Statistics and Computing (2024) 34:110
https://doi.org/10.1007/s11222-024-10422-8

ORIGINAL PAPER

R-VGAL: a sequential variational Bayes algorithm for generalised linear mixed models

Bao Anh Vu[1,2] · David Gunawan[1] · Andrew Zammit-Mangion[1,2]

Received: 1 September 2023 / Accepted: 9 March 2024

Abstract

Models with random effects, such as generalised linear mixed models (GLMMs), are often used for analysing clustered data. Parameter inference with these models is difficult because of the presence of cluster-specific random effects, which must be integrated out when evaluating the likelihood function. Here, we propose a sequential variational Bayes algorithm, called Recursive Variational Gaussian Approximation for Latent variable models (R-VGAL), for estimating parameters in GLMMs. The R-VGAL algorithm operates on the data sequentially, requires only a single pass through the data, and can provide parameter updates as new data are collected without the need of re-processing the previous data. At each update, the R-VGAL algorithm requires the gradient and Hessian of a "partial" log-likelihood function evaluated at the new observation, which are generally not available in closed form for GLMMs. To circumvent this issue, we propose using an importance-sampling-based approach for estimating the gradient and Hessian via Fisher's and Louis' identities. We find that R-VGAL can be unstable when traversing the first few data points, but that this issue can be mitigated by introducing a damping factor in the initial steps of the algorithm. Through illustrations on both simulated and real datasets, we show that R-VGAL provides good approximations to posterior distributions, that it can be made robust through damping, and that it is computationally efficient.

Keywords Fisher's identity · Intractable gradient · Latent variable model · Louis' identity · Damped Newton's method

1 Introduction

Mixed models are useful for analysing clustered data, wherein observations that come from the same cluster/group are likely to be correlated. Example datasets include records of students clustered within schools, and repeated measurements of biomarkers on patients. Mixed models account for intra-group dependencies by incorporating cluster/group-specific "random effects". Inference with these models is made challenging by the fact that the likelihood function involves integrals over the random effects that are not usually tractable except for the few cases where the distribution of the random effects is conjugate to the distribution of the data, such as in the linear mixed model (Verbeke et al. 1997), the beta-binomial model (Crowder 1979), and Rasch's Poisson count model (Jansen 1994). Notably, there is no closed-form expression for the likelihood function in the case of the ubiquitous logistic mixed model.

Maximum-likelihood-based approaches are often used for parameter inference in mixed models. In the case of linear mixed models, parameter inference via maximum likelihood estimation is straightforward (e.g., Wakefield 2013). For mixed models with an intractable likelihood, integrals over random effects need to be numerically approximated, for example by using Gaussian quadrature (Naylor and Smith 1982) or the Laplace approximation (Tierney and Kadane 1986). The likelihood may also be indirectly maximised using an expectation-maximisation type algorithm (Dempster et al. 1977), which treats the random effects as missing, and iteratively maximises the "expected complete-data log-likelihood" of the data and the random effects. Quasi-likelihood approaches such as penalised quasi-likelihood (PQL, Breslow and Clayton 1993) and marginal quasi-likelihood (MQL, Goldstein 1991) approximate nonlinear mixed models with linear mixed models, so that well-developed estimation routines for linear mixed models can be applied; see Tuerlinckx et al. (2006) for a detailed dis-

✉ Bao Anh Vu
bavu@uow.edu.au

[1] School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW, Australia

[2] Securing Antarctica's Environmental Future, University of Wollongong, Wollongong, NSW, Australia

Published online: 06 April 2024

cussion of these methods. These maximum-likelihood-based methods provide point estimates and not full posterior distributions over the parameters.

Full posterior distributions can be obtained using Markov chain Monte Carlo (MCMC, e.g., Zhao et al. 2006; Fong et al. 2010). MCMC provides exact, sample-based posterior distributions, but at a higher computational cost than maximum-likelihood-based methods. Alternatively, variational Bayes (VB) methods (e.g., Ong et al. 2018; Tan and Nott 2018) are becoming increasingly popular for estimating parameters in complex statistical models. These methods approximate the exact posterior distribution with a member from a simple and tractable family of distributions; this family is usually chosen to balance the accuracy of the approximation against the computational cost required to obtain the approximation. VB methods are usually computationally cheaper than MCMC methods. VB approaches can either batch-process the data (e.g., Tran et al. 2016; Ong et al. 2018; Tan and Nott 2018) or sequentially process data points (e.g., Broderick et al. 2013; Gunawan et al. 2021; Lambert et al. 2022). For settings with large amounts of data, a method that targets the posterior distribution via sequential processing of the data offers several advantages. The so-called Recursive Variational Gaussian Approximation (R-VGA, Lambert et al. 2022) algorithm is a recently-developed sequential variational Bayes method that provides a fast and accurate approximation to the posterior distribution with only one pass through the data, making it computationally efficient when compared to MCMC or batch variational Bayes. Lambert et al. (2022) apply the R-VGA algorithm to linear and logistic regression models without random effects.

In this paper, we build on the R-VGA algorithm by proposing a novel recursive variational Gaussian approximation, called Recursive Variational Gaussian Approximation for Latent variable models (R-VGAL), for estimating the parameters in GLMMs. At each update, R-VGAL requires the gradient and Hessian of the "partial" log-likelihood evaluated at the new observation, which are often not available in closed form. To circumvent this issue, we propose an importance-sampling-based approach for estimating the gradient and Hessian that uses Fisher's and Louis' identities (Cappé et al. 2005). This approach was inspired by the work of Nemeth et al. (2016), who used Fisher's and Louis' identities to approximate the gradient and Hessian in a sequential Monte Carlo context. The efficacy of R-VGAL is illustrated using linear, logistic and Poisson mixed effect models on simulated and real datasets. The examples show that R-VGAL provides good approximations to the exact posterior distributions estimated using Hamiltonian Monte Carlo (HMC, Neal 2011; Betancourt and Girolami 2015) and at a low computational cost.

The paper is organised as follows. Section 2 provides some background on the sequential variational Bayes framework and presents the R-VGAL algorithm. Section 3 applies the R-VGAL algorithm to simulated and real datasets. Section 4 concludes with a discussion of our results and an overview of future research directions. This article has an online supplement containing additional technical details, and the code to reproduce results from the simulation and real-data experiments is available on https://github.com/bao-anh-vu/R-VGAL.

2 The R-VGAL algorithm

This section reviews GLMMs (e.g. Demidenko 2013; Faraway 2016) and provides some background on the R-VGA algorithm of Lambert et al. (2022), and then introduces the R-VGAL algorithm for making parameter inference with GLMMs.

2.1 Generalised linear mixed models

GLMMs are statistical models that contain both fixed effects and random effects. Typically, the fixed effects are common across groups, while the random effects are group-specific, and this is the setting we focus on. We briefly discuss the potential application of R-VGAL to models with more complicated random effect structures, such as crossed or nested random effects, in Section S7 of the online supplement.

Denote by y_{ij} the jth response in the ith group, for $i = 1, \dots, N$ groups and $j = 1, \dots, n_i$, where n_i is the number of responses in group i. Let $\mathbf{y} \equiv (\mathbf{y}_1^\top, \dots, \mathbf{y}_N^\top)^\top$ be a vector of observations, where $\mathbf{y}_i \equiv (y_{i1}, \dots, y_{in_i})^\top$ are the responses from the ith group. The GLMMs we consider are constructed by first assigning each y_{ij} a distribution $y_{ij} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}_i, \phi \sim p(\cdot)$, where $p(\cdot)$ is a member of the exponential family with a dispersion parameter ϕ that is usually related to the variance of the datum, $\boldsymbol{\beta}$ are the fixed effect parameters, and $\boldsymbol{\alpha}_i$ are the group-specific random effects for $i = 1, \dots, N$. Then, the mean of the responses, $\mu_{ij} \equiv \mathbb{E}(y_{ij} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}_i, \phi)$, is modelled as

$$g(\mu_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\alpha}_i, \quad i = 1, \dots, N, \quad j = 1, \dots, n_i, \tag{1}$$

where $\mathbf{x}_{ij}$ is a vector of fixed effect covariates corresponding to the jth response in the ith group; $\mathbf{z}_{ij}$ is a vector of predictor variables corresponding to the jth response and the ith random effect; and $g(\cdot)$ is a link function that links the response mean μ_{ij} to the linear predictor $\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\alpha}_i$. We further assume that $\boldsymbol{\alpha}_i \perp\!\!\!\perp \boldsymbol{\alpha}_{i'}$ for $i \neq i'$. The random effects $\boldsymbol{\alpha}_i$, for $i = 1, \dots, N$, are assumed to follow a normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_\alpha$, that is, each $\boldsymbol{\alpha}_i \mid \boldsymbol{\Sigma}_\alpha \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\alpha)$. In practice, some structure is often assumed for the random effects covariance matrix so that it is parameterised in terms of a smaller number of parameters $\boldsymbol{\tau}$, that is, $\boldsymbol{\Sigma}_\alpha = \boldsymbol{\Sigma}_\alpha(\boldsymbol{\tau})$. Inference is then made on the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\tau}^\top, \phi)^\top$.

