Appendix to "Fast approximate Bayesian inference of HIV indicators using PCA adaptive Gauss-Hermite quadrature"

Adam Howes* Alex Stringer† Seth R. Flaxman‡ Jeffrey W. Eaton§

Contents

*Department of Mathematics, Imperial College London
†Department of Statistics and Actuarial Science, University of Waterloo
‡Department of Computer Science, Oxford University
§Harvard T.H. Chan School of Public Health, Harvard University

S1 Simplified Naomi model description

Here, we describe the simplified Naomi model (Eaton et al. 2021) in more detail.

S1.1 Background

S1.1.1 Indexing

Consider the most recent national household survey with HIV testing which has taken place in the country of interest. Let $x \in \{1, \ldots, n\}$ refer to a district located within the Spectrum (Stover et al. 2010) region R_x. Let $s \in \{\text{F}, \text{M}\}$ be sex, and $a \in \{\text{0-5}, \text{5-10}, \ldots, \text{75-80}, \text{80+}\}$ be five-year age groups. As short-hand, we write $a = l$ to refer to the age group with lower bound l, e.g. $a = 20$ for $a = \text{20-25}$. We index the known quantity population size $N_{x,s,a}$ by district, sex and age-band, as well as the following unknown quantities: HIV prevalence $\rho_{x,s,a} \in [0, 1]$, ART coverage $\alpha_{x,s,a} \in [0, 1]$, annual HIV incidence rate $\lambda_{x,s,a} > 0$, and proportion of HIV positive persons recently infected $\kappa_{x,s,a} \in [0, 1]$. Sometimes data are observed at an aggregate level, rather than the more granular modelled level. We then use $\{\cdot\}$ to generically refer to a aggregate set over which an observation is made, e.g. $\{a\} = \{\text{15-19}, \ldots, \text{45-49}\}$ for the adult age range 15-49. Let $\sum_{\{x\}}$ be shorthand for $\sum_{x \in \{x\}}$, analogously for s and a.

S1.1.2 Structured random effects

We use structured random effects to enable partial pooling of information across units assessed as being similar, such as those belonging to neighbouring districts or adjacent age-bands. Let u be a generic random effect, with length $\dim(u)$. Three types of structured random effects are used in the model. First, we specify the first order auto-regressive model by $u \sim \text{AR1}(\sigma, \phi)$ such that

$$
\begin{aligned}
u_1 &\sim \left(0, \frac{1}{1-\rho^2}\right), \\
u_i &= \rho u_{i-1} + \epsilon_i, \quad i = 2, \ldots, \dim(u)
\end{aligned}
$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$ are independent and identically distributed (IID) Gaussian random variables, $\sigma > 0$ is the marginal standard deviation, and $|\rho| < 1$ is the (lag-one) correlation parameter. Second, we use $u \sim \text{ICAR}(\sigma)$ to refer to the Besag intrinsic conditional auto-regressive model (ICAR) (Besag, York, and Mollié 1991) with full conditionals

$$
u_i \,|\, u_{-i} \sim \mathcal{N}\left(\frac{\sum_{j:j\sim i} u_j}{n_{\delta i}}, \frac{\sigma^2}{n_{\delta i}}\right),
$$

where u_{-i} is u with the ith element removed i.e. $(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_{\dim(u)})$, $j \sim i$ if the units i and j are defined as adjacent, $n_{\delta i} = |\{j : j \sim i\}|$ is the total number of adjacent units, and $\sigma > 0$ is the marginal standard deviation. We follow recommendations of Freni-Sterrantino, Ventrucci, and Rue (2018) on scaling of precision matrices, disconnected adjacency graph components, and islands. Third, for the reparameterised Besag-York-Mollie model (BYM2) (Simpson et al. 2017), we write $u \sim \text{BYM2}(\sigma, \phi)$, where u is comprised of a spatially structured ICAR component v with proportion $\phi \in (0, 1)$ and spatially unstructured IID component w with proportion $1 - \phi$, both scaled to have generalised variance equal to one, and $\sigma > 0$ is the marginal standard deviation such that

$$
u = \sigma\left(\sqrt{\phi} \cdot v + \sqrt{1-\phi} \cdot w\right).
$$

S1.1.3 Complex survey design

We assume the household survey was run according to a complex survey design where each individual j in the population U has non-zero probability $\pi_j \in (0, 1)$ of appearing in the sample $S \subseteq U$. Suppose we observe an outcome $\theta_j \in \{0, 1\}$ for $j \in S$. Let $w_j = 1/\pi_j \times 1/\omega_j$ be design weights, where ω_j is a non-response factor, then a survey weighted mean, calculated using the `survey` R package (Lumley 2004), is given by

$$
\hat{\theta} = \frac{\sum_{j \in S} w_j \theta_j}{\sum_{j \in S} w_j}.
$$

We account for survey weighting in the variance via the Kish effective sample size (Kish 1965)

$$m^{\theta} = \frac{\left(\sum_{j \in S} w_j\right)^2}{\sum_{j \in S} w_j^2}.$$

The observed number of indicator cases is then $y^{\theta} = m^{\theta} \cdot \hat{\theta}$.

S1.2 Process specification

	Model component	Latent field	Hyperparameter
S1.2.1	HIV prevalence	$22 + 5n$	9
S1.2.2	ART coverage	$25 + 5n$	9
S1.2.3	HIV incidence rate	$2 + n$	3
S1.2.4	ANC testing	$2 + 2n$	2
S1.2.5	ART attendance	n	1
	Total	$51 + 14n$	24

Table S1: The numer of latent field parameters and hyperparameters in each model section.

We now describe the hyperparameter and latent field process specification for the model. Whereas in the main text process and likelihood specifications are written together, here we consider the likelihood equations separately in Section S1.3 to follow. Table S1 gives the number of latent field parameters and hyperparameters in each component of the model. In the case of Malawi then $n = 32$ such that the total number of latent field parameters is $51 + 14 \cdot 32 = 499$ and the total number of hyperparameters is 24.

S1.2.1 HIV prevalence

We model HIV prevalence $\rho_{x,s,a} \in [0, 1]$ on the logit scale using the linear predictor

$$\text{logit}(\rho_{x,s,a}) = \beta_0^{\rho} + \beta_S^{\rho,s=\text{M}} + u_a^{\rho} + u_a^{\rho,s=\text{M}} + u_x^{\rho} + u_x^{\rho,s=\text{M}} + u_x^{\rho,a<15} + \eta_{R_x,s,a}^{\rho}. \quad (1)$$

Table S2 provides a description of the terms included in Equation 1.

