arg\max_{\boldsymbol{\psi}} \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\psi}).$$

We can also calculate the precision matrix $\boldsymbol{Q}$, for $\boldsymbol{\psi}$ from the Hessian of the log-likelihood at $\boldsymbol{\mu}$. We thus infer, $\boldsymbol{\psi} \sim N(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$. Similar to Section 2 we impose a low-rank implicit correction to $\boldsymbol{\mu}$. We postulate a Gaussian posterior of $\boldsymbol{\psi}$ with the corrected mean

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu} + \boldsymbol{Q}_I^{-1}\boldsymbol{\lambda},$$

and solve for $\boldsymbol{\lambda}$ using (13). This corrected posterior mean then provides the scientist with a Bayesian estimator adapted from MLE, without performing a complete Bayesian analysis.

VBC has the potential to be used also for marginal variance correction and possibly even skewness correction when we move from the Gaussian posterior to a skew-normal posterior family, with a Gaussian copula. As shown in the simulated and real examples, often the variance resulting from the Laplace method is quite accurate when compared with that of MCMC and only a slight correction would be necessary. In the non-latent Gaussian model example the posteriors did not depart significantly from symmetry and the Gaussian posterior appears to be sufficient. In the case latent Gaussian models where we proposed INLA-VBC, the marginal posteriors are not assumed to be symmetric since the integration over the hyperparameter space induces skewness to the Gaussian *conditional* posterior where the VBC was applied, and from the examples considered here the resulting marginal posteriors compare well with that of MCMC. However, scenarios could arise where a variance and skewness correction are beneficial and we are currently exploring these avenues. Initial work in this area is promising, although the task at hand is more demanding.

The work we present herein is based on using the variational concept in an interesting and promising fashion, and we believe that it contributes to the field of approximate Bayesian inference for a large class of models as well as to approximate methods in general by producing accurate results with superior computational efficiency and scalability.

The examples presented herein can be reproduced based on the code available at `https://github.com/JanetVN1201/Code_for_papers/tree/main/Low-rank%20VB%20correction%20to%20GA`.

7. Appendix: Optimization-based view of Variational Bayes

The Variational Bayes framework as proposed by Zellner (1988) can be summarized as follows.

Based on prior information $\mathcal{I}$, data $\boldsymbol{y}$ and parameters $\boldsymbol{\theta}$, define the following:

1. $\pi(\boldsymbol{\theta}|\mathcal{I})$ is the prior model assumed for $\boldsymbol{\theta}$ before observing the data

2. $q(\boldsymbol{\theta}|\mathcal{D})$ is the learned model from the prior information and the data where $\mathcal{D} = \{\mathcal{I}, \boldsymbol{y}\}$

3. $l(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood of state $\boldsymbol{\theta}$ based on the data $\boldsymbol{y}$

4. $p(\boldsymbol{y}|\mathcal{I})$ is the model for the data where $p(\boldsymbol{y}|\mathcal{I}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathcal{I})d\boldsymbol{\theta}$

The input information in the learning of $\boldsymbol{\theta}$ is given by $\pi(\boldsymbol{\theta}|\mathcal{I})$ and $l(\boldsymbol{\theta}|\boldsymbol{y})$. An information processing rule (IPR) then delivers $q(\boldsymbol{\theta}|\mathcal{D})$ and $p(\boldsymbol{y}|\mathcal{I})$ as output information. A stable and efficient IPR would provide the same amount of output information than received through the input information, thus being information conservative. Thus, we learn $q(\boldsymbol{\theta}|\mathcal{D})$ such

