Submitted to the Annals of Applied Statistics

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 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, enabling fast and accurate inference for Naomi and other extended latent Gaussian models. Using data from Malawi, our method improves the accuracy of inferences across a range of metrics, while being substantially faster to run than Hamiltonian Monte Carlo with the No-U-Turn sampler. Our implementation uses the `aghq` R package, facilitating easy, flexible use of the method when provided a `TMB` C++ template for the model's log-posterior.

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 only conducted every five years or so in most countries, with limited sample size at the district level, due to their high costs to run. 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 mitigates the limitations of any single source, increases statistical power, and can prompt investigation into any conflicts of information.

Software (`https://naomi.unaids.org`) has been developed for Naomi, allowing 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 response.

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

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.

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 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, the inference procedure should be fast and low memory usage enough for workshop participants to fit the model. 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, it is preferable that use of the inference method across countries is automatic, and doesn't require 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). Owing to its speed and flexibility, `TMB` is gaining popularity, particularly in spatial statistics (Osgood-Zimmerman and Wakefield, 2022), and via the `glmmTMB` R package (Brooks et al., 2017) which provides a user-friendly formula interface for fitting common models. 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. `TMB` uses automatic differentiation (Baydin et al., 2017) to calculate the derivatives required for numerical optimisation routines and the Laplace approximation, Taking inspiration from the AD Model Builder package (Fournier et al., 2012).

Although the approach of `TMB` is fast, within the empirical Bayes framework hyperparameter uncertainty is not properly accounted for in the latent field posterior. This consideration 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 developed an inference method which extends adaptive Gauss-Hermite quadrature (AGHQ) to handle many 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. 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 Hessian at the mode to find a smaller number of dimensions which explain most of the variance. In our application, this results in a grid which has millions of times fewer nodes the corresponding dense grid. Our method is implemented as an extension of the `aghq` R package (Stringer, 2021). As `aghq` is designed to naturally interface with `TMB`, use is simple when provided a C++ user template for the log-posterior.

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

The remainder of this paper is organised as follows. Section 2 outlines the version of the Naomi model that we consider in this paper, and Section 3 describes how it falls within the ELGM framework. Section 4 outlines our approach to fast, accurate Bayesian inference for ELGMs using simplified INLA and AGHQ. As a case study, we compare the accuracy of our inference method to `TMB` and `tmbstan` for the simplified Naomi model fit to data from Malawi, in Section 5. 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. 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.

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.

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, with individual responses $\theta_j \in \{0, 1\}$ and design weights w_j. The index j is across all individuals in strata $i \in I$ within the relevant denominator i.e. for ART coverage, only those individuals who are HIV positive. The observed number of outcomes are then $y_I^\theta = m_I^\theta \cdot \hat{\theta}_I$ where

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

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

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

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

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

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. Of the number of ANC clients with ascertained status $m_I^{\rho^{\text{ANC}}}$, we 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 above, 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 regressions to model the probabilities of accessing services outside the home district. 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, 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. 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. *Collected together.* Let $\mathbf{y} = (y_I^\theta)$ for $\theta \in \{\rho, \alpha, \kappa, \rho^{\text{ANC}}, \alpha^{\text{ANC}}, N^{\text{ART}}\}$ and $I \in \mathcal{I}$ be a concatenated vector of observations. Here, we could attempt to write the Naomi model fully in a small number of equations, but likely it'd be difficult.

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

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

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,

If Naomi was written out in full mathematical detail (in the section above), it would be easier to make these reasons more concrete. We should also specify how much of an impact each of these reasons is likely to have on the difficulty of inference. Perhaps a table format could be good here.

4. Fast approximate inference method. The joint posterior of the parameters $(\mathbf{x}, \boldsymbol{\theta})$ given data $\mathbf{y}$ 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

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

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

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 $\tilde{p}_{\text{G}}(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}(\mathbf{x} \,|\, \hat{\mathbf{x}}(\boldsymbol{\theta}), \hat{\mathbf{H}}(\boldsymbol{\theta})^{-1})$ be a Gaussian approximation to $p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})$ 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

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

Inference in TMB 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. Latent field joint and marginal inferences then follow directly from the Gaussian approximation $\tilde{p}_{\text{G}}(\mathbf{x} \,|\, \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y})$. Hyperparameter inferences are obtained according to some method which should be specified here as well.

