Submitted to the Annals of Applied Statistics

SIMPLIFIED INTEGRATED NESTED LAPLACE APPROXIMATION FOR EXTENDED LATENT GAUSSIAN MODELS

BY ADAM HOWES [1] , ALEX STRINGER [2]

[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*

Naomi is a spatial evidence synthesis model used by countries in sub-Saharan Africa to produce HIV epidemic indicators. We combine the simplified integrated nested Laplace approximation approach of Wood (2020) with adaptive Gaussian hermite quadrature to enable fast and accurate inference for Naomi and other extended latent Gaussian models. Using data from Malawi, we compare our inference method to other approaches. We provide an easy to use implementation as a part of the `aghq` R package allowing flexible use of the INLA method for a broader class of models, including any with a `TMB` template.

1. Introduction. We are motivated by a challenging inference problem in HIV surveillance. Mounting an effective public health response to the epidemic requires accurate, timely and sufficiently fine-scale estimates of HIV indicators. No single data source is sufficient to base these estimates on. For example, although nationally-representative household surveys are the most reliable data source, in most countries they only occur infrequently. As such, the Naomi small-area estimation model (Eaton et al., 2021) synthesises data from multiple sources to estimate HIV prevalence, HIV incidence, and coverage of antiretroviral treatment (ART) at a district-level. Software (`https://naomi.unaids.org`) has been developed for Naomi, which allows countries to input their data and generate estimates in a yearly process supported by UNAIDS. For this reason, any inference method used must be fast enough to run in production by country teams, ruling out prohibitively slow Markov chain Monte Carlo (MCMC) approaches. Inference is currently conducted using the Template Model Builder (`TMB`) R package (Kristensen et al., 2016). `TMB` implements an empirical Bayes approach, and is gaining popularity in spatial statistics for its speed and flexibility (Osgood-Zimmerman and Wakefield, 2021).

Using simulation we found inferences from `TMB` for the Naomi model to sometimes be inaccurate. The `aghq` R package (Stringer, 2021) for adaptive Gaussian Hermite quadrature (AGHQ), significantly improved performance. We develop a new inference methodology combining the simplified integrated nested Laplace approximation (INLA) approach of Wood (2020) with AGHQ.

2. The Naomi model. Eaton et al. (2019) specify a joint model linking small-area estimation models of HIV prevalence from household surveys, HIV prevalence from antenatal care clinics, and antiretroviral therapy (ART) coverage from routine health data collection. This model forms the basis of the Naomi small-area estimation model, described fully in Eaton et al. (2021). Modelling data from multiple sources concurrently is attractive as it increases statistical power, mitigates the biases of any single source, and prompts investigation into any data conflicts. The model is comprised of three components, described as follows.

Keywords and phrases: Spatial statistics, INLA.

2.1. *Household survey component.* Consider a country partitioned into n areas indexed by i. Suppose a simple random household survey of m_i^{HS} people is conducted in each area, and y_i^{HS} HIV positive cases are observed. Cases may be modelled using a binomial logistic regression model

$$y_i^{\text{HS}} \sim \text{Bin}(m_i^{\text{HS}}, \rho_i^{\text{HS}}), \tag{2.1}$$

$$\text{logit}(\rho_i^{\text{HS}}) \sim \mathcal{N}(\beta_\phi, \sigma_\phi^2), \tag{2.2}$$

where HIV prevalence ρ_i^{HS} is modelled by a Gaussian with mean β_ϕ and standard deviation σ_ϕ.

2.2. *ANC component.* Routinely collected data from pregnant women attending antenatal care clinics (ANCs) is another important source of information about the HIV epidemic. Suppose that of m_i^{ANC} women attending ANC, y_i^{ANC} are HIV positive. Then an analogous binomial logistic regression model

$$y_i^{\text{ANC}} \sim \text{Bin}(m_i^{\text{ANC}}, \rho_i^{\text{ANC}}), \tag{2.3}$$

$$\text{logit}(\rho_i^{\text{ANC}}) = \text{logit}(\rho_i^{\text{HS}}) + b_i, \tag{2.4}$$

$$b_i \sim \mathcal{N}(\beta_b, \sigma_b^2), \tag{2.5}$$

may be used to describe HIV prevalence amongst the sub-population of women attending ANCs. Reflecting the fact that prevalence in ANCs is related but importantly different to prevalence in the general population, bias terms b_i are used to offset ANC prevalence from HIV prevalence on the logit scale.

