Submitted to the Annals of Applied Statistics

FAST APPROXIMATE BAYESIAN INFERENCE FOR THE NAOMI HIV SURVEILLANCE 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 which combines the simplified integrated nested Laplace approximation approach of Wood (2020) with adaptive Gauss-Hermite quadrature 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.

Mounting an effective public health response to the HIV epidemic requires accurate, timely HIV indicator estimates at the level of geographic resolution at which heath systems are planned and delivered. Producing these estimates is challenging because all available data sources have shortcomings which must be overcome. Nationally-representative household surveys provide the most statistically reliable data, but due to their high cost, in most countries they occur 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 meet these 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 mitigates the limitations of any single source, increases statistical power, and prompts investigation into possibly conflicting 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. The 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 estimates.

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

The practical requirements for the model, combined 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 by country teams, ruling out prohibitively slow Markov chain Monte Carlo (MCMC) approaches. 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, 2021). 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, which for the Naomi model, we use for the high-dimensional latent field parameters. 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 has favourable computational properties, it would be preferable to account more fully for hyperparameter uncertainty than is possible in an empirical Bayes framework. This has motivated us to look for an approach closer to full Bayesian inference, which is flexible enough to be compatible with the model, and 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 the simplified integrated nested Laplace approximation (INLA) approach of Wood (2020) with adaptive Gauss-Hermite quadrature (AGHQ). 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 the latent field required for this approximation to be accurate. This extension facilitates inference for models like Naomi, which are not latent Gaussian models (LGMs) and so were not previously amenable to inference with INLA. Instead, due to dependence of observations on multiple structured additive predictors, Naomi is what has been termed by Stringer, Brown and Stafford (2022) an extended latent Gaussian model (ELGM). 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 (2021) 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 of Bachl et al. (2019), and the INLA within MCMC approach of Gómez-Rubio and Rue (2018), both of which leverage the R-INLA R package (Martins et al., 2013). To approximate non-linear predictors, inlabru makes iterated use of R-INLA using linearisation, extending the scope of INLA to LGMs with some suitably small amount of non-linearity. INLA within MCMC is suitable for models which are LGMs, though only conditional on some subset of the parameters being fixed. For these models, Metropolis-Hastings style algorithms be defined to update the fixed parameters, with acceptance probabilities calculated using by repeat calls to R-INLA. It would or would not be possible to fit the Naomi model in these frameworks.

The remainder of this paper is organised as follows. Section 2 describes a simplified version of the Naomi model that we consider in this paper. Section 3 describes how the Naomi model 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. The model is 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 of T_2 and temporal projection to T_3. Below we provide an overview of the simplified model, highlighting the aspects which make it a challenge for existing inferential approaches. A more complete mathematical description of the simplified model, as well as a C++ template for the log-posterior, are provided in the appendix.

2.1. *Household survey component* .

Consider a country in sub-Saharan Africa where a household survey with complex survey design has taken place at time T_1. Let $x \in \mathcal{X}$ index district, $a \in \mathcal{A}$ five-year age band, and $s \in \mathcal{S}$ 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 on such that

$$\text{logit}(\rho_i) = \eta^\rho_i,$$
$$\text{logit}(\alpha_i) = \eta^\alpha_i,$$

for certain choice of linear predictors η^ρ_i and η^α_i. For the HIV incidence rate we model on the log scale $\log(\lambda_i) = \eta^\lambda_i(\{\rho_i, \alpha_i\}_{i \in \mathcal{I}})$, where the linear predictor depends on $\{\rho_i, \alpha_i\}_{i \in \mathcal{I}}$ for some $\mathcal{I}$. 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 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

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

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 $y^{\hat \theta}_{\mathcal{I}} = m^{\hat \theta}_{\mathcal{I}} \cdot \hat \theta_{\mathcal{I}}$ where

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

is the Kish effective sample size (Kish, 1965).

