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 [5,4]

[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*

[5]*Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health, Harvard University,*

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. The model is provided as a tool for countries to input their data to and generate estimates using an empirical Bayes Gaussian approximation via the `TMB` R package. We propose a new inference method extending adaptive Gauss-Hermite quadrature to deal with >20 hyperparameters, thereby enabling 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 substantially faster to run to Hamiltonian Monte Carlo with the No-U-Turn sampler. 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.

1. Introduction. To mount an effective public health response to the HIV epidemic, it is 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, in large part due to limitations of the 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 closer to real-time but based on limited or non-representative samples of the population. To address 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 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.

The complexity of the model, in combination with practical requirements for its operation, present a difficult Bayesian inference problem. First, 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). Further, as well as hundreds of latent field parameters, Naomi has >20 hyperparameters: substantially more than the small number typically required for use of integrated nested Laplace approximations (Rue, Martino and Chopin, 2009). Second, any inferential strategy must be fast enough for interactive review and iteration of modelling results, as well as easy to run in production across a range of countries. In this setting, Markov chain Monte Carlo (MCMC) approaches are prohibitively slow, both 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, leaving the smaller number of hyperparameters. 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, it has the downside that within the empirical Bayes framework hyperparameter uncertainty is not properly accounted for in the latent field posterior. 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 extends adaptive Gauss-Hermite quadrature to handle a higher number of hyperparameters. 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 our method is simple when provided a C++ user template for the log-posterior.

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, defined over three time points $\{T_i\}_{i=1,2,3}$. We consider a simplified version defined only at the time of the most recent household survey with HIV testing (T_1), omitting nowcasting (T_2) and temporal projection (T_3) which involve limited inferences. An overview of the simplified model is given below, and a more complete mathematical description (Appendix S1) as well as a C++ template for the log-posterior (Appendix S2) 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 group, and $s \in \mathcal{S}$ index sex. For ease of notation, let i index the finest district-age-sex division included in the model. Let $I \subseteq \mathcal{X} \times \mathcal{A} \times \mathcal{S}$ be a set of indices i for which an aggregate observation is reported, with $I \in \mathcal{I}$.

Let $N_i \in \mathbb{N}$ be the known, fixed population size. We infer the following unknown HIV indicators using coupled regression equations: HIV prevalence $\rho_i \in [0,1]$, the proportion of individuals 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 $\text{logit}(\rho_i) = \eta_i^{\rho}$ and $\text{logit}(\alpha_i) = \eta_i^{\alpha}$, for certain choice of structured additive predictors. HIV incidence rate is modelled on the log scale as $\log(\lambda_i) = \eta_i^{\lambda}$, and depends on adult HIV prevalence and adult ART coverage. Let κ_i be the proportion recently infected among HIV positive persons. We link this proportion to HIV incidence via

$$\kappa_i = 1 - \exp\left(-\lambda_i \cdot \frac{1-\rho_i}{\rho_i} \cdot (\Omega_T - \beta_T) - \beta_T\right), \tag{2.1}$$

where the mean duration of recent infection Ω_T and the proportion of long-term HIV infections misclassified as recent β_T are strongly informed by priors for the particular survey.

These processes are informed by household survey data. For $\theta \in \{\rho, \alpha, \kappa\}$ let

$$\hat{\theta}_I = \frac{\sum_j w_j \cdot \theta_j}{\sum_j w_j}$$

be weighted, aggregate survey observations, where j indexes individuals across all strata $i \in I$ and w_j are design weights. The observed number of outcomes are $y_I^{\hat{\theta}} = m_I^{\hat{\theta}} \cdot \hat{\theta}_I$ where

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

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

$$y_I^{\hat{\theta}} \sim \text{xBin}(m_I^{\hat{\theta}}, \theta_I)$$

to model these aggregate observations, where θ_I are the following weighted aggregates

$$\rho_I = \frac{\sum_{i \in I} N_i \rho_i}{\sum_{i \in I} N_i}, \quad \alpha_I = \frac{\sum_{i \in I} N_i \rho_i \alpha_i}{\sum_{i \in I} N_i \rho_i}, \quad \kappa_I = \frac{\sum_{i \in I} N_i \rho_i \kappa_i}{\sum_{i \in I} N_i \rho_i}.$$

2.2. *ANC testing component* . HIV prevalence ρ_i^{ANC} and ART coverage α_i^{ANC} among pregnant women are modelled as offset 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 aggregate ANC data from the year of the most recent survey. In particular, the number of ANC clients with ascertained status x_I^{ANC}, the number of those with positive status y_I^{ANC}, and the number of ANC clients already on ART prior to their first ANC visit z_I^{ANC}. We use the binomial working likelihoods

$$y_I^{\text{ANC}} \sim \text{Bin}(x_I^{\text{ANC}}, \rho_I^{\text{ANC}}),$$
$$z_I^{\text{ANC}} \sim \text{Bin}(y_I^{\text{ANC}}, \alpha_I^{\text{ANC}}),$$

where, again, we use weighted aggregates

$$\rho_I^{\text{ANC}} = \frac{\sum_{i\in I} \Psi_i \rho_i^{\text{ANC}}}{\sum_{i\in I} \Psi_i}, \quad \alpha_I^{\text{ANC}} = \frac{\sum_{i\in I} \Psi_i \rho_i^{\text{ANC}} \alpha_i^{\text{ANC}}}{\sum_{i\in I} \Psi_i \rho_i^{\text{ANC}}},$$

with Ψ_i the number of pregnant women, which we assume to be fixed.

