FAST APPROXIMATE BAYESIAN INFERENCE OF HIV INDICATORS USING PCA ADAPTIVE GAUSS-HERMITE QUADRATURE

By Adam Howes [1,5] , Alex Stringer [2]
Seth R. Flaxman [3] , Jeffrey W. Eaton [4,5]

[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] *Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health, Harvard University,* jeaton@hsph.harvard.edu

[5] *MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London,*

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 with during a yearly process supported by UNAIDS. Inference has previously been conducted using empirical Bayes and a Gaussian approximation via the `TMB` R package. We propose a new inference method extending adaptive Gauss-Hermite quadrature to deal with >20 hyperparameters. Using data from Malawi, our method improves the accuracy of inferences for model parameters, while being substantially faster to run than Hamiltonian Monte Carlo with the No-U-Turn sampler. However, for model ouptuts, we found the simpler empirical Bayes approach to achieve similar performance. Our implementation is based on the existing `TMB` C++ template for the model's log-posterior, and is compatible with any model with such a template.

1. Introduction. Accurate estimates of HIV indicators are crucial for mounting an effective public health response to the HIV epidemic. These estimates should be timely, and at a geographic level at which health systems are planned and delivered. Producing granular estimates is challenging, in large part due to limitations of the data from available sources. Nationally-representative household surveys provide the most statistically reliable data, but are costly to run and so only conducted every five years or so in most countries, with limited sample size at the district level. Other data sources, such as routine health surveillance of antenatal care (ANC) clinics, are available in closer to real-time, but are not representative of the entire 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, by age and sex. Modelling multiple data sources jointly in this way mitigates the limitations of any single source, increases statistical power, and can prompt investigation into possible conflicts of information.

Software (`https://naomi.unaids.org`) has been developed for Naomi, allowing over 35 countries to input their data and interactively generate estimates during workshops as a part of a yearly process supported by UNAIDS. Creation of estimates by country teams, rather than external agencies or researchers, is an important and distinctive feature of the HIV

Keywords and phrases: Bayesian statistics, spatial statistics, evidence synthesis, small-area estimation, approximate inference, INLA, AGHQ, HIV epidemiology.

response. Drawing on expertise closest to the data being modelled improves the accuracy of the process, as well as strengthening trust in the resulting estimates, creating a virtuous cycle of data quality, use and ownership (Noor, 2022).

Naomi is a complex model, and as such presents a challenging Bayesian inference problem. As well as hundreds of latent field parameters, Naomi has >20 hyperparameters: substantially more than the small number that can typically be handled by approaches like integrated nested Laplace approximations [INLA; Rue, Martino and Chopin (2009)]. Moreover, observations depend on multiple structured additive predictors, such that Naomi falls into the class of extended latent Gaussian models [ELGMs; Stringer, Brown and Stafford (2022)].

To allow for interactive review and iteration of model results by workshop participants, the inference procedure should be fast and have low memory usage. Due to the scale of the model and features of its posterior geometry (Neal, 2003), Markov chain Monte Carlo (MCMC) approaches are prohibitively slow. Furthermore, use of the inference method across countries should be effortless, without requiring substantial statistical expertise, as would be the case for monitoring MCMC convergence and suitability.

To meet these requirements, 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), which we refer to as TMB (distinguished from the package `TMB`). Owing to its speed and flexibility, `TMB` is gaining popularity, particularly in spatial statistics (Osgood-Zimmerman and Wakefield, 2022) and via the user-friendly `glmmTMB` R package (Brooks et al., 2017). 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 parameters. For the Naomi model, this subset is the high-dimensional latent field, leaving a smaller number of hyperparameters. Taking inspiration from the AD Model Builder 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 the TMB approach is fast, within the empirical Bayes framework hyperparameter uncertainty is not accounted for in the latent field posterior. We suspected this would result in underestimation of posterior variances, which may have implications As such, we were motivated to look for an more fully Bayesian inference approach, which is also flexible enough to be compatible with the model, as well as fast enough to be run in production by country teams. We developed an inference method based on adaptive Gauss-Hermite quadrature (AGHQ) extended to handle integration over large numbers of hyperparameters. AGHQ is a quadrature method based on the theory of polynomial interpolation, which is well suited to statistical estimation problems. Bilodeau, Stringer and Tang (2022) prove stochastic convergence rates for Bayesian posterior quantities when the normalising constant is estimated using AGHQ. However, it is not computationally feasible to use AGHQ in high dimensions directly, as exponentially many nodes would be required. Instead, we use principal components analysis (PCA) of the inverse curvature at the mode to find a smaller number of dimensions which explain most of the variance. For the Naomi model in Malawi, this results in a grid which has millions of times fewer nodes the corresponding dense grid and is tractable to run. Our implementation of the method makes use of the existing Naomi `TMB` template, and is immediately compatible with any model with such a template.

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

The remainder of this paper is organised as follows. Section 2 outlines the version of the Naomi model that we consider in this paper, and Section 3 describes how it falls within the ELGM framework. In Section 4 we review the deterministic inference method for ELGMs used by Stringer, Brown and Stafford (2022) based on nested application of AGHQ and the Laplace approximation, before introducing the PCA-based modification we use to enable application to Naomi. In a case study (Section 5) we evaluate the accuracy of PCA-AGHQ for the simplified Naomi model fit to data from Malawi, as compared with TMB and gold-standard MCMC. Finally, we discuss our conclusions, and directions for future research in Section 6.

2. Simplified Naomi model.

Eaton et al. (2021) specify a joint model linking three small-area estimation models. We consider a simplified version defined only at the time of the most recent household survey with HIV testing, omitting nowcasting and temporal projection, as these time points involve limited inferences. An overview of the simplified model is given below, and a more complete mathematical description is provided in Appendix S1.

2.1. *Household survey component* .

Consider a country in sub-Saharan Africa where a household survey with complex survey design has taken place. 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, and $\mathcal{I}$ be the set of all I such that $I \in \mathcal{I}$.

Let $N_i \in \mathbb{N}$ be the known, fixed population size. We infer the following unknown HIV indicators using linked 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.

We specify independent logistic regression models for HIV prevalence and ART coverage in the general population such that $\text{logit}(\rho_i) = \eta^\rho_i$ and $\text{logit}(\alpha_i) = \eta^\alpha_i$. HIV incidence rate is modelled on the log scale as $\log(\lambda_i) = \eta^\lambda_i$, and depends on adult HIV prevalence and adult ART coverage. The structured additive predictors η^θ_i for $\theta \in \{\rho, \alpha, \lambda\}$ are given in Appendix S1. 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 each informed by household survey data. We first calculate the weighted aggregate survey observations

$$\hat \theta_I = \frac{\sum_j w_j \cdot \theta_j}{\sum_j w_j},$$

with individual responses $\theta_j \in \{0, 1\}$ and design weights w_j for each of $\theta \in \{\rho, \alpha, \kappa\}$. The design weights are provided by the survey and aim to reduce bias by decreasing possible correlation between response and recording mechanism (Meng, 2018). The index j runs across all individuals in strata $i \in I$ within the relevant denominator i.e. for ART coverage, only those individuals who are HIV positive. We take the weighted observed number of outcomes to be $y_I^\theta = m_I^\theta \cdot \hat{\theta}_I$ where

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

is the Kish effective sample size (ESS) (Kish, 1965). As the Kish ESS is maximised by constant design weights, in exchange for reducing bias we reduce our ESS and thereby increase variance. We use a binomial working likelihood defined to operate on the reals

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

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

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

Though our approach accounts for survey weights, it does not take into account sample correlation structure, for example due to cluster sampling (Wakefield, Okonek and Pedersen, 2020).

