Bayesian spatio-temporal methods for small-area estimation of HIV indicators

Imperial College London

Adam Howes

Department of Mathematics

Imperial College London

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

December 2023

Copyright

Statement of Originality

This thesis, and the work presented in it, is work that I conducted myself. In all cases where I describe others' work, I provide appropriate references.

For someone, or something.

Acknowledgements

Thank you to my supervisors Seth Flaxman and Jeff Eaton for their guidance and mentorship. I am grateful for the research environment provided by the Modern Statistics and Statistical Machine Learning Centre for Doctoral Training at Imperial and Oxford, the HIV Inference Group at Imperial, and the Machine Learning and Global Health Network. Thank you to Mike McLaren, Kevin Esvelt, the Nucleic Acid Observatory team, and the Sculpting Evolution lab for hosting my visit to the MIT Media Lab. Thank you to Alex Stringer, and the Department of Statistics and Actuarial Science for hosting my visit to the University of Waterloo. Without Alex, Chapter 6 would never have gotten moving. Thanks to Håvard Rue and Finn Lindgren for helpful answers on the `R-INLA` discussion group. My sense for what matters has been shaped (arguably improved) by the Effective Altruism community. Thank you to the Meridian, Trajan, and LEAH offices for hosting me. This research was made possible by funding provided by the Bill & Melinda Gates Foundation and EPSRC.

Adam Howes
Imperial College London
December 2023

Abstract

Progress towards ending AIDS as a public health threat by 2030 is not being made fast enough. Effective public health response requires accurate, timely, high-resolution estimates of epidemic and demographic indicators. Limitations of available data and statistical methodology make obtaining these estimates difficult. I developed and applied Bayesian spatio-temporal methods to meet this challenge. First, I examined models for area-level spatial structure. Second, I estimated district-level HIV risk group proportions, enabling behavioural prioritisation of prevention services, as put forward in the Global AIDS Strategy. Third, I developed a novel deterministic Bayesian inference method, combining adaptive Gauss-Hermite quadrature with principal component analysis, motivated by the Naomi district-level model of HIV indicators. In developing this method, I implemented integrated nested Laplace approximations using automatic differentiation, enabling inference for a wider class of models. Together, the contributions in this thesis help to guide precision HIV policy in sub-Saharan Africa, as well as advancing Bayesian methods for spatio-temporal data.

Contents

List of Figures **x**

List of Tables **xvii**

List of Abbreviations **xix**

List of Notations **xxi**

1 Introduction **1**

1.1 Chapter overview 2

2 The HIV/AIDS epidemic **4**

2.1 Background 4

2.2 HIV surveillance 9

3 Bayesian spatio-temporal statistics **14**

3.1 Bayesian statistics 14

3.2 Spatio-temporal statistics {st-statistics} 20

3.3 Model structure 25

3.4 Model comparison 28

3.5 Survey methods 28

4 Models for spatial structure **32**

4.1 Models based on adjacency 33

4.2 Models using kernels 41

4.3 Simulation study 45

4.4 HIV prevalence study 45

4.5 Discussion 45

5 A model for risk group proportions **46**

5.1 Background 46

5.2 Data 47

5.3 Model for risk group proportions 51

5.4 Prevalence and incidence by risk group 68

5.5 Discussion 72

6 Fast approximate Bayesian inference **78**

6.1 Inference methods and software 80

6.2 A universal INLA implementation 98

6.3 The Naomi model 110

6.4 AGHQ in moderate dimensions 120

6.5 Malawi case-study 122

6.6 Discussion 125

7 Conclusions **131**

7.1 Strengths 131

7.2 Weaknesses 132

7.3 Future work 132

7.4 Conclusions 132

Appendices

A Models for spatial structure **135**

A.1 Comparison of AGHQ to NUTS 135

A.2 Simulation study 135

A.3 HIV study 135

B A model for risk group proportions **136**

B.1 The Global AIDS Strategy 136

B.2 Household survey data 137

B.3 Spatial analysis levels 139

B.4 Survey questions and risk group allocation 139

C Fast approximate Bayesian inference **141**
C.1 Epilepsy example 141
C.2 Simplified Naomi model description 146
C.3 Model assessment 154
C.4 AGHQ and PCA-AGHQ details 154
C.5 Normalising constant estimation 154
C.6 Inference comparison 154
C.7 MCMC convergence and suitability 154

Works Cited **156**

List of Figures

1.1 HIV/AIDS is the largest cause of annual DALYs among individuals aged >1 year in SSA (Global Burden of Disease Collaborative Network 2019). One DALY represents the loss of the equivalent of one year of full health, and is calculated by the sum of years of life lost and years lost due to disability. Weights used to account for disability vary between 0 (full health) and 1 (death) depending on severity of the condition. 2

2.1 Globally, yearly new HIV infections peaked in 1995, and have since decreased by 59% and yearly AIDS-related deaths peaked in 2004, and have since decreased by 68% (UNAIDS 2023a). Much of the disease burden is concentrated in eastern and southern Africa, as well as western and central Africa. 5

2.2 Adult (15-49) HIV prevalence varies substantially both within and between countries in SSA. The estimates from 2023 were generated by country teams using the Naomi small-area estimation model in a process supported by UNAIDS, and are available from UNAIDS (2023a). White filled points are country-level estimates, and coloured points are district-level estimates. Results from Nigeria were not published. Data collection in the Cabo Delgado province of Mozambique was disrupted by conflict. Obtaining results for the Democratic Republic of the Congo required removing some districts from the model. 7

3.1 An example of Bayesian modelling and computation for a simple one parameter model. Here the likelihood is $y_i \sim \text{Poisson}(\phi)$ for $i = 1, 2, 3$ and prior distribution on the rate parameter $\phi > 0$ is $\phi \sim \text{Gamma}(3, 1)$. Observed data $\mathbf{y} = (1, 2, 3)$ was simulated from the distribution Poisson(2.5). The true data generating process is within the space of models being considered. (This situation is sometimes known (Bernardo and Smith 2001) as the $\mathcal{M}$-closed world, in contrast to the $\mathcal{M}$-open world where the model is said to be misspecified.) Furthermore, the posterior distribution is available in closed form as $\text{Gamma}(9, 4)$. This is because the posterior distribution is in the same family of probability distributions as the prior distribution. Models of this kind are described as being conjugate. Conjugate models are often used because of their convenience. Though other models may be more suitable, they will typically be more computationally demanding. The posterior distribution here is more tightly peaked than the prior distribution. This contraction is typically, but not always, the case. 16

3.2 NUTS can be used to sample from the posterior distribution in the example of Figure 3.1. Panel A shows a histogram of the NUTS samples as compared to the true posterior. The visual appearance of a histogram depends highly on the number of bins chosen. Other visualisations, such as empirical cumulative difference function plots, though less initially intuitive, are preferred for accurate distributinal sample comparisons. Panel B is a traceplot showing the path of the Markov chain $\{\phi_i\}_{i=1}^{1000}$ as it explores the posterior distribution. In this case, the Markov chain moves freely throughout the posterior distribution, without getting stuck in any one location for long, indicating good performance of the sampler. Panel C shows the convergence of the empirical posterior mean $\frac{1}{i}\sum_{j \leq i} \phi_j$ to the true value of $\mathbb{E}(\phi)$ as more iterations of the Markov chain are included in the calculation. 19

3.3 The spatial location of Cape Town in South Africa could be considered a point. The ZF Mgcawu District Municipality on the other hand is an example of an area. World AIDS Day, designated on the 1st of December every year, could be considered a point in time. The second fiscal quarter, running through April, May and June, and denoted by Q2 represents a period of time. (In reality, both Cape Town and World AIDS Day are areas, rather than true point locations. Instances of infinitesimal point locations in everyday life are rare.) . 21

3.4 Simulation of a simple random sample $y_i \sim \text{Bin}(m, p_i)$ with varying sample size $m = 5, 25, 125$ in each of the $i = 1, \ldots, 156$ constituencies of Zambia. Direct estimates were obtained by the empirical ratio of data to sample size. Modelled estimates were obtained using a logistic regression with linear predictor given by an intercept and a spatial random effect. The colour palette used in this figure is called viridis, as implemented by the `viridis` R package (Garnier et al. 2023), and was designed to be perceptually uniform and accessible to colourblind viewers (Smith and van der Walt 2015). This figure was adapted from a presentation given for the Zambia HIV Estimates Technical Working Group, available from `https://github.com/athowes/zambia-unaids`. Estimates of HIV indicators for Zambia have previously been generated at the district-level, comprising 116 spatial units. Moving forward, there is interest in generating estimates at the higher-resolution constituency level, as program planning is devolved locally. 23

3.5 The setting of this figure matches that of Figure 3.4. Estimates from surveys with higher sample size have higher Pearson correlation coefficient R with the underlying truth, illustrating the benefit of collecting more data. For a fixed sample size however, correlation can be improved by using modelled estimates to borrow information across spatial units, rather than using the higher variance direct estimates. Points along the dashed diagonal line correspond to agreement between the estimate obtained from the survey and the underlying truth used to generate the data. For each sample size, using a spatial model increases the correlation between the estimates and underlying truth. The effect is more pronounced for lower sample sizes. 24

4.1 Panel A shows the districts of Zimbabwe. Panel B shows the corresponding adjacency graph structure $\mathcal{G}$, with nodes positioned in alignment with the district that they correspond to. 34

4.2 Though they are quite different, the geometries shown in panels A, B, C, and D each have the same adjacency graph. 39

4.3 A sequence of geometries where the number of neighbours of area one grows by one at each iteration. 39

4.4 Each of the shaded areas are split into two moving from Panel A to Panel B. 40

5.1 Risk of acquiring HIV is depends on both individual-level risk behaviour and population-level HIV incidence. I assume that with no individual-level risk behaviour, there is no risk of acquiring HIV, independent of the population-level HIV incidence. The risk scale is intended to be illustrative, rather than interpreted quantitatively. . 48

5.2 Surveys conducted 1999-2018 that were used in the analysis by year, survey type, sample size, and whether the survey included a specific question about transactional sex. Survey type included AIDS Indicator Surveys (AIS), Demographic and Health Surveys (DHS), the Botswana AIDS Impact Survey 2013 (BAIS), and Population-based HIV Impact Assessment (PHIA) surveys. 49

5.3 Flowchart giving classification of survey respondents to HIV risk groups. 51

5.4 For the multinomial logistic regression model, under the CPO criterion, including Besag spatial random effects rather than IID spatial random effects improved model performance. On the other hand, under the DIC and WAIC, where smaller values are preferred, the opposite was true. Though IID temporal random effects are preferred by all criteria AR1 temporal random effects performed very similarly, likely as there is a limited amount of temporal variation in the data to describe. 59

5.5 For the logistic regression model, the CPO, DIC, and WAIC each agreed that the model containing Besag spatial random effects and the `cfswrecent` covariates was best. Inclusion of Besag spatial random effects consistently improved each criterion, whereas improvements from inclusion of any covariates were marginal. 61

5.6 The disaggregation procedure I used produces an age distribution for FSW peaking in the 20-24 and 25-29 age groups, and declining for older age groups. 63

5.7 Probability integral transform (PIT) histograms (top row) and empirical cumulative distribution function (ECDF) difference plots (bottom row) for the final selected model. 64

5.8 The spatial distribution (posterior mean) of the AGYW risk group proportions in 2018. Estimates are stratified by risk group (columns) and five-year age group (rows). Countries in grey were not included in the analysis. A limitation of this figure is that using a common colour scale, desirable for other reasons, makes it challenging to see spatial variation in the FSW risk group. 65

5.9 National (in white) and subnational (in color) posterior means of the risk group proportions. Estimates are stratified by risk group (columns) and five-year age group (rows). Though the information presented is similar to that of Figure 5.8, this figure presents a clear view of within- and between-country variation in risk group proportions. 66

5.10 Figure caption. 67

5.11 Figure caption. 68

5.12 Percentage of new infections reached across all 13 countries, taking a variety of risk stratification approaches, against the percentage of at risk population required to be reached. 71

5.13 Figure caption. 74

6.1 Demonstration of the Laplace approximation for the simple Bayesian inference example of Figure 3.1. The unnormalised posterior is $p(\phi, \mathbf{y}) = \phi^8 \exp(-4\phi)$, and can be recognised as the unnormalised gamma distribution Gamma$(9, 4)$. The true log normalising constant is $\log p(\mathbf{y}) = \log \Gamma(9) - 9\log(4) = -1.872046$, whereas the Laplace approximate log normalising constant is $\log \tilde{p}_{\texttt{LA}}(\mathbf{y}) = -1.882458$, resulting from the Gaussian approximation $p_{\texttt{G}}(\phi \,|\, \mathbf{y}) = \mathcal{N}(\phi \,|\, \mu = 2, \tau = 2)$. 83

6.2 The trapezoid rule with $k = 5, 10, 20$ equally-spaced ($\epsilon_i = \epsilon > 0$) quadrature nodes can be used to integrate the function $f(z) = z\sin(z)$, shown in green, in the domain $[0, \pi]$. Here, the exact solution is $\pi \approx 3.1416$. As k increases and more nodes are used in the computation, the quadrature estimate becomes closer to the exact solution. The trapezoid rule estimate is given by the sum of the areas of the grey trapezoids. 86

6.3 The Gauss-Hermite quadrature nodes $\mathbf{z} \in \mathcal{Q}(2, 3)$ for a two-dimensional integral with three nodes per dimension (Panel A). Adaption occurs based on the mode (Panel B) and covariance of the integrand via either the Cholesky (Panel C) or spectral (Panel D) decomposition of the inverse curvature at the mode. Here, the integrand is $f(z_1, z_2) = \text{sn}(0.5z_1, \alpha = 2) \cdot \text{sn}(0.8z_1 - 0.5z_2, \alpha = -2)$, where $\text{sn}(\cdot)$ is the standard skewnormal probability density function with shape parameter $\alpha \in \mathbb{R}$. The integral approximation $I \approx \int\int f(z_1, z_2) \mathrm{d}z_1 \mathrm{d}z_2$ obtained by the quadrature rule in each panel are given. 90

6.4 Consider the function $f(z_1, z_2) = \text{sn}(0.5z_1, \alpha = 2) \cdot \text{sn}(0.8z_1 - 0.5z_2, \alpha = -2)$ as described in Figure 6.3. Panel A shows the grid method as used in `R-INLA` and detailed in Section 3.1 of Rue, Martino, and Chopin (2009). Briefly, equally-weighted quadrature points are generated by starting at the mode and taking steps of size δ_z along each eigenvector of the inverse curvature at the mode, scaled by the eigenvalues, until the difference in log-scale function evaluations (compared to the mode) is below a threshold δ_π. Intermediate values are included if they have sufficient log-scale function evaluation. Here, I set $\delta_z = 0.75$ and $\delta_\pi = 2$. Panel B shows a CCD as used in `R-INLA` and detailed in Section 6.5 of Rue, Martino, and Chopin (2009). The CCD was generated using the `rsm` R package (Lenth 2009), and is comprised of: one centre point; four factorial points, used to help estimate linear effects; and four star points, used to help estimate the curvature. 97

6.5 The number of seizures in the treatment group was fewer, on average, than the number of seizures in the control group. This is not sufficient to conclude that the treatment was effective. The GLMM accounts for differences between the treatment and control group, including in baseline seizures and age, and so can be used to help estimate a causal treatment effect. 100

6.6 A submatrix of the full parameter Hessian obtained from `TMB::sdreport` with `getJointPrecision = TRUE` on the log scale. Entries for the latent field parameters $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}$ are omitted due to their respective lengths of 56 and 236. Light grey entries correspond to zeros on the real scale, which cannot be log transformed. 105

6.7 Percentage difference in posterior summary estimate obtained from NUTS as compared to that obtained from a Gaussian or Laplace marginal with quadrature over the hyperparameters. NUTS results were obtained with `tmbstan`. Results from `R-INLA` and `TMB` are similar, especially for the posterior mean, but do differ in places. . . 110

6.8 The ECDF and ECDF difference for the β_0 latent field parameter. For this parameter, the Gaussian results are inaccurate, and corrected by the Laplace. An ECDF difference of zero corresponds to obtaining exactly the same results as NUTS, taken to be the gold-standard. Results obtained using `R-INLA` and `TMB` implementations are highly similar. 111

6.9 The amount of time taken (in seconds) to perform inference with each method and software implementation. 112

6.10 Consider the function $f(z_1, z_2) = \text{sn}(0.5z_1, \alpha = 2) \cdot \text{sn}(0.8z_1 - 0.5z_2, \alpha = -2)$ as described in Figure 6.3. Panel A shows the AGHQ nodes with a spectral matrix decomposition, as usual. Panel B shows the adapted PCA-AGHQ nodes $\mathcal{Q}(2, 1, 3)$. These nodes correspond exactly to those in Panel A along the first eigenvector. The proportion of variation explained by this direction is around 95%, with the remaining 5% explained by the second eigenvector. As before, each panel shows the quadrature estimate of the integral I. . 122

6.11 District-level HIV prevalence, ART coverage, and new HIV cases and HIV incidence for adults 15-49 in Malawi. Inference conducted using a Gaussian approximation and EB via `TMB`. 123

6.12 Monthly R package downloads from the Comprehensive R Archive Network (CRAN) for `brms`, `glmmTMB`, `nimble`, `rstan` and `TMB`, obtained using the Csárdi (2023) R package. Unfortunately, `R-INLA` is not available from CRAN, and so could not be included in this figure. The official `rstan` documentation recommends installation of a development version hosted outside CRAN. 129

C.1 Figure caption. 146

C.2 Figure caption. 147

C.3 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. 155

List of Tables

5.1 HIV risk groups and HIV incidence rate ratios relative to AGYW with one cohabiting sexual partner. The incidence rate ratio for women with non-regular or multiple sexual partner(s) was derived from analysis of longitudinal data by Slaymaker et al. (2020). Among FSW, the incidence rate ratio (25.0, 13.0, 9.0, 6.0, 3.0) depended on the level of HIV incidence among the general population (<0.1%, 0.1-0.3%, 0.3-1.0%, 1.0-3.0%, >3.0%), such that higher local HIV incidence in the general population corresponded to a lower incidence rate ratio for FSW. Estimates of HIV incidence rate ratios for FSW were derived by UNAIDS based on patterns of relative HIV prevalence among FSW compared to general population prevalence. 48

5.2 Four multinomial regression models were considered. Observation random effects θ_{ita}, included in all models, are omitted from this table. 54

5.3 Applying sum-to-zero constraints to interaction effects ensures that the main effect is not interfered with. 58

5.4 CPO, DIC, and WAIC values for the multinomial logistic regression model with corresponding standard errors. 59

5.5 Six logistic regression models were considered. The covariate `cfswever` denotes the proportion of men who have ever paid for sex and `cfswrecent` denotes the proportion of men who have paid for sex in the past 12 months. 60

5.6 CPO, DIC, and WAIC values for the logistic regression model with corresponding standard errors. 61

6.1 The inference methods and software considered. 100

B.1 Prioritisation strata according to HIV incidence in the general population and behavioural risk. 136

B.2 Commitments to be met for each intervention in terms of proportion of the prioritisation strata reached. The symbol "-" represents no commitment. 136

B.3 All of the surveys that used in the analysis and their sample sizes, disaggregated by respondent age. 139
B.4 All of that surveys that were excluded from the analysis. 139
B.5 The numer of areas and analysis levels for each country that were used in the analysis. 139
B.6 The survey questions included in AIDS Indicator Survey (AIS) and Demographic and Health Surveys (DHS). 140
B.7 The survey questions included in Population-Based HIV Impact Assessment (PHIA) surveys. 140

C.1 The Naomi model can be conceptualised as having five processes. This table gives the number of latent field parameters and hyperparameters in each process, where n is the number of districts in the country. . 147
C.2 Each term in Equation (C.1) together with (where applicable) its prior distribution and a written description of its role. 148
C.3 Each term in Equation (C.6) together with (where applicable) its prior distribution and a written description of its role. 149
C.4 Each term in Equations (C.8) and (C.9) together with (where applicable) its prior distribution and a written description of its role. The notation θ is used as stand in for $\theta \in \{\rho, \alpha\}$. 150
C.5 Each term in Equation (C.11) and (C.9) together with (where applicable) its prior distribution and a written description of its role. No terms include x', such that $\gamma_{x,x'}$ is only a function of x. 151
C.6 153

List of Abbreviations

HIV . . . . . Human Immunodeficiency Virus.

AIDS . . . . . Acquired ImmunoDeficiency Syndrome.

PEPFAR . . . President's Emergency Plan for AIDS Relief.

Global Fund . Global Fund to Fight AIDS, Tuberculosis, and Malaria.

HIV . . . . . Demographic and Health Surveys.

AIS . . . . . AIDS Indicator Survey.

PrEP . . . . . Pre-Exposure Prophylaxis.

PEP . . . . . Post-Exposure Prophylaxis.

FSW . . . . . Female Sex Worker(s).

MSM . . . . . Men who have Sex with Men.

PWID . . . . . People Who Inject Drugs.

ANC . . . . . Antenatal Clinic.

UNAIDS . . . United Nations Joint Programme on HIV/AIDS.

CDC . . . . . Centers for Disease Control and Prevention.

UAT . . . . . Unlinked Anonymous Testing.

PMTCT . . . Prevention of Mother-to-Child Transmission.

PLHIV . . . . People Living with HIV.

TaSP . . . . . Treatment as Prevention.

MCMC . . . . Markov Chain Monte Carlo.

VI . . . . . Variational Inference.

INLA . . . . . Integrated Nested Laplace Approximation.

GP . . . . . Gaussian Process.

CAR . . . . . Conditionally Auto-regressive.

ICAR . . . . . Intrisic Conditionally Auto-regressive.

ART . . . . . Antiretroviral Therapy.

SAE . . . . . Small-Area Estimation.

GMRF . . . . Gaussian Markov Random Field.

HMC . . . . . Hamiltonian Monte Carlo.

GMRF . . . . Gauss-Markov Random Field.

HMC . . . . . Hamiltonian Monte Carlo.

LGM . . . . . Latent Gaussian Model.

ELGM . . . . . Extended Latent Gaussian Model.

DIC . . . . . Deviance Information Criterion.

BIC . . . . . Bayesian Information Criterion.

WAIC . . . . . Watanabe-Akaike Information Criterion.

ESS . . . . . Effective Sample Size.

IID . . . . . Independent and Identically Distributed.

PPL . . . . . Probabilistic Programming Language.

CCD . . . . . Central Composite Design.

EB . . . . . Empirical Bayes.

List of Notations

$\propto$	Proportional to.
$\mathbb{R}$	The set of real numbers.
$\mathbb{Z}$	The set of integers.
$\mathbb{Z}^+$	The set of positive integers.
ρ	HIV prevalence.
λ	HIV incidence.
α	ART coverage.
$\mathcal{S}$	Spatial study region $\mathcal{S} \subseteq \mathbb{R}^2$.
$s \in \mathcal{S}$	Point location.
$\mathcal{T}$	Temporal study period $\mathcal{T} \subseteq \mathbb{R}$.
$t \in \mathcal{T}$	Time.
$\mathbf{y}$	Data, a n-vector $(y_1, \ldots, y_n)$.
$\boldsymbol{\phi}$	Parameters, a d-vector $(\phi_1, \ldots, \phi_d)$.
$\mathbf{x}$	Latent field, a N-vector $(x_1, \ldots, x_N)$.
$\boldsymbol{\theta}$	Hyperparameters, a m-vector $(\theta_1, \ldots, \theta_m)$.
$x \sim p(x)$	x has the probability distribution $p(x)$.
A_i	Areal unit.
$A_i \sim A_j$	Adjacency between areal units.
$\mathbf{H}$	Hessian matrix.
$\mathbf{R}$	Structure matrix.
$\mathbf{Q}$	Precision matrix.
$\mathbf{\Sigma}$	Covariance matrix.
$\mathcal{N}$	Gaussian distribution.
$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$	Kernel function on the space $\mathcal{X}$.
$A_i \sim A_j$	Adjacency between areal units.

$\mathcal{Q}$	A set of quadrature nodes.
$\omega : \mathcal{Q} \to \mathbb{R}$ . . .	A quadrature weighting function.
$\mathcal{Q}(m, k)$	Gauss-Hermite quadrature points in m dimensions with k nodes per dimension, constructed according to a product rule.
φ	A standard (multivariate) Gaussian density.

1 Introduction

This thesis is about applied and methodological Bayesian statistics. It is applied and methodological in that I am concerned with real world questions and the means to answer them. The statistical approach is Bayesian because I use probability theory to arrive at conclusions based on models for observed data.

The applied focus of this thesis is in obtaining the strategic information needed to plan the response to the human immunodeficiency virus (HIV) epidemic in sub-Saharan Africa (SSA). Over 40 years since the beginning of the epidemic, HIV is the largest annual cause of disability adjusted life years (DALYs) in SSA among non-infants [Global Burden of Disease Collaborative Network (2019); Figure 1.1]. Quantification of the epidemic using statistics is an important part of the public health response. Effective implementation of HIV prevention and treatment requires strategic information. However, producing suitable estimates of relevant indicators is made difficult by a range of statistical challenges.

The data I use were gathered in national household surveys or routinely collected from healthcare facilities providing HIV services. An important feature of these data are the location and time at which each observation was recorded. Spatio-temporal data occur across a range of application settings. While diverse in setting, they have distinctive recurring commonalities which make their collective study worthwhile.

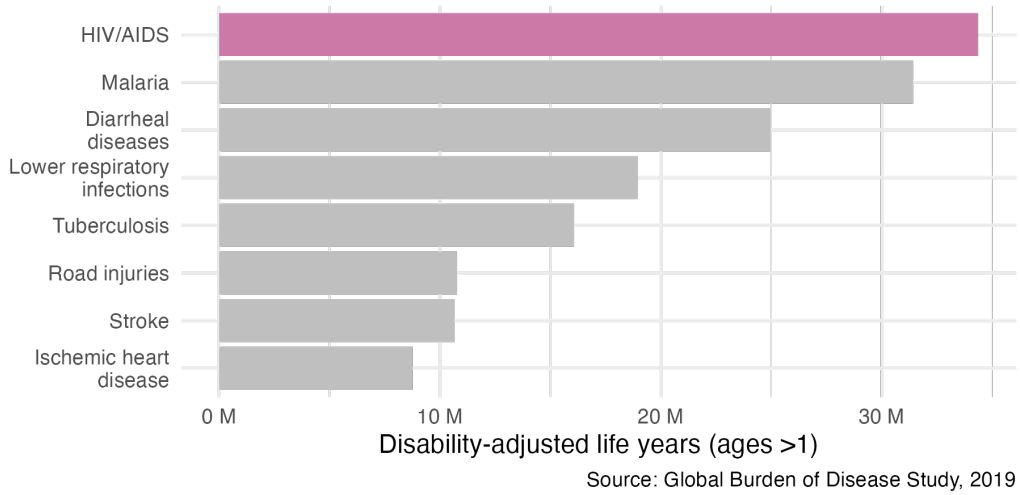

Figure 1.1: HIV/AIDS is the largest cause of annual DALYs among individuals aged >1 year in SSA (Global Burden of Disease Collaborative Network 2019). One DALY represents the loss of the equivalent of one year of full health, and is calculated by the sum of years of life lost and years lost due to disability. Weights used to account for disability vary between 0 (full health) and 1 (death) depending on severity of the condition.

The work conducted in this thesis makes use of, and aspires to contribute to, techniques from spatio-temporal statistics.

Computation is an essential part of modern statistical practice. Each project in this thesis, and the thesis itself, is accompanied by R (R Core Team 2022) code, hosted on GitHub at `https://github.com/athowes`.

1.1 Chapter overview

This thesis is structured as follows:

- Chapter 2 provides an overview of the HIV/AIDS epidemic, and describing the challenges faced by disease surveillance efforts.
- Chapter 3 introduces the statistical concepts and notation used throughout the thesis, focusing on Bayesian modelling and computation, spatio-temporal statistics, and survey methods.
- Chapter 4: The prevailing model for spatial structure used in small-area estimation (Besag et al. 1991) was intended to analyse a grid of pixels. In disease mapping, we work using the districts of a country, which are typically

not a grid. I evaluated the practical consequences of this concern (Howes, Eaton, et al. 2023+).

- Chapter 5: Adolescent girls and young women are a demographic group at disproportionate risk of acquiring HIV infection. The Global AIDS Strategy recommends prioritising interventions on the basis of behaviour to prevent the most new infections using available resources. I estimated the size of behavioural risk groups across priority countries to enable implementation of this strategy, and assessed the potential benefits in terms of numbers of new infections prevented (Howes, Risher, et al. 2023). This work was included in the UNAIDS Global AIDS Update 2022 and 2023.
- Chapter 6: The Naomi small-area estimation model (Eaton et al. 2021) is used by countries to estimate district-level HIV indicators. First, I implemented the integrated nested Laplace approximations using automatic differentiation, such that the method is compatible with Naomi, opening the door to a new class of fast, flexible, and accurate Bayesian inference algorithms. Second, I developed an approximate Bayesian inference method combining adaptive Gauss-Hermite quadrature with principal components analysis (Howes, Stringer, et al. 2023+). I applied these method to data from Malawi, and analysed the consequences of inference method choice for policy relevant outcomes.
- Chapter 7: Finally, I discuss avenues for future work, and my conclusions regarding the research, as well as its strengths and weaknesses.

Though chronological order is recommended, Chapters 4, 5 and 6 may be read in any order.

2 The HIV/AIDS epidemic

2.1 Background

HIV is a retrovirus which infects humans. If untreated, HIV can develop into a more advanced stage known as acquired immunodeficiency syndrome (AIDS). HIV primarily attacks a type of white blood cell vital for the function of the immune system. As a result, AIDS is characterised by increased risk of developing opportunistic infections such as tuberculosis or *Pneumocystis* pneumonias, which can result in death.

The first AIDS cases were reported in Los Angeles in the early 1980s (Gottlieb et al. 1981; Barré-Sinoussi et al. 1983). Since then, HIV has spread globally. Transmission occurs by exposure to specific bodily fluids of an infected person. The most common mode of transmission is via unprotected anal or vaginal sex, though transmission can also occur from a mother to her baby, or when drug injection equipment is shared. Approximately 86 million people have become infected with HIV, and of those 40 million have died of AIDS-related causes.

An ongoing global and multifaceted effort has been made to respond to the epidemic. The response has been shaped by local communities, civil society organisations, national governments, research institutions, pharmaceutical companies, international agencies like the Joint United Nations Programme on HIV/AIDS

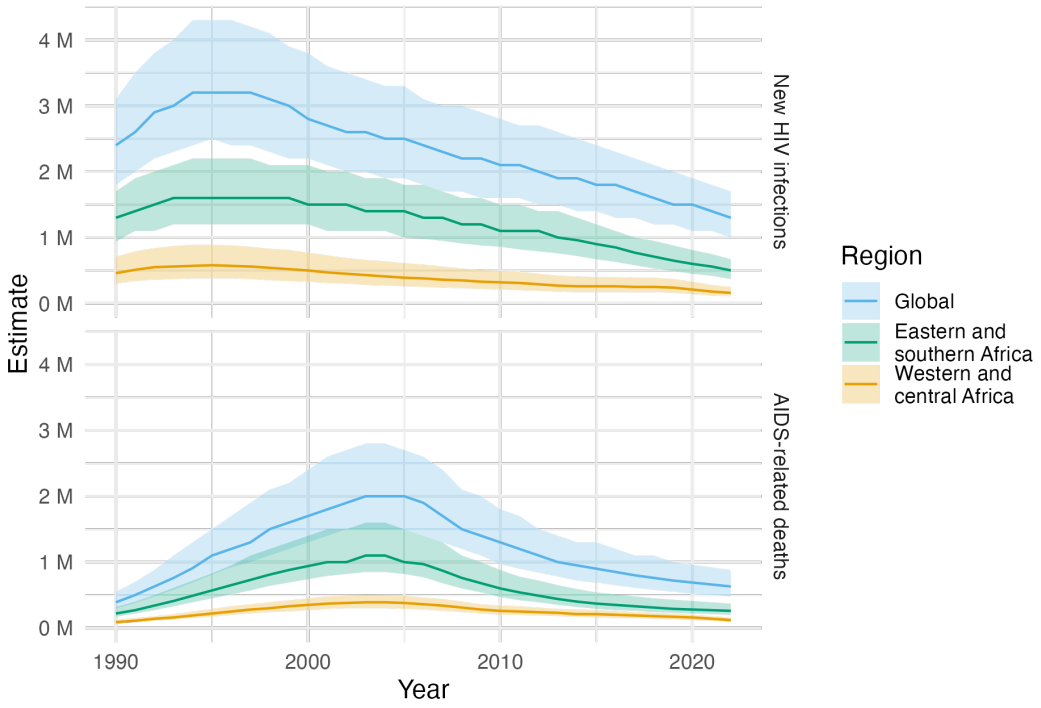

Figure 2.1: Globally, yearly new HIV infections peaked in 1995, and have since decreased by 59% and yearly AIDS-related deaths peaked in 2004, and have since decreased by 68% (UNAIDS 2023a). Much of the disease burden is concentrated in eastern and southern Africa, as well as western and central Africa.

(UNAIDS), and global health initiatives such like the President's Emergency Plan for AIDS Relief (PEPFAR) and the Global Fund to Fight AIDS, Tuberculosis, and Malaria (the Global Fund). To give an indication as to the scale of the response, the investment of $100 billion by PEPFAR constitutes the "largest commitment by a single nation to address a single disease in history" (U.S. Department of State 2022).

Implementation of HIV prevention and treatment has significantly reduced the number of new HIV infections and AIDS-related deaths per year since their peak (Figure 2.1). The most significant evidence-based interventions, in chronological order of their introduction, are described below:

- Condoms are an inexpensive and effective method for prevention of HIV and other sexually transmitted infections (STIs) such as *Chlamydia trachomatis*, *Neisseria gonorrhoeae*, syphilis, and *Trichomonas vaginalis*. Condom usage has increased significantly since 1990, which is estimated to have averted 117

million new HIV infections (Stover and Teng 2021). There remain significant but difficult to close gaps in condom usage.

- Antiretroviral therapy (ART) is a combination of drugs which stop the virus from replicating in the body. A person living with HIV who takes ART daily can live a full and healthy life, transforming what was once a terminal illness to a treatable chronic condition. Of the 39 million people living with HIV (PLHIV) in 2022, around 76% were accessing ART. The number of AIDS-related deaths, 21 million, estimated to have been averted by ART is staggering (UNAIDS 2023b).

  ART reduces the amount of virus in the blood and genital secretions. If the virus is undetectable then there is significant evidence that it cannot be transmitted sexually (Cohen et al. 2011; Broyles et al. 2023). For this reason, in addition to providing life saving treatment, ART also operates as prevention. Approaches to lowering risk of HIV transmission using treatment are referred to as treatment as prevention (TaSP). Particular efforts have been made to provide pregnant women with ART to reduce the chance of mother-to-child transmission (MTCT) (Siegfried et al. 2011).

- Voluntary medical male circumcision (VMMC) partially protects against female-to-male HIV acquisition. Three landmark randomised control trials (Auvert et al. 2005; Gray et al. 2007; Bailey et al. 2007) found complete surgical removal of the foreskin to result a reduction of HIV acquisition in men by 50-60%. Based on this evidence, VMMC has been recommended since 2007 by the World Health Organization (WHO) and UNAIDS as a key HIV intervention in high-prevalence settings. Scale up of VMMC across 15 priority countries between 2008 and 2019 is estimated to have already averted 340 thousand new HIV infections, though the future number of new HIV infections averted is likely to be much higher.

- Pre-exposure prophylaxis (PrEP) and post-exposure prophylaxis (PEP) are antiretrovial drugs which can be taken before and after exposure to prevent

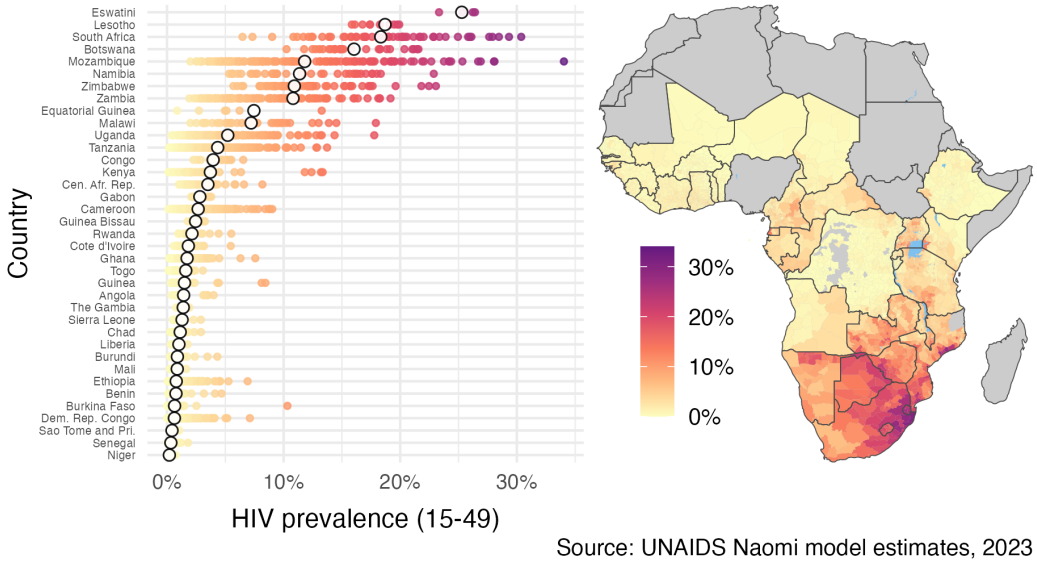

Figure 2.2: Adult (15-49) HIV prevalence varies substantially both within and between countries in SSA. The estimates from 2023 were generated by country teams using the Naomi small-area estimation model in a process supported by UNAIDS, and are available from UNAIDS (2023a). White filled points are country-level estimates, and coloured points are district-level estimates. Results from Nigeria were not published. Data collection in the Cabo Delgado province of Mozambique was disrupted by conflict. Obtaining results for the Democratic Republic of the Congo required removing some districts from the model.

transmission. PrEP and PEP are more costly than some other prevention options, so primarily useful in high risk settings.

Though important progress had been made, facilitated by the interventions above, there remains much more to do. In 2022, 1.3 million people were newly infected with HIV and there were 630 thousand AIDS-related deaths, more than one every minute (UNAIDS 2022). Bold fast-track targets have been set to accelerate the end of AIDS as global public health threat by 2030. To meet these targets in the context of disruption to HIV services caused by the COVID-19 pandemic and a potential shortfall in HIV funding, renewed commitments are required (Economist Impact 2023).

For available resources to have the greatest impact, it is important that the right HIV interventions to be prioritised to the right populations, in the right place, and at the right time. This paradigm has been termed precision public health (Khoury et al. 2016), by analogy to precision medicine. While precision medicine tailors treatment to the individual, precision public health tailors treatment to the population. Differences in the cost-effectiveness of any given intervention can be vast, with some interventions orders of magnitude more impactful than others (Ord 2013).

Disease burden varies substantially across multiple spatial scales. In some countries, the epidemic is concentrated in small populations, and national HIV prevalence is low. In others, the epidemic is sustained by heterosexual transmission, and national HIV prevalence is higher (typically >1%) These two epidemic settings are sometimes described as concentrated and generalised, respectively. Most of the countries severely affected by HIV are in sub-Saharan Africa (SSA). It is estimated that 66% of the 39 million PLHIV worldwide live in SSA. Adult HIV prevalence (ages 15-49) is above than 10% (Figure 2.2) in some countries in southern Africa, with some districts even exceeding 20%. Just as there is variation between countries, there is variation within countries. For example, adult HIV prevalence at the district municipality level in South Africa ranges from 6% in Namakwa to 30% in uMkhanyakude.

In all countries and contexts, some groups of people are at much higher risk than others. Groups of people at increased risk of HIV infection are known as key populations (KPs). Examples include men who have sex with men (MSM), female sex workers (FSW), people who inject drugs (PWID), and transgender people (TGP) (Stevens, Sabin, Garcia, et al. 2022). KPs are often marginalised, and face legal and social barriers. Concentrated settings are defined by the majority of new HIV infections occurring in KPs and their sexual partners. In generalised settings like SSA, risk is more diffuse across the population. For example, in SSA adolescent girls and young women (AGYW) are a large demographic group at increased risk of HIV infection (Risher et al. 2021; Monod et al. 2023) but not

typically considered a KP. That said, concentrated subepidemics often occur within what are nominally generalised settings (Tanser et al. 2014).

Various methods can be used to implement differentiatated HIV treatment and prevention services. These include geographic and demographic prioritisation, key population services, and risk screening based on individual-level risk characteristics.

2.2 HIV surveillance

HIV surveillance refers to the collection, analysis, interpretation and dissemination of data relating to HIV (Pisani et al. 2003). Surveillance can used to track epidemic indicators, identify at-risk populations, find drivers of transmission, and evaluate the impact of prevention and treatment programs. Important indicators include:

- **HIV prevalence** is the proportion $\rho \in [0, 1]$ of a population who have HIV. The number of PLHIV is given by $N\rho$, where N is the (living) population size. Increases in HIV prevalence, or the number of PLHIV, can be caused by either new HIV infections or more PLHIV remaining alive by taking treatment. For this reason care must be taken in directly interpreting changes in HIV prevalence. As a key measure of population disease burden, HIV prevalence is used to calculate all of the other indicators given below.

- **HIV incidence** is the rate $\lambda > 0$ of new HIV infections, often written as number of new infections per 1000 person years. The number of new HIV infections that occur during a given time is the integral of HIV incidence multiplied by the size of the susceptible population. Let ρ_t be the HIV prevalence, and N_t be the population size, at time t. Then the number of new HIV infections which occur in a given period of time are given by

$$I = \int \lambda_t \cdot (1 - \rho_t) \cdot N_t \mathrm{d}t.$$

  Planning, delivery, and evaluation of prevention programming relies on estimates of HIV incidence and the number of new HIV infections.

- **ART coverage** is the proportion $\alpha \in [0, 1]$ of PLHIV who are on ART. Estimates of ART coverage play a direct role in planning the provision of treatment services. The number of people taking ART is given by $N \cdot \rho \cdot \alpha$.

- **Recent infection** is the proportion $\kappa \in [0, 1]$ of PLHIV who have been recently infected. Recent infection can be used to help estimate HIV incidence.

- **Awareness of status** is the proportion $\xi \in [0, 1]$ of PLHIV who have been diagnosed with HIV. Programming of HIV testing and diagnosis is informed by awareness of status.

2.2.1 Data

Data are used to estimate the above HIV indicators, in conjunction with scientific knowledge. Multiple sources of data are used:

- **Household surveys** are large, national cross-sectional studies. The Demographic and Health Surveys (DHS) Program. Population-based HIV Impact Assessments (PHIA). Household surveys provide nationally-representative high quality standardised data about HIV.

- **Programmatic data** refer to data routinely collected during delivery of health services. Examples include data from antenatal care (ANC) HIV testing and ART service delivery. Programmatic data are more regularly available than other data sources. However, the control that can be exercised over collection of programmatic data is limited. As a result, issues of data quality and reliability, as well as bias, are common in working with programmatic data.

- **Cohort studies** follow a group of people over time. Outcomes may be measured more systematically in a cohort study than in other study designs. The data from cohort studies are used to inform otherwise difficult to estimate epidemiological parameters. Examples of such parameters include disease

progression and mortality rates, transmission dynamics, and treatment outcomes. Population-based cohort studies relevant to the SSA setting include Manicaland, Zimbabwe (Gregson et al. 2006); Rakai, Uganda (Grabowski et al. 2017).

2.2.2 Challenges

Obtaining reliable, timely estimates at an appropriate spatial resolution is challenging. The most significant difficulties faced are:

1. **Data sparsity**: Collection of data is costly and time consuming. As a result, limited direct data might be available for the particular time, location, or population of interest. For example, in many countries the last conducted household survey is several years out of date.

2. **Missing data**: The sampling frame of a survey may not correspond to the target population. For example, many KPs are difficult to reach, and may be omitted from sampling frames. Individuals included on the sampling frame may choose not to respond. All surveys are subject to sampling error, as only a subset of the target population are sampled. Each of these issues can be characterised as being problems of missing data. I characterise missing data as referring to the shortfalls of any given study, and data sparsity as referring to limited availability of studies.

3. **Response and measurement biases**: Individuals may be hesitant to disclose their HIV status, or report higher risk behaviours, due to social desirability bias or a fear of discrimination or stigma. When available, biomarker data can be used to overcome under-reporting, but still may be subject to measurement errors.

4. **Denominators and demography**: Many indicators are rates or proportions, which rely on estimates of the population at risk in the denominator. Accurately estimating population denominators is itself a challenging task

(Tatem 2017). Taking a ratio of uncertain quantities amplifies uncertainty, but is rarely properly accounted for.

5. **Inconsistent data collection and reporting**: The types of data that are collected might vary across space and time. Reporting protocols or definitions can also change.

6. **Reliance on epidemiological parameters**: Indicators rely on estimates of epidemiological parameters such as rates of disease progression. These parameters may not generalise to the setting of interest. Further, they are typically applied coarsely, and without proper accounting for uncertainty.

2.2.3 Statistical approaches

The challenges above make direct interpretation of the data often misleading or impossible. Careful statistical modelling is required to overcome these limitations as best as possible.

1. **Borrowing information**: When little direct data are available, data judged to be indirectly related can be used to help improve estimation. For example, if limited data are available for individuals of a certain age, it is likely reasonable to make use of data for individuals of a similar age. As well as over age groups, information can be borrowed between and within countries, and across times.

2. **Evidence synthesis**: Multiple sources of evidence can be combined to overcome the limitations of any one data source. For example, infrequently run household surveys can be complemented by more up-to-date programmatic data.

3. **Expert guidance**: Expert epidemiological, demographic, and local stakeholder guidance can be used to improve estimates. Ensuring the quality of any data used in the estimation process is essential.

4. **Uncertainty quantification**: Conclusions drawn by synthesising multiple incomplete data sources are unlikely to be firm and unanimous. It is therefore particularly that the uncertainties inherent to any statistical analysis are accurately and transparently presented.

3

Bayesian spatio-temporal statistics

3.1 Bayesian statistics

Bayesian statistics is a mathematical paradigm for learning from data. It is especially well suited to facing the challenges presented in Section 2.2 for the following reasons. First, because it allows for principled and flexible integration of prior domain knowledge. Second, because uncertainty over all unknown quantities is handled as an integral part of the Bayesian paradigm. This section provides a brief and at times opinionated overview of Bayesian statistics. For a more complete introduction, I recommend Gelman, Carlin, et al. (2013), McElreath (2020) or Gelman, Vehtari, et al. (2020).

3.1.1 Bayesian modelling

The Bayesian approach to data analysis is based on construction of a probability model for the observed data $\mathbf{y} = (y_1, \ldots, y_n)$. Parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_d)$ are used to describe features of the data. Both the data and parameters are assumed to be random variables, with joint probability distribution written as $p(\mathbf{y}, \boldsymbol{\phi})$. Subsequent calculations, and the conclusions to follow, are made by manipulating the model using probability theory.

