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[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

Summary

- Approximate Bayesian inference method using Laplace approximations and adaptive Gauss-Hermite quadrature
- Motivated by an evidence synthesis model for small-area estimation of HIV indicators in sub-Saharan Africa
- Looking to implement as a part of the `aghq` package (Stringer 2021), allowing flexible use of the method for any model with a `TMB` C++ user template

The Naomi HIV model

- District-level model of HIV indicators (Eaton et al. 2021) which synthesises data from household surveys, antenatal care (ANC) clinics, and 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 (EB) and a Gaussian approximation via Template Model Builder `TMB` (Kristensen et al. 2016)
- It would take days to get accurate answers with MCMC via `tmbstan` (Monnahan and Kristensen 2018), and this is not practical in this setting
- Motivates looking for a fast, approximate approach, that takes uncertainty in hyperparameters into account

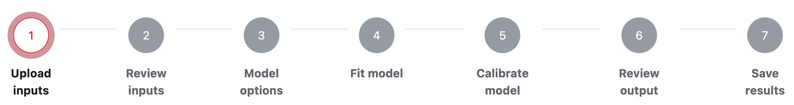

Figure 1: Model fitting occurs interactively in stages.

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 $g : \mathbb{R} \to \mathbb{R}$
  - The `R-INLA` implementation of integrated nested Laplace approximations applies only to LGMs, because ELGM precision matrices are not as sparse
- 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. ANC indicators offset from household survey
  2. Incidence depends on prevalence and ART coverage
  3. Observed data are aggregated finer processes
  4. Allow attendance of ART clinics outside home district
- We extend work of Stringer, Brown, and Stafford (2022) in this setting to the challenging Naomi ELGM
- Though we focus on Naomi here, the HIV Inference Group (`hiv-inference.org`) works on many other complex models, challenging for existing Bayesian inference methods, which require flexible modelling tools

Inference procedure

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

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

  where $\tilde{p}_{\mathrm{G}}(x \mid \theta, y) = \mathcal{N}(x \mid \hat{x}(\theta), \mathbf{H}(\theta)^{-1})$
  - Use automatic differentiation via `CppAD` in `TMB`
- **Adaptive Gauss-Hermite Quadrature**

$$\int_\Theta p(\theta) \mathrm{d}\theta \approx |L| \sum_{z \in \mathcal{Q}(m,k)} p(\hat{\theta} + Lz)\omega(z)$$

  where the Gauss-Hermite quadrature rule $z \in \mathcal{Q}(\dim(\theta), k)$ with k points per dimension is adapted based upon
  - The mode $\hat{\theta} = \mathrm{argmax}_{\theta \in \Theta} p(\theta)$
  - The lower Cholesky $LL^\top = -\partial^2_\theta \log p(\theta)|_{\theta=\hat{\theta}}$
- Algorithm (called `adam` for now) summarized by Figure 2
  - Where possible, previously calculated quantities and quadrature rules are reused