Term	Distribution	Description
β_0^{ρ}	$\mathcal{N}(0, 5)$	Intercept
$\beta_s^{\rho,s=\text{M}}$	$\mathcal{N}(0, 5)$	The difference in logit prevalence for men compared to women
u_a^{ρ}	$\text{AR1}(\sigma_A^{\rho}, \phi_A^{\rho})$	Age random effects for women
$u_a^{\rho,s=\text{M}}$	$\text{AR1}(\sigma_{AS}^{\rho}, \phi_{AS}^{\rho})$	Age random effects for the difference in logit prevalence for men compared to women age a
u_x^{ρ}	$\text{BYM2}(\sigma_X^{\rho}, \phi_X^{\rho})$	Spatial random effects for women
$u_x^{\rho,s=\text{M}}$	$\text{BYM2}(\sigma_{XS}^{\rho}, \phi_{XS}^{\rho})$	Spatial random effects for the difference in logit prevalence for men compared to women in district x
$u_x^{\rho,a<15}$	$\text{ICAR}(\sigma_{XA}^{\rho})$	Spatial random effects for the ratio of paediatric prevalence to adult women prevalence
$\eta_{R_x,s,a}^{\rho}$	–	Fixed offsets specifying assumed odds ratios for prevalence outside the age ranges for which data are available

Table S2: Terms included in the linear predictor for HIV prevalence (Equation 1).

The two BYM2 random effects u_x^{ρ} and $u_x^{\rho,s=\text{M}}$ are comprised of the following unit scaled spatially structured

$\{v_x^\rho, v_x^{\rho,s=\mathrm{M}}\}$ and spatially unstructured $\{w_x^\rho, w_x^{\rho,s=\mathrm{M}}\}$ components, respectively

$$u_x^\rho = \sigma_X^\rho \left(\sqrt{\phi}_X^\rho \cdot v_x^\rho + \sqrt{1-\phi_X^\rho} \cdot w_x^\rho \right),$$
$$u_x^{\rho,s=\mathrm{M}} = \sigma_{XS}^\rho \left(\sqrt{\phi}_{XS}^\rho \cdot v_x^{\rho,s=\mathrm{M}} + \sqrt{1-\phi_{XS}^\rho} \cdot w_x^{\rho,s=\mathrm{M}} \right).$$

We use half-normal priors for the standard deviation terms

$$\{\sigma_A^\rho, \sigma_{AS}^\rho, \sigma_X^\rho, \sigma_{XS}^\rho, \sigma_{XA}^\rho\} \sim \mathcal{N}^+(0, 2.5),$$

uniform priors for the AR1 correlation parameters

$$\{\phi_A^\rho, \phi_{AS}^\rho\} \sim \mathcal{U}(-1, 1),$$

and beta priors for the BYM2 proportion parameters

$$\{\phi_X^\rho, \phi_{XS}^\rho\} \sim \mathrm{Beta}(0.5, 0.5).$$

S1.2.2 ART coverage

We model ART coverage $\alpha_{x,s,a} \in [0,1]$ on the logit scale using the linear predictor

$$\mathrm{logit}(\alpha_{x,s,a}) = \beta_0^\alpha + \beta_S^{\alpha,s=\mathrm{M}} + u_a^\alpha + u_a^{\alpha,s=\mathrm{M}} + u_x^\alpha + u_x^{\alpha,s=\mathrm{M}} + u_x^{\alpha,a<15} + \eta_{R_x,s,a}^\alpha$$

with terms and priors analogous to the HIV prevalence process model in Section S1.2.1 above.

S1.2.3 HIV incidence rate

We model HIV incidence rate $\lambda_{x,s,a} > 0$ on the log scale using the linear predictor

$$\log(\lambda_{x,s,a}) = \beta_0^\lambda + \beta_S^{\lambda,s=\mathrm{M}} + \log(\rho_x^{\text{15-49}}) + \log(1 - \omega \cdot \alpha_x^{\text{15-49}}) + u_x^\lambda + \eta_{R_x,s,a}^\lambda. \quad (2)$$

Table S3 provides a description of the terms included in Equation 2.

We model the proportion recently infected among HIV positive persons $\kappa_{x,s,a} \in [0,1]$ as

$$\kappa_{x,s,a} = 1 - \exp\left(-\lambda_{x,s,a} \cdot \frac{1-\rho_{x,s,a}}{\rho_{x,s,a}} \cdot (\Omega_T - \beta_T) - \beta_T \right),$$

where $\Omega_T \sim \mathcal{N}(\Omega_{T_0}, \sigma^{\Omega_T})$ is the mean duration of recent infection, and $\beta_T \sim \mathcal{N}^+(\beta_{T_0}, \sigma^{\beta_T})$ is the false recent ratio. We use an informative prior for Ω_T based on the characteristics of the recent infection testing algorithm. For PHIA surveys this is $\Omega_{T_0} = 130$ days and $\sigma^{\Omega_T} = 6.12$ days, and further we assume there is no false recency, such that $\beta_{T_0} = 0.0$ and $\sigma^{\beta_T} = 0.0$.

S1.2.4 ANC testing

HIV prevalence $\rho_{x,a}^{\mathrm{ANC}}$ and ART coverage $\alpha_{x,a}^{\mathrm{ANC}}$ among pregnant women are modelled as being offset on the logit scale from the corresponding district-age indicators $\rho_{x,F,a}$ and $\alpha_{x,F,a}$ according to

$$\mathrm{logit}(\rho_{x,a}^{\mathrm{ANC}}) = \mathrm{logit}(\rho_{x,F,a}) + \beta^{\rho^{\mathrm{ANC}}} + u_x^{\rho^{\mathrm{ANC}}} + \eta_{R_x,a}^{\rho^{\mathrm{ANC}}},$$
$$\mathrm{logit}(\alpha_{x,a}^{\mathrm{ANC}}) = \mathrm{logit}(\alpha_{x,F,a}) + \beta^{\alpha^{\mathrm{ANC}}} + u_x^{\alpha^{\mathrm{ANC}}} + \eta_{R_x,a}^{\alpha^{\mathrm{ANC}}},$$

where, for $\theta \in \{\rho, \alpha\}$, $\beta^{\theta^{\mathrm{ANC}}} \sim \mathcal{N}(0,5)$ are the average differences between population and ANC outcomes, $u_x^{\theta^{\mathrm{ANC}}} \sim \mathcal{N}(0, \sigma_X^{\theta^{\mathrm{ANC}}})$ are IID district random effects with $\sigma_X^{\theta^{\mathrm{ANC}}} \sim \mathcal{N}^+(0,1)$, and $\eta_{R_x,a}^{\theta^{\mathrm{ANC}}}$ for are offsets for the log fertility rate ratios for HIV positive women compared to HIV negative women and for women on ART to HIV positive women not on ART, calculated from Spectrum model outputs for region R_x.

