Online Supplement: "R-VGAL: A Sequential Variational Bayes Algorithm for Generalised Linear Mixed Models"

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

[1]School of Mathematics and Statistics, University of Wollongong, Wollongong, New South Wales, Australia

[2]Securing Antarctica's Environmental Future, University of Wollongong, Wollongong, New South Wales, Australia

January 20, 2024

* Corresponding author
E-mail address: bavu@uow.edu.au

S1 Appendix A: Gradient and Hessian derivations

S1.1 Derivation of gradient and Hessian expressions for the linear mixed model

In the linear mixed model, the theoretical gradient and Hessian for the likelihood at observation i for $i = 1, \ldots, N$ are available. This section gives the expressions for both the theoretical gradient and Hessian, as well as their approximations when using Fisher's and Louis' identities.

S1.1.1 Theoretical gradient and Hessian

The "partial" log-likelihood for observations from the ith group is given by

$$\log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{y|\theta}\right| - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_{y|\theta}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \quad i = 1, \ldots, N, \tag{S1}$$

where $\boldsymbol{\Sigma}_{y|\theta} = \exp(\phi_\alpha)\mathbf{z}_i\mathbf{z}_i^\top + \exp(\phi_\epsilon)\mathbf{I}_n$.

The gradient of the log-likelihood with respect to the parameters $\boldsymbol{\theta}$ is given by

$$\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = (\nabla_\beta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})^\top, \nabla_{\phi_\alpha} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}), \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}))^\top,$$

where the components are, respectively,

$$\nabla_\beta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = \mathbf{X}_i^\top \boldsymbol{\Sigma}_{y|\theta}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \tag{S2}$$

$$\nabla_{\phi_\alpha} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = -\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}_{y|\theta}^{-1}\frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha}\right) + \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_{y|\theta}^{-1}\frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha}\boldsymbol{\Sigma}_{y|\theta}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \tag{S3}$$

$$\nabla_{\phi_\epsilon} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = -\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}_{y|\theta}^{-1}\frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\epsilon}\right) + \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_{y|\theta}^{-1}\frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\epsilon}\boldsymbol{\Sigma}_{y|\theta}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \tag{S4}$$

with

$$\frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha} = \exp(\phi_\alpha)\mathbf{z}_i\mathbf{z}_i^\top, \quad \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\epsilon} = \exp(\phi_\epsilon)\mathbf{I}_n. \tag{S5}$$

The Hessian is given by

$$\nabla_\theta^2 \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = \begin{bmatrix} \nabla_\beta^2 \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) & \nabla_\beta \nabla_{\phi_\alpha} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) & \nabla_\beta \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) \\ \nabla_\beta \nabla_{\phi_\alpha} \log p(\mathbf{y}_i \mid \boldsymbol{\theta})^\top & \nabla_{\phi_\alpha}^2 \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) & \nabla_{\phi_\alpha} \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) \\ \nabla_\beta \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i \mid \boldsymbol{\theta})^\top & \nabla_{\phi_\alpha} \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) & \nabla_{\phi_\epsilon}^2 \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) \end{bmatrix}. \tag{S6}$$

The diagonal terms in the Hessian are the second derivatives

$$\nabla^2_{\beta} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = -\mathbf{X}_i^\top \boldsymbol{\Sigma}_{y|\theta}^{-1} \mathbf{X}_i, \tag{S7}$$

$$\nabla^2_{\phi_\alpha} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = -\frac{1}{2} \operatorname{Tr}(\mathbf{G}_{\phi_\alpha}) + \frac{1}{2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})^\top \mathbf{H}_{\phi_\alpha} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}), \tag{S8}$$

where

$$\mathbf{G}_{\phi_\alpha} = -\boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha} \boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha} + \boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial^2 \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha^2} \tag{S9}$$

$$\mathbf{H}_{\phi_\alpha} = -2\boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha} \boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha} \boldsymbol{\Sigma}_{y|\theta}^{-1} + \boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial^2 \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha^2} \boldsymbol{\Sigma}_{y|\theta}^{-1}, \tag{S10}$$

with $\frac{\partial^2 \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha^2} = \exp(\phi_\alpha) \mathbf{z}_i \mathbf{z}_i^\top$. The expression for $\nabla^2_{\phi_\epsilon} \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ is the same as in (S8), but with all derivatives with respect to ϕ_α replaced by those with respect to ϕ_ϵ. Note that $\frac{\partial^2 \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\epsilon^2} = \exp(\phi_\epsilon) \mathbf{I}_n$.

The off-diagonal terms in the Hessian are

$$\nabla_\beta \nabla_{\phi_\alpha} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = -\mathbf{X}_i^\top \boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha} \boldsymbol{\Sigma}_{y|\theta}^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})$$

$$\nabla_\beta \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = -\mathbf{X}_i^\top \boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\epsilon} \boldsymbol{\Sigma}_{y|\theta}^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})$$

$$\nabla_{\phi_\alpha} \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = -\frac{1}{2} \operatorname{Tr}(\mathbf{G}_{\phi_\alpha \phi_\epsilon}) + \frac{1}{2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})^\top \mathbf{H}_{\phi_\alpha \phi_\epsilon} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}),$$

where

$$\mathbf{G}_{\phi_\alpha \phi_\epsilon} = -\boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha} \boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\epsilon} \tag{S11}$$

$$\mathbf{H}_{\phi_\alpha \phi_\epsilon} = -\boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha} \boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\epsilon} \boldsymbol{\Sigma}_{y|\theta}^{-1} - \boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\epsilon} \boldsymbol{\Sigma}_{y|\theta}^{-1} \frac{\partial \boldsymbol{\Sigma}_{y|\theta}}{\partial \phi_\alpha} \boldsymbol{\Sigma}_{y|\theta}^{-1}. \tag{S12}$$

S1.1.2 Expressions for the gradient and Hessian using Fisher's and Louis' identities

Fisher's identity (19) requires the gradient $\nabla_\theta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})$, while Louis' identity (22) requires the Hessian $\nabla^2_\theta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})$. We now give the expression for these quantities for the linear mixed model considered in Section 3.1.

For $i = 1, \ldots, N$, the gradient $\nabla_\theta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})$ can be written as

$$\nabla_\theta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = \nabla_\theta \log p(\mathbf{y}_i \mid \alpha_i, \boldsymbol{\theta}) + \nabla_\theta \log p(\alpha_i \mid \boldsymbol{\theta}), \tag{S13}$$

$$= \nabla_\theta \log p(\mathbf{y}_i \mid \alpha_i, \boldsymbol{\beta}, \phi_\epsilon) + \nabla_\theta \log p(\alpha_i \mid \phi_\alpha), \tag{S14}$$

where

$$\log p(\mathbf{y}_i \mid \alpha_i, \boldsymbol{\beta}, \phi_\epsilon) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\exp(\phi_\epsilon)\mathbf{I}_n| - \frac{1}{2\exp(\phi_\epsilon)}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{z}_i\alpha_i)^\top(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{z}_i\alpha_i), \tag{S15}$$

and

$$\log p(\alpha_i \mid \phi_\alpha) = -\frac{1}{2}\log(2\pi) - \frac{\phi_\alpha}{2} - \frac{\alpha_i^2}{2\exp(\phi_\alpha)}. \tag{S16}$$

Thus, the gradient of the joint $\nabla_\theta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})$ is

$$\nabla_\theta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = \left(\nabla_\beta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})^\top, \nabla_{\phi_\alpha} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}), \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})\right)^\top, \tag{S17}$$

where each component is given by

$$\nabla_\beta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = \nabla_\beta \log p(\mathbf{y}_i \mid \alpha_i, \boldsymbol{\beta}, \phi_\epsilon)$$
$$= \frac{1}{\exp(\phi_\epsilon)}\mathbf{X}_i^\top(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{z}_i\alpha_i), \tag{S18}$$

$$\nabla_{\phi_\alpha} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = \nabla_{\phi_\alpha} \log p(\alpha_i \mid \phi_\alpha)$$
$$= -\frac{1}{2} + \frac{\alpha_i^2}{2\exp(\phi_\alpha)}, \tag{S19}$$

$$\nabla_{\phi_\epsilon} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i \mid \alpha_i, \boldsymbol{\beta}, \phi_\epsilon)$$
$$= -\frac{n}{2} + \frac{1}{2\exp(\phi_\epsilon)}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{z}_i\alpha_i)^\top(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{z}_i\alpha_i). \tag{S20}$$

Similarly, the approximation of the Hessian using Louis' identity requires $\nabla^2_\theta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})$, in particular,

$$\nabla^2_\theta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = \begin{bmatrix} \nabla^2_\beta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) & \nabla_\beta \nabla_{\phi_\alpha} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) & \nabla_\beta \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) \\ \nabla_\beta \nabla_{\phi_\alpha} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})^\top & \nabla^2_{\phi_\alpha} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) & \nabla_{\phi_\alpha} \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) \\ \nabla_\beta \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})^\top & \nabla_{\phi_\alpha} \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})^\top & \nabla^2_{\phi_\epsilon} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) \end{bmatrix}. \tag{S21}$$

The components of (S21) are given by

$$\nabla^2_{\beta} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = -\frac{1}{\exp(\phi_\epsilon)} \mathbf{X}_i^\top \mathbf{X}_i, \tag{S22}$$

$$\nabla^2_{\phi_\alpha} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = -\frac{\alpha_i^2}{2\exp(\phi_\alpha)}, \tag{S23}$$

$$\nabla^2_{\phi_\epsilon} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = -\frac{1}{2\exp(\phi_\epsilon)} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{z}_i\alpha_i)^\top (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{z}_i\alpha_i), \tag{S24}$$

$$\nabla_{\beta} \nabla_{\phi_\alpha} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = 0, \tag{S25}$$

$$\nabla_{\beta} \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = -\frac{1}{\exp(\phi_\epsilon)} \mathbf{X}_i^\top (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{z}_i\alpha_i), \tag{S26}$$

$$\nabla_{\phi_\alpha} \nabla_{\phi_\epsilon} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = 0. \tag{S27}$$

S1.2 Derivation of gradient and Hessian expressions for the logistic mixed model

The parameters of interest are $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi_\tau)^\top$, where $\phi_\tau = \log(\tau^2)$. The likelihood function $p(\mathbf{y}_i \mid \boldsymbol{\theta})$ is not available in closed form for this model, so the gradient and Hessian of the log likelihood with respect to the parameters $\boldsymbol{\theta}$ need to be approximated via Fisher's and Louis' identities.

The evaluation of Fisher's identity requires the gradient $\nabla_\theta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})$, where

$$\begin{aligned} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) &= \log p(\mathbf{y}_i \mid \alpha_i, \boldsymbol{\theta}) + \log p(\alpha_i \mid \boldsymbol{\theta}) \\ &= \log p(\mathbf{y}_i \mid \alpha_i, \boldsymbol{\beta}) + \log p(\alpha_i \mid \phi_\tau). \end{aligned}$$

Individually,

$$\log p(\mathbf{y}_i \mid \alpha_i, \boldsymbol{\beta}) = \sum_{j=1}^{n} y_{ij} \log\left(\frac{1}{1+\exp(-(\mathbf{x}_{ij}^\top\boldsymbol{\beta} + \alpha_i))}\right) + (1-y_{ij}) \log\left(1 - \frac{1}{1+\exp(-(\mathbf{x}_{ij}^\top\boldsymbol{\beta} + \alpha_i))}\right)$$

and

$$\log p(\alpha_i \mid \phi_\tau) = -\frac{1}{2}\log(2\pi) - \frac{\phi_\tau}{2} - \frac{1}{2}\frac{\alpha_i^2}{\exp(\phi_\tau)}.$$

The components of $\nabla_\theta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = (\nabla_\beta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})^\top, \nabla_{\phi_\tau} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}))^\top$ are derived below:

$$\begin{aligned}\nabla_\beta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) &= \nabla_\beta \log p(\mathbf{y}_i \mid \alpha_i, \boldsymbol{\beta}) \quad (\text{since } \log p(\alpha_i \mid \boldsymbol{\theta}) \text{ does not depend on } \boldsymbol{\beta}) \\ &= \sum_{j=1}^{n} y_{ij}[1 + \exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))] \frac{\partial}{\partial \boldsymbol{\beta}} \left(\frac{1}{1 + \exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))} \right) \\ &\quad - \sum_{j=1}^{n} (1 - y_{ij}) \frac{1 + \exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))}{\exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))} \frac{\partial}{\partial \boldsymbol{\beta}} \left(\frac{1}{1 + \exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))} \right), \end{aligned} \tag{S28}$$

where

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\frac{1}{1 + \exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))} \right) = \mathbf{x}_{ij} \frac{\exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))}{[1 + \exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))]^2}. \tag{S29}$$

Substituting (S29) into (S28) and reducing terms gives

$$\nabla_\beta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = \sum_{j=1}^{n} y_{ij} \mathbf{x}_{ij} \frac{\exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))}{1 + \exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))} - \sum_{j=1}^{n} (1 - y_{ij}) \mathbf{x}_{ij} \frac{1}{1 + \exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))} \tag{S30}$$

$$= \sum_{j=1}^{n} \mathbf{x}_{ij} \left[y_{ij} - \frac{1}{1 + \exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))} \right]. \tag{S31}$$

The other component of $\nabla_\theta \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})$ is

$$\begin{aligned}\nabla_{\phi_\tau} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) &= \nabla_{\phi_\tau} \log p(\alpha_i \mid \phi_\tau) \quad (\text{since } \log p(\mathbf{y}_i \mid \alpha_i, \boldsymbol{\beta}) \text{ does not depend on } \phi_\tau) \\ &= -\frac{1}{2} + \frac{\alpha_i^2}{2 \exp(\phi_\tau)}. \end{aligned} \tag{S32}$$

Evaluation of Louis' identity similarly requires

$$\nabla_\theta^2 \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = \begin{bmatrix} \nabla_\beta^2 \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) & \nabla_\beta \nabla_{\phi_\tau} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) \\ \nabla_\beta \nabla_{\phi_\tau} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta})^\top & \nabla_{\phi_\tau}^2 \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) \end{bmatrix}, \tag{S33}$$

the components of which are

$$\nabla_\beta^2 \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = \sum_{j=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}^\top} \left(\frac{\mathbf{x}_{ij}}{1 + \exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))} \right) = -\sum_{j=1}^{n} \mathbf{x}_{ij} \mathbf{x}_{ij}^\top \frac{\exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))}{\left[1 + \exp(-(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_i))\right]^2}, \tag{S34}$$

$$\nabla_{\phi_\tau}^2 \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = -\frac{\alpha_i^2}{2 \exp(\phi_\tau)}, \tag{S35}$$

$$\nabla_\beta \nabla_{\phi_\tau} \log p(\mathbf{y}_i, \alpha_i \mid \boldsymbol{\theta}) = 0. \tag{S36}$$

S1.3 Derivation of gradient and Hessian expressions for the Poisson mixed model

In this section, we derive the gradient and Hessian for the Poisson model with bivariate random effects in Section 3.3. We note that the formula for the gradient in this section can be generalised to an arbitrary number of random effects.