2.3. *ART attendance component* . People living with HIV sometimes choose to access ART services outside of the district that they reside in. To account for this, we use multinomial logistic regression to model the probabilities of accessing services outside the home district. Let $\gamma_{x,x'}$ be the probability that a person on ART residing in district x receives ART in district x', and assume $\gamma_{x,x'} = 0$ unless $x = x'$ or the two districts are neighbouring, denoted by $x \sim x'$. The log-odds $\tilde{\gamma}_{x,x'} = \text{logit}(\gamma_{x,x'})$ are modelled using a structured additive 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. Aggregate ART attendance data $\dot{A}_I$ is modelled using a Gaussian approximation to a sum of binomials. The sum results both by aggregation over $i \in I$ 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.

2.4. *Collected together.* Let $\mathbf{y} = (y_I^{\hat{\theta}})$ for $\theta \in \{\rho, \alpha, \kappa\}$ and $I \in \mathcal{I}$ be the vector of observations. Attempt to write fully the Naomi model in a small number of collected together equations here.

3. Extended Latent Gaussian models. We now describe the popular latent Gaussian class of models, and an extension which encapsulates the complexities of Naomi.

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

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

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 mapping allows for the presence of non-linearity in the model. Put together, ELGMs are then of the form

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

3.2. *Naomi framed as an ELGM.* Naomi has a lot in common with many LGMS: it is a spatio-temporal model with a large latent field, governed by a smaller number of hyperparameters. However, Naomi is not an LGM, and instead falls into the ELGM class, for the following reasons:

1. In the household survey component, HIV incidence depends on district-level adult HIV prevalence and ART coverage, such that $\lambda \propto \rho(1 - \omega \cdot \alpha)$, where $\omega = 0.7$ is a fixed constant. This reflects basic HIV epidemiology: HIV incidence is proportional to unsuppresed viral load. As a result, $\log(\lambda_i)$ depends on 28 structured additive predictors (2 sexes $\times$ 7 age groups $\times$ 2 indicators, HIV prevalence and ART coverage).
2. In the household survey component, HIV incidence and HIV prevalence are linked to the proportion recently infected via Equation 2.1.
3. In the ANC testing component, HIV prevalence and ART coverage depend upon the respective indicators in the household survey component. Although $\text{logit}(\rho_i)$ and $\text{logit}(\alpha_i)$ are Gaussian, nonetheless this introduces dependence of each mean response on two structured additive predictors.
4. Throughout the model components, processes are modelled at the finest distict-age-sex division, but likelihoods are defined for observations aggregated over sets of indices. As such, single observations are related to $|\mathcal{I}|$ structured additive predictors.
5. Individuals taking ART, or who have been recently infected, must be HIV positive.
6. The ART attendance component uses a multinomial model with softmax link function which takes as input $|\{x' : x' \sim x\}| + 1$ structured additive predictors.
7. Multiple link functions are used throughout the model, such that there is no one inverse link function g. Instead,

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

$$\hat{\mathbf{x}}(\boldsymbol{\theta}) = \arg\max_{\mathbf{x}} \log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}),$$

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

Then the Laplace approximation to $p(\boldsymbol{\theta}, \mathbf{y})$ is given by

$$\tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{\tilde{p}_{\text{G}}(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})}\Big|_{\mathbf{x}=\hat{\mathbf{x}}(\boldsymbol{\theta})}. \tag{4.3}$$

Inference in TMB proceeds by optimising Equation 4.3 to obtain $\hat{\boldsymbol{\theta}}_{\text{LA}} = \arg\max_{\boldsymbol{\theta}} \tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y})$. Latent field joint and marginal inferences are then direct from the Gaussian approximation $\tilde{p}_{\text{G}}(\mathbf{x} \mid \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y})$.

4.2. *Gauss-Hermite quadrature.* Quadrature rules can be used to approximate the integral of $\tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y})$ via the weighted sum

$$p(\mathbf{y}) \approx \int_{\boldsymbol{\theta}} \tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y}) \mathrm{d}\boldsymbol{\theta} \approx \sum_{\mathbf{z} \in \mathcal{Q}} p_{\text{LA}}(\mathbf{z}, \mathbf{y}) \omega(\mathbf{z}), \tag{4.4}$$

where $\mathbf{z} \in \mathcal{Q}$ are a set of nodes and $\omega : \mathcal{Q} \to \mathbb{R}$ is a weighting function. Gauss-Hermite quadrature [GHQ; Davis and Rabinowitz (1975)] is a quadrature rule where, in the univariate case, the nodes $\mathcal{Q}(1, k) = \{z : H_k(z) = (-1)^k \exp(z^2/2) \frac{\mathrm{d}}{\mathrm{d}z^k} \exp(-z^2/2) = 0\}$ are selected as zeros of the kth Hermite polynomial. The corresponding weights $\omega : \mathcal{Q}(1, k) \to \mathbb{R}$ are given by

$$\omega(z) = \frac{k!}{[H_{k+1}(z)]^2 \phi(z)},$$

where $\phi(\cdot)$ is a standard univariate Gaussian density. Multivariate GHQ rules are typically obtained using the product rule such that $\mathbf{z} = (z_1, \ldots, z_m) \in \mathcal{Q}(m, k) = \mathcal{Q}(1, k)^m$ and $\omega(\mathbf{z}) = \prod_{j=1}^m \omega(z_j)$. GHQ is attractive because it is exact for functions which are a polynomial of total order no more than $2k - 1$ multiplied by a Gaussian density.