that it minimizes

$$\begin{aligned}
& -\int \left[\log \pi(\boldsymbol{\theta}|\mathcal{I}) + \log l(\boldsymbol{\theta}|\boldsymbol{y})\right] q(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} + \int \left[\log q(\boldsymbol{\theta}|\mathcal{D}) + \log p(\boldsymbol{y}|\mathcal{I})\right] q(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \\
= & -\int \log \pi(\boldsymbol{\theta}|\mathcal{I}) q(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} - \int \log l(\boldsymbol{\theta}|\boldsymbol{y}) q(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} + \int \log q(\boldsymbol{\theta}|\mathcal{D}) q(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} + \log p(\boldsymbol{y}|\mathcal{I}) \\
\propto & \; E_{q(\boldsymbol{\theta}|\mathcal{D})}\left[-\log l(\boldsymbol{\theta}|\boldsymbol{y})\right] + \int \left[-\log \pi(\boldsymbol{\theta}|\mathcal{I}) + \log q(\boldsymbol{\theta}|\mathcal{D})\right] q(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \\
= & \; E_{q(\boldsymbol{\theta}|\mathcal{D})}\left[-\log l(\boldsymbol{\theta}|\boldsymbol{y})\right] + \mathrm{KLD}\left[q(\boldsymbol{\theta}|\mathcal{D})||\pi(\boldsymbol{\theta}|\mathcal{I})\right] \qquad (31)
\end{aligned}$$

where $\mathrm{KLD}\left[a(x)||b(x)\right] = \int \log \frac{a(x)}{b(x)} a(x) dx$ is the Kullback-Leibler divergence measure (or relative entropy).

Zellner (1988) showed that the learned $q(\boldsymbol{\theta}|\mathcal{D})$ corresponds to the posterior density derived from Bayes' theorem, and if we define $q(\boldsymbol{\theta}|\mathcal{D})$ to be the true posterior distribution then the IPR in (31) is 100% efficient. It is optimal in the sense that the amount of the input and output information is as close to each other as possible (also the negative entropy of $q(\boldsymbol{\theta}|\mathcal{D})$ is minimized relative to $\frac{\pi(\boldsymbol{\theta}|\mathcal{I}) l(\boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{y}|\mathcal{I})}$).

Here the Variational concept relates to finding the best candidate based on assumptions of the analytical form of $q(\boldsymbol{\theta}|\mathcal{D})$, that minimizes (31). This view on variational Bayesian inference is beneficial since we do not have to assume that $q(\boldsymbol{\theta}|\mathcal{D})$ is decoupled for $\boldsymbol{\theta}$, like the mean field assumption.

References

Christophe Andrieu and Gareth O Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1):5–43, 2003.

H Attias. Learning parameters and structure of latent variable models by variational Bayes." in proc. In *Uncertainty in Artificial Intelligence*, 1999.

Haakon Bakka, Håvard Rue, Geir-Arne Fuglstad, Andrea Riebler, David Bolin, Janine Illian, Elias Krainski, Daniel Simpson, and Finn Lindgren. Spatial Modeling with R-INLA: A Review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(6):e1443, 2018. doi: 10.1002/wics.1443. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1443`.

Timothy D Barfoot, James R Forbes, and David J Yoon. Exactly sparse Gaussian variational inference with application to derivative-free batch nonlinear state estimation. *The International Journal of Robotics Research*, 39(13):1473–1502, 2020.

Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Michael Betancourt and Mark Girolami. Hamiltonian monte carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79(30):2–4, 2015.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

GEP Box and KB Wilson. On the experimental designs for exploring response surfaces. *Ann Math Stat*, 13:1–45, 1951.

George Casella and Edward I George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Xi Chen, Adityan, Guntuboyina, and Yuchen Zhang. On Bayes Risk Lower Bounds. *Journal of Machine Learning Research*, 17(218):1–58, 2016. URL `http://jmlr.org/papers/v17/16-185.html`.

Guillaume P Dehaene and Simon Barthelmé. Bounding errors of expectation-propagation. *arXiv preprint arXiv:1601.02387*, 2016.

Esmail Abdul Fattah, Janet Van Niekerk, and Håvard Rue. Smart gradient-an adaptive technique for improving gradient estimation. *Foundations of Data Science*, 4(1):123–136, 2022.

Lisa Gaedke-Merzhäuser, Janet van Niekerk, Olaf Schenk, and Håvard Rue. Parallelized integrated nested Laplace approximations for fast Bayesian inference. *Statistics and Computing*, 33(1):25, 2023.