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

$$p(\mathbf{y}) \approx \int_{\boldsymbol{\theta}} \tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y}) \mathrm{d}\boldsymbol{\theta} \approx \sum_{\mathbf{z} \in \mathcal{Q}} p_{\text{LA}}(\mathbf{z}, \mathbf{y}) \omega(\mathbf{z}), \tag{4.6}$$

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 one such 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 $\omega(z) = k!/[H_{k+1}(z)]^2\phi(z)$, where $\phi(\cdot)$ is a standard univariate Gaussian density. 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. Posterior distributions can typically be expected to be relatively well described by this class of functions. Multivariate GHQ rules, required to integrate over $\boldsymbol{\theta}$, 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 $\omega(\mathbf{z}) = \prod_{j=1}^m \omega(z_j)$ (Jäckel, 2005).

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 especially important in statistical quadrature problems, where the integral depends on data $\mathbf{y}$ such that regions of high density are not known in advance. To obtain an AGHQ estimate of Equation 4.6, 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, then

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

That is, the unadapted nodes have been shifted by the mode and rotated by a decomposition of the inverse curvature such that $\mathbf{z} \mapsto \hat{\mathbf{P}}_{\text{LA}}\mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}}$. Two alternatives for the 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). Figure 1 demonstrates GHQ, as well as adaption for these two choices of matrix decomposition. Equation 4.7 may then be used to normalise the Laplace approximation

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

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

$$\tilde{p}(\mathbf{x} \,|\, \mathbf{y}) = |\hat{\mathbf{P}}_{\text{LA}}| \sum_{\mathbf{z}\in\mathcal{Q}(m,k)} \tilde{p}_{\text{G}}(\mathbf{x} \,|\, \hat{\mathbf{P}}_{\text{LA}}\mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y})\tilde{p}_{\text{LA}}(\hat{\mathbf{P}}_{\text{LA}}\mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}} \,|\, \mathbf{y})\omega(\mathbf{z}). \tag{4.8}$$

Samples from this mixture of Gaussians may be obtained by drawing a node $\mathbf{z}$ with multinomial probabilities $\lambda(\mathbf{z}) = |\hat{\mathbf{P}}_{\text{LA}}| p_{\text{LA}}(\hat{\mathbf{P}}_{\text{LA}}\mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}} \,|\, \mathbf{y})\omega(\mathbf{z})$, 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. 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 $\mathcal{Q}(m,\mathbf{k}) = \mathcal{Q}(1,k_1) \times \cdots \times \mathcal{Q}(1,k_m)$ 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 $k_j = k, j \le s$ and $k_j = 1, j > s$ for some $s \le m$. Taken together with use of the spectral decomposition, this choice of levels is analogous to a principal components analysis (PCA) approach to AGHQ

$$\tilde{p}_{\text{PCA}}(\mathbf{y}) = |\hat{\mathbf{E}}_{\text{LA},s}\hat{\boldsymbol{\Lambda}}_{\text{LA},s}^{1/2}| \sum_{\mathbf{z}\in\mathcal{Q}(m,s,k)} \tilde{p}_{\text{LA}}(\hat{\mathbf{E}}_{\text{LA},s}\hat{\boldsymbol{\Lambda}}_{\text{LA},s}^{1/2}\mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y})\omega(\mathbf{z}), \tag{4.9}$$

where $\hat{\mathbf{E}}_{\text{LA},s} = \cdots$ and $\hat{\boldsymbol{\Lambda}}_{\text{LA},s} = \cdots$ are terms which should be described. The final panel of Figure 1 illustrates a case where $m = 2$ and $s = 1$. We refer to use of this quadrature rule as PCA-AGHQ.