The parameters of interest in this model are the fixed effects, $\boldsymbol{\beta}$, and the lower-diagonal elements of the Cholesky factor $\mathbf{L}$ of the random effect covariance matrix $\boldsymbol{\Sigma}_\alpha$. We parameterise $\mathbf{L}$ as

$$\mathbf{L} = \begin{bmatrix} \exp(\zeta_{11}) & 0 \\ \zeta_{21} & \exp(\zeta_{22}) \end{bmatrix}, \tag{S37}$$

and collect the parameters in a vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\zeta})^\top$, where $\boldsymbol{\zeta} = (\zeta_{11}, \zeta_{22}, \zeta_{21})^\top$. The likelihood function $p(\mathbf{y}_i \mid \boldsymbol{\theta})$ is not available in closed form for this model, so the gradient and Hessian of the log likelihood with respect to the parameters $\boldsymbol{\theta}$ need to be approximated via Fisher's and Louis' identities.

The evaluation of Fisher's identity requires the gradient of $\log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})$, where

$$\begin{aligned} \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) &= \log p(\mathbf{y}_i \mid \boldsymbol{\alpha}_i, \boldsymbol{\beta}) + \log p(\boldsymbol{\alpha}_i \mid \boldsymbol{\zeta}) \\ &= \sum_{j=1}^{n} [y_{ij}(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\alpha}_i) - \exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\alpha}_i)] \\ &\quad - \frac{r}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_\alpha| - \frac{1}{2}\boldsymbol{\alpha}_i^\top \boldsymbol{\Sigma}_\alpha^{-1} \boldsymbol{\alpha}_i, \end{aligned} \tag{S38}$$

for $i = 1, \ldots, N$, $j = 1, \ldots, n$, and where r is the number of random effects. The gradient is

$$\nabla_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) = (\nabla_\beta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}), \nabla_\zeta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}))^\top.$$

As $\log p(\boldsymbol{\alpha}_i \mid \boldsymbol{\zeta})$ does not depend on $\boldsymbol{\beta}$, the gradient with respect to $\boldsymbol{\beta}$ is simply

$$\nabla_\beta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) = \nabla_\beta \log p(\mathbf{y}_i \mid \boldsymbol{\alpha}_i, \boldsymbol{\beta}) = \sum_{j=1}^{n} [y_{ij} - \exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\alpha}_i)]\mathbf{x}_{ij}. \tag{S39}$$

Similarly, as $\log p(\mathbf{y}_i \mid \boldsymbol{\alpha}_i, \boldsymbol{\beta})$ does not depend on $\boldsymbol{\zeta}$, the gradient with respect to $\boldsymbol{\zeta}$ reduces to

$$\begin{aligned}\nabla_\zeta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) &= \nabla_\zeta \log p(\boldsymbol{\alpha}_i \mid \boldsymbol{\zeta}) \\ &= \mathrm{Tr}(\mathbf{A}^\top \nabla_{\zeta_{kl}} \mathbf{L}), \quad k = 1, 2, l < k,\end{aligned}$$

where $\mathbf{A} = -\mathbf{L}^{-\top} + \mathbf{L}^{-\top}\mathbf{L}^{-1}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\top\mathbf{L}^{-\top}$, with $\mathbf{M}^{-\top}$ denoting the inverse of the transpose of a matrix $\mathbf{M}$, and

$$\nabla_{\zeta_{kl}} \mathbf{L} = \begin{cases} \mathbf{J}_{kk}^* & \text{if } k = l, \\ \mathbf{J}_{kl} & \text{otherwise,} \end{cases} \tag{S40}$$

for $k = 1, 2$, $l < k$, where $\mathbf{J}_{kl}$ is a 2×2 matrix that has 1 in the (k, l)th element and 0 elsewhere, and $\mathbf{J}_{kk}^*$ is a 2×2 matrix that has $\exp(\zeta_{kk})$ in the (k, k)th element and 0 elsewhere.

The Hessian for the subject-specific log likelihood is

$$\begin{aligned}\nabla_\theta^2 \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) &= \begin{bmatrix} \nabla_\beta^2 \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) & \nabla_\zeta \nabla_\beta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})^\top \\ \nabla_\zeta \nabla_\beta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) & \nabla_\zeta^2 \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) \end{bmatrix} \\ &= \begin{bmatrix} \nabla_\beta^2 \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{0} & \nabla_\zeta^2 \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) \end{bmatrix}.\end{aligned}$$

Here the cross-derivative $\nabla_\zeta \nabla_\beta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})$ is zero because $\nabla_\beta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})$ does not depend on $\boldsymbol{\zeta}$. The remaining components of the Hessian matrix are

$$\nabla_\beta^2 \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) = \sum_{j=1}^{n} [y_{ij} - \exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\alpha}_i)] \mathbf{x}_{ij} \mathbf{x}_{ij}^\top, \tag{S41}$$

and

$$\nabla_\zeta^2 \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) = \begin{bmatrix} \nabla_{\zeta_{11}}^2 \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) & \nabla_{\zeta_{22}} \nabla_{\zeta_{11}} \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})^\top & \nabla_{\zeta_{21}} \nabla_{\zeta_{11}} \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})^\top \\ \nabla_{\zeta_{22}} \nabla_{\zeta_{11}} \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) & \nabla_{\zeta_{22}}^2 \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})^\top & \nabla_{\zeta_{21}} \nabla_{\zeta_{22}} \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})^\top \\ \nabla_{\zeta_{21}} \nabla_{\zeta_{11}} \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) & \nabla_{\zeta_{21}} \nabla_{\zeta_{22}} \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})^\top & \nabla_{\zeta_{21}}^2 \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta})^\top \end{bmatrix}$$

where

$$\nabla_{\zeta_{kl}}^2 \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \boldsymbol{\theta}) = \mathrm{Tr}((\nabla_{\zeta_{kl}} \mathbf{A})^\top \nabla_{\zeta_{kl}} \mathbf{L} + \mathbf{A}^\top \nabla_{\zeta_{kl}}^2 \mathbf{L}), \quad k = 1, 2,\ l < k. \tag{S42}$$

Separately,

$$\nabla^2_{\zeta_{kl}}\mathbf{L} = \begin{cases} \mathbf{J}^*_{kk}, & \text{if } k = l \\ \mathbf{0}_2, & \text{otherwise,} \end{cases} \tag{S43}$$

where $\mathbf{0}_m$ is a matrix of zeros of dimension m, and

$$\begin{aligned} \nabla_{\zeta_{kl}}\mathbf{A} &= \nabla_{\zeta_{kl}}(-\mathbf{L}^{-\top} + \mathbf{L}^{-\top}\mathbf{L}^{-1}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\top\mathbf{L}^{-\top}) \\ &= -\nabla_{\zeta_{kl}}\mathbf{L}^{-\top} + (\nabla_{\zeta_{kl}}\mathbf{L}^{-\top})\mathbf{B} + \mathbf{L}^{-\top}\nabla_{\zeta_{kl}}\mathbf{B}, \end{aligned} \tag{S44}$$

where $\mathbf{B} = \mathbf{L}^{-1}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\top\mathbf{L}^{-\top}$, and $\nabla_{\zeta_{kl}}\mathbf{B} = (\nabla_{\zeta_{kl}}\mathbf{L}^{-\top})\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\top\mathbf{L}^{-\top} + \mathbf{L}^{-1}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\top(\nabla_{\zeta_{kl}}\mathbf{L}^{-\top})$, for $k = 1, 2$, $l < k$. We derive the terms $\nabla_{\zeta_{kl}}\mathbf{L}^{-\top}$ individually:

$$\begin{aligned} \nabla_{\zeta_{11}}\mathbf{L}^{-\top} &= \begin{bmatrix} -\exp(-\zeta_{11}) & 0 \\ \zeta_{21}\exp(-\zeta_{11}-\zeta_{22}) & 0 \end{bmatrix}, \\ \nabla_{\zeta_{22}}\mathbf{L}^{-\top} &= \begin{bmatrix} 0 & 0 \\ \zeta_{21}\exp(-\zeta_{11}-\zeta_{22}) & -\exp(-\zeta_{22}) \end{bmatrix}, \\ \nabla_{\zeta_{21}}\mathbf{L}^{-\top} &= \begin{bmatrix} 0 & 0 \\ -\exp(-\zeta_{11}-\zeta_{22}) & 0 \end{bmatrix}. \end{aligned}$$

S2 Variance of the R-VGAL posterior densities for various Monte Carlo sample sizes

In this section, we study the effects of increasing the number of samples S, for estimating the expectations in the variational mean and precision matrix updates, and S_α, for estimating the gradient of the log-likelihood on the R-VGAL posterior estimates. The results are obtained using the simulated logistic data in Section 3.2 and the Polypharmacy dataset in Section 3.4 of the main paper. Similar results are obtained for other datasets. For all simulations in this section, we use the damped R-VGAL algorithm in Section 2.4 with $n_{damp} = 10$ observations and $K = 4$ damping steps per observation.

S2.1 Simulated logistic data

The simulated data used in this section is the same as that used in Section 3.2 of the main paper. The same values for S and S_α are used and they are taken from the set $\{50, 100, 500, 1000\}$. For each pair of S and S_α values, we independently run R-VGAL 10 times on the simulated dataset, and plot the posterior densities from all 10 runs for each parameter. These posterior densities are shown in Figures S1 to S4. For comparison, the HMC posterior distributions (from a single run, with 2 chains and 20000 total posterior samples after burn-in) are also plotted in each figure.

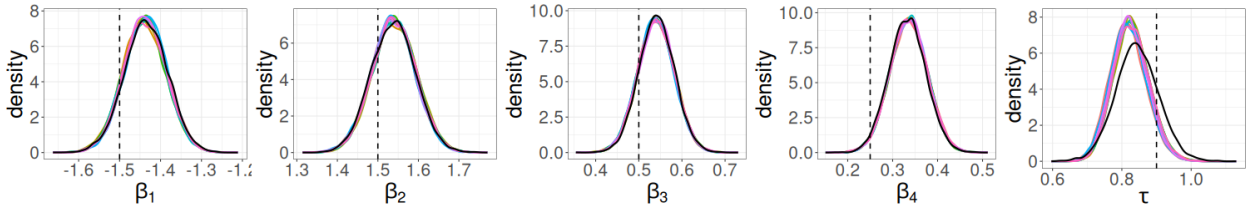

Figure S1: R-VGAL posterior distributions from 10 runs on simulated logistic data are plotted in colour. Damping is done on the first 10 observations, and the Monte Carlo sample sizes are $S = 50$ and $S_\alpha = 50$. HMC posterior distributions are plotted in black for comparison.

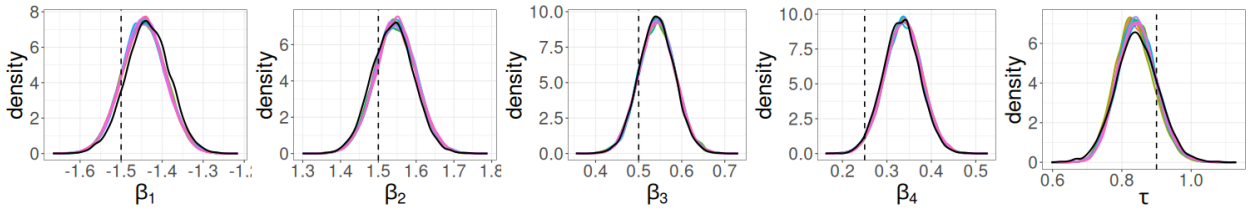

Figure S2: R-VGAL posterior distributions from 10 runs on simulated logistic data are plotted in colour. Damping is done on the first 10 observations, and the Monte Carlo sample sizes are $S = 100$ and $S_\alpha = 100$. HMC posterior distributions are plotted in black for comparison.

As the Monte Carlo sample sizes increase, the R-VGAL posterior density estimates get closer and closer to

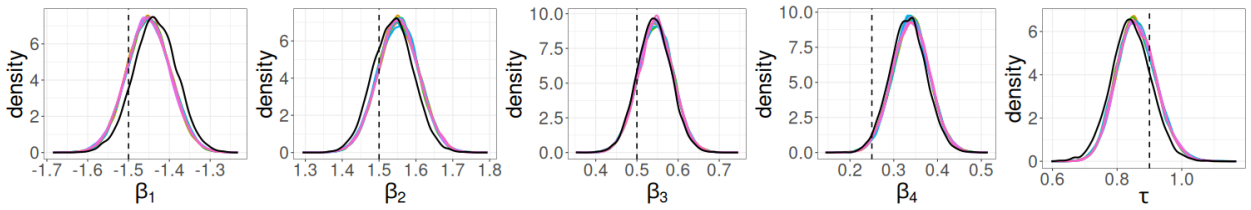

Figure S3: R-VGAL posterior distributions from 10 runs on simulated logistic data are plotted in colour. Damping is done on the first 10 observations, and the Monte Carlo sample sizes are $S = 500$ and $S_\alpha = 500$. HMC posterior distributions are plotted in black for comparison.

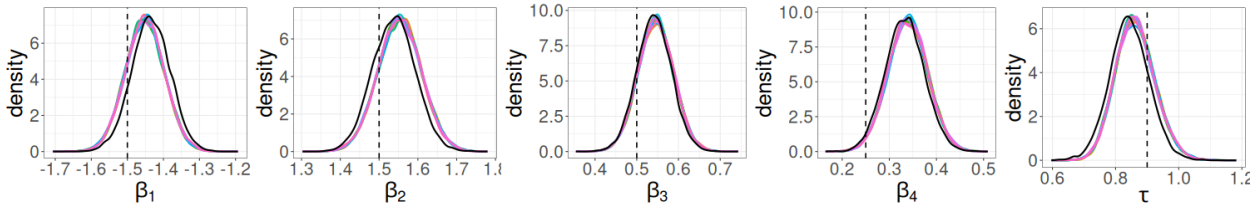

Figure S4: R-VGAL posterior distributions from 10 runs on simulated logistic data are plotted in colour. Damping is done on the first 10 observations, and the Monte Carlo sample sizes are $S = 1000$ and $S_\alpha = 1000$. HMC posterior distributions are plotted in black for comparison.

each other, which shows that increasing the values of S and S_α helps reduce the variability of the R-VGAL posterior estimates across multiple independent runs.

S2.2 Polypharmacy dataset

The Polypharmacy dataset used in this section is the same as that used in Section 3.4 of the main paper. Similar to Section S2.1, the same values for S and S_α are used and they are taken from the set $\{50, 100, 500, 1000\}$. We independently run R-VGAL 10 times on the Polypharmacy dataset for each pair of S and S_α values, and plot the posterior densities from all 10 runs for each parameter. These posteriors are shown in Figures S5 to S8. For comparison, the HMC posterior distributions (from a single run, with 2 chains and 20000 total posterior samples after burn-in) are also plotted.