The main objective of Bayesian inference is to obtain the posterior distribution of the model parameters $\boldsymbol{\theta}$ given the observations $\mathbf{y}$ and the prior distribution $p(\boldsymbol{\theta})$. Through Bayes' rule, the posterior distribution of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = p(\boldsymbol{\beta}, \boldsymbol{\tau}, \phi \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \phi) p(\boldsymbol{\beta}, \boldsymbol{\tau}, \phi). \quad (2)$$

The likelihood function,

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \phi) = \prod_{i=1}^{N} \int p(\mathbf{y}_i \mid \boldsymbol{\alpha}_i, \boldsymbol{\beta}, \phi) p(\boldsymbol{\alpha}_i \mid \boldsymbol{\tau}) \, \mathrm{d}\boldsymbol{\alpha}_i, \quad (3)$$

involves integrals over the random effects $\boldsymbol{\alpha}_i$, $i = 1, \ldots, N$. The likelihood function can be calculated exactly for the linear mixed model with normally distributed random effects, for which

$$\begin{aligned} y_{ij} &= \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\alpha}_i + \epsilon_{ij}, \\ &\boldsymbol{\alpha}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\alpha(\boldsymbol{\tau})), \quad \epsilon_{ij} \sim N(0, \sigma_\epsilon^2), \end{aligned} \quad (4)$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$, where ϵ_{ij} is a zero-mean, independent, normally distributed error term with variance σ_ϵ^2 that is associated with the jth response from the ith group. At the group level, this model can be written as

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\alpha}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\alpha(\boldsymbol{\tau})), \\ &\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{n_i}), \end{aligned}$$

where $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})^\top$, and $\mathbf{Z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{in_i})^\top$, $\boldsymbol{\varepsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^\top$, with n_i being the number of observations in the ith group, for $i = 1, \ldots, N$, and $\mathbf{I}_m$ denotes an identity matrix of size $m \times m$. The likelihood function for this linear mixed model is

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\epsilon^2) &= \prod_{i=1}^{N} p(\mathbf{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\epsilon^2) \\ &= \prod_{i=1}^{N} N(\mathbf{X}_i \boldsymbol{\beta}, \mathbf{Z}_i \boldsymbol{\Sigma}_\alpha(\boldsymbol{\tau}) \mathbf{Z}_i^\top + \sigma_\epsilon^2 \mathbf{I}_{n_i}). \end{aligned} \quad (5)$$

The gradient and Hessian of the log-likelihood for the linear mixed model are also available in closed form. However, the likelihood $p(\mathbf{y}_i \mid \boldsymbol{\alpha}_i, \boldsymbol{\beta}, \phi)$ in (3) cannot be computed exactly for general random effects models. One important case is the logistic mixed model given by

$$\begin{aligned} &y_{ij} \sim \text{Bernoulli}(\pi_{ij}), \quad \text{logit}(\pi_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\alpha}_i, \\ &i = 1, \ldots, N, \quad j = 1, \ldots, n_i, \end{aligned} \quad (6)$$

where $\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)$. The gradient and Hessian of the log-likelihood function for this model can, however, be estimated unbiasedly, as we show in Sects. 2.3.1 and 2.3.2.

2.2 Sequential VB and R-VGA

We begin this section with a review of VB and the sequential VB framework. We then present the main steps in the derivations of the R-VGA algorithm of Lambert et al. (2022), on which our algorithm is based.

2.2.1 Sequential VB

VB is usually used for posterior inference in complex statistical models when inference using asymptotically exact methods such as MCMC is too costly; for a review see, for example, Blei et al. (2017). Let $\boldsymbol{\theta}$ be a vector of model parameters. Here, we consider the class of VB methods where the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$ is approximated by a tractable density $q(\boldsymbol{\theta}; \boldsymbol{\lambda})$ parameterised by $\boldsymbol{\lambda}$. The variational parameters $\boldsymbol{\lambda}$ are optimised by minimising the Kullback–Leibler (KL) divergence between the variational distribution and the posterior distribution, that is, by minimising

$$\text{KL}(q(\boldsymbol{\theta}; \boldsymbol{\lambda}) \parallel p(\boldsymbol{\theta} \mid \mathbf{y})) \equiv \int q(\boldsymbol{\theta}; \boldsymbol{\lambda}) \log \frac{q(\boldsymbol{\theta}; \boldsymbol{\lambda})}{p(\boldsymbol{\theta} \mid \mathbf{y})} \, \mathrm{d}\boldsymbol{\theta}. \quad (7)$$

Many VB algorithms require processing the data as a batch; see, for example, Ong et al. (2018) and Tan and Nott (2018). The variational parameters $\boldsymbol{\lambda}$ are typically updated in an iterative manner using stochastic gradient descent (SGD, Hoffman et al. 2013; Kingma and Welling 2013). In settings with large amounts of data or continuously-arriving data, it is often more practical to use online or sequential variational Bayes algorithms that update the approximation to the posterior distribution sequentially as new observations become available. These online/sequential algorithms are designed to handle data that are too large to fit in memory or that arrive in a continuous stream.

In a sequential VB framework, such as that proposed by Broderick et al. (2013), the observations $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are incorporated sequentially so that at iteration i, $i = 1, \ldots, N$, one targets an approximation $q_i(\boldsymbol{\theta}) \equiv q(\boldsymbol{\theta}; \boldsymbol{\lambda}_i)$ that is closest in a KL sense to the "pseudo-posterior" $p(\mathbf{y}_i \mid \boldsymbol{\theta}) q_{i-1}(\boldsymbol{\theta}) / \mathcal{Z}_i$, where

$$\mathcal{Z}_i \equiv \int p(\mathbf{y}_i \mid \boldsymbol{\theta}) q_{i-1}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}. \quad (8)$$

In this framework, $q_{i-1}(\boldsymbol{\theta})$ is treated as the "prior" for the next iteration i, and the KL divergence between $q_i(\boldsymbol{\theta})$ and the "pseudo-posterior" is minimised at each iteration. Broderick et al. (2013) use a mean field VB approach (e.g., Ormerod and Wand 2010), which assumes no posterior dependence between the elements of $\boldsymbol{\theta}$. The R-VGA algorithm proposed by Lambert et al. (2022) follows closely that of Broderick et al. (2013), but uses a variational distribution of the form $q_i(\boldsymbol{\theta}) = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ is a full covariance matrix, and seeks closed-form updates for $\boldsymbol{\lambda}_i \equiv \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ that minimise the KL divergence between $q_i(\boldsymbol{\theta})$ and $p(\mathbf{y}_i \mid \boldsymbol{\theta})q_{i-1}(\boldsymbol{\theta})/\mathcal{Z}_i$ for $i = 1, \ldots, N$. Another sequential VB algorithm that is similar to that of Broderick et al. (2013) is the Updating Variational Bayes (UVB, Tomasetti et al. 2022) algorithm, which uses SGD (Bottou 2010) at every iteration, $i = 1, \ldots, N$, to minimise the KL divergence between $q_i(\boldsymbol{\theta})$ and $p(\mathbf{y}_i \mid \boldsymbol{\theta})q_{i-1}(\boldsymbol{\theta})/\mathcal{Z}_i$. One advantage of UVB compared to R-VGA is that it does not have to assume that the prior and variational distributions are Gaussian; see Section 5.2 of Tomasetti et al. (2022) for an example of UVB where a beta prior is used for one of the parameters and the variational distribution is a mixture of multivariate normal distributions. However, due to the lack of restrictions on the form of the variational distribution, UVB requires running a full optimisation algorithm at each iteration, whereas the R-VGA updates are available in closed form.

Detailed derivations for the R-VGA algorithm can be found in Lambert et al. (2022). We provide below a sketch of the derivations to aid the exposition of the methodology in subsequent sections.

2.2.2 The R-VGA algorithm

Denote by $\mathbf{y}_{1:i} \equiv (\mathbf{y}_1^\top, \ldots, \mathbf{y}_i^\top)^\top$ a collection of observations from groups 1 to i, $i = 1, \ldots, N$. By assumption of conditional independence between observations $\mathbf{y}_1, \ldots, \mathbf{y}_i$ given the parameters $\boldsymbol{\theta}$, the KL divergence between the variational distribution $q_i(\boldsymbol{\theta})$ and the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y}_{1:i})$ can be expressed as

$$\begin{aligned}\mathrm{KL}(q_i(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{y}_{1:i})) &\equiv \int q_i(\boldsymbol{\theta}) \log \frac{q_i(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \mathbf{y}_{1:i})}\, \mathrm{d}\boldsymbol{\theta} \\ &= \mathbb{E}_{q_i}(\log q_i(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta} \mid \mathbf{y}_{1:i-1}) \\ &\quad - \log p(\mathbf{y}_i \mid \boldsymbol{\theta})) + \log p(\mathbf{y}_{1:i}) \\ &\quad - \log p(\mathbf{y}_{1:i-1}).\end{aligned}$$

The posterior distribution after incorporating the first $i-1$ groups of observations, $p(\boldsymbol{\theta} \mid \mathbf{y}_{1:i-1})$, is approximated by the variational distribution $q_{i-1}(\boldsymbol{\theta})$ to give

$$\begin{aligned}\mathrm{KL}(q_i(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{y}_{1:i})) &\approx \mathbb{E}_{q_i}(\log q_i(\boldsymbol{\theta}) - \log q_{i-1}(\boldsymbol{\theta}) \\ &\quad - \log p(\mathbf{y}_i \mid \boldsymbol{\theta})) + \log p(\mathbf{y}_{1:i}) - \log p(\mathbf{y}_{1:i-1}). \qquad (9)\end{aligned}$$

The R-VGA algorithm assumes a variational distribution of the form $q_i(\boldsymbol{\theta}) = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and seeks parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ that minimise (9). As the last two terms in the right hand side of (9) do not depend on $\boldsymbol{\theta}$, the optimisation problem is equivalent to finding

$$\underset{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i}{\arg\min}\ \mathbb{E}_{q_i}(\log q_i(\boldsymbol{\theta}) - \log q_{i-1}(\boldsymbol{\theta}) - \log p(\mathbf{y}_i \mid \boldsymbol{\theta})). \qquad (10)$$

Differentiating the expectation (10) with respect to $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, setting the derivatives to zero, and rearranging the resulting equations, yields the following recursive updates for the variational mean $\boldsymbol{\mu}_i$ and precision matrix $\boldsymbol{\Sigma}_i^{-1}$:

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_{i-1} + \boldsymbol{\Sigma}_{i-1}\nabla_{\boldsymbol{\mu}_i}\mathbb{E}_{q_i}(\log p(\mathbf{y}_i \mid \boldsymbol{\theta})), \qquad (11)$$

$$\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{\Sigma}_{i-1}^{-1} - 2\nabla_{\boldsymbol{\Sigma}_i}\mathbb{E}_{q_i}(\log p(\mathbf{y}_i \mid \boldsymbol{\theta})). \qquad (12)$$

Then, using Bonnet's Theorem (Bonnet 1964) on (11) and Price's Theorem (Price 1958) on (12), we rewrite the gradient terms as

$$\nabla_{\boldsymbol{\mu}_i}\mathbb{E}_{q_i}(\log p(\mathbf{y}_i \mid \boldsymbol{\theta})) = \mathbb{E}_{q_i}(\nabla_{\boldsymbol{\theta}} \log p(\mathbf{y}_i \mid \boldsymbol{\theta})), \qquad (13)$$

$$\nabla_{\boldsymbol{\Sigma}_i}\mathbb{E}_{q_i}(\log p(\mathbf{y}_i \mid \boldsymbol{\theta})) = \frac{1}{2}\mathbb{E}_{q_i}(\nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{y}_i \mid \boldsymbol{\theta})). \qquad (14)$$

Thus the updates (11) and (12) become

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_{i-1} + \boldsymbol{\Sigma}_{i-1}\mathbb{E}_{q_i}(\nabla_{\boldsymbol{\theta}} \log p(\mathbf{y}_i \mid \boldsymbol{\theta})), \qquad (15)$$

$$\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{\Sigma}_{i-1}^{-1} - \mathbb{E}_{q_i}(\nabla_{\theta}^2 \log p(\mathbf{y}_i \mid \boldsymbol{\theta})). \qquad (16)$$

These updates are implicit as they require the evaluation of expectations with respect to $q_i(\boldsymbol{\theta})$. Under the assumption that $q_i(\boldsymbol{\theta})$ is close to $q_{i-1}(\boldsymbol{\theta})$, Lambert et al. (2022) propose replacing $q_i(\boldsymbol{\theta})$ with $q_{i-1}(\boldsymbol{\theta})$ in (15) and (16), and replacing $\boldsymbol{\Sigma}_{i-1}$ with $\boldsymbol{\Sigma}_i$ on the right hand side of (15), to yield an explicit scheme

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_{i-1} + \boldsymbol{\Sigma}_i\mathbb{E}_{q_{i-1}}(\nabla_{\boldsymbol{\theta}} \log p(\mathbf{y}_i \mid \boldsymbol{\theta})), \qquad (17)$$

$$\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{\Sigma}_{i-1}^{-1} - \mathbb{E}_{q_{i-1}}(\nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{y}_i \mid \boldsymbol{\theta})). \qquad (18)$$

Equations (17) and (18) form the so-called R-VGA algorithm of Lambert et al. (2022).