Models are most naturally constructed from two parts known as the likelihood $p(\mathbf{y} \mid \boldsymbol{\phi})$ and the prior distribution $p(\boldsymbol{\phi})$. The joint distribution is obtained by the product of these two parts

$$p(\mathbf{y}, \boldsymbol{\phi}) = p(\mathbf{y} \mid \boldsymbol{\phi}) p(\boldsymbol{\phi}). \tag{3.1}$$

The likelihood, as a function of $\boldsymbol{\phi}$ with $\mathbf{y}$ fixed, reflects the probability of observing the data when the value of the parameters is $\boldsymbol{\phi}$. The prior distribution encapsulates beliefs about the parameters $\boldsymbol{\phi}$ before the data are observed.

Recommendations for specifying prior distributions vary. A central issue is the extent to which subjective information should be incorporated into the prior distribution. Proponents of the objective Bayesian paradigm (Berger 2006) put forward that the prior distribution should be non-informative, so as not to introduce subjectivity into the analysis. Others see subjectivity as fundamental to scientific inquiry, with no viable alternative (Goldstein 2006). Though subjectivity typically discussed with regard to the prior distribution, we shall see in Section **??** that the distinction between prior distribution and likelihood is not always clear. As such, it may be argued that issues of subjectivity are not unique to prior distribution specification, and ultimately that the challenge of specifying the data generating process is better thought of more holistically (Gelman, Simpson, et al. 2017).

The probability model can be simulated from to obtain samples $(\mathbf{y}, \boldsymbol{\phi}) \sim p(\mathbf{y}, \boldsymbol{\phi})$. If the samples of the data $\mathbf{y}$ differ too greatly from what the analyst would expect to see in reality, then the model does not capture their prior scientific understanding. Models which do not produce plausible data samples can be refined. Checks of this kind [Gelman, Carlin, et al. (2013); Chapter 6] can be used to help iteratively build models, gradually adding complexity as required.

3.1.2 Bayesian computation

Having constructed a model (Equation (3.1)), the primary goal in a Bayesian analysis is to obtain the posterior distribution $p(\boldsymbol{\phi} \mid \mathbf{y})$. This distribution encapsulates probabilistic beliefs about the parameters given the observed data. As such,

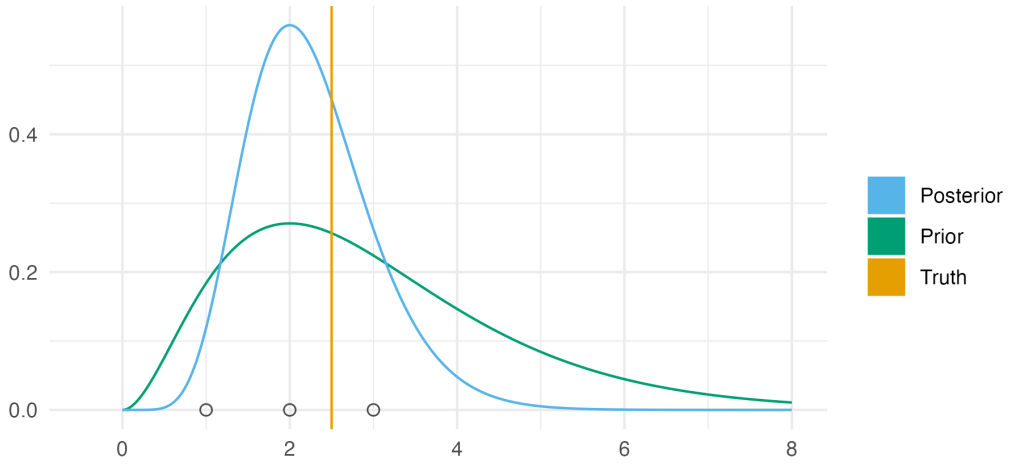

Figure 3.1: An example of Bayesian modelling and computation for a simple one parameter model. Here the likelihood is $y_i \sim \text{Poisson}(\phi)$ for $i = 1, 2, 3$ and prior distribution on the rate parameter $\phi > 0$ is $\phi \sim \text{Gamma}(3, 1)$. Observed data $\mathbf{y} = (1, 2, 3)$ was simulated from the distribution Poisson(2.5). The true data generating process is within the space of models being considered. (This situation is sometimes known (Bernardo and Smith 2001) as the $\mathcal{M}$-closed world, in contrast to the $\mathcal{M}$-open world where the model is said to be misspecified.) Furthermore, the posterior distribution is available in closed form as $\text{Gamma}(9, 4)$. This is because the posterior distribution is in the same family of probability distributions as the prior distribution. Models of this kind are described as being conjugate. Conjugate models are often used because of their convenience. Though other models may be more suitable, they will typically be more computationally demanding. The posterior distribution here is more tightly peaked than the prior distribution. This contraction is typically, but not always, the case.

the posterior distribution has a central role in use of the statistical analysis for decision making.

Using the eponymous Bayes' theorem, the posterior distribution is obtained by

$$p(\boldsymbol{\phi} \,|\, \mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\phi})}{p(\mathbf{y})} = \frac{p(\mathbf{y} \,|\, \boldsymbol{\phi}) p(\boldsymbol{\phi})}{p(\mathbf{y})}. \tag{3.2}$$

Unfortunately, most of the time it is intractable to calculate the posterior distribution analytically. This is because of the potentially high-dimensional integral

$$p(\mathbf{y}) = \int p(\mathbf{y}, \boldsymbol{\phi}) \mathrm{d}\boldsymbol{\phi} \tag{3.3}$$

in the denominator of Equation (3.2). The result of this integral is known as the evidence $p(\mathbf{y})$, and quantifies the probability of obtaining the data under the model. Hence, although it is easy to evaluate a quantity proportional to

the posterior distribution

$$p(\boldsymbol{\phi} \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \boldsymbol{\phi}) p(\boldsymbol{\phi}), \tag{3.4}$$

it is typically difficult to evaluate the posterior distribution itself.

The difficulty in performing Bayesian inference may be thought of as analogous to the difficulty in calculating integrals. As with integration, in specific cases closed form analytic solutions are available. Figure 3.1 illustrates one such case, where the prior distribution and posterior distribution are in the same family of probability distributions. In the more general case no analytic solution is available, and computational methods must be relied on. Computational strategies for approximating the posterior distribution (Martin et al. 2023) may broadly be divided into Monte Carlo algorithms and deterministic approximations.

Monte Carlo algorithms

Monte Carlo algorithms (Robert and Casella 2005) aim to generate samples from the posterior distribution

$$\boldsymbol{\phi}_i \sim p(\boldsymbol{\phi} \,|\, \mathbf{y}), \quad i \in 1, \dots M. \tag{3.5}$$

These samples may be used in any future computations involving functions of the posterior distribution. For example, if $G = G(\boldsymbol{\phi})$ then the expectation of G with respect to the posterior distribution can be approximated by

$$\mathbb{E}(G \,|\, \mathbf{y}) = \int G(\boldsymbol{\phi}) p(\boldsymbol{\phi} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\phi} \approx \frac{1}{M} \sum_{i=1}^{M} G(\boldsymbol{\phi}_i). \tag{3.6}$$

Most quantities of interest can be cast as posterior expectations which can then be approximated empirically using samples in this way.

Markov chain Monte Carlo (MCMC) methods (Roberts and Rosenthal 2004) are the most popular class of sampling algorithms. Using MCMC, samples are generated from by simulating from an ergodic Markov chain with the posterior distribution as its stationary distribution. The Metropolis-Hastings [MH; Metropolis et al. (1953); Hastings (1970)] algorithm uses a proposal distribution $q(\boldsymbol{\phi}_{i+1} \,|\, \boldsymbol{\phi}_i)$

to generate candidate parameters for the next step in the Markov chain. Many MCMC algorithms, including the Gibbs sampler (Geman and Geman 1984), are special cases of MH.

Other notable classes of sampling algorithms include importance sampling [IS; Tokdar and Kass (2010)] methods in which the samples are weighted, sequential Monte Carlo [SMC; Chopin, Papaspiliopoulos, et al. (2020)] methods based on sampling from a sequence of distributions, and approximate Bayesian computation [ABC; Sisson et al. (2018)] Though these methods have found applications in specific domains, MCMC is currently more widely used because of its generality, theoretical reliability, as well as benefiting from more accessible software implementations.

This thesis makes use of the No-U-Turn sampler [NUTS; Hoffman, Gelman, et al. (2014)], a Hamiltonian Monte Carlo [HMC; Duane et al. (1987); Neal et al. (2011)] algorithm, as implemented in the `Stan` (Carpenter et al. 2017) probabilistic programming language (PPL). HMC uses derivatives of the posterior distribution to generate efficient MH proposal distributions based on Hamiltonian dynamics. NUTS automatically adapts the tuning parameters of HMC based local properties of the posterior distribution. Though not a one-size-fits-all solution, NUTS has been shown empirically to be a good choice for sampling from a range of posterior distributions. Figure 3.2 shows an example of using the NUTS MCMC algorithm to sample from a posterior distribution.

After running an MCMC sampler, it is important to check diagnostics to evaluate convergence and assess whether the results of the Markov chain can be used to accurately compute posterior quantities. Panel B of Figure 3.2 shows a traceplot for a Markov chain which has converged. A wide range of convergence diagnostics have been developed for MCMC (Roy 2020).

Deterministic approximations

The Monte Carlo algorithms discussed in Section 3.1.2 make use of stochasticity to generate samples from the posterior distribution. Deterministic approximations are as follows.

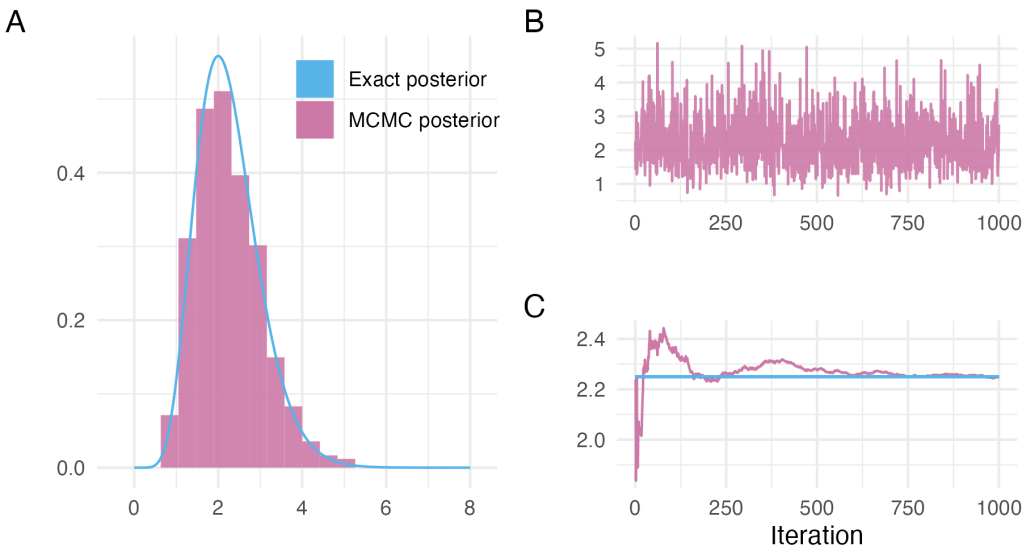

Figure 3.2: NUTS can be used to sample from the posterior distribution in the example of Figure 3.1. Panel A shows a histogram of the NUTS samples as compared to the true posterior. The visual appearance of a histogram depends highly on the number of bins chosen. Other visualisations, such as empirical cumulative difference function plots, though less initially intuitive, are preferred for accurate distributinal sample comparisons. Panel B is a traceplot showing the path of the Markov chain $\{\phi_i\}_{i=1}^{1000}$ as it explores the posterior distribution. In this case, the Markov chain moves freely throughout the posterior distribution, without getting stuck in any one location for long, indicating good performance of the sampler. Panel C shows the convergence of the empirical posterior mean $\frac{1}{i}\sum_{j\leq i}\phi_j$ to the true value of $\mathbb{E}(\phi)$ as more iterations of the Markov chain are included in the calculation.

The Laplace approximation involves approximating the posterior distribution by a Gaussian distribution. The integrated nested Laplace approximation [INLA; Rue, Martino, and Chopin (2009)] combines quadrature with the Laplace approximation. The Laplace approximation and INLA are used extensively throughout this thesis. A complete introduction is provided in Section 6.1.

Another prominent determinisitc approach is variational inference [VI; Blei et al. (2017)], in which the approximate posterior distribution is assumed to belong to a particular family of functions. Optimisation algorithms are then used to choose the best member of that family, typically by minimising the Kullback-Leibler divergence to the posterior distribution. VI is often faster than Monte Carlo methods, especially for large datasets or models. However, it lacks theoretical guarantees and is known to often inaccurately estimate posterior variances (Giordano et al. 2018). Developing

diagnostics to evaluate the accuracy of VI is an important area of ongoing research (Yao et al. 2018). The well-known expectation maximisation [EM; "Maximum likelihood from incomplete data via the EM algorithm" (1977)] and expectation propagation [EP; Minka (2001)] algorithms are closely related to VI.

3.1.3 Interplay between modelling and computation

Modern computational techniques and software like PPLs have succeeded in abstracting away calculation of the posterior distribution from the analyst for many models. However, computation remains intractable in arguably the majority of cases. The analyst need therefore not only to be concerned with choosing a model suitable for the data, but also choosing a model for which the posterior distribution is tractable in reasonable time. As such, there is an important interplay between modelling and computation, wherein models are bound by the limits of computation. As computation improves, the space of models available to the analyst expands.

3.2 Spatio-temporal statistics {st-statistics}

Spatio-temporal statistics (Cressie and Wikle 2015) concerns observations which are indexed by spatial or temporal location. In doing so, it unites the fields of spatial statistics (Bivand et al. 2008) and time series analysis (Shumway and Stoffer 2017).

3.2.1 Properties of spatio-temporal data

Spatio-temporal data have some important properties:

1. **Covariance structure**: According to Tobler's first law of geography "everything is related to everything else, but near things are more related than distant things" (Tobler 1970). In "The Design of Experiments" Fisher (1936) observed that neighbouring crops were more likely to have similar yields than those far apart. This law can be formalised using spatial covariance functions. Spatial covariance functions are called isotropic when they apply equally in all directions, and stationary when they are invariant over space.

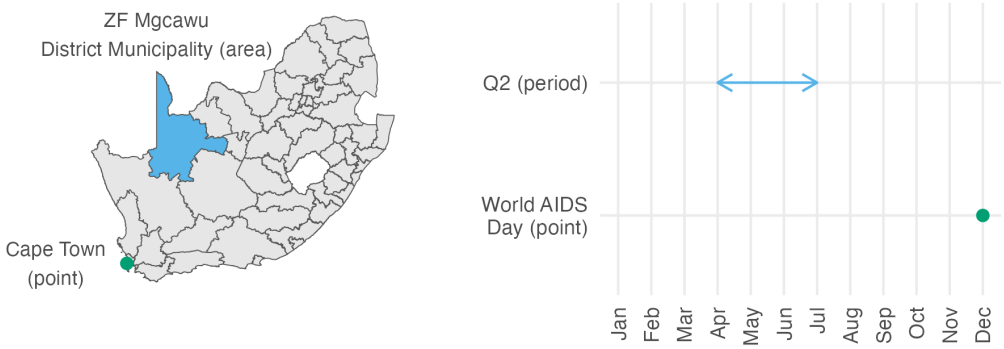

Figure 3.3: The spatial location of Cape Town in South Africa could be considered a point. The ZF Mgcawu District Municipality on the other hand is an example of an area. World AIDS Day, designated on the 1st of December every year, could be considered a point in time. The second fiscal quarter, running through April, May and June, and denoted by Q2 represents a period of time. (In reality, both Cape Town and World AIDS Day are areas, rather than true point locations. Instances of infinitesimal point locations in everyday life are rare.)

As well as space, Tobler's first law applies to time. Observations made close together in time tend to be similar. Temporal covariance structures are often periodic.

The space-time covariance structure (Porcu et al. 2021) is said to be separable when it can be factorised as a product of individual spatial and temporal covariances, and nonseparable when it can't. A separable space-time covariance could have spatial and temporal components which are either independent and identically distributed (IID) or structured (Knorr-Held 2000).

Because of their covariance structure, spatio-temporal data are not IID. Only one observation of a spatio-temporal process is realised.

2. **Scales**: In this thesis the spatial study region $\mathcal{S} \subseteq \mathbb{R}^2$ is assumed to have two dimensions, corresponding to latitude and longitude. Observations may be associated to a point $s \in \mathcal{S}$ or area $A \subseteq \mathcal{S}$ in the study region. The temporal study period $\mathcal{T} \subseteq \mathbb{R}$ can more generally be assumed to be on-dimensional. Together with time moving only in the forward direction, this feature is what distinguishes time from space. As with space, observations may be associated

to a point $t \in \mathcal{T}$ or period of time $T \subseteq \mathcal{T}$. Figure 3.3 illustrates both types of observation for space and time.

Spatio-temporal observations can be made at various possible scales. Sometimes, we may want to model data at a scale it was not observed at. This is known as the change-of-support problem (Gelfand et al. 2001) and includes as special cases the problems of downscaling, upscaling, and dealing with so-called misaligned data. Closely related is the problem of jointly modelling data at different scales simultaneously.

3. **Size**: Data with both spatial and temporal dimensions are often large, making storage and operations on spatio-temporal data potentially difficult. Furthermore, models for spatio-temporal data typically require many parameters. Whereas large IID data can be modelled using a small number of parameters, each observation in a spatio-temporal dataset may need to be characterised by its own parameters. Large data combined with large models make Bayesian inference challenging.

3.2.2 Small-area estimation

Data always has some cost to collect. This cost can be significant and prohibitive, especially for data relating to people where collection is difficult to automate. As a result, given the large number of possible locations in space and time, often no or limited direct observations may be available for any given space-time location. Direct estimates of indicators of interest are either impossible or inaccurate in this setting.

Small-area estimation [SAE; Pfeffermann et al. (2013)] methods aim to overcome the limitations of small data by sharing information. In the spatio-temporal setting sharing of information occurs across space and time. The knowledge that observations in one spatio-temporal location are correlated with those at another can be used to improve estimates. Figures 3.4 and 3.5 illustrate the unreliability of direct estimates from small sample sizes, as well as the way in which a spatial model may be used to overcome this limitation in part.

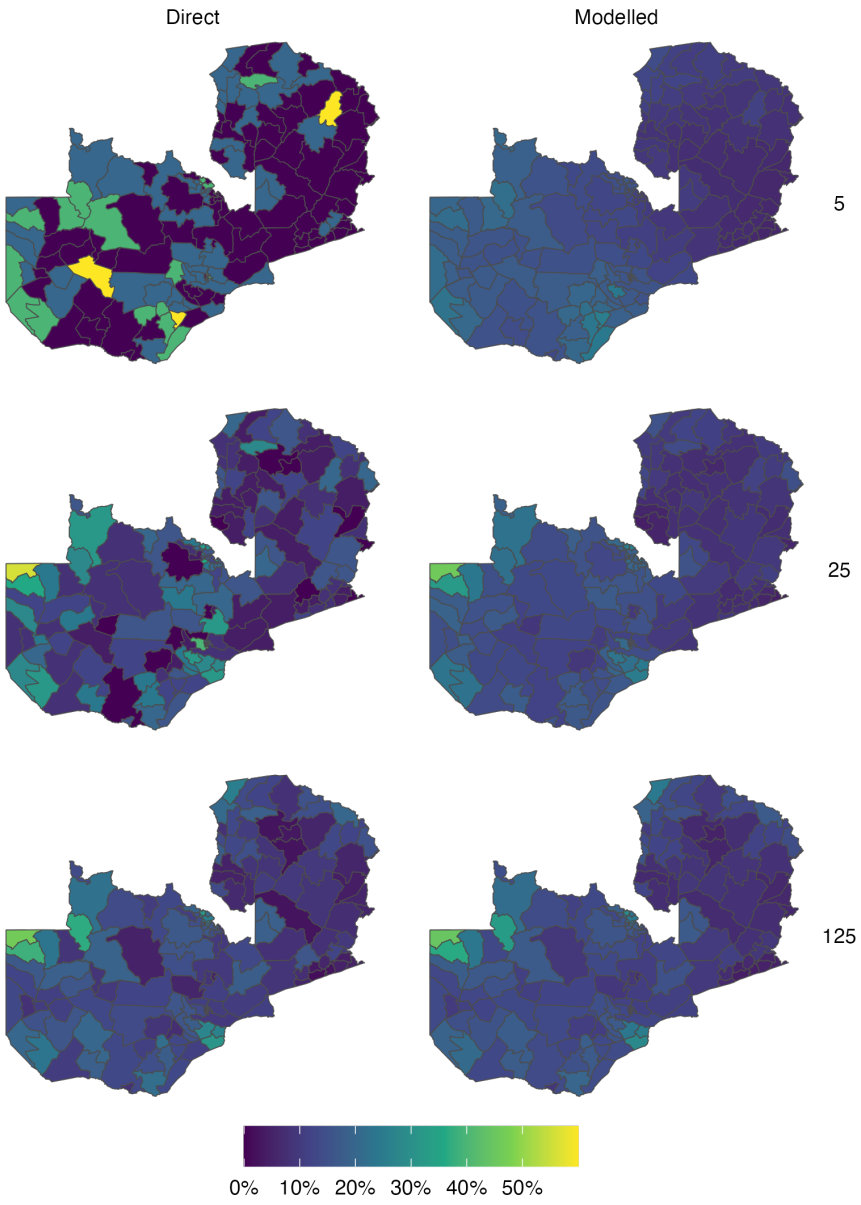

Figure 3.4: Simulation of a simple random sample $y_i \sim \text{Bin}(m, p_i)$ with varying sample size $m = 5, 25, 125$ in each of the $i = 1, \ldots, 156$ constituencies of Zambia. Direct estimates were obtained by the empirical ratio of data to sample size. Modelled estimates were obtained using a logistic regression with linear predictor given by an intercept and a spatial random effect. The colour palette used in this figure is called viridis, as implemented by the `viridis` R package (Garnier et al. 2023), and was designed to be perceptually uniform and accessible to colourblind viewers (Smith and van der Walt 2015). This figure was adapted from a presentation given for the Zambia HIV Estimates Technical Working Group, available from `https://github.com/athowes/zambia-unaids`. Estimates of HIV indicators for Zambia have previously been generated at the district-level, comprising 116 spatial units. Moving forward, there is interest in generating estimates at the higher-resolution constituency level, as program planning is devolved locally.

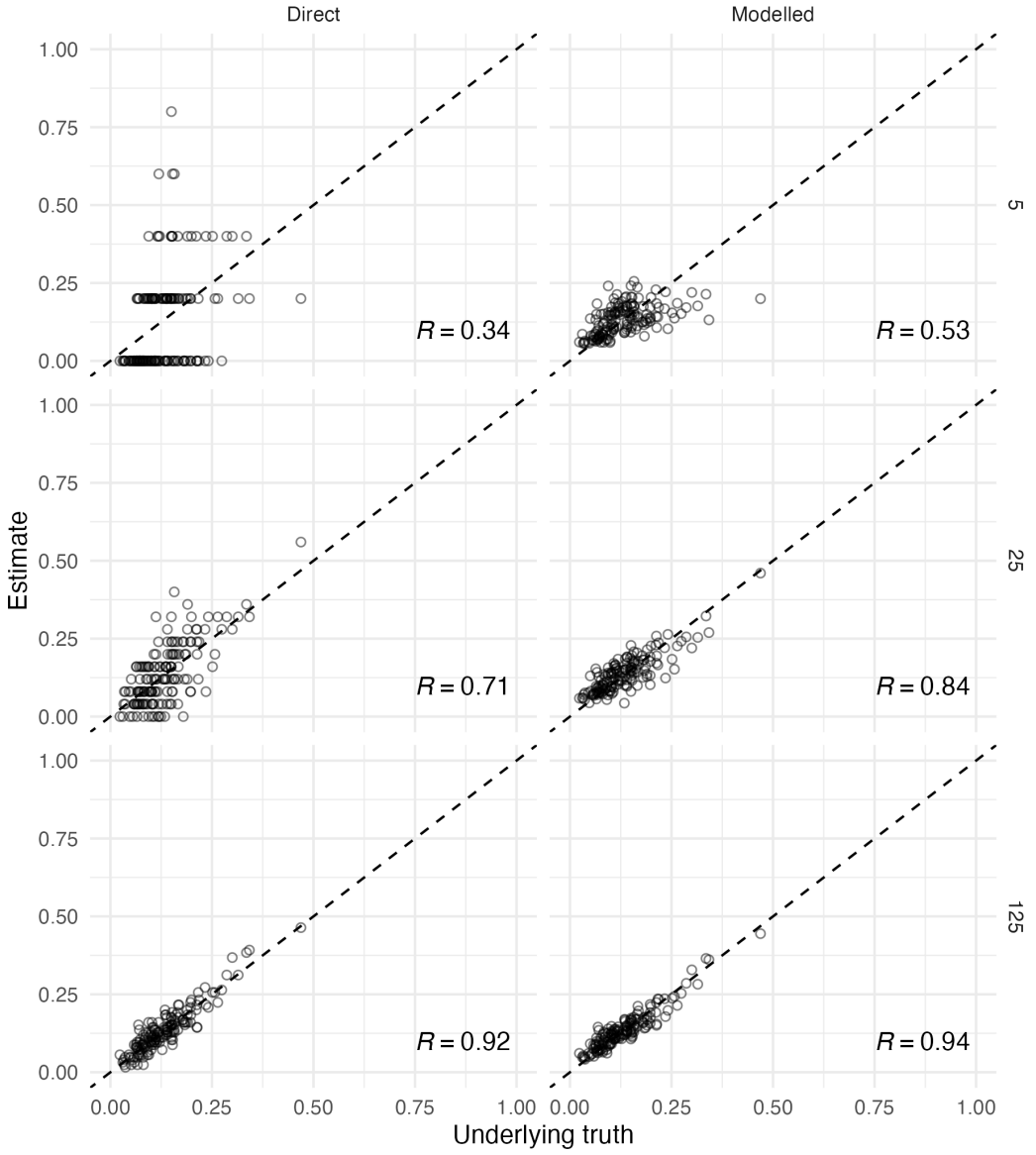

Figure 3.5: The setting of this figure matches that of Figure 3.4. Estimates from surveys with higher sample size have higher Pearson correlation coefficient R with the underlying truth, illustrating the benefit of collecting more data. For a fixed sample size however, correlation can be improved by using modelled estimates to borrow information across spatial units, rather than using the higher variance direct estimates. Points along the dashed diagonal line correspond to agreement between the estimate obtained from the survey and the underlying truth used to generate the data. For each sample size, using a spatial model increases the correlation between the estimates and underlying truth. The effect is more pronounced for lower sample sizes.

More generally, SAE methods are useful when data are limited for subpopulations of interest. These subpopulations could be generated by spatio-temporal variables, as well as by other variables such as demographics. Just as we expect there to be spatio-temporal correlation structure, we also can expect there to be demographic correlation structure. For example, those of the same sex are more likely to be similar, as are those of similar ages or socioeconomic strata.

3.3 Model structure

Section **??** showed that in spatio-temporal statistics observations are related to each other, and should not all be considered as IID. This section discuss ways in which the relations between observations can be encoded mathematically. Beginning with simple approaches, the expressiveness required to model the data in this thesis is built up in stages.

3.3.1 Linear model

In a linear model, each observation $i = 1, \ldots, n$ is modelled using a Gaussian distribution

$$y_i \sim \mathcal{N}(\mu_i, \sigma). \tag{3.7}$$

The conditional mean μ_i is assumed to be linearly related to a collection of l covariates

$$\mu_i = \beta_0 + \sum_{l=1}^{p} \beta_l z_{li}. \tag{3.8}$$

Priors may be placed on the regression coefficients $\beta_l \sim p(\beta_l)$ for $l = 0, \ldots, p$ as well as the observation standard deviation $\sigma \sim p(\sigma)$

3.3.2 Generalised linear model

A generalised linear model (GLM) extends the linear model by allowing the conditional mean to be connected to the linear predictor via a link function

$$y_i \sim p(y_i \mid \eta_i), \tag{3.9}$$

$$\mu_i = \mathbb{E}(y_i \mid \eta_i) = g(\eta_i). \tag{3.10}$$

3.3.3 Generalised linear mixed effects model

In a generalised linear mixed effects model (GLMM) the linear predictor is extended as follows

$$\eta_i = \beta_0 + \sum_{l=1}^{p} \beta_l z_{li} + \sum_{k=1}^{r} u_k(w_{ki}). \tag{3.11}$$

The terms u_k are called random effects, of additional covariates w_{ki}. The terms β_l are then referred to as fixed effects. Unfortunately these terms have notoriously many different and incompatible definitions which can cause confusion (Gelman 2005).

Random effects allow for more complex sharing of information between observations. Complete pooling, no pooling, partial pooling. Random effects can be structured to share information between some observations more than others. In spatio-temporal statistics, structured spatial and temporal random effects are often used to impose smoothness. Spatial random effects are the subject of Chapter 4.

A generalised additive model [GAMs; Wood (2017); Hastie and Tibshirani (1987)] is a type of GLMM in which such and such.

3.3.4 Latent Gaussian model

Latent Gaussian models [LGMs; Rue, Martino, and Chopin (2009)] are a type of GLMMs in which Gaussian priors are used for certain parameters of the model. In particular, the parameters β_0, $\{\beta_j\}$, $\{u_k(\cdot)\}$ are assigned Gaussian prior distributions. These parameters can be collected into a vector $\mathbf{x} \in \mathbb{R}^N$ called the latent field such that

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}(\boldsymbol{\theta}_2)^{-1}), \tag{3.12}$$

where $\boldsymbol{\theta}_2 \in \mathbb{R}^{s_2}$ are hyperparameters, with s_2 assumed small. The vector $\boldsymbol{\theta}_1 \in \mathbb{R}^{s_1}$, with s_1 assumed small, are additional parameters of the likelihood. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}^m$ with $m = s_1 + s_2$ be all hyperparameters, with prior distribution $p(\boldsymbol{\theta})$. Such that $\boldsymbol{\phi} = (\mathbf{x}, \boldsymbol{\theta})$ with posterior distribution

$$p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{3.13}$$

3.3.5 Extended latent Gaussian model

Extended latent Gaussian models [ELGMs; Stringer et al. (2022)] facilitate modelling of data with greater non-linearities than an LGM. In an ELGM, the structured additive predictor is redefined as $\boldsymbol{\eta} = (\eta_1, \dots \eta_{N_n})$, 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 $1 \leq |\mathcal{J}_i| \leq N_n$, where $[N_n] = \{1, \dots, N_n\}$. 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 $\mu_i = g_i(\boldsymbol{\eta}_{\mathcal{J}_i})$. Put together, ELGMs are of the form

$$\begin{aligned}
&y_i \sim p(y_i \,|\, \boldsymbol{\eta}_{\mathcal{J}_i}, \boldsymbol{\theta}_1), \quad i = 1, \dots, n, \\
&\mu_i = \mathbb{E}(y_i \,|\, \boldsymbol{\eta}_{\mathcal{J}_i}) = g_i(\boldsymbol{\eta}_{\mathcal{J}_i}), \\
&\eta_j = \beta_0 + \sum_{l=1}^{p} \beta_l z_{li} + \sum_{k=1}^{r} u_k(w_{ki}), \quad j = 1, \dots, N_n,
\end{aligned}$$

with latent field and hyperparameter prior distributions as in the LGM case.

The ELGM class is well suited to small-area estimation of HIV indicators, and used throughout the thesis. While it can be transformed to an LGM using the Poisson-multinomial transformation (Baker 1994) the multinomial logistic regression model used in Chapter 5 is most naturally written as an ELGM where each observation depends on the set of structured additive predictors corresponding to the set of multinomial observations. In Chapter 6, the Naomi small-area estimation model used to produce estimates of HIV indicators is shown to have the features of an ELGM.

3.4 Model comparison

Talk about: BF, AIC, BIC, WAIC, LOO, LOO-CV, connections.

3.5 Survey methods

Large national household surveys provide the highest quality population-level information about HIV indicators in SSA. Demographic and Health Surveys (DHS) are funded by the United States Agency for International Development (USAID) and run every three to five years in most countries. Population-based HIV Impact Assessment (PHIA) surveys are funded by PEPFAR and run every four to five years in high HIV burden countries.

3.5.1 Survey notation and key terms

Consider a population of individuals $i = 1, \ldots, N$ with outcomes of interest y_i. A census is a type of survey where all individuals are selected. Supposing responses from all individuals were recorded, then all population quantities can be calculated directly. For example, if $G_i = G(y_i)$ then the population mean of G is

$$\bar{G} = \frac{1}{N} \sum_{i=1}^{N} G(y_i). \tag{3.14}$$

In practice, it is usually too expensive to run a census. Instead, only a subset of the individuals are sampled. Furthermore, only a subset of those sampled have their outcome recorded, due to nonresponse or otherwise. Let S_i be an indicator for whether or not individual i is sampled, and R_i be an indicator for whether or not y_i is recorded. If $S_i = 0$ then $R_i = 0$. If $S_i = 1$ then individual i may not respond such that $R_i = 0$. The population mean may be estimated directly based on the recorded subset of the population by

$$\bar{G}_R = \frac{\sum_{i=1}^{N} R_i G(y_i)}{\sum_{i=1}^{N} R_i}, \tag{3.15}$$

where $m_R = \sum_{i=1}^{N} R_i$ is the recorded sample size.

A probability sample refers to the case when individuals are selected to be included in the survey at random. In a non-probability sample, inclusion or exclusion from the survey is deterministic. A simple random sample (SRS) is a probability sample where the sampling probability for each individual is equal, so that $P(S_i = 1) = 1/N$. The survey design is called complex when the sampling probabilities for each individual vary, such that $P(S_i = 1) = \pi_i$ with $\sum_{i=1}^{N} \pi_i = 1$ and $\pi_i > 0$.

Complex survey designs can offer both greater practicality and statistical efficiency than a SRS. However, particular care is required in analysing data collected using complex survey designs. Under a complex design, not accounting for unequal sampling probabilities will result in bias. That said, even for a SRS, nonresponse can cause analogous bias.

3.5.2 Survey design

The DHS (DHS 2012) employs a two-stage sampling procedure. In the first stage, ennumeration areas (EAs) from a recently conducted census are typically used as the primary sampling unit (PSU). The EAs are then stratified by region, as well as urban-rural. After appropriate sample sizes are determined, EAs sampled with probability proportional to size (PPS) measured In the second stage, the secondary sampling units (SSUs) are households. All households in the selected EAs are listed, before being sampled systematically. Finally, each selected household is visited, and all adults are interviewed.

The probability an individual is sampled is equal to the probability their household is sampled. The first-stage sampling probability of the jth cluster in stratum h given by

$$\pi_{1hj} = n_h \times \frac{N_{hj}}{\sum_j N_{hj}}, \tag{3.16}$$

where N_{hj} is the number of households and n_h be the number of clusters selected in stratum h. The second-stage sampling probability each household within the ith cluster in stratum h is

$$\pi_{1hj} = \frac{n_{hj}}{N_{hj}}, \tag{3.17}$$

where n_{hj} is the number of households selected in cluster j and stratum h. That is, each household in the cluster has equal selection probability. The overall selection probability of each household in cluster j of stratum h is $\pi_{hi} = \pi_{1hj} \times \pi_{2hj}$.

3.5.3 Survey analysis

Suppose a complex survey is run with sampling probabilities π_i. The standard method for taking into account that some individuals are more likely to be included in the survey than others is to overweight the responses of those unlikely to be included, and underweight the responses of those likely to be included. This can be achieved using design weights $\delta_i = 1/\pi_i$, which can be thought of as the number of individuals in the population represented by the ith sampled individual. Let $P(R_i = 1 \,|\, S_i = 1) = \upsilon_i$ be the probability of response for sampled individual i. The problem of nonresponse can be treated in the same way using nonresponse weights $\gamma_i = 1/\upsilon_i$, which analogously can be thought of as the number of sampled individuals represented by the ith recorded individual. Multiplying the design and nonresponse weights gives survey weights $\omega_i = \delta_i \times \gamma_i$.

A weighted estimate (Hájek 1971) of the population mean using the survey weights ω_i is given by

$$\bar{G}_\omega = \frac{\sum_{i=1}^N \omega_j R_i G(y_i)}{\sum_{i=1}^N \omega_i R_i}. \tag{3.18}$$

Decomposing the additive error of this estimate provides useful intuition as to the potential benefits of survey weighting. Following Meng (2018) then under SRS

$$\bar{G}_\omega - \bar{G} = \frac{\mathbb{E}(\omega_i R_i G_i)}{\mathbb{E}(\omega_i R_i)} - \mathbb{E}(G_i) = \frac{\mathbb{C}(\omega_i R_i G_i)}{\mathbb{E}(\omega_i R_i)} \tag{3.19}$$

$$= \rho_{R_\omega, G} \times \sqrt{\frac{N - m_{R_\omega}}{m_{R_\omega}}} \times \sigma_G, \tag{3.20}$$

where $R_\omega = \omega R$. The data defect correlation (DDC) $\rho_{R_\omega, G}$ measures the correlation between the weighted recording mechanism and given function of the outcome of interest. To minimise the DDC then $G \perp\!\!\!\perp R_\omega$. The data scarcity $\sigma_{R_\omega} = \sqrt{(N - m_{R_\omega})/m_{R_\omega}}$ measures the effective proportion of the population who have

been recorded. The problem difficultly σ_G measures the intrinsic difficulty of the estimation problem, and is independent of the sampling or analysis method.

For simplicity, let $G(y_i) = y_i$ and each $y_i \in \{0, 1\}$. We weight then model following Chen et al. (2014). While this approach acknowledges the survey design, it has some important limitations. We ignore clustering structure. All of this isn't great and that someone should figure this out (Gelman 2007).

4

Models for spatial structure

This chapter presents an investigation of spatial random effects specifications for areal data. The investigation was motivated by a question that often occurs during model construction. Namely, should the model be expanded to capture a sensible, albeit hypothetical, feature of the data?

The hypothesised feature in this case pertains to the spatial correlation structure between areas. Modelling of spatial variation is particularly important for the small-area estimation of HIV. This is because the covariates which are most strongly associated with HIV are difficult to measure. Examples include sexual risk behaviour. As a result, in previous small-area models of HIV have found including including covariates to only result in a modest improvement in predictive performance (Supplementary Figure 20, Dwyer-Lindgren, Cork, et al. 2019). The lack of predictive covariates foregrounds the role of modelling spatial variation. For mapping of other infectious diseases, such as Malaria where transmission is driven by more predictive and easily-measurable environmental factors, explanatory covariates are more easily available and modelling spatial variation is less pertinent (Weiss et al. 2015).

Spatial variation in areal data are often modelled using spatial random effects (Haining 2003; Cramb et al. 2018). The most common class of models used to specify spatial random effects are Gaussian Markov random fields [GMRFs; Rue and Held

(2005)]. These models combine a Gaussian distribution with Markov conditional independence assumptions between areas. Observations made in areas close together are assumed to be correlated, and more distant relationships are ignored. Perhaps the simplest GMRF model is that of Besag et al. (1991) in which information is borrowed equally from each adjacent area, based on a binary relationship. The Besag model is attractive as it requires minimal additional modelling choices and is accessibly implemented. It has been widely used, including:

- to model bird population dynamics from capute-recapture data (Saracco et al. 2010);
- for the analysis of magnetic resonance images (Gössl et al. 2001; Schmid et al. 2006);
- to model alcohol use patterns (Dwyer-Lindgren, Flaxman, et al. 2015).

The Besag model was designed for use in image analysis, on a regular grid. However, for more irregular geometries, the assumptions made are unrealistic and appear to be violated. The administrative divisions of a country used in small-area estimation are one example of a more irregular geometry. This chapter tests the hypothesis that using more realistic assumptions about spatial structure would improve the performance of small-area estimation models. In doing so, it offers practical recommendations for modelling areal spatial structure. The results are presented in Howes, Eaton, et al. (2023+). Code for the analysis in this chapter is available from `https://github.com/athowes/beyond-borders`.

4.1 Models based on adjacency

This section discusses spatial random effect models based on an symmetric adjacency relation $i \sim j$ between areas A_i and A_j. Adjacency is typically defined by a shared border, though other choices are possible (Paciorek et al. 2013).

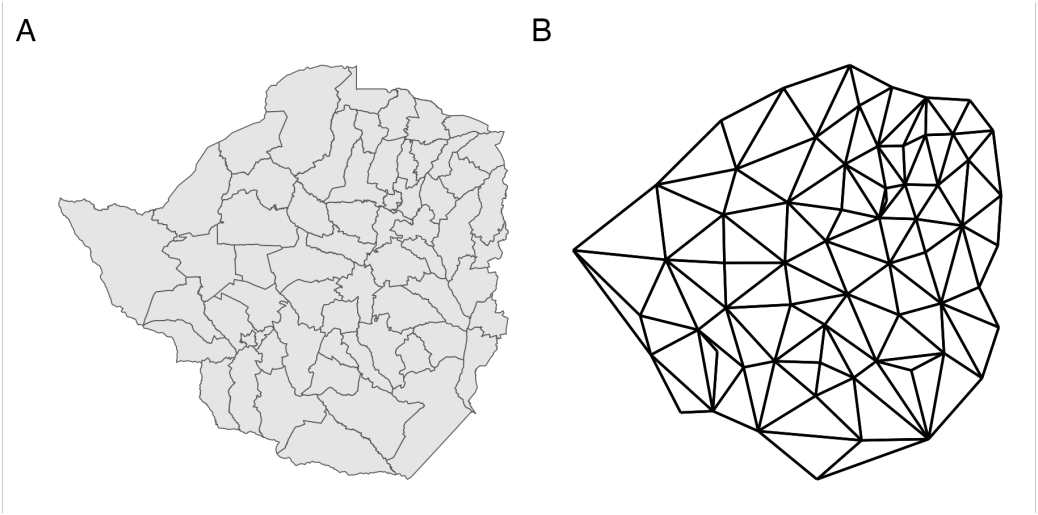

Figure 4.1: Panel A shows the districts of Zimbabwe. Panel B shows the corresponding adjacency graph structure $\mathcal{G}$, with nodes positioned in alignment with the district that they correspond to.

4.1.1 The Besag model

The Besag model (Besag et al. 1991) is an improper conditional auto-regressive (ICAR) model where the full conditional distribution of the ith spatial random effect is given by

$$u_i \,|\, \mathbf{u}_{-i} \sim \mathcal{N}\left(\frac{1}{n_{\delta i}} \sum_{j:j\sim i} u_j, \frac{1}{n_{\delta i}\tau_u}\right), \tag{4.1}$$

where δi is the set of neighbours of A_i with cardinality $n_{\delta i} = |\delta i|$ and $\mathbf{u}_{-i}$ is the vector of spatial random effects with the ith entry removed. The conditional mean of the random effect u_i is the average of its neighbours $\{u_j\}_{j\sim i}$ and the precision $n_{\delta i}\tau_u$ is proportional to the number of neighbours $n_{\delta i}$. By Brook's lemma (Rue and Held 2005) the set of full conditionals of the Besag model are equivalent to the Gaussian Markov random field (GMRF) given by

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \tau_u^{-1}\mathbf{R}^-). \tag{4.2}$$

The matrix $\mathbf{R}^-$ is the generalised inverse of the rank-deficient structure matrix $\mathbf{R}$ with entries

$$R_{ij} = \begin{cases} n_{\delta i}, & i = j \\ -1, & i \sim j \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

The Markov property arises due to the conditional independence structure $p(u_i \mid \mathbf{u}_{-i}) = p(u_i \mid \mathbf{u}_{\delta i})$ whereby each area only depends on its neighbours. This is reflected in the sparsity of $\mathbf{R}$ such that $u_i \perp u_j \mid \mathbf{u}_{-ij}$ if and only if $R_{ij} = 0$. The structure matrix $\mathbf{R}$ may also be expressed as the Laplacian matrix of the adjacency graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $v \in \mathcal{V}$ corresponding to each area and edges $e \in \mathcal{E}$ between vertices i and j when $i \sim j$. Figure 4.1 shows the adjacency graph for the districts of Zimbabwe.

Rewriting Equation (4.2), the probability density function of $\mathbf{u}$ is

$$p(\mathbf{u}) \propto \exp\left(-\frac{\tau_u}{2}\mathbf{u}^\top \mathbf{R}\mathbf{u}\right) \propto \exp\left(-\frac{\tau_u}{2}\sum_{i\sim j}(u_i - u_j)^2\right). \tag{4.4}$$

This density is a function of the pairwise differences $u_i - u_j$ and so is invariant to the addition of a constant c to each entry $p(\mathbf{u}) = p(\mathbf{u} + c\mathbf{1})$. As a result, there is an improper uniform distribution on the average of the u_i. If $\mathcal{G}$ is connected, in that by traversing the edges, any vertex can be reached from any other vertex, then there is only one impropriety in the model and $\text{rank}(\mathbf{R}) = n - 1$, while if $\mathcal{G}$ is disconnected, and composed of $n_c \geq 2$ connected components with index sets $I_1, \ldots, I_{n_c}$, then the corresponding structure matrix $\mathbf{R}$ has rank $n - n_c$ and the density is invariant to the addition of a constant to each of the connected components $p(\mathbf{u}_I) = p(\mathbf{u}_I + c\mathbf{1})$ where $I = I_1, \ldots, I_{n_c}$.

4.1.2 Best practises for the Besag model

Freni-Sterrantino et al. (2018) recommended three best practices:

1. The structure matrix $\mathbf{R}$ should be rescaled to have generalised variance equal to one. The generalised variance is defined by the geometric mean of the diagonal elements of its generalised inverse

$$\sigma^2_{\text{GV}}(\mathbf{R}) = \prod_{i=1}^{n}(\mathbf{R}^-_{ii})^{1/n} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(R^-_{ii})\right). \tag{4.5}$$

   The structure matrix $\mathbf{R}$ may be replaced by

$$\mathbf{R}^\star = \mathbf{R}/\sigma^2_{\text{GV}}(\mathbf{R}). \tag{4.6}$$

As the diagonal elements R_{ii}^{-} correspond to marginal variances, the generalised variance gives a measure of the average marginal variance. This measure, introduced by Sørbye and Rue (2014), ignores off-diagonal entries and more broadly any measure of typical variance could be used. Scaling mitigates the influence of the adjacency graph on the variance of $\mathbf{u}$. Allowing the variance to be controlled by τ_u alone is important as it allows for consistent, interpretable prior selection.

When the adjacency graph is disconnected it is not appropriate to scale the structure matrix $\mathbf{R}$ uniformly for the reason that given the precision τ_u, local smoothing operates on each connected component independently. As such, each connected component should be scaled independently to have generalised variance one giving

$$\mathbf{R}_I^{\star} = \mathbf{R}_I / \sigma^2_{\mathrm{GV}}(\mathbf{R}_I) \tag{4.7}$$

where $\mathbf{R}_I$ is the sub-matrix of the structure matrix corresponding to index set I.