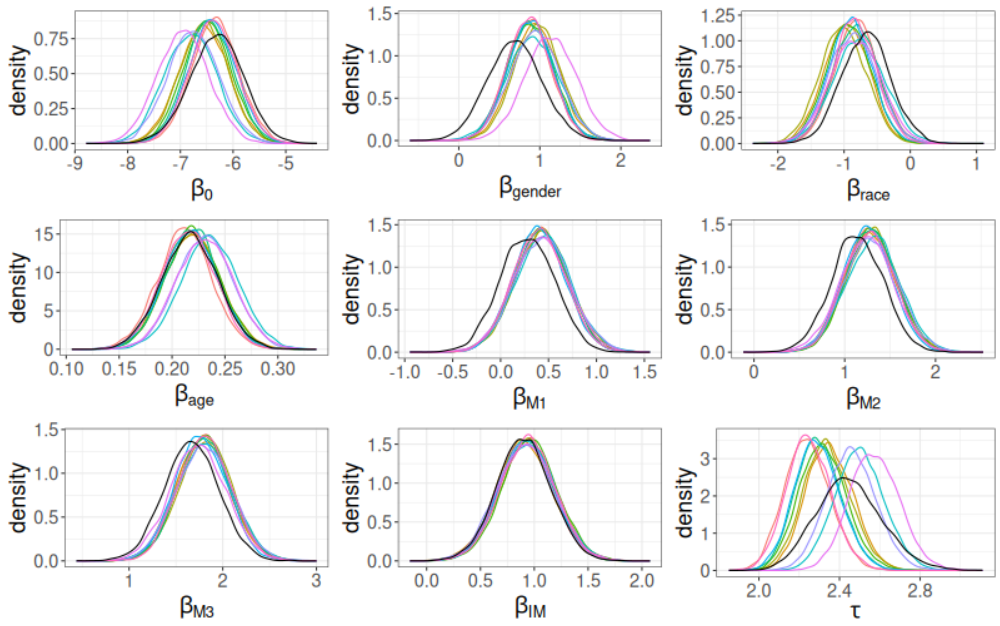

Figure S5: R-VGAL posterior distributions from 10 runs on the Polypharmacy dataset are plotted in colour. Damping is done on the first 10 observations, and the Monte Carlo sample sizes are $S = 50$ and $S_\alpha = 50$. HMC posterior distributions are plotted in black for comparison.

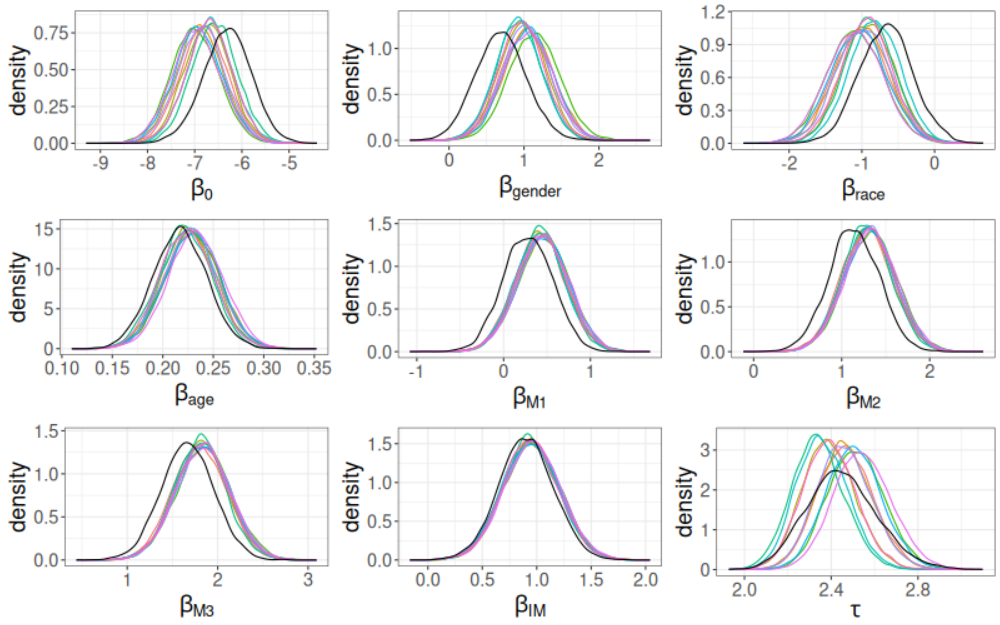

Figure S6: R-VGAL posterior distributions from 10 runs on the Polypharmacy dataset are plotted in colour. Damping is done on the first 10 observations, and the Monte Carlo sample sizes are $S = 100$ and $S_\alpha = 100$. HMC posterior distributions are plotted in black for comparison.

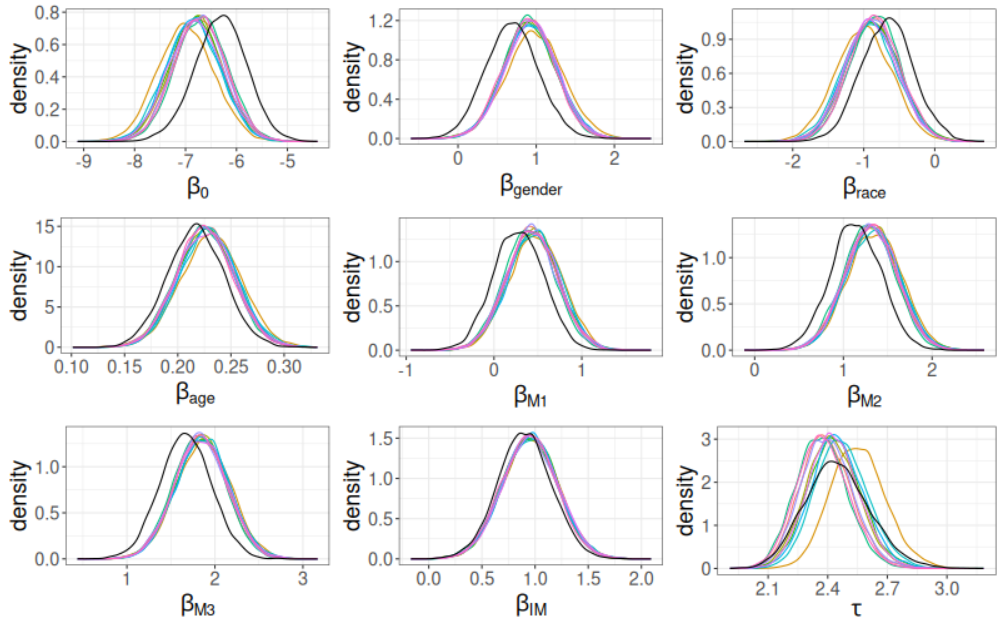

Figure S7: R-VGAL posterior distributions from 10 runs on the Polypharmacy dataset are plotted in colour. Damping is done on the first 10 observations, and the Monte Carlo sample sizes are $S = 500$ and $S_\alpha = 500$. HMC posterior distributions are plotted in black for comparison.

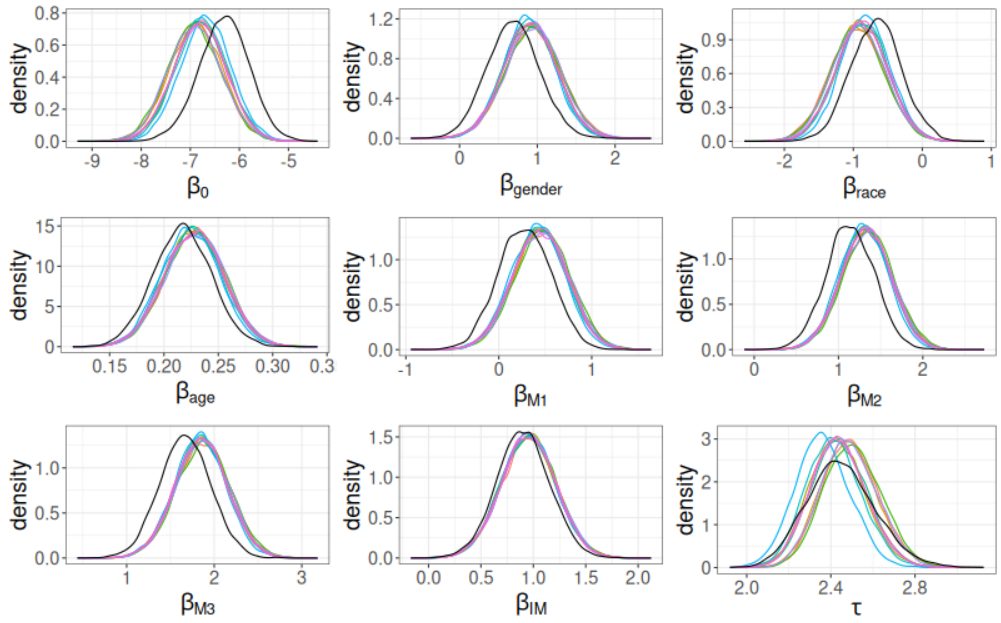

Figure S8: R-VGAL posterior distributions from 10 runs on the Polypharmacy dataset are plotted in colour. Damping is done on the first 10 observations, and the Monte Carlo sample sizes are $S = 1000$ and $S_\alpha = 1000$. HMC posterior distributions are plotted in black for comparison.

As with the previous example, the results in this example also show that increasing the values of S and S_α reduces the variability of the R-VGAL posterior estimates across multiple runs. This phenomenon is particularly pronounced for the random effect standard deviation τ. Suitable values for S and S_α are likely to be application-dependent. However, from our studies, we conclude that S and S_α need to be at least 100 for the Monte Carlo sample sizes not to have a substantial effect on the final estimates.

S3 Robustness check of the R-VGAL algorithm

In this section, we use the Polypharmacy dataset in Section 3.4 to check the robustness of the R-VGAL algorithm given different orderings of the data. The simulations in this section show that R-VGAL can be unstable while traversing the first few observations, which makes it sensitive to the ordering of observations. This instability can, however, be alleviated with the damped R-VGAL algorithm, as described in Section 2.4 of the main paper.

Figures S9 and S10 show the R-VGAL posterior density estimates from 10 independent runs using the original ordering of the data and a random ordering of the data, respectively. In both simulations, the number of Monte Carlo samples S to estimate the expectation with respect to $q_{i-1}(\boldsymbol{\theta})$ and the number of samples S_α to estimate the gradients/Hessians are fixed to 100. In both figures, the HMC posterior densities for each parameter are plotted in black for comparison. The figures show that the R-VGAL estimates are quite far away from those of HMC estimates when using the original ordering of the data, while the R-VGAL estimates are reasonably close to those of HMC when using the random ordering of the data. This suggests that the R-VGAL estimates are not robust with respect to the ordering of the data.

To confirm that the source of variability in the R-VGAL estimates is from different data orderings and not from the low number of Monte Carlo samples, we increase the number of Monte Carlo samples S and S_α. Figures S11 and S12 show the R-VGAL posterior density estimates from 10 independent runs using the original ordering of the data and a random ordering of the data, respectively, with the Monte Carlo sample sizes set to $S = S_\alpha = 1000$. The posterior densities for each parameter are different for the two orderings; for instance, with the original ordering, the posterior of τ is centred around 4.5, while with the random ordering, the posterior of τ is centred around 2.4.

A plot of the trajectory of the variational mean across R-VGAL iterations reveals that R-VGAL is unstable during the first few iterations. The blue lines in Figures S13 and S14 show the trajectories of the variational mean for each of the parameters across 10 independent runs of the R-VGAL algorithm, on the original

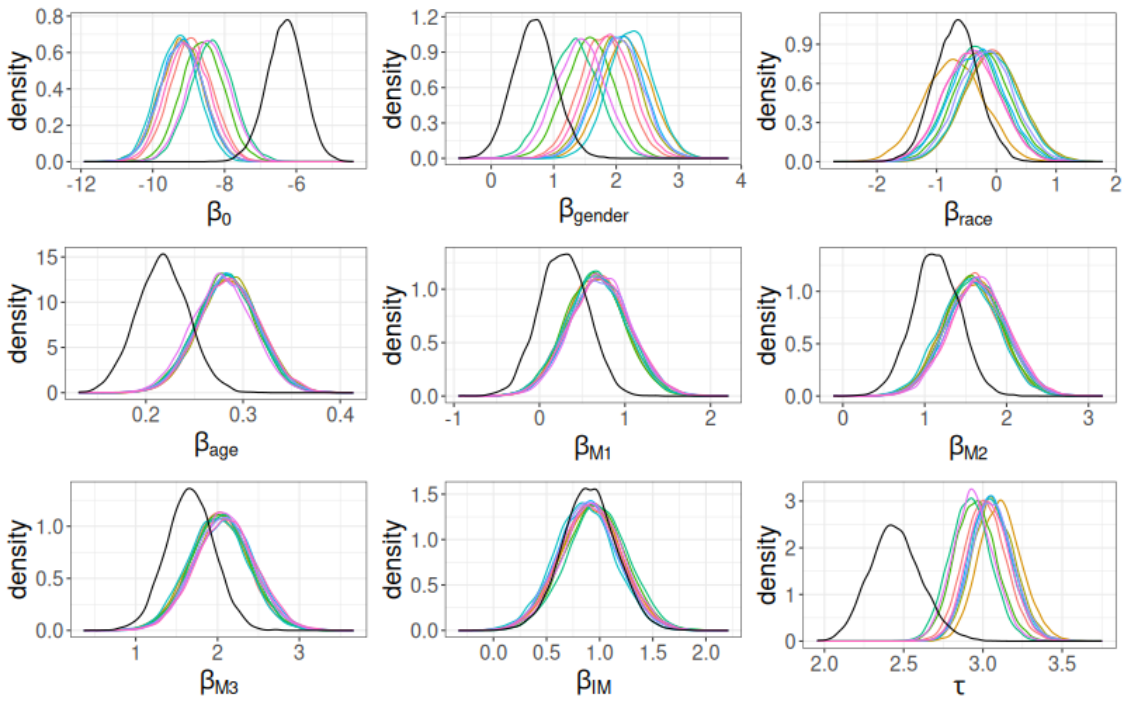

Figure S9: R-VGAL posterior distributions from 10 independent runs using the original ordering of the data are plotted in different colours. The Monte Carlo sample sizes are $S = 100, S_\alpha = 100$. HMC posterior distributions are plotted in black for comparison.

ordering and on a random ordering of the data, respectively. The initial trajectories of the fixed effect parameters in Figure S13 vary significantly (for example, between -50 and 0 for the intercept β_0), and the trajectory of τ is dragged up to nearly 7 before progressively dropping towards 4. This potentially contributes to the biased posterior mean of τ. In Figure S14, where the data were randomly reordered, the trajectories of the parameters are less variable initially, which then allows the variational mean to converge towards the true values more rapidly. This shows that the R-VGAL algorithm is unstable while traversing the first few observations, making the algorithm sensitive to the ordering of the data.

We propose a damping approach (in Section 2.4 of the main paper) to make the R-VGAL algorithm more robust. By damping the first few observations, the R-VGAL posterior estimates become much more consistent across different data orderings. Figures S15 and S16 show the posterior density estimates from 10 independent runs of the R-VGAL algorithm using the original ordering of the data and a random ordering of the data, respectively, with damping done on the first 10 observations, and with Monte Carlo sample sizes $S = S_\alpha = 100$. These figures show that the posterior density estimates of R-VGAL with damping are consistent across two different orderings of the data, and also consistent with those obtained from HMC.