For $\theta \in \{\rho, \alpha, \kappa\}$ we model these aggregate observations using a binomial working likelihood $y^{\hat{\theta}}_{\mathcal{I}} \sim \text{xBin}(m^{\hat{\theta}}_{\mathcal{I}}, \theta_{\mathcal{I}})$, where $\theta_{\mathcal{I}}$ are the following appropriately weighted aggregates

$$\rho_{\mathcal{I}} = \frac{\sum_{i \in \mathcal{I}} N_i \rho_i}{\sum_{i \in \mathcal{I}} N_i},$$

$$\alpha_{\mathcal{I}} = \frac{\sum_{i \in \mathcal{I}} N_i \rho_i \alpha_i}{\sum_{i \in \mathcal{I}} N_i \rho_i},$$

$$\kappa_{\mathcal{I}} = \frac{\sum_{i \in \mathcal{I}} N_i \rho_i \kappa_i}{\sum_{i \in \mathcal{I}} N_i \rho_i}.$$

2.2. *ANC testing component* . We model HIV prevalence ρ^{ANC}_i and ART coverage α^{ANC}_i among pregnant women as being offset on the logit scale from the general population indicator as follows

$$\text{logit}(\rho^{\text{ANC}}_i) = \text{logit}(\rho_i) + \eta^{\rho^{\text{ANC}}}_i,$$

$$\text{logit}(\alpha^{\text{ANC}}_i) = \text{logit}(\alpha_i) + \eta^{\rho^{\text{ANC}}}_i.$$

We inform these processes by specifying likelihoods for the following aggregate ANC data from the year of the most recent survey: the number of ANC clients with ascertained status $x^{\text{ANC}}_{\mathcal{I}}$, the number of those with positive status $y^{\text{ANC}}_{\mathcal{I}}$, and the number of ANC clients already on ART prior to their first ANC visit $z^{\text{ANC}}_{\mathcal{I}}$. We use the binomial working likelihoods

$$y^{\text{ANC}}_{\mathcal{I}} \sim \text{Bin}(y^{\text{ANC}}_{\mathcal{I}}, \alpha^{\text{ANC}}_{\mathcal{I}}),$$

$$z^{\text{ANC}}_{\mathcal{I}} \sim \text{Bin}(y^{\text{ANC}}_{\mathcal{I}}, \alpha^{\text{ANC}}_{\mathcal{I}}),$$

where, again, we use weighted aggregates

$$\rho^{\text{ANC}}_{\mathcal{I}} = \frac{\sum_{i \in \mathcal{I}} \Psi_i \rho^{\text{ANC}}_i}{\sum_{i \in \mathcal{I}} \Psi_i},$$

$$\alpha^{\text{ANC}}_{\mathcal{I}} = \frac{\sum_{i \in \mathcal{I}} \Psi_i \rho^{\text{ANC}} \alpha^{\text{ANC}}_i}{\sum_{i \in \mathcal{I}} \Psi_i \rho^{\text{ANC}}_i},$$

with Ψ_i the number of pregnant women.

2.3. *ART attendance component* . People living with HIV may choose 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 $x \sim x'$, and model the log-odds $\tilde{\gamma}_{x,x'} = \text{logit}(\gamma_{x,x'})$ using a multinomial logistic regression model.

Let $\dot{A}_{\mathcal{I}} = \sum_{x \sim x', x = x'} \dot{A}_{x',x}$ be the number of people receiving ART. Model $\dot{A}_{\mathcal{I}} \sim \mathcal{N}(\tilde{A}_{\mathcal{I}}, \sigma^{\tilde{A}_{\mathcal{I}}})$ using a normal distribution with mean $\sum_{x \sim x', x = x'} N_{x'} \pi_{x',x}$ and standard deviation $\sigma^{\tilde{A}_{\mathcal{I}}}) = \sqrt{\sum_{x \sim x', x = x'} N_{x'} \pi_{x',x} (1 - \pi_{x',x})}$.