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. We take the number of ANC clients with ascertained status to be fixed as $m_I^{\rho^{\text{ANC}}}$. We then model the number of those with positive status $y_I^{\rho^{\text{ANC}}}$, and the number of those already on ART prior to their first ANC visit $y_I^{\alpha^{\text{ANC}}}$ using nested binomial likelihoods

$$y_I^{\rho^{\text{ANC}}} \sim \text{Bin}(m_I^{\rho^{\text{ANC}}}, \rho_I^{\text{ANC}}),$$
$$y_I^{\alpha^{\text{ANC}}} \sim \text{Bin}(y_I^{\rho^{\text{ANC}}}, \alpha_I^{\text{ANC}}).$$

As in the household survey component, 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 model the probabilities of accessing services outside the home district using multinomial logistic regressions. Briefly, 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 such that $x \sim x'$. We model the log-odds $\tilde{\gamma}_{x,x'} = \text{logit}(\gamma_{x,x'})$ using a structured additive predictor $\eta_x^{\tilde{\gamma}}$ which only depends on the home district x. As such, we assume travel to each neighbouring district, for all age-sex strata, is equally likely. We then model aggregate ART attendance data $y_I^{N^{\text{ART}}}$ using a Gaussian approximation to a sum of binomials. This sum is over both strata $i \in I$ and the 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. *Summary.* In all, Naomi is a joint model on the observations $\mathbf{y} = (y_I^\theta)$ for $\theta \in \{\rho, \alpha, \kappa, \rho^{\text{ANC}}, \alpha^{\text{ANC}}, N^{\text{ART}}\}$ and $I \in \mathcal{I}$. The structured additive predictors contain intercept effects, age random effects, and spatial random effects which we collectively describe as the latent field $\mathbf{x}$. The latent field is controlled by hyperparamters $\boldsymbol{\theta}$ which include standard deviations, first-order autoregressive model correlation parameters, and reparameterised Besag-York-Mollie model [BYM2; Simpson et al. (2017)] proportion parameters.

3. Extended Latent Gaussian models.

We now describe the 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 with likelihood

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

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 $\boldsymbol{\theta}_1 \in \mathbb{R}^{s_1}$, 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

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

with latent field and hyperparameter priors as in the LGM case.

3.2. *Naomi as an ELGM.* Naomi is a spatio-temporal model with a large Gaussian latent field, governed by a smaller number of hyperparameters. However, it is an ELGM rather than an LGM, for the reasons below. Note that when dependence on a specific number of structured additive predictors is given, it is for that factor in isolation, and as such should be considered illustrative.

1. In the household survey component, HIV incidence depends on district-level adult HIV prevalence and ART coverage. This reflects basic HIV epidemiology: HIV incidence is proportional to unsuppresed viral load such that such that $\lambda \propto \rho(1 - \omega \cdot \alpha)$, with $\omega = 0.7$ a fixed constant. As a result, each $\log(\lambda_i)$ depends on 28 structured additive predictors (where 28 arises from the product of 2 sexes [male and female], 7 age groups, [{15-19, ..., 45-49}], and 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. Though $\text{logit}(\rho_i)$ and $\text{logit}(\alpha_i)$ are Gaussian, this nonetheless 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 i, but likelihoods are defined for observations aggregated over sets of indices $i \in I$. As such, all observations are related to $|I|$ 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, one for each neighbouring district plus one for remaining in the home district.
7. Multiple link functions are used throughout the model, such that there is no one inverse link function g.

4. Inference methods for Naomi . We first describe (Section 4.1) the inference method for ELGMs based on nested applications of the Laplace approximation and AGHQ used by Stringer, Brown and Stafford (2022). We then propose (Section 4.2) an extension of the method which uses PCA to facilitate inference for Naomi, which otherwise would be intractable.

4.1. *Inference for ELGMs.* The joint posterior of the parameters $(\mathbf{x}, \boldsymbol{\theta})$ given data $\mathbf{y}$ in an ELGM is given by

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

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

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

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

4.1.1. *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

$$\hat{\mathbf{x}}(\boldsymbol{\theta}) = \arg\max_{\mathbf{x}} \log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}), \tag{4.3}$$

$$\hat{\mathbf{H}}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}^\top} \log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) |_{\mathbf{x} = \hat{\mathbf{x}}(\boldsymbol{\theta})}. \tag{4.4}$$

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

$$\tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{\tilde{p}_{\text{G}}(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})}\Big|_{\mathbf{x}=\hat{\mathbf{x}}(\boldsymbol{\theta})} = \sqrt{\frac{|\hat{\mathbf{H}}(\boldsymbol{\theta})|}{(2\pi)^N}} p(\mathbf{y}, \hat{\mathbf{x}}(\boldsymbol{\theta}), \boldsymbol{\theta}). \tag{4.5}$$

Inference proceeds by optimising Equation 4.5 using a gradient-based routine to obtain $\hat{\boldsymbol{\theta}}_{\text{LA}} = \arg\max_{\boldsymbol{\theta}} \tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y})$. Each evaluation in the optimisation requires an inner optimisation to obtain $\hat{\mathbf{x}}(\boldsymbol{\theta})$ via Equation 4.3. Supposing the hyperparameters are to be considered fixed, as with the TMB approach used currently for Naomi, then latent field joint and marginal inferences then follow directly from the Gaussian approximation $\tilde{p}_{\text{G}}(\mathbf{x} \mid \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y})$. Hyperparameter inferences can be obtained according to some method which should be specified here as well.

4.1.2. *Adaptive Gauss-Hermite quadrature.* Let $\mathbf{z} \in \mathcal{Q}(m, k)$ be m-dimensional Gauss-Hermite quadrature [GHQ; Davis and Rabinowitz (1975)] rule with k nodes per dimension constructed using the product rule such that $\mathcal{Q}(m, k) = \mathcal{Q}(1, k) \times \cdots \times \mathcal{Q}(1, k)$ where

$$\mathcal{Q}(1, k) = \{z : H_k(z) = (-1)^k \exp(z^2/2) \frac{\text{d}}{\text{d}z^k} \exp(-z^2/2) = 0\}, \tag{4.6}$$

with $\phi(\cdot)$ is a standard Gaussian density. The corresponding weighting function $\omega : \mathcal{Q}(m, k) \to \mathbb{R}$ is given by $\omega(\mathbf{z}) = \prod_{j=1}^m \omega(z_j)$ where $\omega(z) = k!/[H_{k+1}(z)]^2 \phi(z)$. For $k = 1$ GHQ corresponds to a Laplace approximation. Further GHQ is exact for functions which are a Gaussian density multiplied by a polynomial of total order no more than $2k - 1$, a class of functions we expect the posterior to be close to.

