Submitted to the Annals of Applied Statistics

FAST APPROXIMATE BAYESIAN INFERENCE FOR SMALL-AREA ESTIMATION OF HIV INDICATORS USING THE NAOMI MODEL

BY ADAM HOWES [1,4] , ALEX STRINGER [2]
SETH R. FLAXMAN [3] , JEFFREY W. EATON [4]

[1]*Department of Mathematics, Imperial College London, ath19@ic.ac.uk*

[2]*Department of Statistics and Actuarial Science, University of Waterloo, alex.stringer@uwaterloo.ca*

[3]*Department of Computer Science, University of Oxford, seth.flaxman@cs.ox.ac.uk*

[4]*MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, jeffrey.eaton@imperial.ac.uk*

Naomi is a spatial evidence synthesis model used to produce district-level HIV epidemic indicators in sub-Saharan Africa. Multiple outcomes of policy interest, including HIV prevalence, HIV incidence and antiretroviral therapy treatment coverage, are jointly modelled using both household survey data and routinely reported health system data. Inference for Naomi is currently conducted using an empirical Bayes Gaussian approximation via the `TMB` R package. We propose a new inference method combining adaptive Gauss-Hermite quadrature together with the simplified integrated nested Laplace approximation approach of Wood (2020) to enable fast and accurate inference for Naomi and other extended latent Gaussian models. Using data from Malawi, our method provides more accurate inferences than `TMB`, and is comparable to Hamiltonian Monte Carlo with the No-U-Turn sampler, but faster to run. By extending the `aghq` R package we facilitate easy, flexible use of our method when provided a `TMB` C++ template for the model's log-posterior. In doing so, we enable inference via integrated nested Laplace approximations for a larger class of models than was previously possible.

1. Introduction. To mount an effective public health response to the HIV epidemic, it's crucial to have accurate, timely estimates of HIV indicators at the geographic level at which health systems are planned and delivered. However, producing these estimates is challenging due to the shortcomings of all available data sources. Nationally-representative household surveys provide the most statistically reliable data, but due to their high cost, in most countries they are conducted only every five years or so, with limited sample size at the district level. Other data sources, such as routine health surveillance of antenatal care (ANC) clinics, are available in more real-time but based on limited or non-representative samples of the population. To address challenges, the Naomi small-area estimation model (Eaton et al., 2021) synthesises data from multiple sources to estimate HIV indicators at a district-level. Modelling multiple data sources jointly has many benefits, including mitigating the limitations of any single source, increasing statistical power, and prompting investigation into any conflicts of information between sources. Software (`https://naomi.unaids.org`) has been developed for Naomi, allowing countries to input their data and interactively generate estimates in a yearly process supported by UNAIDS. Creation of estimates by country teams, rather than external agencies, is a noteworthy feature of the HIV response. Drawing on expertise closest to the data being modelled improves the accuracy of the process, as well as strengthening trust and ownership of the resulting estimates.

Keywords and phrases: spatial statistics, small-area estimation, INLA, AGHQ, HIV epidemiology.

Practical requirements for the model, in combination with its relative complexity, present a difficult Bayesian inference problem. Any inferential strategy must be fast enough for interactive review and iteration of modelling results, as well as easy to run in production. Markov chain Monte Carlo (MCMC) approaches are prohibitively slow due to the scale of the model and challenging features of its posterior geometry (Neal, 2003). Inference is currently conducted using an empirical Bayes (EB) approach, with a Gaussian approximation to the latent field, via the Template Model Builder (`TMB`) R package (Kristensen et al., 2016). Owing to its speed and flexibility, `TMB` has recently been gaining popularity more broadly in spatial statistics (Osgood-Zimmerman and Wakefield, 2022). Inference in `TMB` is based on optimisation of a C++ template function, with the option available to use a Laplace approximation to integrate out any subset of the function arguments. In the Naomi model, this subset is the high-dimensional latent field. Taking inspiration from the AD Model Builder (ADMB) package (Fournier et al., 2012), `TMB` uses automatic differentiation (Baydin et al., 2017) to calculate the derivatives required for numerical optimisation routines and the Laplace approximation. Although this approach is fast, we would prefer to account more fully for hyperparameter uncertainty than as is the case in an empirical Bayes framework. This has motivated us to look for an approach closer to full Bayesian inference, which is also flexible enough to be compatible with the model, as well as fast enough to be run in production by country teams.

To obtain fast, accurate Bayesian inferences for the Naomi model we develop an inference methodology which combines adaptive Gauss-Hermite quadrature (AGHQ) with the simplified integrated nested Laplace approximation (INLA) approach of Wood (2020). INLA is an approximate inference approach based on nested Laplace approximations and numerical quadrature. The key innovation of Rue, Martino and Chopin (2009) is an approximation which enables accurate latent field posterior marginals without explicitly computing the full Laplace approximation for each element. Simplified INLA (Wood, 2020) extends INLA by relaxing the sparsity assumptions on precision matrix of the latent field required for this approximation to be accurate. This extension facilitates inference for models like Naomi, which are not quite latent Gaussian models (LGMs) and so were not previously amenable to inference with INLA. In particular, due to dependence of observations on multiple structured additive predictors, Naomi falls into the class of extended latent Gaussian models (ELGMs) (Stringer, Brown and Stafford, 2022). We combine simplified INLA with AGHQ, a quadrature rule based on the theory of polynomial interpolation which adapts to the integrand based on the Hessian at the mode. Though no theory yet exists for the nested case, the first stochastic convergence results for adaptive quadrature rules were recently obtained by Bilodeau, Stringer and Tang (2022) using AGHQ. We implement our method as an extension of the `aghq` R package (Stringer, 2021). Since `aghq` is designed to naturally interface with `TMB`, use of our method is simple when provided a C++ user template for the log-posterior.

Other work aiming to extend the scope of the INLA method includes the `inlabru` R package (Bachl et al., 2019), INLA within MCMC (Gómez-Rubio and Rue, 2018), and importance sampling with INLA (Berild et al., 2022), all of which leverage the `R-INLA` R package (Martins et al., 2013). `inlabru` approximates non-linear predictors using linearisation, which involves making iterative calls to `R-INLA`. INLA within MCMC and importance sampling with INLA are suitable for models which are LGMs conditional on some subset of the parameters being fixed.