Figures S17 and S18 display the R-VGAL posterior densities using the original ordering and the random

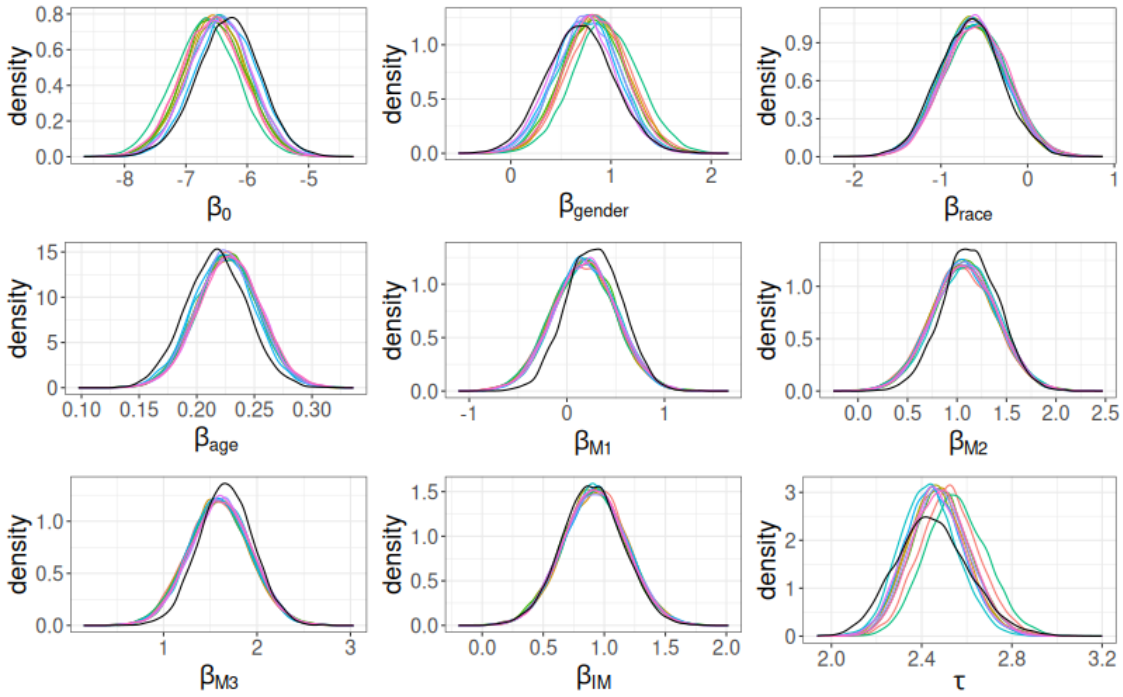

Figure S10: R-VGAL posterior distributions from 10 independent runs using the random ordering of the data are plotted in different colours. The Monte Carlo sample sizes are $S = 100$, and $S_\alpha = 100$. HMC posterior distributions are plotted in black for comparison.

ordering of the data, respectively, with damping done on the first 10 observations, and Monte Carlo sample sizes increased to $S = S_\alpha = 1000$. There is now very little difference between the posterior densities using the original and the random ordering of the dataset. The red lines in Figures S13 and S14 show the parameter trajectories obtained from damped R-VGAL, on the original ordering and the random ordering of the data, respectively. The trajectories with damping (plotted in red) are much more stable than those without damping (plotted in blue), especially during the first few iterations. These figures suggest that damping is effective in reducing the variability of R-VGAL estimates while traversing the first few observations and increases the algorithm's robustness to different data orderings. Other random orderings of the data give similar results.

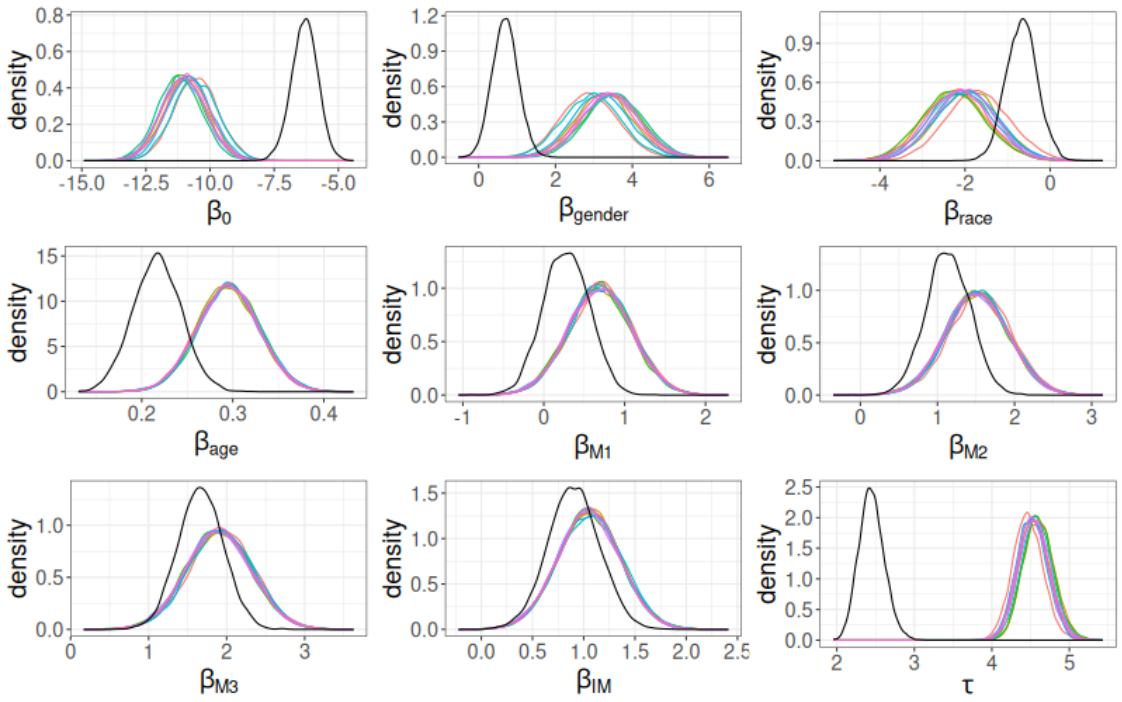

Figure S11: R-VGAL posterior distributions from 10 independent runs using the original ordering of the data are plotted in colour. The Monte Carlo sample sizes are $S = 1000, S_\alpha = 1000$. HMC posterior distributions are plotted in black for comparison.

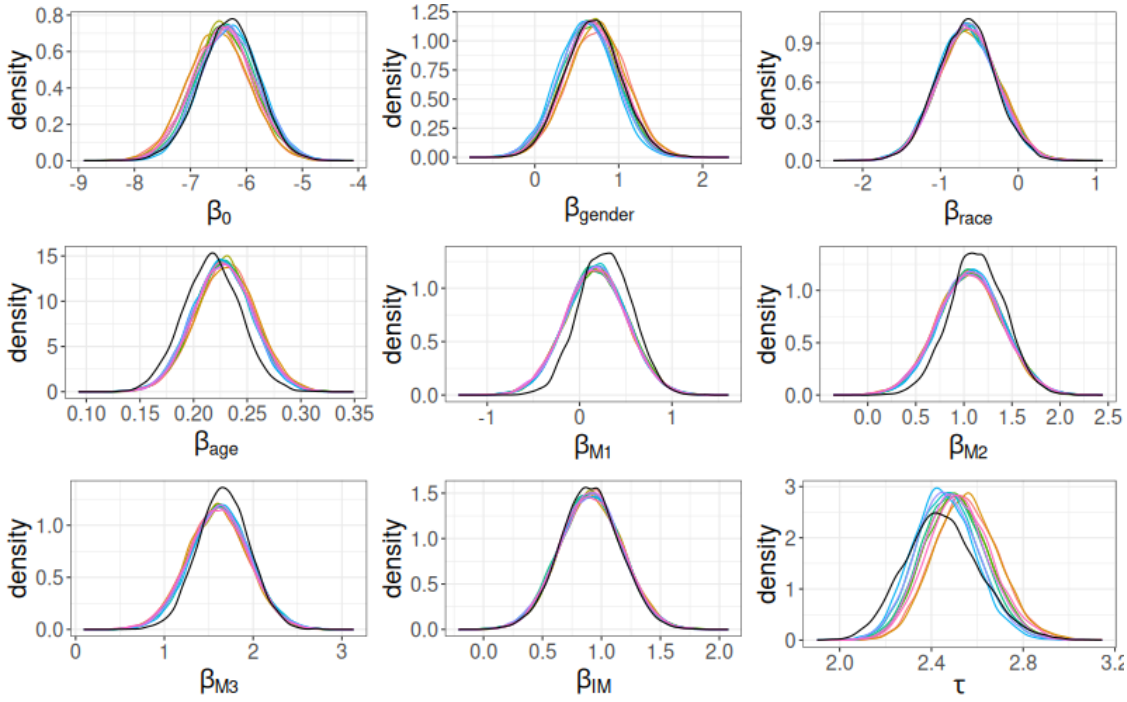

Figure S12: R-VGAL posterior distributions from 10 independent runs using the random ordering of the data are plotted in colour. The Monte Carlo sample sizes are $S = 1000, S_\alpha = 1000$. HMC posterior distributions are plotted in black for comparison.

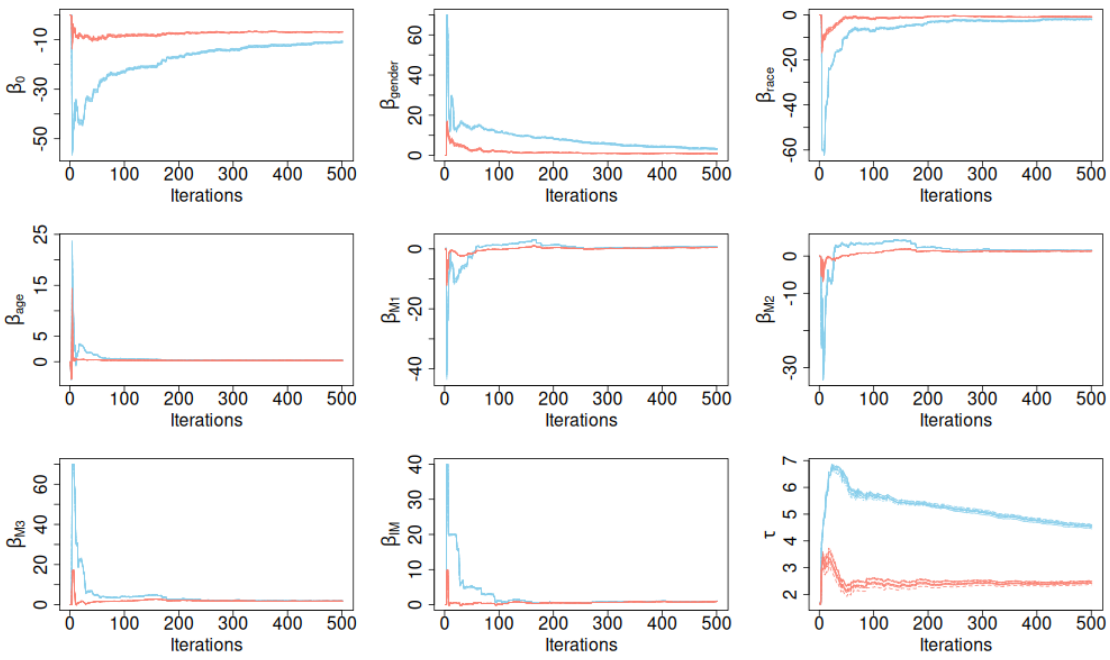

Figure S13: Trajectories of the variational mean without damping (in blue) and with damping (in red) for each parameter across 10 independent runs of R-VGAL on the original ordering of the data. The Monte Carlo sample sizes are $S = 1000$ and $S_\alpha = 1000$.

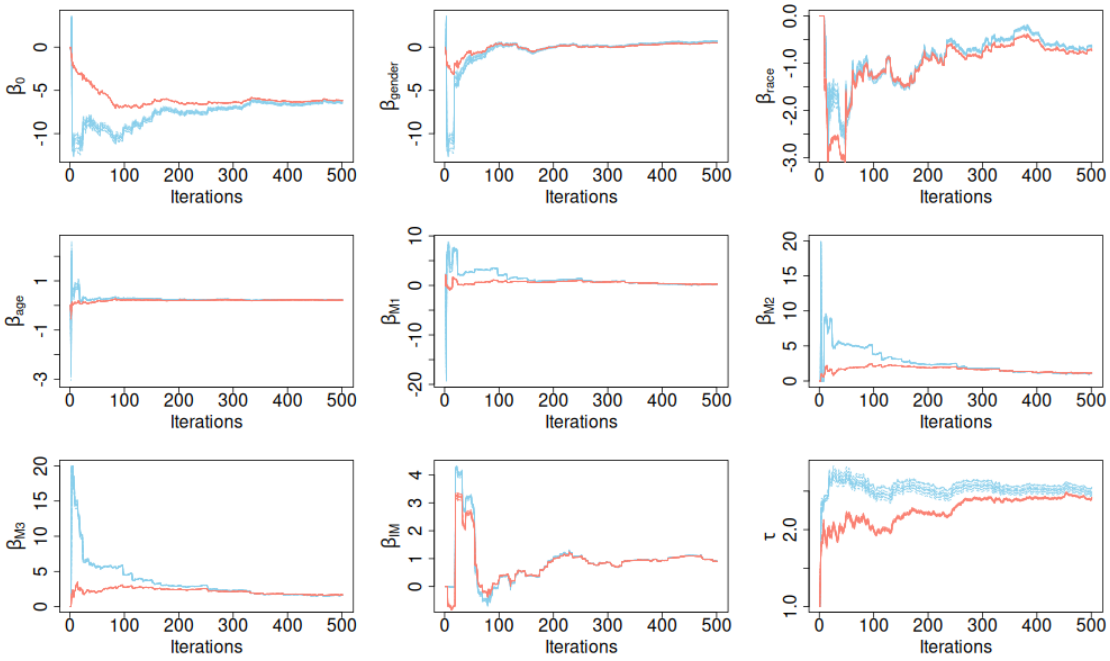

Figure S14: Trajectories of the variational mean without damping (in blue) and with damping (in red) for each parameter across 10 independent runs of R-VGAL on the random ordering of the data. The Monte Carlo sample sizes are $S = 1000$ and $S_\alpha = 1000$.