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 $[\hat{\mathbf{H}}_{\text{LA}}(\hat{\boldsymbol{\theta}}_{\text{LA}})]^{-1} = \hat{\mathbf{P}}_{\text{LA}} \hat{\mathbf{P}}_{\text{LA}}^\top$ be a matrix decomposition of the inverse curvature. An adaptive Gauss-Hermite quadrature [AHGQ; Naylor and Smith (1982); Tierney and Kadane (1986)] estimate of the normalising constant $p(\mathbf{y})$ based on the Laplace approximation is given by

$$p(\mathbf{y}) \approx \int_{\boldsymbol{\theta}} \tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y}) \approx \tilde{p}_{\text{AGHQ}}(\mathbf{y}) = |\hat{\mathbf{P}}_{\text{LA}}| \sum_{\mathbf{z} \in \mathcal{Q}(m,k)} \tilde{p}_{\text{LA}}(\hat{\mathbf{P}}_{\text{LA}} \mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y}) \omega(\mathbf{z}). \tag{4.7}$$

The unadapted nodes are shifted by the mode and rotated by a matrix decomposition of the inverse curvature such that $\mathbf{z} \mapsto \hat{\mathbf{P}}_{\text{LA}} \mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}}$. Repositioning the nodes is crucial for statistical quadrature problems like ours, where the integral depends on data $\mathbf{y}$ and regions of high density are not known in advance. Two alternatives for the matrix decomposition (Jäckel, 2005) are (1) the Cholesky decomposition $\hat{\mathbf{P}}_{\text{LA}} = \hat{\mathbf{L}}_{\text{LA}}$, where $\hat{\mathbf{L}}_{\text{LA}}$ is lower triangular, and (2) the spectral decomposition $\hat{\mathbf{P}}_{\text{LA}} = \hat{\mathbf{E}}_{\text{LA}} \hat{\mathbf{\Lambda}}_{\text{LA}}^{1/2}$, where $\hat{\mathbf{E}}_{\text{LA}} = (\hat{\mathbf{e}}_{\text{LA},1}, \ldots \hat{\mathbf{e}}_{\text{LA},m})$ contains the eigenvectors of $[\hat{\mathbf{H}}_{\text{LA}}(\hat{\boldsymbol{\theta}}_{\text{LA}})]^{-1}$ and $\hat{\mathbf{\Lambda}}_{\text{LA}}$ is a diagonal matrix containing its eigenvalues $(\hat{\lambda}_{\text{LA},1}, \ldots, \hat{\lambda}_{\text{LA},m})$. This estimate may be used to normalise the Laplace approximation

$$\tilde{p}_{\text{LA}}(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_{\text{AGHQ}}(\mathbf{y})}. \tag{4.8}$$

To obtain inferences for the latent field (Equation 4.1) we reuse the adapted nodes and weights (Rue, Martino and Chopin, 2009; Stringer, Brown and Stafford, 2022)

$$\tilde{p}(\mathbf{x} \mid \mathbf{y}) = |\hat{\mathbf{P}}_{\text{LA}}| \sum_{\mathbf{z} \in \mathcal{Q}(m,k)} \tilde{p}_{\text{G}}(\mathbf{x} \mid \hat{\mathbf{P}}_{\text{LA}} \mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y}) \tilde{p}_{\text{LA}}(\hat{\mathbf{P}}_{\text{LA}} \mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}} \mid \mathbf{y}) \omega(\mathbf{z}). \tag{4.9}$$

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

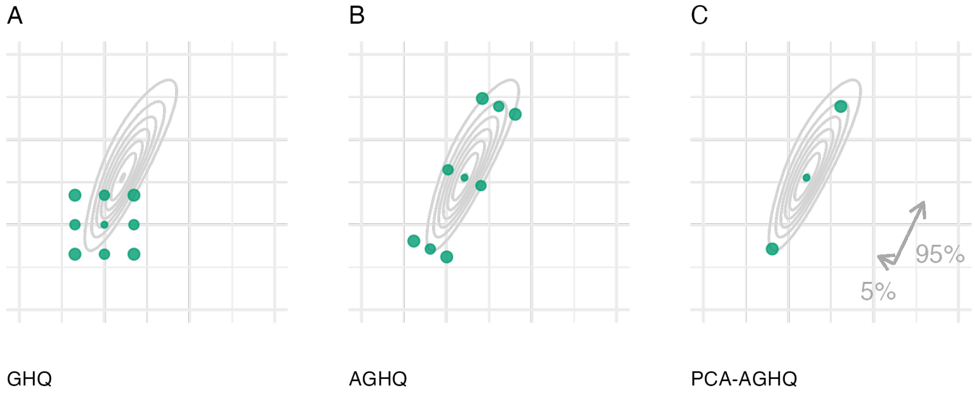

FIGURE 1. *The Gauss-Hermite quadrature nodes* $\mathbf{z} \in \mathcal{Q}(2,3)$ *for a two dimensional integral with three nodes per dimension (A). Adaption occurs based on the mode and covariance matrix of the target via the Cholesky decomposition of the inverse curvature at the mode (B). In PCA-AGHQ (C) only nodes along the first s principal components are kept. Here, 95% of variation is explained by the first principal component. The integrand is* $f(\boldsymbol{\theta}) = sn(0.5\theta_1, \alpha = 2) \cdot sn(0.8\theta_1 - 0.5\theta_2, \alpha = -2)$*, where* $sn(\cdot)$ *is the standard skewnormal probability density function with shape parameter* $\alpha \in \mathbb{R}$*.*