The remainder of this paper is organised as follows. Section 2 outlines the version of the Naomi model that we consider in this paper, and Section 3 describes how it falls within the ELGM framework. Section 4 outlines our approach to fast, accurate Bayesian inference for ELGMs using simplified INLA and AGHQ. As a case-study, we compare the accuracy of our inference method to `TMB` and `tmbstan` for the simplified Naomi model fit to data from Malawi, in Section 2. We also demonstrate a Bayesian workflow, illustrating the applicability

of these tools in a deterministic inference setting. Finally, in Section 6 we discuss our conclusions, how we anticipate our method might be useful for other models, and directions for future research.

2. A simplified Naomi model. Eaton et al. (2021) specify a joint model linking three small-area estimation models, and defined over three time points: T_1 the time of the most recent household survey with HIV testing; T_2, the current time period; and T_3, a short term projection period. We consider a simplified version defined only at T_1 omitting nowcasting and temporal projection, as these time points involve limited additional inference. An overview of this simplified model is given below, highlighting the aspects which make it a challenge for existing inferential approaches. A more complete mathematical description (Appendix S1) as well as a C++ template for the log-posterior (Appendix S3) are provided in the supplementary material.

2.1. *Household survey component* . Consider a country in sub-Saharan Africa where a household survey with complex design has taken place at time T_1. Let $x \in \mathcal{X}$ index district, $a \in \mathcal{A}$ index five-year age band, and $s \in \mathcal{S}$ index sex. For ease of notation, let i index the finest district-age-sex division included in the model. The data we observe may be aggregated over indices i. Let $\mathcal{I} \subseteq \mathcal{X} \times \mathcal{A} \times \mathcal{S}$ be a set of indicies i for which an observation is reported.

Let $N_i \in \mathbb{N}$ be the known, fixed population size. We infer the following unknown HIV indicators: HIV prevalence $\rho_i \in [0,1]$, the proporiton of indviduals who are HIV positive; antiretroviral therapy (ART) coverage $\alpha_i \in [0,1]$, the proportion of people living with HIV who receive ART treatment; and annual HIV incidence rate $\lambda_i > 0$, the yearly rate of new HIV infections occurring. Independent logistic regression models for HIV prevalence and ART coverage in the general population are specified such that

$$\begin{aligned}\text{logit}(\rho_i) &= \eta_i^\rho,\\ \text{logit}(\alpha_i) &= \eta_i^\alpha,\end{aligned}$$

for certain choice of linear predictors η_i^ρ and η_i^α. HIV incidence rate is modelled on the log scale as $\log(\lambda_i) = \eta_i^\lambda(\{\rho_i, \alpha_i\}_{i\in\mathcal{I}})$, where the linear predictor depends on $\{\rho_i, \alpha_i\}_{i\in\mathcal{I}}$ for some $\mathcal{I}$. Finally, let κ_i be the proportion recently infected among HIV positive persons, which we link to HIV incidence via

$$\kappa_i = 1 - \exp\left(-\lambda_i \cdot \frac{1-\rho_i}{\rho_i} \cdot (\Omega_T - \beta_T) - \beta_T\right),$$

where the mean duration of recent infection Ω_T and false recent ratio β_T are strongly informed by priors for the particular survey.

For $\theta \in \{\rho, \alpha, \kappa\}$ we calculate the weighted, aggregated survey observations

$$\hat{\theta}_\mathcal{I} = \frac{\sum_j w_j \cdot \theta_j}{\sum_j w_j},$$

where j indexes individuals across all strata $i \in \mathcal{I}$. The design weights are given by

$$w_j = \frac{1}{\pi_j} \times \frac{1}{\omega_j},$$

where π_j is the probability of inclusion and ω_j is a non-response factor for the particular survey. We calculate the observed number of indicator cases as $y_\mathcal{I}^{\hat{\theta}} = m_\mathcal{I}^{\hat{\theta}} \cdot \hat{\theta}_\mathcal{I}$ where

$$m_\mathcal{I}^{\hat{\theta}} = \frac{\left(\sum_j w_j\right)^2}{\sum_j w_j^2},$$

is the Kish effective sample size (Kish, 1965). We use a binomial working likelihood

$$y_{\mathcal{I}}^{\hat{\theta}} \sim \text{xBin}(m_{\mathcal{I}}^{\hat{\theta}}, \theta_{\mathcal{I}})$$

to model these aggregate observations, where $\theta_{\mathcal{I}}$ are the following weighted aggregates

$$\rho_{\mathcal{I}} = \frac{\sum_{i\in\mathcal{I}} N_i\rho_i}{\sum_{i\in\mathcal{I}} N_i}, \quad \alpha_{\mathcal{I}} = \frac{\sum_{i\in\mathcal{I}} N_i\rho_i\alpha_i}{\sum_{i\in\mathcal{I}} N_i\rho_i}, \quad \kappa_{\mathcal{I}} = \frac{\sum_{i\in\mathcal{I}} N_i\rho_i\kappa_i}{\sum_{i\in\mathcal{I}} N_i\rho_i}.$$

2.2. *ANC testing component* . HIV prevalence ρ_i^{ANC} and ART coverage α_i^{ANC} among pregnant women are modelled as offset on the logit scale from the general population indicators as follows

$$\text{logit}(\rho_i^{\text{ANC}}) = \text{logit}(\rho_i) + \eta_i^{\rho^{\text{ANC}}},$$
$$\text{logit}(\alpha_i^{\text{ANC}}) = \text{logit}(\alpha_i) + \eta_i^{\alpha^{\text{ANC}}}.$$

These processes are informed by likelihoods specified for the following aggregate ANC data from the year of the most recent survey: the number of ANC clients with ascertained status $x_{\mathcal{I}}^{\text{ANC}}$, the number of those with positive status $y_{\mathcal{I}}^{\text{ANC}}$, and the number of ANC clients already on ART prior to their first ANC visit $z_{\mathcal{I}}^{\text{ANC}}$. We use the binomial working likelihoods

$$y_{\mathcal{I}}^{\text{ANC}} \sim \text{Bin}(x_{\mathcal{I}}^{\text{ANC}}, \rho_{\mathcal{I}}^{\text{ANC}}),$$
$$z_{\mathcal{I}}^{\text{ANC}} \sim \text{Bin}(y_{\mathcal{I}}^{\text{ANC}}, \alpha_{\mathcal{I}}^{\text{ANC}}),$$

where, again, we use weighted aggregates

$$\rho_{\mathcal{I}}^{\text{ANC}} = \frac{\sum_{i\in\mathcal{I}} \Psi_i\rho_i^{\text{ANC}}}{\sum_{i\in\mathcal{I}} \Psi_i}, \quad \alpha_{\mathcal{I}}^{\text{ANC}} = \frac{\sum_{i\in\mathcal{I}} \Psi_i\rho_i^{\text{ANC}}\alpha_i^{\text{ANC}}}{\sum_{i\in\mathcal{I}} \Psi_i\rho_i^{\text{ANC}}},$$

with Ψ_i the number of pregnant women.