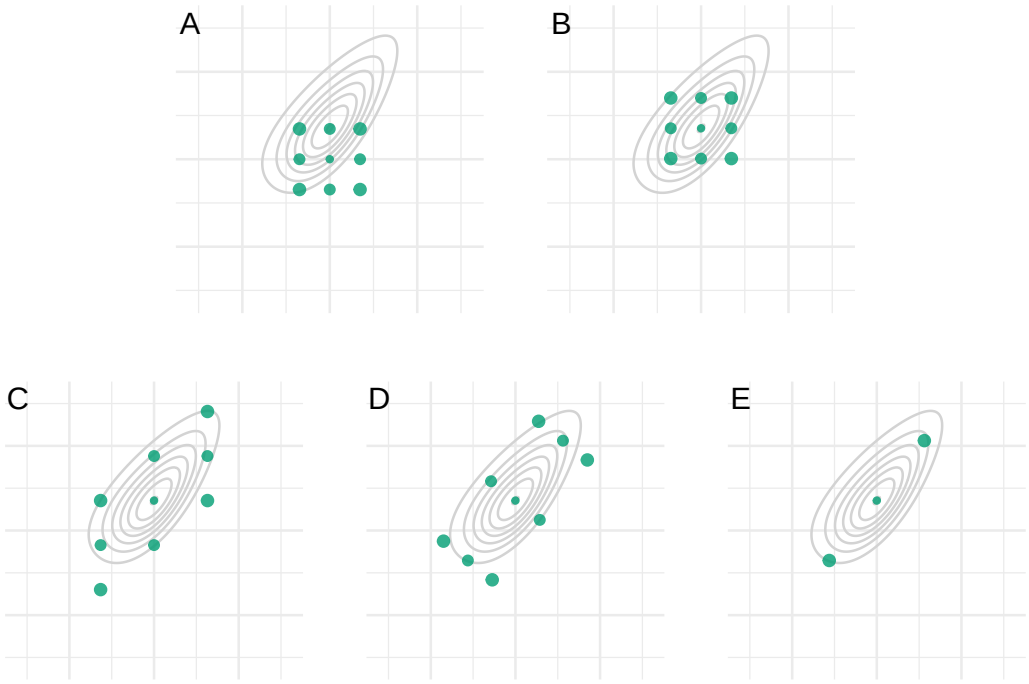

FIGURE 1. *The Gauss-Hermite quadrature nodes* $\mathbf{z} \in \mathcal{Q}(2,3)$ *for a two dimensional integral with three nodes per dimension (A). Adaption occurs based on the mode (B) and covariance matrix of the target via the Cholesky decomposition (C) or spectral decompostion (D) of the inverse curvature at the mode. In PCA-AGHQ (E) only nodes along the first* s *principal components are kept. The integrand is* $f(\boldsymbol{\theta}) = sn(0.3\theta_1, \alpha = 2) \cdot sn(0.5\theta_1 - 0.3\theta_2, \alpha = -2)$*, where* $sn(\cdot)$ *is the standard skewnormal probability density function with shape parameter* $\alpha \in \mathbb{R}$.

4.5. *Sparse rules.* Though there are sparse rules which retain the attractive exactness properties of GHQ, using fewer than k^m quadrature points, they use weighting functions which may be negative $\omega(\mathbf{z}) < 0$. This introduces a problem when trying to produce inferences about the latent field, as the resulting $\lambda(\mathbf{z})$ may be negative. We are not aware of any suitable approaches to obtain multinomial samples in such cases. It is however, still possible to calculate the normalising constant using a sparse rule, which we do so in Appendix S3, finding it to be related to that from PCA-AGHQ as follows.

5. Application to data from Malawi. We fit the simplified Naomi model (Section 2) to data from Malawi using three inferential approaches. 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 `TMB` C++ user-template used to specify the log-posterior was the same for each. 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/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).