Term	Distribution	Description
β_0^λ	$\mathcal{N}(0,5)$	Intercept term proportional to the average HIV transmission rate for untreated HIV positive adults
$\beta_S^{\lambda,s=\mathrm{M}}$	$\mathcal{N}(0,5)$	The log incidence rate ratio for men compared to women
$\rho_x^{\text{15-49}} = \frac{\sum_{s\in\{\mathrm{F,M}\}}\sum_{a=15}^{45} N_{x,s,a}\cdot\rho_{x,s,a}}{\sum_{s\in\{\mathrm{F,M}\}}\sum_{a=15}^{45} N_{x,s,a}}$	–	The HIV prevalence among adults 15-49 calculated by aggregating age-specific HIV prevalences
$\alpha_x^{\text{15-49}} = \frac{\sum_{s\in\{\mathrm{F,M}\}}\sum_{a=15}^{45} N_{x,s,a}\cdot\rho_{x,s,a}\cdot\alpha_{x,s,a}}{\sum_{s\in\{\mathrm{F,M}\}}\sum_{a=15}^{45} N_{x,s,a}\cdot\rho_{x,s,a}}$	–	The ART coverage among adults 15-49 calculated by aggregating age-specific ART coverages
$\omega = 0.7$	–	Average reduction in HIV transmission rate per increase in population ART coverage fixed based on inputs to the Estimation and Projection Package (EPP) model
u_x^λ	$\mathcal{N}(0,\sigma^\lambda)$	IID spatial random effects with $\sigma^\lambda \sim \mathcal{N}^+(0,1)$
$\eta_{R_x,s,a}^\lambda$	–	Fixed log incidence rate ratios by sex and age group calculated from Spectrum model output

Table S3: Terms included in the linear predictor for HIV incidence (Equation 2). Note that the only source age structure for this part of the model are $\eta_{R_x,s,a}^\lambda$. As Spectrum assumes that there are no new infections in children aged 5-9 or 10-14, or adults aged over 80, the posterior over new infections in these age groups is exactly zero, by definition. We remove these identically zero posteriors from any later inference comparisons.

In the full Naomi model, for adult women 15-49 the number of ANC clients $\Psi_{x,a} > 0$ are modelled as

$$\log(\Psi_{x,a}) = \log(N_{x,\mathrm{F},a}) + \psi_{R_x,a} + \beta^\psi + u_x^\psi,$$

where $N_{x,\mathrm{F},a}$ are the female population sizes, $\psi_{R_x,a}$ are fixed age-sex fertility ratios in Spectrum region R_x, β^ψ are log rate ratios for the number of ANC clients relative to the predicted fertility, and $u_x^\psi \sim \mathcal{N}(0,\sigma^\psi)$ are district random effects. Here we fix $\beta^\psi = u_x^\psi = 0$ such that $\Psi_{x,a}$ are simply constants.

S1.2.5 ART attendance

Let $\gamma_{x,x'} \in [0,1]$ be the probability that a person on ART residing in district x receives ART in district x'. We assume that $\gamma_{x,x'} = 0$ for $x \notin \{x, \mathrm{ne}(x)\}$ such that individuals seek treatment only in their residing district or its neighbours $\mathrm{ne}(x) = \{x' : x' \sim x\}$, where $\sim$ is an adjacency relation, and $\sum_{x'\in\{x,\mathrm{ne}(x)\}} \gamma_{x,x'} = 1$. To model $\gamma_{x,x'}$ for $x \sim x'$ we use a multinomial logistic regression model, based on the log-odds ratios

$$\tilde{\gamma}_{x,x'} = \log\left(\frac{\gamma_{x,x'}}{1-\gamma_{x,x'}}\right) = \tilde{\gamma}_0 + u_x^{\tilde{\gamma}}, \tag{3}$$

where $\tilde{\gamma}_0 = -4$ is a fixed intercept, and $u_x^{\tilde{\gamma}} \sim \mathcal{N}(0,\sigma_X^{\tilde{\gamma}})$ are district random effects with $\sigma_X^{\tilde{\gamma}} \sim \mathcal{N}^+(0,2.5)$. Note that Equation 3 does not depend on x', such that $\gamma_{x,x'}$ is only a function of x. Choice of $\tilde{\gamma}_0 = -4$ implies a prior mean on $\gamma_{x,x'}$ of 1.8%, such that $(100 - 1.8 \times \mathrm{ne}(x))$% of ART clients in district x obtain treatment in their home district, a-priori. We fix $\tilde{\gamma}_{x,x} = 0$ and recover the multinomial probabilities using the softmax

$$\gamma_{x,x'} = \frac{\exp(\tilde{\gamma}_{x,x'})}{\sum_{x^\star\in\{x,\mathrm{ne}(x)\}}\exp(\tilde{\gamma}_{x,x^\star})}.$$

Given the total number of PLHIV on ART $A_{x,s,a} = N_{x,s,a}\cdot\rho_{x,s,a}\cdot\alpha_{x,s,a}$, the number of ART clients who reside in district x and obtain ART in district x' are $A_{x,x',s,a} = A_{x,s,a}\cdot\gamma_{x,x'}$, and the total attending ART

facilities in district x' are

$$\tilde{A}_{x',s,a} = \sum_{x \in \{x', \text{ne}(x')\}} A_{x,x',s,a}.$$

S1.3 Likelihood specification

S1.3.1 Household survey data

For HIV prevalence, ART coverage and recent HIV infections, denoted by $\theta \in \{\rho, \alpha, \kappa\}$, the most recent household survey furnishes weighted observations $\hat{\theta}_{\{x\},\{s\},\{a\}}$ with respective Kish effective sample sizes $m^{\theta}_{\{x\},\{s\},\{a\}} \in \mathbb{R}$, and observed number of cases

$$y^{\theta}_{\{x\},\{s\},\{a\}} = m^{\theta}_{\{x\},\{s\},\{a\}} \cdot \hat{\theta}_{\{x\},\{s\},\{a\}} \in \mathbb{R}.$$

To model these observations, we use following three binomial working likelihoods

$$y^{\rho}_{\{x\},\{s\},\{a\}} \sim \text{xBin}(m^{\rho}_{\{x\},\{s\},\{a\}}, \rho_{\{x\},\{s\},\{a\}}), \quad \rho_{\{x\},\{s\},\{a\}} = \frac{\sum_{\{x\}}\sum_{\{s\}}\sum_{\{a\}} N_{x,s,a} \cdot \rho_{x,s,a}}{\sum_{\{x\}}\sum_{\{s\}}\sum_{\{a\}} N_{x,s,a}},$$

$$y^{\alpha}_{\{x\},\{s\},\{a\}} \sim \text{xBin}(m^{\alpha}_{\{x\},\{s\},\{a\}}, \alpha_{\{x\},\{s\},\{a\}}), \quad \alpha_{\{x\},\{s\},\{a\}} = \frac{\sum_{\{x\}}\sum_{\{s\}}\sum_{\{a\}} N_{x,s,a} \cdot \rho_{x,s,a} \cdot \alpha_{x,s,a}}{\sum_{\{x\}}\sum_{\{s\}}\sum_{\{a\}} N_{x,s,a} \cdot \rho_{x,s,a}},$$

$$y^{\kappa}_{\{x\},\{s\},\{a\}} \sim \text{xBin}(m^{\kappa}_{\{x\},\{s\},\{a\}}, \kappa_{\{x\},\{s\},\{a\}}), \quad \kappa_{\{x\},\{s\},\{a\}} = \frac{\sum_{\{x\}}\sum_{\{s\}}\sum_{\{a\}} N_{x,s,a} \cdot \rho_{x,s,a} \cdot \kappa_{x,s,a}}{\sum_{\{x\}}\sum_{\{s\}}\sum_{\{a\}} N_{x,s,a} \cdot \rho_{x,s,a}}.$$

The generalised binomial $y \sim \text{xBin}(m, p)$ is defined for $y, m \in \mathbb{R}^+$ with $y \leq m$ such that

$$\log p(y) = \log\Gamma(m+1) - \log\Gamma(y+1) - \log\Gamma(m-y+1) + y\log p + (m-y)\log(1-p),$$

where the gamma function Γ is such that $\forall n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$.