4.3. *Adaptive quadrature.* In adaptive Gauss-Hermite quadrature [AHGQ; Naylor and Smith (1982); Tierney and Kadane (1986)] the nodes are shifted and rotated to suit the particular integrand. Repositioning the nodes is particularly important for statistical quadrature problems, where the integral depends on data $\mathbf{y}$ and so regions of high density are not known in advance. To obtain an AGHQ estimate of Equation 4.4, let $\hat{\mathbf{H}}_{\text{LA}}(\hat{\boldsymbol{\theta}}_{\text{LA}}) = -\partial^2 \log p_{\text{LA}}(\hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y})$ be the curvature at the mode $\hat{\boldsymbol{\theta}}_{\text{LA}}$ and $[\hat{\mathbf{H}}_{\text{LA}}(\hat{\boldsymbol{\theta}}_{\text{LA}})]^{-1} = \hat{\mathbf{P}}_{\text{LA}} \hat{\mathbf{P}}_{\text{LA}}^\top$ a matrix decomposition of the inverse curvature, then

$$\tilde{p}_{\text{AGHQ}}(\mathbf{y}) = |\hat{\mathbf{P}}_{\text{LA}}| \sum_{\mathbf{z} \in \mathcal{Q}(m,k)} \tilde{p}_{\text{LA}}(\hat{\mathbf{P}}_{\text{LA}} \mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y}) \omega(\mathbf{z}). \tag{4.5}$$

Two possibilities for the matrix decomposition are the Cholesky decomposition $\hat{\mathbf{P}}_{\text{LA}} = \hat{\mathbf{L}}_{\text{LA}}$ and the spectral decomposition $\hat{\mathbf{P}}_{\text{LA}} = \hat{\mathbf{E}}_{\text{LA}} \hat{\boldsymbol{\Lambda}}_{\text{LA}}^{1/2}$ (Jäckel, 2005). Equation 4.5 may be used to normalise the Laplace approximation

$$\tilde{p}_{\text{LA}}(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_{\text{AGHQ}}(\mathbf{y})}.$$

For the latent field, the adapted nodes and weights may be reused (Rue, Martino and Chopin, 2009; Stringer, Brown and Stafford, 2022) in the nested approximation

$$\tilde{p}(\mathbf{x} \,|\, \mathbf{y}) = |\hat{\mathbf{P}}_{\text{LA}}| \sum_{\mathbf{z} \in \mathcal{Q}(m,k)} \tilde{p}_{\text{G}}(\mathbf{x} \,|\, \hat{\mathbf{P}}_{\text{LA}}\mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y}) \tilde{p}_{\text{LA}}(\hat{\mathbf{P}}_{\text{LA}}\mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}} \,|\, \mathbf{y}) \omega(\mathbf{z}). \tag{4.6}$$

Samples from this mixture of Gaussians are obtained by drawing a node $\mathbf{z}$ with multinomial probabilities $\lambda(\mathbf{z}) = |\hat{\mathbf{P}}_{\text{LA}}| p_{\text{LA}}(\hat{\mathbf{P}}_{\text{LA}}\mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}} \,|\, \mathbf{y}) \omega(\mathbf{z})$, then drawing from the Gaussian $\tilde{p}_{\text{G}}(\mathbf{x} \,|\, \hat{\mathbf{P}}_{\text{LA}}\mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y})$.

4.4. *Principal components analysis.* Use of the product rule requires $|\mathcal{Q}(m,k)| = k^m$ quadrature points, which becomes intractable as m increases for $k > 1$. An alternative is to let $\mathbf{k} = (k_1, \ldots, k_m)$ be a vector of levels for each dimension of $\boldsymbol{\theta}$, and define $\mathcal{Q}(m, \mathbf{k}) = \mathcal{Q}(1, k_1) \times \cdots \times \mathcal{Q}(1, k_m)$ of size $|\mathcal{Q}(m, \mathbf{k})| = \prod_{j=1}^{m} k_j$. Let $\mathcal{Q}(m, s, k)$ correspond to $\mathcal{Q}(m, \mathbf{k})$ with choice of levels $k_j = k, j \leq s$ and $k_j = 1, j > s$ for some $s \leq m$. This choice of nodes, taken together with use of the spectral decomposition, is analogous to a principal components analysis (PCA) approach to AGHQ

$$\tilde{p}_{\text{PCA}}(\mathbf{y}) = |\hat{\mathbf{E}}_{\text{LA},s} \hat{\boldsymbol{\Lambda}}_{\text{LA},s}^{1/2}| \sum_{\mathbf{z} \in \mathcal{Q}(m,s,k)} \tilde{p}_{\text{LA}}(\hat{\mathbf{E}}_{\text{LA},s} \hat{\boldsymbol{\Lambda}}_{\text{LA},s}^{1/2} \mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y}) \omega(\mathbf{z}), \tag{4.7}$$

where $\hat{\mathbf{E}}_{\text{LA},s} = \cdots$ and $\hat{\boldsymbol{\Lambda}}_{\text{LA},s} = \cdots$.

4.5. *Sparse rules.* Though there are sparse rules which retain the attractive exactness properties of GHQ, with fewer than k^m quadrature points, they use weighting functions which may be negative $\omega(\mathbf{z}) < 0$. This introduces a problem as we are not aware of any suitable approaches to obtain multinomial samples when $\lambda(\mathbf{z})$ is negative.

5. Application to data from Malawi. We fit the simplified Naomi model (Section 2) to data from Malawi using three inferential approaches. For each approach, the `TMB` C++ user-template (Appendix S2) used to specify the log-posterior was the same. The three approaches were: 1. TMB, 2. PCA-AGHQ, and 3. 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). The dimension of the latent field was $N = 467$ and the dimension of the hyperparameters was $m = 24$. Settings used for each inferential method are provided in Table 1, and, where relevant, discussed further below. 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 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).

5.1. *NUTS convergence.* We obtained satisfactory NUTS results by gradually increasing the amount of computation until all diagnostics were acceptable. These include the potential scale reduction factors $\hat{R}$ (Gelman and Rubin, 1992; Vehtari et al., 2021) for each parameter, bulk and tail effective sample sizes, autocorrelation decay plots, univariate traceplots, pairs density plots, and NUTS specific divergent transition and energy assessments (Betancourt, 2017). For full details see Appendix S3. We treat the NUTS results as a gold-standard.