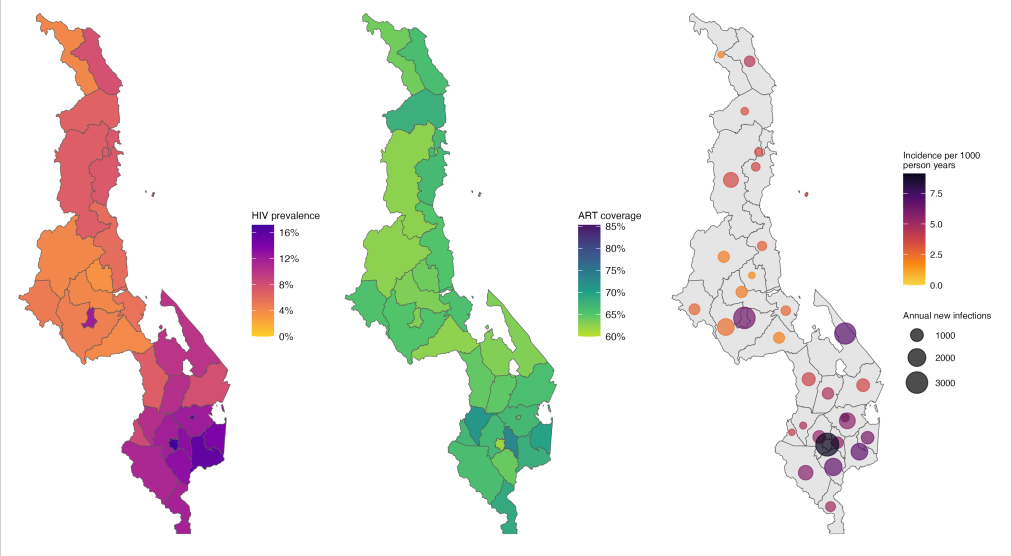

FIGURE 2. *District-level model outputs for adults aged 15-49. Inference conducted with TMB.*

Name	Software	Details
TMB	`TMB`	1000 samples
PCA-AGHQ	`aghq`	$k = 3, s = 8$ (see Section 5.2), 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.

5.1. *NUTS convergence.* To obtain satisfactory NUTS results we increased the Markov chain lengths until all diagnostics were acceptable. This required chains of length 20,000, thinned by a factor of 20 for ease-of-storage. All potential scale reduction factors (Gelman and Rubin, 1992; Vehtari et al., 2021) were then $\hat{R} < 1.05$. For full details see Appendix S2. Though inaccuracies remain possible, we considered the NUTS results to be a gold-standard.

5.2. *PCA-AGHQ settings* . We used a Scree plot (Figure SX) based on the spectral decomposition of $\hat{\mathbf{H}}_{\text{LA}}(\hat{\boldsymbol{\theta}}_{\text{LA}})$ to select the number of principal components $s = 8 < m = 24$ to keep, sufficient to explain close to 90% of total variation. With this choice of s, the reduced rank approximation to the Hessian is visually very similar (Figure SX).

5.2.1. *Visual inspection.* Overlaying the generated $3^8 = 6561$ 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 is an improvment on the 8 that would be naively achieved using a dense grid, there remained many hyperparameters poorly covered. Coverage was associated with marginal standard deviation (Figure SX), which varied particularly according to the hyperparameter scale. 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.