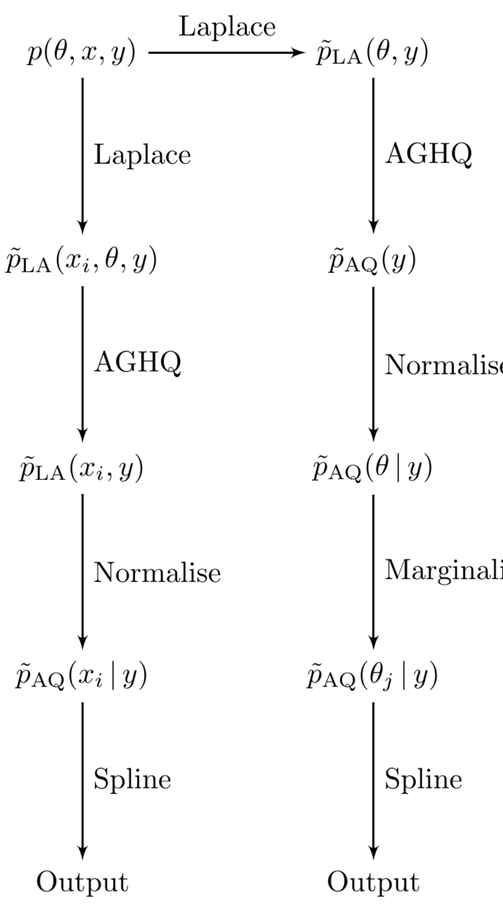

Figure 2: Flowchart describing the algorithm

Application to Malawi data

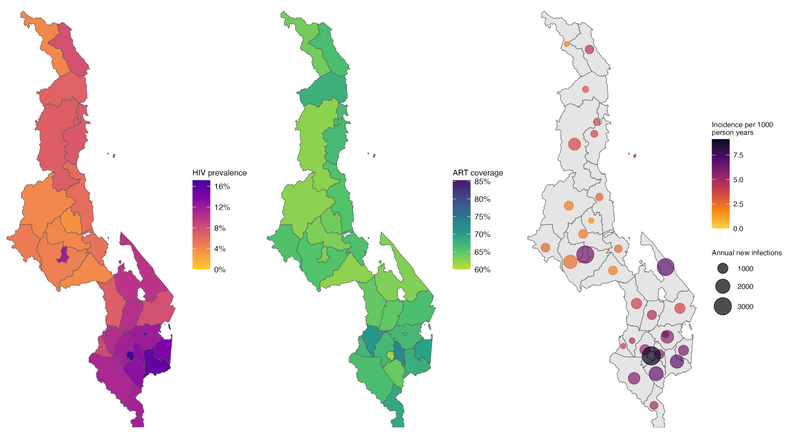

Figure 3: District-level model outputs for adults 15-49 in January 2016. Adapted from Eaton et al. 2021.

- Relatively small country but still a large model: latent field $\dim(x) = 491$, hyperparameters $\dim(\theta) = 24$
- Fit four inference methods (using one [!] C++ template):
  - `TMB` (3 mins)
  - `aghq` (1 mins): $k = 1$
  - `adam` (27 min): $k = 1$
  - `tmbstan` (2.4 days): 4 chains of 100,000 thinned by 40 (required for good diagnostics)
- Figure 3 illustrates example model outputs: HIV prevalence, ART coverage, HIV incidence, and number of new infections, at the district level

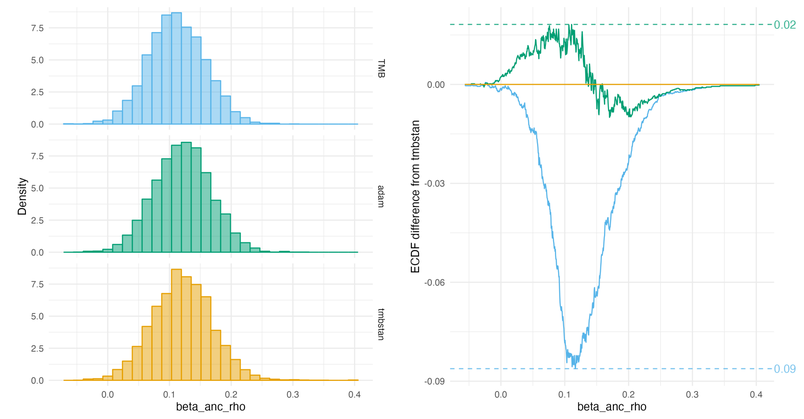

Figure 4: Inference reults and ECDF comparison for one x_i.

- Compare hyperparameter, latent field, and output posterior marginals based on maximum ECDF difference (Kolmogorov-Smirnov test)
- Figure 4 illustrates this approach for one node in the model with $\mathrm{KS}(\texttt{TMB}) = 0.09$ and $\mathrm{KS}(\texttt{adam}) = 0.02$

Future directions

- Scaling up the hyperparameter grid beyond EB $k = 1$
  - Any dense grid would be impractical (k^{24} nodes)
  - Alternatives: sparse grids, dense grids on a subspace
- Add Laplace matrix algebra approximations (Wood 2020) to speed up latent field marginal calculations
- More comprehensive inference comparison
  - Maximum mean discrepancy
  - Pareto-smoothed importance sampling

Interested? Working notebooks and R code available from `github.com/athowes/elgm-inf`. Or get in touch:

- `athowes.github.io`
- `ath19@ic.ac.uk`
- `adamhowes`

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.

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.