2. When one of the connected components is a single area (known either as a singleton or an island) the probability density

$$p(u_i) \propto \exp\left(-\frac{\tau_u}{2}\sum_{i\sim j}(u_i - u_j)^2\right) \tag{4.8}$$

has no dependence on u_i. This is equivalent to using an improper prior $p(u_i) \propto 1$. To avoid this, each singleton can be set to have independent Gaussian noise $p(u_i) \sim \mathcal{N}(0, 1)$.

3. To avoid confounding of the spatial random effects with the intercept, it is recommended to place a sum-to-zero constraint on each non-singleton connected component. In other words,

$$\sum_{i\in I} u_i = 0, \quad |I| > 1. \tag{4.9}$$

4.1.3 The reparameterised Besag-York-Mollié model

Often, as well as spatial structure, there exists IID over-dispersion in the residuals and it is inappropriate to use purely spatially structured random effects in the model. The Besag-York-Mollié (BYM) model of Besag et al. (1991) accounts for this in a natural way by decomposing the spatial random effect $\mathbf{u} = \mathbf{v} + \mathbf{w}$ into a sum of an unstructured IID component $\mathbf{v}$ and a spatially structured Besag component $\mathbf{w}$, each of which with their own respective precision parameters τ_v and τ_w. The resulting distribution is

$$\mathbf{u} \sim \mathcal{N}(0, \tau_v^{-1}\mathbf{I} + \tau_w^{-1}\mathbf{R}^-). \tag{4.10}$$

Including both $\mathbf{v}$ and $\mathbf{w}$ is intended to enable the model to learn the relative extent of the unstructured and structured components via τ_v and τ_w. However, in this specification scaling of the Besag precision matrix $\mathbf{Q}$ is not taken into account despite this issue being particularly pertinent when dealing with multiple sources of noise. In particular, placing a joint prior $(\tau_v, \tau_w) \sim p(\tau_v, \tau_w)$ which doesn't privilege either component is more easily accomplished if $\mathbf{Q}$ and $\mathbf{I}$ have the same scale. Additionally, supposing we have a prior belief that the over-dispersion is primarily IID and $\mathbf{v}$ accounts for the majority of the dispersion, then it is not immediately obvious how to represent this belief using $p(\tau_v, \tau_w)$, without inadvertently altering the prior about the overall variation. This highlights identifiability issues of the parameters (τ_v, τ_w) resulting from them not being orthogonal. Building on the models of Leroux et al. (2000) and Dean et al. (2001) which tackle this identifiability problem, but do not scale the spatially structured noise, Simpson et al. (2017) propose a reparameterisation $(\tau_v, \tau_w) \mapsto (\tau_u, \phi)$ of the BYM model known as the BYM2 model and given by

$$\mathbf{u} = \frac{1}{\tau_u}\left(\sqrt{1-\phi}\,\mathbf{v} + \sqrt{\phi}\,\mathbf{w}^\star\right), \tag{4.11}$$

where τ_u is the marginal precision of $\mathbf{u}$, $\phi \in [0, 1]$ gives the proportion of the marginal variance explained by each component, and $\mathbf{w}^\star$ is a scaled version of $\mathbf{w}$ with precision matrix given by the scaled structure matrix $\mathbf{R}^\star$. When $\phi = 0$ the

random effects are IID, and when $\phi = 1$ the random effects follow the Besag model. To borrow an analogy (Rue 2020) the parameterisation (τ_v, τ_w) is like having one hot water and one cold water tap, whereas the parameterisation (τ_u, ϕ) is like a mixer tap where the amount of water and its temperature can be adjusted separately.

4.1.4 Concerns about the Besag model's spatial representation

The Besag model was originally proposed for use in image analysis, where areas correspond to pixels arranged in a regular lattice structure. Since then, it has seen wider use, including in situations, like small-area estimation of HIV, where the spatial structure is less regular. As such, I have a number of concerns about the model's applicability to this broader setting. This discussion is closely linked to the modifiable areal unit problem (Openshaw and Taylor 1979), whereby statistical conclusions change as a result of seemingly arbitrary changes in data aggregation, as well as the challenge of ecological inference and the ecological fallacy (Wakefield and Lyons 2010).

Adjacency compression

Summarising a geometry by an adjacency graph represents a loss of information. Many geometries share the same adjacency graph, and are as such isomorphic identical under the Besag model (Figure). This is not in itself a problem, but does prompt consideration as to whether the class of geometries with the same adjacency graph is sufficiently similar to merit identical models.[1] Intuitively, the more regular the spatial structure, the less information is lost in compression to an adjacency graph. In image analysis, very little spatial information is lost in compression of a lattice structure to an adjacency graph. On the other hand, the regions of a country, determined by political and geographic forces, tend to display

[1]The regularity of realistic geometries may help to constrain each class to be more similar than it strictly has to be. In other words, although pathological geometries can be constructed, they are implausible in statistical practise and so not of great concern to us here.

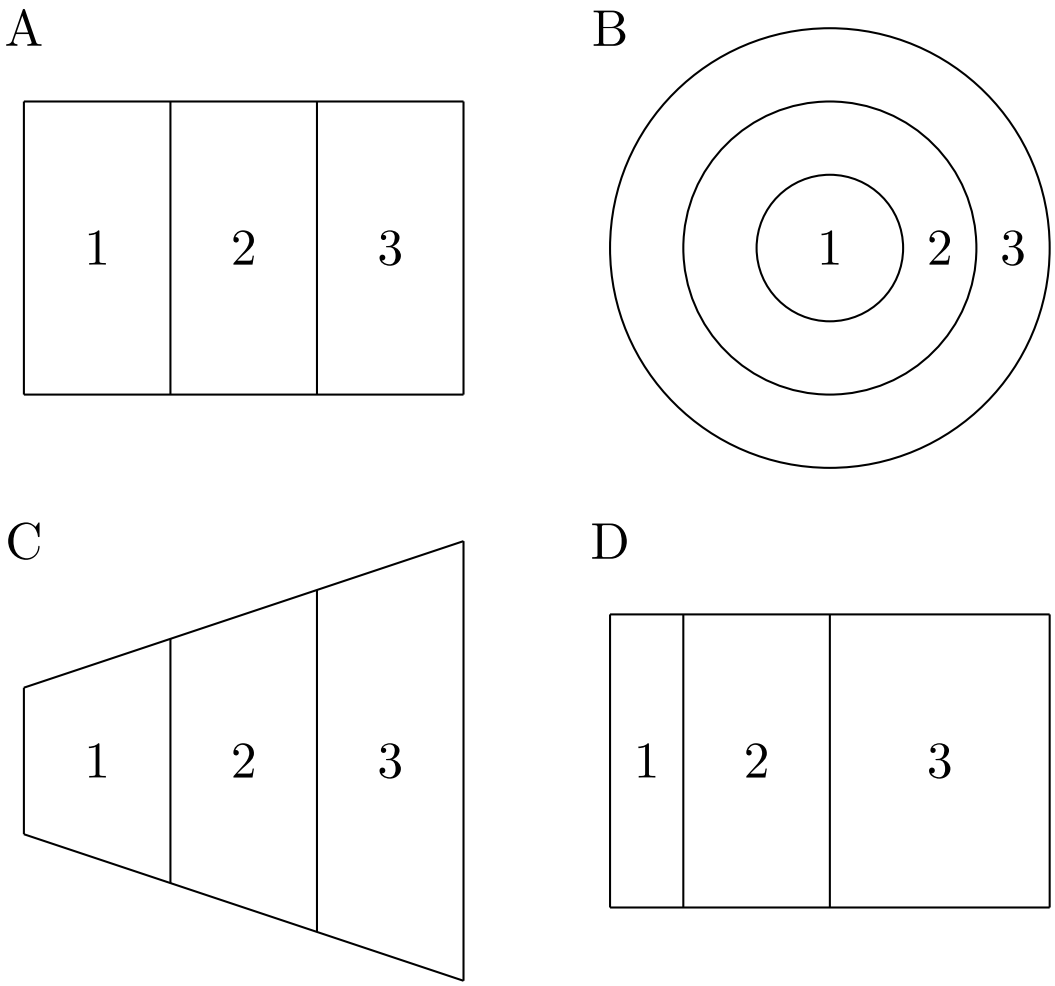

Figure 4.2: Though they are quite different, the geometries shown in panels A, B, C, and D each have the same adjacency graph.

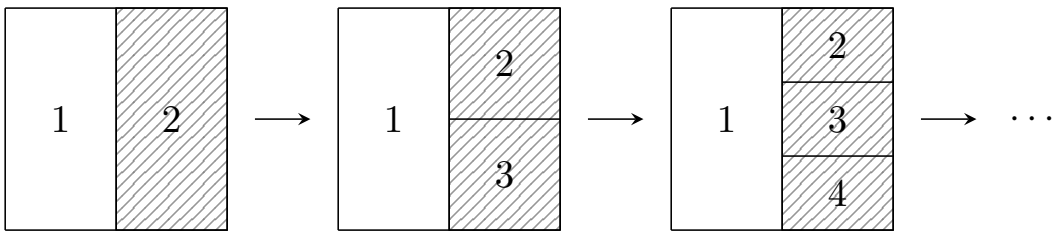

Figure 4.3: A sequence of geometries where the number of neighbours of area one grows by one at each iteration.

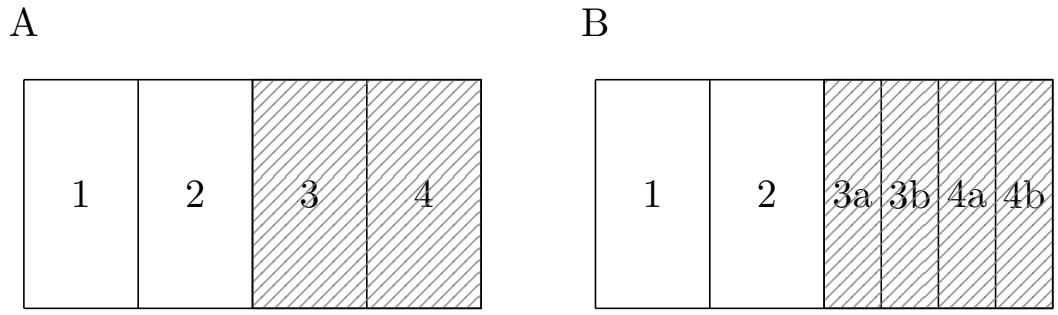

Figure 4.4: Each of the shaded areas are split into two moving from Panel A to Panel B.

greater irregularity. The appropriateness of adjacency compression therefore varies by the type of geometry common to the application setting.

Mean structure

In the Besag model is all adjacent areas count equally. This assumption is unsatisfying: for most geometries, we expect different amounts of correlation between neighbours. Figure illustrates a number of heuristic features for neighbour importance, including length of shared border, and the proximity of centers of mass.

Variance structure

In Equation (4.1) the precision of u_i is proportional to its number of neighbours $n_{\delta i}$. It follow that as $n_{\delta i} \to \infty$ then $\text{Var}(u_i) \to 0$. This is illustrated by Figure where the area on the right is repeatedly divided such that its number of neighbours increases. This property is a consequence of averaging the conditional mean over a greater number of areas, which, in certain situations, can correspond to a greater amount of information. However, if the amount of information in the shaded area remains fixed, it is inappropriate that $\text{Var}(u_1)$ should tend to zero as a result of drawing additional, arbitrary, boundaries. In the image analysis setting this modelling assumption is reasonable: each pixel represents a fixed amount of information and a higher pixel density represents a greater amount of information. On the other hand, in public health and epidemiology, drawing boundaries to create additional areas is not expected to correspond to a greater amount of information.

Suppose we fit a Besag model upon identical data using each of the two geometries in Figure. If the spatial variation is relatively smooth, dividing the shaded areas into two will result in a lower estimated variance σ_u^2 in geometry (ii) as compared with geometry (i) because there will appear to be less variation between neighbouring areas. This problem does not only apply locally: since the effect of σ_u^2 applies everywhere, the smoothing will change even in unaltered parts of the study region.

4.2 Models using kernels

Section 4.1 reviewed ways to construct spatial random effect precision matrices using an adjacency relation. An alternate approach is to define the covariance matrix using an areal kernel function which gives a measure of similarity between two areas $K : \mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S}) \to \mathbb{R}$, where $\mathcal{P}$ denotes the power set such that $\mathcal{P}(\mathcal{S})$ is the space of subsets of the study region. If K is positive semi-definite, then define areal kernel spatial random effects by

$$\phi \sim \mathcal{N}\left(0, \frac{1}{\tau_\phi}\mathbf{K}\right), \tag{4.12}$$

where the $n \times n$ Gram matrix $\mathbf{K}$ with entries $K_{ij} = K(A_i, A_j)$ is a valid covariance matrix. Here, the precision parameter τ_ϕ is placed outside of the Gram matrix, analogous to the relation of the precision and structure matrices. Most well-known spatial process models define the correlation structure between a pair of points using a kernel $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$. A simple method to construct K from k is to average the kernel k computed on some collection of points from within each area.

4.2.1 Centroid kernel

The simplest approach is to use a single point such that $K(A_i, A_j) = k(p_i, p_j)$. A natural choice is the centroid $p_i = c_i$, given by the arithmetic mean of the latitude and longitude, which may be representative of the area.[2] This results in the centroid kernel

$$K(A_i, A_j) = k(c_i, c_j). \tag{4.13}$$

[2]Note it is not guaranteed for the centroid to (even) lie within the area i.e. we may have $c_i \notin A_i$.

The centroid kernel has been used in environmental epidemiology (Wakefield and Morris 1999) and to model the reproduction number of COVID-19 (Teh et al. 2021). A model comparison study [Section 3; Best et al. (2005)] simulated data representing heterogeneous exposure to air pollution, including elevated rates of exposure near two hypothetical point source locations, and found that the centroid kernel tended to over-smooth the high-risk areas. That said, it is unsuprising that a stationary covariance would struggle to recover non-stationary structure.

4.2.2 Integrated kernel

Rather than choosing a single representative point, an alternative is to represent the whole area by integrating the kernel over the areas of interest. This results in the integrated kernel

$$K(A_i, A_j) = \frac{1}{|A_i||A_j|} \int_{A_i} \int_{A_j} k(s, s') \mathrm{d}s\mathrm{d}s'. \tag{4.14}$$

This covariance structure is equivalent to that obtained by aggregating a spatially continuous Gaussian process with kernel k over the areal partition, and has been studied in the machine learning literature under the name aggregated Gaussian processes (Law et al. 2018; Tanaka et al. 2019; Yousefi et al. 2019; Hamelijnck et al. 2019). Unlike for the centroid kernel where $K_{ii} = 1$ for all i, the marginal variance of the ith spatial random effect $K_{ii} = K(A_i, A_i)$ varies depending on the area: becoming smaller for more compact areas and larger for areas which are of greater extent or more spread out.

Accounting for heterogeneity

Additional information accounting for heterogeneity over A_i may be incorporated into the integrated kernel[3] This can be accomplished using weighting distributions $\{W_i\}$ which represent an unequal contribution of each point to the similarity measure, to give a weighted integrated kernel

$$K(A_i, A_j) = \frac{1}{|A_i||A_j|} \int_{A_i} \int_{A_j} w_i(s) w_j(s') k(s, s') \mathrm{d}s\mathrm{d}s', \tag{4.15}$$

[3]Analogously, weighted centroids could also be used in the centroid kernel.

This may be useful in disease mapping, where we expect regions with populations who live close to their shared border to be more strongly correlated than regions whose populations live far apart, which could be accounted for by weighting according to a high resolution measure of population density.

Computation

Most of the time we do not expect to be able to calculate Equation 4.15 analytically. Instead, given n collections of L_i samples $\{s_l^{(i)}\}_{l=1}^{L_i} \sim \mathcal{U}(A_i)$ drawn uniformly from each area then the integral may be approximated using Monte Carlo by the double sum

$$K(A_i, A_j) \approx \frac{1}{L_i L_j} \sum_{l=1}^{L_i} \sum_{m=1}^{L_j} w_i\left(s_l^{(i)}\right) w_j\left(s_m^{(j)}\right) k\left(s_l^{(i)}, s_m^{(j)}\right). \tag{4.16}$$

Equivalently, samples drawn from W_i may be used without weighting by $w_i(s)$. Nodes may also be selected deterministically to give a numerical quadrature estimate of the kernel. These approaches require $\mathcal{O}(\sum_{i=1}^{n} \sum_{j=1}^{n} L_i L_j)$ evaluations of the kernel k to compute the $n \times n$ Gram matrix K. This imposes a significant computational cost if the Gram matrix is often recomputed during inference, as is the case in MCMC when any of the kernel hyperparameters are learnt, placing a limit on the number of samples or nodes it is feasible to use. Kelsall and Wakefield (2002) make inference more feasible by using a discrete hyperparameter prior to reduce the number of Gram matrix constructions and inversions required.

Mismatch to data generating process

Aggregation via the integrated kernel occurs at the level of the latent field rather than at the level of the data. If the link function g is the identity or linear then aggregation at the level of the latent field is equivalent to aggregation at the level of the data. On the other hand, for non-linear link functions g such as the commonly used exponential or logistic, the generative model does not match the proposed data generating process.

Log-Gaussian Cox processes

The log-Gaussian Cox Process framework (Diggle et al. 2013) arrives naturally at the integrated kernel formulation. A Cox process is an inhomogeneous Poisson process with a continuous stochastic intensity function $\{x(s), s \in \mathcal{S}\}$ such that conditional on the realisation of $x(s)$ the number of points in any area A_i follows a Poisson distribution. The rate parameter of this Poisson distribution is explicitly aggregated as follows

$$y_i \,|\, x(s) \sim \text{Poisson}\left(\int_{s \in A_i} x(s)\mathrm{d}s\right). \tag{4.17}$$

In a LGCP the log intensity $\log x(s) = \eta(s)$ is modelled using a Gaussian process prior $\eta(s) \sim \mathcal{GP}(\mu(s), k(s, s'))$. Johnson, Diggle, et al. (2019) obtain Equation 4.15 by considering a discrete Poisson log-linear mixed model approximation to a continuous LGCP, whereby $\eta(s)$ is approximated by a piecewise constant $\eta_i = \mu_i + \phi_i$ in each area A_i. The ith discrete spatial random effect is then $\phi_i = \int_{A_i} w_i(s)\phi(s)\mathrm{d}s$, with covariance structure

$$\text{Cov}\left(\int_{A_i} w_i(s)\phi(s)\mathrm{d}s, \int_{A_j} w_j(s')\phi(s')\mathrm{d}s'\right) = \int_{A_i}\int_{A_j} w_i(s)w_j(s')k(s, s')\mathrm{d}s\mathrm{d}s', \tag{4.18}$$

corresponding to an areal integrated kernel with a logarithmic link function and Poisson likelihood.

Disaggregation regression

Disaggregation regression, also known as downscaling or interpolation, is another closely related approach. Rather than focusing on the aggregate nature of areal observations as primarily a route towards better area-level estimates, disaggregation regression aims to produce high-resolution or point-level estimates from areal observations (Utazi et al. 2019; Nandi et al. 2023).

4.2.3 The stochastic partial differential equation approximation

This is a more computationally efficient way to implement integrated kernels.

4.3 Simulation study

We tested the ability of inferential models with varying spatial random effect specifications to accurately recover small-area quantities. The data and modelling choices were designed with a spatial epidemiology setting in mind.

4.3.1 Synthetic data-sets

4.4 HIV prevalence study

4.5 Discussion

Follestad and Rue (2003)

5

A model for risk group proportions

This chapter describes an application of Bayesian spatio-temporal statistics to small-area estimation of HIV risk group proportions. This work was conducted in collaboration with colleagues from the MRC Centre for Global Infectious Disease Analysis and UNAIDS. I developed the statistical model, building upon an earlier version of the analysis conducted by Dr. Kathryn Risher. The model and results for 13 countries are presented in Howes, Risher, et al. (2023), and implemented in a spreadsheet tool (`https://hivtools.unaids.org/pse/`) for use in national HIV response planning. The tool is being updated by inclusion of more countries to the analysis, and extension of the methodology, including to additional risk groups. Code for the analysis in this chapter is available from `https://github.com/athowes/multi-agyw`.

5.1 Background

In SSA, adolescent girls and young women (AGYW) aged 15-29 are at increased risk of HIV infection. Though AGYW are only 28% of the population, they comprise 44% of new infections (UNAIDS 2021a). HIV incidence for AGYW is 2.4 times higher than for similarly aged (15-29) males. The social and biological reasons for this disparity include structural vulnerabilities and power imbalances, age patterns

of sexual mixing, a younger age at first sex, and increased susceptibility to HIV infection. On this basis, AGYW have been identified as a priority population for HIV prevention services. Significant investments, such as the DREAMS partnership (Saul et al. 2018) and by the Global Fund (The Global Fund 2018), have been made to support prevention programming.

The Global AIDS Strategy 2021-2026 (UNAIDS 2021b) was adopted by the United Nations (UN) General Assembly in June 2021, and "outlines the strategic priorities and actions to be implemented by global, regional, country and community partners to get on-track to ending AIDS". It proposed stratifying HIV prevention packages to AGYW based on two factors

1. local population-level HIV incidence, and
2. individual-level sexual risk behaviour.

Risk of acquiring HIV depends on both factors. As such, prioritisation of prevention services is more efficient if both factors are taken into account. I illustrate this stylistically in Figure 5.1. The strategy encourages programmes to define targets for the proportion of AGYW to be reached with a range of interventions. Implementation of the strategy by national HIV programmes and stakeholders requires data on the population size and HIV incidence in each risk group by location.

5.2 Data

5.2.1 Behavioural data from household surveys

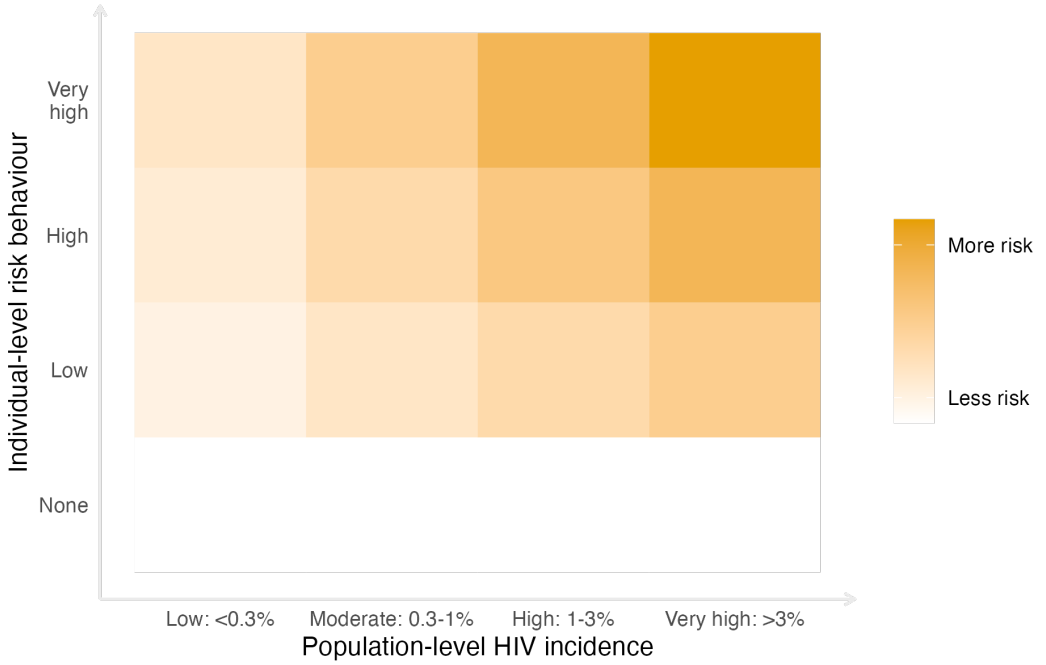

Figure 5.1: Risk of acquiring HIV is depends on both individual-level risk behaviour and population-level HIV incidence. I assume that with no individual-level risk behaviour, there is no risk of acquiring HIV, independent of the population-level HIV incidence. The risk scale is intended to be illustrative, rather than interpreted quantitatively.

Table 5.1: HIV risk groups and HIV incidence rate ratios relative to AGYW with one cohabiting sexual partner. The incidence rate ratio for women with non-regular or multiple sexual partner(s) was derived from analysis of longitudinal data by Slaymaker et al. (2020). Among FSW, the incidence rate ratio (25.0, 13.0, 9.0, 6.0, 3.0) depended on the level of HIV incidence among the general population (<0.1%, 0.1-0.3%, 0.3-1.0%, 1.0-3.0%, >3.0%), such that higher local HIV incidence in the general population corresponded to a lower incidence rate ratio for FSW. Estimates of HIV incidence rate ratios for FSW were derived by UNAIDS based on patterns of relative HIV prevalence among FSW compared to general population prevalence.

Risk group	Description	Incidence rate ratio
None	Not sexually active	0.0
Low	One cohabiting sexual partner	1.0 (baseline)
High	Non-regular or multiple partner(s)	1.72
Very High	Reporting transactional sex (later adjusted to correspond to FSW)	3.0-25.0 (varied depending on local HIV incidence)

I used household survey data from 13 countries identified by the Global Fund

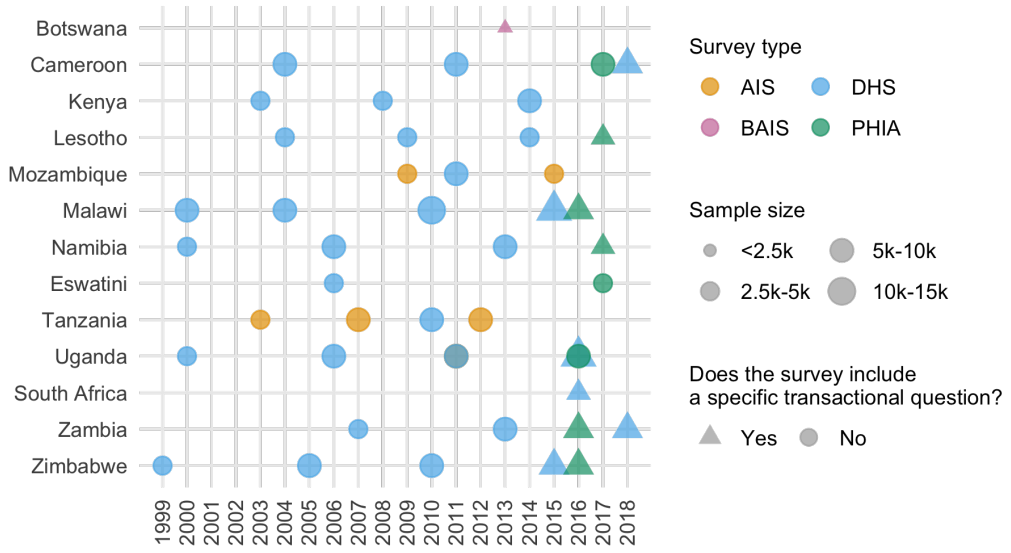

Figure 5.2: Surveys conducted 1999-2018 that were used in the analysis by year, survey type, sample size, and whether the survey included a specific question about transactional sex. Survey type included AIDS Indicator Surveys (AIS), Demographic and Health Surveys (DHS), the Botswana AIDS Impact Survey 2013 (BAIS), and Population-based HIV Impact Assessment (PHIA) surveys.

(The Global Fund 2018) as priority countries for implementation of AGYW HIV prevention. These countries were Botswana, Cameroon, Kenya, Lesotho, Malawi, Mozambique, Namibia, South Africa, Eswatini, Tanzania, Uganda, Zambia and Zimbabwe. Surveys conducted in these countries between 1999 and 2018 were included in which both women were interviewed about their sexual behaviour, and sufficient geographic information was available to locate survey clusters to health districts. There were 46 suitable surveys (Figure 5.2), with a total sample size of 274,970 women aged 15-29 years. Of the respondents, 103,063 were aged 15-19 years, 92,173 were aged 20-24 years, and 79,734 were aged 25-29 years. The median number of surveys per country was four, ranging from one in Botswana and South Africa to six in Uganda.

For each survey, I classified respondents into one of four behavioural risk groups $k = 1, 2, 3, 4$ according to reported sexual risk behaviour in the past 12 months (Figure 5.3). In increasing order of HIV acquisition risk, these risk groups were:

- $k = 1$: Not sexually active

- $k = 2$: One cohabiting sexual partner
- $k = 3$: Non-regular or multiple sexual partner(s), and
- $k = 4$: Reporting transactional sex.

The HIV incidence rate ratio RR_k was assumed to vary by risk group (Table 5.1), with the one cohabiting partner risk group as baseline.

Exact survey questions varied slightly across survey types and between survey phases. Questions captured information about whether the respondent had been sexually active in the past twelve months, and if so with how many partners. For their three most recent partners, respondents were also asked about the type of partnership. Possible partnership types included spouse, cohabiting partner, partner not cohabiting with respondent, friend, sex worker, sex work client, and other. Full survey questions used are in Appendix B.4. In the case of inconsistent responses, women were categorised according to the highest risk group they fell into, ensuring that the categories were mutually exclusive.

Some surveys included a specific question asking if the respondent had received or given money or gifts for sex in the past twelve months. In these surveys, 2.64% of women reported transactional sex. In surveys without such a question, women almost never (0.01%) answered that one of their three most recent partners was a sex work client. This incomparability made it inappropriate to include surveys without a specific transactional sex question when estimating the proportion of the population who engaged in transactional sex. Of the total 46 surveys included in the analysis, 12 had a specific transactional sex question, with a total sample size of 62,853 (28,753 aged 15-19 years, 26,324 aged 20-24 years, and 7,776 aged 25-29 years). The sample size for women aged 25-29 is smaller because there were 6 DHS surveys which excluded women 25-29 from the transactional sex survey question.

5.2.2 Other data

In addition to the household survey behavioural data, I used estimates of population, PLHIV and new HIV infections stratified by district and age group from HIV

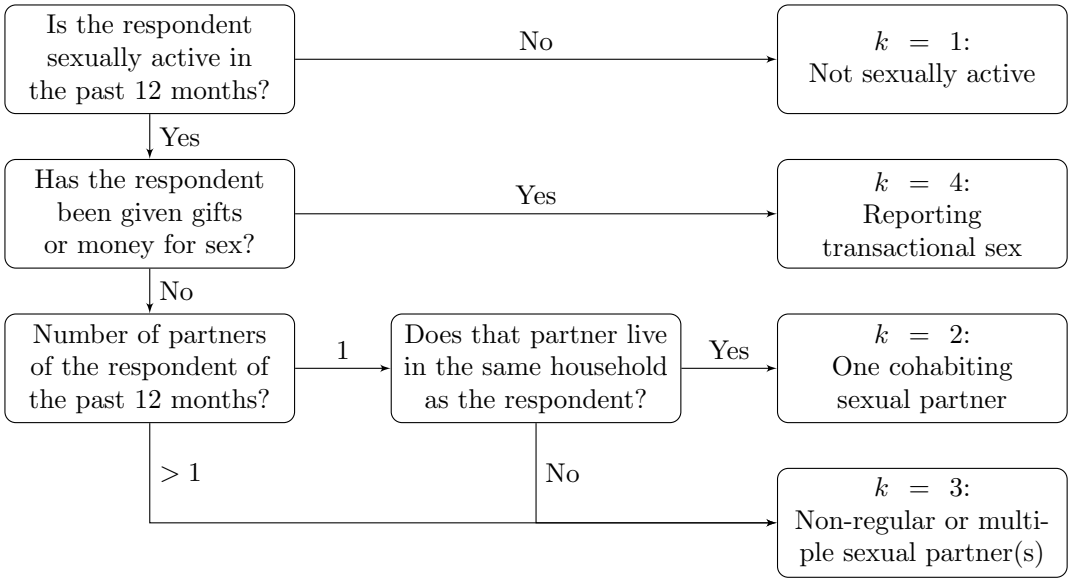

Figure 5.3: Flowchart giving classification of survey respondents to HIV risk groups.

estimates published by UNAIDS that were developed using the Naomi model (Eaton et al. 2021). I used the most recent 2022 estimates for all countries, apart from Mozambique where, due to data accuracy concerns, I used the 2021 estimates (in which the Cabo Delgado province is excluded due to disruption by conflict). I used administrative area hierarchy and geographic boundaries corresponding to those used for health service planning by countries. Exceptions were Cameroon and Kenya, where I conducted analyses one level higher at the department and county levels, respectively.

5.3 Model for risk group proportions

Owing to the incomparability in estimating the $k = 4$ risk gorup across surveys, I took a two-stage modelling approach to estimate the four risk group proportions. Denote being in either the third or fourth risk group as $k = 3^+$. First, using all the surveys, I used a spatio-temporal multinomial logistic regression model to estimate the proportion of AGYW in the risk groups $k \in \{1, 2, 3^+\}$. This model is described in Section 5.3.1. Then, using only those surveys with a specific transactional sex question, I fit a spatial logistic regression model to estimate the proportion of

those in the $k = 3^+$ risk group that were in the $k = 3$ and $k = 4$ risk groups respectively. This model is described in Section 5.3.2.

5.3.1 Spatio-temporal multinomial logistic regression

Let $i \in \{1, \ldots, n\}$ denote districts partitioning the 13 studied AGYW priority countries $c[i] \in \{1, \ldots, 13\}$. Consider the years 1999-2018 denoted as $t \in \{1, \ldots, T\}$, and age groups $a \in \{15\text{-}19, 20\text{-}24, 25\text{-}29\}$. Let $p_{itak} > 0$ with $\sum_{k=1}^{3^+} p_{itak} = 1$, be the probabilities of membership of risk group k.

Multinomial logistic regression

A standard multinomial logistic regression model (e.g. Gelman, Carlin, et al. 2013) is specified by

$$\mathbf{y}_{ita} = (y_{ita1}, \ldots, y_{ita3^+})^\top \sim \text{Multinomial}(m_{ita};\, p_{ita1}, \ldots, p_{ita3^+}), \tag{5.1}$$

$$\log\left(\frac{p_{itak}}{p_{ita1}}\right) = \eta_{itak}, \quad k = 2, 3^+, \tag{5.2}$$

where the number in risk group k is y_{itak}, the fixed sample size is $m_{ita} = \sum_{k=1}^{3^+} y_{itak}$, and $k = 1$ is chosen as the baseline category. This model is not an latent Gaussian model [LGM; Rue, Martino, and Chopin (2009)] because each observation y_{itak} for $k \in \{1, 2, 3^+\}$ depends non-linearly on multiple structured additive predictors $\{\eta_{itak}, k = 1, 2, 3^+\}$.

The model, defined over 940 districts, 20 years, 3 age groups, and 3 risk groups, is too large for MCMC to be tractable in reasonable time. To recast this model as an LGM, I used the multinomial-Poisson transformation (detailed in Section 5.3.1). This modification allowed inference to be performed using the INLA (Rue, Martino, and Chopin 2009) algorithm via the `R-INLA` package (Martins et al. 2013). Inferences from INLA in the LGM setting are accurate and substantially faster than MCMC.

The multinomial-Poisson transformation

The multinomial-Poisson transformation (Baker 1994) reframes a given multinomial logistic regression model, like that described in Equations (5.1) and (5.2), as an equivalent Poisson log-linear model. The equivalent model is of the form

$$y_{itak} \sim \text{Poisson}(\kappa_{itak}), \tag{5.3}$$

$$\log(\kappa_{itak}) = \eta_{itak}. \tag{5.4}$$

The basis of the transformation is that conditional on their sum Poisson counts are jointly multinomially distributed (McCullagh and Nelder 1989) as follows

$$\mathbf{y}_{ita} \,|\, m_{ita} \sim \text{Multinomial}\left(m_{ita}; \frac{\kappa_{ita1}}{\kappa_{ita}}, \ldots, \frac{\kappa_{ita3^+}}{\kappa_{ita}}\right), \tag{5.5}$$

where $\kappa_{ita} = \sum_{k=1}^{3^+} \kappa_{itak}$. The probabilities p_{itak} may then be obtained using the softmax function

$$p_{itak} = \frac{\exp(\eta_{itak})}{\sum_{k=1}^{3^+} \exp(\eta_{itak})} = \frac{\kappa_{itak}}{\sum_{k=1}^{3^+} \kappa_{itak}} = \frac{\kappa_{itak}}{\kappa_{ita}}. \tag{5.6}$$

Under the equivalent model, in Equation (5.3) the sample sizes m_{ita} are treated as random rather than fixed such that

$$m_{ita} = \sum_k y_{itak} \sim \text{Poisson}\left(\sum_k \kappa_{itak}\right) = \text{Poisson}\left(\kappa_{ita}\right). \tag{5.7}$$

Using Equations (5.5) for $p(\mathbf{y}_{ita} \,|\, m_{ita})$ and Equation (5.7) for $p(m_{ita})$, the joint distribution is given by

$$p(\mathbf{y}_{ita}, m_{ita}) = \exp(-\kappa_{ita}) \frac{(\kappa_{ita})^{m_{ita}}}{m_{ita}!} \times \frac{m_{ita}!}{\prod_k y_{itak}!} \prod_k \left(\frac{\kappa_{itak}}{\kappa_{ita}}\right)^{y_{itak}} \tag{5.8}$$

$$= \prod_k \left(\frac{\exp(-\kappa_{itak})\,(\kappa_{itak})^{y_{itak}}}{y_{itak}!}\right) \tag{5.9}$$

$$= \prod_k \text{Poisson}\left(y_{itak} \,|\, \kappa_{itak}\right). \tag{5.10}$$

As expected, Equation (5.10) corresponds to the product of independent Poisson likelihoods defined in Equation (5.3). This exercise demonstrates that the Poisson log-linear model contains within it a multinomial likelihood, with a Poission prior on the sample size.

For this model to be equivalent to a multinomial logistic regression model, the normalisation constants m_{ita} must be recovered exactly. That is to say, their posterior distributions should be as close as possible to a Dirac delta distribution with value zero everywhere but the known value of the sample size. To ensure that this is the case, observation-specific random effects θ_{ita} can be included in the equation for the linear predictor. Multiplying each of $\{\kappa_{itak}\}_{k=1}^{3^+}$ by $\exp(\theta_{ita})$ has no effect on the category probabilities, but does provide the necessary flexibility for κ_{ita} to recover m_{ita} exactly. Although in theory an improper prior distribution $\theta_{ita} \propto 1$ should be used, I found that in practice, by keeping η_{ita} otherwise small using appropriate constraints, so that arbitrarily large values of θ_{ita} are not required, it is sufficient (and practically preferable for inference) to instead use a vague prior distribution.

Model specifications

Table 5.2: Four multinomial regression models were considered. Observation random effects θ_{ita}, included in all models, are omitted from this table.

Category β_k	Country ζ_{ck}	Age α_{ack}	Spatial ϕ_{ik}	Temporal γ_{tk}
M1 IID	IID	IID	IID	IID
M2 IID	IID	IID	Besag	IID
M3 IID	IID	IID	IID	AR1
M4 IID	IID	IID	Besag	AR1

I considered four models (Table 5.2) for η_{ita} in the equivalent Poisson log-linear model of the form

$$\eta_{ita} = \theta_{ita} + \beta_k + \zeta_{c[i]k} + \alpha_{ac[i]k} + u_{ik} + \gamma_{tk}. \tag{5.11}$$

Observation random effects $\theta_{ita} \sim \mathcal{N}(0, 1000^2)$ with a vague prior distribution were included in all models to ensure the multinomial-Poisson transformation was valid. To capture country-specific proportion estimates for each category, I included category random effects $\beta_k \sim \mathcal{N}(0, \tau_\beta^{-1})$ and country-category random effects $\zeta_{ck} \sim \mathcal{N}(0, \tau_\zeta^{-1})$. Heterogeneity in risk group proportions by age was allowed by including age-country-category random effects $\alpha_{ack} \sim \mathcal{N}(0, \tau_\alpha^{-1})$. I

considered several specifications for the space-cateogry u_{ik} and time-cateogry effects γ_{tk}, described in Sections 5.3.1 and 5.3.1.

Use of the multinomial-Poisson transformation required all random effects to include interaction with category k, because any random effects which did not include interaction with category would give no change in category probabilities. The only exception were the observation random effects, which were included as a device to ensure the transformation is valid, rather than to model the data.

Spatial random effects For the space-category random effects u_{ik} I considered two specifications:

1. Independent and identically distributed (IID) $u_{ik} \sim \mathcal{N}(0, \tau_u^{-1})$,
2. The Besag improper conditional autoregressive (ICAR) model (Besag et al. 1991) grouped by category

$$\mathbf{u} = (u_{11}, \ldots, u_{n1}, \ldots, u_{13^+}, \ldots u_{n3^+})^\top \sim \mathcal{N}(\mathbf{0}, (\tau_u \mathbf{R}_u^\star)^-).$$

   The scaled structure matrix $\mathbf{R}_u^\star = \mathbf{R}_b^\star \otimes \mathbf{I}$ is given by the Kronecker product of the scaled Besag structure matrix $\mathbf{R}_b^\star$ and the identity matrix $\mathbf{I}$, and $-$ denotes the generalised matrix inverse. I followed best practices for the Besag model as described in Chapter 4. To implement the Kronecker product I used the `group` option in `R-INLA` [Section 3.5.5; Gómez-Rubio (2020)] setting the random effect to be `f(area_idx, model = "besag", group = cat_idx, control.group = list(model = "iid"), ...)`. Though the Kronecker product is symmetric, performance is better in `R-INLA` when the more complicated effect is written as the first variable rather than the grouping variable.

In preliminary testing I used the BYM2 model (Simpson et al. 2017) in place of the Besag. I found that the proportion parameter posteriors tended to be highly peaked at the value one. For simplicity and to avoid numerical issues, by using Besag random effects I effectively decided to fix this proportion to one.

Temporal random effects For the time-category random effects γ_{tk} I considered two specifications:

1. IID $\gamma_{tk} \sim \mathcal{N}(0, \tau_\gamma^{-1})$,
2. First order autoregressive (AR1) grouped by category

$$\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{13^+}, \ldots, \gamma_{T1}, \ldots, \gamma_{T3^+})^\top \sim \mathcal{N}(\mathbf{0}, (\tau_\gamma \mathbf{R}_\gamma^\star)^-).$$

The scaled structure matrix $\mathbf{R}_\gamma^\star = \mathbf{R}_r^\star \otimes \mathbf{I}$ is given by the Kronecker product of a scaled AR1 structure matrix $\mathbf{R}_r^\star$ and the identity matrix $\mathbf{I}$. The AR1 structure matrix $\mathbf{R}_r$ is obtained by precision matrix of the random effects $\mathbf{r} = (r_1, \ldots, r_T)^\top$ specified by

$$r_1 \sim \left(0, \frac{1}{1-\rho^2}\right), \tag{5.12}$$

$$r_t = \rho r_{t-1} + \epsilon_t, \quad t = 2, \ldots, T, \tag{5.13}$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$ and $|\rho| < 1$. As with the structured spatial random effects, I implemented this Kronecker product using the `group` option via `f(year_idx, model = "ar1", group = cat_idx, control.group = list(model = "iid"), ...)`. Again the more complicated variable was written first.

A note on spatio-temporal interaction random effects I also considered including separable space-time-category random effects δ_{itk} in the model, using the specification

$$\boldsymbol{\delta} = (\delta_{111}, \ldots, \delta_{nT3^+})^\top \sim \mathcal{N}(\mathbf{0}, (\tau_\delta \mathbf{R}_\delta^\star)^-), \tag{5.14}$$

where $\mathbf{R}_\delta^\star$ is a Kronecker product of the relevant space, time and category structure matrices. These specifications were:

1. IID spatial and IID temporal (Type I) $\mathbf{R}_\delta^\star = \mathbf{I} \otimes \mathbf{I} \otimes \mathbf{I}$,
2. Besag spatial and IID temporal (Type II) $\mathbf{R}_\delta^\star = \mathbf{R}_b^\star \otimes \mathbf{I} \otimes \mathbf{I}$,
3. IID spatial and AR1 temporal (Type III) $\mathbf{R}_\delta^\star = \mathbf{I} \otimes \mathbf{R}_a^\star \otimes \mathbf{I}$,
4. Besag spatial and AR1 (Type IV) $\mathbf{R}_\delta^\star = \mathbf{R}_b^\star \otimes \mathbf{R}_a^\star \otimes \mathbf{I}$,

where the first, second and third elements of the Kronecker product represent space, time and category (always IID) structure matrices respectively. The interaction type in brackets (e.g. Type I) is given according to the Knorr-Held (2000) framework.

Though three-way Kronecker products are not directly supported in `R-INLA`, I implemented each specification using a combination of the `group` and `replicate` options [Section 6.5.2; Gómez-Rubio (2020)]. For example, for the Type IV effects the random effects were specified by `f(area_idx_copy, model = "besag", group = year_idx, replicate = cat_idx, control.group = list(model = "ar1"))`. I was able to run these models for single countries, keeping only years at which surveys occurred in those countries. However, when fitting all countries jointly I found inclusion of the space-time-category random effects to be intractable, and as such decided not to include them in the model.

Prior distributions All random effect precision parameters $\tau \in \{\tau_\beta, \tau_\zeta, \tau_\alpha, \tau_u, \tau_\gamma, \tau_\delta\}$ were given independent penalised complexity (PC) prior distributions (Simpson et al. 2017) with base model $\sigma = 0$ given by

$$p(\tau) = 0.5\nu\tau^{-3/2}\exp\left(-\nu\tau^{-1/2}\right) \tag{5.15}$$

where $\nu = -\ln(0.01)/2.5$ such that $\mathbb{P}(\sigma > 2.5) = 0.01$. For the lag-one correlation parameter ρ, I used the PC prior distribution, as derived by Sørbye and Rue (2017), with base model $\rho = 1$ and condition $\mathbb{P}(\rho > 0 = 0.75)$. I chose the base model $\rho = 1$ corresponding to no change in behaviour over time, rather than the alternative $\rho = 0$ corresponding to no correlation in behaviour over time, as I judged the former to be more plausible a priori.

Identifiability constraints

To facilitate interpretability of the posterior inferences, I applied sum-to-zero constraints (Table 5.3) such that none of the category interaction random effects altered overall category probabilities. In testing of the space-time-category random

effects, I applied analogous sum-to-zero constraints to maintain roles of the space-category and time-category random effects. In some cases it was not possible to implement all three sets of constraints for the three-way interactions in `R-INLA`.

Table 5.3: Applying sum-to-zero constraints to interaction effects ensures that the main effect is not interfered with.

Random effects	Constraints
Category	$\sum_k \beta_k = 0$
Country	$\sum_c \zeta_{ck} = 0, \forall\, k$
Age-country	$\sum_a \alpha_{ack} = 0, \forall\, c, k$
Spatial	$\sum_i u_{ik} = 0, \forall\, k$
Temporal	$\sum_t \gamma_{tk} = 0, \forall\, k$
Spatio-temporal	$\sum_i \delta_{itk} = 0, \forall\, t, k; \sum_t \delta_{itk} = 0, \forall\, i, k; \sum_k \delta_{itk} = 0, \forall\, i, t$

Survey weighted likelihood

I accounted for the survey design using a weighted pseudo-likelihood where the observed counts y are replaced by effective counts $y^\star$, as described in Section 3.5. These counts may not be integers, and as such the Poisson likelihood given in Equation (5.3) is not appropriate. Instead, I used a generalised Poisson pseudo-likelihood $y^\star \sim \text{xPoisson}(\kappa)$ given by

$$p(y^\star) = \frac{\kappa^{y^\star}}{\lfloor y^\star! \rfloor} \exp\left(-\kappa\right), \tag{5.16}$$

to extend the Poisson distribution to non-integer weighted counts. This working likelihood is implemented by `family = "xPoisson"` in `R-INLA`.