S1.3.2 ANC testing data

We include ANC testing data for the year of the most recent survey. Let $W^{\text{ANC}}_{\{x\}}$ be the number of ANC clients, $X^{\text{ANC}}_{\{x\}}$ the number of those with ascertained status, $Y^{\text{ANC}}_{\{x\}}$ the number of those with positive status (either known or tested) and $Z^{\text{ANC}}_{\{x\}}$ the number of ANC clients already on ART prior to first ANC, such that

$$W^{\text{ANC}}_{x} \geq X^{\text{ANC}}_{x} \geq Y^{\text{ANC}}_{x} \geq Z^{\text{ANC}}_{x},$$

for all $x \in \{x\}$. When ANC testing data are only available for part of a given year, we denote $m^{\text{ANC}} \in \{1, \ldots, 12\}$ the number of months of reported data reflected in counts for that year. The observed number of HIV positive and already on ART among ANC clients is modelled by

$$Y^{\text{ANC}}_{\{x\}} \sim \text{Bin}\left(X^{\text{ANC}}_{\{x\}}, \rho^{\text{ANC}}_{\{x\},\{15,\ldots 45\}}\right),$$
$$Z^{\text{ANC}}_{\{x\}} \sim \text{Bin}\left(Y^{\text{ANC}}_{\{x\}}, \alpha^{\text{ANC}}_{\{x\},\{15,\ldots 45\}}\right),$$

where prevalence and ART coverage are aggregated by the number of pregnant women $\Psi_{x,a}$

$$\rho^{\text{ANC}}_{\{x\}\{a\}} = \frac{\sum_{\{x\}}\sum_{\{a\}} \Psi_{x,a} \cdot \rho^{\text{ANC}}_{x,a}}{\sum_{\{x\}}\sum_{\{a\}} \Psi_{x,a}},$$

$$\alpha^{\text{ANC}}_{\{x\}\{a\}} = \frac{\sum_{\{x\}}\sum_{\{a\}} \Psi_{x,a} \cdot \rho^{\text{ANC}}_{x,a} \cdot \alpha^{\text{ANC}}_{x,a}}{\sum_{\{x\}}\sum_{\{a\}} \Psi_{x,a} \cdot \rho^{\text{ANC}}_{x,a}}.$$

S1.3.3 Number receiving ART

Let $\dot{A}_{\{x\},\{s\},\{a\}}$ be data for the number receiving ART

$$\dot{A}_{\{x\},\{s\},\{a\}} = \sum_{\{x\}}\sum_{\{s\}}\sum_{\{a\}}\sum_{x\sim x',x=x'} \dot{A}_{x',x,s,a}.$$

We model the unobserved numbers of ART clients travelling from x' to x as

$$\dot{A}_{x',x,s,a} \sim \text{Bin}(N_{x',s,a}, \pi_{x',x,s,a})$$

where $\pi_{x',x,s,a} = \rho_{x',s,a} \cdot \alpha_{x',s,a} \cdot \gamma_{x',x,s,a}$. This likelihood is approximated using a normal for the sum of binomials by

$$\dot{A}_{\{x\},\{s\},\{a\}} \sim \mathcal{N}(\tilde{A}_{\{x\},\{s\},\{a\}}, \sigma^{\tilde{A}}_{\{x\},\{s\},\{a\}})$$

where the mean is

$$\tilde{A}_{\{x\},\{s\}\{a\}} = \sum_{\{x\}}\sum_{\{s\}}\sum_{\{a\}}\sum_{x\sim x',x=x'} N_{x',s,a} \cdot \pi_{x',x,s,a},$$

and the variance is

$$\left(\sigma^{\tilde{A}}_{\{x\},\{s\}\{a\}}\right)^2 = \sum_{\{x\}}\sum_{\{s\}}\sum_{\{a\}}\sum_{x\sim x',x=x'} N_{x',s,a} \cdot \pi_{x',x,s,a} \cdot (1 - \pi_{x',x,s,a}).$$

S1.4 Identifiability constaints

If data are missing, some parameters are fixed to default values to help with identifiability. In particular:

1. If survey data on HIV prevalence or ART coverage by age and sex are not available then we set $u^{\theta}_{a} = 0$ and $u^{\theta}_{a,s=\text{M}} = 0$ and use the average age-sex pattern of from the Spectrum offset $\eta^{\theta}_{R_x,s,a}$. For the Malawi example considered in the main text HIV prevalence and ART coverage data are not available for those aged 65+. As a result, there are $|\{0\text{-}4, \ldots, 50\text{-}54\}| = 13$ age groups included for the age random effects.
2. If no ART data, either survey or ART programme, are available but data on ART coverage among ANC clients are available, the level of ART coverage is not identifiable, but spatial variation is identifiable. In this instance, overall ART coverage is determined by the Spectrum offset, and only area random effects are estimated such that $\text{logit}\,(\alpha_{x,s,a}) = u^{\alpha}_{x} + \eta^{\alpha}_{R_x,s,a}$.
3. If survey data on recent HIV infection are not included in the model, then $\beta^{\lambda}_{0} = \beta^{\lambda,s=\text{M}}_{S} = u^{\lambda}_{x} = 0$. The sex ratio for HIV incidence is determined by the sex incidence rate ratio from Spectrum in the same years and the incidence rate in all districts is modelled assuming the same average HIV transmission rate for untreated adults, but varies according to district estimates of HIV prevalence and ART coverage.

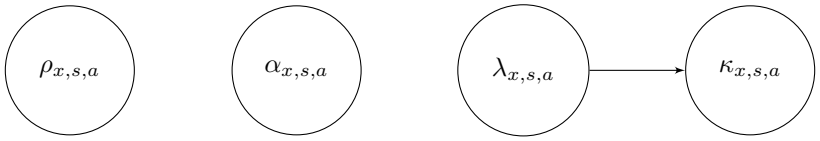

Figure S1: Directed acyclic graph describing the simplified Naomi model (work in progress).

S1.5 Implementation

The C++ `TMB` code for the negative log-posterior of the simplified Naomi model is available from the GitHub repository `athowes/naomi-aghq`. For ease of understanding, Table S4 provides correspondence between the mathematical notation used in Section S1 and the variable names used in the `TMB` code, for all hyperparameters and latent field parameters. For further reference on the `TMB` software see Kristensen (2021).

