Submitted to the Annals of Applied Statistics

INTEGRATED NESTED LAPLACE APPROXIMATIONS FOR EXTENDED LATENT GAUSSIAN MODELS WITH APPLICATION TO THE NAOMI HIV MODEL

BY ADAM HOWES [1] , 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 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. 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 substantially more accurate inferences than the empirical Bayes Gaussian approximation approach used currently, and is comparable to Hamiltonian Monte Carlo with the No-U-Turn sampler. By extending the `aghq` R package we facilitate flexible and easy use of our method when provided a `TMB` C++ template for the model's log-posterior.

1. Introduction. Mounting an effective public health response to the HIV epidemic requires accurate, timely HIV indicator estimates at a sufficiently fine-scale resolution. Producing these estimates is a challenging task, as all available data sources have shortcomings. Nationally-representative household surveys are the most statistically reliable data source, but due to their high cost to run, in most countries they only occur infrequently. Other data sources, such as routine health surveillance of antenatal care clinics, are more real-time but based on a biased sample 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 prevalence, HIV incidence, and coverage of antiretroviral treatment (ART) at a district-level. Software (`https://naomi.unaids.org`) has been developed for Naomi, which allows countries to input their data and generate estimates in a yearly process supported by UNAIDS.

The complexity of the model presents a difficult Bayesian inference problem. Any inferential strategy must be fast, 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 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` is gaining popularity 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. For the Naomi model, we use this option to integrate out the latent field

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

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, we have found the inferences generated for the Naomi model to sometimes be inaccurate.

To obtain fast, accurate inferences for the Naomi model we develop a new inference methodology which combines the simplified integrated nested Laplace approximation (INLA) approach of Wood (2020) with adaptive Gauss-Hermite quadrature (AGHQ). INLA is an approach to approximate Bayesian inference based on nested Laplace approximations and numerical quadrature. The central innovation of Rue, Martino and Chopin (2009) is a way to approximate 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 fall within the extended latent Gaussian model (ELGM) (Stringer, Brown and Stafford, 2022) class, and were not previously amenable to inference with INLA. ELGMs build on latent Gaussian models (LGMs) by allowing each element of the linear predictor to depend on any subset of elements from the latent field. 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). As `aghq` is designed to naturally interface with `TMB`, use of the method is easy when provided a C++ user template for the log-posterior.

The remainder of this paper is organised as follows. In Section 2 we describe a modified version of the Naomi model that we consider. Section 3 outlines our approach to fast, accurate Bayesian inference using simplified INLA and AGHQ. As a case-study, we fit the modified Naomi model on data from Malawi, and compare the accuracy of inferences in Section 2. In this section, we also demonstrating a Bayesian workflow. Finally, in Section 5 we discuss our conclusions, how our method might be used in other models, and directions for future research.

2. A modified Naomi model. Eaton et al. (2021) specify a joint model linking small-area estimation models of HIV prevalence from household surveys, HIV prevalence from antenatal care clinics, and antiretroviral therapy (ART) coverage from routine health data collection. Modelling data from multiple sources concurrently is beneficial as it mitigates the limitations of any single source, increases statistical power and prompts investigation into data conflicts.

Let x index district, a index five-year age band, s index sex and t index time. For each country, the full 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, omiting all temporal projection. In this section, we provide an overview of the modified model, highlighting the aspects which make it a challenge for existing inferential approaches. Full mathematical details of this modified model, as well as a C++ template for the log-posterior, are provided in the appendix.

2.1. *Household survey component.* Consider a household survey occurring at T_1, and let i index the coarsest district-age-sex division included in the model. The data we observe may be aggregated over indices i, so we let $\mathcal{I}$ be a set of i for which observations are reported.