Model selection

I selected the model including Besag spatial random effects and IID temporal random effects based on the conditional predictive ordinate (CPO) criterion (Pettit 1990). For comparison, I also computed the deviance information criterion (DIC) (Spiegelhalter, Best, et al. 2002) and widely applicable information criterion (WAIC) (Watanabe 2013). Each of these criterion can be calculated in `R-INLA` without requiring model refitting. The results are presented in Figure 5.4.

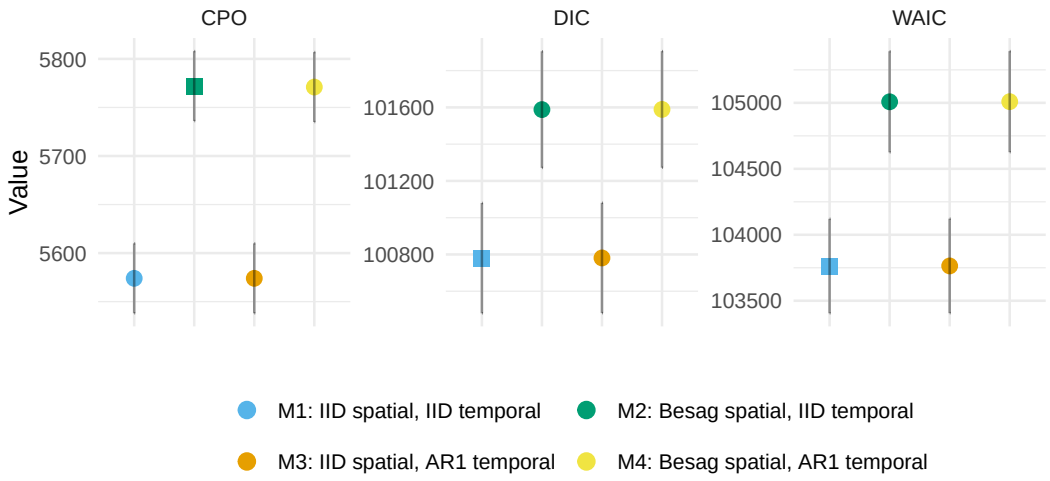

Figure 5.4: For the multinomial logistic regression model, under the CPO criterion, including Besag spatial random effects rather than IID spatial random effects improved model performance. On the other hand, under the DIC and WAIC, where smaller values are preferred, the opposite was true. Though IID temporal random effects are preferred by all criteria AR1 temporal random effects performed very similarly, likely as there is a limited amount of temporal variation in the data to describe.

Table 5.4: CPO, DIC, and WAIC values for the multinomial logistic regression model with corresponding standard errors.

	M1	M2	M3	M4
CPO	5573 (36)	5772 (36)	5574 (36)	5771 (36)
DIC	100780 (300)	101588 (317)	100781 (300)	101589 (317)
WAIC	103763 (358)	105008 (383)	103763 (358)	105009 (383)

5.3.2 Spatial logistic regression

To estimate the proportion of those in the $k = 3^+$ risk group that were in the $k = 3$ and $k = 4$ risk groups respectively, I fit logistic regression models of the form

$$y_{ia4} \sim \text{Binomial}\left(y_{ia3} + y_{ia4}, q_{ia}\right), \tag{5.17}$$

$$q_{ia} = \text{logit}^{-1}\left(\eta_{ia}\right), \tag{5.18}$$

where

$$q_{ia} = \frac{p_{ia4}}{p_{ia3} + p_{ia4}} = \frac{p_{ia4}}{p_{ia3^+}}. \tag{5.19}$$

This two-step approach allowed all surveys to be included in the multinomial regression model, but only those surveys with a specific transactional sex question

to be included in the logistic regression model. As all such surveys occurred in the years 2013-2018 (Figure 5.2), I assumed q_{ia} to be constant with respect to time.

Model specifications

Table 5.5: Six logistic regression models were considered. The covariate `cfswever` denotes the proportion of men who have ever paid for sex and `cfswrecent` denotes the proportion of men who have paid for sex in the past 12 months.

	Intercept β_0	Country ζ_c	Age α_{ac}	Spatial u_i	Covariates
L1	Constant	IID	IID	IID	None
L2	Constant	IID	IID	Besag	None
L3	Constant	IID	IID	IID	`cfswever`
L4	Constant	IID	IID	Besag	`cfswever`
L5	Constant	IID	IID	IID	`cfswrecent`
L6	Constant	IID	IID	Besag	`cfswrecent`

I considered six logistic regression models (Table 5.5). Each included a constant intercept $\beta_0 \sim \mathcal{N}(-2, 1^2)$, country random effects $\zeta_c \sim \mathcal{N}(0, \tau_\zeta^{-1})$, and age-country random effects $\alpha_{ac} \sim \mathcal{N}(0, \tau_\alpha^{-1})$. The Gaussian prior distribution on β_0 placed 95% prior probability on the range 2-50% for the percentage of those with non-regular or multiple partners who report transactional sex. I considered two specifications (IID, Besag) for the spatial random effects u_i. To aid estimation with sparse data, I also considered national-level covariates for the proportion of men who have paid for sex ever `cfswever` or in the last twelve months `cfswrecent` (Hodgins et al. 2022). For both random effect precision parameters $\tau \in \{\tau_\alpha, \tau_\zeta\}$ I used the PC prior distribution with base model $\sigma = 0$ and $\mathbb{P}(\sigma > 2.5 = 0.01)$. For both regression parameters $\beta \in \{\beta_{\texttt{cfswever}}, \beta_{\texttt{cfswrecent}}\}$ I used the prior distribution $\beta \sim \mathcal{N}(0, 2.5^2)$.

Survey weighted likelihood

As with the multinomial regression model, I used survey weighted counts $y^\star$ and sample sizes $m^\star$. I used a generalised binomial pseudo-likelihood $y^\star \sim$ xBinomial$(m^\star, q)$ given by

$$p(y^\star \mid m^\star, q) = \binom{\lfloor m^\star \rfloor}{\lfloor y^\star \rfloor} q^{y^\star} (1-q)^{m^\star - y^\star} \tag{5.20}$$

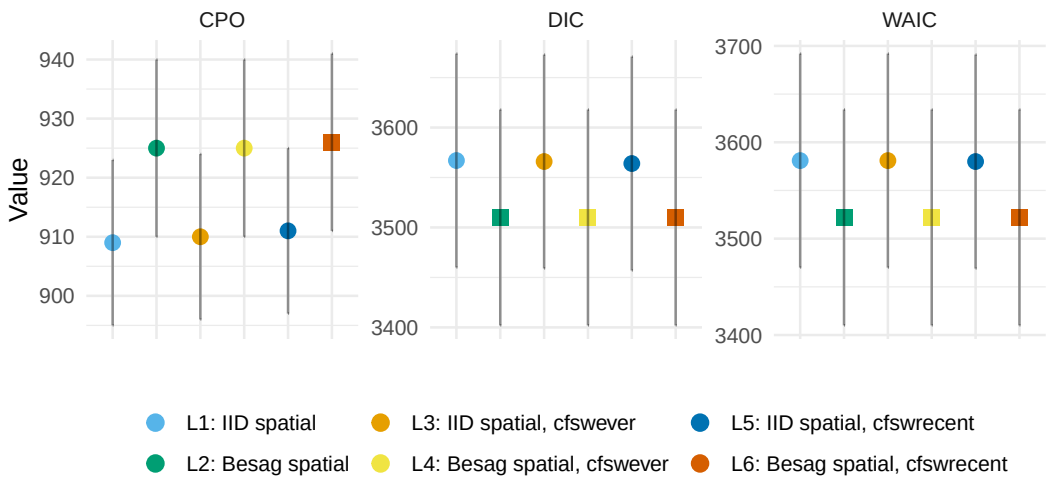

Figure 5.5: For the logistic regression model, the CPO, DIC, and WAIC each agreed that the model containing Besag spatial random effects and the `cfswrecent` covariates was best. Inclusion of Besag spatial random effects consistently improved each criterion, whereas improvements from inclusion of any covariates were marginal.

to extend the binomial distribution to non-integer weighted counts and sample sizes. This working likelihood is implemented by `family = "xBinomial"` in `R-INLA`.

Model selection

I selected the model including Besag spatial effects and `cfswrecent` covariates according to the CPO criterion. All results, including DIC and WAIC, are presented in Table and Figure 5.5. Inclusion of Besag spatial random effects, rather than IID, consistently improved performance. Benefits from inclusion of covariates were more marginal. As some countries had no suitable surveys, I nonetheless preferred to include covariate information so that estimates in these countries would be based on some country-specific data.

Table 5.6: CPO, DIC, and WAIC values for the logistic regression model with corresponding standard errors.

	L1	L2	L3	L4	L5	L6
DIC	4662 (110)	4605 (111)	4662 (110)	4605 (111)	4662 (110)	4605 (111)
WAIC	4692 (115)	4624 (115)	4692 (115)	4624 (115)	4692 (115)	4624 (115)
CPO	950 (15)	969 (15)	951 (15)	970 (15)	950 (15)	970 (15)

5.3.3 Model combination

How were the models combined? Using samples.

5.3.4 Female sex worker population size adjustment

Having had sex "in return for gifts, cash or anything else in the past 12 months" is not considered sufficient to constitute sex work. As such, I adjusted the estimates obtained based on the transactional sex survey question to match FSW population size estimates obtained using an alternative method, which I describe below. The estimates of the non-regular or multiple sexual partner(s) population size were changed to facilitate changing of the FSW population size. This approach retained subnational variation informed by the transactional sex survey question.

I used the estimates adult (15-49) FSW population size by country from a Bayesian meta-analysis of key population specific data sources (Stevens, Sabin, Arias Garcia, et al. 2022). To disaggregate these estimates by age, I took the following steps. First, I calculated the total sexually debuted population in each age group, by country. To describe the distribution of age at first sex, I used skew logistic distributions (Nguyen and Eaton 2022) with cumulative distribution function given by

$$F(x) = (1 + \exp(\kappa_c(\mu_c - x)))^{-\gamma_c}, \tag{5.21}$$

where $\kappa_c, \mu_c, \gamma_c > 0$ are country-specific shape, shape and skewness parameters respectively. Next, I used the assumed Gamma($\alpha = 10.4, \beta = 0.36$) FSW age distribution in South Africa from the Thembisa model (Johnson and Dorrington 2020) to calculate the implied ratio between the number of FSW and the sexually debuted population in each age group. I assumed the South African ratios were applicable to every country, allowing calculation of the number of FSW by age group in all 13 countries. The resulting age trends obtained (Figure 5.6) reflect country-level variation in demographics and age-at-first-sex.

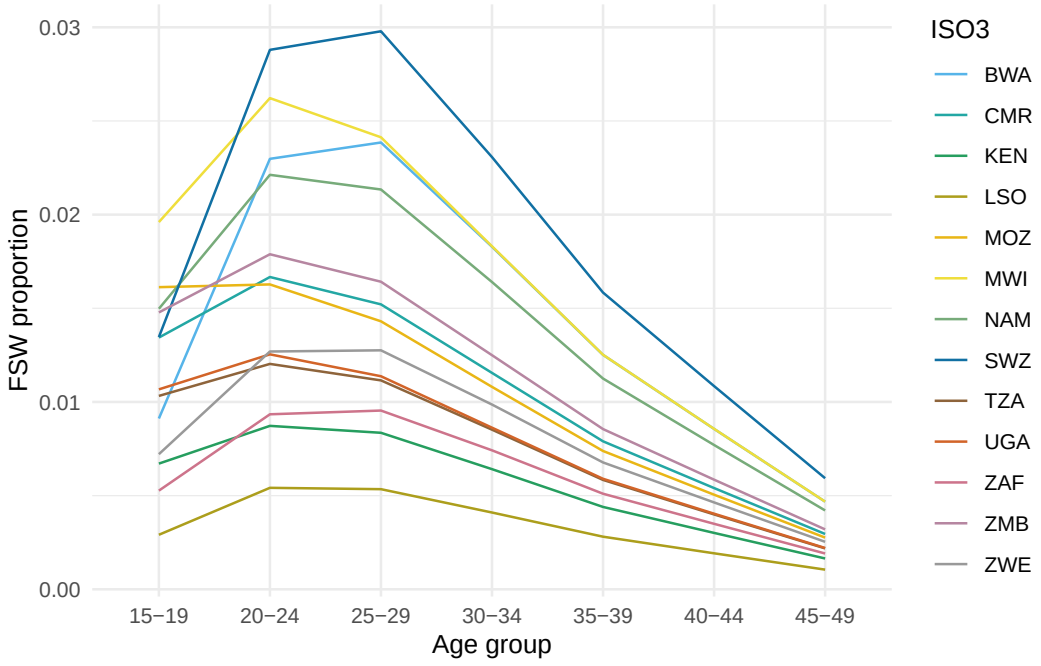

Figure 5.6: The disaggregation procedure I used produces an age distribution for FSW peaking in the 20-24 and 25-29 age groups, and declining for older age groups.

5.3.5 Results

Coverage assessment

To assess the calibration of the fitted model, I calculated the quantile q of each observation within the posterior predictive distribution. For calibrated models, these quantiles, known as probability integral transform (PIT) values (Dawid 1984; Bosse et al. 2022), should follow a uniform distribution $q \sim \mathcal{U}[0, 1]$. To generate samples from the posterior predictive distribution, I applied the multinomial likelihood to samples from the latent field, setting the sample size to be the floor of the Kish effective sample size. Using the PIT values, it is possible to calculate the empirical coverage of all $(1-\alpha)100\%$ equal-tailed posterior predictive credible intervals. These empirical coverages can be compared to the nominal coverage $(1-\alpha)$ for each value of $\alpha \in [0, 1]$ to give empirical cumulative distribution function (ECDF) difference values. This approach has the advantage of considering all possible confidence values at once. To test for uniformity, I used the binomial distribution based simultaneous confidence bands for ECDF difference values developed by Säilynoja et al. (2021). I found the only significant deviation from uniformity occurred in

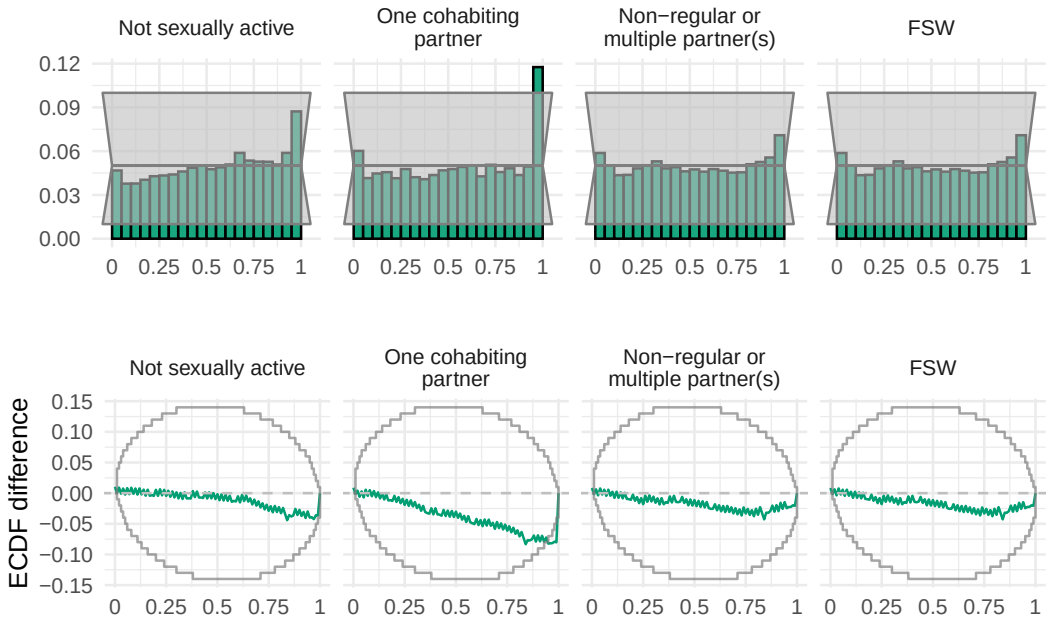

Figure 5.7: Probability integral transform (PIT) histograms (top row) and empirical cumulative distribution function (ECDF) difference plots (bottom row) for the final selected model.

the right-hand tail of the one cohabiting partner risk group. That is to say, the proportion of the PIT values which were greater than 0.95 was significantly more than would be expected under a calibrated model.

Estimates

Figure 5.11 and Figure 5.9 show posterior mean estimates for the proportion in each risk group for the final model in 2018, the most recent year included in our analysis. I focused on the most recent estimates because they are the most relevant to inform ongoing HIV policy. In subsequent results, all estimates refer to 2018, unless otherwise indicated.

The median national FSW proportion was 1.1% (95% CI 0.4–1.9) for the 15-19 age group, 1.6% (95% CI 0.6–2.8) for the 20-24 age group and 1.9% (95% CI 0.5–3.5) for the 25-29 age group, in line with the results displayed in Figure 5.6.

In the 20-24 and 25-29 year age groups, the majority of women were either cohabiting or had non-regular or multiple partner(s). Countries in eastern and

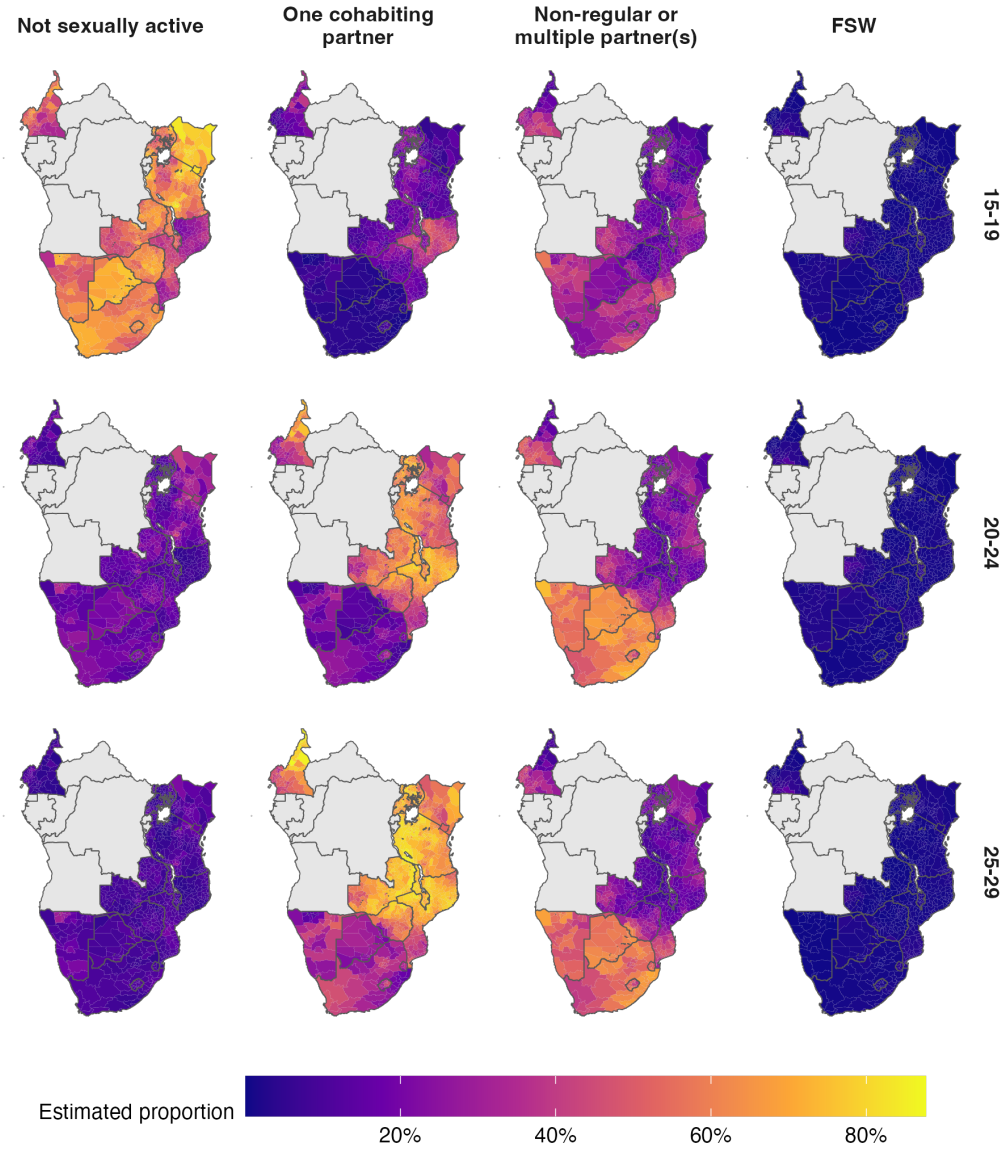

Figure 5.8: The spatial distribution (posterior mean) of the AGYW risk group proportions in 2018. Estimates are stratified by risk group (columns) and five-year age group (rows). Countries in grey were not included in the analysis. A limitation of this figure is that using a common colour scale, desirable for other reasons, makes it challenging to see spatial variation in the FSW risk group.

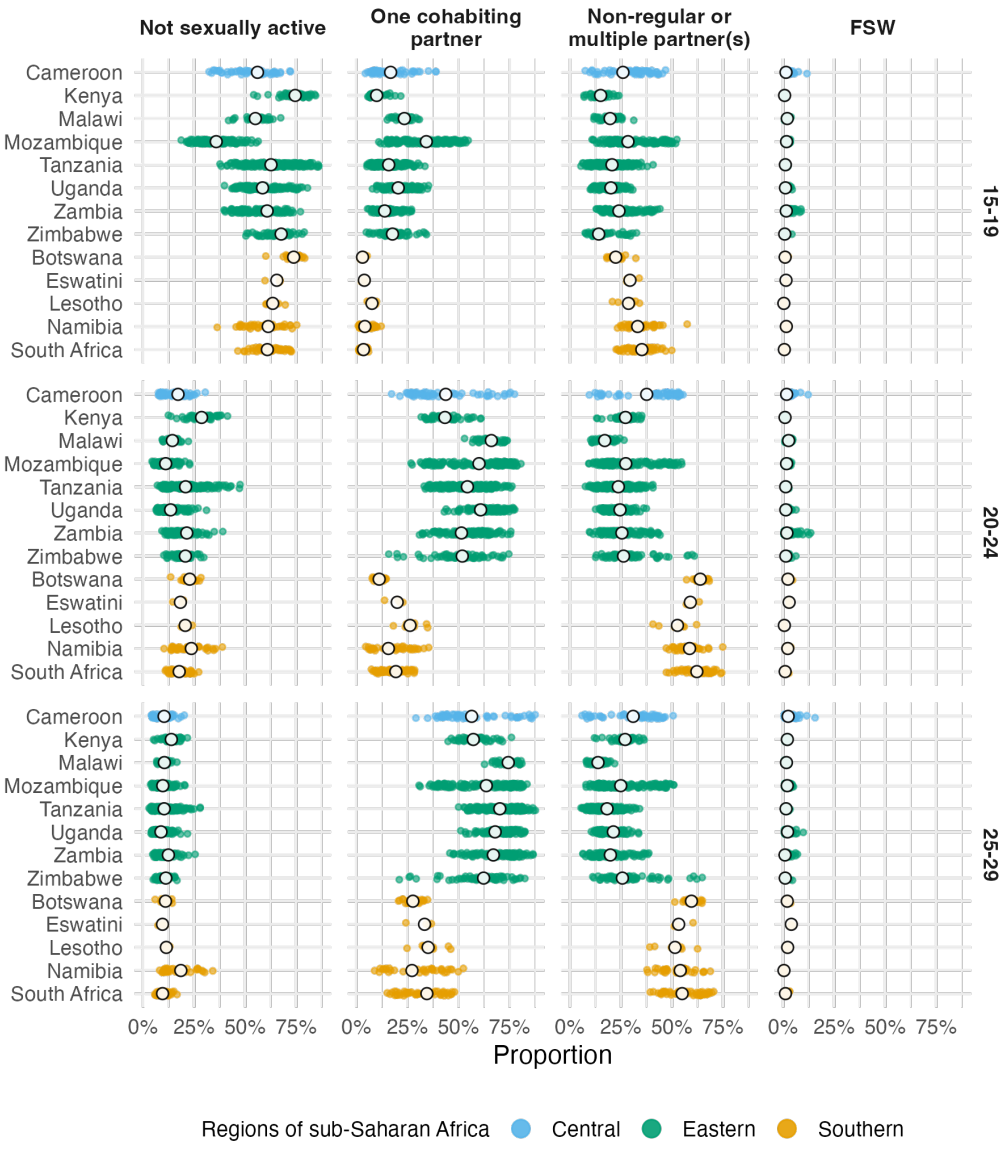

Figure 5.9: National (in white) and subnational (in color) posterior means of the risk group proportions. Estimates are stratified by risk group (columns) and five-year age group (rows). Though the information presented is similar to that of Figure 5.8, this figure presents a clear view of within- and between-country variation in risk group proportions.

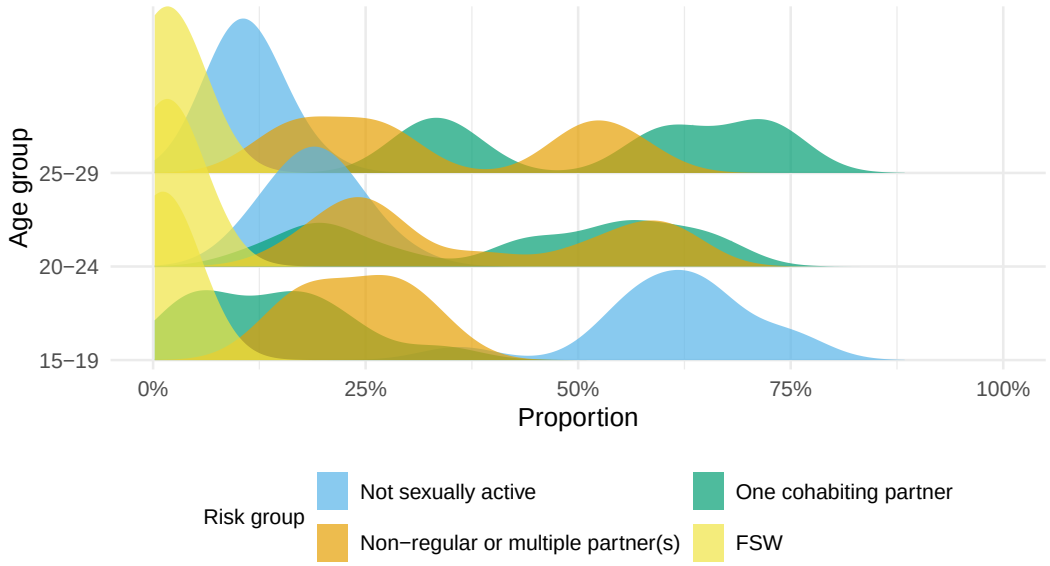

Figure 5.10: Figure caption.

central Africa (Cameroon, Kenya, Malawi, Mozambique, Tanzania, Uganda, Zambia and Zimbabwe) had a higher proportion of women in these age groups cohabiting (63.1% [95% CI 35–78.7%] compared with 21.3% [95% CI 10.1–48.8%] with non-regular partner[s]). In contrast, countries in southern Africa (Botswana, Eswatini, Lesotho, Namibia and South Africa) had a higher proportion with non-regular or multiple partner(s) (58.9% [95% CI 43.2–70.5%], compared with 23.4% [95% CI 9.7–39.1%] cohabiting). This finding is the most notable feature of between-country variation shown in Figure 5.9. Figure 5.8 shows the geographic delineation to pass along the border of Mozambique, through the interior of Zimbabwe and along the border of Zambia.

In most districts (57.9%; 95% credible interval [CI] 27.7–79.7) adolescent girls aged 15-19 were not sexually active. The exception was Mozambique, where the majority (64.23%) were sexually active in the past year and close to a third (34.17%) were cohabiting with a partner.

Variance decomposition

Age group was the most important factor explaining variation in risk group proportions, accounting for 65.9% (95% CI 54.1–74.9%) of total variation. The

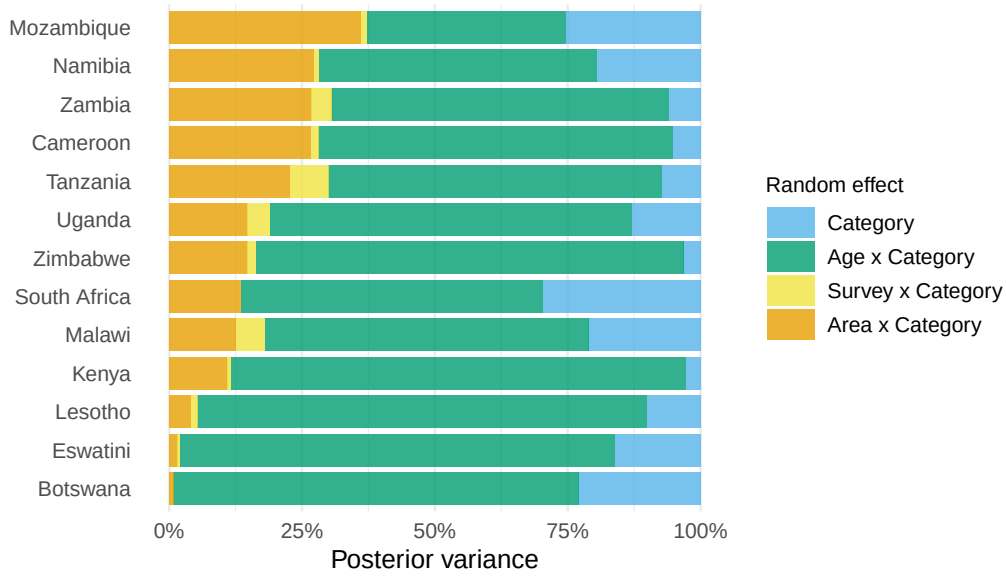

Figure 5.11: Figure caption.

primary change in risk group proportions by age group occurs between the 15-19 age group and 20-29 age group (Figure 5.8). The next most important factor was location. Country-level differences explained 20.9% (95% CI 11.9–34.5%) of variation, while district-level variation within countries explained 11.3% (95% CI 8.2–15.3%). Temporal changes only explained 0.9% (95% CI 0.6–1.4%) of variation, indicating very little change in risk group proportions over time. I found similar variance decomposition results fitting each country individually (Figure 5.11) and using other model specifications.

5.4 Prevalence and incidence by risk group

Using the most recent risk group proportion estimates, I calculated the following indicators stratified according to district, age group and risk group:

1. HIV prevalence ρ_{iak},
2. the number of people living with HIV (PLHIV) H_{iak},
3. HIV incidence λ_{iak}, and
4. the number of new HIV infections I_{iak}.

To do so, I disaggregated district, age group specific Naomi estimates by risk group.

5.4.1 Disaggregation of Naomi prevalence estimates

To disaggregate HIV prevalence, I began by estimating HIV prevalence log odds ratios $\log(\text{OR}_k)$ relative to the general population. To do so, I fit a logistic regression model using age, country and risk group specific HIV prevalence bio-marker survey data. I also included general population HIV prevalence data. The logistic regression model included an indicator function for each risk group, and an indicator for being in the general population, such that the regression coefficients in this model correspond to log odds. The log odds ratios may then be easily obtained by taking the difference in odds ratios.

To allow the log odds ratio for the highest risk group to vary based on general population prevalence I fit a linear regression of the FSW log odds against the general population log odds. I ensured that log odds ratios for the FSW risk group were at least as large as those for the multiple or non-regular partner(s) risk group.

Given the fitted log odds ratios, I disaggregated Naomi estimates of PLHIV H_{ia} on the logit scale using numerical optimisation. I did this by by finding the values of θ_{ia} which minimise the equation

$$f(\theta_{ia}) = \sum_{k=1}^{4} (\text{logistic}(\theta_{ia} + \log(\text{OR}_k)) \cdot N_{iak}) - H_{ia}, \tag{5.22}$$

where

$$\text{logistic}(x) = \exp(x)/(1 + \exp(x)), \tag{5.23}$$

such that

$$\text{logistic}(\hat{\theta}_{ia} + \log(\text{OR}_k)) = \rho_{iak}. \tag{5.24}$$

These values are given by

$$\hat{\theta}_{ia} = \arg \min_{\theta_{ia} \in [-10,10]} f(\theta_{ia})^2. \tag{5.25}$$

The number of PLHIV were obtained by $H_{iak} = \rho_{iak} N_{iak}$, where N_{iak} is the risk group population size.

5.4.2 Disaggregation of Naomi incidence estimates

I calculated the number of new HIV infections by risk group using linear disaggregation

$$I_{ia} = \sum_k I_{iak} = \sum_k \lambda_{iak}(1-\rho_{iak})N_{iak} \tag{5.26}$$

$$= 0 + \lambda_{ia2}(1-\rho_{ia2})N_{ia2} + \lambda_{ia3}(1-\rho_{ia3})ia3 + \lambda_{ia4}(1-\rho_{ia4})N_{ia4} \tag{5.27}$$

$$= \lambda_{ia2}\left((1-\rho_{ia2})N_{ia2} + \mathrm{RR}_3(1-\rho_{ia3})N_{ia3} + \mathrm{RR}_4(\lambda_{ia})(1-\rho_{i4})N_{ia4}\right), \tag{5.28}$$

where RR_2, RR_3 and $\mathrm{RR}_4(\cdot)$ are the HIV risk ratios given in Table 5.1, and $(1-\rho_{iak})N_{iak}$ are the susceptible population sizes in each risk group. The risk ratio for FSW was defined as a function of district-level incidence in the general population λ_{ia}.

Risk group specific HIV incidence estimates were then given by

$$\lambda_{ia1} = 0, \tag{5.29}$$

$$\lambda_{ia2} = I_{ia}/\left((1-\rho_{ia2})N_{ia2} + \mathrm{RR}_3(1-\rho_{ia3})N_{ia3} + \mathrm{RR}_4(\lambda_{ia})(1-\rho_{ia4})N_{ia4}\right), \tag{5.30}$$

$$\lambda_{ia3} = \mathrm{RR}_3\lambda_{ia2}, \tag{5.31}$$

$$\lambda_{ia4} = \mathrm{RR}_4(\lambda_{ia})\lambda_{ia2}. \tag{5.32}$$

I evaluated these equations using Naomi model estimates of the number of new HIV infections $I_{ia} = \lambda_{ia}N_{ia}$. The number of new HIV infections were $I_{iak} = \lambda_{iak}N_{iak}$.

5.4.3 Expected new infections reached

I calculated the number of new infections that would be reached prioritising according to each possible stratification of the population. That is, for all $2^3 = 8$ possible combinations of stratification by location, age, and risk group.

To illustrate this approach, consider stratification by age. I first aggregated the number of new HIV infections and HIV incidence such that

$$I_a = \sum_{ik} I_{iak}, \tag{5.33}$$

$$\lambda_a = I_a/\sum_{ik}(1-\rho_{iak})N_{iak}. \tag{5.34}$$

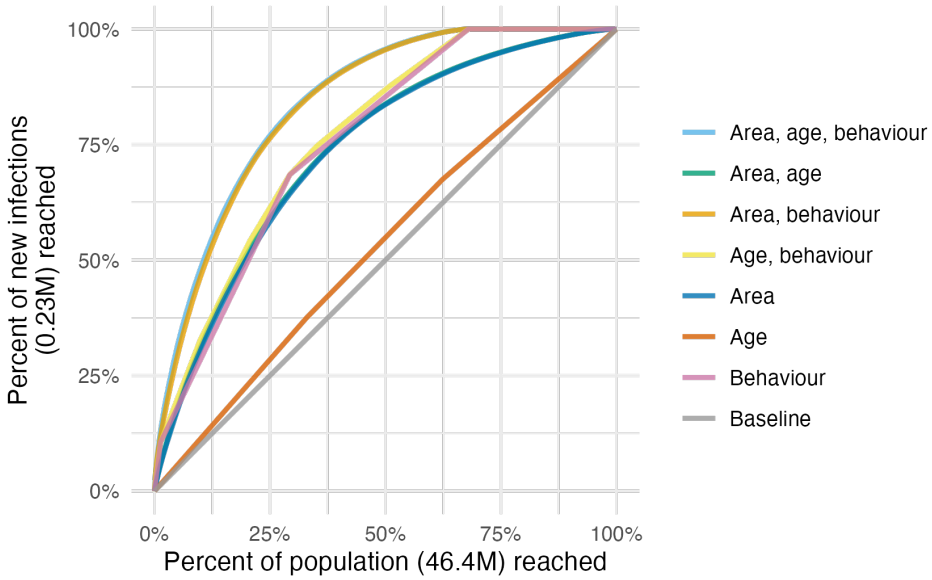

Figure 5.12: Percentage of new infections reached across all 13 countries, taking a variety of risk stratification approaches, against the percentage of at risk population required to be reached.

I then considered prioritisation individuals by age group a according to the highest HIV incidence λ_a. By cumulatively summing the expected infections, for each fraction of the total population reached I calculated the fraction of total expected new infections that would be reached. As there are three age groups, the resulting function was piecewise linear with three segments.

5.4.4 Results

For any given fraction of AGYW prioritised, substantially more new infections were reached by strategies that included behavioural risk stratification. Reaching half of all expected new infections required reaching 19.4% of the population when stratifying by subnational area and age, but only 10.6% when behavioural stratification was included (Figure 5.12). The majority of this benefit came from reaching FSW, who were 1.3% of the population but 10.6% of all new infections.

Considering each country separately, on average, reaching half of new infections in each country required reaching 14.6% (range 8.7-21.8%) of the population when stratifying by area and age, reducing to 5.1% (range 2.1-13.2%) when behaviour

was included. The relative importance of stratifying by age, location and behaviour varied between countries, analogous to the varying contribution of each to the total variance (Section 5.3.5).

5.5 Discussion

In this chapter, I estimated the proportion of AGYW who fall into different risk groups at a district level in 13 sub-Saharan African countries. These estimates support consideration of differentiated prevention programming according to geographic locations and risk behaviour, as outlined in the Global AIDS Strategy. Systematic differences in risk by age groups, and variation within and between countries, explained the large majority of variation in risk group proportions. Changes over time were negligible in the overall variation in risk group proportions. The proportion of 15-19 year olds who are sexually active, and among women aged 20-29 years, norms around cohabitation especially varied across districts and countries. This variation underscores the need for these granular data to implement HIV prevention options aligned to local norms and risk behaviours.

I considered four risk groups based on sexual behaviour, the most proximal determinant of risk. Other factors, such as condom usage or type of sexual act, may account for additional heterogeneity in risk from sexual behaviour. However, I did not include these factors in view of measurement difficulties, concerns about consistency across contexts, and the operational benefits of describing risk parsimoniously.

Sexual behaviour confers risk only when AGYW reside in geographic locations where there is unsuppressed viral load among their potential partners. I did not include more distal determinants, such as school attendance, orphanhood, or gender empowerment, as I expect their effects on risk to largely be mediated by more proximal determinants. However, to effectively implement programming, it is crucial to understand these factors, as well as the broader structural barriers and limits

to personal agency faced by AGYW. Importantly, programs must ensure that intervention prioritisation occurs without stigmatising or blaming AGYW.

By considering a range of possible risk stratification strategies, I showed that successful implementation of a risk-stratified approach would allow substantially more of those at risk for infections to be identified before infection occurs. A considerable proportion of estimated new infections were among FSW, supporting the case for HIV programming efforts focused on key population groups (Baral et al. 2012). There is substantial variation in the importance of prioritisation by age, location and behaviour within each country. This highlights the importance of understanding and tailoring HIV prevention efforts to country-specific contexts. By standardising the analysis across all 13 countries, I showed the additional efficiency benefits of resource allocation between countries.

I found a geographic delineation in the proportion of women cohabiting between southern and eastern Africa, calling attention to a divide attributable to many cultural, social, and economic factors. The delineation does not represent a boundary between predominately Christian and Muslim populations, which is further north. I also note that the high numbers of adolescent girls aged 15-19 cohabiting in Mozambique is markedly different from the other countries (UNICEF 2019).

Brugh et al. (2021) previously geographically mapped AGYW HIV risk groups using biomarker and behavioural data from the most recent surveys in Eswatini, Haiti and Mozambique to define and subsequently map risk groups with a range of machine learning techniques. My work builds on Brugh et al. (2021) by including more countries, integrating a greater number of surveys, and connecting risk group proportions with HIV epidemic indicators to help inform programming.

My modelled estimates of risk group proportions improve upon direct survey results for three reasons. First, by taking a modular modelling approach, I integrated all relevant survey information from multiple years, allowing estimation of the FSW proportion for surveys without a specific transactional sex question. Second, whereas direct estimates exhibit large sampling variability at a district level, I alleviated this issue using spatio-temporal smoothing (Figure 5.13). Third, I provided estimates

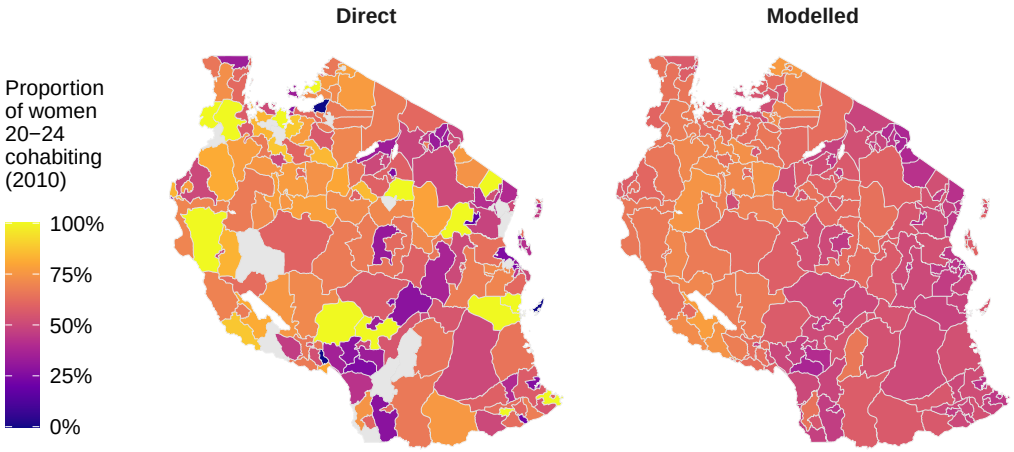

Figure 5.13: Figure caption.

in all district-years, including those not directly sampled by surveys, allowing estimates to be consistently fed into further analysis and planning pipelines such as my analysis of risk group specific prevalence and incidence (Figure 5.13).

The final surveys included in the risk model model were conducted in 2018. The analysis may be updated with more surveys as they become available. I do not anticipate that the risk group proportions will change substantially, as I found that they did not change significantly over time.

My analysis focused on females aged 15-29 years, and could be extended to consider optimisation of prevention more broadly, accounting for the 0% of new infections among adults 15-49 which occur in women 30-49 and men 15-49. Estimating sexual risk behaviour in adults 15-49 would be a crucial step toward greater understanding of the dynamics of the HIV epidemic in sub-Saharan Africa, and would allow incidence models to include stratification of individuals by sexual risk.

5.5.1 Limitations

This analysis was subject to challenges shared by most approaches to monitoring sexual behaviour in the general population (Cleland et al. 2004). In particular,

under-reporting of higher risk sexual behaviours among AGYW could affect the validity of my risk group proportion estimates. Due to social stigma or disapproval, respondents may be reluctant to report non-marital partners (Nnko et al. 2004; Helleringer et al. 2011) or may bias their reporting of sexual debut (Zaba et al. 2004; Wringe et al. 2009; Nguyen and Eaton 2022). For guidance of resource allocation, differing rates of under-reporting by country, district, year or age group are particularly concerning to the applicability of my results; and, while it may be reasonable to assume a constant rate over space-time, the same cannot be said for age, where aspects of under-reporting have been shown to decline as respondents age (Glynn et al. 2011), suggesting that the elevated risks I found faced by younger women are likely a conservative estimate. If present, these reporting biases will also have distorted the estimates of infection risk ratios and prevalence ratios I used in my analysis, likely over-attributing risk to higher risk groups.

I have the least confidence in my estimates for the FSW risk group. As well as having the smallest sample sizes, my transactional sex estimates do not overcome the difficulties of sampling hard to reach groups. I inherent any limitations of the national FSW estimates (Stevens, Sabin, Arias Garcia, et al. 2022) which I adjust my estimates of transactional sex to match. Furthermore, I do not consider seasonal migration patterns, which may particularly affect FSW population size. More generally, I did not consider covariates potentially predictive of risk group proportions (such as sociodemographic characteristics, education, local economic activity, cultural and religious norms and attitudes), which are typically difficult to measure spatially. Identifying measurable correlates of risk, or particular settings in which time-concentrated HIV risk occurs, is an important area for further research to improve risk prioritisation and precision HIV programme delivery.

The efficiency of each stratified prevention strategy depends on the ability of programmes to identify and effectively reach those in each strata. My analysis of new infections potentially averted assumed a "best-case" scenario where AGYW of every strata can be reached perfectly, and should therefore be interpreted as illustrating the potentially obtainable benefits rather than benefits which would be

obtained from any specific intervention strategy. In practice, stratified prevention strategies are likely to be substantially less efficient than this best-case scenario. Factors I did not consider include the greater administrative burden of more complex strategies, variation in difficulty or feasibility of reaching individuals in each strata, variation in the range or effectiveness of interventions by strata, and changes in strata membership that may occur during the course of a year. Identifying and reaching behavioural strata may be particularly challenging. Empirical evaluations of behavioural risk screening tools have found only moderate discriminatory ability (Jia et al. 2022), and risk behaviour may change rapidly among young populations, increasing the challenge to effectively deliver appropriately timed prevention packages. This consideration may motivate selecting risk groups based on easily observable attributes, such as attendance of a particular service or facility, rather than sexual behaviour.

In conducting this work, there was insufficient engagement with country experts or civil society organisations. As a result, in early use of the risk group tool the FSW population size estimates were met with some disagreement in Malawi. In that instance, the cause of the disagreement was external model inputs used. In future, estimates should be generated and reviewed by country teams.

5.5.2 Conclusion

I estimated HIV risk group proportions, HIV prevalences and HIV incidences for AGYW aged 15-19, 20-24 and 25-29 years at a district-level in 13 priority countries. Using these estimates, I analysed the number of infections that could be reached by prioritisation based upon location, age and behaviour. Though subject to limitations, these estimates provide data that national HIV programmes can use to set targets and implement differentiated HIV prevention strategies as outlined in the Global AIDS Strategy. Successfully implementing this approach would result in more efficiently reaching a greater number of those at risk of infection.