Variable name	Notation	Type	Size	Domain	ρ input?	α input?	λ input?
`logit_phi_rho_x`	$\text{logit}(\phi^{\rho}_{X})$	Hyper	1	$\mathbb{R}$	✓		
`log_sigma_rho_x`	$\log(\sigma^{\rho}_{X})$	Hyper	1	$\mathbb{R}$	✓		
`logit_phi_rho_xs`	$\text{logit}(\phi^{\rho}_{XS})$	Hyper	1	$\mathbb{R}$	✓		
`log_sigma_rho_xs`	$\log(\sigma^{\rho}_{XS})$	Hyper	1	$\mathbb{R}$	✓		
`logit_phi_rho_a`	$\text{logit}(\phi^{\rho}_{A})$	Hyper	1	$\mathbb{R}$	✓		
`log_sigma_rho_a`	$\log(\sigma^{\rho}_{A})$	Hyper	1	$\mathbb{R}$	✓		
`logit_phi_rho_as`	$\text{logit}(\phi^{\rho}_{AS})$	Hyper	1	$\mathbb{R}$	✓		
`log_sigma_rho_as`	$\log(\sigma^{\rho}_{AS})$	Hyper	1	$\mathbb{R}$	✓		
`log_sigma_rho_xa`	$\log(\sigma^{\rho}_{XA})$	Hyper	1	$\mathbb{R}$	✓		
`logit_phi_alpha_x`	$\text{logit}(\phi^{\alpha}_{X})$	Hyper	1	$\mathbb{R}$		✓	
`log_sigma_alpha_x`	$\log(\sigma^{\alpha}_{X})$	Hyper	1	$\mathbb{R}$		✓	
`logit_phi_alpha_xs`	$\text{logit}(\phi^{\alpha}_{XS})$	Hyper	1	$\mathbb{R}$		✓	
`log_sigma_alpha_xs`	$\log(\sigma^{\alpha}_{XS})$	Hyper	1	$\mathbb{R}$		✓	
`logit_phi_alpha_a`	$\text{logit}(\phi^{\alpha}_{A})$	Hyper	1	$\mathbb{R}$		✓	
`log_sigma_alpha_a`	$\log(\sigma^{\alpha}_{A})$	Hyper	1	$\mathbb{R}$		✓	
`logit_phi_alpha_as`	$\text{logit}(\phi^{\alpha}_{AS})$	Hyper	1	$\mathbb{R}$		✓	
`log_sigma_alpha_as`	$\log(\sigma^{\alpha}_{AS})$	Hyper	1	$\mathbb{R}$		✓	
`log_sigma_alpha_xa`	$\log(\sigma^{\alpha}_{XA})$	Hyper	1	$\mathbb{R}$		✓	
`OmegaT_raw`	Ω_T	Hyper	1	$\mathbb{R}$			✓
`log_betaT`	$\log(\beta_T)$	Hyper	1	$\mathbb{R}$			✓
`log_sigma_lambda_x`	$\log(\sigma^{\lambda})$	Hyper	1	$\mathbb{R}$			✓
`log_sigma_ancrho_x`	$\log(\sigma^{\rho^{\text{ANC}}}_{X})$	Hyper	1	$\mathbb{R}$			
`log_sigma_ancalpha_x`	$\log(\sigma^{\alpha^{\text{ANC}}}_{X})$	Hyper	1	$\mathbb{R}$			
`log_sigma_or_gamma`	$\log(\sigma^{\tilde{\gamma}}_{X})$	Hyper	1	$\mathbb{R}$			
`beta_rho`	$(\beta^{\rho}_{0}, \beta^{\rho,s=\text{M}}_{s})$	Latent	2	$\mathbb{R}^2$	✓		
`beta_alpha`	$(\beta^{\alpha}_{0}, \beta^{\alpha,s=\text{M}}_{S})$	Latent	2	$\mathbb{R}^2$		✓	
`beta_lambda`	$(\beta^{\lambda}_{0}, \beta^{\lambda,s=\text{M}}_{S})$	Latent	2	$\mathbb{R}^2$			✓
`beta_anc_rho`	$\beta^{\rho^{\text{ANC}}}$	Latent	1	$\mathbb{R}$			
`beta_anc_alpha`	$\beta^{\alpha^{\text{ANC}}}$	Latent	1	$\mathbb{R}$			
`u_rho_x`	w^{ρ}_{x}	Latent	n	$\mathbb{R}^n$	✓		
`us_rho_x`	v^{ρ}_{x}	Latent	n	$\mathbb{R}^n$	✓		
`u_rho_xs`	$w^{\rho,s=\text{M}}_{x}$	Latent	n	$\mathbb{R}^n$	✓		
`us_rho_xs`	$v^{\rho,s=\text{M}}_{x}$	Latent	n	$\mathbb{R}^n$	✓		
`u_rho_a`	u^{ρ}_{a}	Latent	10	$\mathbb{R}^{10}$	✓		
`u_rho_as`	$u^{\rho,s=\text{M}}_{a}$	Latent	10	$\mathbb{R}^{10}$	✓		
`u_rho_xa`	$u^{\rho,a<15}_{x}$	Latent	n	$\mathbb{R}^n$	✓		
`u_alpha_x`	w^{α}_{x}	Latent	n	$\mathbb{R}^n$		✓	
`us_alpha_x`	v^{α}_{x}	Latent	n	$\mathbb{R}^n$		✓	
`u_alpha_xs`	$w^{\alpha,s=\text{M}}_{x}$	Latent	n	$\mathbb{R}^n$		✓	
`us_alpha_xs`	$v^{\alpha,s=\text{M}}_{x}$	Latent	n	$\mathbb{R}^n$		✓	
`u_alpha_a`	u^{α}_{a}	Latent	13	$\mathbb{R}^{13}$		✓	
`u_alpha_as`	$u^{\alpha,s=\text{M}}_{a}$	Latent	10	$\mathbb{R}^{10}$		✓	
`u_alpha_xa`	$u^{\alpha,a<15}_{x}$	Latent	n	$\mathbb{R}^n$		✓	
`ui_lambda_x`	u^{λ}_{x}	Latent	n	$\mathbb{R}^n$			✓
`ui_anc_rho_x`	$u^{\rho^{\text{ANC}}}_{x}$	Latent	n	$\mathbb{R}^n$			
`ui_anc_alpha_x`	$u^{\alpha^{\text{ANC}}}_{x}$	Latent	n	$\mathbb{R}^n$			
`log_or_gamma`	$u^{\tilde{\gamma}}_{x}$	Latent	n	$\mathbb{R}^n$			

Table S4: Correspondence between mathematical notation and variable names used in our `TMB` code. The total number of hyperparameters is 24, and the total number of latent field parameters is $51 + 14n$, where n is the number of districts. We use the notation ✓ to refer to direct dependence of the parameter on the variable, ✗ to refer to no dependence, and a blank entry to refer to dependence conditional on the data.

S2 MCMC convergence and suitability

We assessed MCMC convergence and suitability using a range of graphical and numerical tests. All potential scale reduction factor $\hat{R}$ statistics (Vehtari et al. 2021) were below 1.05 and therefore acceptable (Figure S2). However, even thinning by a factor of 20, samples were not obtained very efficiently, resulting in the majority of effective sample size (ESS) ratios being below 0.5, with some as low as 0.1 (Figure S3). As a result, the number of obtained ESS varied substantially by parameter (Figure S4). There were no divergent transitions. Could add energy plot from Betancourt (2017) here.