We note that an "order 1 form" of the R-VGA algorithm exists, which allows the variational precision matrix to be updated using the first order derivatives of the log-likelihood without the need for the Hessian matrix. However, these updates are implicit and not directly implementable. Corollary 1 of Lambert et al. (2022) provides more details on this Hessian-free form.

Algorithm 1 R-VGAL

```
Input: observations $\mathbf{y}_1, \dots, \mathbf{y}_N$, initial values $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$.
Output: variational parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, for $i = 1, \dots, N$.
Set $q_0(\boldsymbol{\theta}) = \text{Gau}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.
for $i = 1, \dots, N$ do
    $\boldsymbol{\mu}_i = \boldsymbol{\mu}_{i-1} + \boldsymbol{\Sigma}_i \mathbb{E}_{q_{i-1}} \widehat{(\nabla_{\boldsymbol{\theta}} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}))}$
    $\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{\Sigma}_{i-1}^{-1} - \mathbb{E}_{q_{i-1}} \widehat{(\nabla^2_{\boldsymbol{\theta}} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}))}$
end for
```

2.3 R-VGAL

The R-VGA updates in (17) and (18) require the gradient $\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ and Hessian $\nabla^2_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ of the "partial" log-likelihood for the ith observation. However, for the GLMMs discussed in Sect. 2.1, there are usually no closed-form expressions for said quantities, as evaluation of the partial log-likelihood involves an intractable integral over the random effects $\boldsymbol{\alpha}_i$. Our R-VGAL algorithm circumvents this issue by replacing $\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ and $\nabla^2_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ with their unbiased estimates, $\widehat{\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})}$ and $\widehat{\nabla^2_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})}$, respectively. These unbiased estimates are obtained by using an importance-sampling-based approach applied to Fisher's and Louis' identities (Cappé et al. 2005), which we discuss in more detail in Sects. 2.3.1 and 2.3.2. We summarise the R-VGAL algorithm in Algorithm 1.

To approximate the expectations with respect to $q_{i-1}(\boldsymbol{\theta})$ in the updates of the variational mean and precision matrix in Algorithm 1, we generate Monte Carlo samples, $\boldsymbol{\theta}^{(l)} \sim q_{i-1}(\boldsymbol{\theta})$, $l = 1, \dots, S$, and compute:

$$\mathbb{E}_{q_{i-1}}\widehat{(\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}))} \approx \frac{1}{S} \sum_{l=1}^{S} \widehat{\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}^{(l)})},$$

$$\mathbb{E}_{q_{i-1}}\widehat{(\nabla^2_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}))} \approx \frac{1}{S} \sum_{l=1}^{S} \widehat{\nabla^2_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}^{(l)})},$$

for $i = 1, \dots, N$.

The following sections discuss approaches to obtain unbiased estimates of the gradient and the Hessian of the log-likelihood with respect to the parameters.

2.3.1 Approximation of the gradient with Fisher's identity

Consider the ith iteration. Fisher's identity (Cappé et al. 2005) for the gradient of $\log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ is

$$\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = \int p(\boldsymbol{\alpha}_i \mid \mathbf{y}_i, \boldsymbol{\theta}) \nabla_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\alpha}_i. \tag{19}$$

If it is possible to sample directly from $p(\boldsymbol{\alpha}_i \mid \mathbf{y}_i, \boldsymbol{\theta})$ (e.g., as it is with the linear random effects model in Sect. 3.1), the above identity can be approximated by

$$\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) \approx \frac{1}{S_\alpha} \sum_{s=1}^{S_\alpha} \nabla_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i^{(s)} \mid \boldsymbol{\theta}),$$
$$\boldsymbol{\alpha}_i^{(s)} \sim p(\boldsymbol{\alpha}_i \mid \mathbf{y}_i, \boldsymbol{\theta}). \tag{20}$$

In the case where direct sampling from $p(\boldsymbol{\alpha}_i \mid \mathbf{y}_i, \boldsymbol{\theta})$ is difficult, we use importance sampling (e.g., Tokdar and Kass 2010) to estimate the gradient of the log-likelihood in (19). Specifically, we draw samples $\{\boldsymbol{\alpha}_i^{(s)} : s = 1, \dots, S_\alpha\}$ from an importance distribution $r(\boldsymbol{\alpha}_i \mid \mathbf{y}_i, \boldsymbol{\theta})$, and then compute the weights

$$w_i^{(s)} = \frac{p(\mathbf{y}_i \mid \boldsymbol{\alpha}_i^{(s)}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}_i^{(s)} \mid \boldsymbol{\theta})}{r(\boldsymbol{\alpha}_i^{(s)} \mid \mathbf{y}_i, \boldsymbol{\theta})}, \quad s = 1, \dots, S_\alpha.$$

The gradient of the log-likelihood is then approximated as

$$\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) \approx \sum_{s=1}^{S_\alpha} \bar{w}_i^{(s)} \nabla_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i^{(s)} \mid \boldsymbol{\theta}), \tag{21}$$

where $\mathcal{W}_i \equiv \{\bar{w}_i^{(s)} : s = 1, \dots, S_\alpha\}$ are the normalised weights given by

$$\bar{w}_i^{(s)} = \frac{w_i^{(s)}}{\sum_{q=1}^{S_\alpha} w_i^{(q)}}, \quad s = 1, \dots, S_\alpha.$$

One possible choice for the importance distribution is the distribution of the random effects, that is, $p(\boldsymbol{\alpha}_i \mid \boldsymbol{\theta})$. In this case, the weights $\mathcal{W}_i$ reduce to

$$w_i^{(s)} = p(\mathbf{y}_i \mid \boldsymbol{\alpha}_i^{(s)}, \boldsymbol{\theta}), \quad s = 1, \dots, S_\alpha.$$

We use this importance distribution in all of the case studies illustrated in Sect. 3.

2.3.2 Approximation of the Hessian with Louis' identity

Consider again the ith iteration. Louis' identity (Cappé et al. 2005) for the Hessian $\nabla^2_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ is

$$-\nabla^2_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = \nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) \nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})^\top - \frac{\nabla^2_\theta p(\mathbf{y}_i \mid \boldsymbol{\theta})}{p(\mathbf{y}_i \mid \boldsymbol{\theta})}, \tag{22}$$

where

$$\frac{\nabla^2_\theta p(\mathbf{y}_i \mid \boldsymbol{\theta})}{p(\mathbf{y}_i \mid \boldsymbol{\theta})} = \int p(\boldsymbol{\alpha}_i \mid \mathbf{y}_i, \boldsymbol{\theta}) \nabla_\theta \log p$$

$$(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})\nabla_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})^\top \,\mathrm{d}\boldsymbol{\alpha}_i + \int p(\boldsymbol{\alpha}_i \mid \mathbf{y}_i, \boldsymbol{\theta})\nabla^2_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\alpha}_i. \tag{23}$$

The first term on the right-hand side of (22) is obtained using Fisher's identity, as discussed in Sect. 2.3.1. The second term consists of two integrals (see (23)), which can also be approximated using samples. Specifically,

$$\frac{\nabla^2_\theta p(\mathbf{y}_i \mid \boldsymbol{\theta})}{p(\mathbf{y}_i \mid \boldsymbol{\theta})} \approx \frac{1}{S_\alpha}\sum_{s=1}^{S_\alpha}\left(\nabla_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i^{(s)} \mid \boldsymbol{\theta})\nabla_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i^{(s)} \mid \boldsymbol{\theta})^\top + \nabla^2_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i^{(s)} \mid \boldsymbol{\theta})\right),$$

where $\boldsymbol{\alpha}_i^{(s)} \sim p(\boldsymbol{\alpha}_i \mid \mathbf{y}_i, \boldsymbol{\theta})$ for $s = 1, \ldots, S_\alpha$. If obtaining samples from $p(\boldsymbol{\alpha}_i \mid \mathbf{y}_i, \boldsymbol{\theta})$ is not straightforward, importance sampling (as in Sect. 2.3.1) can be used instead. Following Nemeth et al. (2016), for computational efficiency, we use the same samples $\{\boldsymbol{\alpha}_i^{(s)} : s = 1, \ldots, S_\alpha\}$ that were used to approximate the score using Fisher's identity and their corresponding normalised weights $\mathcal{W}_i$ to obtain the estimates of the second term in Louis' identity. Then

$$\frac{\nabla^2_\theta p(\mathbf{y}_i \mid \boldsymbol{\theta})}{p(\mathbf{y}_i \mid \boldsymbol{\theta})} \approx \sum_{s=1}^{S_\alpha}\bar{w}_i^{(s)}\left(\nabla_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i^{(s)} \mid \boldsymbol{\theta})\nabla_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i^{(s)} \mid \boldsymbol{\theta})^\top + \nabla^2_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i^{(s)} \mid \boldsymbol{\theta})\right).$$

2.4 Damped R-VGAL

A possible problem with R-VGAL is its instability in the first few observations, making it sensitive to the ordering of the observations. In Section S3 of the online supplement, we run the R-VGAL algorithm on a dataset in its original order, and also on a random reordering of the observations, and find that the R-VGAL parameter estimates from these two runs differ. Figures S13 and S14 in Section S3 show that the first few observations can heavily influence the trajectory of the variational mean. Here, we propose a damping approach to stabilise the R-VGAL algorithm during the initial few steps.