2.3. *ART component.* The number of people receiving treatment at district health facilities A_i provides further information about HIV prevalence. Districts with high prevalence are likely to have a greater number of people receiving treatment, and vice versa. ART coverage, defined to be the proportion of people living with HIV (PLHIV) currently on ART on district i, is given by $\alpha_i = A_i/\rho_i^{\text{HS}} N_i$, where N_i is the total population of district i and assumed to be constant. As such, ART coverage may also be modelled using a binomial logistic regression model

$$A_i \sim \text{Bin}(N_i, \rho_i^{\text{HS}} \alpha_i), \tag{2.6}$$

$$\text{logit}(\alpha_i) \sim \mathcal{N}(\beta_\alpha, \sigma_\alpha^2), \tag{2.7}$$

where the proportion of people receiving ART is $\rho_i^{\text{HS}} \alpha_i$. Here we assume no travel between districts to receive treatment.

2.4. *Joint model.* Let $\mathbf{y} = (\mathbf{y}^{\text{HS}}, \mathbf{y}^{\text{ANC}}, \boldsymbol{A})$ be the complete response vector. In this section, we would like to set-up notation for the joint model, helping to show that it isn't a latent Gaussian model, but is an extended latent Gaussian model. When introducing LGM and ELGM below, we may refer back to this model as example. For example, we can note that each of the components individually might be latent Gaussian (perhaps for the ART component it depends if you consider ρ_i^{HS} to be fixed) but when combined they are no longer.

3. Fast inference methods.

3.1. *Integrated nested Laplace approximation.* INLA (Rue, Martino and Chopin, 2009) is an approximate Bayesian inference method based on nested Gaussian approximations and numerical integration, and designed for use with latent Gaussian models (LGMs) of the form

$$\text{(3.1)} \qquad \text{(Observations)} \qquad y_i \sim p(y_i \,|\, x_i, \boldsymbol{\theta}), \quad i = 1, \ldots, n,$$

$$\text{(3.2)} \qquad \text{(Latent field)} \qquad \mathbf{x} \sim \mathcal{N}(\mathbf{x} \,|\, \mathbf{0}, \mathbf{Q}(\boldsymbol{\theta})^{-1}),$$

$$\text{(3.3)} \qquad \text{(Parameters)} \qquad \boldsymbol{\theta} \sim p(\boldsymbol{\theta}),$$

where $\dim(\mathbf{y}) = \dim(\mathbf{x}) = n$ and $\dim(\boldsymbol{\theta}) = m$, and $m < n$. For such models, the joint posterior of $(\mathbf{x}, \boldsymbol{\theta})$ is given by

$$\text{(3.4)} \qquad p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}) \propto p(\boldsymbol{\theta}) |\mathbf{Q}(\boldsymbol{\theta})|^{n/2} \exp\left(-\frac{1}{2} \mathbf{x}^\top \mathbf{Q}(\boldsymbol{\theta}) \mathbf{x} + \sum_{i=1}^{n} \log p(y_i \,|\, x_i, \boldsymbol{\theta}) \right).$$

Rather than approximating the above full posterior, the INLA method instead approximates the posterior marginals of each latent random variable x_i and parameter θ_j given by

$$\text{(3.5)} \qquad p(x_i \,|\, \mathbf{y}) = \int p(x_i, \boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta} = \int p(x_i \,|\, \boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta}, \quad i = 1, \ldots, n,$$

$$\text{(3.6)} \qquad p(\theta_j \,|\, \mathbf{y}) = \int p(\boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta}_{-j} \quad j = 1, \ldots, m.$$

An approximation is made to each of the two quantities, $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ and $p(x_i \,|\, \boldsymbol{\theta}, \mathbf{y})$, nested inside the above integrals: (i) $p(\boldsymbol{\theta} \,|\, \mathbf{y}) \approx \tilde{p}(\boldsymbol{\theta} \,|\, \mathbf{y})$ and (ii) $p(x_i \,|\, \boldsymbol{\theta}, \mathbf{y}) \approx \tilde{p}(x_i \,|\, \boldsymbol{\theta}, \mathbf{y})$. We discuss these two approximations in turn below.