2.3. *ART attendance component* . We use a multinomial logistic regression model to account for people living with HIV choosing to access ART services outside of the district that they reside in. Let $\gamma_{x,x'} \in [0,1]$ be the probability that a person on ART residing in district x receives ART in district x'. We assume that $\gamma_{x,x'} = 0$ unless $x = x'$ or the two districts are adjacent such that $x \sim x'$. We model the log-odds $\tilde{\gamma}_{x,x'} = \text{logit}(\gamma_{x,x'})$ using a linear predictor $\eta_x^{\tilde{\gamma}}$ which only depends on the home district x, such that travel to each neighbouring district, for all age-sex strata, is equally likely. We model aggregate ART attendance data $\dot{A}_{\mathcal{I}}$ using a Gaussian approximation to a sum of binomials. The sum results both by aggregation over $i \in \mathcal{I}$ and by number of ART clients travelling from district x' to x. More details regarding this part of the model are provided in Appendix S1.

3. Extended Latent Gaussian models. Latent Gaussian models (LGMs) (Rue, Martino and Chopin, 2009) are three-stage hierarchical models of the form

$$y_i \sim p(y_i \,|\, \eta_i, \boldsymbol{\theta}_1), \quad i \in [n]$$
$$\mu_i = \mathbb{E}(y_i \,|\, \eta_i) = g(\eta_i),$$
$$\eta_i = \beta_0 + \sum_{l=1}^{p} \beta_j z_{ji} + \sum_{k=1}^{r} f_k(u_{ki}),$$

where $[n] = \{1, \ldots, n\}$. The response variable is $\mathbf{y} = (y)_{i\in[n]}$ with likelihood $p(\mathbf{y} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}_1) = \prod_{i=1}^{n} p(y_i \,|\, \eta_i, \boldsymbol{\theta}_1)$, where $\boldsymbol{\eta} = (\eta)_{i\in[n]}$. Each response has conditional mean μ_i with inverse

link function $g : \mathbb{R} \to \mathbb{R}$ such that $\mu_i = g(\eta_i)$. The vector $\boldsymbol{\theta}_1 \in \mathbb{R}^s$, with s_1 assumed small, are additional parameters of the likelihood. The structured additive predictor η_i may include an intercept β_0, linear effects β_j of the covariates z_{ji}, and unknown functions $f_k(\cdot)$ of the covariates u_{ki}. The parameters β_0, $\{\beta_j\}$, $\{f_k(\cdot)\}$ are each assigned Gaussian priors. It is convenient to collect these parameters into a vector $\mathbf{x} \in \mathbb{R}^N$ called the latent field such that $\mathbf{x} \sim \mathcal{N}(0, \boldsymbol{Q}(\boldsymbol{\theta}_2)^{-1})$ where $\boldsymbol{\theta}_2 \in \mathbb{R}^{s_2}$ are further parameters, again with s_2 assumed small. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}^s$ with $m = s_1 + s_2$ be all hyperparameters, with prior $p(\boldsymbol{\theta})$.

Extended latent Gaussian models (ELGMs) (Stringer, Brown and Stafford, 2022) relax the restriction that there is a one-to-one mapping between the mean response $\boldsymbol{\mu}$ and structured additive predictor $\boldsymbol{\eta}$. Instead, the structured additive predictor is redefined as $\boldsymbol{\eta} = (\eta)_{i \in [N_n]}$, where $N_n \in \mathbb{N}$ is a function of n, and it is possible that $N_n \neq n$. Each mean response μ_i now depends on some subset $\mathcal{J}_i \subseteq [N_n]$ of indices of $\boldsymbol{\eta}$, with $\cup_{i=1}^n \mathcal{J}_i = [N_n]$ and $1 \leq |\mathcal{J}_i| \leq N_n$. The inverse link function $g(\cdot)$ is redefined for each observation to be a possibly many-to-one mapping $g_i : \mathbb{R}^{|\mathcal{J}_i|} \to \mathbb{R}$, such that $\mu_i = g_i(\boldsymbol{\eta}_{\mathcal{J}_i})$. Importantly, this mapping allows for the presence of more non-linearity in the model. ELGMs are then of the form

$$
\begin{aligned}
y_i &\sim p(y_i \,|\, \boldsymbol{\eta}_{\mathcal{J}_i}, \boldsymbol{\theta}_1), \quad i \in [n] \\
\mu_i &= \mathbb{E}(y_i \,|\, \boldsymbol{\eta}_{\mathcal{J}_i}) = g_i(\boldsymbol{\eta}_{\mathcal{J}_i}), \\
\eta_j &= \beta_0 + \sum_{l=1}^{p} \beta_j z_{ji} + \sum_{k=1}^{r} f_k(u_{ki}), \quad j \in [N_n].
\end{aligned}
$$

Naomi is not an LGM, and instead falls into the ELGM class, for the following reasons:

1. In the household survey component (Section 2.1), the HIV incidence rate depends on the HIV prevalence and ART coverage linear predictors.
2. In the ANC testing component (Section 2.2), the HIV prevalence and ART coverage depend upon the household survey component. Specifically, $|\mathcal{J}_i| = 2$ such that for $\theta \in \{\rho, \alpha\}$
$$\mu_i = g_i(\eta_i^{\theta}, \eta_i^{\theta^{\text{ANC}}}) = \text{logit}^{-1}(\eta_i^{\theta} + \eta_i^{\theta^{\text{ANC}}}).$$
3. In the ART attendance component (Section 2.3), the multinomial logistic regression.
4. In all components the data we model is aggregated. Processes are modelled at the finest distict-age-sex division i, but likelihoods are defined for observations over sets of indices $\mathcal{I}$.