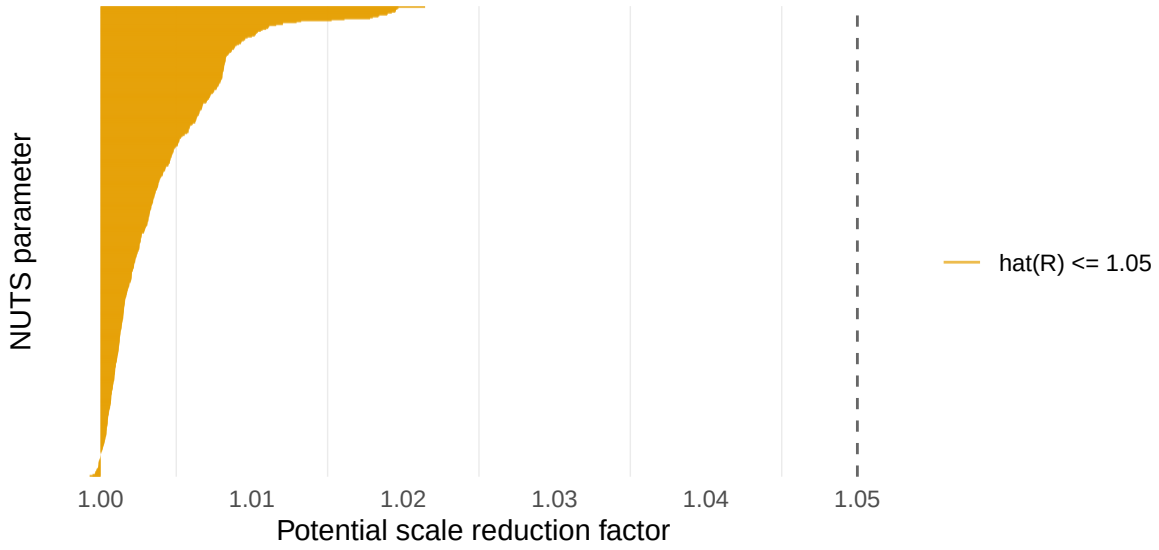

Figure S2: The potential scale reduction factor compares between- and within- estimates of univariate parameters. It is recommended only to use NUTS results if the value is less than 1.05, which it is for all parameters.

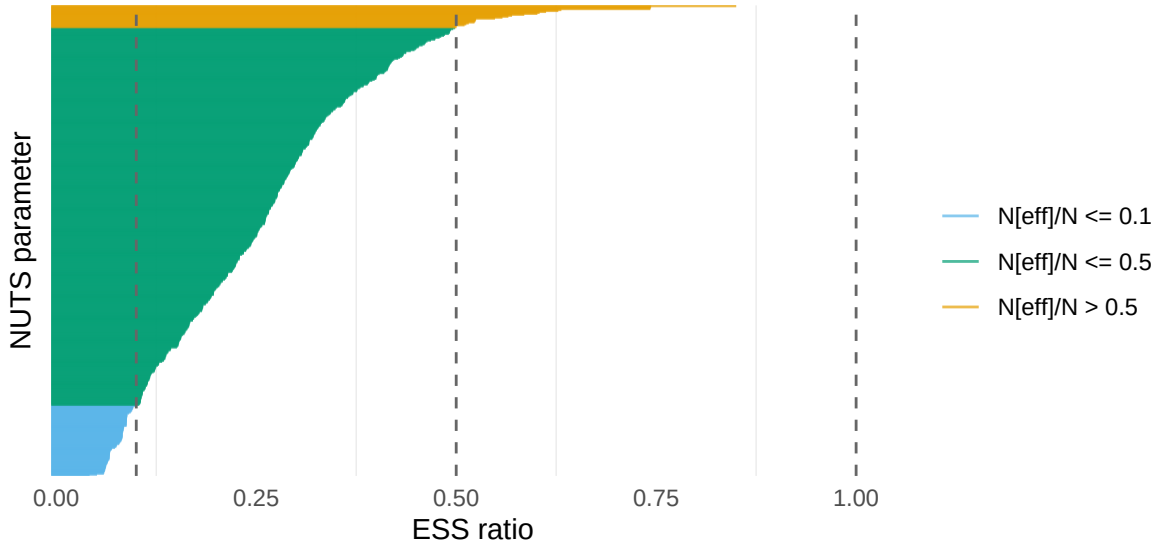

Figure S3: The efficiently, as measured by the ESS ratio, of the NUTS sampler was poor.

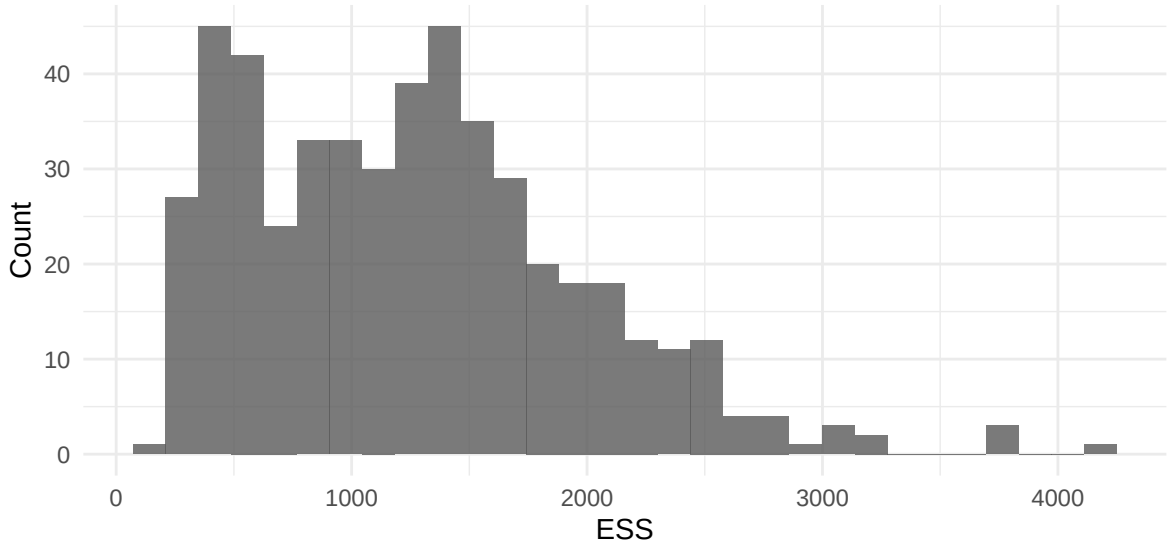

Figure S4: The effective number of samples we obtained varied substantially between parameters.