4.2. *Principal components analysis.* Use of the product rule grid described above requires $|\mathcal{Q}(m,k)| = k^m$ quadrature points. This quickly 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 $\boldsymbol{\theta}$. We may then define $\mathcal{Q}(m, \mathbf{k}) = \mathcal{Q}(1, k_1) \times \cdots \times \mathcal{Q}(1, k_m)$ to be a GHQ grid with possible variable levels 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 $s \leq m$. For example, for $m = 2$ and $s = 1$ then $\mathbf{k} = (k, 1)$. In combination with use of the spectral decomposition, this choice of levels is analogous to a principal components analysis (PCA) approach to AGHQ. We refer to this approach as PCA-AGHQ, with corresponding estimate of the normalising constant given by

$$\tilde{p}_{\texttt{PCA}}(\mathbf{y}) = |\hat{\mathbf{E}}_{\texttt{LA}} \hat{\mathbf{\Lambda}}_{\texttt{LA}}^{1/2}| \sum_{\mathbf{z} \in \mathcal{Q}(m,s,k)} \tilde{p}_{\texttt{LA}}(\hat{\mathbf{E}}_{\texttt{LA},s} \hat{\mathbf{\Lambda}}_{\texttt{LA},s}^{1/2} \mathbf{z} + \hat{\boldsymbol{\theta}}_{\texttt{LA}}, \mathbf{y}) \omega(\mathbf{z}), \tag{4.10}$$

where $\hat{\mathbf{E}}_{\texttt{LA},s}$ is an $m \times s$ matrix containing the first s eigenvectors, $\hat{\mathbf{\Lambda}}_{\texttt{LA},s}$ is the $s \times s$ diagonal matrix containing the first s eigenvalues, and $\omega(\mathbf{z}) = \prod_{j=1}^{s} \omega_s(z_j) \times \prod_{j=s+1}^{d} \omega_1(z_j)$. Panel C of Figure 1 illustrates PCA-AGHQ for a case when $m = 2$ and $s = 1$. As AGHQ with $k = 1$ corresponds to the Laplace approximation, PCA-AGHQ can be interpreted as performing AGHQ on the first s principal components of the inverse curvature, and a Laplace approximation on the remaining $m - s$ principal components. Inference for the latent field follows analogously to Equation 4.9.

5. Application to data from Malawi.

We fit the simplified Naomi model (Section 2) to data from Malawi using three inferential approaches. These were:

1. TMB (54 seconds), based on a Gaussian approximation at $\hat{\boldsymbol{\theta}}_{\texttt{LA}}$.
2. PCA-AGHQ (1.2 hours), based on a Gaussian approximation mixture at the adapted nodes $\mathbf{z} \in \mathcal{Q}(m, s, k)$, as described in Section 4.2, and implemented via extension of the `aghq` package (Stringer, 2021).

Method	Software	Details
TMB	`TMB`	1000 samples
PCA-AGHQ	`aghq`	$k = 3, s = 8$ (see Section 5.2), 1000 samples
NUTS	`tmbstan`	4 chains of 100000 iterations, with the first 50000 iterations of each chain discarded as warmup, thinned by a factor of 40, to give a total of 5000 samples kept. Default NUTS tuning parameters (Hoffman et al., 2014).

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

3. NUTS (3.3 days), the Hamiltonian Monte Carlo (HMC) algorithm No-U-Turn Sampling using Stan (Carpenter et al., 2017) implemented via the `tmbstan` package (Monnahan and Kristensen, 2018).

Our goal was to determine the accuracy of the approximate methods (TMB and PCA-AGHQ) as compared with the gold-standard (NUTS). Settings used for each inferential method are provided in Table 1, and, where relevant, discussed further below. The `TMB` C++ user-template used to specify the log-posterior was the same for each approach. The dimension of the latent field was $N = 467$ and the dimension of the hyperparameters was $m = 24$. 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. To provide intuition, model outputs from TMB are illustrated in Figure 2.

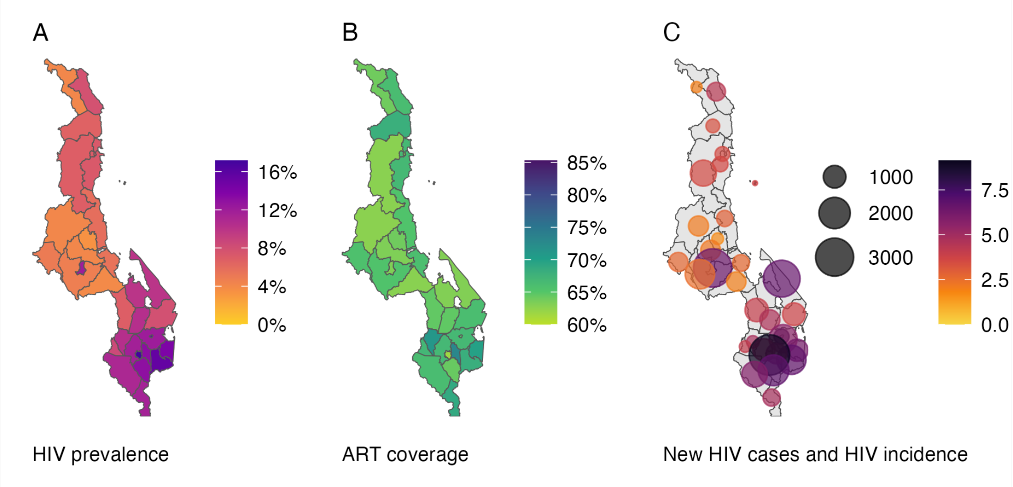

FIGURE 2. *District-level HIV prevalence (A), ART coverage (B), and new HIV cases and HIV incidence (C) for adults 15-49 in Malawi. Inference conducted using TMB.*

The R (R Core Team, 2021) code used to produce all results we describe below is available at `github.com/athowes/naomi-aghq`. 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.* Due to low effective sample size ratios (Figure S9), obtaining acceptable NUTS diagnostics required four chains run in parallel for 100000 iterations, thinned by a factor of 20 for ease-of-storage. There were no divergent transitions, and the largest potential scale reduction factor (Gelman and Rubin, 1992; Vehtari et al., 2021) was $\hat{R} = 1.021$ (Figure S8). We considered the NUTS results a gold-standard, though inaccuracies remain possible. For full details see Appendix S5.

5.2. *Use of PCA-AGHQ .* We used a Scree plot based on the spectral decomposition of $\hat{\mathbf{H}}_{\texttt{LA}}(\hat{\boldsymbol{\theta}}_{\texttt{LA}})^{-1}$ to select the number of principal components to keep (Figure S3). We found that $s = 8$ principal components were sufficient to explain 87% of total variation. This choice of s gave a visually similar reduced rank approximation to the inverse curvature (Figure S4).

5.2.1. *Visual inspection.* Overlaying the resulting $3^8 = 6561$ PCA-AGHQ nodes onto the hyperparameter marginal posteriors obtained using NUTS, we found approximately 12 of the 24 hyperparameters had well covered marginals (Figure 3). Though 12 does improve on the 8 naively obtained with a dense grid, there remained poorly covered hyperparameters. Coverage was associated with marginal standard deviation (Figure S5). All constrained hyperparameters θ were transformed to the real line, using either a log ($\theta > 0$) or logit ($\theta \in [0, 1]$) transformation. As a result, marginal standard deviations for log transformed hyperparameters were systematically smaller than those which were logit transformed (Figure S6).

5.2.2. *Interpreation of eigenvectors.* The ordered eigenvectors correspond to the directions of greatest variation in the inverse curvature. The first and second eigenvectors each contain coupled AR1 standard deviation and correlation parameters (Figure S7). These parameters are weakly identified, and have high correlation in the the NUTS posterior pairs plot (Figure S12, average absolute correlation across all four pairs of 0.81). The reason why is that the same amount of variation can equally be explained by high standard deviation and high correlation or low standard deviation and low correlation. The BYM2 standard deviation and proportion parameters on the other hand are designed to be orthogonal, and as such did not display posterior correlation (Figure S13, average absolute correlation across all four pairs of 0.17) or appear prominently in the eigenvectors.