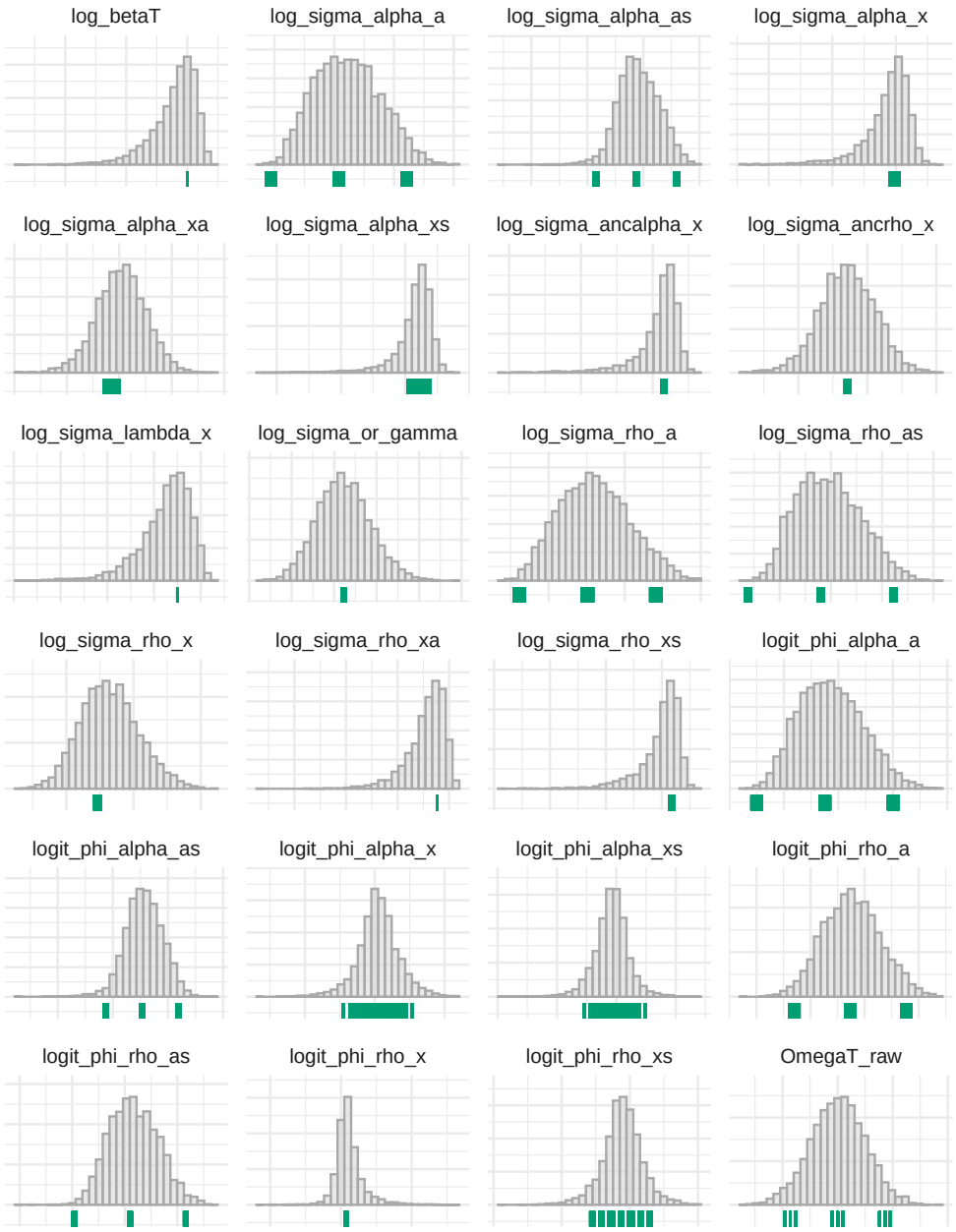

FIGURE 3. *The 6561 PCA-AGHQ node positions (green, rug plot) projected onto the hyperparameter marginal posteriors (grey, histogram) for each of the 24 hyperparameters. Some hyperparameters, such as logit_phi_alpha_x, are well covered where as others, such as log_sigma_lambda_x, are near only being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

5.2.2. *Normalising constant assessment.* We assessed appropriateness of the grid by comparing the estimate of $\log p_{\text{PCA}}(\mathbf{y})$ for a range of settings. Convergence in $\log p_{\text{PCA}}(\mathbf{y})$ as s and k are increased may suggest a suitable grid has been reached. Appendix S3 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.* Let ϕ be a generic model parameter. To assess the informativeness of the data we compared the prior variance $\sigma^2_{\text{prior}}(\phi)$ to the posterior variance $\sigma^2_{\text{posterior}}(\phi)$ via the posterior contraction $c(\phi) = 1 - (\sigma^2_{\text{posterior}}(\phi)/\sigma^2_{\text{prior}}(\phi))$ (Schad, Betancourt and Vasishth, 2021). We found that (Figure 4)) something something. For greater interpretability, facet parameters in this plot according to model component.