3.1.1. *Approximation (i) .* The posterior marginal of the parameters $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ appears in both Equations (3.5) and (3.6). This distribution is approximated by $\tilde{p}(\boldsymbol{\theta} \,|\, \mathbf{y})$ and represented by a set of K integration points $\{\boldsymbol{\theta}^{(k)}\}$ and area-weights $\{\Delta^{(k)}\}$. The first step is to rewrite $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ as

$$\text{(3.7)} \qquad p(\boldsymbol{\theta} \,|\, \mathbf{y}) = \frac{p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y})}{p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})} \propto \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})}.$$

Approximation (i) then uses a Gaussian approximation to the denominator given by

$$\text{(3.8)} \qquad p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) \approx p_G(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) \triangleq \mathcal{N}(\mathbf{x} \,|\, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}), \hat{\mathbf{Q}}(\boldsymbol{\theta})^{-1}).$$

This approximation is accurate as the Gaussian prior on the latent field $\mathbf{x}$ makes the posterior distribution, given by

$$\text{(3.9)} \qquad p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) \propto \exp\left(-\frac{1}{2} \mathbf{x}^\top \mathbf{Q}(\boldsymbol{\theta}) \mathbf{x} + \sum_{i=1}^{n} \log p(y_i \,|\, x_i, \boldsymbol{\theta}) \right),$$

close to being Gaussian because $\mathbf{y}$ tends not to be strongly informative and the observation distribution $p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta})$ is usually well-behaved (Blangiardo and Cameletti, 2015). As $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ does not depend on $\mathbf{x}$, any value may be chosen to evaluate the right hand side of Equation 3.7. Taking $\mathbf{x} = \hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$, the value where the Gaussian approximation is most accurate, gives the final approximation as

$$\text{(3.10)} \qquad \tilde{p}(\boldsymbol{\theta} \,|\, \mathbf{y}) \propto \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{p_G(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x} = \hat{\boldsymbol{\mu}}(\boldsymbol{\theta})} = \frac{p(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}), \boldsymbol{\theta})}{\det(\hat{\boldsymbol{Q}}(\boldsymbol{\theta}))^{1/2}},$$

where the equality is because $p_G(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})$ is evaluated at its mode $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$.

3.1.2. *Approximation (ii).* We now move on to the approximation $p(x_i \mid \boldsymbol{\theta}, \mathbf{y}) \approx \tilde{p}(x_i \mid \boldsymbol{\theta}, \mathbf{y})$. Having used used the Gaussian approximation $p(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y}) \approx p_G(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})$ in Section 3.1.1 above, a natural approach, and that taken by Rue and Martino (2007), is to marginalise this distribution directly to obtain

$$\tilde{p}(x_i \mid \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}(x_i \mid \hat{\mu}_i(\boldsymbol{\theta}), 1/\hat{q}_i(\boldsymbol{\theta})), \tag{3.11}$$

where the marginal mean $\hat{\mu}_i(\boldsymbol{\theta})$ and precision $\hat{q}_i(\boldsymbol{\theta})$ are recovered directly from the relevant entries of $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{Q}}(\boldsymbol{\theta})$ respectively. However, although this approximation is fast, it tends not to be accurate, as it involves evaluating the Gaussian approximation away from its mode. As a result, although this method is available in `R-INLA` it is generally not advised. Instead, Rue, Martino and Chopin (2009) propose two methods, a Laplace approximation and a simplified version which is less computationally demanding. The full Laplace approximation is

$$p(x_i \mid \boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y}) \times \frac{1}{p(\mathbf{x}_{-i} \mid x_i, \boldsymbol{\theta}, \mathbf{y})} \tag{3.12}$$

$$= \frac{p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})}{p(\boldsymbol{\theta} \mid \mathbf{y})} \times \frac{1}{p(\mathbf{x}_{-i} \mid x_i, \boldsymbol{\theta}, \mathbf{y})} \tag{3.13}$$

$$\propto \frac{p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})}{p(\mathbf{x}_{-i} \mid x_i, \boldsymbol{\theta}, \mathbf{y})} \tag{3.14}$$