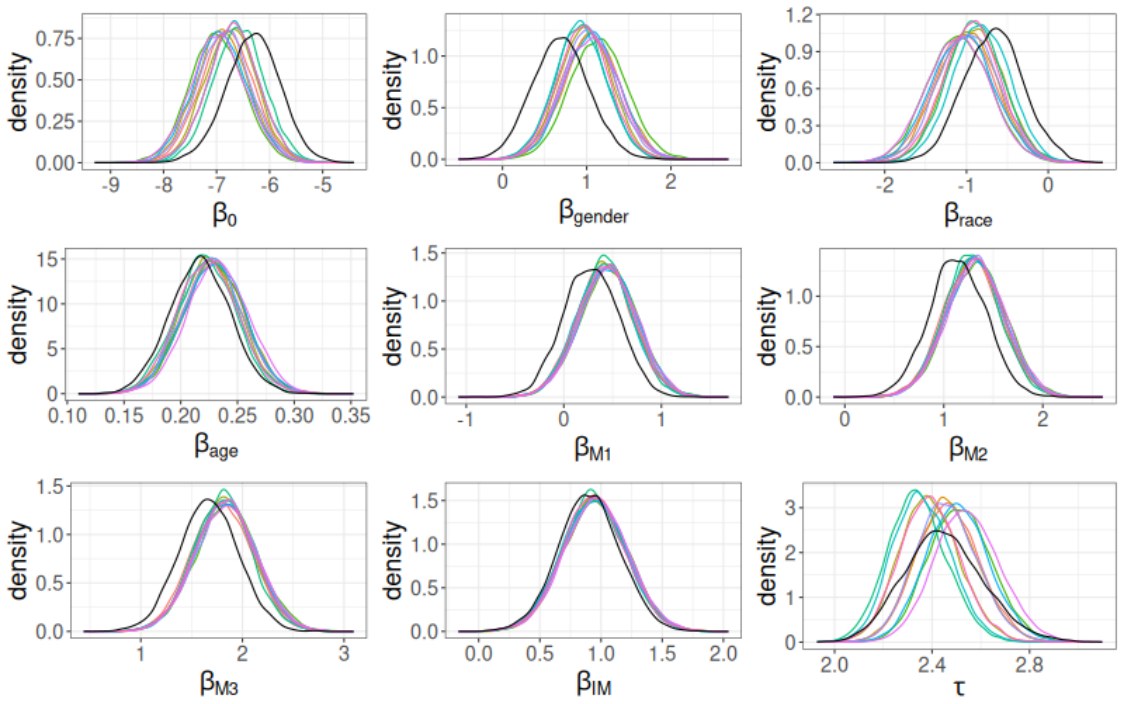

Figure S15: R-VGAL posterior distributions from 10 independent runs using the original ordering of the data are plotted in colour. Damping is done on the first 10 observations. The Monte Carlo sample sizes are $S = 100, S_\alpha = 100$. HMC posterior distributions are plotted in black for comparison.

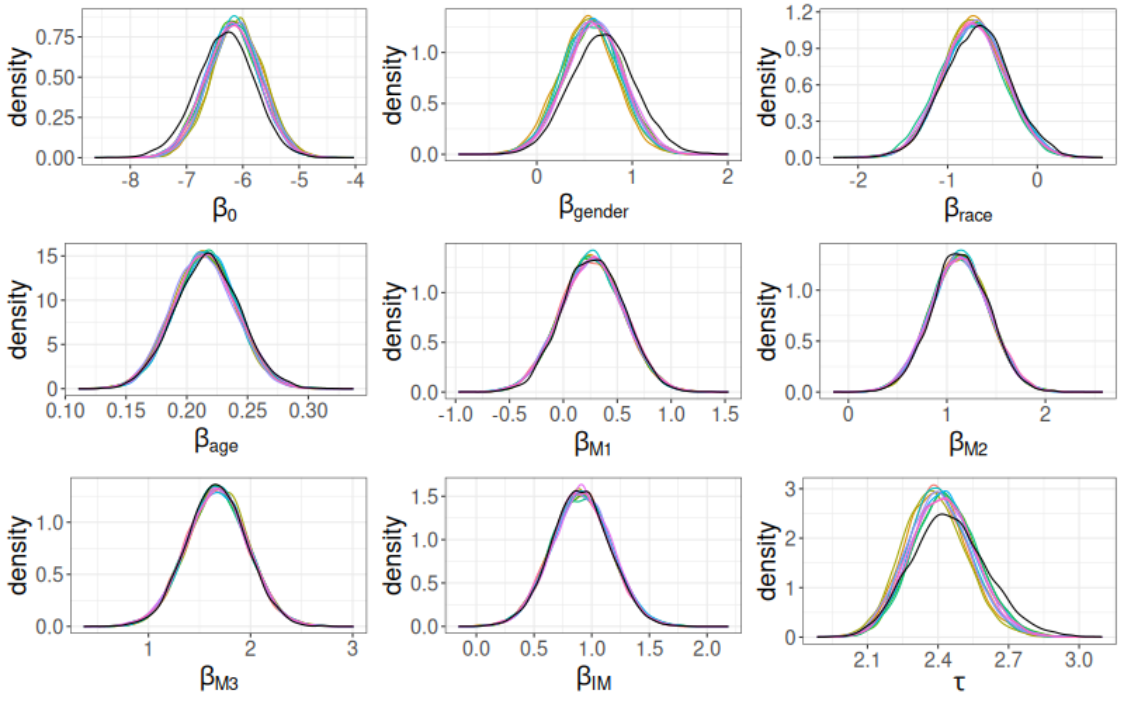

Figure S16: R-VGAL posterior distributions from 10 independent runs using a random reordering of the data are plotted in colour. Damping is done on the first 10 observations. The Monte Carlo sample sizes are $S = 100, S_\alpha = 100$. HMC posterior distributions are plotted in black for comparison.

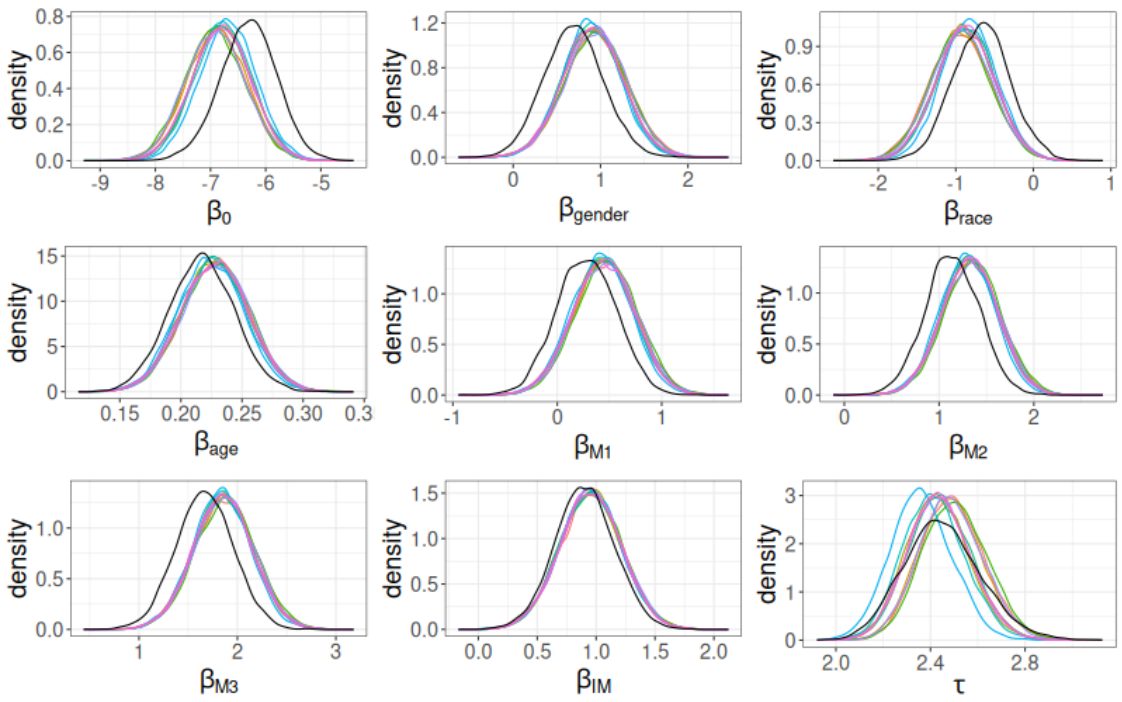

Figure S17: R-VGAL posterior distributions from 10 independent runs using the original ordering of the data are plotted in colour. Damping is done on the first 10 observations. The Monte Carlo sample sizes are $S = 1000, S_\alpha = 1000$. HMC posterior distributions are plotted in black for comparison.

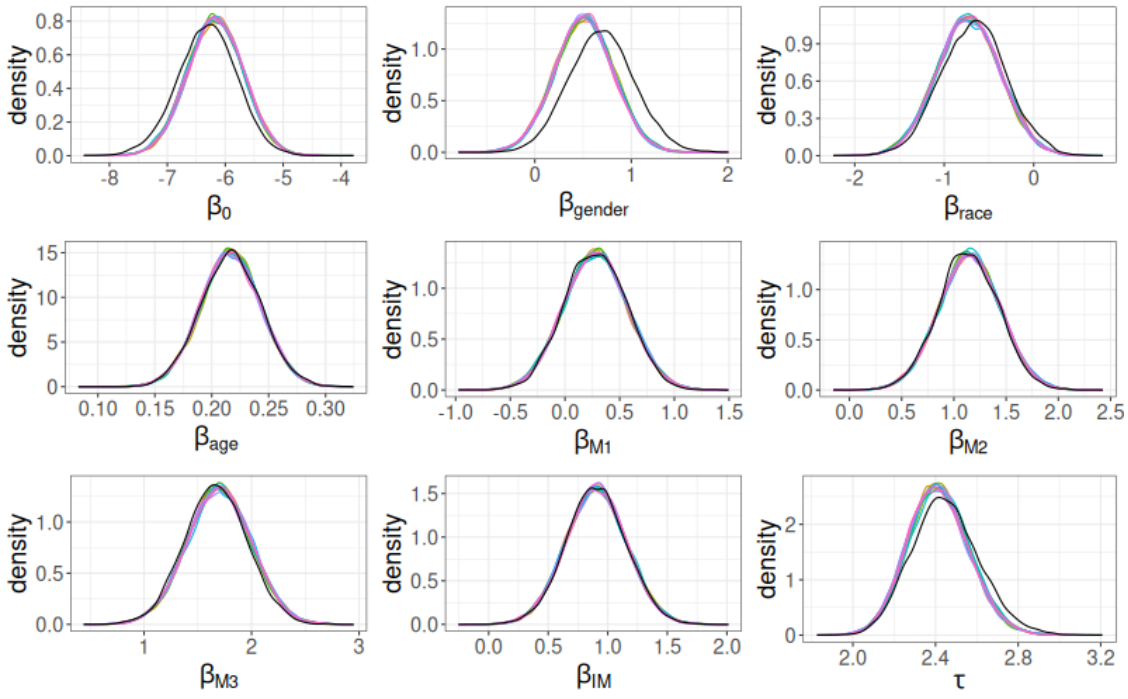

Figure S18: R-VGAL posterior distributions from 10 independent runs using a random reordering of the data are plotted in colour. Damping is done on the first 10 observations. The Monte Carlo sample sizes are $S = 1000, S_\alpha = 1000$. HMC posterior distributions are plotted in black for comparison.

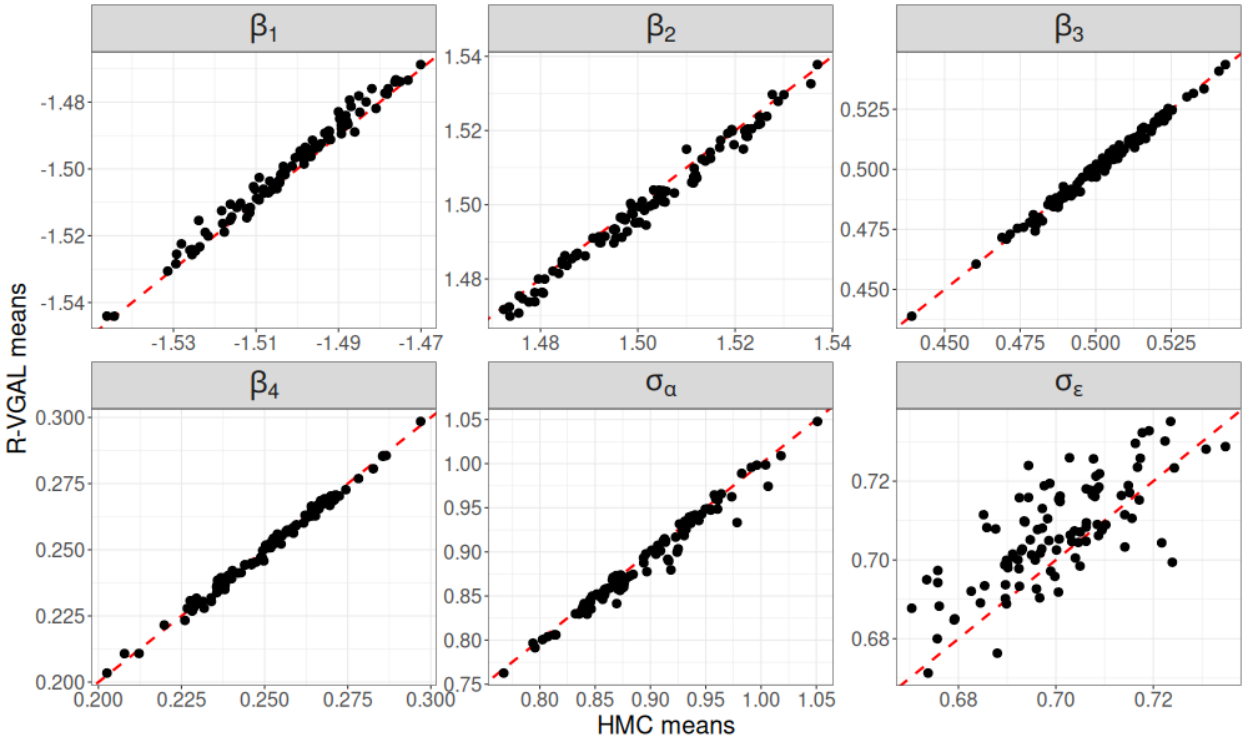

Figure S19: Plot of the R-VGAL posterior means against those from HMC for 100 datasets simulated from the linear mixed model.

S4 Repeated simulations

In this section, for each of the linear, logistic and Poisson mixed models, we simulate 100 datasets with the same parameter settings. We run R-VGAL and HMC on each of the 100 datasets, and compare their posterior density estimates.

S4.1 Repeated simulations from the linear mixed model

The synthetic datasets in this section are simulated according to the number of observations and parameter values detailed in Section 3.1. Figure S19 plots the R-VGAL posterior means against those from HMC for each of the 100 simulated datasets. The red dotted diagonal line marks the "ideal" scenario where the posterior means from the two methods are equal. Figure S20 plots the ratio between the R-VGAL and HMC posterior standard deviations, with the dotted horizontal line marking the ideal ratio of one. Figure S21 compares the distribution of the differences between the R-VGAL means and the true parameter values to the distribution of the differences between the HMC means and the true parameter values. We see that posterior means and standard deviations from the two methods are very similar across the 100 replicated datasets.