In damped R-VGAL, the updates of the mean and precision matrix for each observation are split into K steps, where K is selected on a case by case basis. In each step, we multiply the gradient and the Hessian of $\log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ by a factor $a = \frac{1}{K}$ (which acts as a "step size"), and then update the variational parameters K times during the ith iteration. Intuitively, in this way, one observation is split into K "parts" and incorporated into the updates one part at a time. Using a smaller step size helps stabilise the R-VGAL algorithm, particularly for the first few observations. Section S3 of the online supplement shows that damping the first few iterations makes the R-VGAL algorithm more robust to different orderings of the data.

The damped R-VGAL approach we present here is inspired by the so-called *damped Newton's method*. In the case where the model is linear and the likelihood is Gaussian, the original R-VGA algorithm, upon which R-VGAL is based, can be shown to be equivalent to an online version of Newton's method; see Appendix 8.2 of Lambert et al. (2022) for a proof. Newton's method seeks the minimiser of a continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}, d \in \mathbb{N}$, by beginning with some starting value $\mathbf{u}_0 \in \mathbb{R}^d$ and sequentially minimising the quadratic approximation of the function $f(\cdot)$ around the current value in order to find the next value:

$$\mathbf{u}_{k+1} = \arg\min_{\mathbf{u}} f(\mathbf{u}_k) + \nabla_{\mathbf{u}} f(\mathbf{u}_k)^\top(\mathbf{u} - \mathbf{u}_k) + \frac{1}{2}(\mathbf{u} - \mathbf{u}_k)^\top \nabla^2_{\mathbf{u}} f(\mathbf{u}_k)(\mathbf{u} - \mathbf{u}_k), \quad k = 0, 1, 2, \ldots.$$

Provided that $\nabla^2 f(\mathbf{u}_k)$ is positive definite, the minimiser of $f(\cdot)$ is unique and can be computed iteratively as

$$\mathbf{u}_{k+1} = \mathbf{u}_k - (\nabla^2_{\mathbf{u}} f(\mathbf{u}_k))^{-1}\nabla_{\mathbf{u}} f(\mathbf{u}_k), \quad k = 0, 1, 2, \ldots. \tag{24}$$

These iterations stop when $\|\nabla_{\mathbf{u}} f(\mathbf{u}_{k+1})\| \leq \epsilon_0$, where ϵ_0 is some small tolerance parameter. Often, in practice, Newton's method is modified to include a step size $0 < \rho \leq 1$ to improve convergence:

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \rho(\nabla^2_{\mathbf{u}} f(\mathbf{u}_k))^{-1}\nabla_{\mathbf{u}} f(\mathbf{u}_k), \quad k = 0, 1, 2, \ldots, \tag{25}$$

resulting in the damped Newton's method. This step size ρ is similar to the multiplicative factor a in our damped R-VGAL approach.

We also note that, in the case where the model is linear or when the likelihood function comes from an exponential family and the model is linearised, the R-VGA algorithm of Lambert et al. (2022) is equivalent to an online natural gradient algorithm with step size $\frac{1}{1+t}$, where t denotes the iteration. A proof of this equivalence can be found in Appendix 8.3 of Lambert et al. (2022). Viewed from the perspective of natural gradient optimisation, the damping factor a in damped R-VGAL can be interpreted as a reduction of the step size in natural gradient updates.

We summarise the damped R-VGAL algorithm in Algorithm 2.

Algorithm 2 Damped R-VGAL

```
Input: observations $\mathbf{y}_1, \ldots, \mathbf{y}_N$, initial values $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$, number of observations to damp $n_{damp}$, number of damping steps $K$.
Output: variational parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, for $i = 1, \ldots, N$.
Set $q_0(\boldsymbol{\theta}) = N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.
for $i = 1, \ldots, N$ do
    if $i \le n_{damp}$ then
        Set $a = 1/K$, $\boldsymbol{\mu}_{i,0} = \boldsymbol{\mu}_{i-1}$, $\boldsymbol{\Sigma}_{i,0} = \boldsymbol{\Sigma}_{i-1}$
        for $k = 1, \ldots, K$ do
            Set $q_{i,k-1}(\boldsymbol{\theta}) = N(\boldsymbol{\mu}_{i,k-1}, \boldsymbol{\Sigma}_{i,k-1})$.
            $\boldsymbol{\mu}_{i,k} = \boldsymbol{\mu}_{i,k-1} + a\boldsymbol{\Sigma}_{i,k}\widehat{\mathbb{E}_{q_{i,k-1}}(\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}))}$
            $\boldsymbol{\Sigma}_{i,k}^{-1} = \boldsymbol{\Sigma}_{i,k-1}^{-1} - a\widehat{\mathbb{E}_{q_{i,k-1}}(\nabla_\theta^2 \log p(\mathbf{y}_i \mid \boldsymbol{\theta}))}$
        end for
        Set $\boldsymbol{\mu}_i = \boldsymbol{\mu}_{i,K}$, $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_{i,K}$, $q_i(\boldsymbol{\theta}) = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.
    else
        $\boldsymbol{\mu}_i = \boldsymbol{\mu}_{i-1} + \boldsymbol{\Sigma}_i\widehat{\mathbb{E}_{q_{i-1}}(\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}))}$
        $\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{\Sigma}_{i-1}^{-1} - \widehat{\mathbb{E}_{q_{i-1}}(\nabla_\theta^2 \log p(\mathbf{y}_i \mid \boldsymbol{\theta}))}$
    end if
end for
```

3 Applications of R-VGAL

In this section, we apply R-VGAL to estimate parameters in linear, logistic and Poisson mixed models using three simulated datasets and two real datasets: the Six City dataset from Fitzmaurice and Laird (1993), and the Polypharmacy dataset from Hosmer et al. (2013). The linear and logistic models have univariate random effects, while the Poisson model has bivariate random effects. There are two additional examples in Section S6 of the online supplement: a real data example with the Poisson model applied to the Epilepsy dataset from Thall and Vail (1990), and a synthetic data example with a high number of observations simulated from the logistic mixed model.

We validate R-VGAL against Hamiltonian Monte Carlo (HMC, Neal 2011; Betancourt and Girolami 2015), which is implemented using the Stan programming language (Stan Development Team 2023) in R (Core 2022). In examples with real data, the true parameters are unknown. We instead compute the maximum likelihood estimates for the parameters using the R package `lme4` (Bates et al. 2015), and also treat results from HMC as the "ground truth", as HMC provides samples from the true posterior distributions. For all examples, we run 2 HMC chains for 15,000 iterations each, and discard the first 5000 from each chain as burn in. We find that the effective sample sizes are high and the $\hat{R}$ statistics are close to 1 for all examples, indicating that the HMC chains are well-mixed and have converged; see Section S5 of the online supplement for further details. Reproducible R code for all examples is available on https://github.com/bao-anh-vu/R-VGAL.

For all applications in this paper, we use the damped R-VGAL algorithm described in Sect. 2.4. We show that damping makes the algorithm more robust to different orderings of the observations in Section S3 of the online supplement. The values of n_{damp} and K used in damping observations should be kept as small as possible to limit the extra computational overhead, while also be sufficiently large to reduce the instability observed with the R-VGAL algorithm in the initial stages. In our applications, we experimented with a few different settings of n_{damp} and K and plotted the trajectories of the variational mean under those settings. We found that the trajectories were most unstable during the first 10 observations, so we chose $n_{damp} = 10$ observations and the number of steps $K = 4$ to reduce the initial instability at the expense of a small additional computational cost. These values are used throughout our examples. Adaptive schemes for selecting the values of n_{damp} and K are left as future research directions.

3.1 Linear mixed effect model

In this example, we generate data from a linear mixed model with $N = 200$ groups and $n = 10$ responses per group. The jth response from the ith group is modelled as

$$y_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + z_{ij}\alpha_i + \epsilon_{ij}, \quad \alpha_i \sim N(0, \sigma_\alpha^2),$$
$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2), \tag{26}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, n$, where $\mathbf{x}_{ij}$ is drawn from a $N(\mathbf{0}, \mathbf{I}_4)$ distribution and z_{ij} is drawn from a $N(0, 1)$ distribution. For this example, we did not include an intercept term, but it can be added if necessary. The true parameter values are $\boldsymbol{\beta} = (-1.5, 1.5, 0.5, 0.25)^\top$, $\sigma_\alpha = 0.9$, and $\sigma_\epsilon = 0.7$. Since R-VGAL uses a multivariate normal distribution as the variational approximation, we consider the log-transformed variables $\phi_\alpha \equiv \log(\sigma_\alpha^2)$ and $\phi_\epsilon \equiv \log(\sigma_\epsilon^2)$ so that ϕ_α and ϕ_ϵ are unconstrained. We then make inference on the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi_\alpha, \phi_\epsilon)^\top$ using R-VGAL.

At the group level, the linear mixed model is

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{z}_i\alpha_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, N, \tag{27}$$

where $\mathbf{y}_i \equiv (y_{i1}, \ldots, y_{in})^\top$, $\mathbf{X}_i \equiv (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in})^\top$, $\mathbf{z}_i \equiv (z_{i1}, \ldots, z_{in})^\top$, and $\boldsymbol{\varepsilon}_i \equiv (\epsilon_{i1}, \ldots, \epsilon_{in})^\top$. At each iteration, $i = 1, \ldots, N$, the R-VGAL algorithm makes use of the "partial" likelihood of the observations from the ith group, $p(\mathbf{y}_i \mid \boldsymbol{\theta}) = N(\boldsymbol{\mu}_{y|\theta}, \boldsymbol{\Sigma}_{y|\theta})$, where $\boldsymbol{\mu}_{y|\theta} = \mathbf{X}_i\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{y|\theta} = \sigma_\alpha^2\mathbf{z}_i\mathbf{z}_i^\top + \sigma_\epsilon^2\mathbf{I}_n$. For this model, the gradient and Hessian of $\log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ with respect to each of the parameters are available in closed form; see Section S1.1 of the online supplement. In this case, we are therefore able to compare the accuracy of R-VGAL implemented using approximate gradients and Hessians with that of R-VGAL implemented using exact gradients and Hessians.