Among AGYW, there was systematic variation in sexual behaviour by age and location, but not over time. Age group variation was primarily attributable to age

of sexual debut (ages 15-24). Spatial variation was particularly present between those who reported one cohabiting partner versus non-regular or multiple partners. Risk group proportions did not change substantially over time, indicating that norms relating to sexual behaviour are relatively static. These findings underscore the importance of providing effective HIV prevention options tailored to the needs of particular age groups, as well as local norms around sexual partnerships.

6

Fast approximate Bayesian inference

This chapter describes the development of a novel Bayesian inference method, motivated by the Naomi small-area estimation model (Eaton et al. 2021), tackling both implementation and methological challenges. Over 35 countries (UNAIDS 2023b) have used the Naomi model web interface (`https://naomi.unaids.org` to produce subnational estimates of HIV indicators. Evidence is synthesised from household surveys and routinely collected health data to generate estimates of HIV indicators by district, age, and sex. The complexity and size of the model makes obtaining fast and accurate Bayesian inferences challenging.

The methods developed in this chapter combine Laplace approximations with adaptive quadrature, and are descended from the integrated nested Laplace approximation (INLA) method pioneered by Rue, Martino, and Chopin (2009). The INLA method has enabled fast and accurate Bayesian inferences for a vast array of models, across a large number of scientific fields (Rue, Riebler, et al. 2017). The success of INLA is in large part due to its accessible implementation in the `R-INLA` software. Use of the INLA method and the `R-INLA` software are close to ubiquitous in applied settings. However, the Naomi model is not compatible with `R-INLA`, foremost because it is too complex to be expressed using a formula interface. As a result, inferences for the Naomi model have previously been obtained using an empirical

Bayes [EB; Casella (1985)] approximation to full Bayesian inference, using a Laplace approximation implemented by the more flexible Template Model Builder [`TMB`; Kristensen et al. (2016)] R package. In the EB approximation hyperparameters are fixed by optimisation of the marginal posterior. This is undesirable as it results in underestimation of uncertainty, which may ultimately lead to worse policy decisions which stem from overconfidence.

Most methodological work relating to INLA has taken place using the `R-INLA` software package. There are two notable exceptions. First, the simplified INLA approach of Wood (2020), implemented in the `mgcv` R package, proposed an fast Laplace approximation approach which does not rely on Markov structure of the latent field. Second, Stringer et al. (2022) extended the scope and scalability of INLA by avoiding augmenting the latent field with the noisy structured additive predictors. This enables the application of INLA to a wider class of extended latent Gaussian models, which includes Naomi. Van Niekerk et al. (2023) refer to this as the "modern" formulation of the INLA method, as opposed to the "classic" formulation of Rue, Martino, and Chopin (2009), and it is now included in `R-INLA` using `inla.mode = "experimental"`. Stringer et al. (2022) also propose use of the adaptive Gauss-Hermite quadrature [AGHQ; Naylor and Smith (1982)] rule to perform integration with respect to the hyperparameters.

The methodological contributions of this chapter extend Stringer et al. (2022) in two directions:

1. First, a universally applicable implementation of INLA with Laplace marginals, where automatic differentiation via `TMB` is used to obtain the derivatives required for the Laplace approximation. Section 6.2 demonstrates the implementation using two examples, one compatible with `R-INLA` and one incompatible.
2. Second, a quadrature rule which combines AGHQ with principal components analysis to enable integration over moderate-dimensional spaces, described in Section 6.4. This quadrature rule is used to perform inference for the

Naomi model by integrating the marginal Laplace approximation with respect to the moderate-dimensional hyperparameters within an INLA algorithm implemented in `TMB` in Section 6.5.

This work was conducted in collaboration with Prof. Alex Stringer, whom I visited at the University of Waterloo during the fall term of 2022. The results are presented in Howes, Stringer, et al. (2023+). Code for the analysis in this chapter is available from `https://github.com/athowes/naomi-aghq`.

6.1 Inference methods and software

This section reviews existing deterministic Bayesian inference methods (Sections 6.1.1, 6.1.2, 6.1.3) and the software implementing them (Section 6.1.4). Inference comprises obtaining the posterior distribution

$$p(\boldsymbol{\phi} \,|\, \mathbf{y}) = \frac{p(\boldsymbol{\phi}, \mathbf{y})}{p(\mathbf{y})}, \tag{6.1}$$

or some way to compute relevant functions of it. The posterior distribution encapsulates beliefs about the parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_d)$ having observed data $\mathbf{y} = (y_1, \ldots, y_n)$. Here I assume these quantities are expressible as vectors.

Inference is a sensible goal because (under Bayesian decision theory) the posterior distribution is sufficient for use in decision making. More specifically, given a loss function $l(a, \boldsymbol{\phi})$, the expected posterior loss of a decision a depends on the data only via the posterior distribution

$$\mathbb{E}(l(a, \boldsymbol{\phi}) \,|\, \mathbf{y}) = \int_{\mathbb{R}^d} l(a, \boldsymbol{\phi}) p(\boldsymbol{\phi} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\phi}. \tag{6.2}$$

For example, historic data about treatment demand are only required for planning of HIV treatment service provision in so far as they alter the posterior distribution of current demand. The information provided for strategic response to the HIV epidemic may therefore be thought of as functions of some posterior distribution.

It is usually intractable to obtain the posterior distribution. This is because the denominator in Equation (6.1) contains a potentially high-dimensional integral over the $d \in \mathbb{Z}^+$ -dimensional parameters

$$p(\mathbf{y}) = \int_{\mathbb{R}^d} p(\mathbf{y}, \boldsymbol{\phi}) \mathrm{d}\boldsymbol{\phi}. \tag{6.3}$$

This quantity is sometimes called the evidence or posterior normalising constant. As a result, approximations to the posterior distribution $\tilde{p}(\boldsymbol{\phi} \,|\, \mathbf{y})$ are typically used in place of the exact posterior distribution.

Some approximate Bayesian inference methods, like Markov chain Monte Carlo (MCMC), avoid directly calculating the posterior normalising constant. Instead they find ways to work with the unnormalised posterior distribution

$$p(\boldsymbol{\phi} \,|\, \mathbf{y}) \propto p(\boldsymbol{\phi}, \mathbf{y}), \tag{6.4}$$

where $p(\mathbf{y})$ is not a function of $\boldsymbol{\phi}$ and so can be removed as a constant. Other approximate Bayesian inference methods can more directly be thought of as ways to estimate the posterior normalising constant (Equation (6.3)). The methods in this chapter fall into this latter category, and are sometimes described as deterministic Bayesian inference methods because they do not make fundamental use of randomness.

6.1.1 The Laplace approximation

Laplace's method (Laplace 1774) is a technique used to approximate integrals of the form

$$\int \exp(Ch(\mathbf{z})) \mathrm{d}\mathbf{z}, \tag{6.5}$$

where $C > 0$ is a constant, h is a function which is twice-differentiable, and $\mathbf{z}$ are generic variables. The Laplace approximation (Tierney and Kadane 1986) is obtained by application of Laplace's method to calculate the posterior normalising constant (Equation (6.3)). Let $h(\boldsymbol{\phi}) = \log p(\boldsymbol{\phi}, \mathbf{y})$ such that

$$p(\mathbf{y}) = \int_{\mathbb{R}^d} p(\mathbf{y}, \boldsymbol{\phi}) \mathrm{d}\boldsymbol{\phi} = \int_{\mathbb{R}^d} \exp(h(\boldsymbol{\phi})) \mathrm{d}\boldsymbol{\phi}. \tag{6.6}$$

Laplace's method involves approximating the function h by its second order Taylor expansion. This expansion is then evaluated at a maxima of h to eliminate the first order term. Let

$$\hat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}} h(\boldsymbol{\phi}) \tag{6.7}$$

be the posterior mode, and

$$\hat{\mathbf{H}} = -\frac{\partial^2}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^\top} h(\boldsymbol{\phi})|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}} \tag{6.8}$$

be the Hessian matrix evaluated at the posterior mode. The Laplace approximation to the posterior normalising constant (Equation (6.3)) is then

$$\tilde{p}_{\mathtt{LA}}(\mathbf{y}) = \int_{\mathbb{R}^d} \exp\left(h(\hat{\boldsymbol{\phi}}) - \frac{1}{2}(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})^\top \hat{\mathbf{H}} (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})\right) \mathrm{d}\boldsymbol{\phi} \tag{6.9}$$

$$= p(\hat{\boldsymbol{\phi}}, \mathbf{y}) \cdot \frac{(2\pi)^{d/2}}{|\hat{\mathbf{H}}|^{1/2}}. \tag{6.10}$$

The result above is calculated using the known normalising constant of the Gaussian distribution

$$p_{\mathtt{G}}(\boldsymbol{\phi} \,|\, \mathbf{y}) = \mathcal{N}(\boldsymbol{\phi} \,|\, \hat{\boldsymbol{\phi}}, \hat{\mathbf{H}}^{-1}) = \frac{|\hat{\mathbf{H}}|^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})^\top \hat{\mathbf{H}} (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})\right). \tag{6.11}$$

The Laplace approximation may be thought of as approximating the posterior distribution by a Gaussian distribution $p(\boldsymbol{\phi} \,|\, \mathbf{y}) \approx p_{\mathtt{G}}(\boldsymbol{\phi} \,|\, \mathbf{y})$ such that

$$\tilde{p}_{\mathtt{LA}}(\mathbf{y}) = \left.\frac{p(\boldsymbol{\phi}, \mathbf{y})}{p_{\mathtt{G}}(\boldsymbol{\phi} \,|\, \mathbf{y})}\right|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}. \tag{6.12}$$

Calculation of the Laplace approximation requires obtaining the second derivative of h with respect to $\boldsymbol{\phi}$ (Equation (6.8)). Derivatives may also be used to improve the performance of the optimisation algorithm used to obtain the maxima of h (Equation (6.7)) by providing access to the gradient of h with respect to $\boldsymbol{\phi}$.

The marginal Laplace approximation

Approximating the full joint posterior distribution using a Gaussian distribution may be inaccurate. An alternative is to approximate the marginal posterior distribution of some subset of the parameters, referred to as the marginal Laplace

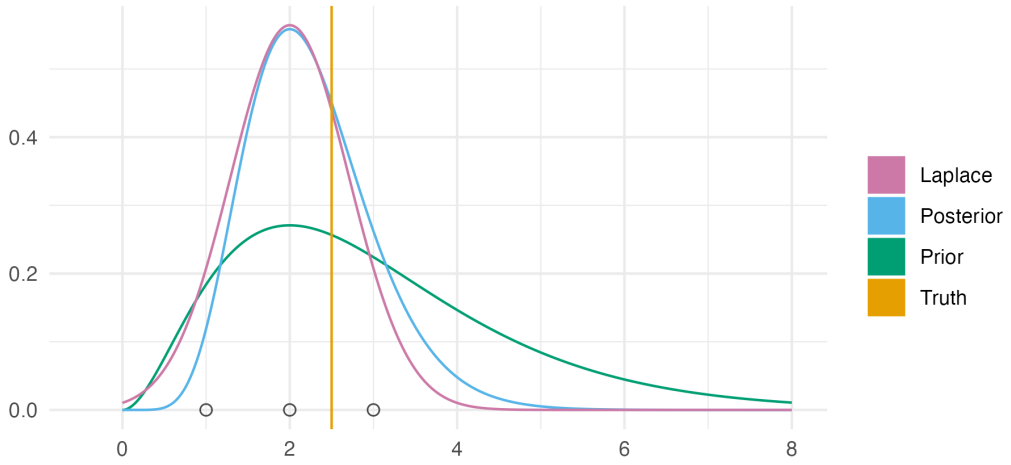

Figure 6.1: Demonstration of the Laplace approximation for the simple Bayesian inference example of Figure 3.1. The unnormalised posterior is $p(\phi, \mathbf{y}) = \phi^8 \exp(-4\phi)$, and can be recognised as the unnormalised gamma distribution Gamma$(9, 4)$. The true log normalising constant is $\log p(\mathbf{y}) = \log \Gamma(9) - 9 \log(4) = -1.872046$, whereas the Laplace approximate log normalising constant is $\log \tilde{p}_{\texttt{LA}}(\mathbf{y}) = -1.882458$, resulting from the Gaussian approximation $p_{\texttt{G}}(\phi \,|\, \mathbf{y}) = \mathcal{N}(\phi \,|\, \mu = 2, \tau = 2)$.

approximation. It remains to integrate out the remaining parameters, using another more suitable method. This approach is the basis of the INLA method.

Let $\boldsymbol{\phi} = (\mathbf{x}, \boldsymbol{\theta})$ and consider a three-stage hierarchical model

$$p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}), \tag{6.13}$$

where $\mathbf{x} = (x_1, \ldots, x_N)$ is the latent field, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ are the hyperparameters. Applying a Gaussian approximation to the latent field, we have $h(\mathbf{x}, \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ with N-dimensional posterior mode

$$\hat{\mathbf{x}}(\boldsymbol{\theta}) = \arg\max_{\mathbf{x}} h(\mathbf{x}, \boldsymbol{\theta}) \tag{6.14}$$

and $(N \times N)$-dimensional Hessian matrix evaluated at the posterior mode

$$\hat{\mathbf{H}}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}^\top} h(\mathbf{x}, \boldsymbol{\theta})|_{\mathbf{x} = \hat{\mathbf{x}}(\boldsymbol{\theta})}. \tag{6.15}$$

Dependence on the hyperparameters $\boldsymbol{\theta}$ is made explicit in both Equation (6.14) and (6.15) such that there is a Gaussian approximation to the marginal posterior of the latent field $\tilde{p}_{\texttt{G}}(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}(\mathbf{x} \,|\, \hat{\mathbf{x}}(\boldsymbol{\theta}), \hat{\mathbf{H}}(\boldsymbol{\theta})^{-1})$ at each value $\boldsymbol{\theta}$ in the space

$\mathbb{R}^m$. The resulting marginal Laplace approximation, for a particular value of the hyperparameters, is then

$$\tilde{p}_{\mathtt{LA}}(\boldsymbol{\theta}, \mathbf{y}) = \int_{\mathbb{R}^N} \exp\left(h(\hat{\mathbf{x}}(\boldsymbol{\theta}), \boldsymbol{\theta}) - \frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}}(\boldsymbol{\theta}))^\top \hat{\mathbf{H}}(\boldsymbol{\theta})(\mathbf{x} - \hat{\mathbf{x}}(\boldsymbol{\theta}))\right) \mathrm{d}\mathbf{x} \tag{6.16}$$

$$= \exp(h(\hat{\mathbf{x}}(\boldsymbol{\theta}), \mathbf{y})) \cdot \frac{(2\pi)^{d/2}}{|\hat{\mathbf{H}}(\boldsymbol{\theta})|^{1/2}} \tag{6.17}$$

$$= \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{\tilde{p}_{\mathtt{G}}(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})}\Big|_{\mathbf{x}=\hat{\mathbf{x}}(\boldsymbol{\theta})}. \tag{6.18}$$

The marginal Laplace approximation is most accurate when the marginal posterior $p(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})$ is accurately approximated by a Gaussian distribution. For the class of latent Gaussian models (Rue, Martino, and Chopin 2009) the prior distribution on the latent field is Gaussian

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x} \mid \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x} \mid \mathbf{0}, \mathbf{Q}(\boldsymbol{\theta})), \tag{6.19}$$

with assumed zero mean $\mathbf{0}$, and precision matrix $\mathbf{Q}(\boldsymbol{\theta})$. The resulting marginal posterior distribution

$$p(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y}) \propto \mathcal{N}(\mathbf{x} \mid \boldsymbol{\theta}) p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) \tag{6.20}$$

$$\propto \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{Q}(\boldsymbol{\theta})\mathbf{x} + \log p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})\right) \tag{6.21}$$

is not exactly Gaussian. However, its deviation can be expected to be small if $\log p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})$ is small.

6.1.2 Quadrature

Quadrature is a method used to approximate integrals using a weighted sum of function evaluations. As with the Laplace approximation, it is deterministic in that the computational procedure is not intrinsically random. Let $\mathcal{Q}$ be a set of quadrature nodes $\mathbf{z} \in \mathcal{Q}$ and $\omega : \mathbb{R}^d \to \mathbb{R}$ be a weighting function. Then, quadrature can be used to estimate the posterior normalising constant (Equation (6.3)) by

$$\tilde{p}_{\mathcal{Q}}(\mathbf{y}) = \sum_{\mathbf{z} \in \mathcal{Q}} p(\mathbf{y}, \mathbf{z}) \omega(\mathbf{z}). \tag{6.22}$$

To illustrate quadrature for a simple example, consider integrating the univariate function $f(z) = z\sin(z)$ between $z = 0$ and $z = \pi$. This integral can be calculated analytically using integration by parts and evaluates to π. A quadrature approximation of this integral is

$$\pi = \sin(z) - z\cos(z)\Big|_0^\pi = \int_0^\pi z\sin(z)\mathrm{d}z \approx \sum_{z\in\mathcal{Q}} z\sin(z)\omega(z), \tag{6.23}$$

where $\mathcal{Q} = \{z_1, \dots z_k\}$ are a set of k quadrature nodes and $\omega : \mathbb{R} \to \mathbb{R}$ is a weighting function.

The trapezoid rule is an example of a quadrature rule, in which quadrature nodes are spaced throughout the domain with $\epsilon_i = z_i - z_{i-1} > 0$ for $1 < i < k$. The weighting function is

$$\omega(z_i) = \begin{cases} \epsilon_i & 1 < i < k, \\ \epsilon_i/2 & i \in \{1, k\}. \end{cases} \tag{6.24}$$

Figure 6.2 shows application of the trapezoid rule to integration of $z\sin(z)$ as described in Equation (6.23). The more quadrature nodes are used, the more accurate the estimate of the integrand is. Under some regularity conditions on f, as the spacing between quadrature nodes $\epsilon \to 0$ the estimate obtained using the trapezoid rule converges to the true value of the integral. Indeed, this approach was used by Riemann to provide the first rigorous definition of the integral.

Quadrature methods are most effective when integrating over small dimensions, say three or less. This is because the number of quadrature nodes at which the function is required to be evaluated in the computation grows exponentially with the dimension. For even moderate dimension, this quickly makes computation intractable. For example, using 5, 10, or 20 quadrature nodes per dimension, as in Figure 6.2, in five-dimensions (rather than one, as shown) would require 3125, 100000 or 3200000 quadrature nodes respectively. Though quadrature is embarrassingly parallel, in that function evaluation at each node is entirely independent, solutions requiring the evaluation of millions quadrature nodes are unlikely to be tractable.

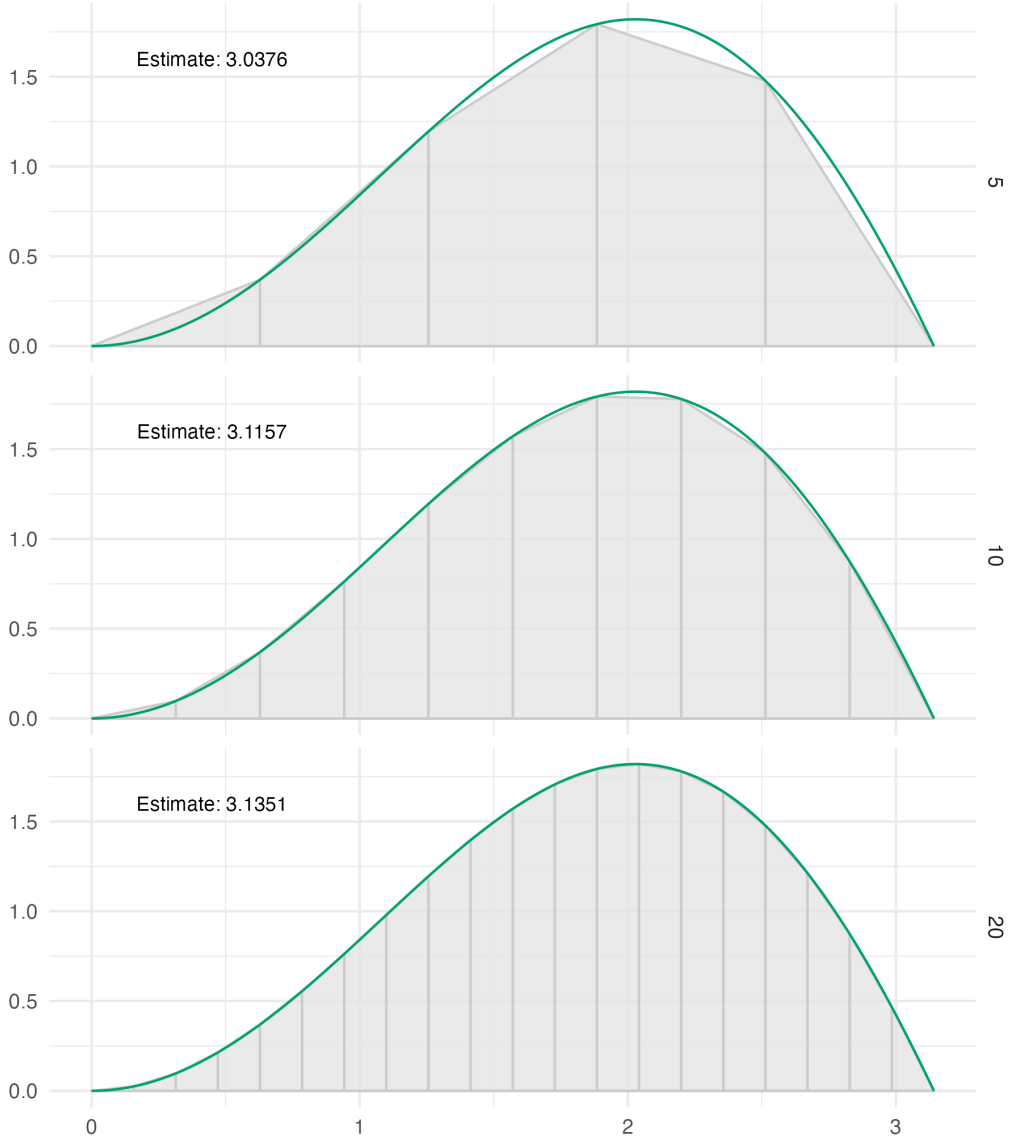

Figure 6.2: The trapezoid rule with $k = 5, 10, 20$ equally-spaced ($\epsilon_i = \epsilon > 0$) quadrature nodes can be used to integrate the function $f(z) = z\sin(z)$, shown in green, in the domain $[0, \pi]$. Here, the exact solution is $\pi \approx 3.1416$. As k increases and more nodes are used in the computation, the quadrature estimate becomes closer to the exact solution. The trapezoid rule estimate is given by the sum of the areas of the grey trapezoids.

Gauss-Hermite quadrature

It is possible to construct quadrature rules which use relatively few nodes and are highly accurate when the integrand adheres to certain assumptions [Chapter 4; Press et al. (2007)]. Gauss-Hermite quadrature [GHQ; Davis and Rabinowitz (1975)] is a quadrature rule designed to integrate functions of the form $f(\mathbf{z}) = \varphi(\mathbf{z})P_\alpha(\mathbf{z})$ exactly, that is with no error, such that

$$\int \varphi(\mathbf{z})P_\alpha(\mathbf{z})\mathrm{d}\mathbf{z} = \sum_{\mathbf{z}\in\mathcal{Q}} \varphi(\mathbf{z})P_\alpha(\mathbf{z})\omega(\mathbf{z}). \tag{6.25}$$

In this equation, the term $\varphi(\cdot)$ is a standard multivariate normal density $\mathcal{N}(\cdot \mid \mathbf{0}, \mathbf{I})$, where $\mathbf{0}$ and $\mathbf{I}$ are the zero-vector and identify matrix of relevant dimension, and the term $P_\alpha(\cdot)$ is a polynomial with highest degree monomial $\alpha \leq 2k-1$, where k is the number of quadrature nodes per dimension. GHQ is attractive for Bayesian inference problems because posterior distributions are typically well approximated by functions of this form. Support for this statement is provided by the Bernstein–von Mises theorem, which states that, under some regularity conditions, as the number of data points increases the posterior distribution convergences to a Gaussian.

I follow the notation for GHQ established by Bilodeau et al. (2022). First, to construct the univariate GHQ rule for $z \in \mathbb{R}$, let $H_k(z)$ be the kth (probabilist's) Hermite polynomial

$$H_k(z) = (-1)^k \exp(z^2/2)\frac{\mathrm{d}}{\mathrm{d}z^k}\exp(-z^2/2) \tag{6.26}$$

The Hermite polynomials are defined to be orthogonal with respect to the standard Gaussian probability density function

$$\int H_k(z)H_l(z)\varphi(z)\mathrm{d}z = \delta_{kl}, \tag{6.27}$$

where $\delta_{kl} = 1$ if $k = l$ and $\delta_{kl} = 0$ otherwise. The GHQ nodes $z \in \mathcal{Q}(1,k)$ are given by the k zeroes of the kth Hermite polynomial. For $k = 1, 2, 3$ these zeros, up to three decimal places, are

$$H_1(z) = z = 0 \implies \mathcal{Q}(1,1) = \{0\}, \tag{6.28}$$

$$H_2(z) = z^2 - 1 = 0 \implies \mathcal{Q}(1,2) = \{-0.707, 0.707\}, \tag{6.29}$$

$$H_3(z) = z^3 - 3z = 0 \implies \mathcal{Q}(1,3) = \{-1.225, 0, 1.225\}. \tag{6.30}$$

The quadrature nodes are symmetric about zero, and include zero when k is odd. The corresponding weighting function $\omega : \mathcal{Q}(1,k) \to \mathbb{R}$ chosen to satisfy Equation (6.25) is given by

$$\omega(z) = \frac{k!}{\varphi(z)[H_{k+1}(z)]^2}. \tag{6.31}$$

Multivariate GHQ rules are usually constructed using the product rule with identical univariate GHQ rules in each dimension. As such, in d dimensions, the multivariate GHQ nodes $\mathbf{z} \in \mathcal{Q}(d,k)$ are defined by

$$\mathcal{Q}(d,k) = \mathcal{Q}(1,k)^d = \mathcal{Q}(1,k) \times \cdots \times \mathcal{Q}(1,k). \tag{6.32}$$

The corresponding weighting function $\omega : \mathcal{Q}(d,k) \to \mathbb{R}$ is given by a product of the univariate weighting functions $\omega(\mathbf{z}) = \prod_{j=1}^{d} \omega(z_j)$.

Adaptive quadrature

In adaptive quadrature, the quadrature nodes and weights selected depend on the specific integrand being considered. For example, adaptive use of the trapezoid rule requires specifying a rule for the start point, end point, and spacing between quadrature nodes. It is particularly important to use an adaptive quadrature rule for Bayesian inference problems because the posterior normalising constant $p(\mathbf{y})$ is a function of the data. No fixed quadrature rule can be expected to effectively integrate all possible posterior distributions.

In adaptive GHQ [AGHQ; Naylor and Smith (1982)] the quadrature nodes are shifted by the mode of the integrand, and rotated based on a matrix decomposition of the inverse curvature at the mode. To demonstrate AGHQ, consider its application to calculation of the posterior normalising constant. The relevant transformation of the GHQ nodes $\mathcal{Q}(d,k)$ is

$$\boldsymbol{\phi}(\mathbf{z}) = \hat{\mathbf{P}}\mathbf{z} + \hat{\boldsymbol{\phi}}, \tag{6.33}$$

where $\hat{\mathbf{P}}$ is a matrix decomposition of $\hat{\mathbf{H}}^{-1} = \hat{\mathbf{P}}\hat{\mathbf{P}}^{\top}$. To account for the transformation, the weighting function may be redefined to include a matrix determinant,

analogous to the Jacobian determinant, or more simply the matrix determinant may be written outside the integral. Taking the later approach, the resulting adaptive quadrature estimate of the posterior normalising constant is

$$\tilde{p}_{\texttt{AQ}}(\mathbf{y}) = |\hat{\mathbf{P}}| \sum_{\mathbf{z}\in\mathcal{Q}(d,k)} p(\mathbf{y}, \boldsymbol{\phi}(\mathbf{z}))\omega(\mathbf{z}) \tag{6.34}$$

$$= |\hat{\mathbf{P}}| \sum_{\mathbf{z}\in\mathcal{Q}(d,k)} p(\mathbf{y}, \hat{\mathbf{P}}\mathbf{z} + \hat{\boldsymbol{\phi}})\omega(\mathbf{z}). \tag{6.35}$$

The quantities $\hat{\boldsymbol{\phi}}$ and $\hat{\mathbf{H}}$ are exactly those given in Equations (6.7) and (6.8) and used in the Laplace approximation. Indeed, when $k = 1$ then AGHQ corresponds to the Laplace approximation. To see this, we have $H_1(z)$ with univariate zero $z = 0$ such that the adapted node is given by the mode $\boldsymbol{\phi}(\mathbf{z} = \mathbf{0} = 0 \times \cdots \times 0) = \hat{\boldsymbol{\phi}}$. The weighting function is given by

$$\omega(0)^d = \left(\frac{1!}{\varphi(0)H_2(0)^2}\right)^d = \left(\frac{1}{\varphi(0)}\right)^d = (2\pi)^{d/2}. \tag{6.36}$$

The AGHQ estimate of the normalising constant for $k = 1$ is then given by

$$\tilde{p}_{\texttt{AQ}}(\mathbf{y}) = p(\mathbf{y}, \hat{\boldsymbol{\phi}}) \cdot |\hat{\mathbf{P}}| \cdot (2\pi)^{d/2} = p(\mathbf{y}, \hat{\boldsymbol{\phi}}) \cdot \frac{(2\pi)^{d/2}}{|\hat{\mathbf{H}}|^{1/2}}, \tag{6.37}$$

which corresponds to the Laplace approximation $\tilde{p}_{\texttt{LA}}(\mathbf{y})$ given in Equation (6.10). This connection supports AGHQ being a natural extension of the Laplace approximation when greater accuracy than $k = 1$ is required.

Two alternatives for the matrix decomposition $\hat{\mathbf{H}}^{-1} = \hat{\mathbf{P}}\hat{\mathbf{P}}^\top$ are the Cholesky and spectral decomposition (Jäckel 2005). For the Cholesky decomposition $\hat{\mathbf{P}} = \hat{\mathbf{L}}$, where

$$\hat{\mathbf{L}} = \begin{pmatrix} L_{11} & 0 & \cdots & 0 \\ \hat{L}_{12} & \hat{L}_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \hat{L}_{1d} & \dots & \hat{L}_{(d-1)d} & \hat{L}_{dd} \end{pmatrix} \tag{6.38}$$

is a lower triangular matrix. For the spectral decomposition $\hat{\mathbf{P}} = \hat{\mathbf{E}}\hat{\boldsymbol{\Lambda}}^{1/2}$, where $\hat{\mathbf{E}} = (\hat{\mathbf{e}}_1, \dots \hat{\mathbf{e}}_d)$ contains the eigenvectors of $\hat{\mathbf{H}}^{-1}$ and $\hat{\boldsymbol{\Lambda}}$ is a diagonal matrix containing its eigenvalues $(\hat{\lambda}_1, \dots, \hat{\lambda}_d)$. Figure 6.3 demonstrates GHQ and AGHQ for a two-dimensional example, using both decomposition approaches. Using the

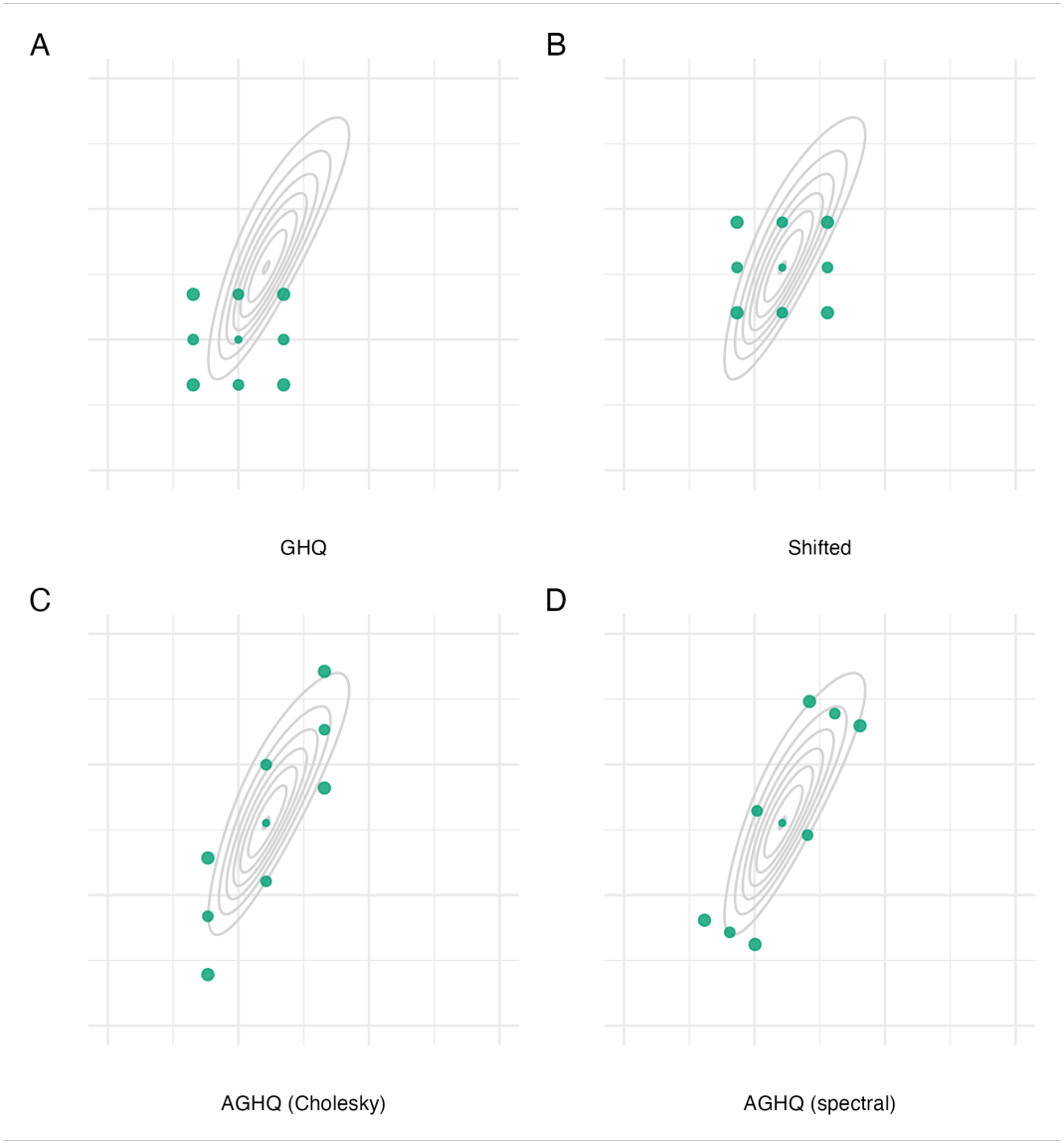

Figure 6.3: The Gauss-Hermite quadrature nodes $\mathbf{z} \in \mathcal{Q}(2, 3)$ for a two-dimensional integral with three nodes per dimension (Panel A). Adaption occurs based on the mode (Panel B) and covariance of the integrand via either the Cholesky (Panel C) or spectral (Panel D) decomposition of the inverse curvature at the mode. Here, the integrand is $f(z_1, z_2) = \text{sn}(0.5z_1, \alpha = 2) \cdot \text{sn}(0.8z_1 - 0.5z_2, \alpha = -2)$, where $\text{sn}(\cdot)$ is the standard skewnormal probability density function with shape parameter $\alpha \in \mathbb{R}$. The integral approximation $I \approx \int\int f(z_1, z_2)\text{d}z_1\text{d}z_2$ obtained by the quadrature rule in each panel are given.

Cholesky decomposition results in adapted quadrature nodes which collapse along one of the dimensions, as a result of the matrix $\hat{\mathbf{L}}$ being lower triangular. On the other hand, using the spectral decomposition results in adapted quadrature nodes which lie along the orthogonal eigenvectors of $\hat{\mathbf{H}}^{-1}$.

Using AGHQ, Bilodeau et al. (2022) provide the first stochastic convergence rate for adaptive quadrature applied to Bayesian inference.

6.1.3 Integrated nested Laplace approximation

The integrated nested Laplace approximation (INLA) method (Rue, Martino, and Chopin 2009) combines marginal Laplace approximations with quadrature to enable approximation of posterior marginal distributions.

Consider the marginal Laplace approximation (Section 6.1.1) for a three-stage hierarchical model given by

$$\tilde{p}_{\texttt{LA}}(\boldsymbol{\theta}, \mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{\tilde{p}_{\texttt{G}}(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})}\bigg|_{\mathbf{x}=\hat{\mathbf{x}}(\boldsymbol{\theta})}. \tag{6.39}$$

To complete approximation of the posterior normalising constant, the marginal Laplace approximation can be integrated over the hyperparameters using a quadrature rule (Section 6.1.2)

$$\tilde{p}(\mathbf{y}) = \sum_{\mathbf{z} \in \mathcal{Q}} \tilde{p}_{\texttt{LA}}(\mathbf{z}, \mathbf{y}) \omega(\mathbf{z}). \tag{6.40}$$

Though any choice of quadrature rule is possible, following Stringer et al. (2022) here I consider use of AGHQ. Let $\mathbf{z} \in \mathcal{Q}(m, k)$ be the m-dimensional GHQ nodes constructed using the product rule with k nodes per dimension, and $\omega : \mathbb{R}^m \to \mathbb{R}$ the corresponding weighting function. These nodes are adapted by $\boldsymbol{\theta}(\mathbf{z}) = \hat{\mathbf{P}}_{\texttt{LA}} \mathbf{z} + \hat{\boldsymbol{\theta}}_{\texttt{LA}}$ where

$$\hat{\boldsymbol{\theta}}_{\texttt{LA}} = \arg\max_{\boldsymbol{\theta}} \log \tilde{p}_{\texttt{LA}}(\boldsymbol{\theta}, \mathbf{y}), \tag{6.41}$$

$$\hat{\mathbf{H}}_{\texttt{LA}} = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \log \tilde{p}_{\texttt{LA}}(\boldsymbol{\theta}, \mathbf{y})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\texttt{LA}}}, \tag{6.42}$$

$$\hat{\mathbf{H}}_{\texttt{LA}}^{-1} = \hat{\mathbf{P}}_{\texttt{LA}} \hat{\mathbf{P}}_{\texttt{LA}}^\top. \tag{6.43}$$

The nested AGHQ estimate of the posterior normalising constant is then

$$\tilde{p}_{\mathtt{AQ}}(\mathbf{y}) = |\hat{\mathbf{P}}_{\mathtt{LA}}| \sum_{\mathbf{z} \in \mathcal{Q}(m,k)} \tilde{p}_{\mathtt{LA}}(\boldsymbol{\theta}(\mathbf{z}), \mathbf{y}) \omega(\mathbf{z}). \tag{6.44}$$

This estimate can be used to normalise the marginal Laplace approximation as follows

$$\tilde{p}_{\mathtt{LA}}(\boldsymbol{\theta} \,|\, \mathbf{y}) = \frac{\tilde{p}_{\mathtt{LA}}(\boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_{\mathtt{AQ}}(\mathbf{y})}. \tag{6.45}$$

The posterior marginals $\tilde{p}(\theta_j \,|\, \mathbf{y})$ may be obtained by

$$\tilde{p}(\theta_j \,|\, \mathbf{y}) = \int \tilde{p}(\theta_j \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta}_{-j}. \tag{6.46}$$

These integrals may be computed by reusing the AGHQ rule. More recent methods are discussed in Section 3.2 of Martins et al. (2013).

Multiple methods have been proposed for obtaining the $\tilde{p}(\mathbf{x} \,|\, \mathbf{y})$ or individual marginals $\tilde{p}(x_i \,|\, \mathbf{y})$ Four methods are presented below, trading-off accuracy with computational expense.

Gaussian marginals

Most easily, inferences for the latent field can be obtained by approximation of $p(\mathbf{x} \,|\, \mathbf{y})$ using another application of the quadrature rule (Rue and Martino 2007)

$$p(\mathbf{x} \,|\, \mathbf{y}) = \int p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta} = \int p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta} \tag{6.47}$$

$$\approx |\hat{\mathbf{P}}_{\mathtt{LA}}| \sum_{\mathbf{z} \in \mathcal{Q}(m,k)} \tilde{p}_{\mathtt{G}}(\mathbf{x} \,|\, \boldsymbol{\theta}(\mathbf{z}), \mathbf{y}) \tilde{p}_{\mathtt{LA}}(\boldsymbol{\theta}(\mathbf{z}) \,|\, \mathbf{y}) \omega(\mathbf{z}). \tag{6.48}$$

The quadrature rule $\mathbf{z} \in \mathcal{Q}(m,k)$ is used both internally to normalise the marginal Laplace approximation, and externally to perform integration with respect to the hyperparameters. Equation (6.48) is a mixture of Gaussian distributions

$$p_{\mathtt{G}}(\mathbf{x} \,|\, \boldsymbol{\theta}(\mathbf{z}), \mathbf{y}), \tag{6.49}$$

each with multinomial probabilities

$$\lambda(\mathbf{z}) = |\hat{\mathbf{P}}_{\mathtt{LA}}| \tilde{p}_{\mathtt{LA}}(\boldsymbol{\theta}(\mathbf{z}) \,|\, \mathbf{y}) \omega(\mathbf{z}), \tag{6.50}$$

where $\sum \lambda(\mathbf{z}) = 1$ and $\lambda(\mathbf{z}) > 0$. Samples may therefore be naturally obtained for the complete vector $\mathbf{x}$ jointly by first drawing a node $\mathbf{z} \in \mathcal{Q}(m, k)$ with multinomial probabilities $\lambda(\mathbf{z})$ then drawing a sample from the corresponding Gaussian distribution in Equation (6.49). Algorithms for fast and exact simulation from a Gaussian distribution have been developed, including by Rue (2001). The posterior marginals for any subset of the complete vector can simply be obtained by keeping the relevant entries of $\mathbf{x}$.

Laplace marginals

An alternative higher accuracy, but more computationally expensive, approach is to calculate a Laplace approximation to the marginal posterior

$$\tilde{p}_{\mathtt{LA}}(x_i, \boldsymbol{\theta}, \mathbf{y}) = \frac{p(x_i, \mathbf{x}_{-i}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_{\mathtt{G}}(\mathbf{x}_{-i} \,|\, x_i, \boldsymbol{\theta}, \mathbf{y})}\bigg|_{\mathbf{x}_{-i}=\hat{\mathbf{x}}_{-i}(x_i,\boldsymbol{\theta})}. \tag{6.51}$$

Here, the variable x_i is excluded from the Gaussian approximation such that

$$p_{\mathtt{G}}(\mathbf{x}_{-i} \,|\, x_i, \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}(\mathbf{x}_{-i} \,|\, \hat{\mathbf{x}}_{-i}(x_i, \boldsymbol{\theta}), \hat{\mathbf{H}}_{-i,-i}(x_i, \boldsymbol{\theta})), \tag{6.52}$$

with $(N-1)$-dimensional posterior mode

$$\hat{\mathbf{x}}_{-i}(x_i, \boldsymbol{\theta}) = \arg\max_{\mathbf{x}_{-i}} \log p(\mathbf{y}, x_i, \mathbf{x}_{-i}, \boldsymbol{\theta}), \tag{6.53}$$

and $[(N-1)\times(N-1)]$-dimensional Hessian matrix evaluated at the posterior mode

$$\hat{\mathbf{H}}_{-i,-i}(x_i, \boldsymbol{\theta}) = -\frac{\partial^2}{\partial \mathbf{x}_{-i} \partial \mathbf{x}_{-i}^\top} \log p(\mathbf{y}, x_i, \mathbf{x}_{-i}, \boldsymbol{\theta})|_{\mathbf{x}_{-i}=\hat{\mathbf{x}}_{-i}(x_i,\boldsymbol{\theta})}. \tag{6.54}$$

The approximate posterior marginal $\tilde{p}(x_i \,|\, \mathbf{y})$ may be obtained by normalising the marginal Laplace approximation (Equation (6.51)) before performing integration with respect to the hyperparameters (as in Equation (6.48)). The normalised Laplace approximation is

$$\tilde{p}_{\mathtt{LA}}(x_i, \boldsymbol{\theta} \,|\, \mathbf{y}) = \frac{\tilde{p}_{\mathtt{LA}}(x_i, \boldsymbol{\theta}, \mathbf{y})}{\tilde{p}(\mathbf{y})}. \tag{6.55}$$

where either the estimate of the evidence in Equation (6.44) may be reused or a de novo estimate can be computed. Integration with respect to the hyperparameters is performed via

$$p(x_i \mid \mathbf{y}) = \int p(x_i, \boldsymbol{\theta} \mid \mathbf{y}) \mathrm{d}\boldsymbol{\theta} \tag{6.56}$$

$$\approx |\hat{\mathbf{P}}_{\mathtt{LA}}| \sum_{\mathbf{z} \in \mathcal{Q}(m,k)} \tilde{p}_{\mathtt{LA}}(x_i, \boldsymbol{\theta}(\mathbf{z}) \mid \mathbf{y}) \tilde{\omega}(\mathbf{z}). \tag{6.57}$$

Equation (6.57) is a mixture of the normalised Laplace approximations $\tilde{p}_{\mathtt{LA}}(x_i, \boldsymbol{\theta} \mid \mathbf{y})$ over the hyperparameter quadrature nodes. However, unlike the Gaussian case (Section 6.1.3) it is not easy to directly sample each Laplace approximation. As such, Equation (6.57) may instead be represented by its evaluation at a number of nodes. The nodes may be based on a one-dimensional AGHQ rule, using the mode and standard deviation of the Gaussian approximation to avoid computation of the Laplace marginal mode and standard deviation. The probability density function of the marginal posterior may be recovered using a Lagrange polynomial or spline interpolant to the log probabilities.

Simplified Laplace marginals

When the latent field $\mathbf{x}$ is a Gauss-Markov random fields [GMRF; Rue and Held (2005)] it is possible to efficiently approximate the Laplace marginals in Section 6.1.3. The simplified approximation is achieved by a Taylor expansion on the numerator and denominator of Equation (6.51) up to third order. The approach is analogous to correcting the Gaussian approximation in Section 6.1.3 for location and skewness. Details are left to Section 3.2.3 of Rue, Martino, and Chopin (2009).

Simplified INLA

Wood (2020) describe a method for approximating the Laplace marginals without depending on the Markov structure, while still achieving equivalent efficiency. This work was motivated by the spline setting, which similar to the extended latent Gaussian models [ELGMs; Stringer et al. (2022)] setting, results in precision matrices which are not as sparse. Details are left to Wood (2020).