4. Fast approximate inference method. The joint posterior of $(\mathbf{x}, \boldsymbol{\theta})$ for an ELGM is given by

$$
p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}) \propto p(\boldsymbol{\theta}) |\mathbf{Q}(\boldsymbol{\theta})|^{n/2} \exp\left(-\frac{1}{2} \mathbf{x}^\top \mathbf{Q}(\boldsymbol{\theta}) \mathbf{x} + \sum_{i=1}^{n} \log p(y_i \,|\, \mathbf{x}_{\mathcal{J}_i}, \boldsymbol{\theta}) \right).
$$

We consider approximations to the posterior marginals of each latent random variable x_i and hyperparameter θ_j given by

$$
\tilde{p}(x_i \,|\, \mathbf{y}) \approx p(x_i \,|\, \mathbf{y}) = \int p(x_i, \boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta} = \int p(x_i \,|\, \boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta}, \quad i \in [N], \tag{4.1}
$$

$$
\tilde{p}(\theta_j \,|\, \mathbf{y}) \approx p(\theta_j \,|\, \mathbf{y}) = \int p(\boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta}_{-j} \quad j \in [m]. \tag{4.2}
$$

Given the negative unnormalised log posterior $-\log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$, we obtain the above posterior marginal approximations $\{\tilde{p}(x_i \,|\, \mathbf{y})\}_{i=1}^n$ and $\{\tilde{p}(\theta_j \,|\, \mathbf{y})\}_{j=1}^m$ via nested applications of the Laplace approximation and AGHQ.

4.1. *Laplace approximation.*

4.2. *AGHQ.* A quadrature rule can be used to approximate the integral of a function $f : \mathcal{Z} \to \mathbb{R}$ by a weighted sum

$$\int_{\mathbf{z}\in\mathcal{Z}} f(\mathbf{z})\mathrm{d}\mathbf{z} \approx \sum_{\mathbf{z}\in\mathcal{Q}} f(\mathbf{z})w(\mathbf{z})$$

where $\mathbf{z} \in \mathcal{Q} \subset \mathcal{Z}$ are a set of nodes and $\omega : \mathcal{Q} \to \mathbb{R}$ is a weighting function. In Gauss-Hermite quadrature (Davis and Rabinowitz, 1975) the nodes are selected as zeros of the Hermite polynomials. These nodes and weights can be shifted and scaled to suit the particular integrand using adaptive Gauss-Hermite quadrature (Naylor and Smith, 1982; Tierney and Kadane, 1986).

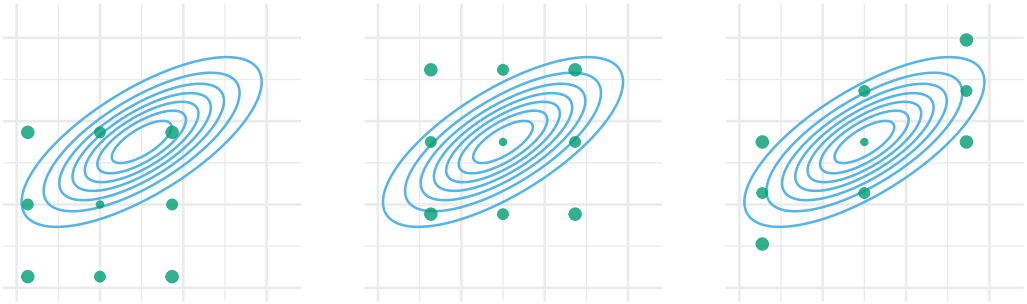

FIGURE 1. *The Gauss-Hermite quadrature nodes* $\mathbf{z} \in \mathcal{Q}(2, 3)$ *for this two dimensional integral with three nodes per dimension are adapted based on the mean and covariance matrix of the target via* $\boldsymbol{\theta}(\mathbf{z}) = \hat{\boldsymbol{\theta}} + \mathbf{L}\mathbf{z}$.

4.3. *Algorithm.*

1. Calculate the mode, Hessian at the mode, and lower Cholesky

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \tilde{p}_{\mathrm{LA}}(\boldsymbol{\theta}, \mathbf{y}),$$
$$\mathbf{H} = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} - \log\tilde{p}_{\mathrm{LA}}(\boldsymbol{\theta}, \mathbf{y})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$
$$\mathbf{H}^{-1} = \mathbf{L}\mathbf{L}^\top,$$

   of the Laplace approximation

$$\tilde{p}_{\mathrm{LA}}(\boldsymbol{\theta}, \mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{\tilde{p}_{\mathrm{G}}(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})}\Big|_{\mathbf{x}=\hat{\mathbf{x}}(\boldsymbol{\theta})}$$

   where $\tilde{p}_{\mathrm{G}}(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}(\mathbf{x} \,|\, \hat{\mathbf{x}}(\boldsymbol{\theta}), \mathbf{H}(\boldsymbol{\theta})^{-1})$ is a Gaussian approximation to $p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})$ with mode and precision matrix given by

$$\hat{\mathbf{x}}(\boldsymbol{\theta}) = \arg\min_{\mathbf{x}} -\log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}),$$
$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial x \partial x^\top} - \log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})|_{\mathbf{x}=\hat{\mathbf{x}}(\boldsymbol{\theta})}.$$

2. Generate a set of nodes $\mathbf{u} \in \mathcal{Q}(m, k)$ and weights $\omega : \mathbf{u} \to \mathbb{R}$ from a Gauss-Hermite quadrature rule with k nodes per dimension, which are then adapted based on the mode

and lower Choleksy via $\boldsymbol{\theta}(\mathbf{u}) = \hat{\boldsymbol{\theta}} + \mathbf{L}\mathbf{u}$. If possible $k \geq 3$ is preferred, though the number of grid points scales exponentially with choice of k. Then, use this quadrature rule to calculate the normalising constant $\tilde{p}_{\texttt{AQ}}(\mathbf{y})$ as follows

$$\tilde{p}_{\texttt{AQ}}(\mathbf{y}) = \sum_{\mathbf{u} \in \mathcal{Q}(m,k)} \tilde{p}_{\texttt{LA}}(\boldsymbol{\theta}(\mathbf{u}), \mathbf{y}) \omega(\mathbf{u}). \tag{4.3}$$