The prior distribution we use, which is also the "initial" variational distribution, is

$$p(\boldsymbol{\theta}) = q_0(\boldsymbol{\theta}) = N\left(\begin{bmatrix}\mathbf{0}\\ \log(0.5^2)\\ \log(0.5^2)\end{bmatrix}, \begin{bmatrix}10\mathbf{I}_4 & \mathbf{0} & \mathbf{0}\\ \mathbf{0}^\top & 1 & 0\\ \mathbf{0}^\top & 0 & 1\end{bmatrix}\right). \tag{28}$$

A $N(\log(0.5^2), 1)$ prior distribution for ϕ_α and ϕ_ϵ is equivalent to a log-normal prior distribution with mean 0.41 and variance 0.29 for both σ_α^2 and σ_ϵ^2. Using this prior distribution, the 2.5th and 97.5th percentiles for both σ_α^2 and σ_ϵ^2 are $(0.035, 1.775)$.

At each iteration $i = 1, \ldots, 200$, we use $S_\alpha = 100$ Monte Carlo samples (of α_i) to approximate the gradient and Hessian of $\log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ using Fisher's and Louis' identities. We use $S = 100$ Monte Carlo samples (of $\boldsymbol{\theta}$) to approximate the expectations with respect to $q_{i-1}(\boldsymbol{\theta})$ in the R-VGAL updates of the mean and precision matrix. These values were chosen based on an experimental study on the effect of S and S_α on the posterior estimates of R-VGAL in Section S2 of the online supplement.

We validate R-VGAL against HMC, which we implemented in Stan. Figure 1 shows the marginal posterior distributions of the parameters, along with bivariate posterior distributions as estimated using R-VGAL with approximate gradients and Hessians, R-VGAL with exact gradients and Hessians, and HMC. The posterior distributions obtained using R-VGAL are clearly very similar to those obtained using HMC, irrespective of whether exact or approximate gradients and Hessians are used.

3.2 Logistic mixed effect model

In this example, we generate simulated data from a random effects logistic regression model with $N = 500$ groups and $n = 10$ responses per group. The random effect logistic regression model we use is

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}(\pi_{ij}),\\ \pi_{ij} &= p(y_{ij} = 1 \mid \boldsymbol{\beta}, \tau^2)\\ &= \frac{\exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i)}{1 + \exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i)},\\ \alpha_i &\sim N(0, \tau^2), \end{aligned} \tag{29}$$

where $\mathbf{x}_{ij}$ is drawn from a $N(\mathbf{0}, \mathbf{I}_4)$ distribution, for $i = 1, \ldots, N$ and $j = 1, \ldots, n$. For this example, we did not include an intercept term, but it can be added if necessary. The true parameter values are $\boldsymbol{\beta} = (-1.5, 1.5, 0.5, 0.25)^\top$ and $\tau = 0.9$.

As in the linear case, although the parameters of the model are $\boldsymbol{\beta}$ and τ, we work with $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi_\tau)^\top$ where $\phi_\tau \equiv \log(\tau^2)$. The gradient and Hessian of the "partial" log-likelihood $\log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ in this model are not analytically tractable, but can be estimated unbiasedly using Fisher's and Louis' identities as discussed in Sects. 2.3.1 and 2.3.2. These identities require the expressions for $\nabla_\theta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})$ and $\nabla_\theta^2 \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})$, which are provided in Section S1.2 of the online supplement.

The prior distribution we use, which is also the "initial" variational distribution, is

$$p(\boldsymbol{\theta}) = q_0(\boldsymbol{\theta}) = N\left(\begin{bmatrix}\mathbf{0}\\ \log(0.5^2)\end{bmatrix}, \begin{bmatrix}10\mathbf{I}_4 & \mathbf{0}\\ \mathbf{0}^\top & 1\end{bmatrix}\right). \tag{30}$$

A $N(\log(0.5^2), 1)$ prior distribution for ϕ_τ is equivalent to a log-normal prior distribution with mean 0.41 and variance 0.29 for τ^2. The prior 2.5th and 97.5th percentiles for τ^2 are $(0.035, 1.775)$. At each iteration $i = 1, \ldots, 500$, we use $S_\alpha = 100$ Monte Carlo samples (of α_i) to approximate the gradient and Hessian of $\log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ during importance sampling, and $S = 100$ samples (of $\boldsymbol{\theta}$) to approximate the expectations with respect to $q_{i-1}(\boldsymbol{\theta})$ in the R-VGAL updates of the mean and precision matrix.

Figure 2 shows the marginal posterior distributions of the parameters, along with bivariate posterior distributions as estimated using R-VGAL and HMC. The posterior distributions obtained using R-VGAL are again very similar to those obtained using HMC.

3.3 Poisson mixed model

We now apply R-VGAL to a model with bivariate random effects. For this example, we simulate data with $N = 200$ groups and $n = 10$ responses per group from the following Poisson mixed effect regression model:

$$\begin{aligned} y_{ij} &\sim \text{Poisson}(\lambda_{ij}),\\ \lambda_{ij} &= \exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\alpha}_i),\\ \boldsymbol{\alpha}_i &\sim N(\mathbf{0}, \boldsymbol{\Sigma}_\alpha), \end{aligned}$$

where $\mathbf{x}_{ij} \equiv (1, x_{ij,1})^\top$, with $x_{ij,1}$ drawn from a $N(0, 1)$ distribution, and $\mathbf{z}_{ij} \equiv (1, z_{ij,1})^\top$, with $z_{ij,1}$ drawn from a $N(0, 1)$ distribution, for $i = 1, \ldots, N$ and $j = 1, \ldots, n$. We denote the fixed and random effects as $\boldsymbol{\beta} \equiv (\beta_0, \beta_1)^\top$ and $\boldsymbol{\alpha}_i \equiv (\alpha_{i,1}, \alpha_{i,2})^\top$, respectively. The true parameter values are

$$\boldsymbol{\beta} = (-1.5, -0.5)^\top, \quad \boldsymbol{\Sigma}_\alpha = \begin{bmatrix}0.15 & 0.05\\ 0.05 & 0.20\end{bmatrix}.$$

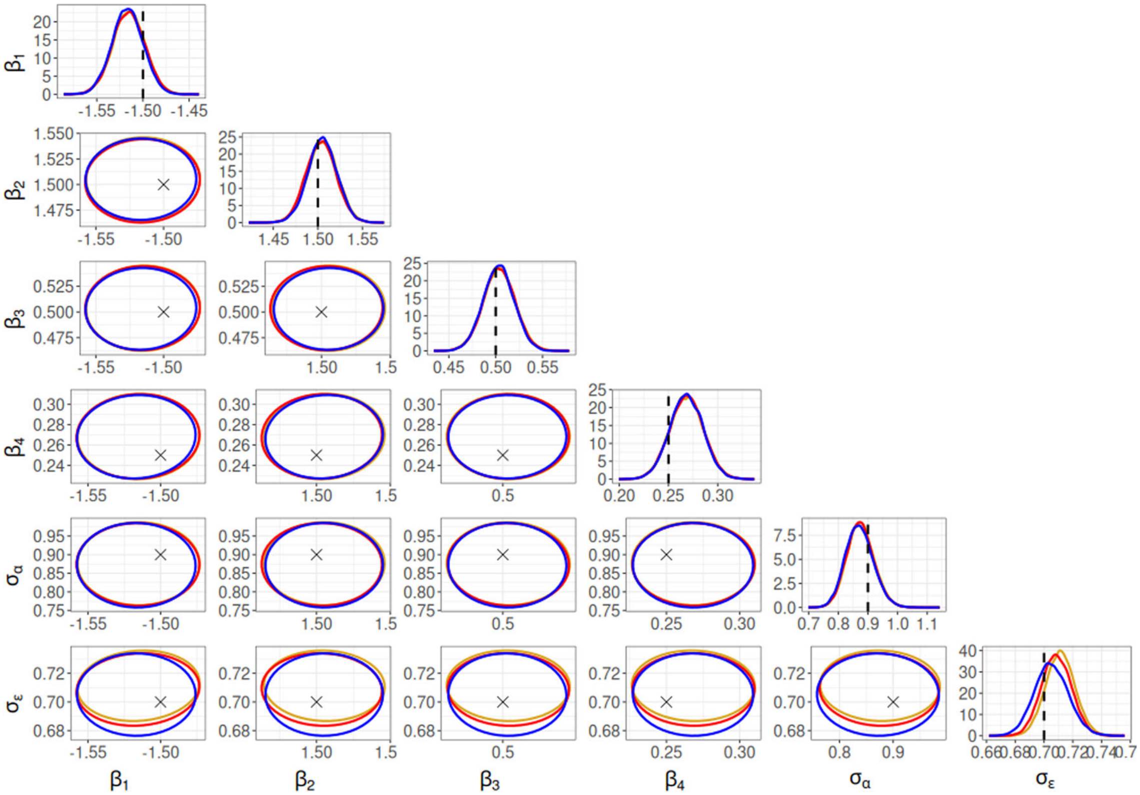

Fig. 1 Exact posterior distributions (from HMC, in blue) and approximate posterior distributions (from R-VGAL with estimated gradients and Hessians in red, and from R-VGAL with exact gradients and Hessians in yellow) for the linear mixed model experiment. Diagonal panels: Marginal posterior distributions with true parameters denoted using dotted lines. Off-diagonal panels: Bivariate posterior distributions with true parameters denoted using the symbol ×. (Color figure online)

We parameterise $\mathbf{\Sigma}_\alpha = \mathbf{L}\mathbf{L}^\top$, where $\mathbf{L}$ denotes the lower Cholesky factor of $\mathbf{\Sigma}_\alpha$ and takes the form

$$\mathbf{L} = \begin{bmatrix} \exp(\zeta_{11}) & 0 \\ \zeta_{21} & \exp(\zeta_{22}) \end{bmatrix}.$$

In the algorithm, we consider the unconstrained parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\zeta}^\top)^\top$, where $\boldsymbol{\zeta} \equiv (\zeta_{11}, \zeta_{22}, \zeta_{21})^\top$. The gradient $\nabla_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})$ and Hessian $\nabla^2_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})$, which are necessary in the computation of the gradient and Hessian of the group-specific log likelihood $\log p(\mathbf{y}_i \mid \boldsymbol{\theta})$, are provided in Section S1.3 of the online supplement.

