Fast approximate Bayesian inference for small-area estimation of HIV indicators using the Naomi model

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

[1] Department of Mathematics, Imperial College London
[2] MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London
[3] Department of Statistics and Actuarial Science, University of Waterloo
[4] Department of Computer Science, University of Oxford
[5] Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T.H. Chan School of Public Health

Summary

- We developed an approximate Bayesian inference method using Laplace approximation, adaptive Gauss-Hermite quadrature and principal component analysis
- Motivated by an evidence synthesis model for small-area estimation of HIV indicators in sub-Saharan Africa
- Implemented as a part of the `aghq` package (Stringer 2021), allowing flexible use of the method for any model with a Template Model Builder `TMB` (Kristensen et al. 2016) C++ user template

1. The Naomi HIV model

- District-level model of HIV indicators (Eaton et al. 2021) which synthesises data from 1) household surveys, 2) antenatal care (ANC) clinics, and 3) routine service provision of antiretroviral therapy (ART)
  - Combining evidence from multiple data sources helps overcome the limitations of any one
  - Small-area estimation methods to overcome small district-level sample sizes
- Yearly estimation process: model run interactively by country teams using a web-app `naomi.unaids.org`
  - Figure 1 illustrates the seven stages of using the app
- Inference conducted in minutes using empirical Bayes and a Gaussian approximation
- It would take days to get accurate answers with MCMC via `tmbstan` (Monnahan and Kristensen 2018), and this is not practical in this setting
- We are looking for a fast, approximate approach, that properly takes uncertainty in hyperparameters into account

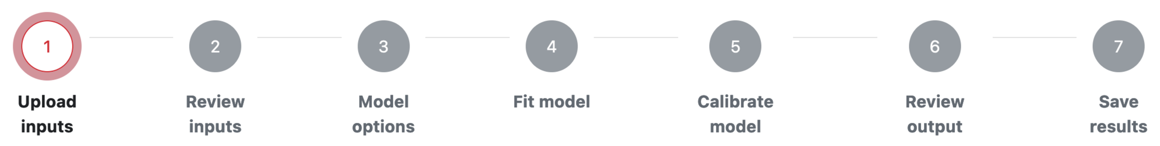

Figure 1: Model fitting occurs interactively in stages.

2. Extended latent Gaussian models

- Latent Gaussian models (LGMs) (Rue, Martino, and Chopin 2009) are three stage hierarchical models with observations y, Gaussian latent field x and hyperparameters θ
- In an LGM the conditional mean depends on exactly one structured additive predictor $\mu_i = g(\eta_i)$ with $g : \mathbb{R} \to \mathbb{R}$
- Extended latent Gaussian models (ELGM) remove this requirement such that $\mu_i = g(\eta_{\mathcal{J}_i})$ where $g_i : \mathbb{R}^{|\mathcal{J}_i|} \to \mathbb{R}$ and $\mathcal{J}_i$ is some set of indices
  - Allows a higher degree of non-linearity in the model
- Naomi is an ELGM, not an LGM, because it includes complex dependency structures:
  1. Incidence depends on prevalence and ART coverage
  2. Incidence ane prevalence linked to recent infection
  3. ANC offset from household survey
  4. ART coverage and recent infection are products
  5. Observed data are aggregated finer processes
  6. ART attendance uses the multinomial
  7. Multiple link functions
- We extend work of Stringer, Brown, and Stafford (2022) in this setting to the challenging Naomi ELGM
- Though we focus on Naomi, the HIV Inference Group (`hiv-inference.org`) works on many other complex models, challenging for existing Bayesian inference methods, which require flexible modelling tools

3. Inference procedure

- **Laplace approximation** Integrate out latent field using a Gaussian approximation to the denominator

$$p(\theta, y) \approx \tilde p_{\texttt{LA}}(\theta, y) = \frac{p(y, x, \theta)}{\tilde p_{\texttt{G}}(x \,|\, \theta, y)}\Big|_{x = \hat x(\theta)},$$

where $\tilde p_{\texttt{G}}(x \,|\, \theta, y) = \mathcal{N}(x \,|\, \hat x(\theta), \hat H(\theta)^{-1})$
  - Use automatic differentiation via `CppAD` in `TMB`