3. For $i \in [n]$ generate l nodes $x_i(\mathbf{v})$ via a Gauss-Hermite quadrature rule $\mathbf{v} \in \mathcal{Q}(1, l)$ adapted based on the mode $\hat{\mathbf{x}}(\boldsymbol{\theta})_i$ and standard deviation $\sqrt{\text{diag}[\mathbf{H}(\boldsymbol{\theta})^{-1}]_i}$ of the Gaussian marginal. A value of $l \geq 4$ is recommended to enable B-spline interpolation. Then, for $x_i \in \{x_i(\mathbf{v})\}_{\mathbf{v} \in \mathcal{Q}(1,l)}$ and $\boldsymbol{\theta} \in \{\boldsymbol{\theta}(\mathbf{u})\}_{\mathbf{u} \in \mathcal{Q}(m,k)}$ calculate the modes and Hessians

$$\hat{\mathbf{x}}_{-i}(x_i, \boldsymbol{\theta}) = \arg\min_{\mathbf{x}_{-i}} - \log p(\mathbf{y}, x_i, \mathbf{x}_{-i}, \boldsymbol{\theta}),$$

$$\mathbf{H}_{-i,-i}(x_i, \boldsymbol{\theta}) = \frac{\partial^2}{\partial \mathbf{x}_{-i} \partial \mathbf{x}_{-i}^\top} - \log p(\mathbf{y}, x_i, \mathbf{x}_{-i}, \boldsymbol{\theta})|_{\mathbf{x}_{-i} = \hat{\mathbf{x}}_{-i}(x_i, \boldsymbol{\theta})},$$

where optimisation to obtain $\hat{\mathbf{x}}_{-i}(x_i, \boldsymbol{\theta})$ is initialised at $\hat{\mathbf{x}}(\boldsymbol{\theta})_{-i}$.

4. For $x_i \in \{x_i(\mathbf{v})\}_{\mathbf{v} \in \mathcal{Q}(1,l)}$ calculate

$$\tilde{p}_{\texttt{AQ}}(x_i \,|\, \mathbf{y}) = \frac{\tilde{p}_{\texttt{LA}}(x_i, \mathbf{y})}{\tilde{p}_{\texttt{AQ}}(\mathbf{y})}. \tag{4.4}$$

where

$$\tilde{p}_{\texttt{LA}}(x_i, \mathbf{y}) = \sum_{\mathbf{u} \in \mathcal{Q}(m,k)} \tilde{p}_{\texttt{LA}}(x_i, \boldsymbol{\theta}(\mathbf{u}), \mathbf{y}) \omega(\mathbf{u}).$$

and

$$\tilde{p}_{\texttt{LA}}(x_i, \boldsymbol{\theta}, \mathbf{y}) = \frac{p(x_i, \mathbf{x}_{-i}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_{\texttt{G}}(\mathbf{x}_{-i} \,|\, x_i, \boldsymbol{\theta}, \mathbf{y})}\Big|_{\mathbf{x}_{-i} = \hat{\mathbf{x}}_{-i}(x_i, \boldsymbol{\theta})}.$$

Although Equation 4.4 can be calculated using the estimate of the evidence in Equation 4.3 it is more numerically accurate to use the estimate

$$\tilde{p}_{\texttt{AQ}}(\mathbf{y}) = \sum_{\mathbf{v} \in \mathcal{Q}(1,l)} \tilde{p}_{\texttt{LA}}(x_i(\mathbf{v}), \mathbf{y}) \omega(\mathbf{v})$$

5. Given $\{x_i(\mathbf{v}), \tilde{p}_{\texttt{AQ}}(x_i(\mathbf{v}) \,|\, \mathbf{y})\}_{\mathbf{v} \in \mathcal{Q}(1,l)}$ create a spline interpolant to each posterior marginal on the log-scale. Samples, and thereby relevant posterior marginal summaries, may be obtained using inverse transform sampling.

5. Application to data from Malawi. Using one `TMB` C++ user-template (available in the appendix), we fit the simplified Naomi model (Section 2) to data from Malawi using four inferential approaches. These were: 1. `TMB`: EB combined with a Gaussian approximation via `TMB`, 2. `aghq`: AGHQ combined with a Gaussian approximation via `aghq`, 3. `adam`: AGHQ combined with a Laplace approximation by extending `aghq`, and 4. `NUTS`: the Hamiltonian Monte Carlo (HMC) algorithm No-U-Turn Sampling (NUTS) using Stan (Carpenter et al., 2017) via the `tmbstan` package (Monnahan and Kristensen, 2018). In this instance, the dimension of the latent field $N = 491$ and the dimension of the hyperparameters $m = 24$. Settings used for each inferential method are provided in Table 1. For the deterministic methods, following inference we simulated hyperparameter and latent field samples. For all methods, we simulated age-sex-district specific HIV prevalence, ART coverage and HIV incidence from the latent field and hyperparameter posteriors. Example model outputs

Inferential method	Details
1. `TMB`: EB, Gaussian	1000 samples
2. `aghq`: AGHQ, Gaussian	$k = 1$, 1000 samples
3. `adam`: AGHQ, Laplace	$k = 1$, $l = 5$, 1000 samples
4. `NUTS`: NUTS	4 chains of 20000 iterations with the first 10000 iterations of each chain discarded as warmup, then thinned by a factor of 20. HMC parameters set to default for `rstan`.

TABLE 1
A summary of settings used for each inferential method.

from `TMB` are illustrated in Figure 2. The R (R Core Team, 2021) code used to produce all results we describe below is available at `github.com/athowes/elgm-inf`. We used `orderly` (FitzJohn et al., 2022) for reproducible research, `ggplot2` for data visualisation (Wickham, 2016) and `rticles` (Allaire et al., 2022a) for reporting via `rmarkdown` (Allaire et al., 2022b).