Inclusion of the structured additive predictor in the latent field

The discussion of INLA is concluded by briefly outlining the cause for and against inclusion of the structured additive predictor from the latent field. The issue relates to the sparsity structure of the Hessian matrix

$$\hat{\mathbf{H}}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}^\top} h(\mathbf{x}, \boldsymbol{\theta})|_{\mathbf{x}=\hat{\mathbf{x}}(\boldsymbol{\theta})}, \tag{6.58}$$

where, as before, $\hat{\mathbf{x}}(\boldsymbol{\theta}) = \arg\max_{\mathbf{x}} h(\mathbf{x}, \boldsymbol{\theta})$. The modified model used by Rue, Martino, and Chopin (2009) is

$$\boldsymbol{\eta}^\star = \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{6.59}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \tau^{-1}\mathbf{I}_n), \tag{6.60}$$

$$\mathbf{x}^\star = (\boldsymbol{\eta}^\star, \mathbf{x}) \sim \mathcal{N}(\cdots). \tag{6.61}$$

Stringer et al. (2022) note that this doesn't work as well for ELGMs. Discussed in Van Niekerk et al. (2023).

6.1.4 Software

`R-INLA`

The `R-INLA` software (Martins et al. 2013) implements the INLA method, as well as the stochastic partial differential equation (SPDE) approach of Lindgren et al. (2011). `R-INLA` is the R interface to the core `inla` program, which is written in C (Martino and Rue 2009). Algorithms for sampling from GMRFs are used from the `GMRFLib` C library (Rue and Follestad 2001). First and second derivatives are either hard coded, or computed numerically using central finite differences (Fattah et al. 2021). For a review recent computational features of `R-INLA`, including parallelism via `OpenMP` (Diaz et al. 2018) and use of the PARDISO sparse linear equation solver (Bollhöfer et al. 2020), see Gaedke-Merzhäuser et al. (2023). Further information about `R-INLA`, including recent developments, can be found at `https://r-inla.org`.

The connection between the latent field $\mathbf{x}$ and structured additive predictor $\boldsymbol{\eta}$ is specified in `R-INLA` using a formula interface of the form `y ~ ...`. The interface is

similar to that used in the `lm` function in the core `stats` R package. For example, a model with one fixed effect `a` and one IID random effect `b`, has the formula `y ~ a + f(b, model = "iid")`. This interface is easy to engage with for new users, but can be limiting for more advanced users.

The approach used to compute the marginals $\tilde{p}(x_i \mid \mathbf{y})$ can chosen by setting `method` to `"gaussian"` (Section 6.1.3), `"laplace"` (Section 6.1.3) or `simplified.laplace` (Section 6.1.3). The quadrature grid used can be chosen by setting `int.strategy` to `"eb"` (empirical Bayes, one quadrature node), `"grid"` (a dense grid), or `"ccd"` [Box-Wilson central composite design; Box and Wilson (1992)]. Figure 6.4 demonstrates the latter two integration strategies. By default, the `"grid"` strategy is used for $m \leq 2$ and the `"ccd"` strategy is used for $m > 2$.

Various software packages have been built using `R-INLA`. Perhaps the most substantial of which is the `inlabru` R package (Bachl et al. 2019). As well as a simplified syntax, `inlabru` provides capabilities for fitting more general non-linear structured additive predictor expressions via linearisation and repeat use of `R-INLA`. These complex model components are specified in `inlabru` using the `bru_mapper` system.

`TMB`

Template Model Builder [`TMB`; Kristensen et al. (2016)] is an R package which implements the Laplace approximation. In `TMB` derivatives are obtained using automatic differentiation, also known as algorithmic differentiation [AD; Baydin et al. (2017)]. The approach of AD is to decompose any function into a sequence of elementary operations with known derivatives. The known derivatives of the elementary operations may then be composed by repeat use of the chain rule to obtain the function's derivative. `TMB` uses the C++ package `CppAD` (Bell 2023) for AD [Section 3; Kristensen et al. (2016)]. The development of `TMB` was strongly inspired by the Automatic Differentiation Model Builder [ADMB; Fournier et al. (2012); Bolker et al. (2013)] project. An algorithm is used in `TMB` to automatically determine matrix sparsity structure [Section 4.2; Kristensen et al. (2016)]. The R

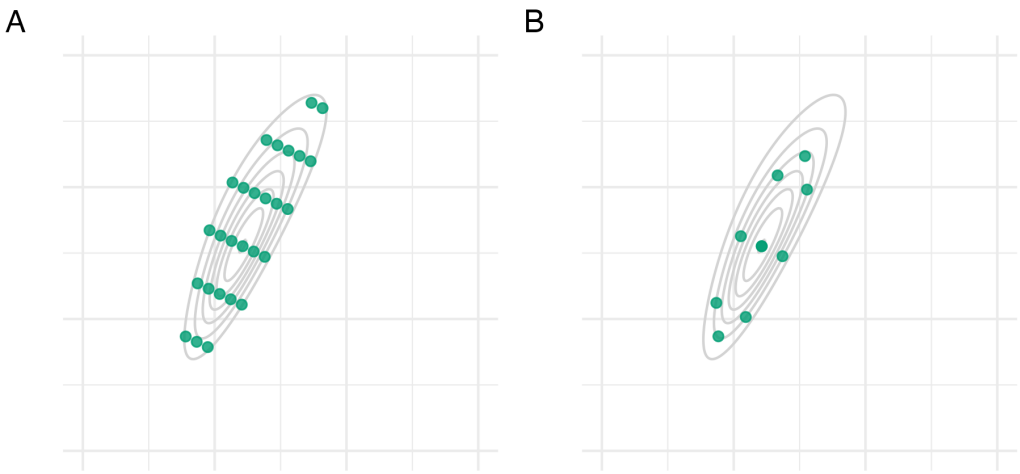

Figure 6.4: Consider the function $f(z_1, z_2) = \text{sn}(0.5z_1, \alpha = 2) \cdot \text{sn}(0.8z_1 - 0.5z_2, \alpha = -2)$ as described in Figure 6.3. Panel A shows the grid method as used in `R-INLA` and detailed in Section 3.1 of Rue, Martino, and Chopin (2009). Briefly, equally-weighted quadrature points are generated by starting at the mode and taking steps of size δ_z along each eigenvector of the inverse curvature at the mode, scaled by the eigenvalues, until the difference in log-scale function evaluations (compared to the mode) is below a threshold δ_π. Intermediate values are included if they have sufficient log-scale function evaluation. Here, I set $\delta_z = 0.75$ and $\delta_\pi = 2$. Panel B shows a CCD as used in `R-INLA` and detailed in Section 6.5 of Rue, Martino, and Chopin (2009). The CCD was generated using the `rsm` R package (Lenth 2009), and is comprised of: one centre point; four factorial points, used to help estimate linear effects; and four star points, used to help estimate the curvature.

package `Matrix` and C++ package `Eigen` are then used for sparse and dense matrix calculations. Kristensen et al. (2016) highlight the modular design philosophy of `TMB`.

Models are specified in `TMB` using a C++ template file which evaluates $\log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ in a Bayesian context or $\log p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})$ in a frequentist setting. Other software packages have been developed which also use `TMB` C++ templates. The `tmbstan` R package (Monnahan and Kristensen 2018) allows running the Hamiltonian Monte Carlo (HMC) algorithm via `Stan`. The `aghq` R package (Stringer 2021b) allows use of AGHQ, and AGHQ over the marginal Laplace approximation, via the `mvQuad` R package (Weiser 2016). The `glmmTMB` R package (Brooks et al. 2017) allows specification of common GLMM models via a formula interface. It is also possible to extract the `TMB` objective function used by `glmmTMB`, which may then be passed into `aghq` or `tmbstan`.

A review of the use of `TMB` for spatial modelling, including comparison to `R-INLA`, is provided by Osgood-Zimmerman and Wakefield (2023).

Other software

The `mgcv` [Mixed GAM computation vehicle; Wood (2017)] R package estimates generalised additive models (GAMs) specified using a formula interface. This package is briefly mentioned so as to note that the function `mgcv::ginla` implements the simplified INLA approach of Wood (2020) as described in Section 6.1.3.

6.2 A universal INLA implementation

This section is about implementation of the INLA method using AD via the `TMB` package. Both the Gaussian and Laplace latent field marginal approximations are implemented. The implementation is universal in that it is compatible with any model with a `TMB` C++ template, rather than being based on a restrictive formula interface. The `TMB` probabilistic programming language is described as "universal" in that it is an extension of the Turing-complete general purpose language C++.

Martino and Riebler (2019) note that "implementing INLA from scratch is a complex task" and as a result "applications of INLA are limited to the (large class of) models implemented [in `R-INLA`]". A universal INLA implementation facilitates application of the method to models which are not compatible with `R-INLA`. The Naomi model is one among many examples. Section 5 of Osgood-Zimmerman and Wakefield (2023) notes that "`R-INLA` is capable of using higher-quality approximations than `TMB`" (hyperparameter integration and latent field Laplace marginals) and "in return `TMB` is applicable to a wider class of models". Yet there is no inherent reason for these capabilities to be in conflict: it is possible to have both high-quality approximations and flexibility. The potential benefits of a more flexible INLA implementation based on AD were noted by Skaug (2009) (a coauthor of `TMB`) in discussion of Rue, Martino, and Chopin (2009), who noted that such a system would be "fast, flexible, and easy-to-use", as well as "automatic

from a user's perspective". As this suggestion was made close to 15 years ago, it is surprising that its potential remains unrealised.

I demonstrate the universal implementation with two examples:

1. Section 6.2.1 considers a generalised linear mixed model (GLMM) of an epilepsy drug. This example was used in Section 5.2 of Rue, Martino, and Chopin (2009), and the model is compatible with `R-INLA`. For some parameters there is a notable difference in approximation error depending on use of Gaussian or Laplace marginals. This example is therefore used to demonstrate the correspondence between my Laplace marginal implementation, and that of `R-INLA` with `method` set to `"laplace"`.
2. Section 6.2.2 considers a model which is not is not compatible with `R-INLA`. This example is used as a demonstration of the benefit of a more generally applicable INLA implementation.

6.2.1 Epilepsy GLMM

Consider a GLMM for an epilepsy drug double-blind clinical trial (Leppik et al. 1985). The original GLMM of Thall and Vail (1990) was modified by Breslow and Clayton (1993) and widely disseminated as a part of the BUGS [Bayesian inference using Gibbs sampling; Spiegelhalter, Thomas, et al. (1996)] manual.

Patients $i = 1, \ldots, 59$ were each assigned either a new drug $\texttt{Trt}_i = 1$ or a placebo $\texttt{Trt}_i = 0$. Each patient made four visits the clinic $j = 1, \ldots, 4$, and the observations y_{ij} are the number of seizures of the ith person in the two weeks preceding their jth clinic visit (Figure 6.5). The covariates used in the model were baseline seizure counts $\texttt{Base}_i$, treatment $\texttt{Trt}_i$, age $\texttt{Age}_i$, and an indicator for the final clinic visit $\texttt{V}_{4j}$. Each of the covariates were centred. The observations were modelled using a Poisson distribution

$$y_{ij} \sim \text{Poisson}(e^{\eta_{ij}}), \tag{6.62}$$

Figure 6.5: The number of seizures in the treatment group was fewer, on average, than the number of seizures in the control group. This is not sufficient to conclude that the treatment was effective. The GLMM accounts for differences between the treatment and control group, including in baseline seizures and age, and so can be used to help estimate a causal treatment effect.

with structured additive predictor

$$\eta_{ij} = \beta_0 + \beta_{\texttt{Base}} \log(\texttt{Base}_i/4) + \beta_{\texttt{Trt}} \texttt{Trt}_i + \beta_{\texttt{Trt}\times\texttt{Base}}(\texttt{Trt}_i \times \log(\texttt{Base}_i/4)) \quad (6.63)$$

$$+ \beta_{\texttt{Age}} \log(\texttt{Age}_i) + \beta_{\texttt{V}_4} \texttt{V}_{4j} + \epsilon_i + \nu_{ij}, \quad i \in [59], \quad j \in [4]. \quad (6.64)$$

The prior distribution on each of the regression parameters, including the intercept β_0, was $\mathcal{N}(0, 100^2)$. The patient $\epsilon_i \sim \mathcal{N}(0, 1/\tau_\epsilon)$ and patient-visit $\nu_{ij} \sim \mathcal{N}(0, 1/\tau_\nu)$ random effects were IID with gamma precision prior distributions $\tau_\epsilon, \tau_\nu \sim \Gamma(0.001, 0.001)$.

Table 6.1: The inference methods and software considered.

	Method	Software
Section 6.2.1	Gaussian, EB	`R-INLA`
Section 6.2.1	Gaussian, grid	`R-INLA`
Section 6.2.1	Laplace, EB	`R-INLA`
Section 6.2.1	Laplace, grid	`R-INLA`
Section 6.2.1	Gaussian, EB	`TMB`
Section 6.2.1	Gaussian, AGHQ	`TMB` and `aghq`
Section 6.2.1	Laplace, EB	`TMB`

	Method	Software
Section 6.2.1	Laplace, AGHQ	`TMB` and `aghq`
Section 6.2.1	NUTS	`rstan`

Inference for the epilepsy GLMM was conducted using a range of approaches (Table 6.1). Section 6.2.1 compares the results. The foremost objective of this exercise is to demonstrate correspondence between inferences obtained from `R-INLA` and those from `TMB`. Furthermore, illustrative code is used throughout this section to enhance understanding of the methods and software used. As such, this section is more verbose than future sections.

INLA with `R-INLA`

The epilepsy data are available from the `R-INLA` package. The covariates may be obtained and their transformations centred by:

```
centre <- function(x) (x - mean(x))


Epil <- Epil %>%
  mutate(CTrt    = centre(Trt),
         ClBase4 = centre(log(Base/4)),
         CV4     = centre(V4),
         ClAge   = centre(log(Age)),
         CBT     = centre(Trt * log(Base/4)))
```

The structured additive predictor in Equation (6.64) is then specified by:

```
formula <- y ~ 1 + CTrt + ClBase4 + CV4 + ClAge + CBT +
  f(rand, model = "iid", hyper = tau_prior) +
  f(Ind,  model = "iid", hyper = tau_prior)
```

The object `tau_prior` specifies the $\Gamma(0.001, 0.001)$ precision prior:

```
tau_prior <- list(prec = list(
  prior = "loggamma",
```

```
  param = c(0.001, 0.001),
  initial = 1,
  fixed = FALSE)
)
```

The prior is specified as `loggamma` because `R-INLA` represents the precision internally on the log scale, to avoid any $\tau > 0$ constraints. Inference may then be performed, specifying the latent field posterior marginals approach `strat` and quadrature approach `int_strat`:

```
beta_prior <- list(mean = 0, prec = 1 / 100^2)

epil_inla <- function(strat, int_strat) {
  inla(
    formula,
    control.fixed = beta_prior,
    family = "poisson",
    data = Epil,
    control.inla = list(strategy = strat, int.strategy = int_strat),
    control.predictor = list(compute = TRUE),
    control.compute = list(config = TRUE)
  )
}
```

The object `beta_prior` specifies the $\mathcal{N}(0, 100^2)$ regression coefficient prior. The Poisson likelihood is specified via the `family` argument. Inferences may be then obtained via the `fit` object:

```
fit <- epil_inla(strat = "gaussian", int_strat = "grid")
```

As described in Section 6.1.4, `strat` may be set to one of `"gaussian"`, `"laplace"`, or `"simplified.laplace"` and `int_strat` may be set to one of `"eb"`, `"grid"`, or `"ccd"`.

Gaussian marginals and EB with `TMB`

With `TMB`, the log-posterior of the model is specified using a C++ template. For simple models, writing this template is usually a more involved task then specifying the formula object required for `R-INLA`. The `TMB` C++ template `epil.cpp` for the epilepsy GLMM is in Appendix C.1.1. This template specifies exactly the same model as `R-INLA` in Section 6.2.1. It is not trivial to do this, because each detail of the model must match.

Lines with a `DATA` prefix specify the fixed data inputs to be passed to `TMB`.

```
DATA
```

Examples of data inputs include the data $\mathbf{y}$ and the covariate design matrix. Lines with a `PARAMETER` prefix specify the parameters $\boldsymbol{\phi} = (\mathbf{x}, \boldsymbol{\theta})$ to be estimated.

```
PARAMETER
```

It is recommended to specify all parameters on the real scale to help performance of the optimisation procedure. More familiar versions of parameters, such as the precision rather than log precision, may be created outside the `PARAMETER` section. Lines of the form `nlp -= ddist(...)` increment the negative log-posterior, where `dist` is the name of a distribution.

```
nlp -= ddist(...)
```

In R, the `TMB` user template may now be compiled and linked:

```
compile("epil.cpp")
dyn.load(dynlib("epil"))
```

An objective function `obj` implementing $\tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y})$ and its first and second derivatives may then be created:

```
obj <- TMB::MakeADFun(
  data = dat,
  parameters = param,
  random = c("beta", "epsilon", "nu"),
```

```
  DLL = "epil"
)
```

The object `dat` is a list of data inputs passed to `TMB`. The object `param` is a list of parameter starting values passed to `TMB`. The argument `random` determines which parameters are to be integrated out with a Gaussian approximation, here set to `c("beta", "epsilon", "nu")`. Mathematically, these parameters correspond to the latent field

$$(\beta_0, \beta_{\texttt{Base}}, \beta_{\texttt{Trt}}, \beta_{\texttt{Trt}\times\texttt{Base}}, \beta_{\texttt{Age}}, \beta_{\texttt{V}_4}, \epsilon_1, \ldots, \epsilon_{59}, \nu_{1,1}, \ldots, \nu_{59,4}) = (\boldsymbol{\beta}, \boldsymbol{\epsilon}, \boldsymbol{\nu}) = \mathbf{x}. \quad (6.65)$$

The objective function `obj` may then be optimised using a gradient based optimiser to obtain $\hat{\boldsymbol{\theta}}_{\texttt{LA}}$. Here I use a quasi-Newton method (Dennis Jr et al. 1981) as implemented by `nlminb` from the `stats` R package, making use of the first derivative `obj$gr` of the objective function:

```
opt <- nlminb(
  start = obj$par,
  objective = obj$fn,
  gradient = obj$gr,
  control = list(iter.max = 1000, trace = 0)
)
```

The `sdreport` function is used to evaluate the Hessian matrix of the parameters at a particular value. Typically, these Hessian matrices are for the hyperparameters, and based on the marginal Laplace approximation. Setting `par.fixed` to the previously obtained `opt$par` returns $\hat{\mathbf{H}}_{\texttt{LA}}$. However, by setting `getJointPrecision = TRUE` the the full Hessian matrix for the hyperparameters and latent field together is returned:

```
sd_out <- TMB::sdreport(
  obj,
  par.fixed = opt$par,
```

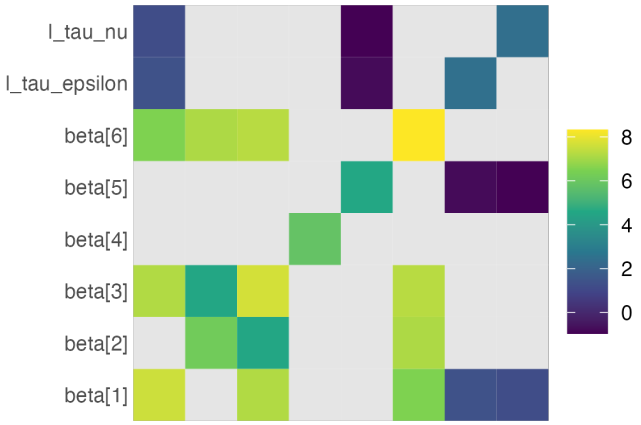

Figure 6.6: A submatrix of the full parameter Hessian obtained from `TMB::sdreport` with `getJointPrecision = TRUE` on the log scale. Entries for the latent field parameters $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}$ are omitted due to their respective lengths of 56 and 236. Light grey entries correspond to zeros on the real scale, which cannot be log transformed.

```
  getJointPrecision = TRUE
)
```

The epilepsy GLMM may also be succinctly fit in a frequentist setting (that is, using improper hyperparameter priors $p(\boldsymbol{\theta}) \propto 1$) using the formula interface provided by `glmmTMB`:

```
fit <- glmmTMB(
  y ~ 1 + CTrt + ClBase4 + CV4 + ClAge + CBT + (1 | rand) + (1 | Ind),
  data = Epil,
  family = poisson(link = "log")
)
```

Gaussian marginals and AGHQ with `TMB`

The objective function `obj` created in Section 6.2.1 may be directly passed to `aghq` to perform inference by integrating the marginal Laplace approximation over the hyperparameters using AGHQ. The argument `k` specifies the number of quadrature nodes to be used per hyperparameter dimension. Here there are two hyperparameters $\boldsymbol{\theta} = (\tau_\epsilon, \tau_\nu)$, and `k` is set to three, such that in total there are $3^2 = 9$ quadrature nodes:

```
init <- c(param$l_tau_epsilon, param$l_tau_nu)
fit <- aghq::marginal_laplace_tmb(obj, k = 3, startingvalue = init)
```

Draws from the mixture of Gaussians approximating the latent field posterior distribution (Equation (6.48)) can be obtained by:

```
samples <- aghq::sample_marginal(aghq, M = 4000)$samps
```

For a more complete `aghq` vignette, see Stringer (2021b).

Laplace marginals and EB with `TMB`

The Laplace latent field marginal $\tilde{p}_{\text{LA}}(x_i, \boldsymbol{\theta}, \mathbf{y})$ may be obtained using `TMB` by setting `random` to $\mathbf{x}_{-i}$ in the `MakeADFun` function call to approximate $p(\mathbf{x}_{-i} \,|\, x_i, \boldsymbol{\theta}, \mathbf{y})$ with a Gaussian distribution. However, it is not directly possible to do this, because the `random` argument takes a vector of strings as input (e.g. `c("beta", "epsilon", "nu")`) and does not have a native method for indexing. Instead, I took the following steps to modify the `TMB` C++ template and enable the desired indexing:

1. Include `DATA_INTEGER(i)` to pass the index i to `TMB` via the `data` argument of `MakeADFun`.
2. Concatenate the latent field to `PARAMETER_VECTOR(x_minus_i)` and `PARAMETER(x_i)` such that `random` can be set to `x_minus_i` in the call to `MakeADFun`.
3. Include `DATA_IVECTOR(x_lengths)` and `DATA_IVECTOR(x_starts)` to pass the (integer) start point and lengths of each subvector of $\mathbf{x}$ via the `data` argument of `MakeADFun`. The jth subvector may then be obtained from within the `TMB` template via `x.segment(x_starts(j), x_lengths(j))`.

The modified `TMB` C++ template `epil_modified.cpp` for the epilepsy GLMM is in Appendix C.1.2, and may be compared to the unmodified version to provide an example of implementing the above steps. After suitable alterations are made to `dat` and `param`, it is then possible to obtain the desired objective function in `TMB` via:

```
compile("epil_modified.cpp")
dyn.load(dynlib("epil_modified.cpp"))

obj_i <- MakeADFun(
  data = dat,
  parameters = param,
  random = "x_minus_i",
  DLL = "epil_modified",
  silent = TRUE,
)
```

This section takes an EB approach, fixing the hyperparameters to their modal value $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\texttt{LA}}$ obtained previously in `opt`. The latent field marginals approximation is then directly proportional to the unnormalised Laplace approximation obtained above as `obj_i`, evaluated at $(x_i, \hat{\boldsymbol{\theta}}_{\texttt{LA}})$

$$\tilde{p}(x_i \,|\, \mathbf{y}) \approx \tilde{p}_{\texttt{LA}}(x_i \,|\, \hat{\boldsymbol{\theta}}_{\texttt{LA}}, \mathbf{y}) \tilde{p}_{\texttt{LA}}(\hat{\boldsymbol{\theta}}_{\texttt{LA}} \,|\, \mathbf{y}) \tag{6.66}$$

$$\propto \tilde{p}_{\texttt{LA}}(x_i, \hat{\boldsymbol{\theta}}_{\texttt{LA}}, \mathbf{y}). \tag{6.67}$$

This expression may be evaluated at a set of GHQ nodes $z \in \mathcal{Q}(1, l)$ adapted $z \mapsto x_i(z)$ based on the mode and standard deviation of the Gaussian marginal. Here, $l = 5$ quadrature nodes were chosen to allow spline interpolation of the resulting log-posterior. Each evaluation of `obj_i`, which involves an inner optimisation loop to compute the Laplace approximation, can be initialised by $\mathbf{x}_{-i}$ set to the mode of the full N-dimensional Gaussian approximation $p_{\texttt{G}}(\mathbf{x} \,|\, \hat{\boldsymbol{\theta}}_{\texttt{LA}}, \mathbf{y})$ with the ith entry removed $\hat{\mathbf{x}}(\boldsymbol{\theta})_{-i}$. This is an efficient approach because the $(N-1)$-dimensional posterior mode, with x_i fixed, is likely to be similar to the N-dimensional posterior mode with the ith entry removed. A normalised posterior can be obtained by computing a de novo posterior normalising constant based on the set of evaluated l quadrature nodes.

This approach requires creation of the objective function `obj_i` for $i = 1, \ldots, N$. Each of these functions are then evaluated at a set of l quadrature nodes. It is inefficient to run `MakeADFun` from scratch for each i, when only one data input `i`

is changing. `TMB` does have a `DATA_UPDATE` macro, which would allow changing of data "on the R side" without retaping via:

```
obj_i$env$data$i <- i
```

Although this approach would be more efficient, if else statements on updateable data items (as used in `epil_modified.cpp`) are not supported, so this is not yet possible.

Laplace marginals and AGHQ with `TMB`

The approach taken in Section 6.2.1 may be extended by integrating the marginal Laplace approximation with respect to the hyperparameters. To perform this integration, the quadrature nodes used to integrate $p_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y})$ may be reused. The latent field marginal approximation is then

$$\tilde{p}(x_i \,|\, \mathbf{y}) \propto \sum_{\mathbf{z} \in \mathcal{Q}(m,k)} \tilde{p}_{\text{LA}}(x_i, \boldsymbol{\theta}(\mathbf{z}), \mathbf{y}) \omega(\mathbf{z}). \tag{6.68}$$

As in Section 6.2.1 this expression may be evaluated at a set of l quadrature nodes, and normalised de novo. Each objective function inner optimisation can be initialised using the mode $\hat{\mathbf{x}}(\boldsymbol{\theta}(\mathbf{z}))_{-i}$ of $p_{\text{G}}(\mathbf{x} \,|\, \boldsymbol{\theta}(\mathbf{z}), \mathbf{y})$. Integration over the hyperparameters requires each of the N objective functions to be evaluated at $k \times l$ points, rather than the $1 \times l$ points required in the EB approach.

NUTS with `tmbstan`

Running NUTS with `tmbstan` using the objective function `obj` is easy to do:

```
fit <- tmbstan::tmbstan(obj = obj, chains = 4, laplace = FALSE)
```

As specified above, the objective function with no marginal Laplace approximation is used. To instead use the marginal Laplace approximation, set `laplace = TRUE`. Convergence diagnostics for NUTS with `tmbstan` are in Appendix C.1.4.

NUTS with `rstan`

For interest in the relative inefficiency of `tmbstan`, the epilepsy model was also implemented in `Stan`. The `Stan` C++ template `epil.stan` for the epilepsy GLMM is in Appendix C.1.3. This may be of interest to users familiar with `Stan` syntax, to help provide context for `TMB`. The `Stan` template was validated as be equivalent to the `TMB` template up to a constant of proportionality. Inferences from `Stan` may be obtained by

```
fit <- rstan::stan(file = "epil.stan", data = dat, chains = 4)
```

Convergence diagnostics for NUTS with `rstan` are in Appendix C.1.4.

Comparison

Posterior means and standard deviations for the the six regression parameters $\boldsymbol{\beta}$ from the inference methods implemented in `TMB` (Section 6.2.1, 6.2.1, 6.2.1, 6.2.1) were highly similar to their `R-INLA` analogs in Section 6.2.1 (Figure 6.7). Furthermore, the posterior distributions, not only point estimates, were also highly similar. Figure 6.8 shows highly similar ECDF difference plots for Gaussian or Laplace marginals from `TMB` and `R-INLA` (as compared with results from NUTS implemented in `tmbstan`) for β_0. These results provide strong evidence that the approaches developed above to implement INLA in `TMB` are sound.

6.2.2 An example that `R-INLA` doesn't work for

An example of a model that `R-INLA` can't fit but INLA with `TMB` can. Demonstrate Laplace marginals making a difference to accuracy as compared with Gaussian marginals. The model structure was as follows. Inference was performed using:

1. Gaussian marginals and EB with `TMB`
2. Gaussian marginals and AGHQ with `TMB`
3. Laplace marginals and EB with `TMB`
4. Laplace marginals and AGHQ with `TMB`
5. NUTS with `tmbstan`

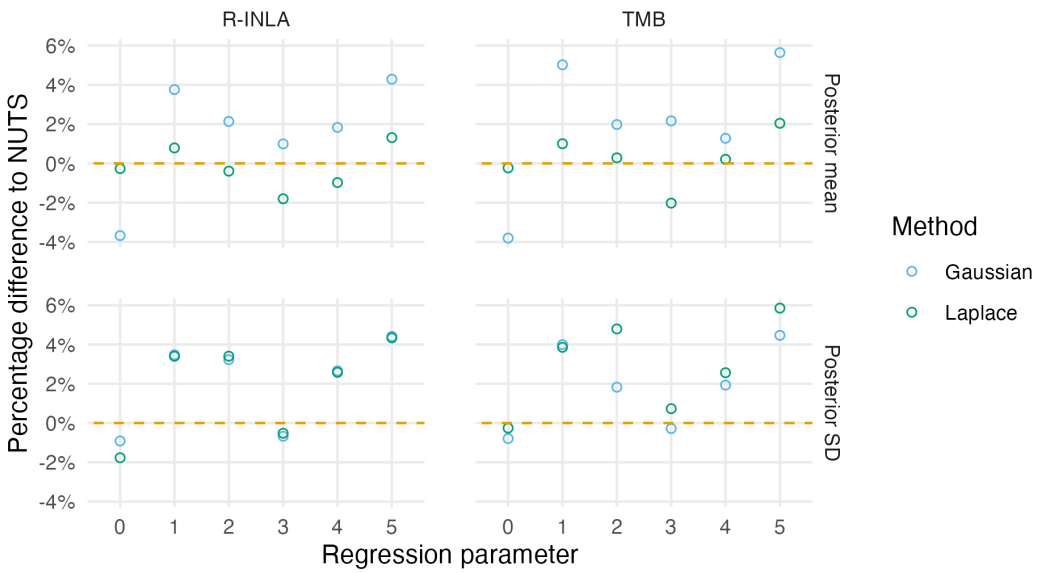

Figure 6.7: Percentage difference in posterior summary estimate obtained from NUTS as compared to that obtained from a Gaussian or Laplace marginal with quadrature over the hyperparameters. NUTS results were obtained with `tmbstan`. Results from `R-INLA` and `TMB` are similar, especially for the posterior mean, but do differ in places.

These methods were implemented as in Section 6.2.1. The results obtained were as follows.

6.3 The Naomi model

The work in this chapter was conducted in search of a fast and accurate Bayesian inference method for the Naomi model (Eaton et al. 2021). This section begins (Section 6.3.1) by describing of the simplified version of Naomi considered in this chapter. The model is simplified in that it is defined only at the time of the most recent household survey with HIV testing is considered. The nowcasting and temporal projection components of the complete model are omitted. However, these time points play a limited role in inference as they correspond to a small proportion of the total data. As such, findings about inference for the simplified model are very likely transferable to the complete model. Some features of the simplified model are left to more exhaustive description in Appendix C.2. After outlining the model, Section 6.3.2 explains why it is not an extended latent Gaussian

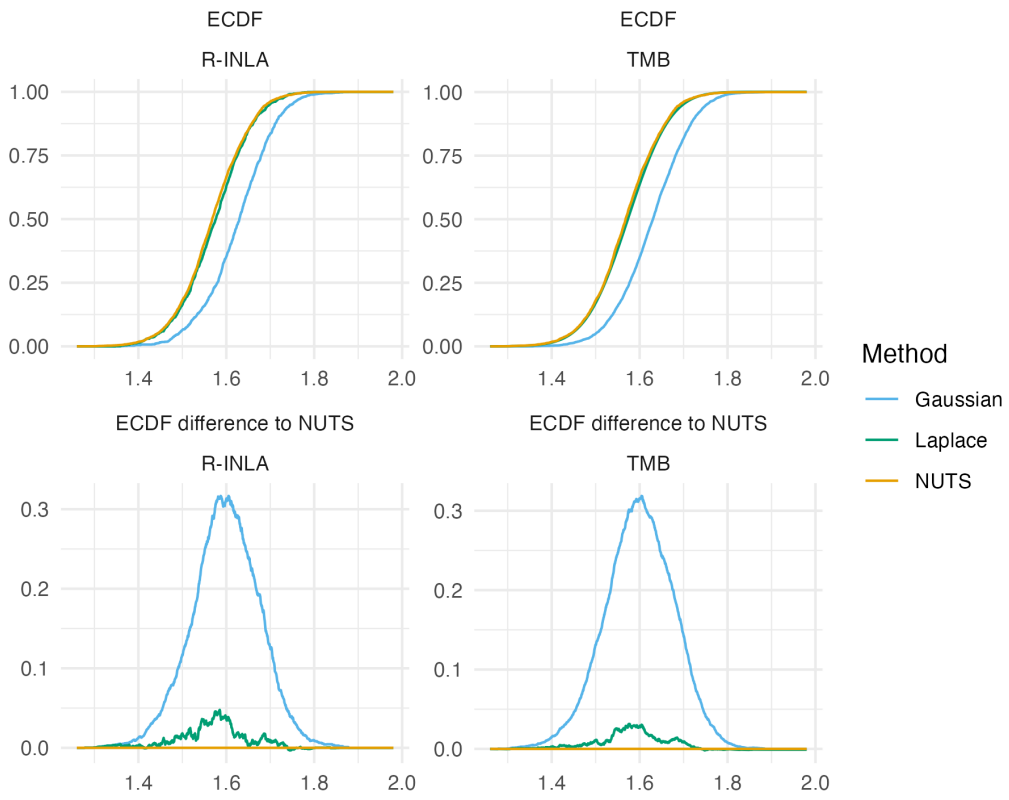

Figure 6.8: The ECDF and ECDF difference for the β_0 latent field parameter. For this parameter, the Gaussian results are inaccurate, and corrected by the Laplace. An ECDF difference of zero corresponds to obtaining exactly the same results as NUTS, taken to be the gold-standard. Results obtained using `R-INLA` and `TMB` implementations are highly similar.

model [ELGM; Stringer et al. (2022)] rather than a latent Gaussian model [LGM; Rue, Martino, and Chopin (2009)].

6.3.1 Model structure

Naomi synthesises data from three different sources to estimate HIV indicators at a district-level, by age and sex. It may be described as having three components, corresponding to these three data sources. The model components are:

- the household survey component (Section 6.3.1);
- the antenatal care (ANC) clinic testing component (Section 6.3.1);
- the antiretroviral therapy (ART) attendance component (Section 6.3.1).

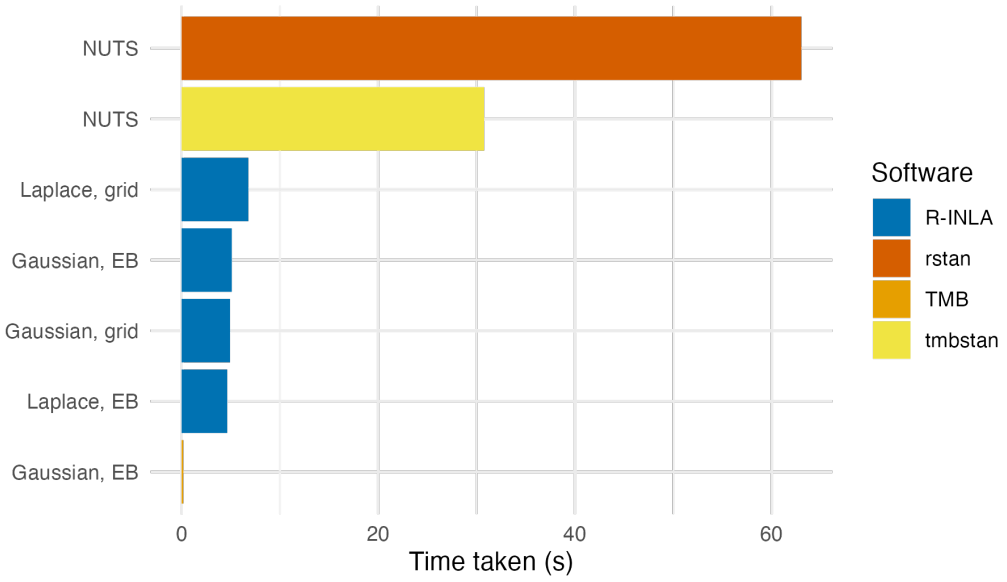

Figure 6.9: The amount of time taken (in seconds) to perform inference with each method and software implementation.

After specifying common notation used throughout the model (Section 6.3.1) each of these components is described in turn.

Notation

Consider a country in sub-Saharan Africa where a household survey with complex 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. (A district-age-sex specific quantity $z_{x,a,s}$ may then be written as z_i. When required the district, age, and sex corresponding to the index i may be recovered by $x(i) = x$, $a(i) = a$, and $s(i) = s$.)

Let:

- $N_i \in \mathbb{N}$ be the known, fixed population size;
- $\rho_i \in [0, 1]$ be the HIV prevalence;
- $\alpha_i \in [0, 1]$ be the ART coverage;
- $\kappa_i \in [0, 1]$ be the proportion recently infected among HIV positive persons;
- $\lambda_i > 0$ be the annual HIV incidence rate.

Some observations are made at an aggregate level over a collection of strata i rather than for a single i. Let $I \subseteq \mathcal{X} \times \mathcal{A} \times \mathcal{S}$ be a set of indices i for which an aggregate observation is reported. The set of all I is denoted $\mathcal{I}$ such that $I \in \mathcal{I}$.

Household survey component

Independent logistic regression models are specified for HIV prevalence and ART coverage in the general population. Without giving the linear predictors in detail, these models are specified by

$$\text{logit}(\rho_i) = \eta_i^{\rho}, \tag{6.69}$$

and

$$\text{logit}(\alpha_i) = \eta_i^{\alpha}. \tag{6.70}$$

HIV incidence rate is modelled on the log scale as

$$\log(\lambda_i) = \eta_i^{\lambda}. \tag{6.71}$$

The structured additive predictor η_i^{λ} includes terms for adult HIV prevalence and adult ART coverage. The proportion recently infected among HIV positive persons is linked to HIV incidence via

$$\kappa_i = 1 - \exp\left(-\lambda_i \cdot \frac{1-\rho_i}{\rho_i} \cdot (\Omega_T - \beta_T) - \beta_T\right), \tag{6.72}$$

where the mean duration of recent infection Ω_T and the proportion of long-term HIV infections misclassified as recent β_T are set based on informative priors for the particular HIV test used.

These three (Equations (6.69), (6.70), (6.71)) processes are each primarily informed by household survey data. Let j denote a surveyed individual, in strata $i(j)$. Weighted aggregate survey observations are calculated based on individual responses $\theta_j \in \{0, 1\}$ as

$$\hat{\theta}_I = \frac{\sum_{i(j)\in I} w_j \cdot \theta_j}{\sum_{i(j)\in I} w_j}, \tag{6.73}$$

Survey weights w_j for each of $\theta \in \{\rho, \alpha, \kappa\}$ are supplied by the survey provider. These weights aim to reduce bias by decreasing possible correlation between response and recording mechanism (Meng 2018). The weighted aggregate number of outcomes are obtained by multiplying Equation (6.73) by the Kish effective sample size [ESS; Kish (1965)]

$$y_I^\theta = m_I^\theta \hat{\theta}_I, \tag{6.74}$$

where

$$m_I^\theta = \frac{\left(\sum_{i(j)\in I} w_j\right)^2}{\sum_{i(j)\in I} w_j^2}. \tag{6.75}$$

As the Kish ESS is maximised by constant survey weights, in exchange for reducing bias, survey weighting increases variance. Equations (6.73) and (6.75) are slightly imprecise in the notation used does not reflect the fact that j only runs over individuals within the relevant denominator. In particular, for ART coverage α and the proportion recently infected among HIV positive persons κ, only those individuals who are HIV positive are included in the set. The denominator for HIV prevalence ρ includes all individuals.

The weighted aggregate number of outcomes are modelled using a binomial working likelihood (Chen et al. 2014) defined to operate on the reals

$$y_I^\theta \sim \text{xBin}(m_I^\theta, \theta_I). \tag{6.76}$$

The terms θ_I are the following weighted aggregates

$$\rho_I = \frac{\sum_{i\in I} N_i \rho_i}{\sum_{i\in I} N_i}, \quad \alpha_I = \frac{\sum_{i\in I} N_i \rho_i \alpha_i}{\sum_{i\in I} N_i \rho_i}, \quad \kappa_I = \frac{\sum_{i\in I} N_i \rho_i \kappa_i}{\sum_{i\in I} N_i \rho_i}, \tag{6.77}$$

where the denominators of α_I and κ_I reflect their restriction to HIV positive persons.

ANC testing component

Women attending ANC clinics are routinely tested for HIV, to help prevent mother-to-child transmission.

HIV prevalence $\rho_i^{\text{ANC}} \in [0, 1]$ and ART coverage $\alpha_i^{\text{ANC}} \in [0, 1]$ among pregnant women are modelled as offset from the general population indicators. (For $s(i)$

male, these quantities are not defined.) Again not detailing the linear predictors, the model is of the form

$$\text{logit}(\rho_i^{\text{ANC}}) = \text{logit}(\rho_i) + \eta_i^{\rho^{\text{ANC}}}, \tag{6.78}$$

$$\text{logit}(\alpha_i^{\text{ANC}}) = \text{logit}(\alpha_i) + \eta_i^{\alpha^{\text{ANC}}}. \tag{6.79}$$

The terms $\eta_i^{\rho^{\text{ANC}}}$ and $\eta_i^{\alpha^{\text{ANC}}}$ can be interpreted as the differences in HIV prevalence and ART coverage between pregnant women attending ANC, and the general population. As such, both the household survey data informs ANC indicators, and the ANC indicator informs general population indicators.

These two processes are informed by likelihoods specified for aggregate ANC clinic data from the year of the most recent survey. Let:

- the number of ANC clients with ascertained status be fixed as $m_I^{\rho^{\text{ANC}}}$;
- the number of those with positive status are $y_I^{\rho^{\text{ANC}}} \leq m_I^{\rho^{\text{ANC}}}$;
- the number of those already on ART prior to their first ANC visit are $y_I^{\alpha^{\text{ANC}}} \leq y_I^{\rho^{\text{ANC}}}$.

These data are modelled using nested binomial likelihoods

$$y_I^{\rho^{\text{ANC}}} \sim \text{Bin}(m_I^{\rho^{\text{ANC}}}, \rho_I^{\text{ANC}}),$$
$$y_I^{\alpha^{\text{ANC}}} \sim \text{Bin}(y_I^{\rho^{\text{ANC}}}, \alpha_I^{\text{ANC}}).$$

It is not necessary to use an extended binomial working likelihood, as in Section 3.5, because the ANC data are not survey weighted and therefore are integer valued. Analogous to Equation (6.77) in the household survey component, the weighted aggregates used here are

$$\rho_I^{\text{ANC}} = \frac{\sum_{i \in I} \Psi_i \rho_i^{\text{ANC}}}{\sum_{i \in I} \Psi_i}, \quad \alpha_I^{\text{ANC}} = \frac{\sum_{i \in I} \Psi_i \rho_i^{\text{ANC}} \alpha_i^{\text{ANC}}}{\sum_{i \in I} \Psi_i \rho_i^{\text{ANC}}},$$

where Ψ_i are the number of pregnant women, which are assume to be fixed.

ART attendance component

Data on attendance of ART clinics are routinely collected. These data provide helpful information about HIV prevalence and coverage of ART, but are challenging to use because people living with HIV sometimes choose to access ART services outside of the district that they reside in. (Indeed, this section of the model remains a challenge, and is under active development (Esra 2023).)

Multinomial logistic regression equations are used to model the probabilities of individuals accessing treatment outside their home district. Briefly, let $\gamma_{x,x'}$ be the probability that a person on ART residing in district x receives ART in district x'. These probabilities are set to $\gamma_{x,x'} = 0$ unless $x = x'$ or the two districts are neighbouring such that $x \sim x'$. As such, it is assumed that no one travels beyond their district or its immediate neighbours to receive ART services. (Of course, in reality this assumption is violated.) The log-odds are modelled using a structured additive predictor which only depends on the home district x

$$\tilde{\gamma}_{x,x'} = \text{logit}(\gamma_{x,x'}) = \eta_x^{\tilde{\gamma}}. \tag{6.80}$$

As a result, it is assumed that travel to each neighbouring district, for all age-sex strata, is equally likely.

Let the number of people observed receiving ART in strata i be y_i^A with corresponding aggregate

$$y_I^A = \sum_{i \in I} y_i^A. \tag{6.81}$$

Let the probability of a person in strata i travelling from district $x(i) = x$ to x' to receive ART be

$$\pi_{i,x(i)=x,x'} = \rho_i \alpha_i \gamma_{x(i)=x,x'}. \tag{6.82}$$

These probabilities are the product of three probabilities, each for a person in strata i:

1. the probability of a having HIV ρ_i,
2. the probability of taking ART α_i,

3. the probability of travelling from district $x(i) = x$ to district x' to receive ART $\gamma_{x(i)=x,x'}$.

Let the unobserved count of people in strata i who travel to x' to receive ART be $A_{i,x(i)=x,x'}$, such that

$$A_i = \sum_{x' \sim x, x'=x} A_{i,x(i)=x',x}. \tag{6.83}$$

Each unobserved count can be considered as arising from a binomial distribution, with sample size given by the population in strata i, here with $x(i) = x'$ such that

$$A_{i,x(i)=x',x} \sim \text{Bin}(N_{i,x(i)=x'}, \pi_{i,x(i)=x',x}). \tag{6.84}$$

Each aggregate attendance observation (Equation (6.81)) is modelled using a Gaussian approximation to a sum of binomials. This sum is over both the strata $i \in I$ and the number of ART clients travelling from district $x(i) = x'$ to x to receive treatment. The Gaussian approximation is

$$y_I^A \sim \mathcal{N}(\mu_I^A, \sigma_I^{A^2}), \tag{6.85}$$

where the mean is

$$\mu_I^A = \sum_{i \in I} \sum_{x' \sim x, x'=x} N_{i,x(i)=x'} \cdot \pi_{i,x(i)=x',x}, \tag{6.86}$$

and the variance is

$$\sigma_I^{A^2} = \sum_{i \in I} \sum_{x' \sim x, x'=x} N_{i,x(i)=x'} \cdot \pi_{i,x(i)=x',x} \cdot (1 - \pi_{i,x(i)=x',x}). \tag{6.87}$$

Equations (6.86) and (6.87) are based on a Gaussian approximation to the binomial distribution $\text{Bin}(n, p)$ with mean np and variance $np(1 - p)$, together with the equations for a linear combination of Gaussian random variables.