5.2.3. *Normalising constant estimation.* We assessed appropriateness of the quadrature grid by comparing the estimate of $\log p_{\texttt{PCA}}(\mathbf{y})$ for a range of settings. Convergence in $\log p_{\texttt{PCA}}(\mathbf{y})$ as s and k are increased may suggest a suitable grid has been reached. Appendix S3.2 shows those values which we could compute in a reasonable time (less than 24 hours using a high performance computing cluster).

5.3. *Model assessment.*

5.3.1. *Posterior contraction.* To assess the informativeness of the data we compared the prior variance $\sigma^2_{\text{prior}}(\psi)$ to the posterior variance $\sigma^2_{\text{posterior}}(\psi)$ via the posterior contraction (Schad, Betancourt and Vasishth, 2021)

$$c(\psi) = 1 - (\sigma^2_{\text{posterior}}(\psi)/\sigma^2_{\text{prior}}(\psi)), \tag{5.1}$$

where ψ is a model parameter. We found that (Figure S2) something something. For greater interpretability, facet parameters in this plot according to model component.

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

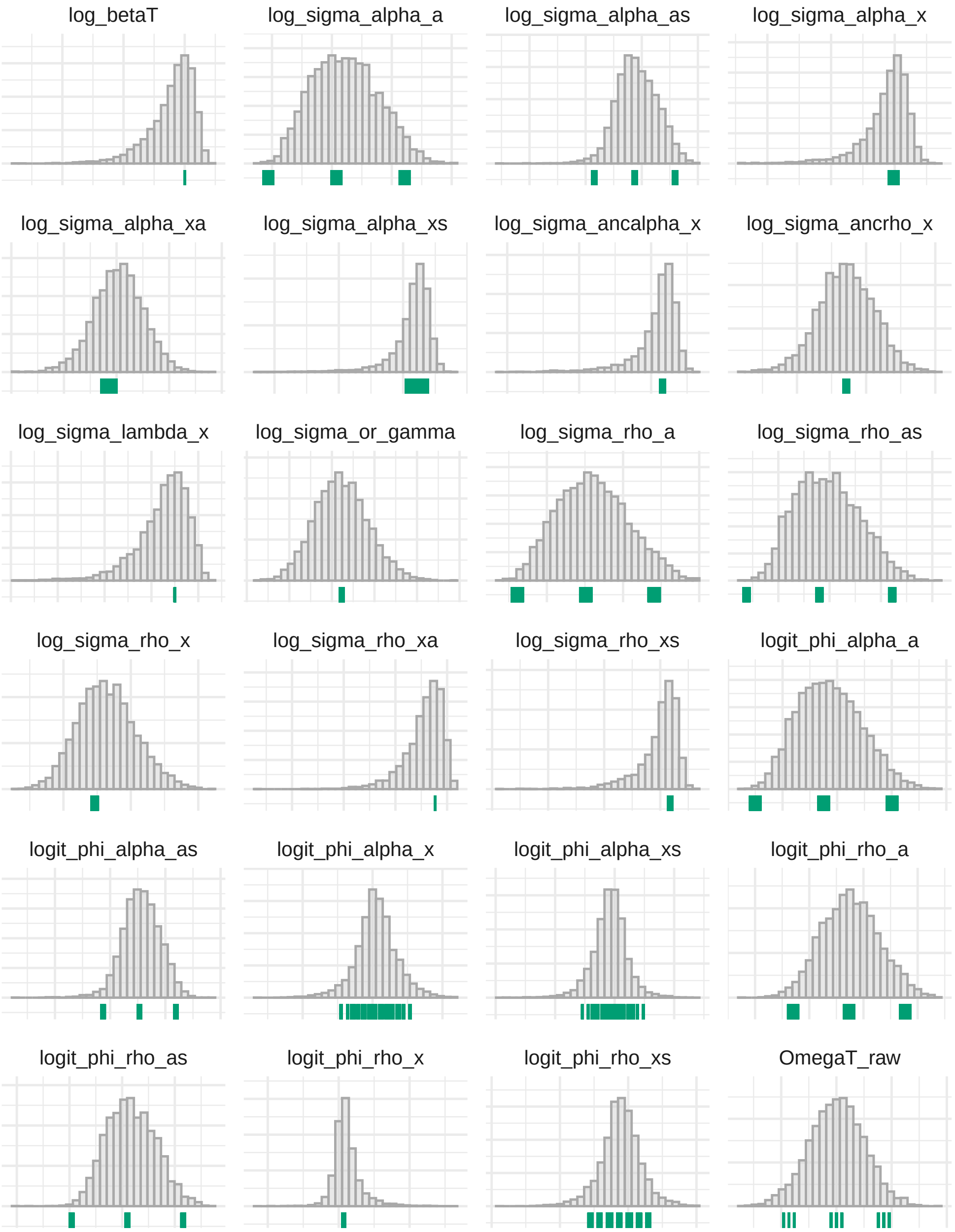

FIGURE 3. *The 6561 PCA-AGHQ node positions (green, rug plot) projected onto the hyperparameter marginal posteriors (grey, histogram) for each of the 24 hyperparameters. Some hyperparameters, such as* `logit_phi_alpha_x`*, are well covered where as others, such as* `log_sigma_lambda_x`*, are near only being covered by one unique node.*

5.4. *Inference comparison.* We compared the accuracy of posterior distributions produced by TMB and PCA-AGHQ as compared with those from NUTS for latent field parameters and model outputs. The metrics we used were (1) marginal point estimates, (2) marginal Kolmogorov-Smirnov and Anderson-Darling tests using the empirical cumulative distribution function (ECDF), (3) joint Pareto-smoothed importance sampling results, and (4) joint maximum mean discrepancy results.

5.4.1. *Point estimates.* For the latent field, the root mean square error (RMSE) between posterior mean estimates from PCA-AGHQ and NUTS (0.063) was 20% lower than that between TMB and NUTS (0.078). For the posterior standard deviation estimates, there was a substantial 60% reduction in RMSE: from 0.14 (TMB) to 0.05 (PCA-AGHQ). These results, alongside those for the mean absolute error (MAE), are presented in Figure 4.

These improvements did not transfer to the model outputs (Figure 5).