$$\approx \frac{p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})}{p_G(\mathbf{x}_{-i} \mid x_i, \boldsymbol{\theta}, \mathbf{y})}\Big|_{\mathbf{x}_{-i} = \hat{\boldsymbol{\mu}}_{-i}(x_i, \boldsymbol{\theta})} = \tilde{p}_{LA}(x_i \mid \boldsymbol{\theta}, \mathbf{y}), \tag{3.15}$$

where

$$p_G(\mathbf{x}_{-i} \mid x_i, \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}(\mathbf{x}_{-i} \mid \hat{\boldsymbol{\mu}}_{-i}(x_i, \boldsymbol{\theta}), \hat{\boldsymbol{Q}}_{-i}(x_i, \boldsymbol{\theta})), \tag{3.16}$$

is the Gaussian approximation to $\mathbf{x}_{-i} \mid x_i, \boldsymbol{\theta}, \mathbf{y}$ and $\hat{\boldsymbol{\mu}}_{-i}(x_i, \boldsymbol{\theta})$ is its modal configuration.[1] The set of distributions $\{p(\mathbf{x}_{-i} \mid x_i, \boldsymbol{\theta}, \mathbf{y})\}_{i=1}^n$ are usually reasonably Gaussian so this approximation tends to work well. However, the Gaussian approximation $p_G(\mathbf{x}_{-i} \mid x_i, \boldsymbol{\theta}, \mathbf{y})$ must be recomputed for each value of i, which is often computationally prohibitive. Therefore, two modifications to Equation (3.15) are proposed by Rue, Martino and Chopin (2009) to reduce the computational cost:

1. Avoiding having to find the mode via optimisation by using the approximation $\hat{\boldsymbol{\mu}}_{-i}(x_i, \boldsymbol{\theta}) \approx \mathbb{E}_{p_G(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})}(\mathbf{x}_{-i} \mid x_i)$
2. As only those x_j close to x_i should have an impact on the marginal of x_i, then by selecting some subset $R_i(\boldsymbol{\theta})$ of nodes j to impact j the matrix which needs to be factorised can be reduced in dimension to be $|R_i(\boldsymbol{\theta})| \times |R_i(\boldsymbol{\theta})|$ rather than $n \times n$

3.2. *Simplified INLA and connection to ELGMs.* Wood (2020) propose a simplified version of the INLA algorithm which relaxes the sparsity assumptions on the latent field required for the modifications to Equation (3.15) to be efficient. Doing so facilitates inference for ELGMs (Stringer, Brown and Stafford, 2022) which we define below. Informally, ELGMs build on LGMs by allowing each element of the linear predictor μ_i to depend on any subset of elements from the latent field.

3.3. *Implementation within* `aghq`.

[1] Note that $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$ is the mode of the Gaussian approximation to the full latent field given $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\mu}}_{-i}(x_i, \boldsymbol{\theta})$ is not the same as $\hat{\boldsymbol{\mu}}_{-i}(\boldsymbol{\theta})$.

3.3.1. *Adaptive Gaussian Hermite quadrature.*

3.3.2. *Template Model Builder.* `TMB` (Kristensen et al., 2016) is an R package for fitting random effect models, also known as latent variable models, hierarchical models or a host of other names. In `TMB` inference is based upon optimisation of a target function. This makes it very flexible, and able to handle non-linear, non-Gaussian random effect models.

The approach of `TMB` is inspired by the AD Model Builder (ADMB) package (Fournier et al., 2012), where the "AD" stands for automatic differentiation, a technique for calculating derivates of functions by repeated application of the chain rule. AD is popular in machine learning (Baydin et al., 2017), including as the basis for backpropagation algorithm, and is beginning to gain popularity in statistics, including as a part of Stan (Carpenter et al., 2017). `TMB` uses the derivatives from AD for multiple purposes including calculation of the Hessian used in Gaussian approximations and for numerical optimisation routines.