6.3.2 Naomi as an ELGM

In all, Naomi is a joint model on the observations

$$\mathbf{y} = (y_I^\theta), \quad \theta \in \{\rho, \alpha, \kappa, \rho^{\text{ANC}}, \alpha^{\text{ANC}}, A\}, \quad I \in \mathcal{I}. \tag{6.88}$$

The observations are modelled using the structured additive predictor $\boldsymbol{\eta}$, which includes intercept effects, age random effects, and spatial random effects which may be concatenated into the latent field $\mathbf{x}$. The latent field is controlled by hyperparameters $\boldsymbol{\theta}$ which include standard deviations, first-order autoregressive model correlation parameters, and reparameterised Besag-York-Mollie model [BYM2; Simpson et al. (2017)] proportion parameters. These features are described in more detail in Appendix C.2.

Naomi has a large Gaussian latent field, governed by a smaller number of hyperparameters $m < N$. However, it has complexities which place it outside the class of LGMs, as defined in Section 3.3.4. Instead, it is an ELGM, as defined in Section 3.3.5. In an ELGM, each mean response is allowed to depend non-linearly upon more than one structured additive predictor. The departures of Naomi from the LGM framework are enumerated below. When dependence on a specific number of structured additive predictors is given, it is in isolation, rather than in conjunction.

1. Throughout Naomi, processes are modelled at the finest district-age-sex division i, but likelihoods are defined for observations aggregated over sets of indices $i \in I$. As such, these aggregate observations are related to $|I|$ structured additive predictors, rather than just one.

2. Multiple link functions are used in Naomi, such that there is no one inverse link function g as specified in definition of an LGM. This is a relatively minor point, and it is possible to specify models with several likelihoods in `R-INLA` by setting `family` to be vector valued [Section 6.4; Gómez-Rubio (2020)].

3. In Section 6.3.1, HIV incidence depends on district-level adult HIV prevalence and ART coverage (Equation (C.6))). Each $\log(\lambda_i)$ therefore depends on

28 structured additive predictors, where 28 arises by the product of 2 sexes (male and female), 7 age groups ($\{15\text{-}19, \ldots, 45\text{-}49\}$), and 2 indicators, HIV prevalence and ART coverage. This reflects basic HIV epidemiology: incidence of sexually transmitted HIV is proportional to unsuppressed viral load among an individual's potential sexual partners. The district-level adult averages are used as a proxy.

4. In Section 6.3.1, the proportion recently infected κ_i is given by a non-linear function (Equation (6.72)) of HIV incidence λ_i, HIV prevalence ρ_i, mean duration of recent infection Ω_T and proportion of long-term HIV infections misclassified as recent β_T. Though arguably a contorting of the ELGM framework, by considering Ω_T and β_T as (Gaussian) linear predictors, then each κ_i depends on four structured additive predictors.

5. In Section 6.3.1, HIV prevalence and ART coverage among pregnant women are modelled as offset from their respective indicators in the general population. Thus each mean response depends on two structured additive predictors. The `copy` feature in `R-INLA` [Section 6.5; Gómez-Rubio (2020)] allows for this type of model structure.

6. In Section 6.3.1, nested binomial likelihoods are used.

7. In Section 6.3.1 a multinomial model with softmax link function is used. The multinomial likelihood takes as input $|\{x' : x' \sim x\}| + 1$ structured additive predictors, one for each neighbouring district plus one for remaining in the home district.

8. In Section 6.3.1 the probability of an individual receiving ART in a given district is the product of three probabilities.

Though intended for use with LGMs, the advanced features of `R-INLA` [Chapter 6; Gómez-Rubio (2020)] allow for fitting of some ELGMs as described above. In some sense then, the above exercise is mostly academic rather than practical. The

crux is that Naomi cannot be fit using `R-INLA` because it is not possible to specify such a complex model using a formula interface. The limitations of modelling with formula interfaces are not unique to `R-INLA`. Indeed, any such statistical software will see requests for users for additional features. The practical impossibility of meeting all feature requests motivates a more universal INLA implementation (Section 6.2) for advanced users.

6.4 AGHQ in moderate dimensions

Inference for the Naomi model has previously been conducted using a marginal Laplace approximation, and optimisation over the hyperparameters, implemented using `TMB`. This approach is illustrated for the epilepsy example in Section 6.2.1 and analogous for Naomi. It would be instead desirable to integrate with respect to the hyperparameters, taking an INLA-like approach as described in Section 6.1.3.

Section 6.2 attends to this challenge, by developing INLA methods which compatible with the Naomi model log-posterior as implemented in `TMB`. However, the quadrature methods previously used are not directly applicable to Naomi. The reason is that Naomi has $m = 24$ hyperparameters. Although $m = 24$ cannot be described as high-dimensional, it is certainly more than the $m < 4$ or so (low-dimensional) hyperparameters which are typical for use of INLA, and hence referred to here as moderate-dimensional. Naive use of AGHQ with the product rule requires evaluation of $|\mathcal{Q}(m, k)| = k^m$ quadrature points. This would be intractable for $m = 24$ and any $k > 1$. As a result, integrating out the hyperparameters for Naomi requires a quadrature rule which does not scale exponentially.

This section focuses on the development of an AGHQ rule for moderate dimension, for use within an inference procedure for the Naomi model. Though the rule is to be applied within a nested Laplace approximation approach, it is not limited to this setting.

6.4.1 AGHQ with variable levels

Rather than having the same number of quadrature nodes for each dimension of $\boldsymbol{\theta}$, it is possible to use a variable number of nodes per dimension. In line with the terminology used in the `mvQuad` package, I refer to the number of nodes per dimension as levels. Let $\mathbf{k} = (k_1, \ldots, k_m)$ be a vector of levels, where each $k_j \in \mathbb{Z}^+$. A GHQ grid with (potentially) variable levels is then given by

$$\mathcal{Q}(m, \mathbf{k}) = \mathcal{Q}(1, k_1) \times \cdots \times \mathcal{Q}(1, k_m). \tag{6.89}$$

The size of this grid is given by the product of the levels $|\mathcal{Q}(m, \mathbf{k})| = \prod_{j=1}^{m} k_j$. The corresponding weighting function is given by

$$\omega(\mathbf{z}) = \prod_{j=1}^{m} \omega_{k_j}(z_j). \tag{6.90}$$

This expression is a product of the univariate weighting functions for the relevant GHQ rule with k_j nodes.

6.4.2 Principal components analysis

A special case of the variable levels approach above is to set the first $s \leq m$ levels to be k and the remaining $m - s \geq 0$ levels to be one. Denote $\mathcal{Q}(m, s, k)$ to be $\mathcal{Q}(m, \mathbf{k})$ with levels $k_j = k, j \leq s$ and $k_j = 1, j > s$ for some $s \leq m$. For example, for $m = 2$ and $s = 1$ then $\mathbf{k} = (k, 1)$.

When the spectral decomposition is used to adapt the quadrature nodes, this choice of levels is analogous to principal components analysis (PCA). Figure 6.10 illustrates PCA-AGHQ for a case when $m = 2$ and $s = 1$. Since AGHQ with $k = 1$ corresponds to the Laplace approximation, PCA-AGHQ can be interpreted as performing AGHQ on the first s principal components of the inverse curvature, and a Laplace approximation on the remaining $m - s$ principal components. As such, it may be argued that PCA-AGHQ provides a natural compromise between the EB and AGHQ integration strategies.

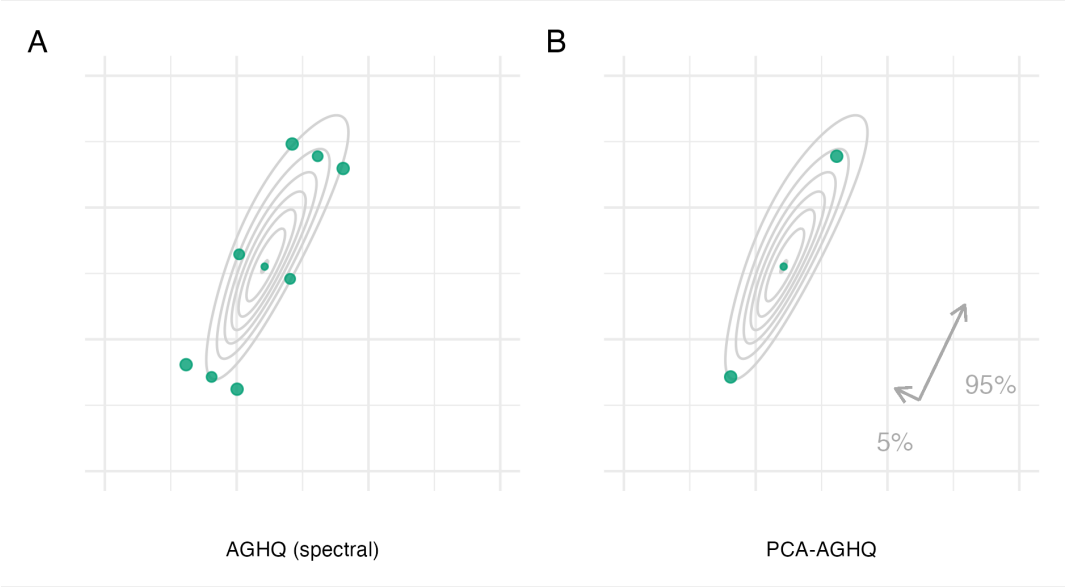

Figure 6.10: Consider the function $f(z_1, z_2) = \text{sn}(0.5z_1, \alpha = 2) \cdot \text{sn}(0.8z_1 - 0.5z_2, \alpha = -2)$ as described in Figure 6.3. Panel A shows the AGHQ nodes with a spectral matrix decomposition, as usual. Panel B shows the adapted PCA-AGHQ nodes $\mathcal{Q}(2, 1, 3)$. These nodes correspond exactly to those in Panel A along the first eigenvector. The proportion of variation explained by this direction is around 95%, with the remaining 5% explained by the second eigenvector. As before, each panel shows the quadrature estimate of the integral I.

For concreteness, the normalising constant obtained by application of PCA-AGHQ to integration of the marginal Laplace approximation (Equation (6.40)) is given by

$$\tilde{p}_{\texttt{PCA}}(\mathbf{y}) = |\hat{\mathbf{E}}_{\texttt{LA}}\hat{\mathbf{\Lambda}}_{\texttt{LA}}^{1/2}| \sum_{\mathbf{z}\in\mathcal{Q}(m,s,k)} \tilde{p}_{\texttt{LA}}(\hat{\mathbf{E}}_{\texttt{LA},s}\hat{\mathbf{\Lambda}}_{\texttt{LA},s}^{1/2}\mathbf{z} + \hat{\boldsymbol{\theta}}_{\texttt{LA}}, \mathbf{y})\omega(\mathbf{z}), \tag{6.91}$$

where $\hat{\mathbf{E}}_{\texttt{LA},s}$ is an $m \times s$ matrix containing the first s eigenvectors, $\hat{\mathbf{\Lambda}}_{\texttt{LA},s}$ is the $s \times s$ diagonal matrix containing the first s eigenvalues, and

$$\omega(\mathbf{z}) = \prod_{j=1}^{s} \omega_s(z_j) \times \prod_{j=s+1}^{d} \omega_1(z_j). \tag{6.92}$$

6.5 Malawi case-study

This section presents a case study of Bayesian inference methods applied to the Naomi model in Malawi. Data from Malawi has previously been used to demonstrate

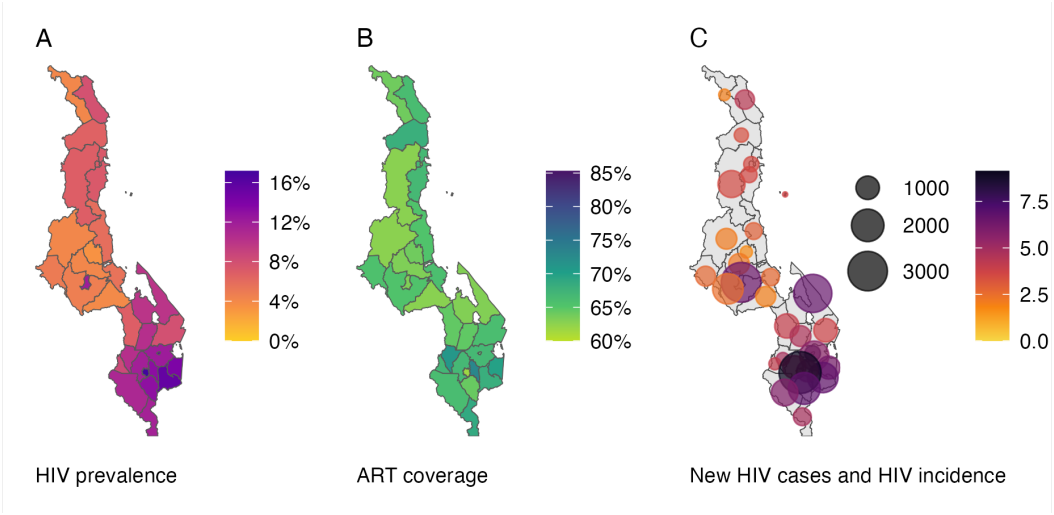

Figure 6.11: District-level HIV prevalence, ART coverage, and new HIV cases and HIV incidence for adults 15-49 in Malawi. Inference conducted using a Gaussian approximation and EB via `TMB`.

the Naomi model, including as a part of the `naomi` R package vignette available from `https://github.com/mrc-ide/naomi`. Malawi was chosen in part because it has a small number of districts, $n = 30$, limiting the computational demand of the model.

Four Bayesian inference approaches were considered:

1. Gaussian marginals and EB with `TMB`. This is the method that has previously been used for Naomi. As short-hand, this method is referred to as GEB.
2. Laplace marginals and EB with `TMB`. This is a novel method. As short-hand, this method is referred to as LEB.
3. Gaussian marginals and PCA-AGHQ with `TMB` This is a novel method. As short-hand, this method is referred to as PCA-AGHQ.
4. NUTS with `tmbstan`. This represents a gold-standard.

The `TMB` C++ user-template used to specify the log-posterior was the same for each approach, and described in Appendix C.2.4. The dimension of the latent field was $N = 467$ and the dimension of the hyperparameters was $m = 24$. For EB and PCA-AGHQ, hyperparameter and latent field samples were simulated following deterministic inference. For all methods, age-sex-district specific HIV

prevalence, ART coverage and HIV incidence were simulated from the latent field and hyperparameter posterior samples. Model outputs are illustrated in Figure 6.11.

6.5.1 NUTS convergence

The effective sample size ratios generated sampling with NUTS were low. Four chains run in parallel for 100000 iterations were required to obtain acceptable NUTS diagnostics. For ease-of-storage, these samples were thinned by a factor of 20 . There were no divergent transitions, and the largest potential scale reduction factor (Gelman and Rubin 1992; Vehtari et al. 2021) was $\hat{R} = 1.02$. These diagnostics were sufficient to treat results from NUTS as a gold-standard, though inaccuracies remain possible.

6.5.2 Use of PCA-AGHQ

A Scree plot based on the spectral decomposition of $\hat{\mathbf{H}}_{\texttt{LA}}(\boldsymbol{\theta}_{\texttt{LA}})^{-1}$ was used to select the number of principal components to keep. Keeping $s = 8$ principal components was sufficient to explain some proportion of total variation. The reduced rank approximation to the inverse curvature with this choice of s was visually similar to the full rank matrix.

6.5.3 Model assessment

6.5.4 Inference comparison

6.5.5 Exceedance probabilities

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".

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. The Naomi model can be used to assess the probability of a strata having high incidence by evaluating $\mathbb{P}(\lambda_i > 0.01)$.

6.6 Discussion

This chapter made two main contributions. First, the universal INLA implementation of Section 6.2. Second, the PCA-AGHQ rule (Sections 6.4) with application to INLA for Naomi (Section 6.5). This section discusses these contributions in turn, before outlining suggestions for future related work.

6.6.1 A universal INLA implementation

After requesting for the generalised binomial distribution used in Equation (6.76) to be included in `R-INLA` a prototype version was shortly made available. However, it is more sustainable to enable users to implement their own distributions and models. Discussion of Bayesian inference software in Štrumbelj et al. (2023).

6.6.2 PCA-AGHQ with application to INLA for Naomi

For the simplified Naomi model applied to data from Malawi, INLA with Gaussian marginals and the PCA-AGHQ quadrature rule was more accurate at inferring latent field posterior marginal distributions than with an EB quadrature rule. However, model output posterior marginals did not see the same improvements. The approximate posterior exceedance probabilities from both EB and PCA-AGHQ had systematic inaccuracies as compared with NUTS. EB and PCA-AGHQ were substantially faster than NUTS, which took over two days to reach convergence.

The inaccuracies in model outputs from EB and PCA-AGHQ have the potential to meaningfully mislead policy. As such, where possible gold-standard NUTS results should be computed. Though NUTS is to slow too run during a workshop, it could be run afterwards. That said, Malawi is one of the countries with the fewest number of districts. As NUTS took days to run in Malawi, for larger countries, with hundreds of districts, it may not be possible to run NUTS to convergence.

PCA-AGHQ and NUTS could be added to the Naomi web interface (`https://naomi.unaids.org`) as an alternative to EB. Analysts would be able to quickly iterate over model

options using EB, before switching to a more accurate approach once they are happy with the results.

PCA-AGHQ can be adjusted to suit the computational budget available by choice of the number of dimensions kept in the PCA s and the number of points per dimension k. The scree plot is a well established heuristic for choosing s. Heuristics for choosing k are less well established. Whether it is preferable for a given computational budget to increase s or increase k is an open question. Further strategies, such as gradually lowering k over the principal components, could also be considered.

6.6.3 Suggestions for future work

Finally, this section presents suggestions for future work based on this chapter. Some suggestions relate more to individual contributions, others take a broader view, or relate to multiple contributions.

Further comparisons

Comparison to additional Bayesian inference methods could be included in Section 6.5. Three further comparisons stand out as being particularly valuable:

1. There exist other quadrature rules for moderate dimension, such as the CCD. It would be of interest to compare INLA with a PCA-AGHQ rule to INLA with other such quadrature rules.
2. NUTS is not especially well suited to sampling from Gaussian latent field models like Naomi. Other MCMC algorithms, such as blocked Gibbs sampling (Geman and Geman 1984) or slice sampling (Neal 2003), could be considered. Both of these MCMC algorithms are implemented and can be customised, including the choice of block structure, within the `NIMBLE` probabilistic programming language (de Valpine et al. 2017).
3. Rather than use quadrature to integrate the marginal Laplace approximation, an alternative approach, implemented in `tmbstan` by setting `laplace = TRUE`,

is to run HMC (Monnahan and Kristensen 2018; Margossian et al. 2020). When run to convergence, inferential error of this method would solely be due to the Laplace approximation, helping to clarify the extent to which the inferential error of INLA is attributable to the quadrature grid.

Investigation of quadrature grid

Gaussian times polynomial kernel fit to the 24 dimensional NUTS hyperparameter samples. Assess whether PCA-AGHQ is suitable. Or whether AGHQ is suitable. Could be generalised.

Better quadrature grids

PCA-AGHQ is a sensible approach to allocating more computational to dimensions which contribute more to the integral in question. However, its application to Naomi surfaced instances where it overlooked potential benefits, or otherwise did not behave as one might wish:

1. The amount of variation explained in the Hessian matrix may not be of direct interest. For the Naomi model, interest is in the effect of including each dimension on the relevant model outputs. As such, using alternative measures of importance from sensitivity analysis, such as Shapley values (Shapley et al. 1953) or Sobol indicies, could be preferable.
2. Use of PCA is challenging when the dimensions have different scales. For the Naomi model, logit-scale hyperparameters were systematically favoured over those on the log-scale.
3. When the quadrature rule is used within an INLA algorithm, it is more important to allocate quadrature nodes to those hyperparameter marginals which are non-Gaussian. This is because the Laplace approximation is exact when the integrand is Gaussian, so a single quadrature node is sufficient. 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 EB. The measures of fit would have to be for marginals, ruling out approaches like PSIS (Yao et al. 2018) which operate on joint distributions.

Computational improvements

1. Approximation: Implement the simplified Laplace marginals of Wood (2020) (Section 6.1.3).
2. Parallelisation: Integration over a moderate number of hyperparameters resulted in use of quadrature grids with a large number of nodes. Computation at each node is independent, so algorithm run-time could potentially be significantly improved using parallel computing. This point is discussed by Kristensen et al. (2016) who highlight that `TMB` could applied to perform function evaluations in parallel, for example using the `parallel` R package.
3. Hardware: Further computational speed-ups might be obtained using graphics processing units (GPUs) specialised for the relevant matrix operations.

Statistical theory

The class of functions which are integrated exactly by PCA-AGHQ remains to be shown. Theorem 1 of Stringer et al. (2022) bounds the total variation error of AGHQ, establishing convergence in probability of coverage probabilities under the approximate posterior distribution to those under the true posterior distribution. Similar theory could be established for PCA-AGHQ, or more generally AGHQ with varying levels. The challenge of connecting this theory to nested use of any quadrature rule, like that in the INLA algorithm, remains an important open question.

Exploration of the accuracy of INLA for complex models

A universal INLA implementation could be used to assess the accuracy of INLA for a wider range of models. Among the ELGM-type structures of particular

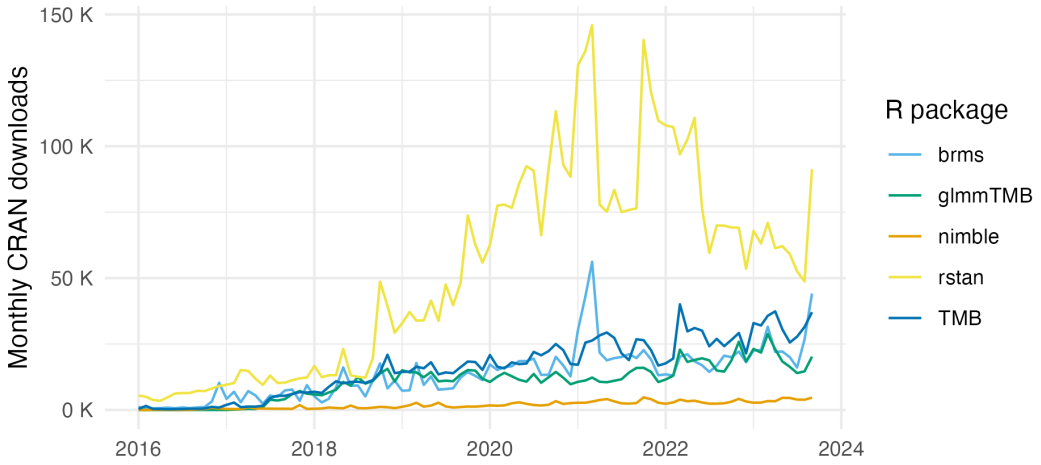

Figure 6.12: Monthly R package downloads from the Comprehensive R Archive Network (CRAN) for `brms`, `glmmTMB`, `nimble`, `rstan` and `TMB`, obtained using the Csárdi (2023) R package. Unfortunately, `R-INLA` is not available from CRAN, and so could not be included in this figure. The official `rstan` documentation recommends installation of a development version hosted outside CRAN.

interest are aggregated Gaussian process models (Nandi et al. 2023) and evidence synthesis models (Amoah et al. 2020).

Methods dissemination

The approach used to implement Laplace marginals with `TMB` was relatively ad-hoc, and involved modification of the `TMB` C++ template (Section 6.2.1). For wider dissemination of this method, it is important that the user is not burdened with making these modifications themself. One possibility would be to change the `random` argument in `TMB::MakeADFun` to allow for indexing. Another (less desirable) option would be to algorithimically generate the modified `TMB` C++ template based on the original template.

Though gaining in popularity, the user-base of `TMB` is relatively small, and package downloads are in large part driven by use of the more easy-to-use `glmmTMB` package (Figure 6.12). For users unfamiliar with C++, it can be challenging to use `TMB` directly. One possibility is to look to disseminate methods via the users of `glmmTMB`. Another approach would be to implement methods in other probabilistic programming languages, such as `Stan` or `NIMBLE`. Implementation in `Stan` is made

possible by the `bridgestan` package (Ward 2023), which provides access to the methods of a `Stan` model, and could be combined with the prototyping of an adjoint-differentiated Laplace approximation done in `Stan` by Margossian et al. (2020). The ratio of downloads of `rstan` as compared with `brms` suggests a larger proportion of `Stan` users are interested in specifying their own model. Implementation in `NIMBLE` is also possible as of version >1.0.0 which includes functionality for automatic differentiation and Laplace approximation [Part V; de Valpine et al. (2023)] like `TMB` built using `CppAD`. Both `NIMBLE` and `Stan` developers are actively looking into implementation of algorithms combining the Laplace approximation and quadrature.

7 Conclusions

7.1 Strengths

- Chapter 4 conducted thorough experiments to compare models for spatial structure using tools for model assessment such as proper scoring rules and posterior predictive checks.

- Chapter 5 estimated HIV risk group proportions for AGYW, facilitating countries in SSA to prioritise their delivery of HIV prevention services. The number of new infections that might be reached under a variety of risk stratification strategies were analyzed. The `R-INLA` software was used to specify multinomial spatio-temporal models via the Poisson-multinomial transformation, including complex two- and three-way Kronecker product interactions.

- Chapter 6 developed a novel Bayesian inference method, motivated by a challenging and practically important problem in HIV surveillance.

7.2 Weaknesses

7.3 Future work

Avenues for future work include:

- Extending the risk group model described in Chapter 5 to include all adults 15-49. This may involve modelling of age-stratified sexual partnerships (Wolock et al. 2021). Such a model would likely fall out of the scope of `R-INLA`, but would be possible to write with `TMB` and therefore amenable to the methods discussed in Chapter 6.

7.4 Conclusions

- Modelling complex data often pushes on the boundaries of the available statistical toolkit.
- A challenge encountered while conducting this research was the difficulty of implementing identical models across multiple frameworks, looking to study inference methods. Or, of a similarly fraught nature, comparing different models implemented in different frameworks, looking to study model differences. The frequently asked questions section of the `R-INLA` website (Rue 2023) notes that "the devil is in the details". I have resolved this challenge by using a given `TMB` model template to fit models using multiple inference methodologies. The benefits of such a ecosystem of packages are noted by Stringer (2021a). I particularly highlight the benefit of enabling analysts to easily vary their choice of inference method based on the stage of model development that they are in.
- To the best of my abilities, this thesis, and the work described within it, was written in keeping with the principles of open science. I hope that doing so allows my work to be scrutinised, and optimistically built upon. This would not have been possible without a range of tools from the R ecosystem such as

`rmarkdown` and `rticles`, as well as those developed within the MRC Centre for Global Infectious Disease Analysis such as `orderly` and `didehpc`.

Appendices

A

Models for spatial structure

A.1 Comparison of AGHQ to NUTS

A.2 Simulation study

A.3 HIV study

B

A model for risk group proportions

B.1 The Global AIDS Strategy

Prioritisation strata	Criterion
Low	0.3-1.0% incidence and low-risk behaviour, or <0.3% incidence and high-risk behaviour
Moderate	1.0-3.0% incidence and low-risk behaviour, or 0.3-1.0% incidence and high-risk behaviour
High	1.0-3.0% incidence and high-risk behaviour
Very high	>3.0% incidence

Table B.1: Prioritisation strata according to HIV incidence in the general population and behavioural risk.

Intervention	Low	Moderate	High	Very High
Condoms and lube for those with non-regular partners(s) with unknown STI status and not on PrEP	50%	70%	95%	95%
STI screening and treatment	10%	10%	80%	80%
Access to PEP	-	-	50%	90%
PrEP use	-	5%	50%	50%
Economic empowerment	-	-	20%	20%

Table B.2: Commitments to be met for each intervention in terms of proportion of the prioritisation strata reached. The symbol "-" represents no commitment.

B.2 Household survey data

	Type	Year	Transactional sex question	Sample size			
				15-19	20-24	25-29	Total
Botswana							
	BAIS	2013	✓	557	588	649	1794
Total				557	588	649	1794
Cameroon							
	DHS	2004	✗	2675	2207	1732	6614
	DHS	2011	✗	3588	3115	2655	9358
	PHIA	2017	✗	2620	2339	2259	7218
	DHS	2018	✓	3349	2463	2345	8157
Total				12232	10124	8991	31347
Kenya							
	DHS	2003	✗	1819	1709	1391	4919
	DHS	2008	✗	1767	1743	1419	4929
	DHS	2014	✗	2861	2534	2858	8253
Total				6447	5986	5668	18101
Lesotho							
	DHS	2004	✗	1761	1455	1026	4242
	DHS	2009	✗	1833	1543	1194	4570
	DHS	2014	✗	1537	1292	1067	3896
	PHIA	2017	✓	1156	1202	1054	3412
Total				6287	5492	4341	16120
Mozambique							
	AIS	2009	✗	1031	1106	987	3124
	DHS	2011	✗	2932	2299	2206	7437
	AIS	2015	✗	1552	1389	1080	4021
Total				5515	4794	4273	14582
Malawi							
	DHS	2000	✗	2914	2998	2358	8270
	DHS	2004	✗	2407	2823	2135	7365
	DHS	2010	✗	5031	4387	4309	13727
	DHS	2015	✓	5273	5094	3976	14343
	PHIA	2016	✓	1646	1934	1511	5091
Total				17271	17236	14289	48796
Namibia							
	DHS	2000	✗	1427	1313	1098	3838

	DHS	2006	✗	2203	1869	1544	5616
	DHS	2013	✗	1852	1709	1481	5042
	PHIA	2017	✓	1491	1525	1370	4386
Total				6973	6416	5493	18882
Eswatini							
	DHS	2006	✗	1265	1027	731	3023
	PHIA	2017	✗	1031	895	811	2737
Total				2296	1922	1542	5760
Tanzania							
	AIS	2003	✗	1466	1377	1270	4113
	AIS	2007	✗	2137	1676	1509	5322
	DHS	2010	✗	2221	1860	1613	5694
	AIS	2012	✗	2474	1923	1815	6212
	PHIA	2016	✓	2999	2845	2521	8365
Total				11297	9681	8728	29706
Uganda							
	DHS	2000	✗	1687	1541	1326	4554
	DHS	2006	✗	1948	1660	1404	5012
	AIS	2011	✗	2451	2164	1921	6536
	DHS	2011	✗	2025	1664	1614	5303
	DHS	2016	✓	4276	3782	3014	11072
	PHIA	2016	✗	3289	3059	2574	8922
Total				15676	13870	11853	41399
South Africa							
	DHS	2016	✓	1505	1408	1397	4310
Total				1505	1408	1397	4310
Zambia							
	DHS	2007	✗	1598	1405	1373	4376
	DHS	2013	✗	3685	3036	2789	9510
	PHIA	2016	✓	2120	2045	1619	5784
	DHS	2018	✓	3112	2687	2166	7965
Total				10515	9173	7947	27635
Zimbabwe							
	DHS	1999	✗	1467	1230	1011	3708
	DHS	2005	✗	2128	1943	1438	5509
	DHS	2010	✗	1963	1796	1679	5438
	DHS	2015	✓	2154	1777	1646	5577
	PHIA	2016	✓	2114	1817	1573	5504
Total				9826	8563	7347	25736

Total	106397	95253	82518	284168

Table B.3: All of the surveys that used in the analysis and their sample sizes, disaggregated by respondent age.

Survey	Exclusion reason
MOZ2003DHS	No GPS coordinates available to place survey clusters within districts.
TZA2015DHS	Insufficient sexual behaviour questions.
UGA2004AIS	Unable to download region boundaries.
ZMB2002DHS	No GPS coordinates available to place survey clusters within districts.

Table B.4: All of that surveys that were excluded from the analysis.

B.3 Spatial analysis levels

Country	Number of areas	Analysis level
Botswana	27	3
Cameroon	58	2
Kenya	47	2
Lesotho	10	1
Mozambique	161	3
Malawi	33	5
Namibia	38	2
Eswatini	4	1
Tanzania	195	4
Uganda	136	3
South Africa	52	2
Zambia	116	2
Zimbabwe	63	2

Table B.5: The numer of areas and analysis levels for each country that were used in the analysis.

B.4 Survey questions and risk group allocation

Variable(s)	Description
`v501`	Current marital status of the respondent.
`v529`	Computed time since last sexual intercourse.
`v531`	Age at first sexual intercourse–imputed.
`v766b`	Number of sexual partners during the last 12 months (including husband).
`v767[a, b, c]`	Relationship with last three sexual partners. Options are: spouse, boyfriend not living with respondent, other friend, casual acquaintance, relative, commercial sex worker, live-in partner, other.
`v791a`	Had sex in return for gifts, cash or anything else in the past 12 months. Asked only to women 15-24 who are not in a union.

Table B.6: The survey questions included in AIDS Indicator Survey (AIS) and Demographic and Health Surveys (DHS).

Variable(s)	Description
`part12monum`	Number of sexual partners during the last 12 months (including husband).
`part12modkr`	Reason for leaving `part12monum` blank.
`partlivew[1, 2, 3]`	Does the person you had sex with live in this household?
`partrelation[1, 2, 3]`	Relationship with last three sexual partners. Options are: husband, live-in partner, partner (not living with), ex-spouse/partner, friend/acquaintance, sex worker, sex worker client, stranger, other, don't know, refused.
`sellsx12mo`	Had sex for money and/or gifts in the last 12 months.
`buysx12mo`	Paid money or given gifts for sex in the last 12 months.

Table B.7: The survey questions included in Population-Based HIV Impact Assessment (PHIA) surveys.

C

Fast approximate Bayesian inference

C.1 Epilepsy example

C.1.1 TMB C++ template

```
// epil.cpp

#include <TMB.hpp>

template <class Type>
Type objective_function<Type>::operator()()
{
  DATA_INTEGER(N);
  DATA_INTEGER(J);
  DATA_INTEGER(K);
  DATA_MATRIX(X);
  DATA_VECTOR(y);
  DATA_MATRIX(E); // Epsilon matrix

  PARAMETER_VECTOR(beta);
```

```
  PARAMETER_VECTOR(epsilon);
  PARAMETER_VECTOR(nu);
  PARAMETER(l_tau_epsilon);
  PARAMETER(l_tau_nu);

  Type tau_epsilon = exp(l_tau_epsilon);
  Type tau_nu = exp(l_tau_nu);
  Type sigma_epsilon = sqrt(1 / tau_epsilon);
  Type sigma_nu = sqrt(1 / tau_nu);
  vector<Type> eta(X * beta + nu + E * epsilon); // Linear predictor
  vector<Type> lambda(exp(eta));

  Type nll;
  nll = Type(0.0);

  // Note: dgamma() is parameterised as (shape, scale)
  // R-INLA is parameterised as (shape, rate)
  nll -= dlgamma(l_tau_epsilon, Type(0.001), Type(1.0 / 0.001), true);
  nll -= dlgamma(l_tau_nu, Type(0.001), Type(1.0 / 0.001), true);
  nll -= dnorm(epsilon, Type(0), sigma_epsilon, true).sum();
  nll -= dnorm(nu, Type(0), sigma_nu, true).sum();
  nll -= dnorm(beta, Type(0), Type(100), true).sum();

  nll -= dpois(y, lambda, true).sum();

  ADREPORT(tau_epsilon);
  ADREPORT(tau_nu);

  return(nll);
}
```

C.1.2 Modified TMB C++ template

```cpp
// epil_modified.cpp

#include <TMB.hpp>

template <class Type>
Type objective_function<Type>::operator()()
{
  DATA_INTEGER(N);
  DATA_INTEGER(J);
  DATA_INTEGER(K);
  DATA_MATRIX(X);
  DATA_VECTOR(y);
  DATA_MATRIX(E); // Epsilon matrix

  DATA_IVECTOR(x_starts);  // Start index of each subvector of x
  DATA_IVECTOR(x_lengths); // Length of each subvector of x
  DATA_INTEGER(i);         // Index i

  PARAMETER(x_i);
  PARAMETER_VECTOR(x_minus_i);

  vector<Type> x(301);
  int k = 0;
  for (int j = 0; j < 301; j++) {
    if (j + 1 == i) { // +1 because C++ does zero-indexing
      x(j) = x_i;
    } else {
      x(j) = x_minus_i(k);
```

```
    k++;
  }
}

vector<Type> beta = x.segment(x_starts(0), x_lengths(0));
vector<Type> epsilon = x.segment(x_starts(1), x_lengths(1));
vector<Type> nu = x.segment(x_starts(2), x_lengths(2));

PARAMETER(l_tau_epsilon);
PARAMETER(l_tau_nu);

Type tau_epsilon = exp(l_tau_epsilon);
Type tau_nu = exp(l_tau_nu);
Type sigma_epsilon = sqrt(1 / tau_epsilon);
Type sigma_nu = sqrt(1 / tau_nu);
vector<Type> eta(X * beta + nu + E * epsilon); // Linear predictor
vector<Type> lambda(exp(eta));

Type nll;
nll = Type(0.0);

// Note: dgamma() is parameterised as (shape, scale)
// R-INLA is parameterised as (shape, rate)
nll -= dlgamma(l_tau_epsilon, Type(0.001), Type(1.0 / 0.001), true);
nll -= dlgamma(l_tau_nu, Type(0.001), Type(1.0 / 0.001), true);
nll -= dnorm(epsilon, Type(0), sigma_epsilon, true).sum();
nll -= dnorm(nu, Type(0), sigma_nu, true).sum();
nll -= dnorm(beta, Type(0), Type(100), true).sum();

nll -= dpois(y, lambda, true).sum();
```

```
  ADREPORT(tau_epsilon);
  ADREPORT(tau_nu);

  return(nll);
}
```

C.1.3 Stan C++ template

```
// epil.stan

data {
  int<lower=0> N;             // Number of patients
  int<lower=0> J;             // Number of clinic visits
  int<lower=0> K;             // Number of predictors (inc. intercept)
  matrix[N * J, K] X;         // Design matrix
  int<lower=0> y[N * J];      // Outcome variable
  matrix[N * J, N] E;         // Epsilon matrix
}

parameters {
  vector[K] beta;             // Vector of coefficients
  vector[N] epsilon;          // Patient specific errors
  vector[N * J] nu;           // Patient-visit errors
  real<lower=0> tau_epsilon; // Precision of epsilon
  real<lower=0> tau_nu;       // Precision of nu
}

transformed parameters {
  vector[N * J] eta = X * beta + nu + E * epsilon;  // Linear predictor
}
```

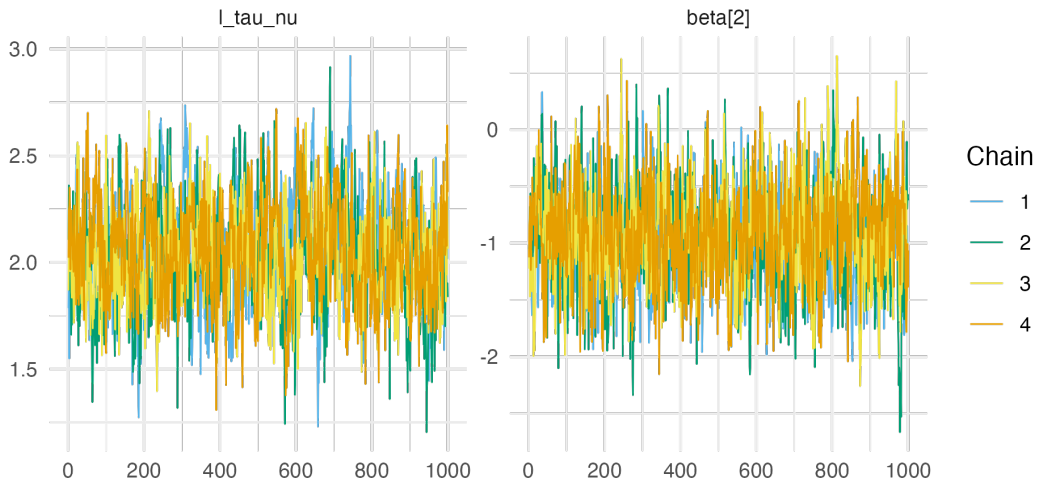

Figure C.1: Figure caption.

```
model {
  beta ~ normal(0, 100);
  tau_epsilon ~ gamma(0.001, 0.001);
  tau_nu ~ gamma(0.001, 0.001);
  epsilon ~ normal(0, sqrt(1 / tau_epsilon));
  nu ~ normal(0, sqrt(1 / tau_nu));
  y ~ poisson_log(eta);
}
```

C.1.4 NUTS convergence diagnostics

tmbstan
rstan

C.2 Simplified Naomi model description

This section describes the simplified version of the Naomi model (Eaton et al. 2021) in more detail. The concise i indexing used in Section 6.3 is replaced by a more complete x, s, a indexing. There are four sections:

1. Section C.2.1 gives the process specifications, giving the terms in each structured additive predictor, along with their distributions.

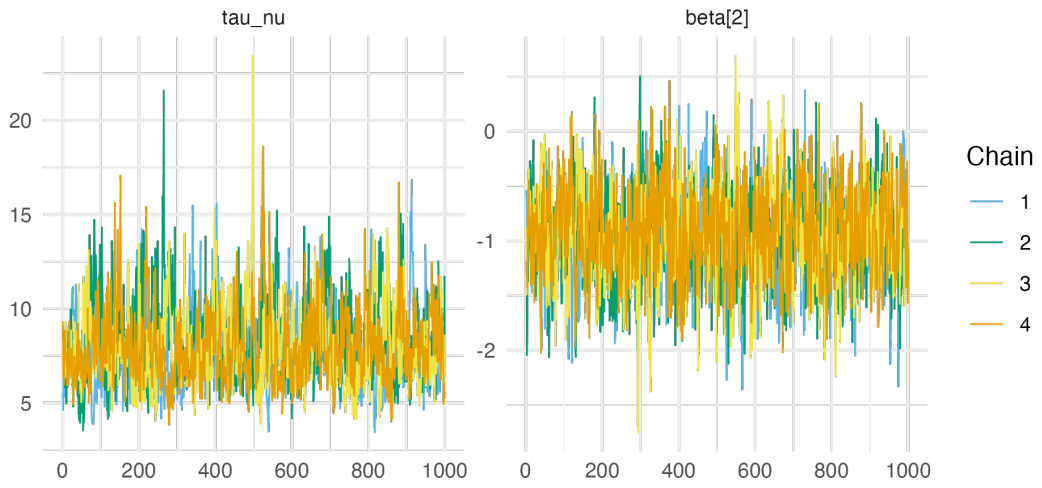

Figure C.2: Figure caption.

2. Section C.2.2 gives additional details about the likelihood terms not provided in Section 6.3.
3. Section C.2.3 gives identifiability constraints used in circumstances where incomplete data is available for the country.
4. Section C.2.4 provides details of the `TMB` implementation.

C.2.1 Process specification

Table C.1: The Naomi model can be conceptualised as having five processes. This table gives the number of latent field parameters and hyperparameters in each process, where n is the number of districts in the country.

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

HIV prevalence

HIV prevalence $\rho_{x,s,a} \in [0,1]$ was modelled on the logit scale using the structured additive predictor

$$\text{logit}(\rho_{x,s,a}) = \beta_0^\rho + \beta_S^{\rho,s=\text{M}} + \mathbf{u}_a^\rho + \mathbf{u}_a^{\rho,s=\text{M}} + \mathbf{u}_x^\rho + \mathbf{u}_x^{\rho,s=\text{M}} + \mathbf{u}_x^{\rho,a<15} + \boldsymbol{\eta}_{R_x,s,a}^\rho. \quad \text{(C.1)}$$

Table C.2 provides a description of the terms included in Equation (C.1). Independent half-normal prior distributions were chosen for the five standard deviation terms

$$\{\sigma_A^\rho, \sigma_{AS}^\rho, \sigma_X^\rho, \sigma_{XS}^\rho, \sigma_{XA}^\rho\} \sim \mathcal{N}^+(0, 2.5), \quad \text{(C.2)}$$

independent uniform prior distributions for the two AR1 correlation parameters

$$\{\phi_A^\rho, \phi_{AS}^\rho\} \sim \mathcal{U}(-1, 1), \quad \text{(C.3)}$$

and independent beta prior distributions for the two BYM2 proportion parameters

$$\{\phi_X^\rho, \phi_{XS}^\rho\} \sim \text{Beta}(0.5, 0.5). \quad \text{(C.4)}$$

Table C.2: Each term in Equation (C.1) together with (where applicable) its prior distribution and a written description of its role.

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
$\mathbf{u}_a^\rho$	$\text{AR1}(\sigma_A^\rho, \phi_A^\rho)$	Age random effects for women
$\mathbf{u}_a^{\rho,s=\text{M}}$	$\text{AR1}(\sigma_{AS}^\rho, \phi_{AS}^\rho)$	Age random effects for the difference in logit prevalence for men compared to women age a
$\mathbf{u}_x^\rho$	$\text{BYM2}(\sigma_X^\rho, \phi_X^\rho)$	Spatial random effects for women
$\mathbf{u}_x^{\rho,s=\text{M}}$	$\text{BYM2}(\sigma_{XS}^\rho, \phi_{XS}^\rho)$	Spatial random effects for the difference in logit prevalence for men compared to women in district x
$\mathbf{u}_x^{\rho,a<15}$	$\text{ICAR}(\sigma_{XA}^\rho)$	Spatial random effects for the difference in logit paediatric prevalence to adult women prevalence in district x
$\boldsymbol{\eta}_{R_x,s,a}^\rho$	–	Fixed offsets specifying assumed odds ratios for prevalence outside the age ranges for which data were available. Calculated from Spectrum model (Stover, Glaubius, et al. 2019) outputs for region R_x

ART coverage

ART coverage $\alpha_{x,s,a} \in [0,1]$ was modelled on the logit scale using the structured additive predictor

$$\text{logit}(\alpha_{x,s,a}) = \beta_0^{\alpha} + \beta_S^{\alpha,s=\text{M}} + \mathbf{u}_a^{\alpha} + \mathbf{u}_a^{\alpha,s=\text{M}} + \mathbf{u}_x^{\alpha} + \mathbf{u}_x^{\alpha,s=\text{M}} + \mathbf{u}_x^{\alpha,a<15} + \boldsymbol{\eta}_{R_x,s,a}^{\alpha} \quad \text{(C.5)}$$

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

HIV incidence rate

HIV incidence rate $\lambda_{x,s,a} > 0$ was modelled on the log scale using the structured additive predictor

$$\log(\lambda_{x,s,a}) = \beta_0^{\lambda} + \beta_S^{\lambda,s=\text{M}} + \log(\rho_x^{\text{15-49}}) + \log(1 - \omega \cdot \alpha_x^{\text{15-49}}) + \mathbf{u}_x^{\lambda} + \boldsymbol{\eta}_{R_x,s,a}^{\lambda}. \quad \text{(C.6)}$$

Table C.3 provides a description of the terms included in Equation (C.6).

Table C.3: Each term in Equation (C.6) together with (where applicable) its prior distribution and a written description of its role.