S3 AGHQ and PCA-AGHQ details

S3.1 Additional figures

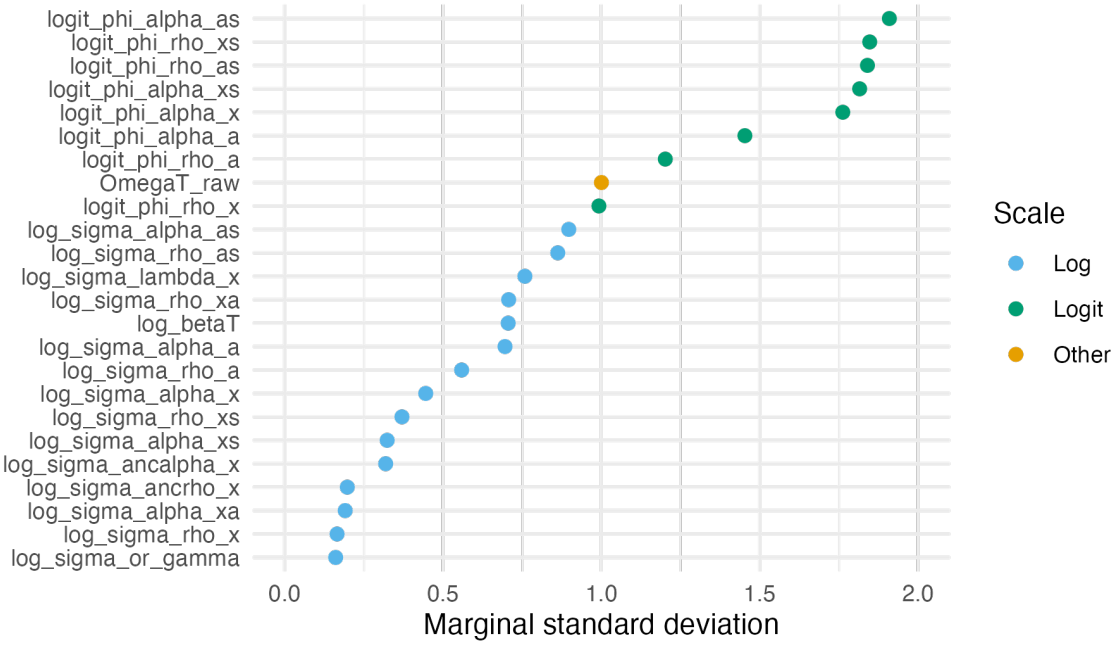

Figure S5: Hyperparameters on the logit-scale had systematically higher marginal standard deviations than those on the log-scale. This is because variation on the real scale is compressed by the inverse logit and expanded by the inverse log (exponential).

S3.2 Estimated normalising constant comparison

Add here plots and description of the estimated normalising constant for different PCA-AGHQ settings.

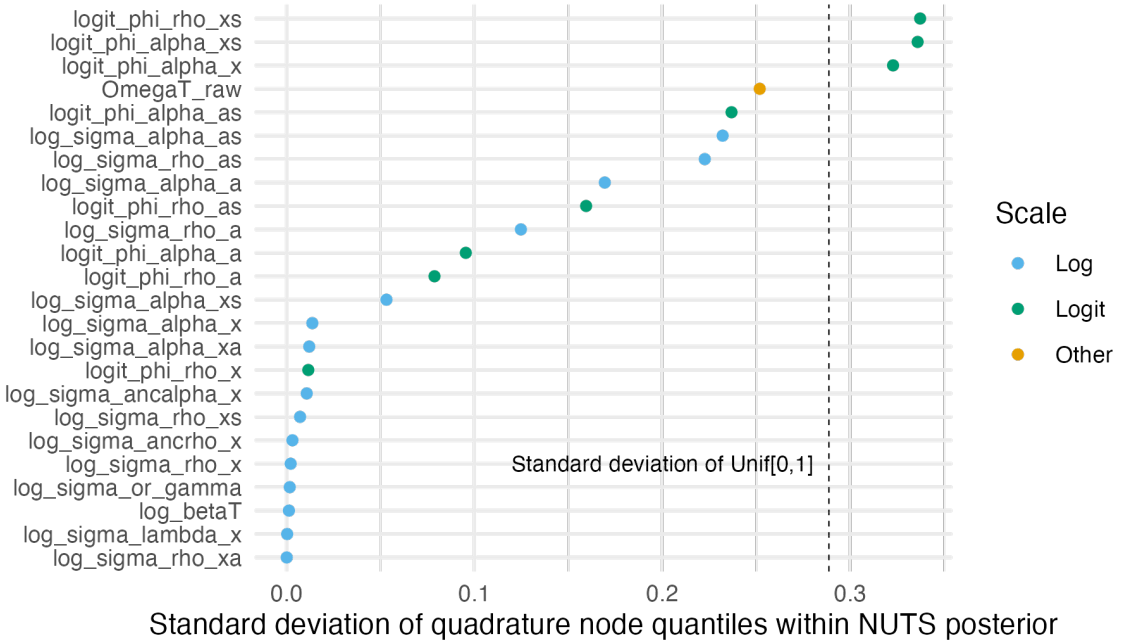

Figure S6: Standard deviations of the quantiles of the quadrature nodes within the NUTS posterior draws varied substantially, in accordance with the marginal standard deviations shown in Figure S5.

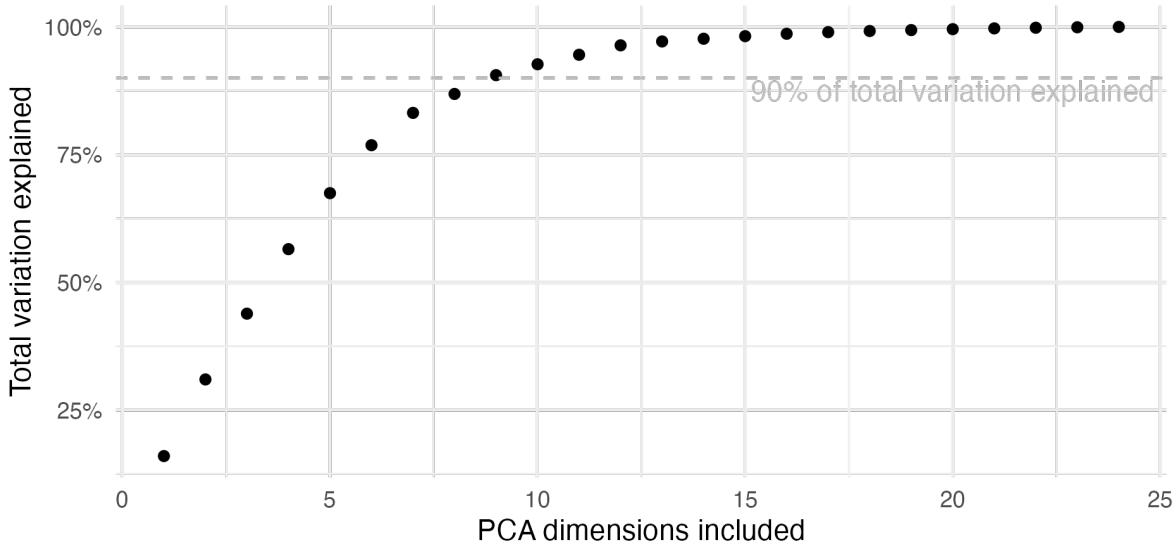

Figure S7: Most variation could be explained by far fewer than the full 24 dimensions.