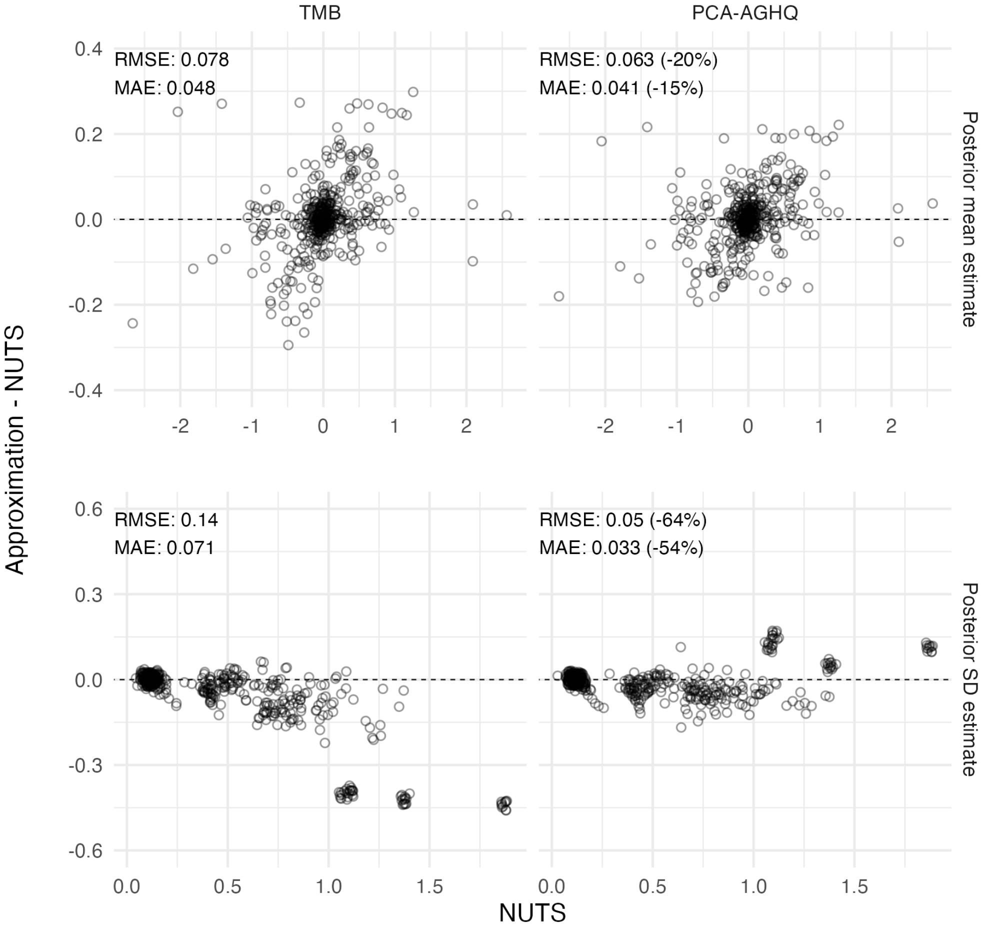

FIGURE 4. *For the latent field PCA-AGHQ modestly improves estimation of the posterior mean, and substantially improves estimation of the posterior standard deviation, as compared with TMB.*

5.4.2. *Distribution tests.* The two-sample Kolmogorov-Smirnov (KS) test statistic (Smirnov, 1948) is the maximum absolute difference between two ECDFs $F(\omega) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\psi_i \leq \omega}$. See Figure 6 for an example. Additionally, we found that KS test results were negatively correlated with ESS (Figure S14).

5.4.3. *Pareto-smoothed importance sampling.* Let $\{\psi_i\}_{i=1}^{n}$ be joint posterior samples. Pareto-smoothed importance sampling [PSIS; Vehtari et al. (2015), Yao et al. (2018)] is a method for stabilising the ratios used in importance sampling. Results for the PSIS analysis are pending.

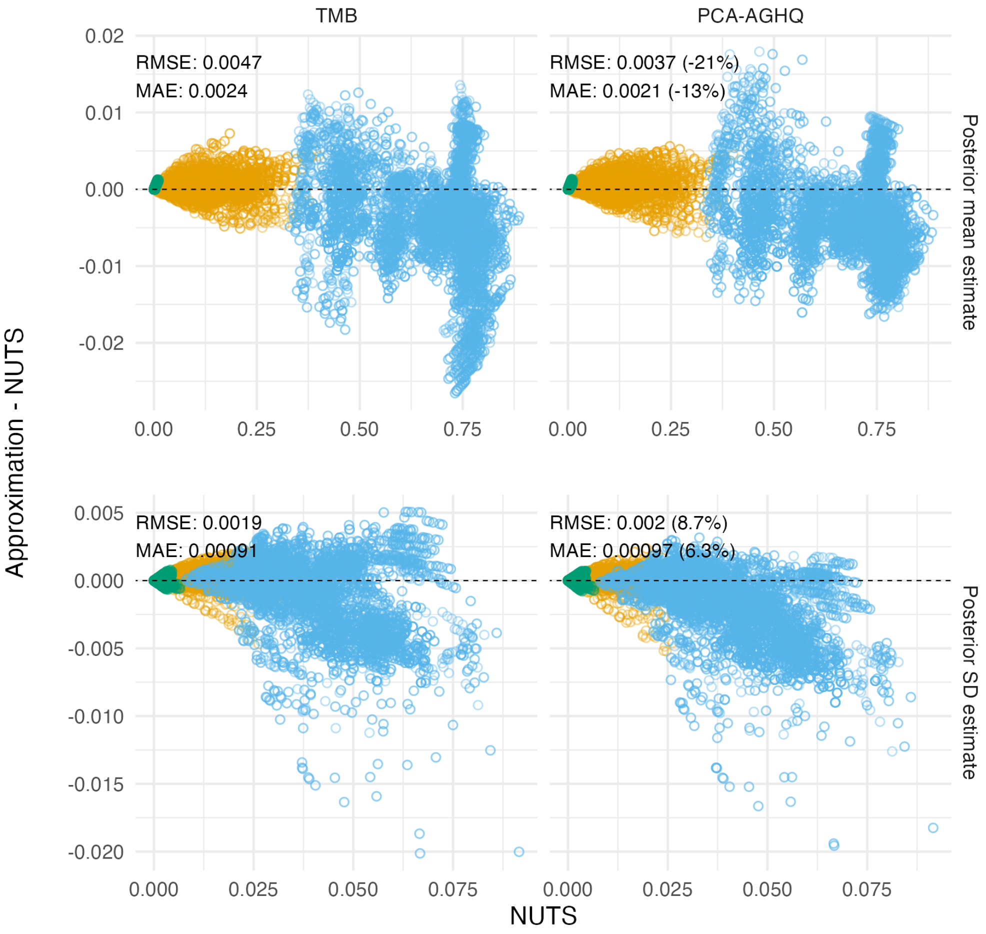

FIGURE 5. *PCA-AGHQ doesn't do so well for the model outputs.*

5.4.4. *Maximum mean discrepancy.* Let $\Psi^1 = \{\psi_i^1\}_{i=1}^n$ and $\Psi^2 = \{\psi_i^2\}_{i=1}^n$ be two sets of joint posterior samples, and k be a kernel. The maximum mean discrepancy [MMD; Gretton et al. (2006)] can be empirically estimated by

$$\text{MMD}(\Psi^1, \Psi^2) = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} k(\psi_i^1, \psi_j^1) - \frac{2}{n^2}\sum_{i,j=1}^{n} k(\psi_i^1, \psi_j^2) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(\psi_i^2, \psi_j^2)}.$$

We set $k(\psi^1, \psi^2) = \exp(-\sigma\|\psi^1 - \psi^2\|^2)$ with σ estimated from data using the `kernlab` R package (Karatzoglou et al., 2019). As compared with NUTS, the MMD from PCA-AGHQ (0.071) was 11% smaller than that of TMB (0.080).