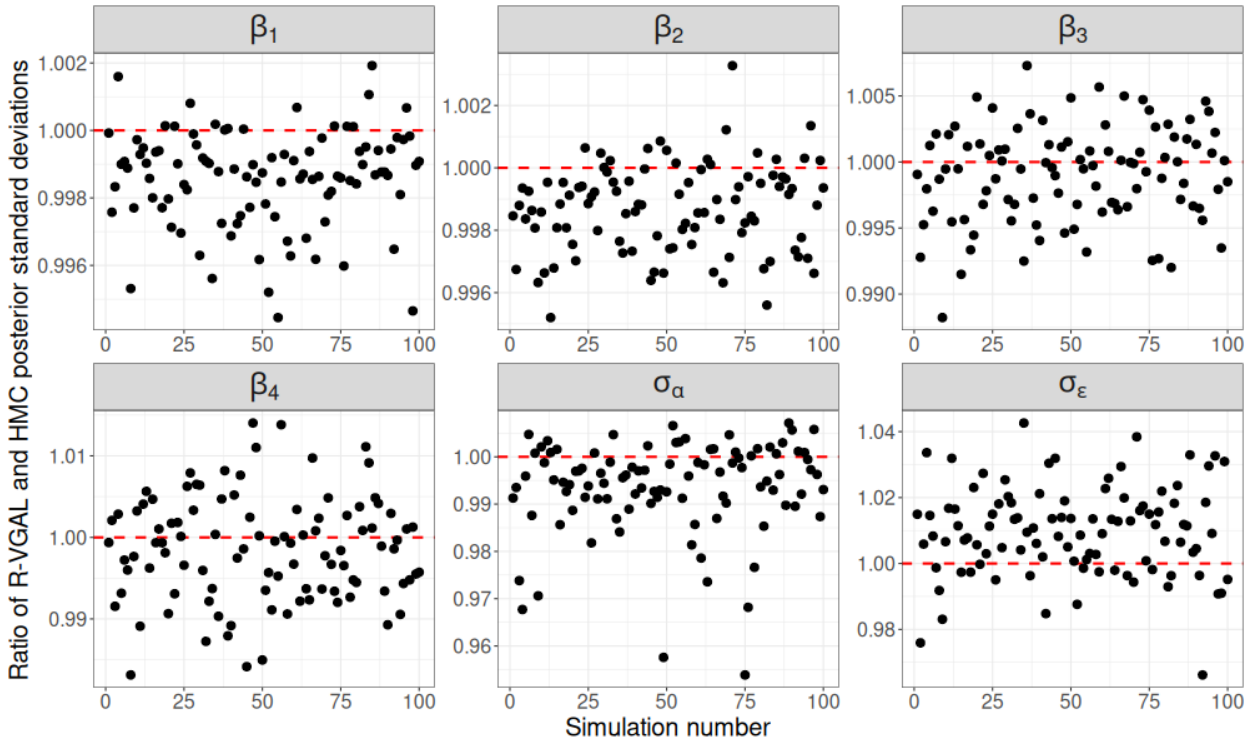

Figure S20: Plot of the ratio between the R-VGAL posterior standard deviations and those from HMC for 100 datasets simulated from the linear mixed model.

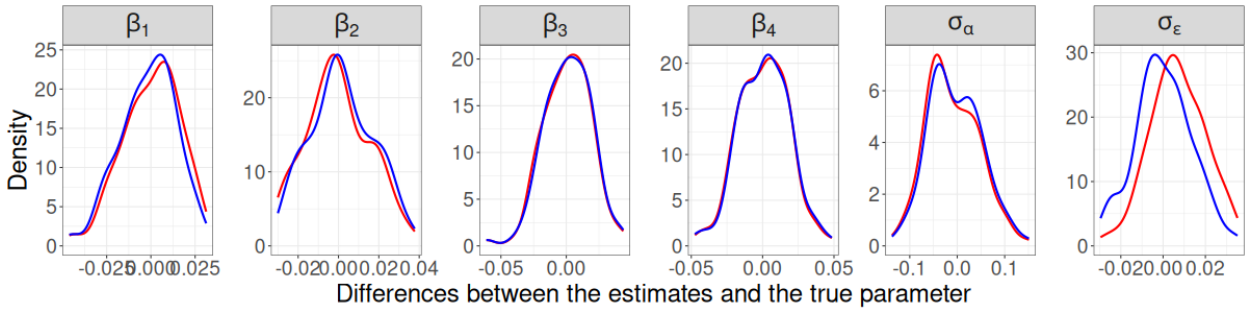

Figure S21: Comparison of the distribution of the differences between the R-VGAL posterior means and the true parameters (in red), against the distribution of the differences between the HMC posterior means and the true parameters (in blue), for 100 datasets simulated from the linear mixed model.

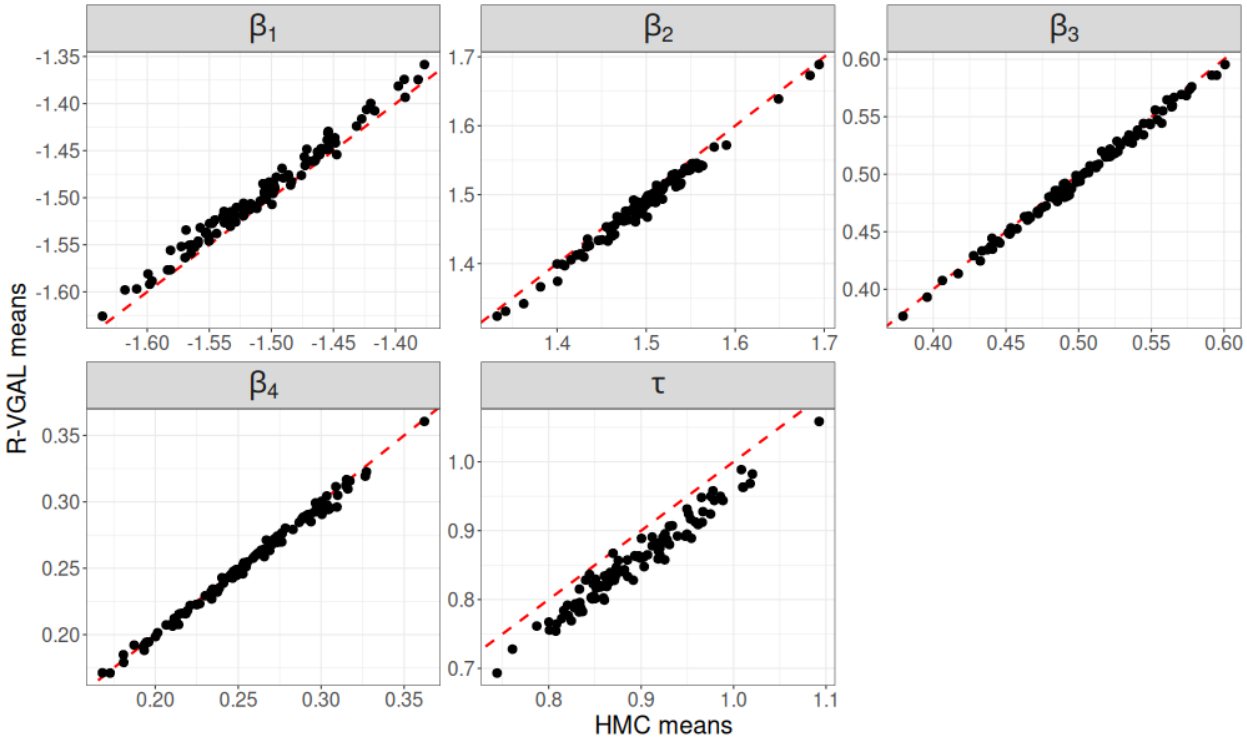

Figure S22: Plot of the R-VGAL posterior means against those from HMC for 100 datasets simulated from the logistic mixed model.

S4.2 Repeated simulations from the logistic mixed model

The synthetic datasets in this section are simulated according to the number of observations and parameter values detailed in Section 3.2. As with the previous section, we present plots comparing the R-VGAL and HMC posterior means in Figure S22, the ratio between the R-VGAL and HMC posterior standard deviations in Figure S23, and the distributions of the differences between each method's posterior mean estimates and the true parameters in Figure S24. The R-VGAL and HMC posterior means are similar, though R-VGAL tends to slightly underestimate the random effect variance when compared to HMC.

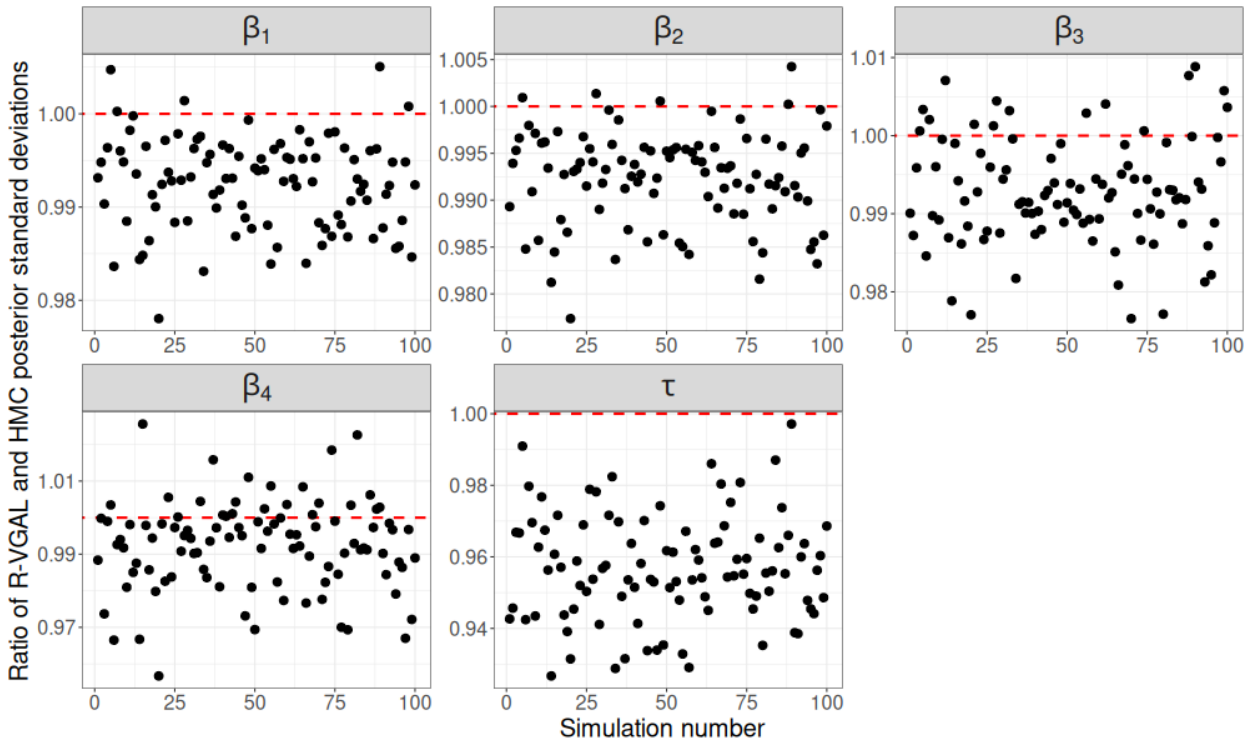

Figure S23: Plot of the ratio between the R-VGAL posterior standard deviations and those from HMC for 100 datasets simulated from the logistic mixed model.

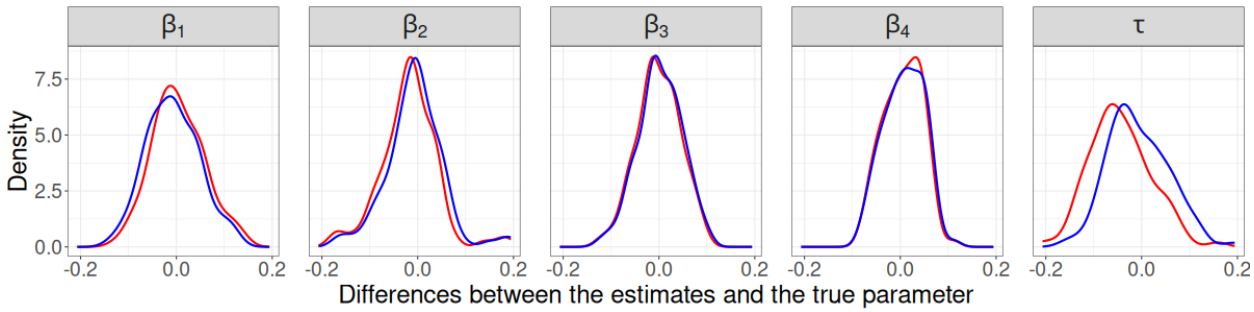

Figure S24: Comparison of the distribution of the differences between the R-VGAL posterior means and the true parameters (in red), against the distribution of the differences between the HMC posterior means and the true parameters (in blue), for 100 datasets simulated from the logistic mixed model.

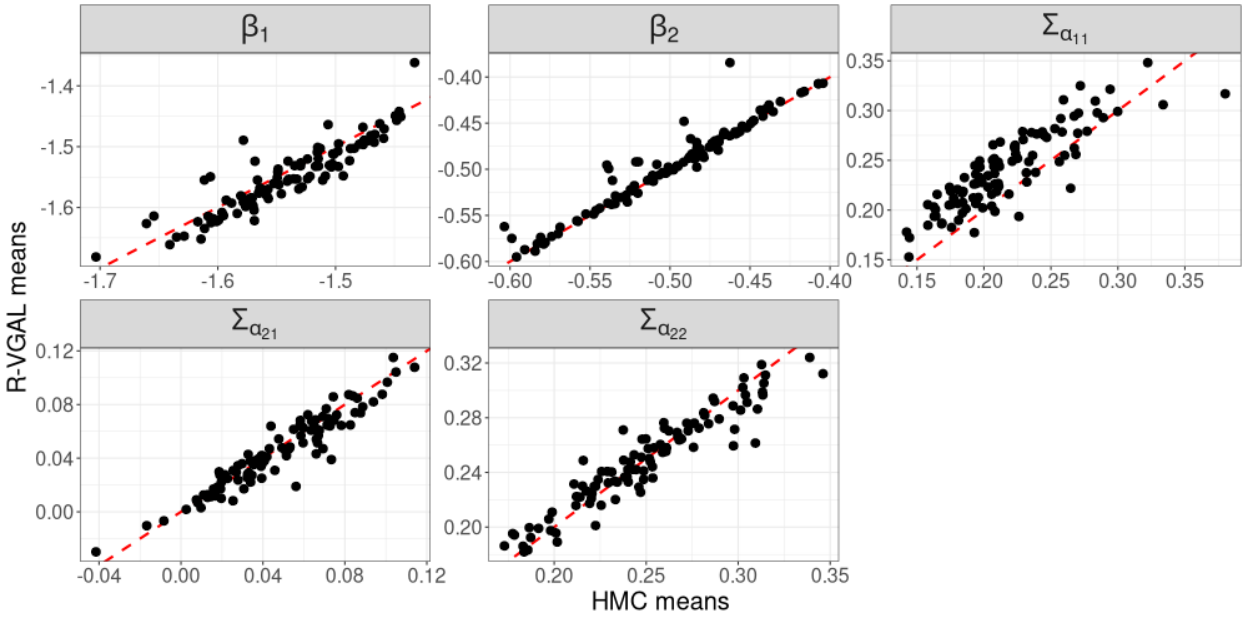

Figure S25: Plot of the R-VGAL and HMC posterior means for 100 datasets simulated from the Poisson mixed model.