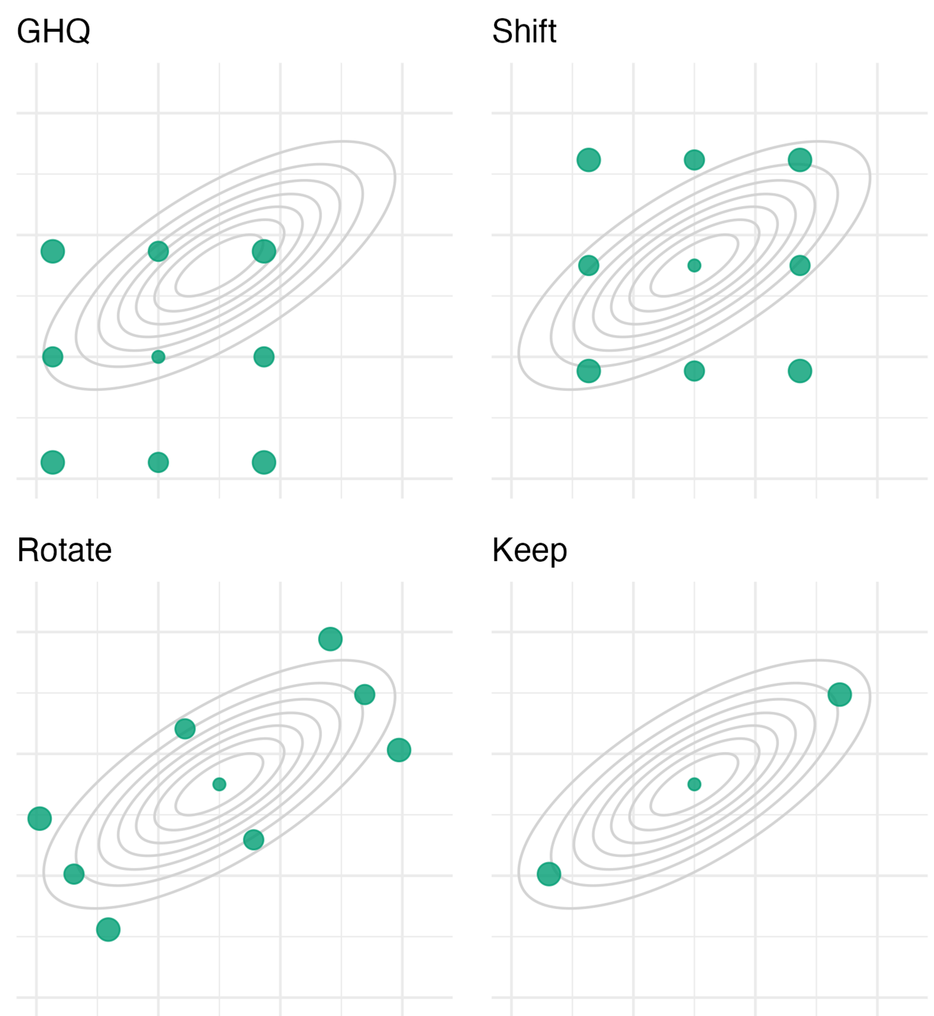

Figure 2: Demonstration of PCA-AGHQ.

- **Adaptive Gauss-Hermite Quadrature** (AGHQ) perform quadrature over the hyperparameters

$$\int_\Theta p_{\texttt{LA}}(\theta, y) \text{d}\theta \approx |L| \sum_{z \in \mathcal{Q}(m,k)} p_{\texttt{LA}}(\hat\theta + Lz, y)\omega(z),$$

where the Gauss-Hermite quadrature rule $\{z \in \mathcal{Q}(m, k), \omega\}$ with $m = \dim(\theta)$ and k points per dimension is adapted based upon
  - The mode $\hat\theta = \text{argmax}_{\theta \in \Theta} p_{\texttt{LA}}(\theta, y)$
  - A matrix decomposition $LL^\top = -\partial^2_\theta \log p_{\texttt{LA}}(\theta, y)|_{\theta = \hat\theta}$
- Use the spectral decomposition $L = E\Lambda^{1/2}$ and keep only the first $s < m$ **principal components** (PCA-AGHQ)