We use the following prior/initial variational distribution:

$$p(\boldsymbol{\theta}) = q_0(\boldsymbol{\theta}) = N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{I}_2 & \mathbf{0} \\ \mathbf{0}^\top & 0.1\mathbf{I}_3\end{bmatrix}\right).$$

A $N(0, 0.1)$ prior distribution for ζ_{11}, ζ_{22} and ζ_{21} leads to having 2.5th and 97.5th percentiles of (0.290, 3.485) for $\mathbf{\Sigma}_{\alpha_{11}}$, (0.342, 3.577) for $\mathbf{\Sigma}_{\alpha_{22}}$, and (−0.713, 0.713) for the off-diagonal entries $\mathbf{\Sigma}_{\alpha_{21}}$ and $\mathbf{\Sigma}_{\alpha_{12}}$.

As with the linear and logistic examples, we use $S_\alpha = 100$ for the importance sampling step and $S = 100$ samples for approximating the expectations with respect to $q_{i-1}(\boldsymbol{\theta})$ in the R-VGAL updates. Figure 3 shows the marginal posterior distributions of the parameters, along with bivariate posterior distributions as estimated using R-VGAL and HMC. For all parameters, the R-VGAL and HMC posterior densities are very similar, though the posterior densities of $\mathbf{\Sigma}_{\alpha_{11}}$ from both methods appear a bit biased.

To assess the robustness of the results in these simulation studies, we also include repeated simulation studies on the linear, logistic and Poisson mixed models in Section S4 of the online supplement. For each of these models, we simulate 100 datasets using the same parameter settings, and compare the posterior estimates from R-VGAL and HMC on these simulated datasets. We find that the R-VGAL and HMC posterior estimates are very similar across simulations for the linear and logistic models, while for the Poisson model, the estimates from the two methods are close for most simulations, with only a few cases where estimates are slightly different. We also find that the posterior standard deviations from R-VGAL tend to be slightly smaller than those from HMC.

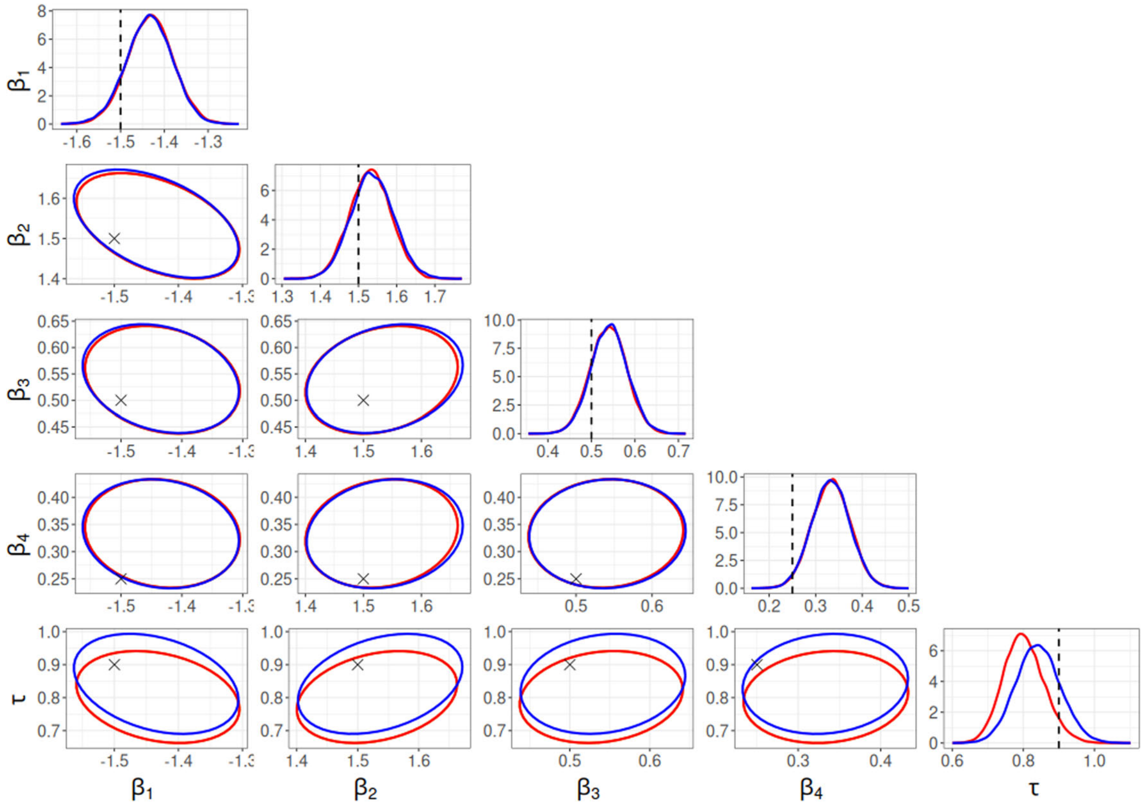

Fig. 2 Exact posterior distributions from HMC (in blue) and approximate posterior distributions from R-VGAL with estimated gradients and Hessians (in red) for the logistic mixed model experiment. Diagonal panels: Marginal posterior distributions with true parameters denoted using dotted lines. Off-diagonal panels: Bivariate posterior distributions with true parameters denoted using the symbol ×. (Color figure online)

3.4 Real data examples

We now apply R-VGAL to two real datasets: the Six City dataset from Fitzmaurice and Laird (1993), and the Polypharmacy dataset from Hosmer et al. (2013).

For the Six City dataset, we follow Tran et al. (2017) and consider the random intercept logistic regression model

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_{age}\texttt{Age}_{ij} + \beta_{smoke}\texttt{Smoke}_{ij} + \alpha_i,$$
$$\alpha_i \sim N(0, \tau^2), \qquad (31)$$

where $\pi_{ij} \equiv p(y_{ij} = 1 \mid \boldsymbol{\beta}, \tau^2)$, with $\boldsymbol{\beta} \equiv (\beta_0, \beta_{age}, \beta_{smoke})^\top$, for $i = 1, \ldots, 537$ and $j = 1, \ldots, 4$. The binary response variable $y_{ij} = 1$ if child i is wheezing at time point j, and 0 otherwise. The covariate $\texttt{Age}_{ij}$ is the age of child i at time point j, centred at 9 years, while the covariate $\texttt{Smoke}_{ij} = 1$ if the mother of child i is smoking at time point j, and 0 otherwise. Finally, α_i is the random effect associated with the ith child. The parameters of the model are $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi_\tau)^\top$, where $\phi_\tau \equiv \log(\tau^2)$.

For the Polypharmacy dataset, we consider the random intercept logistic regression model from Tan and Nott (2018):

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_{gender}\texttt{Gender}_i + \beta_{race}\texttt{Race}_i$$
$$+ \beta_{age}\texttt{Age}_{ij} + \beta_{M1}\texttt{MHV1}_{ij} + \beta_{M2}\texttt{MHV2}_{ij}$$
$$+ \beta_{M3}\texttt{MHV3}_{ij} + \beta_{IM}\texttt{INPTMHV}_{ij} + \alpha_i,$$
$$\alpha_i \sim N(0, \tau^2), \qquad (32)$$

where $\pi_{ij} \equiv \Pr(y_{ij} = 1 \mid \boldsymbol{\beta}, \tau^2)$, $\boldsymbol{\beta} \equiv (\beta_0, \beta_{gender}, \beta_{race}, \beta_{age}, \beta_{M1}, \beta_{M2}, \beta_{M3}, \beta_{IM})^\top$, for $i = 1, \ldots, 500$ and $j = 1, \ldots, 7$. The response variable y_{ij} is 1 if subject i in year j is taking drugs from three or more different classes (of drugs), and 0 otherwise. The covariate $\texttt{Gender}_i = 1$ if subject i is male, and 0 if female, while $\texttt{Race}_i = 0$ if the race of subject i is white, and 1 otherwise. The covariate $\texttt{Age}_{ij}$ is the age (in years and months, to two decimal places) of subject i in year j. The number of outpatient mental health visits (MHV) for subject i in year j is split into three dummy variables: $\texttt{MHV1}_{ij} = 1$ if $1 \leq \texttt{MHV}_{ij} \leq 5$, and 0 otherwise; $\texttt{MHV2}_{ij} = 1$ if $6 \leq \texttt{MHV}_{ij} \leq 14$, and 0 otherwise; and $\texttt{MHV3}_{ij} = 1$ if $\texttt{MHV}_{ij} \geq 15$, and 0 otherwise. The covariate

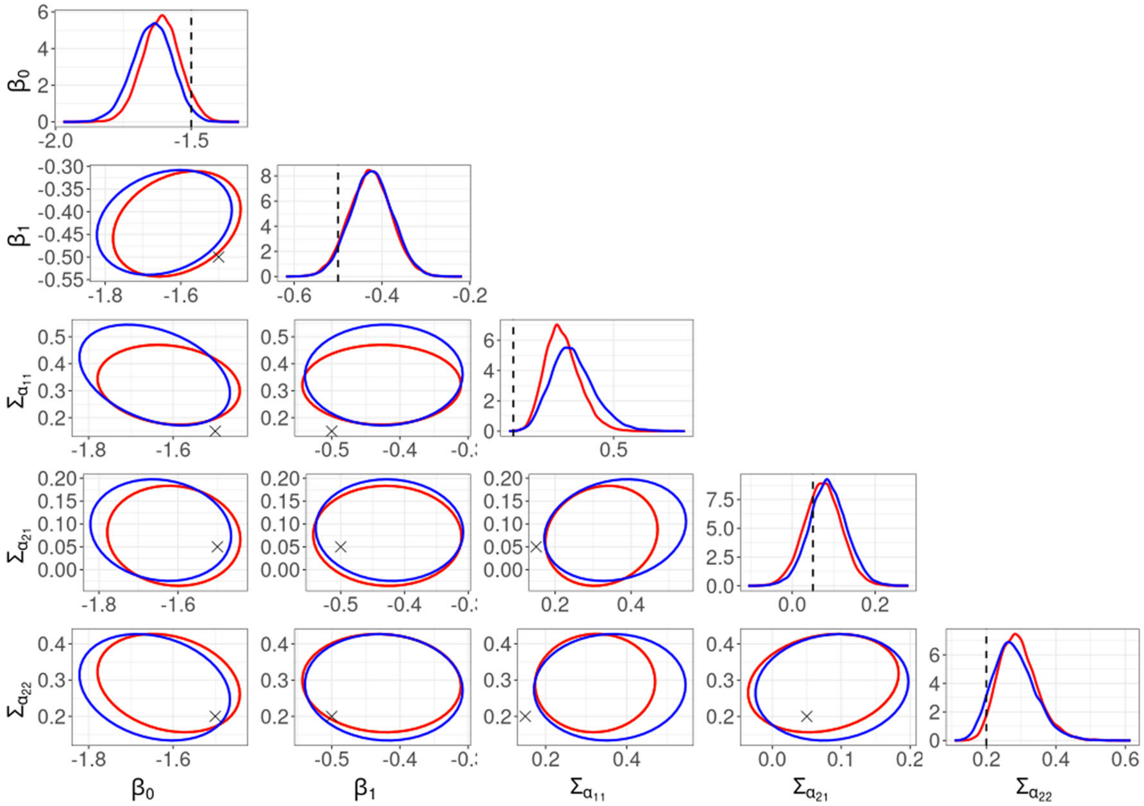

Fig. 3 Exact posterior distributions from HMC (in blue) and approximate posterior distributions from R-VGAL with estimated gradients and Hessians (in red) for the Poisson mixed model experiment. Diagonal panels: Marginal posterior distributions with true parameters denoted using dotted lines. Off-diagonal panels: Bivariate posterior distributions with true parameters denoted using the symbol ×. (Color figure online)

$\texttt{INPTMHV}_{ij} = 0$ if there were no inpatient mental health visits for subject i in year j, and 1 otherwise. Finally, α_i is a subject-level random effect for subject i. The parameters of the model are $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi_\tau)^\top$, where $\phi_\tau \equiv \log(\tau^2)$.