S4.3 Repeated simulations on the Poisson mixed model

The synthetic datasets in this section are simulated according to the number of observations and parameter values detailed in Section 3.3. As with the previous sections, we present plots comparing the R-VGAL and HMC posterior means in Figure S25, the ratio between the R-VGAL and HMC posterior standard deviations in Figure S26, and the distributions of the differences between each method's posterior mean estimates and the true parameters in Figure S27. The R-VGAL and HMC posterior means and standard deviations are quite similar across simulations, although R-VGAL has a mild tendency for overestimating the parameter $\mathbf{\Sigma}_{\alpha_{11}}$ and underestimating the posterior variance of $\mathbf{\Sigma}_{\alpha_{11}}$ and $\mathbf{\Sigma}_{\alpha_{22}}$ compared to HMC.

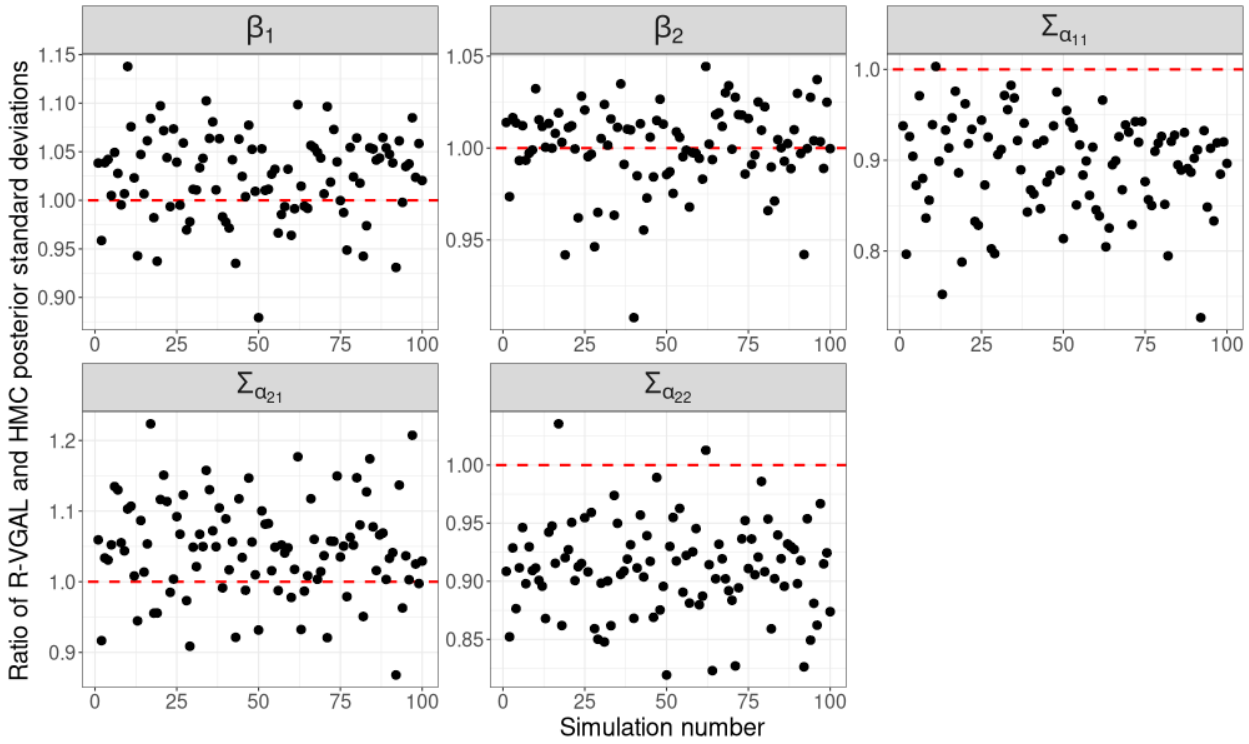

Figure S26: Plot of the ratio between the R-VGAL and HMC posterior standard deviations for 100 datasets simulated from the Poisson mixed model.

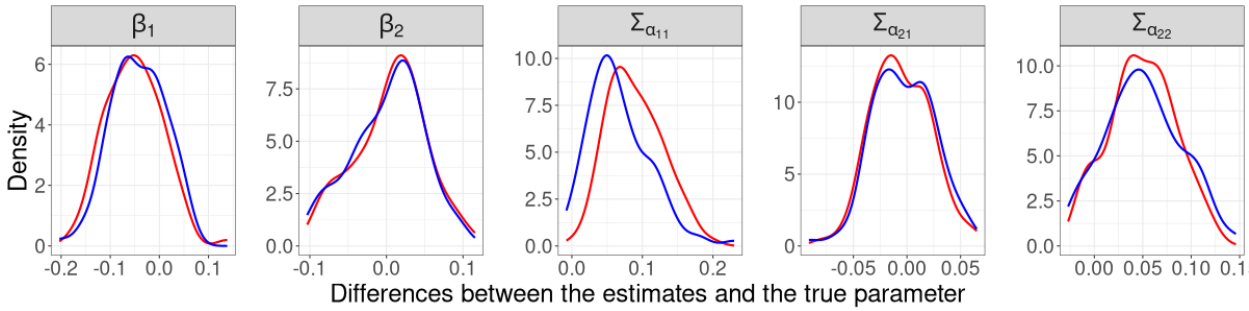

Figure S27: Comparison of the distribution of the differences between the R-VGAL posterior means and the true parameters (in red), against the distribution of the differences between the HMC posterior means and the true parameters (in blue), for 100 datasets simulated from the Poisson mixed model.

S5 Convergence statistics for HMC

This section contains some statistics on the convergence of HMC in the simulated and real data examples in Section 3. In particular, we show the effective sample size (ESS) and $\hat{R}$ values obtained from the `RStan` diagnostics. The ESS is a measure of the number of uncorrelated posterior samples, and the higher the ESS the better; see Gelman et al. (2013) for further details on how STAN calculates ESS. The $\hat{R}$ statistic proposed by Gelman and Rubin (1992) measures the ratio of the average variance of samples within each chain to the variance of the pooled samples across chains; an $\hat{R}$ close to 1 indicates that the chains have converged. Gelman and Rubin (1992) recommends that the independent Markov chains be initialized with diffuse starting values for the parameters and sampled until all values for $\hat{R}$ are below 1.1.

The ESS and $\hat{R}$ values for each parameter in each of the examples we considered in Section 3 are shown in Table S1.

Example	Parameter	ESS	$\hat{R}$
Linear	β_1	42590.45	0.9999355
Linear	β_2	46091.76	0.9999275
Linear	β_3	45614.94	0.9999371
Linear	β_4	44814.06	0.9999292
Linear	σ_α	39572.87	0.9999377
Linear	σ_ϵ	36754.35	0.9999034
Logistic	β_1	14116.662	1.0002822
Logistic	β_2	12974.649	1.0000641
Logistic	β_3	26374.149	1.0001275
Logistic	β_4	31828.214	0.9999238
Logistic	τ	3700.836	1.0004196
Poisson	β_1	14313.12	1.0000291
Poisson	β_2	14433.41	1.0001494
Poisson	$\Sigma_{\alpha_{11}}$	20022.85	1.0000592
Poisson	$\Sigma_{\alpha_{21}}$	17313.42	1.0004479
Poisson	$\Sigma_{\alpha_{22}}$	19660.59	0.9999396
Six City	β_1	2041.219	1.0001195
Six City	β_2	21906.059	0.9999014
Six City	β_3	6902.207	0.9999044
Six City	τ	1406.197	1.0007235
Polypharmacy	β_0	8724.835	1.0001990
Polypharmacy	β_{gender}	7868.562	1.0003835
Polypharmacy	β_{race}	8810.641	1.0000568
Polypharmacy	β_{age}	17864.738	1.0000620
Polypharmacy	β_{M1}	17248.489	0.9999448
Polypharmacy	β_{M2}	15052.711	1.0000169
Polypharmacy	β_{M3}	14499.271	0.9999639
Polypharmacy	β_{IM}	31828.214	0.9999535
Polypharmacy	τ	37289.351	1.0001344

Table S1: Effective sample size and $\hat{R}$ values for the parameters in each model.

S6 Additional examples

This section contains two additional examples: the first one applies the Poisson model in Section 3.3 of the main paper to the Epilepsy dataset from Thall and Vail (1990), and the second involves a bigger synthetic dataset simulated from the logistic mixed model in Section 3.2 of the main paper.

S6.1 Real data example: Poisson mixed model

In this example, we apply R-VGAL on the well-known Epilepsy dataset from Thall and Vail (1990). This dataset includes $N = 59$ epileptic patients who were treated with either a new drug (Progabide) or placebo in a clinical trial. The response variable is the number of seizures patients have during $n = 4$ follow-up periods. We index the patients as $i = 1, \ldots, N$ and the responses for each patient as $j = 1, \ldots, n$. We follow Tan and Nott (2018) and use the following covariates: the logarithm of 1/4 the number of baseline seizures (`Base`); the `Treatment`, coded as 1 for Progabide and 0 for placebo; the log-transformed and centred age, $\widetilde{\texttt{Age}}_i = \texttt{Age}_i - \frac{1}{N}\sum_{i=1}^{N}(\texttt{Age}_i)$, where $\texttt{Age}_i$ is the logarithm of the age of the ith individual; and the follow up period, `Visit`, coded as $\texttt{Visit} = -0.3$ for the first visit, $\texttt{Visit} = -0.1$ for the second, $\texttt{Visit} = 0.1$ for the third and $\texttt{Visit} = 0.3$ for the fourth.

We consider the following model with random slope and random intercept:

$$y_{ij} \sim \text{Poisson}(\lambda_{ij}), \tag{S45}$$

$$\log(\lambda_{ij}) = \beta_0 + \beta_{base}\texttt{Base}_i + \beta_{treatment}\texttt{Treatment}_i + \beta_{age}\widetilde{\texttt{Age}}_i + \beta_{visit}\texttt{Visit}_{ij} + \alpha_{i,1} + \alpha_{i,2}\texttt{Visit}_{ij}, \tag{S46}$$

where $\boldsymbol{\alpha} = (\alpha_{i,1}, \alpha_{i,2})^\top \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\alpha)$, with $\boldsymbol{\Sigma}_\alpha = \mathbf{L}\mathbf{L}^\top$ and $\mathbf{L} = \begin{bmatrix} \exp(\zeta_{11}) & 0 \\ \zeta_{21} & \exp(\zeta_{22}) \end{bmatrix}$. In the algorithm, we consider the unconstrained parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\zeta}^\top)^\top$, where $\boldsymbol{\beta} = (\beta_0, \beta_{base}, \beta_{treatment}, \beta_{age}, \beta_{visit})^\top$ and $\boldsymbol{\zeta} = (\zeta_{11}, \zeta_{22}, \zeta_{21})^\top$. The gradient $\nabla_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \theta)$ and Hessian $\nabla^2_\theta \log p(\mathbf{y}_i, \boldsymbol{\alpha}_i \mid \theta)$, which are necessary in the computation of the gradient and Hessian of the subject-specific log likelihood $\log p(\mathbf{y}_i \mid \boldsymbol{\theta})$, are provided in Section S1.3 of the online supplement.

The initial variational distribution we use is

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

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,

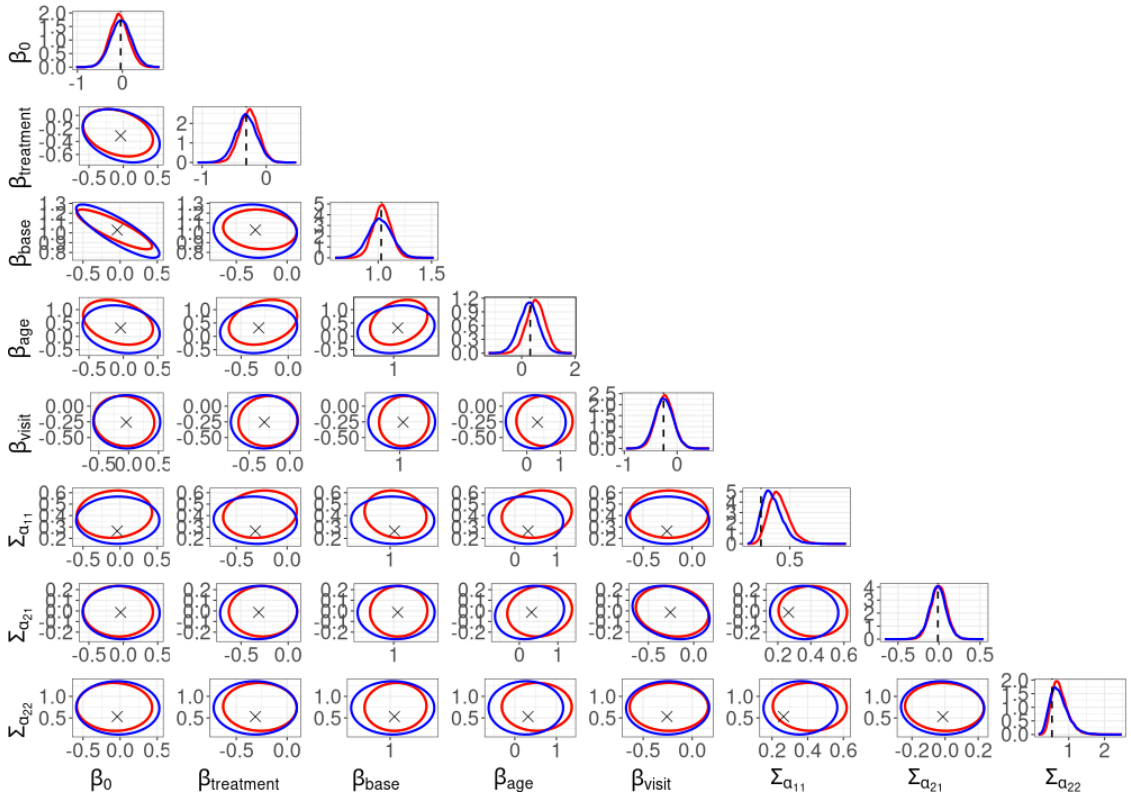

Figure S28: Exact posterior distributions from HMC (in blue) and approximate posterior distributions from R-VGAL with estimated gradients and Hessians (in red) for the Poisson mixed model experiment. Diagonal panels: Marginal posterior distributions with the maximum likelihood estimates marked using dotted lines. Off-diagonal panels: Bivariate posterior distributions with the maximum likelihood estimates marked using the symbol $\times$.

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 simulated Poisson model in Section 3.3, we run damped R-VGAL with $n_{damp} = 10$ observations and $K = 4$ steps per observation. We use $S_\alpha = 200$ Monte Carlo samples in the importance sampling step, and $S = 200$ samples to approximate the expectations with respect to $q_{i-1}(\boldsymbol{\theta})$ in the R-VGAL updates of the variational mean and the precision matrix. Figure S28 shows the marginal posterior distributions with maximum likelihood estimates of the parameters, along with bivariate posterior distributions as estimated using R-VGAL and HMC. We find that the estimates from R-VGAL and HMC agree quite well for all parameters, and posterior modes for all parameters from both methods are close to the maximum likelihood estimates.