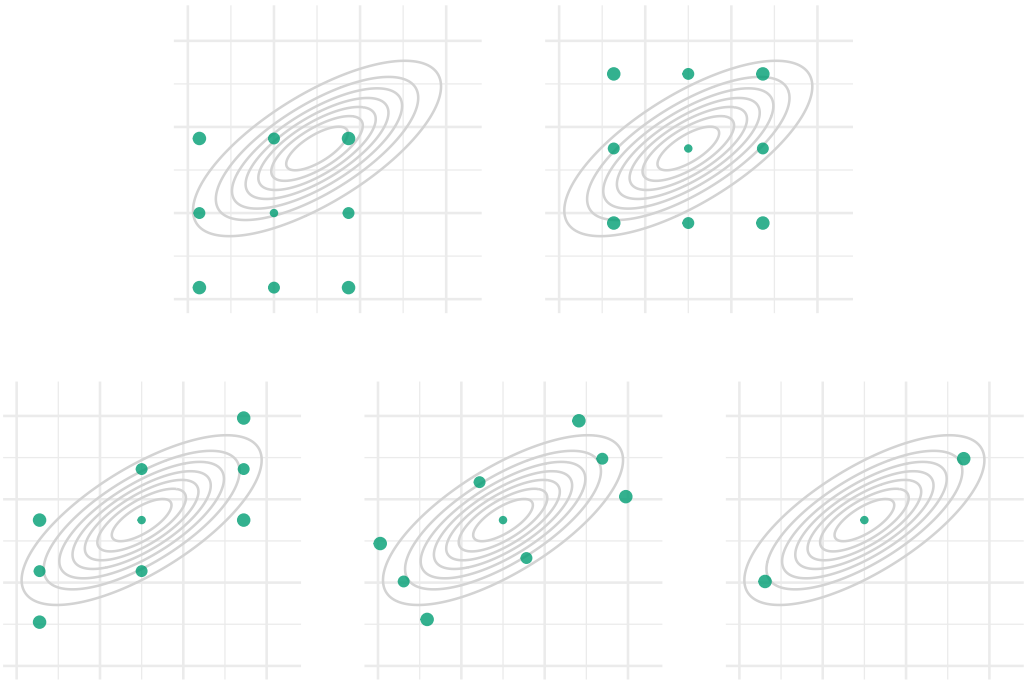

Figure 1: The Gauss-Hermite quadrature nodes $\mathbf{z} \in \mathcal{Q}(2, 3)$ for this two dimensional integral with three nodes per dimension are adapted based on the mean and covariance matrix of the target via the Cholesky decomposition or spectral decompostion of the inverse curvature at the mode.

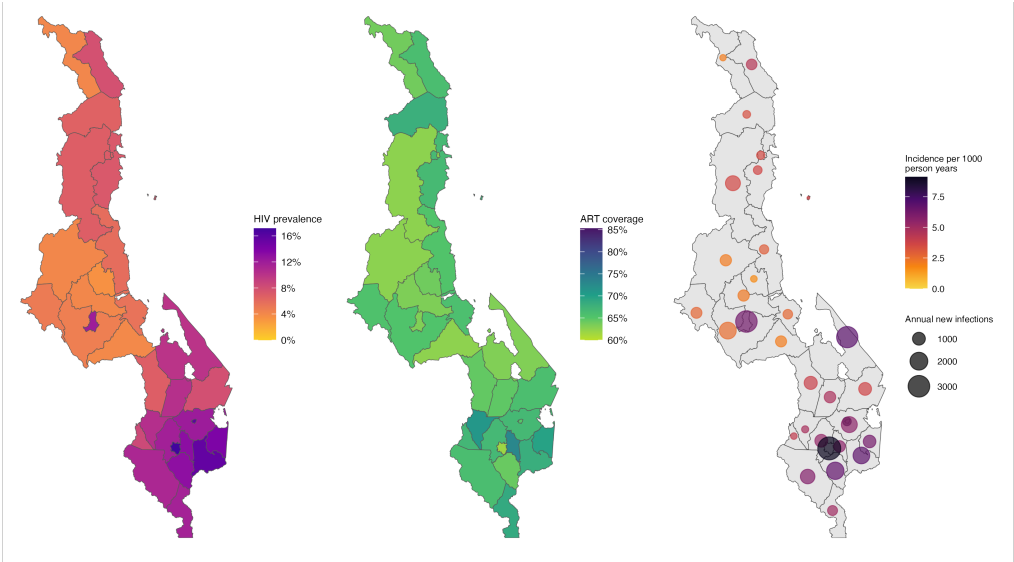

Figure 2: District-level model outputs for adults aged 15-49. Inference conducted with TMB.

5.2. *PCA-AGHQ settings.* We used a Scree plot to select the number of principal components $s < m$ to keep. Figure showing Scree plot. Figure showing approximation to precision matrix.

Name	Software	Details
TMB	`TMB`	1000 samples
PCA-AGHQ	`aghq`	$k = 3, s = 8$, 1000 samples
NUTS	`tmbstan`	4 chains of 20000 iterations, with the first 10000 iterations of each chain discarded as warmup, thinned by a factor of 20. Default NUTS tuning parameters (Hoffman et al., 2014).

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

Figure **??** shows the positions of the generated PCA-AGHQ nodes overlaid onto the hyperparameter marginal posteriors obtained NUTS.

5.3. *Model assessment* .

5.3.1. *Posterior contraction.* Let ϕ be a generic model parameter. To assess the informativeness of the data we compared (Figure 4) the prior variance $\sigma^2_{\text{prior}}(\phi)$ to the posterior variance $\sigma^2_{\text{posterior}}(\phi)$ via the posterior contraction (Schad, Betancourt and Vasishth, 2021)

$$c(\phi) = 1 - \frac{\sigma^2_{\text{posterior}}(\phi)}{\sigma^2_{\text{prior}}(\phi)}.$$

For the parameters of length greater than one, we averaged the posterior contraction.