3. Extended Latent Gaussian models.

Latent Gaussian models (LGMs) (Rue, Martino and Chopin, 2009) are of the form

$$y_i \sim p(y_i \,|\, \eta_i, \boldsymbol{\theta}_1), \quad i \in [n]$$

$$\begin{aligned}\mu_i &= \mathbb{E}(y_i \,|\, \eta_i) = g(\eta_i), \\ \eta_i &= \beta_0 + \sum_{l=1}^{p} \beta_j z_{ji} + \sum_{k=1}^{r} f_k(u_{ki}),\end{aligned}$$

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 allows for the presence of more non-linearity. 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. Fast approximate inference methods. 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

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

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

4.1. *Algorithm.* Given the negative unnormalised log posterior $-\log p(\mathbf{y},\mathbf{x},\boldsymbol{\theta})$, we obtain the posterior marginal approximations $\{\tilde{p}(x_i \,|\, \mathbf{y})\}_{i=1}^{n}$ and $\tilde{p}(\theta_j \,|\, \mathbf{y})_{j=1}^{m}$ via the following algorithm, comprised of nested applications of Laplace approximation and adaptive Gauss-Hermite quadrature.

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

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

   of the Laplace approximation

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

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

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

2. Generate a set of nodes $\mathbf{z} \in \mathcal{Q}(m,k)$ and weights $\omega : \mathbf{z} \in \mathcal{Q}(m,k) \to \mathbb{R}$ from a Gauss-Hermite quadrature rule with k nodes per dimension, which are then adapted based on the mode and lower Choleksy via $\boldsymbol{\theta}(\mathbf{z}) = \hat{\boldsymbol{\theta}} + \mathbf{L}\mathbf{z}$. 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 $\tilde{p}_{\text{AQ}}(\mathbf{y})$ as follows

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

3. For $i \in [n]$ generate l nodes $x_i(\mathbf{z})$ via a Gauss-Hermite quadrature rule $\mathbf{z} \in \mathcal{Q}(1,l)$ 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 $x_i \in \{x_i(\mathbf{z})\}_{\mathbf{z}\in\mathcal{Q}(1,l)}$ and $\boldsymbol{\theta} \in \{\boldsymbol{\theta}(\mathbf{z})\}_{\mathbf{z}\in\mathcal{Q}(m,k)}$ calculate the modes and Hessians

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

   where optimisation to obtain $\hat{\mathbf{x}}_{-i}(x_i,\boldsymbol{\theta})$ is initialised at $\hat{\mathbf{x}}(\boldsymbol{\theta})_{-i}$.
4. For $x_i \in \{x_i(\mathbf{z})\}_{\mathbf{z}\in\mathcal{Q}(1,l)}$ calculate

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

   where

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

Inferential method	Details
1. EB, Gaussian	1000 samples
2. AGHQ, Gaussian	$k = 1$, 1000 samples
3. AGHQ, Laplace	$k = 1$, $l = 5$, 1000 samples
4. 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.

and

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

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

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

5. Given $\{x_i, \tilde{p}_{\text{AQ}}(x_i \,|\, \mathbf{y})\}_{x_i \in \mathcal{X}_i}$ create a spline interpolant to each posterior marginal on the log-scale, from which samples or relevant posterior marginal summaries may be obtained.

Note that clearer notation is required for $\mathbf{z}$ and $\omega(\mathbf{z})$.

5. Application to data from Malawi. Using one `TMB` template (available in the appendix), we fit the simplified Naomi model to data from Malawi using four inferential approaches: (1) EB combined with a Gaussian approximation using `TMB`, (2) AGHQ combined with a Gaussian approximation using `aghq`, (3) AGHQ combined with a Laplace approximation by extending `aghq` and (4) the Hamiltonian Monte Carlo (HMC) algorithm No-U-Turn Sampling (NUTS) using `tmbstan`. Details regarding particular settings used for each inferential method are provided in Table 1.

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).

5.1. *NUTS convergence.* MCMC can be used to obtain accurate inferential results only once convergence has been reached and the Markov chain length is sufficiently long. 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 can be compared to.