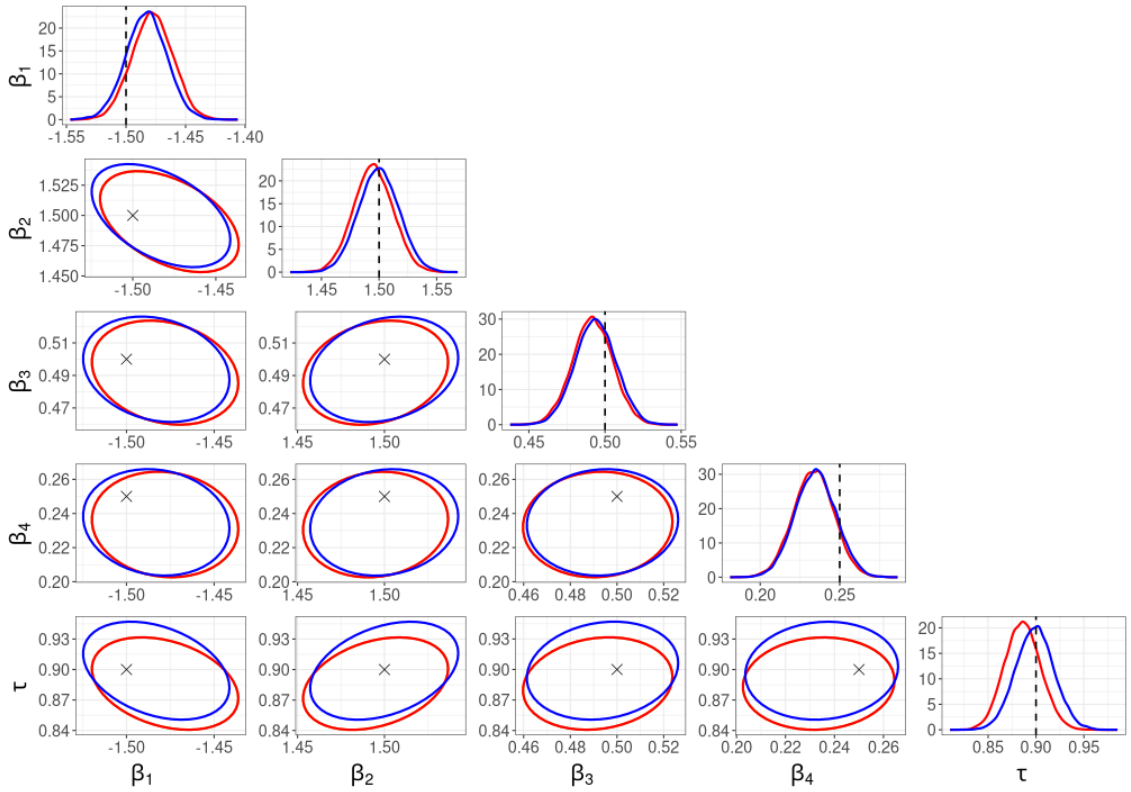

Figure S29: Exact posterior distributions from HMC (in blue) and approximate posterior distributions from R-VGAL with estimated gradients and Hessians (in red) for the logistic mixed model experiment with 50000 observations. Diagonal panels: Marginal posterior distributions with true parameters denoted using dotted lines. Off-diagonal panels: Bivariate posterior distributions with true parameters denoted using the symbol $\times$.

S6.2 Big data example: Logistic mixed model

For this experiment, we simulate data from the logistic mixed model in Section 3.2 with $N = 5000$ and $n = 10$, resulting in a total of 50000 observations. We find that it is necessary to increase the number of Monte Carlo samples to $S = 200$ and $S_\alpha = 200$ to achieve accurate posterior estimates in this example. The prior distribution and the settings for damping in R-VGAL are kept the same as in the example in Section 3.2.

Figure S29 shows the marginal posterior densities, along with bivariate posterior plots from R-VGAL and HMC. As with previous simulations, the posterior distributions obtained using R-VGAL are very similar to those obtained using HMC.

S7 Application of R-VGAL to models with other random effect structures

In the main paper, we considered the application of R-VGAL to models where the random effects are correlated within subjects (individuals), but independent between subjects. In practice, there are many cases where other random effect structures, such as crossed or nested random effects, are needed; see Pinheiro and Bates (2006); Gelman and Hill (2007); West et al. (2014) or Papaspiliopoulos et al. (2023) for examples. Here, we briefly discuss the application of R-VGAL to some classes of models with crossed or nested random effects. The implementation of R-VGAL to these models is left for future endeavours.

Models with crossed effects are often used to model data that can be organised in the form of contingency tables between categorical variables. For example, consider a study of annual income, in which a number of characteristics from participants are recorded using categorical variables, such as their age range, sex, ethnicity, and highest level of education. In this case, the data can be organised into a multi-dimensional contingency table between age range, sex, ethnicity, and level of education. Participants who have the same combination of characteristics may have similar income levels, and the correlation between people in the same set of categories may be modelled with the addition of category-specific random effects.

The notation we use in the following crossed effect model follows that in Chapter 11 of Gelman and Hill (2007) and Section 2 of Papaspiliopoulos et al. (2023). Suppose that there are K categorical variables, and the kth variable has L_k levels, for $k = 1, \ldots, K$. The logarithm of the income of the ith individual may then be modelled as

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{k=1}^{K} \alpha^{(k)}_{l_k[i]} + \epsilon_i, \quad \alpha^{(k)}_l \sim N(0, \sigma^2_\alpha), \quad \epsilon_i \sim N(0, \sigma^2_\epsilon), \tag{S47}$$

for $i = 1, \ldots, N$, where $\mathbf{x}_i$ denotes a vector of covariates associated with the fixed effects $\boldsymbol{\beta}$, and $\alpha^{(k)}_l$ denotes the random effect associated with the lth level of the kth categorical variable. The notation $l_k[i]$ denotes the level of the kth category that the ith individual falls into; for example, if the first categorical variable in the model is age range, where the categories are 1 for 18–30 years old, 2 for 30–50 years old, and 3 for 50 years old and above, then $l_1[4] = 2$ means that the 4th individual in the dataset is in level 2 of the "age" variable (between 30 and 50 years old). Here we have assumed that the random effects $\alpha^{(k)}_l$ have the same variance for all $l = 1, \ldots, L$ and $k = 1, \ldots, K$. Thus the parameters of interest in this model are $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2_\alpha, \sigma^2_\epsilon)^\top$.

In this model, there are $G = L_1 \times L_2 \times \ldots L_K$ combinations of levels. The R-VGAL updates would therefore be based on the log-likelihood of all individuals within the same combination (group). For each combination

(group) $g = 1, \ldots, G$, the vector of responses of people in this group is denoted as $\mathbf{y}_g$. Then for each group $g = 1, \ldots, G$, the gradient of the group-specific log likelihood can be expressed via Fisher's identity as

$$\nabla_\theta \log p(\mathbf{y}_g \mid \boldsymbol{\theta}) = \int \nabla_\theta \log p(\mathbf{y}_g, \alpha^{(1)}_{l_{1,g}}, \ldots, \alpha^{(K)}_{l_{K,g}} \mid \boldsymbol{\theta})\, p(\alpha^{(1)}_{l_{1,g}}, \ldots, \alpha^{(K)}_{l_{K,g}} \mid \mathbf{y}_g, \boldsymbol{\theta})\, \mathrm{d}\alpha^{(1)}_{l_{1,g}} \ldots \mathrm{d}\alpha^{(K)}_{l_{K,g}}, \tag{S48}$$

where, here, the notation $l_{k,g}$ denotes the level of the kth categorical variable associated with group g. The Hessian of the group log likelihood can be similarly expressed via Louis' identity (22), which we do not restate here. The gradient and Hessian of the group log likelihood can then be approximated using the importance-sampling-based approach described in Sections 2.3.1 and 2.3.2. Note that there are analytical formulae for the gradient (and Hessian) in this case, as the model is linear; but this approach is applicable to a wide class of GLMMs with crossed random effects.

R-VGAL is also applicable to models with nested random effects. One popular example of such models is that of students being nested within classes, which are then nested within schools (see, for example, Chapter 4 of West et al. (2014)). Suppose that we are interested in the final exam marks of Year 12 students across a number of schools in the year 2023. If we write the final mark of the kth student in the jth class at the ith school as y_{ijk}, for $i = 1, \ldots, N$, $j = 1, \ldots, n_i$ and $k = 1, \ldots, n_{ij}$, then a simple linear model with one random intercept at both the school level and the class level is

$$y_{ijk} = \mathbf{x}_{ijk}^\top \boldsymbol{\beta} + \alpha_i + \gamma_{ij} + \epsilon_{ijk}, \quad \alpha_i \sim N(0, \sigma_\alpha^2), \quad \gamma_{ij} \sim N(0, \sigma_\gamma^2), \quad \epsilon_{ijk} \sim N(0, \sigma_\epsilon^2), \tag{S49}$$

where $\mathbf{x}_{ijk}$ is a vector of fixed covariates for the kth student (which may include, for instance, the average number of hours they spend studying per week, or the average number of hours slept per night), $\boldsymbol{\beta}$ are the corresponding fixed effects, α_i is the random effect associated with the school that the student attends, γ_{ij} is the random effect associated with the class they are in at their school, and ϵ_{ijk} is an error term associated with each student.

At the class level, the model (S49) may be written as

$$\mathbf{y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{1}_{n_{ij}}\alpha_i + \mathbf{1}_{n_{ij}}\gamma_{ij} + \boldsymbol{\varepsilon}_{ij}, \quad \alpha_i \sim N(0, \sigma_\alpha^2), \quad \gamma_{ij} \sim N(0, \sigma_\gamma^2), \quad \boldsymbol{\varepsilon}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\epsilon,i}), \tag{S50}$$

where $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijn_{ij}})^\top$ is a vector containing final exam marks for all students in class j of school i, $\mathbf{X}_i = (\mathbf{x}_{ij1}^\top, \ldots, \mathbf{x}_{ijn_{ij}}^\top)^\top$ is a design matrix containing fixed covariates from all students in class j of school i, $\mathbf{1}_d$ denotes a $d \times 1$ vector of ones, and $\boldsymbol{\varepsilon}_{ij} = (\epsilon_{ij1}, \ldots, \epsilon_{ijn_{ij}})^\top$, for classes $j = 1, \ldots, n_i$ and schools $i = 1, \ldots, N$. For simplicity, we assume here that the error terms $\epsilon_{ijk}, k = 1, \ldots, n_{ij}$, are uncorrelated, so

that $\boldsymbol{\Sigma}_{\epsilon,i} = \sigma_\epsilon^2 \mathbf{I}_{n_{class,i}}$, where $\mathbf{I}_d$ denotes a $d \times d$ identity matrix, and $n_{class,i}$ is the number of students in each class at school i, which we assume does not change with class; that is, we let $n_{ij} = n_{class,i}$, for $j = 1, ..., n_i$ and $i = 1, ..., N$.

We will now go one step further and write the model at the school level as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\alpha_i + \mathbf{W}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i, \quad \alpha_i \sim N(0, \sigma_\alpha^2), \quad \boldsymbol{\gamma}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\gamma_i}), \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma}_{\epsilon,i}), \tag{S51}$$

where $\otimes$ denotes the Kronecker product between two matrices, $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i})^\top$, $\mathbf{Z}_i = \mathbf{1}_{n_i n_{class,i}}$, $\mathbf{W}_i = \mathbf{I}_{n_i} \otimes \mathbf{1}_{n_{class,i}}$, $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{in_i})^\top$ is a vector of all random effects from classes $j = 1, \ldots, n_i$ in school i, and $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{i1}^\top, \ldots, \boldsymbol{\varepsilon}_{in_i}^\top)^\top$ is a vector of random error terms for all students in school i, $i = 1, \ldots, N$. Here we allow the random effects $\gamma_{ij}, j = 1, \ldots, n_i$ to be correlated between classes in the same school, so that for each school $i = 1, \ldots, N$, the covariance matrix $\boldsymbol{\Sigma}_{\gamma_i}$ is potentially dense. We also assume that the school-specific random effects α_i are independent and identically distributed for all schools $i = 1, \ldots, N$.

The parameters of interest in this model are $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma_\alpha^2, \boldsymbol{\Sigma}_{\gamma_1}, \ldots, \boldsymbol{\Sigma}_{\gamma_N}, \sigma_\epsilon^2\}$. To make inference on these parameters, the R-VGAL updates can be performed by processing the data one school at a time. That is, for $i = 1, \ldots, N$, the R-VGAL updates are made in terms of the gradient and Hessian of the school-specific log-likelihood, $\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ and $\nabla_\theta^2 \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$. Using Fisher's identity (19), the gradient $\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ can be written as

$$\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta}) = \int p(\alpha_i, \boldsymbol{\gamma}_i \mid \boldsymbol{\theta}) \nabla_\theta \log p(\mathbf{y}_i, \alpha_i, \boldsymbol{\gamma}_i \mid \boldsymbol{\theta}) \, \mathrm{d}\alpha_i \, \mathrm{d}\boldsymbol{\gamma}_i, \quad i = 1, \ldots, N, \tag{S52}$$

where the term $\log p(\mathbf{y}_i, \alpha_i, \boldsymbol{\gamma}_i \mid \boldsymbol{\theta})$ can be expanded further as

$$\log p(\mathbf{y}_i, \alpha_i, \boldsymbol{\gamma}_i \mid \boldsymbol{\theta}) = \log p(\mathbf{y}_i \mid \alpha_i, \boldsymbol{\gamma}_i, \boldsymbol{\theta}) + \log p(\alpha_i, \boldsymbol{\gamma}_i \mid \boldsymbol{\theta}), \tag{S53}$$

with

$$p(\mathbf{y}_i \mid \alpha_i, \boldsymbol{\gamma}_i, \boldsymbol{\theta}) = N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\alpha_i + \mathbf{W}_i\boldsymbol{\gamma}_i, \boldsymbol{\Sigma}_{\epsilon,i} \otimes \mathbf{I}_{n_i}),$$
$$p(\alpha_i, \boldsymbol{\gamma}_i \mid \boldsymbol{\theta}) = N\left(\begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_\alpha^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\gamma_i} \end{bmatrix}\right), \quad i = 1, \ldots, N,$$

based on (S51). Approximation of the gradient $\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ can then proceed via the importance-sampling-based approach in Section 2.3.1, and a similar procedure for the Hessian $\nabla_\theta \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$ can be

done as shown in Section 2.3.2. As the model is linear in this case, we note that the log-likelihood, gradient, and Hessian are available analytically. However, this approach is applicable to a wide class of GLMMs with nested random effects.

References

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, New York, NY, 3rd edition.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Papaspiliopoulos, O., Stumpf-Fétizon, T., and Zanella, G. (2023). Scalable Bayesian computation for crossed and nested hierarchical models. *Electronic Journal of Statistics*, 17(2):3575–3612.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects Models in S and S-PLUS*. Springer Science and Business Media, New York, NY.

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

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

West, B. T., Welch, K. B., and Galecki, A. T. (2014). *Linear Mixed Models: A Practical Guide Using Statistical Software*. CRC Press, Boca Raton, FL.