5.3.2. *Coverage.* We assessed the coverage of our estimates via the uniformity of the data within each posterior marginal distribution. Let $\{\phi_i\}_{i=1}^n$ be posterior marginal samples.

5.4. *Inference comparison* . We used three statistical approaches to assess the accuracy of posterior distributions produced by TMB and PCA-AGHQ as compared with those from NUTS: (1) Kolmogorov-Smirnov tests, (2) Pareto-smoothed importance sampling, and (3) maximum mean discrepancy.

5.4.1. *Kolmogorov-Smirnov tests.* The two-sample Kolmogorov-Smirnov (KS) test statistic (Smirnov, 1948) is the maximum absolute difference between two empirical cumulative distribution (ECDF) functions $F(\varphi) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\phi_i \leq \varphi}$. We compare the KS statistics

$$D_{\text{TMB}} = \sup_{\varphi} |F_{\text{NUTS}}(\varphi) - F_{\text{TMB}}(\varphi)|,$$

$$D_{\text{PCA-AGHQ}} = \sup_{\varphi} |F_{\text{NUTS}}(\varphi) - F_{\text{PCA-AGHQ}}(\varphi)|.$$

See an illustration of the KS test in Figure 5 and a summary of the results in Table.

	D(TMB)	D(PCA-AGHQ)
beta_alpha	0.077	0.076
beta_anc_alpha	0.081	0.077
beta_anc_rho	0.100	0.113
beta_lambda	0.070	0.070
beta_rho	0.071	0.072

log_or_gamma	0.056	0.053
u_alpha_a	0.070	0.035
u_alpha_as	0.074	0.079
u_alpha_x	0.116	0.098
u_alpha_xa	0.071	0.061
u_alpha_xs	0.094	0.079
u_rho_a	0.072	0.083
u_rho_as	0.068	0.063
u_rho_x	0.071	0.067
u_rho_xs	0.147	0.139
ui_anc_alpha_x	0.078	0.075
ui_anc_rho_x	0.055	0.059
ui_lambda_x	0.110	0.113
us_alpha_x	0.088	0.062
us_alpha_xs	0.097	0.063
us_rho_x	0.081	0.083
us_rho_xs	0.039	0.039
Average	0.081	0.075

5.4.2. *Pareto-smoothed importance sampling.* As well as marginal distributions, we are interested in assessing the accuracy of joint distributions. Let $\{\boldsymbol{\phi}_i\}_{i=1}^n$ be generic samples from a joint posterior. Pareto-smoothed importance sampling (PSIS) (Vehtari et al., 2015, Yao et al. (2018)) is a method for stabilising the ratios used in importance sampling. See a summary of the results in Table and Figure 6.

5.4.3. *Maximum mean discrepancy.* Another way to compare joint distributions is via the maximum mean discrepancy [MMD; Gretton et al. (2006)]. Let $\Phi = \{\boldsymbol{\phi}_i\}_{i=1}^n$ and $\Psi = \{\boldsymbol{\psi}_i\}_{i=1}^n$ be two sets of samples from a joint posterior obtained using different inference methods. The MMD can be empirically estimated by

$$\text{MMD}(\Phi, \Psi) = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} k(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j) - \frac{2}{n^2}\sum_{i,j=1}^{n} k(\boldsymbol{\phi}_i, \boldsymbol{\psi}_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(\boldsymbol{\psi}_i, \boldsymbol{\psi}_j)}.$$

See a summary of the results in Table and Figure 7.

5.5. *Case study on exceedance probabilites .*

5.5.1. *Meeting the second 90.* The Joint United Nations Programme on HIV/AIDS has developed ambitious fast-track targets for scaling up ART treatment with the goal of "ending the AIDS epidemic by 2030". Specifically, the "90-90-90 by 2020" fast-track target is that:

- 90% of PLHIV know their status,
- 90% of those are on antiretroviral therapy (ART), and
- 90% of those have suppressed viral load.

Naomi can be used to identify treatment gaps by calculating the probability that the second 90 target has been met $\mathbb{P}(\alpha_i > 0.81)$ for each strata i. We found that both TMB and PCA-AGHQ underestimate the probability that the second 90 target has been met in women (Figure 8). This difficulty may be related to interactions between the household survey and ANC components of the model.