Let N_i be the population size, and consider the following three indicators: HIV prevalence $\rho_i \in [0,1]$, ART coverage $\alpha_i \in [0,1]$, and annual HIV incidence rate $\lambda_i > 0$. We specify independent mixed effects models for HIV prevalence and ART coverage in the general population on the logit scale such that

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

for certain choice of linear predictors η_i^{ρ} and η_i^{α}. For the HIV incidence rate we use a mixed effects model on the log scale

$$\log(\lambda_i) = \eta_i^{\lambda}(\{\rho_i, \alpha_i\}_{i \in \mathcal{I}}),$$

where the linear predictor depends on $\{\rho_i, \alpha_i\}_{i \in \mathcal{I}}$.

Let κ_i be the proportion recently infected among HIV positive persons. For each set of observed strata indices $\mathcal{I}$, we calculate the weighted observations $\hat{\theta}_{\mathcal{I}}$ for $\theta \in \{\rho, \alpha, \kappa\}$ with respective Kish effective sample sizes

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

where j index individuals in all strata $i \in \mathcal{I}$, with corresponding survey weights w_j. The observed number of indicator cases is then

$$Y_{\mathcal{I}}^{\hat{\theta}} = M_{\mathcal{I}}^{\hat{\theta}} \cdot \hat{\theta}_{\mathcal{I}}.$$

For $\theta \in \{\rho, \alpha, \kappa\}$ we model these aggregate observations using a binomial working likelihood

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

where $\theta_{\mathcal{I}}$ are the following weighted aggregates

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

We link the proportion recently infected among HIV positive persons κ_i to HIV incidence λ_i by

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

where Ω_T is the mean duration of recent infection and β_T is the false recent ratio.

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

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

Process section to be written.

Likelihood section to be written.

2.3. *ART attendance component.* Let $\gamma_{x,x'} \in [0,1]$ be the probability that a person on ART residing in district x recieves ART in district x'. Process section to be written.

Let $\dot{A}_i$ be the number of people receiving ART. Likelihood section to be written.

3. Fast approximate inference method. Consider a latent Gaussian model (LGM) of the form

$$\begin{aligned} &\text{(Observations)} && \mathbf{y} \sim p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}), \\ &\text{(Latent field)} && \mathbf{x} \sim \mathcal{N}(\mathbf{x} \,|\, \mathbf{0}, \mathbf{Q}(\boldsymbol{\theta})^{-1}), \\ &\text{(Parameters)} && \boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \end{aligned}$$

where $\dim(\mathbf{y}) = \dim(\mathbf{x}) = n$ and $\dim(\boldsymbol{\theta}) = m$, and $m < n$. The joint posterior of $(\mathbf{x}, \boldsymbol{\theta})$ is given by

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

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

$$\tilde{p}(x_i \,|\, \mathbf{y}) \approx p(x_i \,|\, \mathbf{y}) = \int p(x_i, \boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta} = \int p(x_i \,|\, \boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta}, \quad i = 1, \ldots, n, \tag{3.1}$$

$$\tilde{p}(\theta_j \,|\, \mathbf{y}) \approx p(\theta_j \,|\, \mathbf{y}) = \int p(\boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta}_{-j} \quad j = 1, \ldots, m. \tag{3.2}$$

3.1. *Forwards algorithm.* Given a C++ user template `model.cpp` for 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. Use a Laplace approximation to obtain the unnormalised $\tilde{p}_{\mathrm{LA}}(\boldsymbol{\theta}, \mathbf{y})$

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

   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

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

2. Normalise $\tilde{p}_{\mathrm{LA}}(\boldsymbol{\theta}, \mathbf{y})$ using adaptive Gauss-Hermite quadrature to obtain

$$\tilde{p}_{\mathrm{AQ}}(\boldsymbol{\theta} \,|\, \mathbf{y}) = \frac{\tilde{p}_{\mathrm{LA}}(\boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_{\mathrm{AQ}}(\mathbf{y})},$$

   where the normalising constant is calculated using nodes from a Gauss-Hermite quadrature rule $\mathbf{z} \in \mathcal{Q}(m, k)$ with $m = \dim(\boldsymbol{\theta})$, k nodes per dimension, and weights $\omega : \mathbf{z} \in \mathcal{Q}(m, k) \to \mathbb{R}$ as

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

The nodes $\mathbf{z}$ are adapted based on the mode and curvature at the mode of the Laplace approximation as follows

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

We typically set $k = 3$ such that there are 3^m nodes in total.