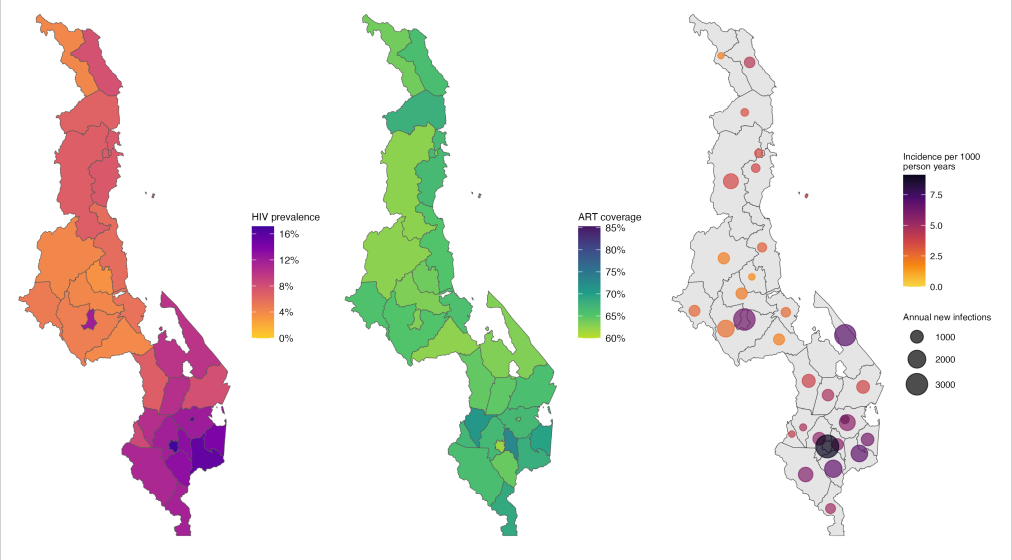

FIGURE 2. *District-level model outputs for adults aged 15-49. Inference conducted with* `TMB`.

5.1. *NUTS convergence.* Inferential results from MCMC are only accurate once convergence has been reached and the chain lengths are sufficient. We assessed the quality of our MCMC results using the potential scale reduction factor $\hat{R}$, bulk and tail effective sample size (ESS), autocorrelation decay plots, univariate traceplots, pairs density plots, and NUTS specific divergent transition and energy assessments. Full details are provided in the appendix. We treat these results from NUTS as a gold-standard to which other inferential methods are compared to.

5.2. *Model assessment*. We performed posterior predictive checks to assess the coverage of our estimates via the uniformity of the data within each posterior marginal distribution.

5.3. *Inference comparison.* We used three methods to assess the accuracy of posterior distributions produced by each inferential method as compared with those from NUTS: (1) Kolmogorov-Smirnov tests, (2) maximum mean discrepancy, and (3) Pareto-smoothed importance sampling. Our primary applied interest is in comparing the accuracy of model outputs, rather than the internal hyperparameters or latent field parameters. That said, obtaining accurate inferences for these internal parameters is also relevant.

5.3.1. *Kolmogorov-Smirnov tests.* Let $\{\theta_i\}_{i=1}^n$ be posterior marginal samples from some quantity with empirical cumulative distribution (ECDF) function $F(\vartheta) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\theta_i \leq \vartheta}$. The two-sample Kolmogorov-Smirnov (KS) test statistic (Smirnov, 1948) is given by the maximum absolute difference between two ECDFs. For each method we compare the KS statistics

$$D_\bullet = \sup_\vartheta |F_{\texttt{NUTS}}(\vartheta) - F_\bullet(\vartheta)|.$$

See a summary of the results in Table and Figure 3, and full results available in the appendix.

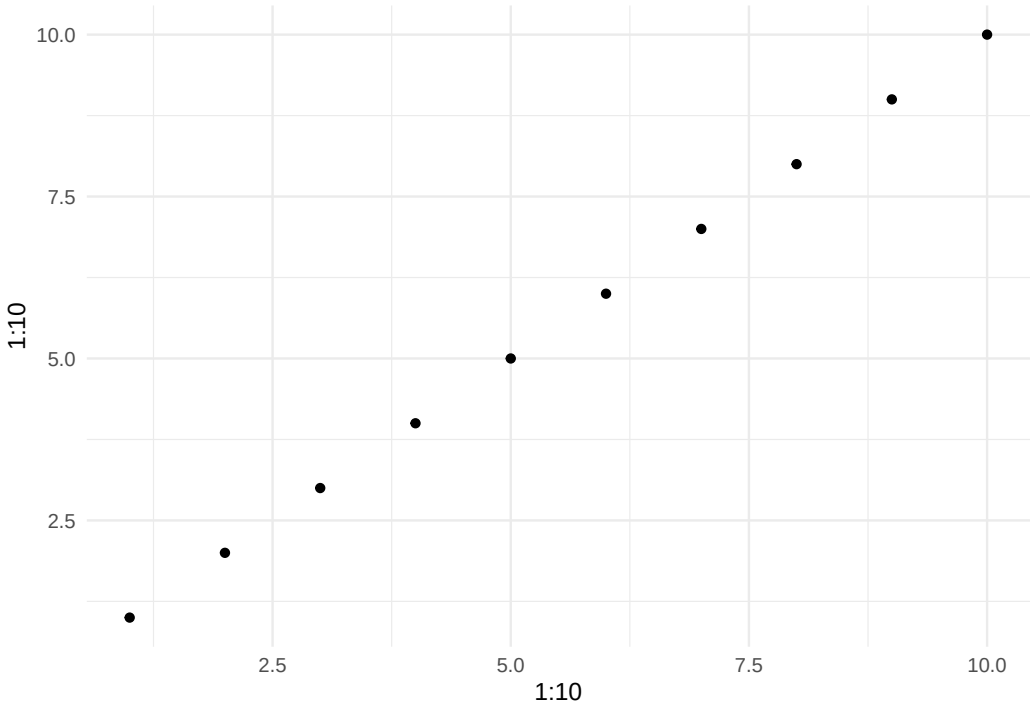

FIGURE 3. *Results of Kolmogorov-Smironv tests analysis.*

5.3.2. *Maximum mean discrepancy.* To write. See a summary of the results in Table and Figure 4, and full results available in the appendix.

5.3.3. *Pareto-smoothed importance sampling.* To write. See a summary of the results in Table and Figure 5, and full results available in the appendix.