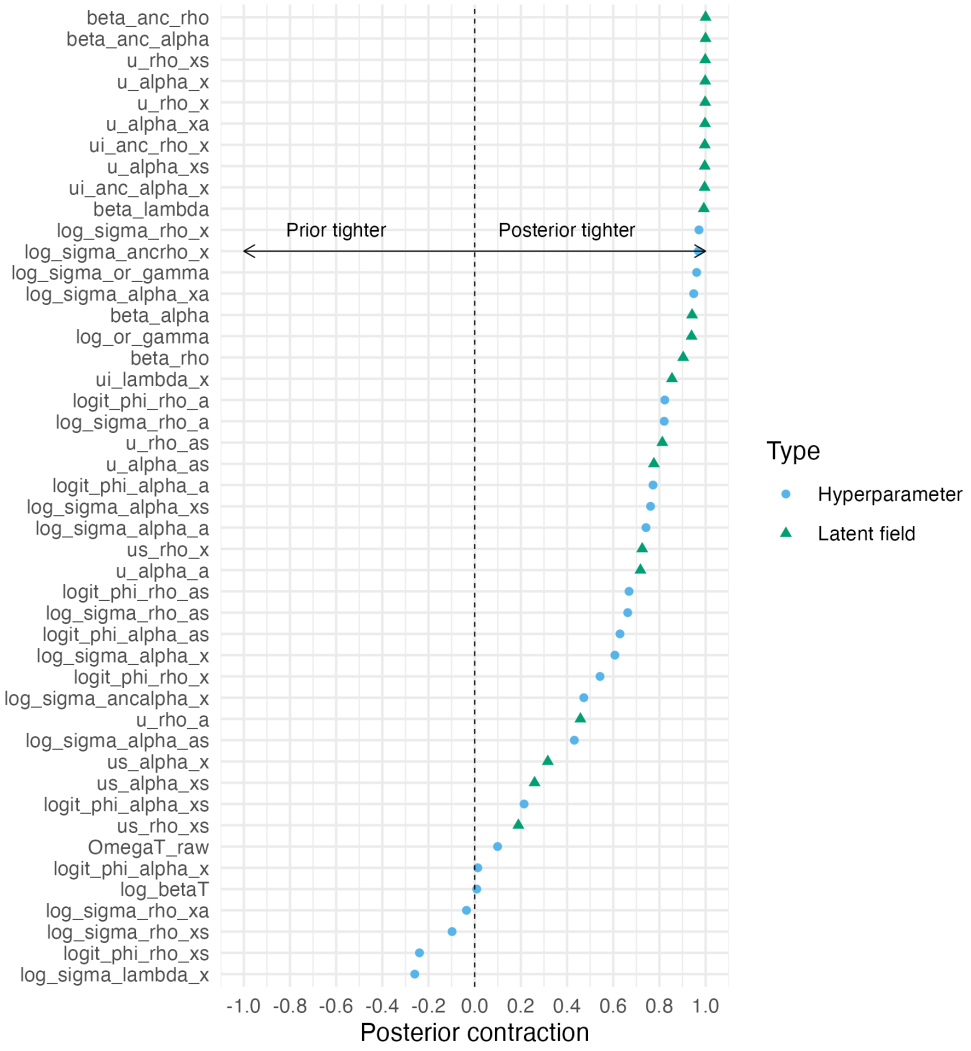

FIGURE 4. *A posterior contraction tending toward 1.0 corresponds to a posterior tending towards a Dirac delta function. A posterior contraction of 0.0 corresponds to no change in standard deviation between prior and posterior. A posterior contraction less than 0.0 corresponds to a wider posterior than prior.*

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* . To compare the accuracy of posterior distributions produced by TMB and PCA-AGHQ as compared with those from NUTS we assessed (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.* The root mean square error (RMSE) between posterior mean estimates from PCA-AGHQ and NUTS (0.063) was 19% lower than that between TMB and NUTS (0.077). For the posterior standard deviation estimates, there was a substantial 70% reduction in RMSE: from 0.15 (TMB) to 0.05 (PCA-AGHQ). These results, alongside those for the mean absolute error (MAE), are presented in Figure 5.