5.5. *Case study on exceedance probabilities.*

5.5.1. *Meeting the second 90.* Ambitious fast-track targets for scaling up ART treatment have been developed by UNAIDS, with the goal of "ending the AIDS epidemic by 2030". Meeting the "90-90-90" fast-track target requires that 90% of people living with HIV know their status, 90% of those are on 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

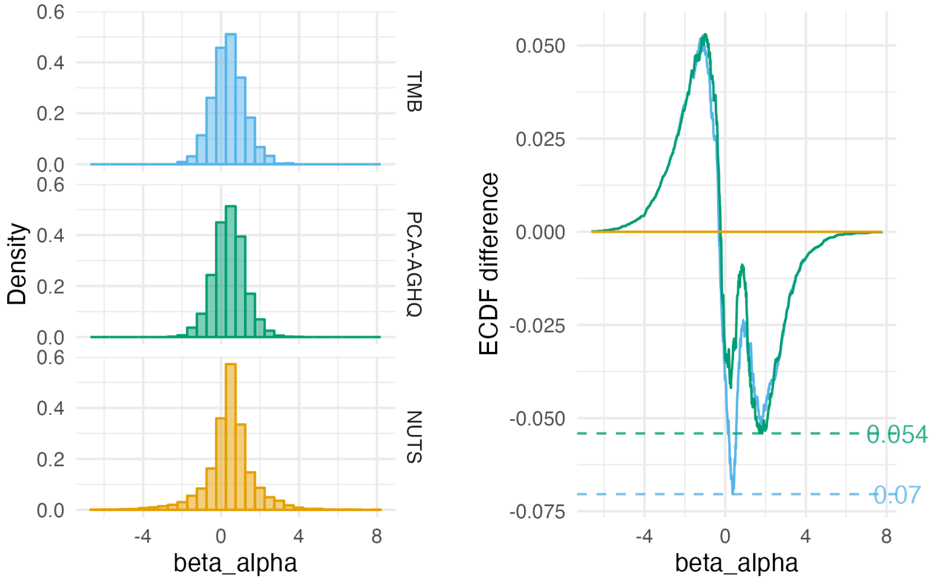

FIGURE 6. *Example KS test for one parameter.*

has been met, that is $\mathbb{P}(\alpha_i > 0.9^2 = 0.81)$ for each strata i. We found that for women both TMB and PCA-AGHQ underestimate these exceedance probabilities (Figure 7, first row). We hypothesise this discrepancy in accuracy by sex is related to interactions between the household survey and ANC components of the model creating a more challenging posterior geometry.

5.5.2. *Finding strata with high incidence.* Some HIV interventions are cost-effective only within high HIV incidence settings, typically defined as higher 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 7, second row). This is surprising, in that we expect inferences from NUTS to be more heavy-tailed than those from TMB or PCA-AGHQ.

6. Discussion. We developed an approximate Bayesian inference algorithm, combining AGHQ with PCA, motivated by a challenging problem in small-area estimation of HIV. For the simplified Naomi model in Malawi (Section 5) we demonstrated the method to be more accurate for model parameters, across a broad range of metrics, than TMB, and substantially faster than NUTS. However, this improvement in accuracy for model parameters did not translate into model outputs. Indeed, we found posterior exceedance probabilities (Section 5.5) from both TMB and PCA-AGHQ to have systematically inaccurates, with the potential to meaningfully mislead policy.

PCA-AGHQ could be added to the Naomi web interface as an alternative to TMB. Analysts may then quickly iterate over model options using a fast inference approach, before switching to a more accurate approach once they are happy with the results. By selecting s and k, PCA-AGHQ can be adjusted to suit the computational budget available. We selected s based on the Scree plot, and for the most part fixed $k = 3$. Whether it is preferable, for a given computational budget, to increase s or increase k remains an open question. Further strategies, such as gradually lowering k over the principal components, could also be considered.