Term	Distribution	Description
β_0^{λ}	$\mathcal{N}(0,5)$	Intercept term proportional to the average HIV transmission rate for untreated HIV positive adults
$\beta_S^{\lambda,s=\text{M}}$	$\mathcal{N}(0,5)$	The log incidence rate ratio for men compared to women
$\rho_x^{\text{15-49}}$	–	The HIV prevalence among adults 15-49 in district x calculated by aggregating age-specific HIV prevalences
$\alpha_x^{\text{15-49}}$	–	The ART coverage among adults 15-49 in district x 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
$\mathbf{u}_x^{\lambda}$	$\mathcal{N}(0,\sigma^{\lambda})$	IID spatial random effects with $\sigma^{\lambda} \sim \mathcal{N}^{+}(0,1)$
$\boldsymbol{\eta}_{R_x,s,a}^{\lambda}$	–	Fixed log incidence rate ratios by sex and age group calculated from Spectrum model outputs for region R_x

The proportion recently infected among HIV positive persons $\kappa_{x,s,a} \in [0,1]$ was modelled as

$$\kappa_{x,s,a} = 1 - \exp\left(-\lambda_{x,s,a} \cdot \frac{1-\rho_{x,s,a}}{\rho_{x,s,a}} \cdot (\Omega_T - \beta_T) - \beta_T\right), \quad \text{(C.7)}$$

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. The prior distribution for Ω_T was informed by the characteristics of the recent infection testing algorithm. For PHIA surveys this was $\Omega_{T_0} = 130$ days and $\sigma^{\Omega_T} = 6.12$ days. For PHIA surveys there was assumed to be no false recency, such that $\beta_{T_0} = 0.0$, $\sigma^{\beta_T} = 0.0$, and $\beta_T = 0$.

ANC testing

HIV prevalence $\rho^{\text{ANC}}_{x,a}$ and ART coverage $\alpha^{\text{ANC}}_{x,a}$ among pregnant women were 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

$$\text{logit}(\rho^{\text{ANC}}_{x,a}) = \text{logit}(\rho_{x,F,a}) + \beta^{\rho^{\text{ANC}}} + \mathbf{u}^{\rho^{\text{ANC}}}_{x} + \boldsymbol{\eta}^{\rho^{\text{ANC}}}_{R_x,a}\,, \tag{C.8}$$

$$\text{logit}(\alpha^{\text{ANC}}_{x,a}) = \text{logit}(\alpha_{x,F,a}) + \beta^{\alpha^{\text{ANC}}} + \mathbf{u}^{\alpha^{\text{ANC}}}_{x} + \boldsymbol{\eta}^{\alpha^{\text{ANC}}}_{R_x,a}\,. \tag{C.9}$$

Table C.4 provides a description of the terms included in Equation (C.8) and Equation (C.9).

Table C.4: Each term in Equations (C.8) and (C.9) together with (where applicable) its prior distribution and a written description of its role. The notation θ is used as stand in for $\theta \in \{\rho, \alpha\}$.

Term	Distribution	Description
$\beta^{\theta^{\text{ANC}}}$	$\mathcal{N}(0, 5)$	Intercept giving the average difference between population and ANC outcomes
$\mathbf{u}^{\theta^{\text{ANC}}}_{x}$	$\mathcal{N}(0, \sigma^{\theta^{\text{ANC}}}_{X})$	IID district random effects with $\sigma^{\theta^{\text{ANC}}}_{X} \sim \mathcal{N}^+(0, 1)$
$\boldsymbol{\eta}^{\theta^{\text{ANC}}}_{R_x,a}$	$-$	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

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

$$\log(\Psi_{x,a}) = \log(N_{x,\text{F},a}) + \psi_{R_x,a} + \beta^{\psi} + \mathbf{u}^{\psi}_{x}, \tag{C.10}$$

where $N_{x,\text{F},a}$ 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 $\mathbf{u}_x^{\psi} \sim \mathcal{N}(0, \sigma^{\psi})$ are district random effects. Here these terms are fixed to $\beta^{\psi} = 0$ and $\mathbf{u}_x^{\psi} = \mathbf{0}$ such that $\Psi_{x,a}$ are simply constants.

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'. Assume that $\gamma_{x,x'} = 0$ for $x \notin \{x, \text{ne}(x)\}$ such that individuals seek treatment only in their residing district or its neighbours $\text{ne}(x) = \{x' : x' \sim x\}$, where $\sim$ is an adjacency relation, and $\sum_{x' \in \{x, \text{ne}(x)\}} \gamma_{x,x'} = 1$.

The probabilities $\gamma_{x,x'}$ for $x \sim x'$ were modelled using multinomial logistic regression model, based on the log-odds ratios

$$\tilde{\gamma}_{x,x'} = \log\left(\frac{\gamma_{x,x'}}{1 - \gamma_{x,x'}}\right) = \tilde{\gamma}_0 + \mathbf{u}_x^{\tilde{\gamma}}. \tag{C.11}$$

Table C.5 provides a description of the terms included in Equation (C.11). Fixing $\tilde{\gamma}_{x,x} = 0$ then the multinomial probabilities may be recovered using the softmax

$$\gamma_{x,x'} = \frac{\exp(\tilde{\gamma}_{x,x'})}{\sum_{x^\star \in \{x, \text{ne}(x)\}} \exp(\tilde{\gamma}_{x,x^\star})}. \tag{C.12}$$

Table C.5: Each term in Equation (C.11) and (C.9) together with (where applicable) its prior distribution and a written description of its role. No terms include x', such that $\gamma_{x,x'}$ is only a function of x.

Term	Distribution	Description
$\tilde{\gamma}_0$	$-$	Fixed intercept $\tilde{\gamma}_0 = -4$. Implies a prior mean on $\gamma_{x,x'}$ of 1.8%, such that a-priori $(100 - 1.8 \times \text{ne}(x))\%$ of ART clients in district x obtain treatment in their home district
$\mathbf{u}_x^{\tilde{\gamma}}$	$\mathcal{N}(0, \sigma_X^{\tilde{\gamma}})$	District random effects, with $\sigma_X^{\tilde{\gamma}} \sim \mathcal{N}^+(0, 2.5)$

C.2.2 Additional likelihood specification

Though Section 6.3 provides a complete description of Naomi's likelihood specification, additional useful details are provided here.

Household survey data

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

$$\log p(y) = \log \Gamma(m+1) - \log \Gamma(y+1) - \log \Gamma(m-y+1) + y \log p + (m-y) \log(1-p), \tag{C.13}$$

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

ANC testing data

Put any additional details here.

ART attendence data

Put any additional details here.

C.2.3 Identifiability constraints

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 $\mathbf{u}^{\theta}_{a} = 0$ and $\mathbf{u}^{\theta}_{a,s=\text{M}} = 0$. In this case, the average age-sex pattern from the Spectrum is used. For the Malawi case-study (Section 6.5), HIV prevalence and ART coverage data are not available for those aged 65+. As a result, there are $|\{\text{0-4}, \ldots, \text{50-54}\}| = 13$ 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

$$\text{logit}\,(\alpha_{x,s,a}) = \mathbf{u}^{\alpha}_{x} + \boldsymbol{\eta}^{\alpha}_{R_x,s,a}. \tag{C.14}$$

3. If survey data on recent HIV infection are not included in the model, then $\beta_0^\lambda = \beta_S^{\lambda,s=\mathrm{M}} = 0$ and $\mathbf{u}_x^\lambda = \mathbf{0}$. The sex ratio for HIV incidence is determined by the sex incidence rate ratio from Spectrum, and the incidence rate in all districts is modelled assuming the same average HIV transmission rate for untreated adults, but varies according to district-level estimates of HIV prevalence and ART coverage.

C.2.4 Implementation

The `TMB` C++ code for the negative log-posterior of the simplified Naomi model is available from `https://github.com/athowes/naomi-aghq`. (It is not given here, as it is over 500 lines.) For ease of understanding, Table C.6 provides correspondence between the mathematical notation used in Section C.2 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).

Table C.6

Variable name	Notation	Type	Size	Domain	ρ input?	α input?	λ input?
`logit_phi_x`	$\log(\phi_X^\rho)$	Hyper	1	$\mathbb{R}$	Yes		
`log_sigma_x`	$\log(\sigma_X^\rho)$	Hyper	1	$\mathbb{R}$	Yes		
`logit_phi_xs`	$\log(\phi_{XS}^\rho)$	Hyper	1	$\mathbb{R}$	Yes		
`log_sigma_xs`	$\log(\sigma_{XS}^\rho)$	Hyper	1	$\mathbb{R}$	Yes		
`logit_phi_a`	$\log(\phi_A^\rho)$	Hyper	1	$\mathbb{R}$	Yes		
`log_sigma_a`	$\log(\sigma_A^\rho)$	Hyper	1	$\mathbb{R}$	Yes		
`logit_phi_as`	$\log(\phi_{AS}^\rho)$	Hyper	1	$\mathbb{R}$	Yes		
`log_sigma_as`	$\log(\sigma_{AS}^\rho)$	Hyper	1	$\mathbb{R}$	Yes		
`log_sigma_xa`	$\log(\sigma_{XA}^\rho)$	Hyper	1	$\mathbb{R}$	Yes		
`logit_phi_alpha_x`	$\mathrm{logit}(\phi_X^\alpha)$	Hyper	1	$\mathbb{R}$		Yes	
`log_sigma_alpha_x`	$\log(\sigma_X^\alpha)$	Hyper	1	$\mathbb{R}$		Yes	
`logit_phi_alpha_xs`	$\mathrm{logit}(\phi_{XS}^\alpha)$	Hyper	1	$\mathbb{R}$		Yes	
`log_sigma_alpha_xs`	$\log(\sigma_{XS}^\alpha)$	Hyper	1	$\mathbb{R}$		Yes	
`logit_phi_alpha_a`	$\mathrm{logit}(\phi_A^\alpha)$	Hyper	1	$\mathbb{R}$		Yes	
`log_sigma_alpha_a`	$\log(\sigma_A^\alpha)$	Hyper	1	$\mathbb{R}$		Yes	
`logit_phi_alpha_as`	$\mathrm{logit}(\phi_{AS}^\alpha)$	Hyper	1	$\mathbb{R}$		Yes	
`log_sigma_alpha_as`	$\log(\sigma_{AS}^\alpha)$	Hyper	1	$\mathbb{R}$		Yes	
`log_sigma_alpha_xa`	$\log(\sigma_{XA}^\alpha)$	Hyper	1	$\mathbb{R}$		Yes	
`OmegaT_raw`	Ω_T	Hyper	1	$\mathbb{R}$			Yes
`log_betaT`	$\log(\beta_T)$	Hyper	1	$\mathbb{R}$			Yes

Variable name	Notation	Type	Size	Domain	ρ input?	α input?	λ input?
log_sigma_lambda_x	$\log(\sigma_x^\lambda)$	Hyper	1	$\mathbb{R}$			Yes
log_sigma_anc_rho_x	$\log(\sigma_x^{\rho,\text{ANC}})$	Hyper	1	$\mathbb{R}$		Yes	
log_sigma_anc_alpha_x	$\log(\sigma_x^{\alpha,\text{ANC}})$	Hyper	1	$\mathbb{R}$		Yes	
log_sigma_log_gamma	$\log(\sigma_x^{\tilde{\gamma}})$	Hyper	1	$\mathbb{R}$			
beta_rho	$(\beta_0^\rho, \beta_s^{\rho,s=\text{M}})$	Latent	2	$\mathbb{R}^2$	Yes		
beta_alpha	$(\beta_0^\alpha, \beta_S^{\alpha,s=\text{M}})$	Latent	2	$\mathbb{R}^2$		Yes	
beta_lambda	$(\beta_0^\lambda, \beta_S^{\lambda,s=\text{M}})$	Latent	2	$\mathbb{R}^2$			Yes
beta_anc_rho	$\beta^{\rho,\text{ANC}}$	Latent	1	$\mathbb{R}$		Yes	
beta_anc_alpha	$\beta^{\alpha,\text{ANC}}$	Latent	1	$\mathbb{R}$		Yes	
u_rho_x	$\mathbf{w}_x^\rho$	Latent	n	$\mathbb{R}^n$	Yes		
us_rho_x	$\mathbf{v}_x^\rho$	Latent	n	$\mathbb{R}^n$	Yes		
u_rho_xs	$\mathbf{w}_x^{\rho,s=\text{M}}$	Latent	n	$\mathbb{R}^n$	Yes		
us_rho_xs	$\mathbf{v}_x^{\rho,s=\text{M}}$	Latent	n	$\mathbb{R}^n$	Yes		
u_rho_a	$\mathbf{u}_a^\rho$	Latent	10	$\mathbb{R}^{10}$	Yes		
u_rho_as	$\mathbf{u}_a^{\rho,s=\text{M}}$	Latent	10	$\mathbb{R}^{10}$	Yes		
u_rho_xa	$\mathbf{u}_x^{\rho,a<15}$	Latent	n	$\mathbb{R}^n$	Yes		
u_alpha_x	$\mathbf{w}_x^\alpha$	Latent	n	$\mathbb{R}^n$		Yes	
us_alpha_x	$\mathbf{v}_x^\alpha$	Latent	n	$\mathbb{R}^n$		Yes	
u_alpha_xs	$\mathbf{w}_x^{\alpha,s=\text{M}}$	Latent	n	$\mathbb{R}^n$		Yes	
us_alpha_xs	$\mathbf{v}_x^{\alpha,s=\text{M}}$	Latent	n	$\mathbb{R}^n$		Yes	
u_alpha_a	$\mathbf{u}_a^\alpha$	Latent	13	$\mathbb{R}^{13}$		Yes	
u_alpha_as	$\mathbf{u}_a^{\alpha,s=\text{M}}$	Latent	10	$\mathbb{R}^{10}$		Yes	
u_alpha_xa	$\mathbf{u}_x^{\alpha,a<15}$	Latent	n	$\mathbb{R}^n$		Yes	
ui_lambda_x	$\mathbf{u}_x^\lambda$	Latent	n	$\mathbb{R}^n$			Yes
ui_anc_rho_x	$\mathbf{u}_x^{\rho,\text{ANC}}$	Latent	n	$\mathbb{R}^n$		Yes	
ui_anc_alpha_x	$\mathbf{u}_x^{\alpha,\text{ANC}}$	Latent	n	$\mathbb{R}^n$		Yes	
log_or_gamma	$\mathbf{u}_x^{\tilde{\gamma}}$	Latent	n	$\mathbb{R}^n$			

C.3 Model assessment

C.4 AGHQ and PCA-AGHQ details

C.5 Normalising constant estimation

C.6 Inference comparison

C.7 MCMC convergence and suitability

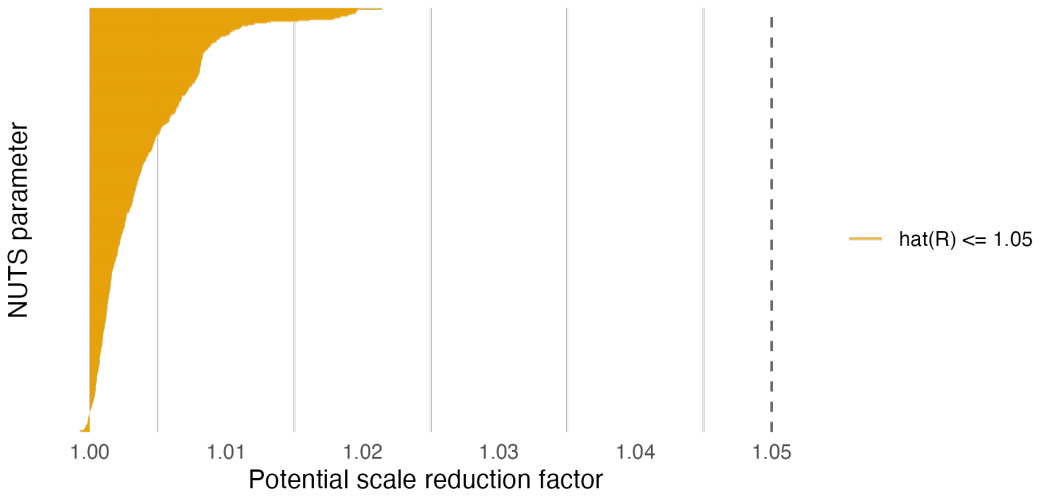

Figure C.3: The potential scale reduction factor compares between- and within- estimates of univariate parameters. It is recommended only to use NUTS results if the value is less than 1.05, which it is for all parameters.

Works Cited

Amoah, Benjamin, Peter J Diggle, and Emanuele Giorgi (2020). "A geostatistical framework for combining spatially referenced disease prevalence data from multiple diagnostics". In: *Biometrics* 76.1, pp. 158–170.

Auvert, Bertran et al. (2005). "Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial". In: *PLoS medicine* 2.11, e298.

Bachl, Fabian E et al. (2019). "inlabru: an R package for Bayesian spatial modelling from ecological survey data". In: *Methods in Ecology and Evolution* 10.6, pp. 760–766.

Bailey, Robert C et al. (2007). "Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial". In: *The Lancet* 369.9562, pp. 643–656.

Baker, Stuart G (1994). "The multinomial-Poisson transformation". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 43.4, pp. 495–504.

Baral, Stefan et al. (2012). "Burden of HIV among female sex workers in low-income and middle-income countries: a systematic review and meta-analysis". In: *The Lancet Infectious Diseases* 12.7, pp. 538–549.

Barré-Sinoussi, Françoise et al. (1983). "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)". In: *Science* 220.4599, pp. 868–871.

Baydin, Atılım Günes et al. (2017). "Automatic differentiation in machine learning: a survey". In: *The Journal of Machine Learning Research* 18.1, pp. 5595–5637.

Bell, Bradley (2023). *CppAD: a package for C++ algorithmic differentiation.* `http://www.coin-or.org/CppAD`. Accessed: September 25, 2023.

Berger, James (2006). "The case for objective Bayesian analysis". In.

Bernardo, José M and Adrian FM Smith (2001). *Bayesian theory.* John Wiley & Sons.

Besag, Julian, Jeremy York, and Annie Mollié (1991). "Bayesian image restoration, with two applications in spatial statistics". In: *Annals of the Institute of Statistical Mathematics* 43.1, pp. 1–20.

Best, Nicky, Sylvia Richardson, and Andrew Thomson (2005). "A comparison of Bayesian spatial models for disease mapping". In: *Statistical Methods in Medical Research* 14.1, pp. 35–59.

Bilodeau, Blair, Alex Stringer, and Yanbo Tang (2022). "Stochastic convergence rates and applications of adaptive quadrature in Bayesian inference". In: *Journal of the American Statistical Association*, pp. 1–11.

Bivand, Roger S et al. (2008). *Applied spatial data analysis with R.* Vol. 747248717. Springer.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518, pp. 859–877.

Bolker, Benjamin M et al. (2013). "Strategies for fitting nonlinear ecological models in R, AD M odel B uilder, and BUGS". In: *Methods in Ecology and Evolution* 4.6, pp. 501–512.

Bollhöfer, Matthias et al. (2020). "State-of-the-art sparse direct solvers". In: *Parallel algorithms in computational science and engineering*, pp. 3–33.

Bosse, Nikos I. et al. (2022). *Evaluating Forecasts with scoringutils in R*. DOI: 10.48550/ARXIV.2205.07090. URL: https://arxiv.org/abs/2205.07090.

Box, George EP and Kenneth B Wilson (1992). "On the experimental attainment of optimum conditions". In: *Breakthroughs in statistics: methodology and distribution*. Springer, pp. 270–310.

Breslow, Norman E and David G Clayton (1993). "Approximate inference in generalized linear mixed models". In: *Journal of the American statistical Association* 88.421, pp. 9–25.

Brooks, Mollie E et al. (2017). "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling". In: *The R journal* 9.2, pp. 378–400.

Broyles, Laura N et al. (2023). "The risk of sexual transmission of HIV in individuals with low-level HIV viraemia: a systematic review". In: *The Lancet*.

Brugh, Kristen N et al. (2021). "Characterizing and mapping the spatial variability of HIV risk among adolescent girls and young women: A cross-county analysis of population-based surveys in Eswatini, Haiti, and Mozambique". In: *PLoS One* 16.12, e0261520.

Carpenter, Bob et al. (2017). "Stan: A probabilistic programming language". In: *Journal of Statistical Software* 76.1.

Casella, George (1985). "An introduction to empirical Bayes data analysis". In: *The American Statistician* 39.2, pp. 83–87.

Chen, Cici, Jon Wakefield, and Thomas Lumely (2014). "The use of sampling weights in Bayesian hierarchical models for small area estimation". In: *Spatial and spatio-temporal epidemiology* 11, pp. 33–43.

Chopin, Nicolas, Omiros Papaspiliopoulos, et al. (2020). *An introduction to sequential Monte Carlo*. Vol. 4. Springer.

Cleland, John et al. (2004). "Monitoring sexual behaviour in general populations: a synthesis of lessons of the past decade". In: *Sexually Transmitted Infections* 80.suppl 2, pp. ii1–ii7.

Cohen, Myron S et al. (2011). "Prevention of HIV-1 infection with early antiretroviral therapy". In: *New England journal of medicine* 365.6, pp. 493–505.

Cramb, SM et al. (2018). "Investigation of Bayesian spatial models". In.

Cressie, Noel and Christopher K Wikle (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Csárdi, Gábor (2023). *cranlogs: Download Logs from the 'RStudio' 'CRAN' Mirror*. https://github.com/r-hub/cranlogs, https://r-hub.github.io/cranlogs.

Davis, Philip J and Philip Rabinowitz (1975). *Methods of numerical integration*. Academic Press.

Dawid, A Philip (1984). "Present position and potential developments: Some personal views statistical theory the prequential approach". In: *Journal of the Royal Statistical Society: Series A (General)* 147.2, pp. 278–290.

de Valpine, Perry et al. (2023). *NIMBLE User Manual*. Version 1.0.1. R package manual version 1.0.1. DOI: 10.5281/zenodo.1211190. URL: https://r-nimble.org.

Dean, CB, MD Ugarte, and AF Militino (2001). "Detecting interaction between random region and fixed age effects in disease mapping". In: *Biometrics* 57.1, pp. 197–202.

Dennis Jr, John E, David M Gay, and Roy E Walsh (1981). "An adaptive nonlinear least-squares algorithm". In: *ACM Transactions on Mathematical Software (TOMS)* 7.3, pp. 348–368.

De Valpine, Perry et al. (2017). "Programming with models: writing statistical algorithms for general model structures with NIMBLE". In: *Journal of Computational and Graphical Statistics* 26.2, pp. 403–413.

DHS (2012). *Sampling and Household Listing Manual: Demographic and Health Surveys Methodology*.

Diaz, Jose Monsalve et al. (2018). "Openmp 4.5 validation and verification suite for device offload". In: *Evolving OpenMP for Evolving Architectures: 14th International Workshop on OpenMP, IWOMP 2018, Barcelona, Spain, September 26–28, 2018, Proceedings 14*. Springer, pp. 82–95.

Diggle, Peter J et al. (2013). "Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm". In: *Statistical Science* 28.4, pp. 542–563.

Duane, Simon et al. (1987). "Hybrid Monte Carlo". In: *Physics letters B* 195.2, pp. 216–222.

Dwyer-Lindgren, Laura, Michael A Cork, et al. (2019). "Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017". In: *Nature* 570.7760, pp. 189–193.

Dwyer-Lindgren, Laura, Abraham D Flaxman, et al. (2015). "Drinking patterns in US counties from 2002 to 2012". In: *American Journal of Public Health* 105.6, pp. 1120–1127.

Eaton, Jeffrey W et al. (2021). "Naomi: A New Modelling Tool for Estimating HIV Epidemic Indicators at the District Level in Sub-Saharan Africa". In.

Economist Impact (2023). "A triple dividend: the health, social and economic gains from financing the HIV response in Africa". In.

Esra, Rachel (2023). "Improved indicators for subnational unmet antiretroviral therapy need in the health system: updates to the Naomi model in 2023". In.

Fattah, Esmail Abdul, Janet Van Niekerk, and Håvard Rue (2021). "Smart Gradient–An Adaptive Technique for Improving Gradient Estimation". In: *arXiv preprint arXiv:2106.07313*.

Fisher, Ronald Aylmer (1936). "Design of experiments". In: *British Medical Journal* 1.3923, p. 554.

Follestad, Turid and Håvard Rue (2003). *Modelling spatial variation in disease risk using Gaussian Markov random field proxies for Gaussian random fields*. Tech. rep. SIS-2003-305.

Fournier, David A et al. (2012). "AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models". In: *Optimization Methods and Software* 27.2, pp. 233–249.

Freni-Sterrantino, Anna, Massimo Ventrucci, and Håvard Rue (2018). "A note on intrinsic conditional autoregressive models for disconnected graphs". In: *Spatial and spatio-temporal epidemiology* 26, pp. 25–34.

Gaedke-Merzhäuser, Lisa et al. (2023). "Parallelized integrated nested Laplace approximations for fast Bayesian inference". In: *Statistics and Computing* 33.1, p. 25.

Garnier et al. (2023). *viridis(Lite) - Colorblind-Friendly Color Maps for R*. viridis package version 0.6.4. DOI: 10.5281/zenodo.4679423. URL: https://sjmgarnier.github.io/viridis/.

Gelfand, Alan E, Li Zhu, and Bradley P Carlin (2001). "On the change of support problem for spatio-temporal data". In: *Biostatistics* 2.1, pp. 31–45.

Gelman, Andrew (2005). "Analysis of variance—why it is more important than ever". In.

— (2007). "Struggles with survey weighting and regression modeling". In.

Gelman, Andrew, John B Carlin, et al. (2013). *Bayesian data analysis*. CRC press.

Gelman, Andrew and Donald B Rubin (1992). "Inference from iterative simulation using multiple sequences". In: *Statistical science*, pp. 457–472.

Gelman, Andrew, Daniel Simpson, and Michael Betancourt (2017). "The prior can often only be understood in the context of the likelihood". In: *Entropy* 19.10, p. 555.

Gelman, Andrew, Aki Vehtari, et al. (2020). "Bayesian workflow". In: *arXiv preprint arXiv:2011.01808*.

Geman, Stuart and Donald Geman (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on pattern analysis and machine intelligence* 6, pp. 721–741.

Giordano, Ryan, Tamara Broderick, and Michael I. Jordan (2018). "Covariances, Robustness, and Variational Bayes". In: *Journal of Machine Learning Research* 19.51, pp. 1–49. URL: `http://jmlr.org/papers/v19/17-670.html`.

Global Burden of Disease Collaborative Network (2019). *Global Burden of Disease Study 2019 (GBD 2019) Results.* URL: `https://vizhub.healthdata.org/gbd-results/`.

Glynn, Judith R et al. (2011). "Assessing the validity of sexual behaviour reports in a whole population survey in rural Malawi". In: *PLoS One* 6.7, e22840.

Goldstein, Michael (2006). "Subjective Bayesian analysis: principles and practice". In.

Gómez-Rubio, Virgilio (2020). *Bayesian inference with INLA*. CRC Press.

Gössl, Christoff, Dorothee P Auer, and Ludwig Fahrmeir (2001). "Bayesian spatiotemporal inference in functional magnetic resonance imaging". In: *Biometrics* 57.2, pp. 554–562.

Gottlieb, Michael S et al. (1981). "Pneumocystis pneumonia—Los Angeles". In: *Mmwr* 30.21, pp. 1–3.

Grabowski, M Kate et al. (2017). "HIV prevention efforts and incidence of HIV in Uganda". In: *New England Journal of Medicine* 377.22, pp. 2154–2166.

Gray, Ronald H et al. (2007). "Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial". In: *The Lancet* 369.9562, pp. 657–666.

Gregson, Simon et al. (2006). "HIV decline associated with behavior change in eastern Zimbabwe". In: *Science* 311.5761, pp. 664–666.

Haining, Robert P (2003). *Spatial data analysis: theory and practice*. Cambridge university press.

Hájek, Jaroslav (1971). "Discussion of 'An essay on the logical foundations of survey sampling, part I'". In: *Foundations of Statistical Inference (Proc. Sympos., Univ. Waterloo, Ontario, 1970)*, p. 236.

Hamelijnck, O et al. (2019). "Multi-resolution multi-task Gaussian processes". In: *Advances in Neural Information Processing Systems* 32.

Hastie, Trevor and Robert Tibshirani (1987). "Generalized additive models: some applications". In: *Journal of the American Statistical Association* 82.398, pp. 371–386.

Hastings, W Keith (1970). "Monte Carlo sampling methods using Markov chains and their applications". In.

Helleringer, Stéphane et al. (2011). "The reliability of sexual partnership histories: implications for the measurement of partnership concurrency during surveys". In: *AIDS (London, England)* 25.4, p. 503.

Hodgins, Caroline et al. (2022). "Population sizes, HIV prevalence, and HIV prevention among men who paid for sex in sub-Saharan Africa (2000–2020): A meta-analysis of 87 population-based surveys". In: *PLoS Medicine* 19.1, e1003861.

Hoffman, Matthew D, Andrew Gelman, et al. (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *J. Mach. Learn. Res.* 15.1, pp. 1593–1623.

Howes, Adam, Jeffrey W. Eaton, and Seth R. Flaxman (2023+). "Beyond borders: evaluating the suitability of spatial adjacency for small-area estimation". In.

Howes, Adam, Kathryn A. Risher, et al. (Apr. 2023). "Spatio-temporal estimates of HIV risk group proportions for adolescent girls and young women across 13 priority countries in sub-Saharan Africa". In: *PLOS Global Public Health* 3.4, pp. 1–14. DOI: 10.1371/journal.pgph.0001731. URL: https://doi.org/10.1371/journal.pgph.0001731.

Howes, Adam, Alex Stringer, et al. (2023+). "Fast approximate Bayesian inference of HIV indicators using PCA adaptive Gauss-Hermite quadrature". In.

Jäckel, Peter (2005). "A note on multivariate Gauss-Hermite quadrature". In: *London: ABN-Amro. Re.*

Jia, Katherine M et al. (2022). "Risk scores for predicting HIV incidence among adult heterosexual populations in sub-Saharan Africa: a systematic review and meta-analysis". In: *Journal of the International AIDS Society* 25.1, e25861.

Johnson, L and RE Dorrington (2020). "Thembisa version 4.3: A model for evaluating the impact of HIV/AIDS in South Africa". In: *View Article.*

Johnson, Olatunji, Peter Diggle, and Emanuele Giorgi (2019). "A spatially discrete approximation to log-Gaussian Cox processes for modelling aggregated disease count data". In: *Statistics in Medicine* 38.24, pp. 4871–4887.

Kelsall, Julia and Jonathan Wakefield (2002). "Modeling spatial variation in disease risk: a geostatistical approach". In: *Journal of the American Statistical Association* 97.459, pp. 692–701.

Khoury, Muin J, Michael F Iademarco, and William T Riley (2016). "Precision public health for the era of precision medicine". In: *American journal of preventive medicine* 50.3, pp. 398–401.

Kish, Leslie (1965). *Survey sampling.* 04; HN29, K5.

Knorr-Held, Leonhard (2000). "Bayesian modelling of inseparable space-time variation in disease risk". In: *Statistics in medicine* 19.17-18, pp. 2555–2567.

Kristensen, Kasper (2021). *The comprehensive TMB documentation.* https://kaskr.github.io/adcomp/_book/Introduction.html. Accessed: June 2, 2023.

Kristensen, Kasper et al. (2016). "TMB: Automatic Differentiation and Laplace Approximation". In: *Journal of Statistical Software* 70.i05.

Laplace, P. S. (1774). "Memoire sur la probabilite de causes par les evenements". In: *Memoire de l'Academie Royale des Sciences.*

Law, Ho Chung et al. (2018). "Variational learning on aggregate outputs with Gaussian processes". In: *Advances in Neural Information Processing Systems* 31.

Lenth, Russell (2009). "Response-Surface Methods in R, Using rsm". In: *Journal of Statistical Software* 32.7, pp. 1–17. DOI: 10.18637/jss.v032.i07.

Leppik, IE et al. (1985). "A double-blind crossover evaluation of progabide in partial seizures". In: *Neurology* 35.4, p. 285.

Leroux, Brian G, Xingye Lei, and Norman Breslow (2000). "Estimation of disease rates in small areas: a new mixed model for spatial dependence". In: *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. Springer, pp. 179–191.

Lindgren, Finn, Håvard Rue, and Johan Lindström (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 73.4, pp. 423–498.

Margossian, Charles et al. (2020). "Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for latent Gaussian models and beyond". In: *Advances in Neural Information Processing Systems* 33, pp. 9086–9097.

Martin, Gael M, David T Frazier, and Christian P Robert (2023). "Computing Bayes: From then 'til now". In: *Statistical Science* 1.1, pp. 1–17.

Martino, Sara and Andrea Riebler (2019). "Integrated nested Laplace approximations (INLA)". In: *arXiv preprint arXiv:1907.01248*.

Martino, Sara and Håvard Rue (2009). "Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the inla program". In: *Department of Mathematical Sciences, NTNU, Norway*.

Martins, Thiago G et al. (2013). "Bayesian computing with INLA: new features". In: *Computational Statistics & Data Analysis* 67, pp. 68–83.

"Maximum likelihood from incomplete data via the EM algorithm" (1977). In: *Journal of the royal statistical society: series B (methodological)* 39.1, pp. 1–22.

McCullagh, Peter and John A Nelder (1989). *Generalized linear models*. Routledge.

McElreath, Richard (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Meng, Xiao-Li (2018). "Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election". In: *The Annals of Applied Statistics* 12.2, pp. 685–726.

Metropolis, Nicholas et al. (1953). "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6, pp. 1087–1092.

Minka, Thomas P (2001). "Expectation Propagation for approximate Bayesian inference". In: *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pp. 362–369.

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". In: *PloS one* 13.5, e0197954.

Monod, Mélodie et al. (2023). "Growing gender disparity in HIV infection in Africa: sources and policy implications". In: *medRxiv*, pp. 2023–03.

Nandi, Anita K et al. (2023). "disaggregation: An R Package for Bayesian Spatial Disaggregation Modeling". In: *Journal of Statistical Software* 106, pp. 1–19.

Naylor, John C and Adrian FM Smith (1982). "Applications of a method for the efficient computation of posterior distributions". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 31.3, pp. 214–225.

Neal, Radford M (2003). "Slice sampling". In: *The Annals of Statistics* 31.3, pp. 705–767.

Neal, Radford M et al. (2011). "MCMC using Hamiltonian dynamics". In: *Handbook of Markov chain Monte Carlo* 2.11, p. 2.

Nguyen, Van Kính and Jeffrey W. Eaton (2022). "Trends and country-level variation in age at first sex in sub-Saharan Africa among birth cohorts entering adulthood

between 1985 and 2020". In: *BMC Public Health* 22.1, p. 1120. DOI: 10.1186/s12889-022-13451-y. URL: https://doi.org/10.1186/s12889-022-13451-y.

Nnko, Soori et al. (2004). "Secretive females or swaggering males?: An assessment of the quality of sexual partnership reporting in rural Tanzania". In: *Social Science & Medicine* 59.2, pp. 299–310.

Openshaw, S and P.J. Taylor (1979). "A million or so correlation coefficients, three experiments on the modifiable areal unit problem". In: *Statistical Applications in the Spatial Science*, pp. 127–144.

Ord, Toby (2013). "The moral imperative toward cost-effectiveness in global health". In: *Center for Global Development* 12.

Osgood-Zimmerman, Aaron and Jon Wakefield (2023). "A Statistical Review of Template Model Builder: A Flexible Tool for Spatial Modelling". In: *International Statistical Review* 91.2, pp. 318–342.

Paciorek, Christopher J et al. (2013). "Spatial models for point and areal data using Markov random fields on a fine grid". In: *Electronic Journal of Statistics* 7, pp. 946–972.

Pettit, LI (1990). "The conditional predictive ordinate for the normal distribution". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 52.1, pp. 175–184.

Pfeffermann, Danny et al. (2013). "New Important Developments in Small Area Estimation". In: *Statistical Science* 28.1, pp. 40–68.

Pisani, Elizabeth et al. (2003). "HIV surveillance: a global perspective". In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 32, S3–S11.

Porcu, Emilio, Reinhard Furrer, and Douglas Nychka (2021). "30 Years of space–time covariance functions". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 13.2, e1512.

Press, William H et al. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org.

Risher, Kathryn A et al. (2021). "Age patterns of HIV incidence in eastern and southern Africa: a modelling analysis of observational population-based cohort studies". In: *The Lancet HIV* 8.7, e429–e439.

Robert, Christian P and George Casella (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*.

Roberts, Gareth O and Jeffrey S Rosenthal (2004). "General state space Markov chains and MCMC algorithms". In.

Roy, Vivekananda (2020). "Convergence diagnostics for markov chain monte carlo". In: *Annual Review of Statistics and Its Application* 7, pp. 387–412.

Rue, Håvard (2001). "Fast sampling of Gaussian Markov random fields". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2, pp. 325–338.

— (2020). "Comment on R-INLA Discussion Group thread". In.

Rue, Havard (2023). "'R-INLA' Project - FAQ". Accessed 23/01/2023. URL: https://www.r-inla.org/faq.

Rue, Havard and Leonhard Held (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Rue, Håvard and Turid Follestad (2001). *GMRFLib: a C-library for fast and exact simulation of Gaussian Markov random fields*. Tech. rep. SIS-2002-236.

Rue, Håvard and Sara Martino (2007). "Approximate Bayesian inference for hierarchical Gaussian Markov random field models". In: *Journal of Statistical Planning and Inference* 137.10, pp. 3177–3192.

Rue, Håvard, Sara Martino, and Nicolas Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, pp. 319–392.

Rue, Håvard, Andrea Riebler, et al. (2017). "Bayesian computing with INLA: a review". In: *Annual Review of Statistics and Its Application* 4, pp. 395–421.

Säilynoja, Teemu, Paul-Christian Bürkner, and Aki Vehtari (2021). "Graphical Test for Discrete Uniformity and its Applications in Goodness of Fit Evaluation and Multiple Sample Comparison". In: *arXiv preprint arXiv:2103.10522*.

Saracco, James F et al. (2010). "Modeling spatial variation in avian survival and residency probabilities". In: *Ecology* 91.7, pp. 1885–1891.

Saul, Janet et al. (2018). "The DREAMS core package of interventions: a comprehensive approach to preventing HIV among adolescent girls and young women". In: *PLoS One* 13.12, e0208167.

Schmid, Volker J et al. (2006). "Bayesian methods for pharmacokinetic models in dynamic contrast-enhanced magnetic resonance imaging". In: *IEEE Transactions on Medical Imaging* 25.12, pp. 1627–1636.

Shapley, Lloyd S et al. (1953). "A value for n-person games". In.

Shumway, Robert H and David S Stoffer (2017). "Time Series Analysis and Its Applications With R Examples". In.

Siegfried, Nandi et al. (2011). "Antiretrovirals for reducing the risk of mother-to-child transmission of HIV infection". In: *Cochrane database of systematic reviews* 7.

Simpson, Daniel et al. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors". In: *Statistical Science* 32.1, pp. 1–28.

Sisson, Scott A, Yanan Fan, and Mark Beaumont (2018). *Handbook of approximate Bayesian computation*. CRC Press.

Skaug, Hans J. (2009). "Discussion of "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations"". In: vol. 71. 2. Wiley Online Library, pp. 319–392.

Slaymaker, Emma et al. (2020). "Risk factors for new HIV infections in the general population in sub-Saharan Africa". In.

Smith, Nathaniel and Stéfan van der Walt (2015). "A Better Default Colormap for Matplotlib". In: *Proceedings of the 14th Python in Science Conference (SciPy)*.

Sørbye, Sigrunn Holbek and Håvard Rue (2014). "Scaling intrinsic Gaussian Markov random field priors in spatial modelling". In: *Spatial Statistics* 8, pp. 39–51.

— (2017). "Penalised complexity priors for stationary autoregressive processes". In: *Journal of Time Series Analysis* 38.6, pp. 923–935.

Spiegelhalter, David, Andrew Thomas, et al. (1996). "BUGS 0.5 Examples". In: *MRC Biostatistics Unit, Institute of Public health, Cambridge, UK* 256.

Spiegelhalter, David J, Nicola G Best, et al. (2002). "Bayesian measures of model complexity and fit". In: *Journal of the Royal Statistical Society: Series B (statistical Methodology)* 64.4, pp. 583–639.

Stevens, Oliver, Keith Sabin, Sonia Arias Garcia, et al. (2022). "Estimating key population size, HIV prevalence, and ART coverage for sub-Saharan Africa at the national level". In.

Stevens, Oliver, Keith Sabin, Sonia Arias Garcia, et al. (2022). "Key population size, HIV prevalence, and ART coverage in sub-Saharan Africa: systematic collation and synthesis of survey data". In: *medRxiv*, pp. 2022–07.

Stover, John, Robert Glaubius, et al. (2019). "Updates to the Spectrum/AIM model for estimating key HIV indicators at national and subnational levels". In: *AIDS (London, England)* 33.Suppl 3, S227.

Stover, John and Yu Teng (2021). "The impact of condom use on the HIV epidemic". In: *Gates Open Research* 5.

Stringer, Alex (2021a). "Implementing Approximate Bayesian Inference Using Adaptive Quadrature". Statistics Graduate Student Research Day 2021, The Fields Institute for Research in Mathematical Sciences. URL: `http://www.fields.utoronto.ca/talks/Implementing-Approximate-Bayesian-Inference-Using-Adaptive-Quadrature`.

— (2021b). "Implementing Approximate Bayesian Inference using Adaptive Quadrature: the aghq Package". In: *arXiv preprint arXiv:2101.04468*.

Stringer, Alex, Patrick Brown, and Jamie Stafford (2022). "Fast, scalable approximations to posterior distributions in extended latent Gaussian models". In: *Journal of Computational and Graphical Statistics*, pp. 1–15.

Štrumbelj, Erik et al. (2023). "Past, Present, and Future of Software for Bayesian Inference". In.

Tanaka, Yusuke et al. (2019). "Spatially aggregated Gaussian processes with multivariate areal outputs". In: *Advances in Neural Information Processing Systems*, pp. 3005–3015.

Tanser, Frank et al. (2014). "Concentrated HIV sub-epidemics in generalized epidemic settings". In: *Current Opinion in HIV and AIDS* 9.2, p. 115.

Tatem, Andrew J (2017). "WorldPop, open data for spatial demography". In: *Scientific data* 4.1, pp. 1–4.

Teh, Yee Whye et al. (2021). "Efficient Bayesian Inference of Instantaneous Re-production Numbers at Fine Spatial Scales, with an Application to Mapping and Nowcasting the Covid-19 Epidemic in British Local Authorities". In: *URL https://rss. org. uk/RSS/media/File-library/News/2021/WhyeBhoopchand. pdfhttps://localcovid. info/2*.

Thall, Peter F and Stephen C Vail (1990). "Some covariance models for longitudinal count data with overdispersion". In: *Biometrics*, pp. 657–671.

The Global Fund (2018). *The Global Fund Measurement Framework for Adolescent Girls and Young Women Programs*. Accessed 30/08/2021. URL: `https://www.theglobalfund.org/media/8076/me_adolescentsgirlsandyoungwomenprograms_frameworkmeasurement_en.pdf`.

Tierney, Luke and Joseph B Kadane (1986). "Accurate approximations for posterior moments and marginal densities". In: *Journal of the American Statistical Association* 81.393, pp. 82–86.

Tobler, Waldo R (1970). "A computer movie simulating urban growth in the Detroit region". In: *Economic geography* 46.sup1, pp. 234–240.

Tokdar, Surya T and Robert E Kass (2010). "Importance sampling: a review". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.1, pp. 54–60.

U.S. Department of State (2022). *Latest Global Program Results*. https://www.state.gov/wp-content/uploads/2022/11/PEPFAR-Latest-Global-Results_December-2022.pdf. Accessed: 10/08/2023.

UNAIDS (2021a). *2021 UNAIDS Global AIDS Update - Confronting Inequalities - Lessons for pandemic responses from 40 Years of AIDS*. Accessed: June 2023.

— (2021b). "Global AIDS strategy 2021–2026. End inequalities. End AIDS". In: Accessed: June 2023.

— (2022). *In Danger: UNAIDS Global AIDS Update 2022*. https://www.unaids.org/en/resources/documents/2022/in-danger-global-aids-update. Accessed: June 2023.

— (2023a). *AIDSinfo: Global data on HIV epidemiology and response*. https://aidsinfo.unaids.org/. Accessed: August 2023.

— (2023b). *The path that ends AIDS: UNAIDS Global AIDS Update 2023*. https://www.unaids.org/en/resources/documents/2023/global-aids-update-2023. Accessed: August 2023.

UNICEF (2019). *Adolescent & social norms situation in Mozambique*. Accessed 25/03/2022. URL: https://www.unicef.org/mozambique/en/adolescent-social-norms.

Utazi, C Edson et al. (2019). "A spatial regression model for the disaggregation of areal unit based data to high-resolution grids with application to vaccination coverage mapping". In: *Statistical Methods in Medical Research* 28.10-11, pp. 3226–3241.

Van Niekerk, Janet et al. (2023). "A new avenue for Bayesian inference with INLA". In: *Computational Statistics & Data Analysis* 181, p. 107692.

Vehtari, Aki et al. (2021). "Rank-normalization, folding, and localization: an improved R for assessing convergence of MCMC (with discussion)". In: *Bayesian analysis* 16.2, pp. 667–718.

Wakefield, J and S Morris (1999). "Spatial dependence and errors-in-variables in environmental epidemiology". In: *Bayesian statistics* 6, pp. 657–684.

Wakefield, Jonathan and Hilary Lyons (Mar. 2010). "Spatial Aggregation and the Ecological Fallacy". In: vol. 2010, pp. 541–558. DOI: 10.1201/9781420072884-c30.

Ward, Brian (2023). *bridgestan: BridgeStan, Accessing Stan Model Functions in R*. R package version 1.0.1.

Watanabe, Sumio (2013). "A widely applicable Bayesian information criterion". In: *Journal of Machine Learning Research* 14.Mar, pp. 867–897.

Weiser, Constantin (2016). *mvQuad: Methods for Multivariate Quadrature.* (R package version 1.0-6). URL: http://CRAN.R-project.org/package=mvQuad.

Weiss, Daniel J et al. (2015). "Re-examining environmental correlates of Plasmodium falciparum malaria endemicity: a data-intensive variable selection approach". In: *Malaria journal* 14.1, pp. 1–18.

Wolock, Timothy M et al. (June 2021). "Evaluating distributional regression strategies for modelling self-reported sexual age-mixing". In: *eLife* 10. Ed. by Eduardo Franco, Talía Malagón, and Adam Akullian, e68318. DOI: 10.7554/eLife.68318. URL: https://doi.org/10.7554/eLife.68318.

Wood, Simon N (2017). *Generalized additive models: an introduction with R*. CRC press.

— (2020). "Simplified integrated nested Laplace approximation". In: *Biometrika* 107.1, pp. 223–230.

Wringe, A et al. (2009). "Comparative assessment of the quality of age-at-event reporting in three HIV cohort studies in sub-Saharan Africa". In: *Sexually transmitted infections* 85.Suppl 1, pp. i56–i63.

Yao, Yuling et al. (2018). "Yes, but did it work?: Evaluating variational inference". In: *International Conference on Machine Learning*. PMLR, pp. 5581–5590.

Yousefi, Fariba, Michael T Smith, and Mauricio Alvarez (2019). "Multi-task learning for aggregated data using Gaussian processes". In: *Advances in Neural Information Processing Systems* 32.

Zaba, Basia et al. (2004). "Age at first sex: understanding recent trends in African demographic surveys". In: *Sexually transmitted infections* 80.suppl 2, pp. ii28–ii35.