Consider unobserved latent random effects $\mathbf{x} \in \mathbb{R}^n$ and parameters $\boldsymbol{\theta} \in \mathbb{R}^m$.[2] Let $\ell(\mathbf{x}, \boldsymbol{\theta}) \triangleq -\log p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta})$ be the negative joint log-likelihood. In `TMB`, the user writes C++ code to evaluate this negative log-likelihood function ℓ. A standard maximum likelihood approach is to optimise

$$L_\ell(\boldsymbol{\theta}) \triangleq \int_{\mathbb{R}^n} p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}) \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^n} \exp(-\ell(\mathbf{x}, \boldsymbol{\theta})) \mathrm{d}\mathbf{x} \tag{3.17}$$

with respect to $\boldsymbol{\theta}$ to find the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$. Taking a superficially more Bayesian approach than above, instead of ℓ, the user may instead write a function to evaluate the negative joint penalised log-likelihood given by

$$f(\mathbf{x}, \boldsymbol{\theta}) \triangleq -\log p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}, \boldsymbol{\theta}) = \ell(\mathbf{x}, \boldsymbol{\theta}) - \log p(\mathbf{x}, \boldsymbol{\theta}), \tag{3.18}$$

equivalent up to an additive constant to the negative log-posterior. Using f in place of ℓ, then the penalised likelihood is proportional to the posterior marginal of $\boldsymbol{\theta}$

$$L_f(\boldsymbol{\theta}) \triangleq \int_{\mathbb{R}^n} \exp(-f(\mathbf{x}, \boldsymbol{\theta})) \mathrm{d}\mathbf{x} \propto \int_{\mathbb{R}^n} p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\mathbf{x} = p(\boldsymbol{\theta} \,|\, \mathbf{y}). \tag{3.19}$$

Integrating out the random effects directly, as in Equation 3.19 above, is usually intractable because $\mathbf{x}$ is high-dimensional, so Kristensen et al. (2016, Equation 3) use a Laplace approximation $L^\star_f(\boldsymbol{\theta})$ based instead upon integrating out a Gaussian approximation to the random effects. This Laplace approximation is analogous to the INLA approximation $\tilde{p}(\boldsymbol{\theta} \,|\, \mathbf{y})$ given in Section 3.1.1.

$$f''_{\mathbf{xx}}(\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}), \boldsymbol{\theta}) = -\frac{\partial^2}{\partial \mathbf{x}^2} \log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) \Big|_{\mathbf{x}=\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})} = -\frac{\partial^2}{\partial \mathbf{x}^2} \log p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) \Big|_{\mathbf{x}=\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})} = \hat{\boldsymbol{Q}}(\boldsymbol{\theta}).$$

Inference proceeds by optimising $L^\star_f(\boldsymbol{\theta})$ via minimisation of

$$-\log L^\star_f(\boldsymbol{\theta}) \propto \frac{1}{2} \log \det(\hat{\boldsymbol{Q}}(\boldsymbol{\theta})) + f(\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}), \boldsymbol{\theta}), \tag{3.20}$$

where $\propto$ is used to mean proportional up to an additive constant. The parameters of the Gaussian approximation (Equation 3.8), are found in terms of f via $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}) = \arg\min_{\mathbf{x}} f(\mathbf{x}, \boldsymbol{\theta})$ and $\hat{\boldsymbol{Q}}(\boldsymbol{\theta}) = f''_{\mathbf{xx}}(\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}), \boldsymbol{\theta})$ and must be recomputed for each value of $\boldsymbol{\theta}$. Obtaining $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$ is known as the inner optimisation step.

[2] Kristensen et al. (2016) use the notation u for random effects and θ for parameters. We aim for consistency with Section 3.1.

4. Application to the Naomi model.

5. Discussion.

Appendix.

Acknowledgements. AH was supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1).

REFERENCES

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.

BLANGIARDO, M. and CAMELETTI, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

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**.

EATON, J. W., BAJAJ, S., JAHN, A., KALUA, T., MGANGA, A., AULD, A. F., KIM, E., PAYNE, D., SHIRAISHI, R. W., GUTREUTER, S., HALLETT, T. B. and JOHNSON, L. F. (2019). Joint small-area estimation of HIV prevalence, ART coverage and HIV incidence. *Working paper*.

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.

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.

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**.

OSGOOD-ZIMMERMAN, A. and WAKEFIELD, J. (2021). A Statistical Introduction to Template Model Builder: A Flexible Tool for Spatial Modeling.

RUE, H. and MARTINO, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference* **137** 3177–3192.

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.

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.

WOOD, S. N. (2020). Simplified integrated nested Laplace approximation. *Biometrika* **107** 223–230.