We use similar priors/initial variational distributions for both examples. For the Six City dataset, the prior/initial variational distribution we use is

$$p(\boldsymbol{\theta}) = q_0(\boldsymbol{\theta}) = N\left(\begin{bmatrix}\mathbf{0}\\ 1\end{bmatrix}, \begin{bmatrix}10\mathbf{I}_3 & \mathbf{0}\\ \mathbf{0}^\top & 1\end{bmatrix}\right), \tag{33}$$

and for the Polypharmacy dataset, we use

$$p(\boldsymbol{\theta}) = q_0(\boldsymbol{\theta}) = N\left(\begin{bmatrix}\mathbf{0}\\ 1\end{bmatrix}, \begin{bmatrix}10\mathbf{I}_8 & \mathbf{0}\\ \mathbf{0}^\top & 1\end{bmatrix}\right). \tag{34}$$

A $N(1, 1)$ prior distribution for ϕ_τ leads to a log-normal prior distribution with mean 4.48 and variance 34.51 for τ^2. Using this prior distribution, the 2.5th and 97.5th percentiles for τ^2 are (0.383, 19.297), which cover most values of τ^2 in practice. At each R-VGAL iteration, the gradient and Hessian of $\log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ are approximated using $S_\alpha = 200$ Monte Carlo samples (of α_i), and the expectations with respect to $q_{i-1}(\boldsymbol{\theta})$ in the R-VGAL updates are approximated using $S = 200$ Monte Carlo samples (of $\boldsymbol{\theta}$).

As there are no ground truths to these examples, we compare the posterior density estimates from R-VGAL to those from HMC. In addition, we also compute the maximum likelihood estimates using the `lme4` package in R. Figures 4 and 5 show the marginal posterior distributions with maximum likelihood estimates of the parameters, along with bivariate posterior distributions estimated using R-VGAL and HMC for the Six City and Polypharmacy datasets, respectively. In the Six City example, there is a slight difference in the marginal and bivariate posterior densities from R-VGAL and HMC for the fixed effect β_{smoke}, but the posterior densities for other parameters are very similar between the two methods. For the intercept β_0 and the random effect standard deviation τ, the posterior modes of HMC are closer to the maximum likelihood estimates than the posterior modes of R-VGAL, but for the other parameters, the posterior modes from both R-VGAL and HMC are close to the maximum likelihood estimates. For the Polypharmacy example, there are slight differences between the R-VGAL and HMC marginal and bivariate posterior densities for the intercept β_0

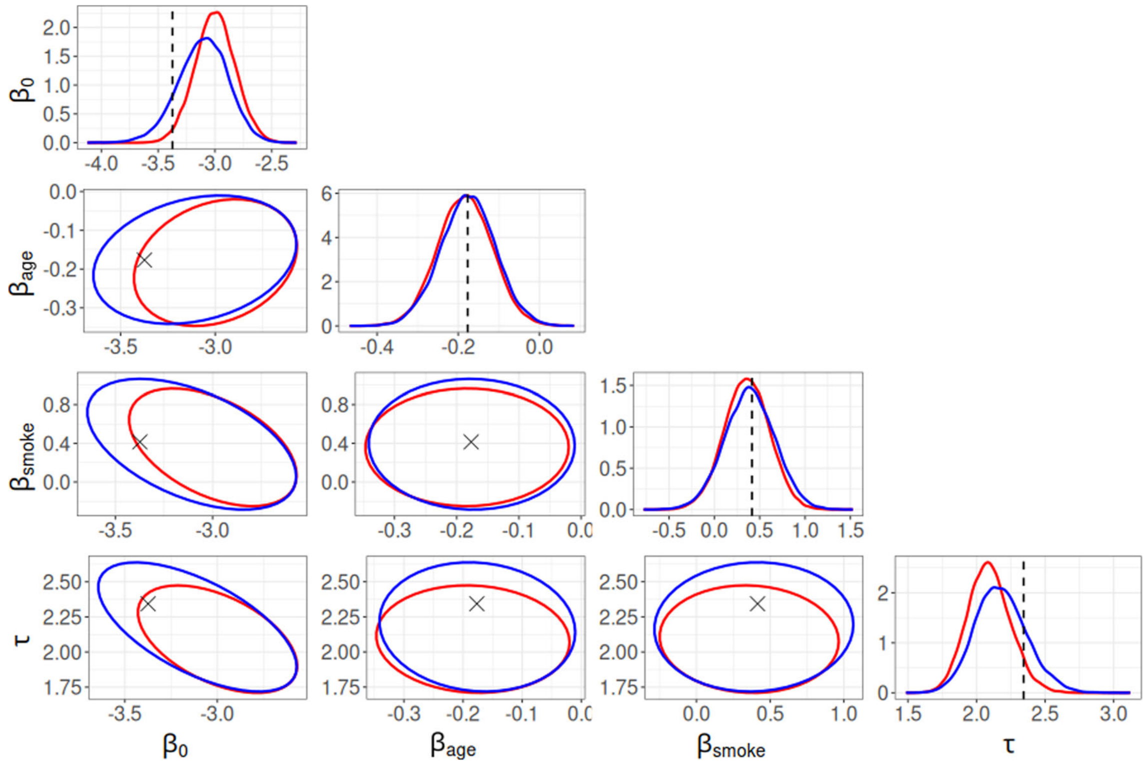

Fig. 4 Exact posterior distributions from HMC (in blue) and approximate posterior distributions from R-VGAL with estimated gradients and Hessians (in red) for the experiment with the Six City dataset. Diagonal panels: Marginal posterior distributions with the maximum likelihood estimates marked using dotted lines. Off-diagonal panels: Bivariate posterior distributions with the maximum likelihood estimates marked using the symbol ×. (Color figure online)

and the fixed effects β_{gender} and β_{race}, but for other parameters, the posterior densities are comparable between the two methods. The posterior modes of both R-VGAL and HMC are close to the maximum likelihood estimates for all parameters in this example.

3.5 Computing time

Table 1 compares the computing time (in minutes) of R-VGAL and HMC for all simulated and real data examples that we have discussed in Sect. 3 and Section S6 of the online supplement, and includes the corresponding dataset size for each example. The last column in the table shows the average time taken (in seconds) for a single iteration of R-VGAL. For the linear example, where we run R-VGAL with both the theoretical and estimated gradients/Hessians, the displayed time is that of R-VGAL with the estimated gradients/Hessians. All experiments were carried out on the High Performance Computer system of the National Institute for Applied Statistics Research Australia, with an NVIDIA Tesla V100 PCIe 32GB graphics processing unit (GPU). The GPU was used to parallelise the computations in the importance sampling step, so that the gradient and Hessian of the joint log-likelihood $\log p(\mathbf{y}_i, \boldsymbol{\alpha}_i^{(s)} \mid \boldsymbol{\theta})$, $s = 1, \ldots, S_\alpha$, and their corresponding weights $\mathcal{W}_i$, are computed all at once. The GPU was also used to parallelise over the Monte Carlo samples used in the estimation of the expectations with respect to $q_{i-1}(\cdot)$ in Algorithm 1. We use the R interface to Tensorflow (Abadi et al. 2015) to facilitate GPU computations.

The table shows that the R-VGAL algorithm is generally 3 to 8 times faster than HMC. This is substantial given that our code is not as highly optimised as that in Stan. The difference in computing times also becomes more notable with a bigger dataset: in the logistic example with 50000 synthetic observations (see Section S6.2 of the online supplement), R-VGAL takes only 17 min to produce posterior estimates, while HMC takes more than 2 h. Furthermore, since R-VGAL is a sequential algorithm, posterior approximations from R-VGAL can be easily updated as new observations become available. To incorporate an additional observation, R-VGAL needs to perform a single update, while an algorithm like HMC requires rerunning the entire sampling procedure.