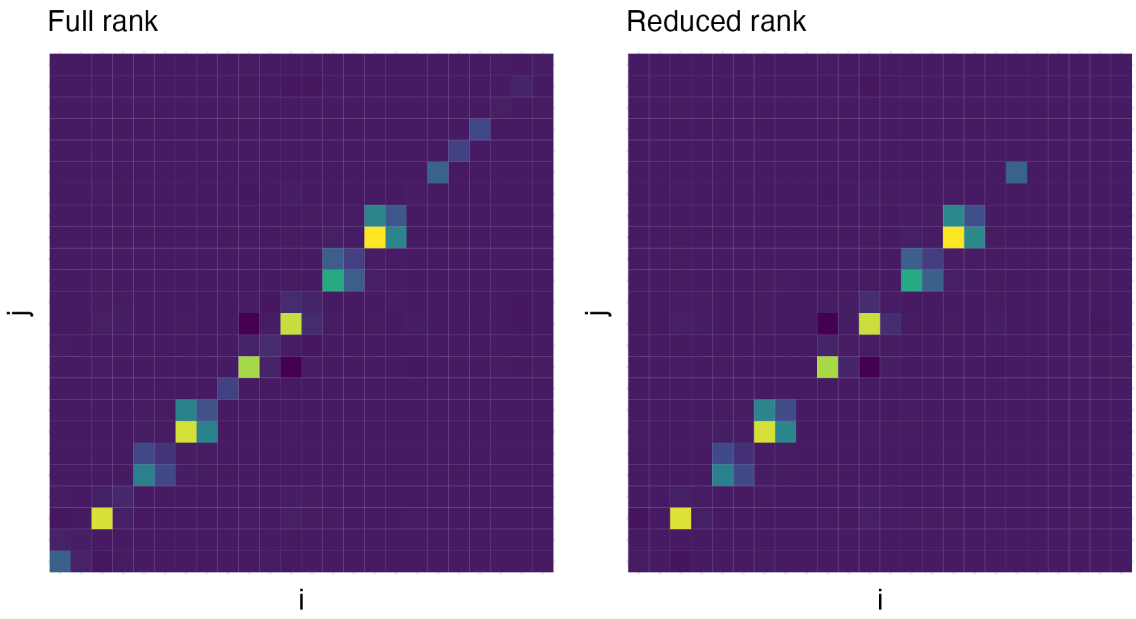

Figure S8: The reduced rank (8) matrix approximation to the Hessian is visually similar to the full rank matrix.

S4 Further inference comparison results

S4.1 Point estimates

S4.2 Distribution tests

	D(TMB)	D(PCA-AGHQ)
beta_alpha	0.077	0.076
beta_anc_alpha	0.081	0.077
beta_anc_rho	0.100	0.113
beta_lambda	0.070	0.070
beta_rho	0.071	0.072
log_or_gamma	0.056	0.053
u_alpha_a	0.070	0.035
u_alpha_as	0.074	0.079
u_alpha_x	0.116	0.098
u_alpha_xa	0.071	0.061
u_alpha_xs	0.094	0.079
u_rho_a	0.072	0.083
u_rho_as	0.068	0.063
u_rho_x	0.071	0.067
u_rho_xs	0.147	0.139
ui_anc_alpha_x	0.078	0.075
ui_anc_rho_x	0.055	0.059
ui_lambda_x	0.110	0.113
us_alpha_x	0.088	0.062
us_alpha_xs	0.097	0.063
us_rho_x	0.081	0.083

us_rho_xs	0.039	0.039
Average	0.081	0.075

S4.3 Pareto-smoothed importance sampling

S4.4 Maximum mean discrepancy

S5 Algorithm using Laplace latent field marginals

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

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

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

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

2. Generate a set of nodes $\mathbf{u} \in \mathcal{Q}(m, k)$ and weights $\omega : \mathbf{u} \to \mathbb{R}$ from a Gauss-Hermite quadrature rule with k nodes per dimension. Adapt these nodes based on the mode and lower Choleksy via $\boldsymbol{\theta}(\mathbf{u}) = \hat{\boldsymbol{\theta}} + \mathbf{L}\mathbf{u}$. Use this quadrature rule to calculate the normalising constant $\tilde{p}_{\texttt{AQ}}(\mathbf{y})$ as follows

$$\tilde{p}_{\texttt{AQ}}(\mathbf{y}) = \sum_{\mathbf{u}\in\mathcal{Q}(m,k)} \tilde{p}_{\texttt{LA}}(\boldsymbol{\theta}(\mathbf{u}), \mathbf{y})\omega(\mathbf{u}). \tag{4}$$

3. For $i \in [N]$ generate l nodes $x_i(\mathbf{v})$ via a Gauss-Hermite quadrature rule $\mathbf{v} \in \mathcal{Q}(1, l)$ adapted based on the mode $\hat{\mathbf{x}}(\boldsymbol{\theta})_i$ and standard deviation $\sqrt{\text{diag}[\hat{\mathbf{H}}(\boldsymbol{\theta})^{-1}]_i}$ of the Gaussian marginal. A value of $l \geq 4$ is recommended to enable B-spline interpolation. For $x_i \in \{x_i(\mathbf{v})\}_{\mathbf{v}\in\mathcal{Q}(1,l)}$ and $\boldsymbol{\theta} \in \{\boldsymbol{\theta}(\mathbf{u})\}_{\mathbf{u}\in\mathcal{Q}(m,k)}$ calculate the modes and Hessians

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

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

4. For $x_i \in \{x_i(\mathbf{v})\}_{\mathbf{v}\in\mathcal{Q}(1,l)}$ calculate

(5)

   where

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

   and

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

   Although Equation **??** can be calculated using the estimate of the evidence given in Equation 4 it is more numerically accurate, and requires little extra computation, to use the estimate

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

5. Given $\{x_i(\mathbf{v}), \tilde{p}_{\texttt{AQ}}(x_i(\mathbf{v})\,|\,\mathbf{y})\}_{\mathbf{v}\in\mathcal{Q}(1,l)}$ create a spline interpolant to each posterior marginal on the log-scale. Samples, and thereby relevant posterior marginal summaries, may be obtained using inverse transform sampling.

References

Besag, Julian, Jeremy York, and Annie Mollié. 1991. "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics* 43 (1): 1–20.

Betancourt, Michael. 2017. "A Conceptual Introduction to Hamiltonian Monte Carlo." *arXiv Preprint arXiv:1701.02434*.

Eaton, Jeffrey W., Laura Dwyer-Lindgren, Steve Gutreuter, Megan O'Driscoll, Oliver Stevens, Sumali Bajaj, Rob Ashton, et al. 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 (S5): e25788. https://doi.org/https://doi.org/10.1002/jia2.25788.

Freni-Sterrantino, Anna, Massimo Ventrucci, and Håvard Rue. 2018. "A Note on Intrinsic Conditional Autoregressive Models for Disconnected Graphs." *Spatial and Spatio-Temporal Epidemiology* 26: 25–34.

Kish, Leslie. 1965. "Survey Sampling."

Kristensen, Kasper. 2021. "The Comprehensive TMB Documentation." https://kaskr.github.io/adcomp/_book/Introduction.html.

Lumley, Thomas. 2004. "Analysis of Complex Survey Samples." *Journal of Statistical Software* 9: 1–19.

Simpson, Daniel, Håvard Rue, Andrea Riebler, Thiago G Martins, and Sigrunn H Sørbye. 2017. "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." *Statistical Science* 32 (1): 1–28.

Stover, J, P Johnson, T Hallett, M Marston, R Becquet, and IM Timaeus. 2010. "The Spectrum Projection Package: Improvements in Estimating Incidence by Age and Sex, Mother-to-Child Transmission, HIV Progression in Children and Double Orphans." *Sexually Transmitted Infections* 86 (Suppl 2): ii16–21.

Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2021. "Rank-normalization, folding, and localization: an improved R for assessing convergence of MCMC (with discussion)." *Bayesian Analysis* 16 (2): 667–718.