4. Application to Malawi

- Malawi is a relatively small country but still has latent field $\dim(x) = 491$ and hyperparameters $\dim(\theta) = 24$

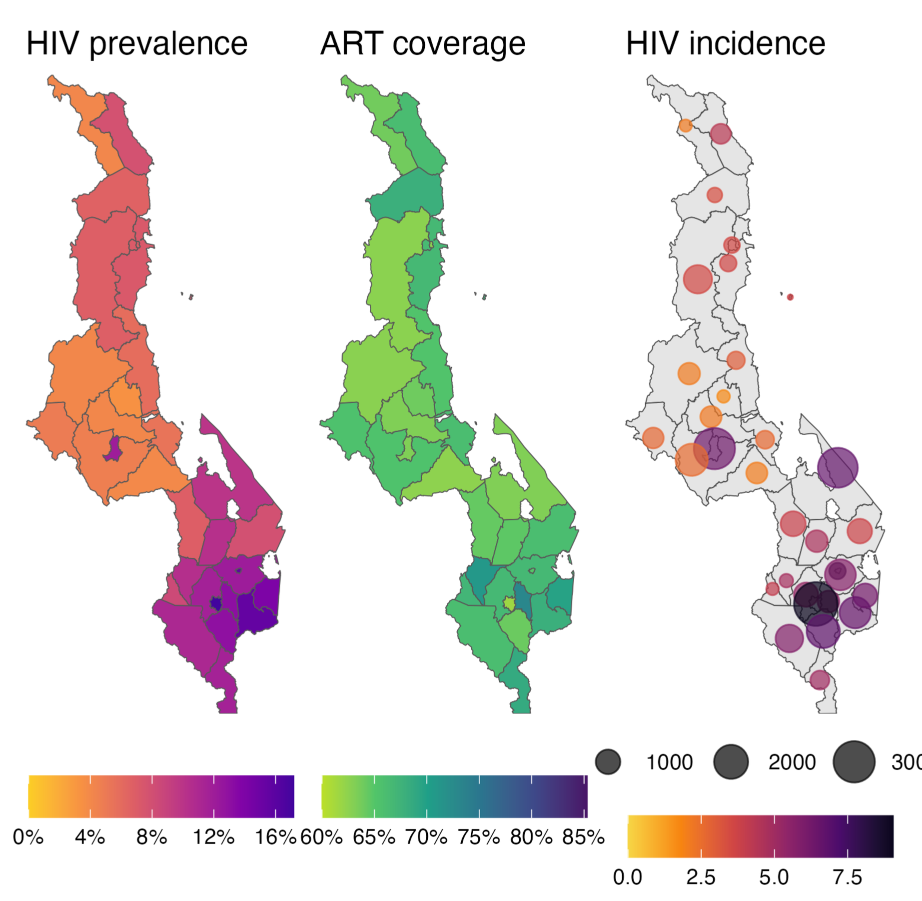

Figure 3: District-level model outputs for adults 15-49 in January 2016. Adapted from Eaton et al. 2021.

Method	Description	Time
TMB	Baseline	42 secs
PCA-AGHQ	Ours	1 hour
NUTS	Gold-standard	3.3 days

- For PCA-AGHQ $k = 3$ and $s = 8$ chosen using Scree plot to explain ~90% of variance
- For NUTS 4 chains of 100,000 thinned by 40 were required for good diagnostics
- Kolmogorov-Smirnov (KS) test based on the maximum difference between marginal ECDFs
  - Average KS distance from NUTS reduced by 10%
  - Also considering joint posteriors via Pareto-smoothed importance sampling and maximum mean discrepancy
- Naomi can be used to assess probabilities targets have been met e.g. 90% of those who know their HIV status are on ART ("second 90"). Both TMB and PCA-AGHQ have biased inferences (Figure 4)
  - Reduced RMSE for estimating second 90 exceedance probabilities by 9%