Théo Galy-Fajou, Valerio Perrone, and Manfred Opper. Flexible and Efficient Inference with Particles for the Variational Gaussian Approximation. *Entropy*, 23(8):990, 2021.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

Robin Henderson, Silvia Shimakura, and David Gorst. Modeling spatial variation in leukemia survival data. *Journal of the American Statistical Association*, 97(460):965–972, 2002.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

Theodore R Holford. The analysis of rates and of survivorship using log-linear models. *Biometrics*, pages 299–305, 1980.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *The Journal of Machine Learning Research*, 23(1):5789–5897, 2022.

Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33, 2020.

Elias T Krainski, Virgilio Gómez-Rubio, Haakon Bakka, Amanda Lenzi, Daniela Castro-Camilo, Daniel Simpson, Finn Lindgren, and Håvard Rue. *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC, 2018.

Nan Laird and Donald Olivier. Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374):231–240, 1981.

Marc Lambert, Silvere Bonnabel, and Francis Bach. The recursive variational Gaussian approximation (R-VGA). 2020.

Pierre Simon Laplace. Memoir on the probability of the causes of events. *Statistical science*, 1(3):364–378, 1986.

Finn Lindgren and Håvard Rue. Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63(1):1–25, 2015.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73 (4):423–498, 2011.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *arXiv preprint arXiv:1608.04471*, 2016.

Jianfeng Lu, Yulong Lu, and James Nolen. Scaling limit of the Stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671, 2019.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

T Minka. Expectation propagation for approximate Bayesian inference, Doctoral Dissertation. 2001.

Henry B. Moss, David S. Leslie, Javier Gonzalez, and Paul Rayson. GIBBON: General-purpose Information-Based Bayesian Optimisation. *Journal of Machine Learning Research*, 22(235):1–49, 2021. URL `http://jmlr.org/papers/v22/21-0120.html`.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Manfred Opper and Ole Winther. Gaussian processes for classification: Mean-field algorithms. *Neural computation*, 12(11):2655–2684, 2000.

Victor Picheny, Tobias Wagner, and David Ginsbourger. A benchmark of kriging-based infill criteria for noisy optimization. *Structural and multidisciplinary optimization*, 48: 607–626, 2013.

Nicholas G Polson and Vadim Sokolov. Deep learning: A Bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304, 2017.

Tanzeel U Rehman, Md Sultan Mahmud, Young K Chang, Jian Jin, and Jaemyung Shin. Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. *Computers and electronics in agriculture*, 156:585–605, 2019.

Eitan Richardson and Yair Weiss. A Bayes-Optimal View on Adversarial Examples. *Journal of Machine Learning Research*, 22(221):1–28, 2021. URL `http://jmlr.org/papers/v22/20-567.html`.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Peter J Rossky, Jimmie D Doll, and Harold L Friedman. Brownian dynamics as smart Monte Carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, 1978.

Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

Rajiv Sambasivan, Sourish Das, and Sujit K Sahu. A Bayesian perspective of statistical machine learning for big data. *Computational Statistics*, 35(3):893–930, 2020.

Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

Sergios Theodoridis. *Machine learning: a Bayesian and optimization perspective*. Academic press, 2015.

Luke Tierney, Robert E Kass, and Joseph B Kadane. Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84(407):710–716, 1989.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Janet Van Niekerk, Haakon Bakka, Haavard Rue, and Olaf Schenk. New frontiers in Bayesian modeling using the INLA package in R. *Journal of Statistical Software*, 100 (2):1–28, 2021. doi: 10.18637/jss.v100.i02.

Aki Vehtari, Andrew Gelman, Tuomas Sivula, Pasi Jylänki, Dustin Tran, Swupnil Sahai, Paul Blomstedt, John P. Cunningham, David Schiminovich, and Christian P. Robert. Expectation Propagation as a Way of Life: A Framework for Bayesian Inference on Partitioned Data. *Journal of Machine Learning Research*, 21(17):1–53, 2020. URL `http://jmlr.org/papers/v21/18-817.html`.

Arnold Zellner. Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):278–280, 1988.

Jingwei Zhuo, Chang Liu, Jiaxin Shi, Jun Zhu, Ning Chen, and Bo Zhang. Message passing Stein variational gradient descent. In *International Conference on Machine Learning*, pages 6018–6027. PMLR, 2018.