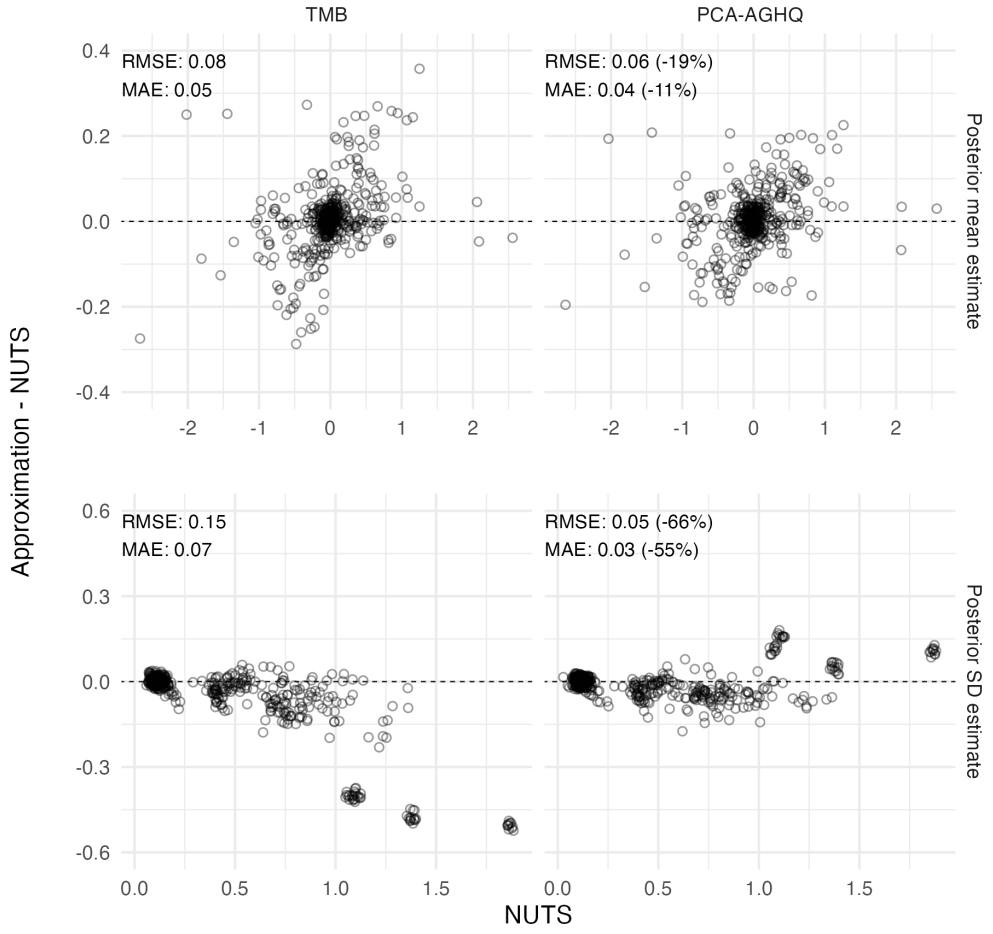

FIGURE 5. *PCA-AGHQ modestly improves estimation of the posterior mean, and substantially improves estimation of the posterior standard deviation, as compared with TMB.*

5.4.2. *Distribution tests.* The two-sample Kolmogorov-Smirnov (KS) test statistic (Smirnov, 1948) is the maximum absolute difference between two ECDFs $F(\varphi) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\phi_i \leq \varphi}$. See an illustration of the KS test in Figure 6 and a summary of the results in Table.

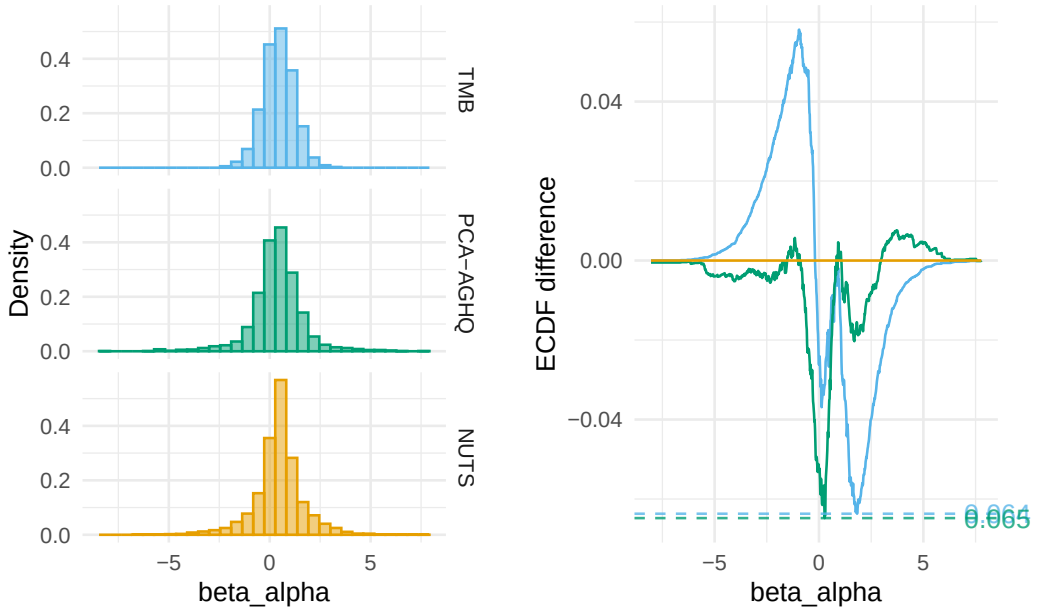

FIGURE 6. *Example KS test for one parameter.*

5.4.3. *Pareto-smoothed importance sampling.* Let $\{\boldsymbol{\phi}_i\}_{i=1}^n$ 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.