6. Discussion. We developed an approximate Bayesian inference algorithm motivated by a challenging problem in small-area estimation of HIV in low resource settings. For the simplified Naomi model in Malawi (Section 5) our method is demonstrated to be more accurate than the EB Gaussian approximation currently in use, and substantially faster than

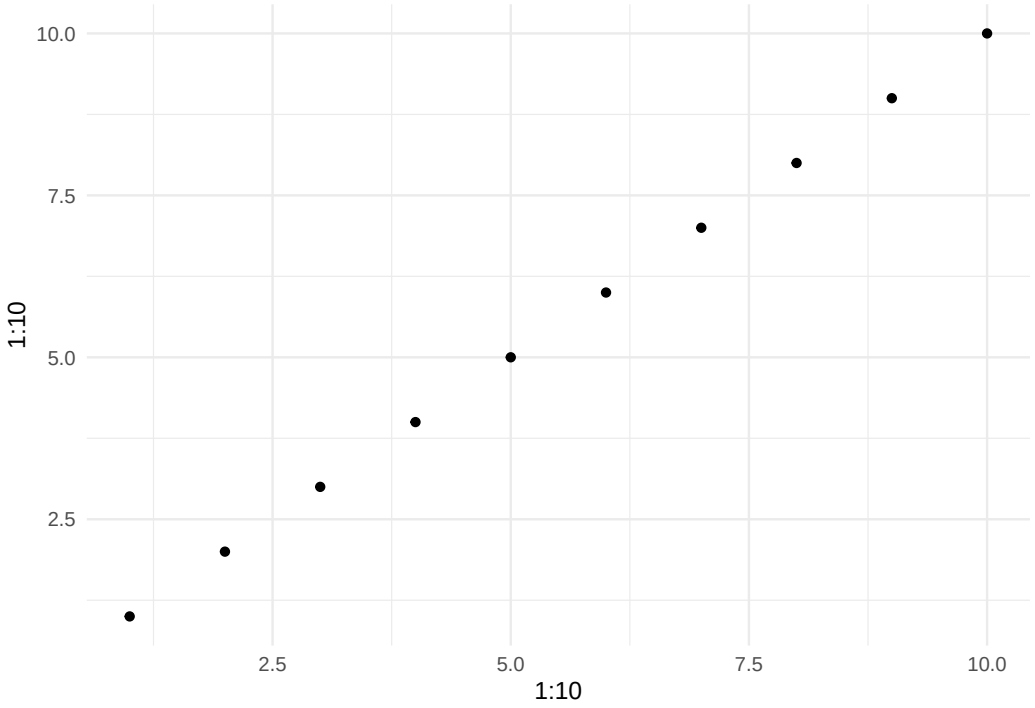

FIGURE 4. *Results of maximum mean discrepancy analysis.*

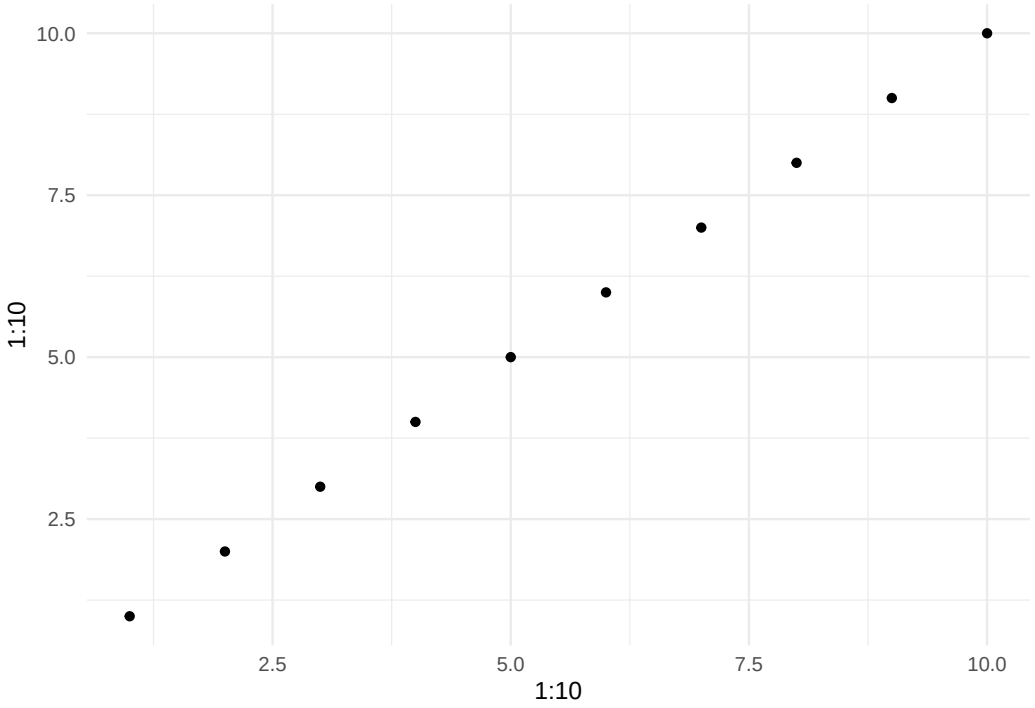

FIGURE 5. *Results of Pareto-smoothed importance sampling analysis.*

NUTS. We anticipate that our method could be added to the Naomi web interface as an alternative to `TMB`. Analysts might quickly iterate over model options using the faster, less accurate inference approach, switching to the slower, more accurate approach once they are happy with the results.

We provide a flexible implementation of the algoritm, building on the`TMB` and `aghq` R packages. In doing so, we hope our work enables use of deterministic inference algorithms for ELGMs in applied settings, as well as further methodological exploration of their accuracy and limitations. Among the ELGMs structures of particular interest in spatial epidemiology, many of which are included in the structure of Naomi, are: aggregated Gaussian process models (Nandi et al., 2020), evidence synthesis models (Amoah, Diggle and Giorgi, 2020). Although our method is designed for ELGMs, it is possible to use it outside this class, as it is compatible with any model with a `TMB` C++ template.

In our case study we demonstrated a Bayesian workflow for deterministic inference methods. We retained the ability to draw samples from the posterior distributions of interest, facilitating use of posterior predictive checks (Section 5.2).

Future work could look to implement our algorithm (Section 4.3) within probabilistic programming languages, facilitating access by a broader user-base. This might be possible in Stan by use of the `bridgestan` package (Ward, 2023) together with the adjoint-differentiated Laplace approximation of Margossian et al. (2020). As well, statistical theory for the algorithm could be established by extension of Theorem 1 in Stringer, Brown and Stafford (2022).

Acknowledgements. AH was supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1), and conducted part of this research while an International Visiting Graduate Student at the University of Waterloo. AH and JWE were supported by the Bill and Melinda Gates Foundation (OPP1190661, OPP1164897). SRF was supported by the EPSRC (EP/V002910/2). JWE was supported by UNAIDS and National Institute of Allergy and Infectious Disease of the National Institutes of Health (R01AI136664). This research was supported by the MRC Centre for Global Infectious Disease Analysis (MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat program and is also part of the EDCTP2 programme supported by the European Union.