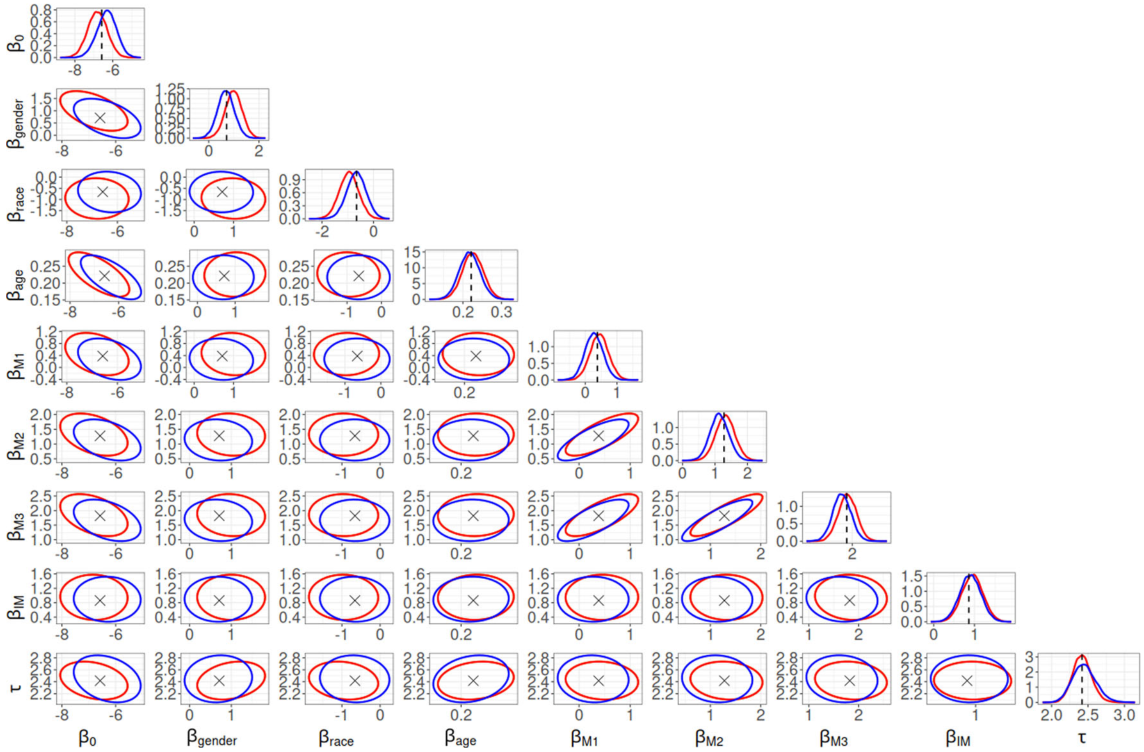

Fig. 5 Exact posterior distributions from HMC (in blue) and approximate posterior distributions from R-VGAL with estimated gradients and Hessians (in red) for the experiment with the Polypharmacy dataset. Diagonal panels: Marginal posterior distributions with the maximum likelihood estimates marked using dotted lines. Off-diagonal panels: Bivariate posterior distributions with the maximum likelihood estimates marked using the symbol ×. (Color figure online)

Table 1 Computing time (in minutes) for the R-VGAL and HMC methods for the simulated and real datasets, with accompanying dataset sizes

	N	n	HMC (min)	R-VGAL (min)	One R-VGAL iteration (s)
Linear (simulated data)	200	10	2.5	0.6	0.17
Logistic (simulated data)	500	10	7.2	1.1	0.13
Poisson (simulated data)	200	10	11.3	3.1	1.05
Logistic (Six City)	537	4	3.4	1.2	0.13
Logistic (Polypharmacy)	500	7	18.5	2.4	0.29
Poisson (Epilepsy)*	59	4	3.3	1.2	1.25
Logistic (simulated data)*	5000	10	133.6	16.8	0.20

Timings for one R-VGAL update is shown (in seconds). Examples with the * symbol are in the online supplement

4 Conclusion

In this article, we propose a sequential variational Bayes algorithm for estimating parameters in GLMMs based on an extension of the R-VGA algorithm of Lambert et al. (2022). The original R-VGA algorithm requires the gradient and Hessian of the partial log-likelihood at each observation, which are computationally intractable for most GLMMs. To overcome this, we use Fisher's and Louis' identities to obtain unbiased estimates of the gradient and Hessian, which can be used in place of the closed form gradient and Hessian in the R-VGAL algorithm.

We apply R-VGAL to the linear, logistic and Poisson mixed effect models with simulated and real datasets. In all examples, we compare the posterior distributions of the parameters estimated using R-VGAL to those obtained using HMC (Neal 2011; Betancourt and Girolami 2015). The examples show that R-VGAL yields comparable posterior estimates to HMC while being substantially faster, and the R-VGAL posterior modes are very close to the maximum likelihood estimates for most parameters in the models we consider. R-VGAL would be especially useful in situations where new observations are being continuously collected.

In the current paper, we assume that the random effects are independent and identically distributed between subjects or groups. We discuss the potential application of R-VGAL to models with more complicated random effect structures, such as crossed or nested effects, in Section S7 of the online supplement. Future work will attempt to extend R-VGAL to cases where the random effects are temporally correlated. This will expand the set of models on which R-VGAL can be used to include time series and state space models.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-024-10422-8.

Acknowledgements This work was supported by ARC SRIEAS Grant SR200100005 Securing Antarctica's Environmental Future. Andrew Zammit-Mangion's research was also supported by ARC Discovery Early Career Research Award DE180100203. The authors would also like to thank two anonymous reviewers for their insightful and constructive comments.

Author Contributions B.A.V. and D.G. conceived the methodology, B.A.V. wrote the manuscript and software in close collaboration with both D.G. and A.Z.-M. All authors reviewed the manuscript.

Declarations

Competing interests The authors declare no competing interests.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: large-scale machine learning on heterogeneous systems. Software available from tensorflow.org

Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. J. Stat. Softw. **67**(1), 1–48 (2015)

Betancourt, M., Girolami, M.: Hamiltonian Monte Carlo for hierarchical models. Curr. Trends Bayesian Methodol. Appl. **79**(30), 2–4 (2015)

Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017)

Bonnet, G.: Transformations des signaux aléatoires a travers les systemes non linéaires sans mémoire. Annales des Télécommunications **19**, 203–220 (1964)

Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics, pp. 177–186. Springer, New York (2010)

Breslow, N.E., Clayton, D.G.: Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. **88**(421), 9–25 (1993)

Broderick, T., Boyd, N., Wibisono, A., Wilson, A.C., Jordan, M.I. (2013). Streaming variational Bayes. In: Proceedings of the 26th International Conference on Neural Information Processing Systems—Volume 2, NIPS'13, pp. 1727–1735. Curran Associates Inc., Red Hook

Cappé, O., Moulines, E., Rydén, T.: Inference in Hidden Markov Models. Springer, New York (2005)

Crowder, M.J.: Inference about the intraclass correlation coefficient in the beta-binomial ANOVA for proportions. J. R. Stat. Soc. B **41**(2), 230–234 (1979)

Demidenko, E.: Mixed Models: Theory and Applications with R. Wiley, Hoboken (2013)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39**(1), 1–22 (1977)

Faraway, J.J.: Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, 2nd edn. CRC Press, New York (2016)

Fitzmaurice, G.M., Laird, N.M.: A likelihood-based method for analysing longitudinal binary responses. Biometrika **80**(1), 141–151 (1993)

Fong, Y., Rue, H., Wakefield, J.: Bayesian inference for generalized linear mixed models. Biostatistics **11**(3), 397–412 (2010)

Goldstein, H.: Nonlinear multilevel models, with an application to discrete response data. Biometrika **78**(1), 45–51 (1991)

Gunawan, D., Kohn, R., Nott, D.: Variational Bayes approximation of factor stochastic volatility models. Int. J. Forecast. **37**(4), 1355–1375 (2021)

Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. J. Mach. Learn. Res. **14**, 1303–1347 (2013)

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, 3rd edn. Wiley, Hoboken (2013)

Jansen, M.G.H.: Parameters of the latent distribution in Rasch's Poisson counts model. In: Fischer, G.H., Laming, D. (eds.) Contributions to Mathematical Psychology, Psychometrics, and Methodology, pp. 319–326. Springer, New York (1994)

Kingma, D.P., Welling, M.: Auto-encoding variational Bayes (2013). arXiv preprint arXiv:1312.6114

Lambert, M., Bonnabel, S., Bach, F.: The recursive variational Gaussian approximation (R-VGA). Stat. Comput. **32**(1) (2022)

Naylor, J.C., Smith, A.F.: Applications of a method for the efficient computation of posterior distributions. J. R. Stat. Soc. C **31**(3), 214–225 (1982)

Neal, R.: MCMC Using Hamiltonian dynamics. In: Brooks, S., Gelman, A., Jones, G.J., Meng, X.-L. (eds.) Handbook of Markov Chain Monte Carlo. CRC Press, Boca Raton (2011)

Nemeth, C., Fearnhead, P., Mihaylova, L.: Particle approximations of the score and observed information matrix for parameter estimation in state-space models with linear computational cost. J. Comput. Graph. Stat. **25**(4), 1138–1157 (2016)

Ong, V.M.-H., Nott, D.J., Smith, M.S.: Gaussian variational approximation with a factor covariance structure. J. Comput. Graph. Stat. **27**(3), 465–478 (2018)

Ormerod, J.T., Wand, M.P.: Explaining variational approximations. Am. Stat. **64**(2), 140–153 (2010)

Price, R.: A useful theorem for nonlinear devices having Gaussian inputs. IRE Trans. Inf. Theory **4**(2), 69–72 (1958)

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2022)

Stan Development Team: RStan: the R interface to Stan. R package version **2**(21), 8 (2023)

Tan, L.S., Nott, D.J.: Gaussian variational approximation with sparse precision matrices. Stat. Comput. **28**(2018), 259–275 (2018)

Thall, P.F., Vail, S.C.: Some covariance models for longitudinal count data with overdispersion. Biometrics **46**(3), 657–671 (1990)

Tierney, L., Kadane, J.B.: Accurate approximations for posterior moments and marginal densities. J. Am. Stat. Assoc. **81**(393), 82–86 (1986)

Tokdar, S.T., Kass, R.E.: Importance sampling: a review. WIREs Comput. Stat. **2**(1), 54–60 (2010)

Tomasetti, N., Forbes, C., Panagiotelis, A.: Updating variational Bayes: Fast sequential posterior inference. Stat. Comput. 32(1)

Tran, M.-N., Nott, D.J., Kuk, A.Y., Kohn, R.: Parallel variational Bayes for large datasets with an application to generalized linear mixed models. J. Comput. Graph. Stat. **25**(2), 626–646 (2016)

Tran, M.-N., Nott, D.J., Kohn, R.: Variational Bayes with intractable likelihood. J. Comput. Graph. Stat. **26**(4), 873–882 (2017)

Tuerlinckx, F., Rijmen, F., Verbeke, G., De Boeck, P.: Statistical inference in generalized linear mixed models: a review. Br. J. Math. Stat. Psychol. **59**(2), 225–255 (2006)

Verbeke, G., Molenberghs, G., Verbeke, G.: Linear Mixed Models for Longitudinal Data. Springer, New York (1997)

Wakefield, J.: Bayesian and Frequentist Regression Methods. Springer, New York (2013)

Zhao, Y., Staudenmayer, J., Coull, B.A., Wand, M.P.: General design Bayesian generalized linear mixed models. Stat. Sci. **21**(1), 35–51 (2006)