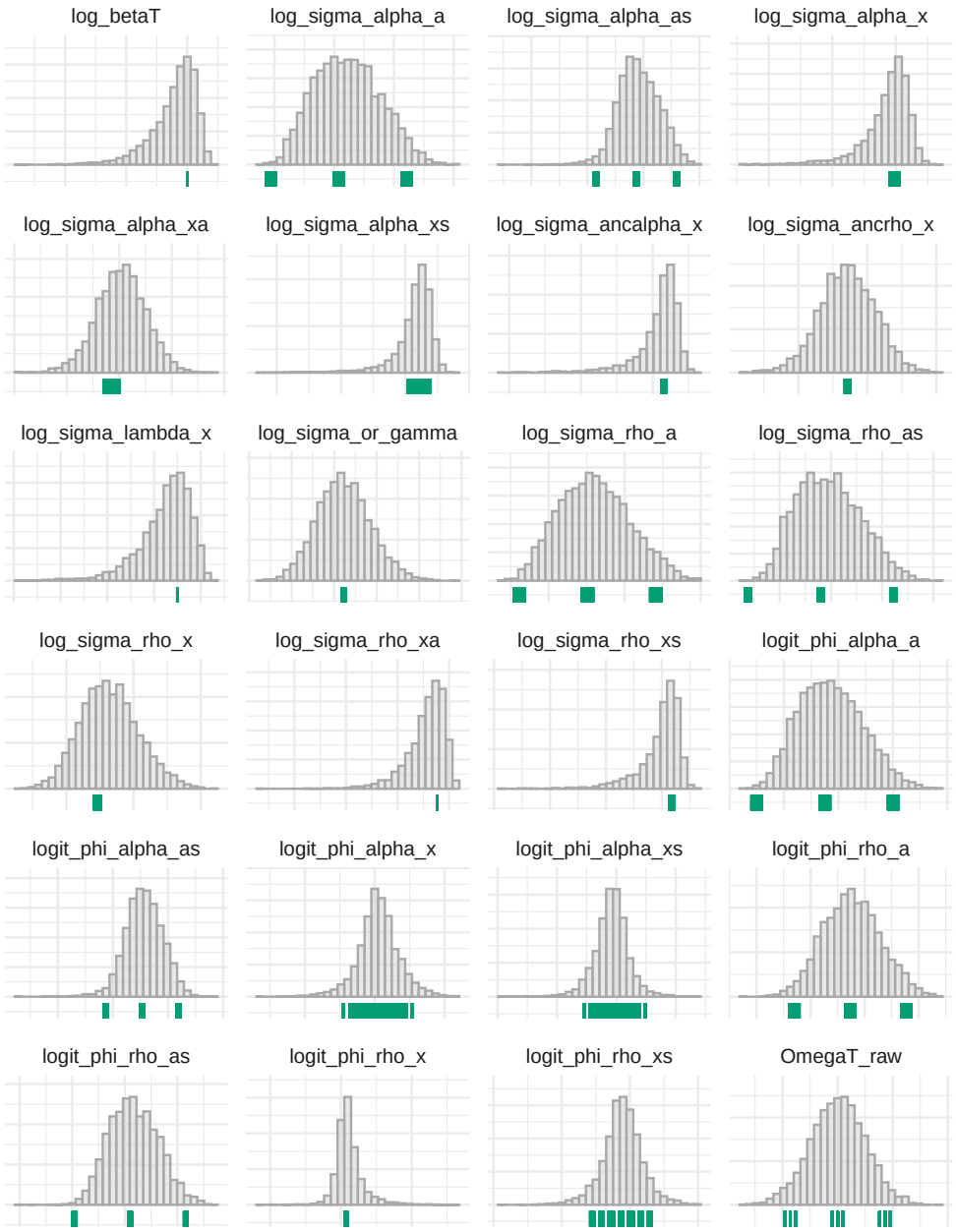

Figure 3: PCA-AGHQ node positions (green, rug plot) overlaid onto the hyperparameter marginal posteriors (grey, histogram) for each of the 24 hyperparameters.

5.5.2. *Finding strata with high incidence.* Some HIV interventions are cost-effective only within high HIV incidence settings, typically defined as greater than 1% incidence per year. Naomi can be used to assess the probability of a strata having high incidence by evaluating $\mathbb{P}(\lambda_i > 0.01)$. We found that both TMB and PCA-AGHQ overestimate these exceedance probabilities (Figure 9). We do not yet have a working hypothesis as to why this is.

6. Discussion. We developed an approximate Bayesian inference algorithm motivated by a challenging problem in small-area estimation of HIV. For the simplified Naomi model

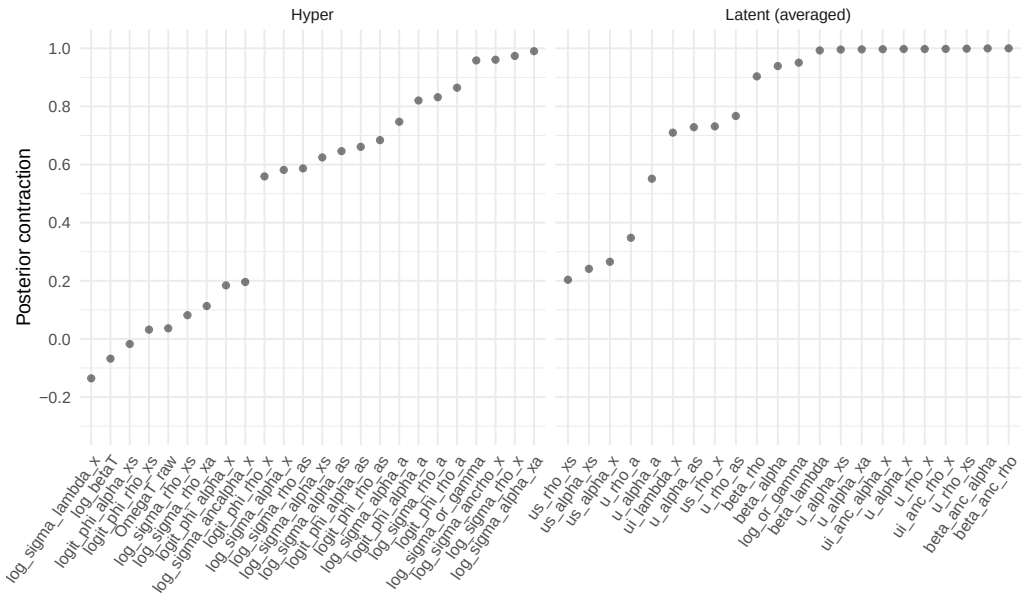

Figure 4: Posterior contraction.