5.4.4. *Maximum mean discrepancy.* Let $\Phi = \{\boldsymbol{\phi}_i\}_{i=1}^n$ and $\Psi = \{\boldsymbol{\psi}_i\}_{i=1}^n$ 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

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

We set $k(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j) = \exp(-\sigma\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_j\|^2)$ 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 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.

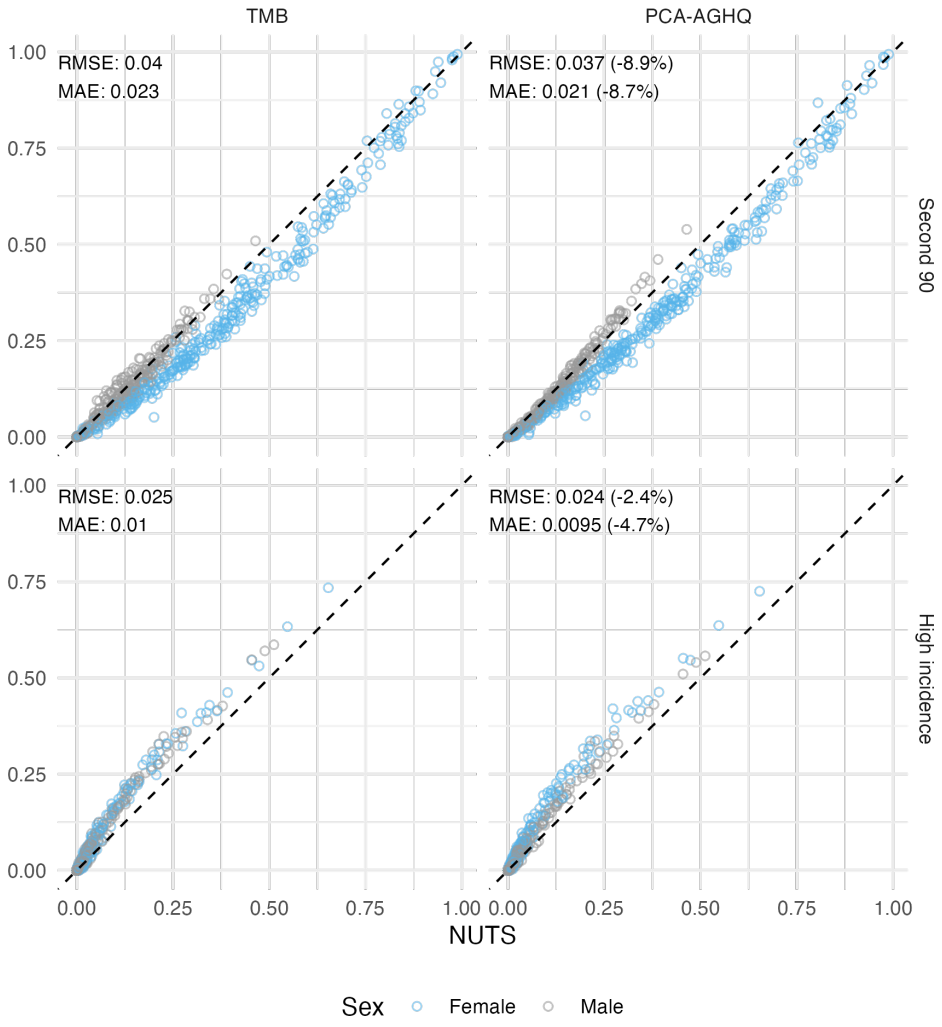

FIGURE 7. *Though PCA-AGHQ does perform slightly better, both approximate inference methods are meaningfully inaccurate as compared with NUTS for estimating exceedance probabilities. For the second 90 target the inaccuracy varies substantially by sex.*

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, across a broad range of metrics, than TMB, and substantially faster than NUTS.

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.

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

We found that posterior exceedance probabilities (Section 5.5) from the approximate methods to be systematically inaccurate, with the potential to meaningfully mislead policy.

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 S4.

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.

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.

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

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.