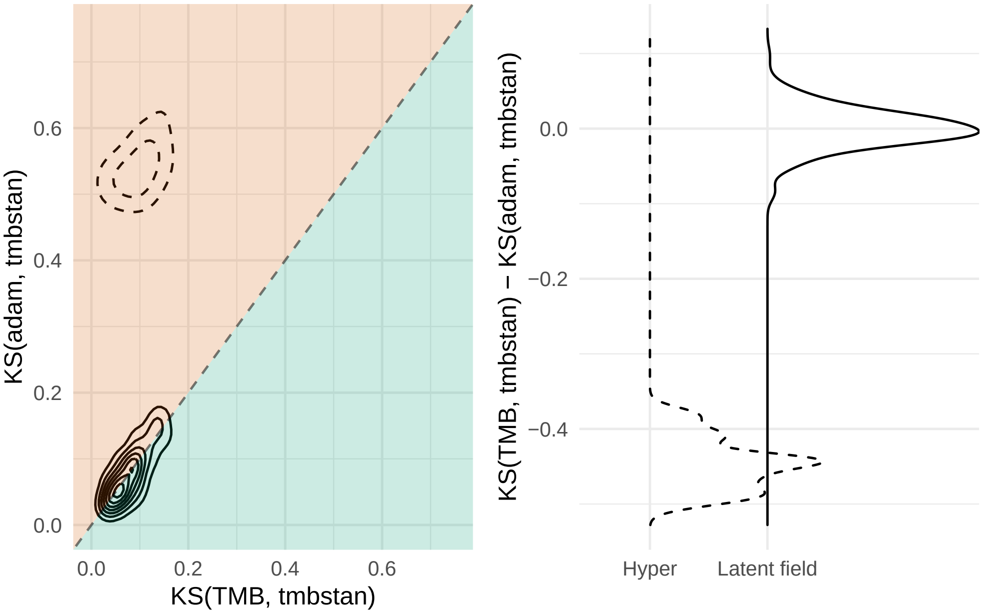
5.2. *Model assessment .* We performed posterior predictive checks to assess the coverage of our estimates via the uniformity of the data within each posterior marginal distribution.

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.

5.3.1. *Kolmogorov-Smirnov tests.* Let $\{\theta_i\}_{i=1}^n$ be posterior marginal samples 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 is given by the maximum absolute difference between two ECDFs. For each method $\{\texttt{TMB}, \texttt{aghq}, \texttt{adam}\}$ we compare the KS statistics

$$D_{\texttt{method}} = \sup_{\vartheta} |F_{\texttt{tmbstan}}(\vartheta) - F_{\texttt{method}}(\vartheta)|.$$

See a summary of the results in Table 1 and Plot 1, and full results available in the appendix.



Type	TMB	aghq	adam
Hyper	0.12204167	0.56843750	0.5684375
Latent field	0.08378051	0.08461456	0.0837955

5.3.2. *Maximum mean discrepancy.* To write. See a summary of the results in Table 2 and Plot 2, and full results available in the appendix.

5.3.3. *Pareto-smoothed importance sampling.* To write. See a summary of the results in Table 3 and Plot 3, and full results available in the appendix.

6. Discussion. We developed an approximate Bayesian inference algorithm to solve a challenging problem in small-area estimation of HIV for low resource settings. Our method is demonstrated to be more accurate than the EB Gaussian approximation and substantially faster than NUTS for the simplified Naomi model in Malawi (Section 5). We anticipate that our method could be added to the Naomi software as an alternative option to the, still substantially faster, EB Gaussian approximation. Analysts might quickly iterate over model options using the faster, less accurate inference approach, only switching to the slower, more accurate approach once they are happy with the results.

We provide a flexible implementation, which builds on the aghq R package. In doing so, we hope our work enables use of INLA for ELGMs in applied settings, as well as further methodological exploration of the algorithms accuracy and limits. Among the ELGMs structures of greatest interest are: aggregated Gaussian process models (Nandi et al., 2020). Although our method is designed for ELGMs, it may even be used outside this class, and is compatible with any model with a TMB C++ template.

We would be excited to see statistical theory for our algorithm by extension of Theorem 1 of Stringer, Brown and Stafford (2022).

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).

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 (GRANT). AH and JWE were supported by the Bill and Melinda Gates Foundation (OPP1190661, OPP1164897). AS was supported by… 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.

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.

Bilodeau, B., Stringer, A. and Tang, Y. (2021). Stochastic convergence rates and applications of adaptive quadrature in Bayesian inference. *arXiv preprint arXiv:2102.06801*.

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**.

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

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*.

OSGOOD-ZIMMERMAN, A. and WAKEFIELD, J. (2021). A Statistical Introduction to Template Model Builder: A Flexible Tool for Spatial Modeling.

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.

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.

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.