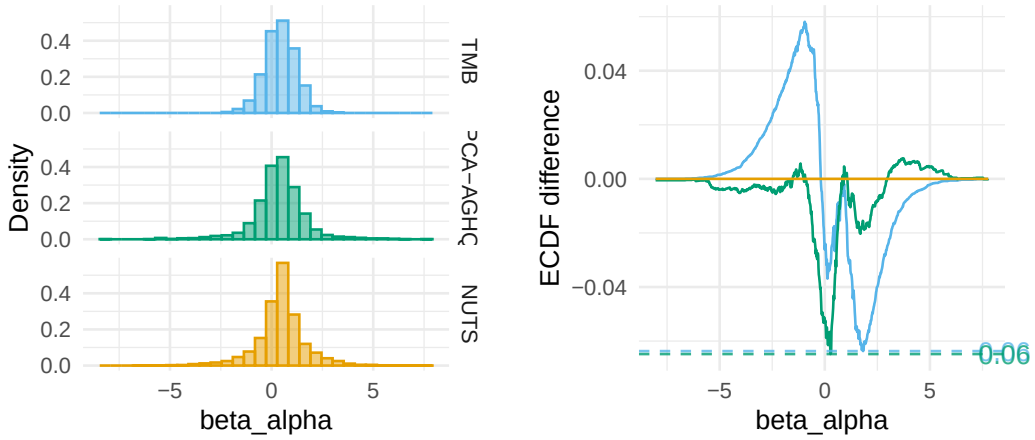

Figure 5: Example KS test for one parameter.

in Malawi (Section 5) our method is demonstrated to be more accurate than TMB, and substantially faster than NUTS. We anticipate that our method could be added to the Naomi web interface as an alternative to TMB. This would enable analysts to quickly iterate over model options using a faster, less accurate inference approach, before switching to a slower, more accurate approach once they are happy with the results.

We provide a flexible implementation of the algorithm, 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 ELGM-type structures of particular interest in spatial epidemiology, many of which feature in Naomi, are: aggregated Gaussian process models (Nandi et al., 2020), evidence synthesis models (Amoah, Diggle and Giorgi, 2020). Although our method

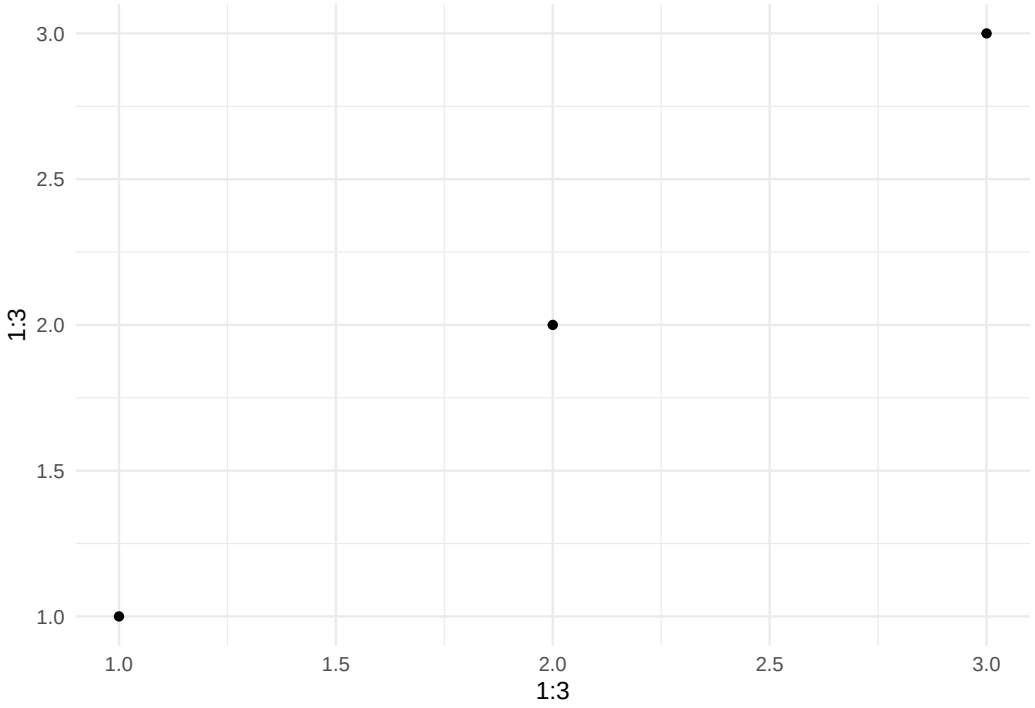

Figure 6: Results of Pareto-smoothed importance sampling analysis.

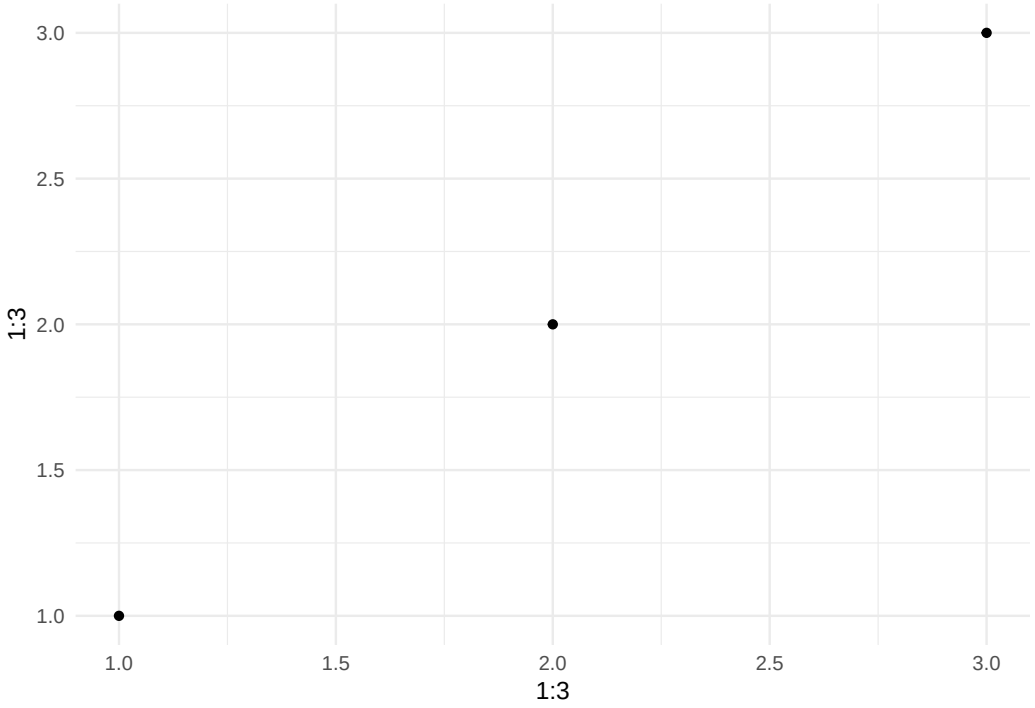

Figure 7: Results of maximum mean discrepancy analysis.