REFERENCES

ALLAIRE, J., XIE, Y., DERVIEUX, C., R FOUNDATION, WICKHAM, H., JOURNAL OF STATISTICAL SOFTWARE, VAIDYANATHAN, R., ASSOCIATION FOR COMPUTING MACHINERY, BOETTIGER, C., ELSEVIER, BROMAN, K., MUELLER, K., QUAST, B., PRUIM, R., MARWICK, B., WICKHAM, C., KEYES, O., YU, M., EMAASIT, D., ONKELINX, T., GASPARINI, A., DESAUTELS, M.-A., LEUTNANT, D., MDPI, TAYLOR AND FRANCIS, ÖĞREDEN, O., HANCE, D., NÜST, D., UVESTEN, P., CAMPITELLI, E., MUSCHELLI, J., HAYES, A., KAMVAR, Z. N., ROSS, N., CANNOODT, R., LUGUERN, D., KAPLAN, D. M., KREUTZER, S., WANG, S., HESSELBERTH, J. and HYNDMAN, R. (2022a). rticles: Article Formats for R Markdown R package version 0.23.6.

ALLAIRE, J., XIE, Y., MCPHERSON, J., LURASCHI, J., USHEY, K., ATKINS, A., WICKHAM, H., CHENG, J., CHANG, W. and IANNONE, R. (2022b). rmarkdown: Dynamic Documents for R R package version 2.14.

AMOAH, B., DIGGLE, P. J. and GIORGI, E. (2020). A geostatistical framework for combining spatially referenced disease prevalence data from multiple diagnostics. *Biometrics* **76** 158–170.

BACHL, F. E., LINDGREN, F., BORCHERS, D. L. and ILLIAN, J. B. (2019). inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution* **10** 760–766.

BAYDIN, A. G., PEARLMUTTER, B. A., RADUL, A. A. and SISKIND, J. M. (2017). Automatic differentiation in machine learning: a survey. *The Journal of Machine Learning Research* **18** 5595–5637.

BERILD, M. O., MARTINO, S., GÓMEZ-RUBIO, V. and RUE, H. (2022). Importance sampling with the integrated nested Laplace approximation. *Journal of Computational and Graphical Statistics* **31** 1225–1237.

BILODEAU, B., STRINGER, A. and TANG, Y. (2022). Stochastic convergence rates and applications of adaptive quadrature in Bayesian inference. *Journal of the American Statistical Association* 1–11.

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. and RIDDELL, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* **76**.

DAVIS, P. J. and RABINOWITZ, P. (1975). *Methods of numerical integration*. Academic Press.

EATON, J. W., DWYER-LINDGREN, L., GUTREUTER, S., O'DRISCOLL, M., STEVENS, O., BAJAJ, S., ASHTON, R., HILL, A., RUSSELL, E., ESRA, R., DOLAN, N., ANIFOWOSHE, Y. O., WOODBRIDGE, M., FELLOWS, I., GLAUBIUS, R., HAEUSER, E., OKONEK, T., STOVER, J., THOMAS, M. L., WAKEFIELD, J., WOLOCK, T. M., BERRY, J., SABALA, T., HEARD, N., DELGADO, S., JAHN, A., KALUA, T., CHIMPANDULE, T., AULD, A., KIM, E., PAYNE, D., JOHNSON, L. F., FITZJOHN, R. G., WANYEKI, I., MAHY, M. I. and SHIRAISHI, R. W. (2021). Naomi: a new modelling tool for estimating HIV epidemic indicators at the district level in sub-Saharan Africa. *Journal of the International AIDS Society* **24** e25788.

FITZJOHN, R., ASHTON, R., HILL, A., EDEN, M., HINSLEY, W., RUSSELL, E. and THOMPSON, J. (2022). orderly: Lightweight Reproducible Reporting https://www.vaccineimpact.org/orderly/, https://github.com/vimc/orderly.

FOURNIER, D. A., SKAUG, H. J., ANCHETA, J., IANELLI, J., MAGNUSSON, A., MAUNDER, M. N., NIELSEN, A. and SIBERT, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software* **27** 233–249.

GÓMEZ-RUBIO, V. and RUE, H. (2018). Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing* **28** 1033–1051.

KISH, L. (1965). Survey sampling.

KRISTENSEN, K., NIELSEN, A., BERG, C. W., SKAUG, H., BELL, B. M. et al. (2016). TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software* **70**.

MARGOSSIAN, C., VEHTARI, A., SIMPSON, D. and AGRAWAL, R. (2020). Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for latent Gaussian models and beyond. *Advances in Neural Information Processing Systems* **33** 9086–9097.

MARTINS, T. G., SIMPSON, D., LINDGREN, F. and RUE, H. (2013). Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis* **67** 68–83.

MONNAHAN, C. C. and KRISTENSEN, K. (2018). No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages. *PloS one* **13** e0197954.

NANDI, A. K., LUCAS, T. C., ARAMBEPOLA, R., GETHING, P. and WEISS, D. J. (2020). Disaggregation: an R package for Bayesian spatial disaggregation modelling. *arXiv preprint arXiv:2001.04847*.

NAYLOR, J. C. and SMITH, A. F. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society Series C: Applied Statistics* **31** 214–225.

NEAL, R. M. (2003). Slice sampling. *The Annals of Statistics* **31** 705–767.

OSGOOD-ZIMMERMAN, A. and WAKEFIELD, J. (2022). A Statistical Review of Template Model Builder: A Flexible Tool for Spatial Modelling. *International Statistical Review*.

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 319–392.

SMIRNOV, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics* **19** 279–281.

STRINGER, A. (2021). Implementing Approximate Bayesian Inference using Adaptive Quadrature: the aghq Package. *arXiv preprint arXiv:2101.04468*.

STRINGER, A., BROWN, P. and STAFFORD, J. (2022). Fast, scalable approximations to posterior distributions in extended latent Gaussian models. *Journal of Computational and Graphical Statistics* 1–15.

R CORE TEAM (2021). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

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

WARD, B. (2023). bridgestan: BridgeStan, Accessing Stan Model Functions in R R package version 1.0.1.

WICKHAM, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

WOOD, S. N. (2020). Simplified integrated nested Laplace approximation. *Biometrika* **107** 223–230.