3. Obtain an unnormalised nested approximation to the posterior marginal of the ith latent effect by

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

The nodes and weights $\{\mathcal{Q}(m, k), \omega\}$ used to obtain $\tilde{p}_{\text{AQ}}(\mathbf{y})$ are reused to perform integration with respect to the hyperparameters above. For each of the k^m values of $\boldsymbol{\theta}(\mathbf{z})$ we obtain $\tilde{p}_{\text{LA}}(x_i, \boldsymbol{\theta}(\mathbf{z}), \mathbf{y})$ by setting $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{z})$ in the following Laplace approximation

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

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

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

Optimisation to obtain $\hat{\mathbf{x}}_{-i}(x_i, \boldsymbol{\theta})$ may be initialised at $\hat{\mathbf{x}}(\boldsymbol{\theta})_{-i}$.

4. Normalise $\tilde{p}_{\text{LA}}(x_i, \mathbf{y})$ using $\tilde{p}_{\text{AQ}}(\mathbf{y})$ to obtain

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

which may be evaluated for some choice of values $x_i \in \{\ldots\}$.

3.2. *Backward algorithm.*

1. Calculate

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

where

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

where $\tilde{p}_{\text{G}}(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}(\mathbf{x} \mid \hat{\mathbf{x}}(\boldsymbol{\theta}), \mathbf{H}(\boldsymbol{\theta})^{-1})$ is a Gaussian approximation to $p(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})$ 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 $\mathbf{z} \in \mathcal{Q}(m, k)$ and weights $\omega : \mathbf{z} \in \mathcal{Q}(m, k) \to \mathbb{R}$ from a Gauss-Hermite quadrature rule with $m = \dim(\boldsymbol{\theta})$, k nodes per dimension, as follows

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

We typically set $k = 3$ such that there are 3^m nodes in total.

3. Use this quadrature rule to calculate $\tilde{p}_{\text{AQ}}(\mathbf{y})$ as follows

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

4. For $x_i \in \{\ldots\}$ and $\boldsymbol{\theta} \in \{\boldsymbol{\theta}(\mathbf{z})\}_{\mathbf{z} \in \mathcal{Q}(m,k)}$ 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 $\hat{\mathbf{x}}_{-i}(x_i, \boldsymbol{\theta})$ may be initialised at $\hat{\mathbf{x}}(\boldsymbol{\theta})_{-i}$.

5. For each x_i calculate

$$\tilde{p}_{\text{AQ}}(x_i \mid \mathbf{y}) = \frac{\tilde{p}_{\text{LA}}(x_i, \mathbf{y})}{p_{\text{AQ}}(\mathbf{y})}.$$

where

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

and

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

4. Application to the Naomi model.

- The R (R Core Team, 2021) code used to produce all results we describe is available at `github.com/athowes/elgm-inf`. The inference method is available in versions 0.5.0. onwards of the `aghq` package
- Using the `TMB` template, we fit the model using four inferential approaches: (1) empirical Bayes combined with a Gaussian approximation, (2) AGHQ combined with a Gaussian approximation, (3) AGHQ combined with a Laplace approximation and (4) the Hamiltonian Monte Carlo (HMC) algorithm No-U-Turn Sampling (NUTS). We treat results from NUTS as the gold-standard

- We used the Kolmogorov-Smirnov test for the maximum difference between two empirical cumulative distribution functions to compare posterior marginal distributions
- We performed posterior predictive checks to assess the coverage of our estimates via the uniformity of the data within each posterior marginal distribution

5. Discussion.

- We developed an approximate Bayesian inference algorithm to solve a challenging problem in the small-area estimation of HIV in low resource settings
- The flexibility of our method implementation, including compatibility with any `TMB` C++ template, allows broader use, as well as investigation of, deterministic inference methods than had previously been possible
- We demonstrated a Bayesian workflow for deterministic inference methods

Acknowledgements. AH was supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1).

REFERENCES

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.

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.

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**.

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.

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