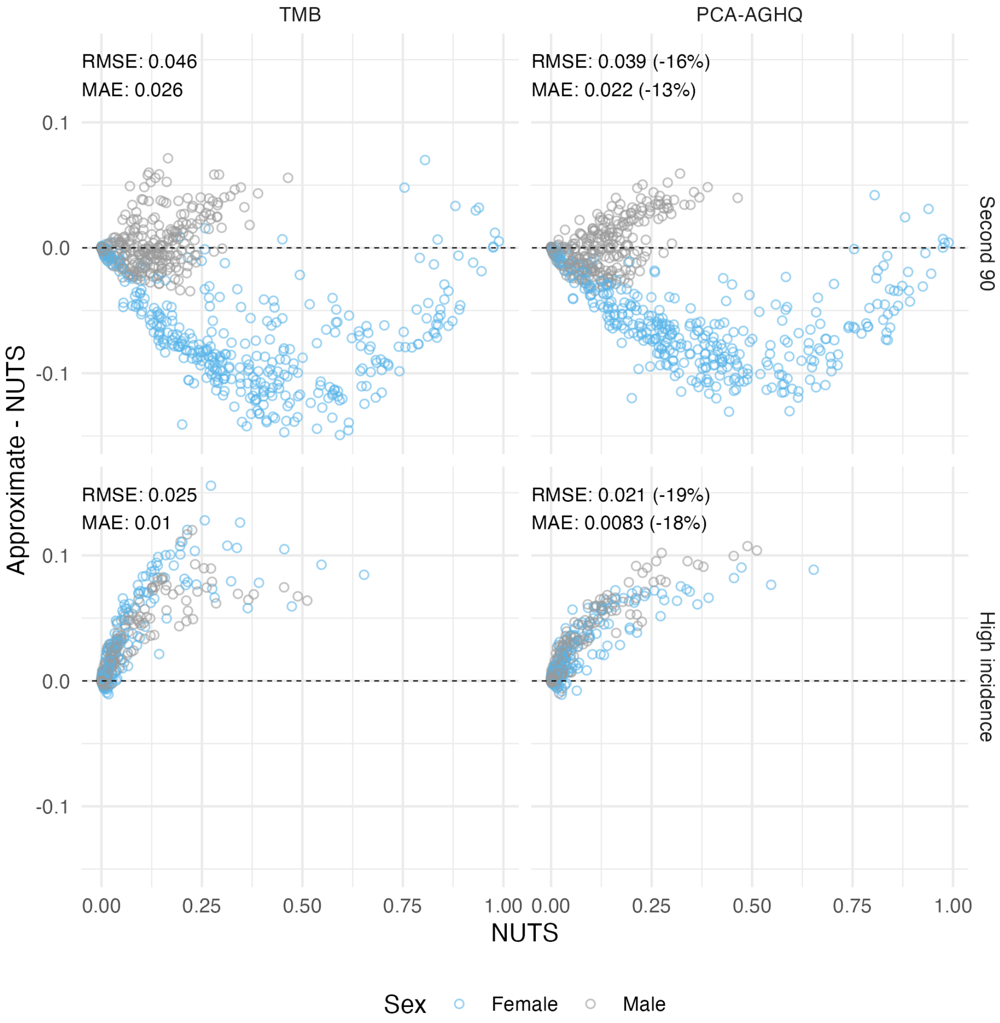

FIGURE 7. *Though PCA-AGHQ was marginally better, both approximate inference methods were meaningfully inaccurate as compared with NUTS for estimating exceedance probabilities. For the second 90 target the inaccuracy varied substantially by sex.*

Bilodeau, Stringer and Tang (2022) highlight developing computationally feasible quadrature methods for high dimensions as a challenging open problem.

PCA-AGHQ was implemented using the `TMB` and `aghq` R packages.

We hope that our work further encourages use of deterministic inference algorithms for ELGMs in applied settings, as well as methodological exploration of their accuracy and limitations. Among the ELGM-type structures of particular interest in spatial epidemiology are aggregated Gaussian process models (Nandi et al., 2020) and evidence synthesis models (Amoah, Diggle and Giorgi, 2020).

6.1. *Future directions.*

6.1.1. *Improving the quadrature grid.* We aimed to develop a quadrature grid which allocates more effort to more important dimensions. While PCA is a sensible approach, there are avenues where it does not behave as one might hope, or otherwise overlooks potential benefits. The first challenge we identified was using PCA when the dimensions have different scales. Specifically, we found logit-scale hyperparamters to be systematically favoured over those on the log-scale. Second, the amount of variation explained for the Hessian matrix is not of directly interest, rather the effect of the different dimensions on the relevant outputs. Using measures of importance from sensitivity analysis, such as Shapley values (Shapley et al., 1953) may be preferable. Third, it is more important to allocate quadrature nodes to those marginals which are non-Gaussian. This is because the Laplace approximation is exact when the integrand is Gaussian, so a single quadrature node is sufficiently. The difficulty is, of course, knowing in advance which marginals will be non-Gaussian. This could be done if there were a cheap way to obtain posterior means, which could then be compared to posterior modes obtained using optimisation. Another approach would be to measure the fit of marginal samples from a cheap approximation, like TMB. The main challenge is that the measurements have to for marginals, ruling out approaches like PSIS which operate on joint distributions (Yao et al., 2018).

6.1.2. *Computational speed-ups.* Integration over a moderate number of hyperparameters posed a challenge, and led us to use a quadrature grids with a large number of nodes. However, computation at each node is independent, such that the run-time of the algorithm could potentially be significantly improved by parallel computing. Further computational speed-ups might be obtained using graphics processing units (GPUs) speciallised for the relevant matrix operations.

6.1.3. *Comparison to other MCMC algorithms.* Blocked Gibbs sampling (Geman and Geman, 1984) or slice sampling (Neal, 2003), may be better suited than NUTS to sampling from Naomi. These algorithms are available, and customisable, including e.g. choice of block structure within the `NIMBLE` probabilistic programming language (de Valpine et al., 2017).

6.1.4. *Implementation into probabilistic programming languages.* Though gaining in popularity, the user-base of `TMB` remains relatively small. Furthermore, for users unfamiliar with C++, it can be challenging to use. As such, it could be beneficial to implement AGHQ within other probabilistic programming languages. Implementation in `NIMBLE` could be relatively straightforward, as it (for version >1.0.0) includes functionality for automatic differentiation and Laplace approximation, built using `CppAD` like `TMB`. Similarly, implementation in Stan could be possible by use of the `bridgestan` package (Ward, 2023) together with the adjoint-differentiated Laplace approximation of Margossian et al. (2020).

6.1.5. *Statistical theory.* Stringer, Brown and Stafford (2022) (Theorem 1) bound the total variation error of AGHQ, establishing convergence in probability of coverage probabilities under the approximate posterior to those under the true posterior. It's possible that similar theory could be established for PCA-AGHQ, or more generally AGHQ with varying numbers of nodes per dimension.

6.1.6. *Laplace marginals.* See Appendix S6.

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

REFERENCES

Allaire, J., Xie, Y., Dervieux, C., R Foundation, Wickham, H., Journal of Statistical Software, Vaidyanathan, R., Association for Computing Machinery, Boettiger, C., Elsevier, Broman, K., Mueller, K., Quast, B., Pruim, R., Marwick, B., Wickham, C., Keyes, O., Yu, M., Emaasit, D., Onkelinx, T., Gasparini, A., Desautels, M.-A., Leutnant, D., MDPI, Taylor and Francis, Öğreden, O., Hance, D., Nüst, D., Uvesten, P., Campitelli, E., Muschelli, J., Hayes, A., Kamvar, Z. N., Ross, N., Cannoodt, R., Luguern, D., Kaplan, D. M., Kreutzer, S., Wang, S., Hesselberth, J. and Hyndman, R. (2022a). rticles: Article Formats for R Markdown R package version 0.23.6.

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W. and Iannone, R. (2022b). rmarkdown: Dynamic Documents for R R package version 2.14.

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

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

Baydin, A. G., Pearlmutter, B. A., Radul, A. A. and Siskind, J. M. (2017). Automatic differentiation in machine learning: a survey. *The Journal of Machine Learning Research* **18** 5595–5637.

Berild, M. O., Martino, S., Gómez-Rubio, V. and Rue, H. (2022). Importance sampling with the integrated nested Laplace approximation. *Journal of Computational and Graphical Statistics* **31** 1225–1237.

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

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M. and Bolker, B. M. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal* **9** 378–400.

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.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T. and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* **26** 403–413.

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.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* **6** 721–741.

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

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*.

KARATZOGLOU, A., SMOLA, A., HORNIK, K. and KARATZOGLOU, M. A. (2019). Package 'kernlab'. *CRAN R Project*.

KISH, L. (1965). Survey sampling.

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

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

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

MENG, X.-L. (2018). Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics* **12** 685–726.

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.

NOOR, A. M. (2022). Country ownership in global health. *PLOS Global Public Health* **2** e0000113.

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.

SHAPLEY, L. S. et al. (1953). A value for n-person games.

SIMPSON, D., RUE, H., RIEBLER, A., MARTINS, T. G. and SØRBYE, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science* **32** 1–28.

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.

WAKEFIELD, J., OKONEK, T. and PEDERSEN, J. (2020). Small area estimation for disease prevalence mapping. *International Statistical Review* **88** 398–418.

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.