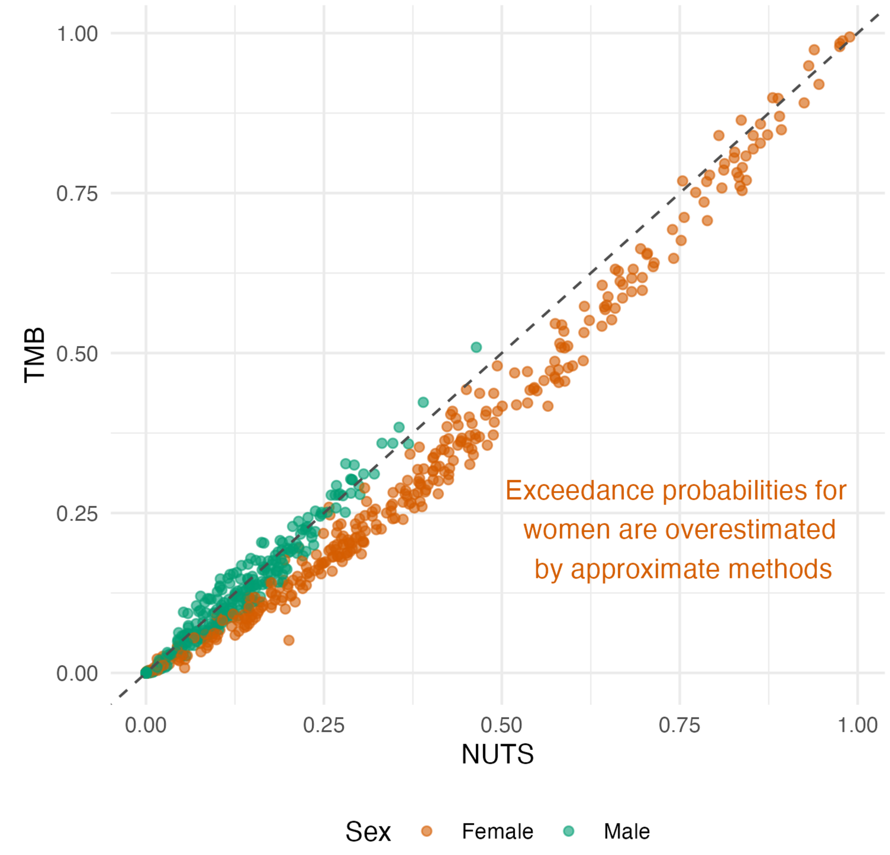

Figure 4: Both approximate methods are meaningfully incorrect for policy.

5. Future directions

- **Can we do any better than modest improvements?**
- Laplace marginals with matrix algebra approximations (Wood 2020) to speed up calculations
- Further methods for allocation of effort to "important" dimensions of hyperparameter grid

Funding AH was supported by the EPSRC and Bill & Melinda Gates Foundation. This research was supported by the MRC Centre for Global Infectious Disease Analysis.

Contact `athowes.github.io`; `ath19@ic.ac.uk`

References

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.

Kristensen, Kasper, Anders Nielsen, Casper W Berg, Hans Skaug, Bradley M Bell, et al. 2016. "TMB: Automatic Differentiation and Laplace Approximation." *Journal of Statistical Software* 70 (i05).

Monnahan, Cole C, and Kasper Kristensen. 2018. "No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages." *PloS One* 13 (5): e0197954.

Rue, Håvard, Sara Martino, and Nicolas Chopin. 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 (2): 319–92.

Stringer, Alex. 2021. "Implementing Approximate Bayesian Inference using Adaptive Quadrature: the aghq Package." *arXiv Preprint arXiv:2101.04468*.

Stringer, Alex, Patrick Brown, and Jamie Stafford. 2022. "Fast, Scalable Approximations to Posterior Distributions in Extended Latent Gaussian Models." *Journal of Computational and Graphical Statistics*, 1–15.

Wood, Simon N. 2020. "Simplified integrated nested Laplace approximation." *Biometrika* 107 (1): 223–30.