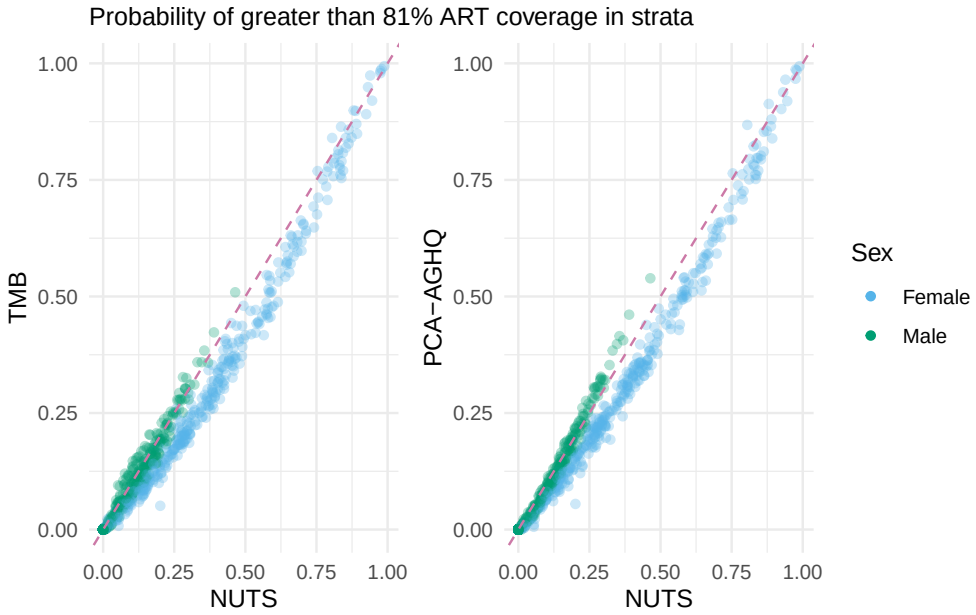

Figure 8: Results of second 90 case-study.

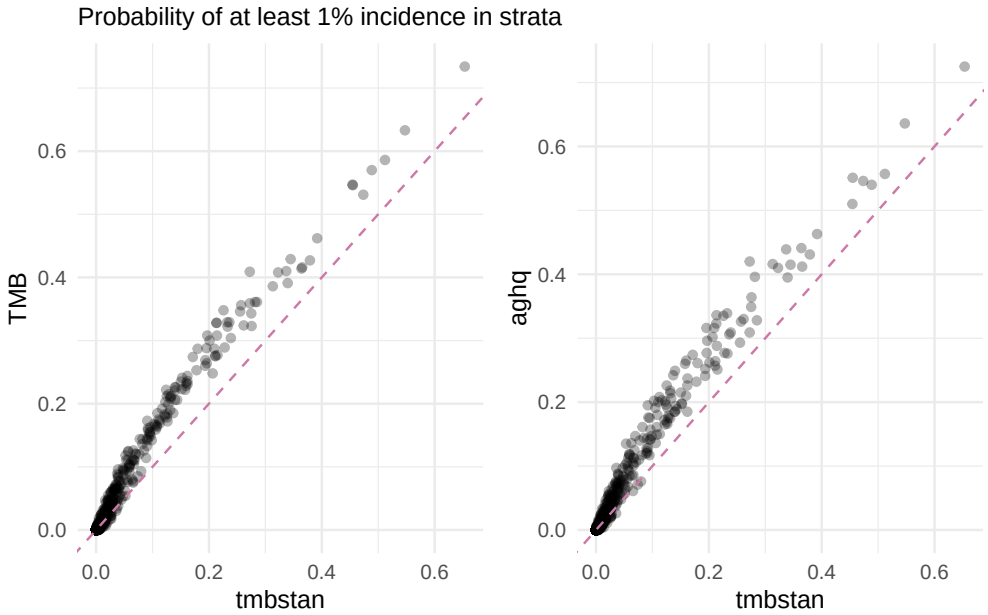

Figure 9: Results of high incidence case-study.

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.

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 5.3) and methods for inference comparison such as Pareto-

smoothed importance sampling (Section 5.4). The exceedance probabilities case-study (Section 5.5) demonstrates the importance of accurate posterior inferences for realistic use-cases of the Naomi model.

Future work could look to implement our algorithm 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.

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.

BETANCOURT, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

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.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science* 457–472.

GRETTON, A., BORGWARDT, K., RASCH, M., SCHÖLKOPF, B. and SMOLA, A. (2006). A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems* **19**.

HOFFMAN, M. D., GELMAN, A. et al. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15** 1593–1623.

JÄCKEL, P. (2005). A note on multivariate Gauss-Hermite quadrature. *London: ABN-Amro. Re*.

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.

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.

SCHAD, D. J., BETANCOURT, M. and VASISHTH, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological methods* **26** 103.

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.

VEHTARI, A., SIMPSON, D., GELMAN, A., YAO, Y. and GABRY, J. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*.

VEHTARI, A., GELMAN, A., SIMPSON, D., CARPENTER, B. and BÜRKNER, P.-C. (2021). Rank-normalization, folding, and localization: an improved R for assessing convergence of MCMC (with discussion). *Bayesian analysis* **16** 667–718.

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.

YAO, Y., VEHTARI, A., SIMPSON, D. and GELMAN, A. (2018). Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning* 5581–5590. PMLR.