

# **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

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

# 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

I would first like to express my gratitude to Seth Flaxman and Jeff Imai-Eaton for their mentorship. Their guidance has been crucial in shaping this thesis, and my development as a scientist. Thanks to the HIV Inference Group at Imperial for exposing me to impact driven research, helping me to learn to present my work, and tolerating a statistician. I am grateful to have been a part of the Modern Statistics and Statistical Machine Learning Centre for Doctoral Training at Imperial and Oxford, and the Machine Learning and Global Health Network. Thanks to Antoine, Chris, Enrico, Phil, Yanni, Tim, Liza, and Theo for conversations, some of which were about research. This work was made possible by funding provided by the EPSRC and Bill & Melinda Gates Foundation. There are many worse ways to spend billions of dollars than fighting poverty and disease.

Thanks to Mike McLaren, Kevin Esvelt, the Nucleic Acid Observatory team, and the Sculpting Evolution lab for hosting my visit to the MIT Media Lab. I left Cambridge with appropriately raised aspirations, Google document templates, and only a little terrified about the future. Thanks to Trenton, Lenni, Lenny, Geetha, Janika, Simon, Phil, Frances, Leilani and Tammy.

Thanks 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 not have been possible, and I'd still be waiting Markov chains began in Chapter 4 to converge. Tim Lucas and Patrick Brown for put me in touch with Alex, and Håvard Rue and Finn Lindgren gave helpful answers on the R-INLA discussion group. Thanks also to Kate, my tour guide in Waterloo, and Midtown Yoga for helping me stay balanced.

My sense for what matters has been shaped, and arguably improved, by the Effective Altruism community. Thank you to the Meridian, Trajan, and LEAH offices for hosting me this final year. Thanks to my housemates in Hackney: August, Dewi, Henry, Jerome, Johnny, and Tamara. Not to be all Bay area, but I'm proud of the community we've built. Pinar believed in me and my research at times when I didn't. Thanks to Mr Sam, and attendees of the Manshead grit salt, for conferring upon me the status of stats man. No thanks to Simon Marshall, he didn't help, if anything he held me back. I extend my deepest thanks to my parents, Deborah and Karl, and my grandparents, Kath and Tony, whose love and support have granted me the privilege to pursue my interests.

# 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 used scoring rules to compare models for area-level spatial structure with both simulated and real data. Second, I estimated district-level HIV risk group proportions, enabling behavioural prioritisation of prevention services, as put forward in the UNAIDS 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 use of this algorithm 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

<b>List of Figures</b>	<b>x</b>
<b>List of Tables</b>	<b>xxviii</b>
<b>List of Abbreviations</b>	<b>xxxi</b>
<b>List of Notations</b>	<b>xxxiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Chapter overview . . . . .	2
<b>2 The HIV/AIDS epidemic</b>	<b>4</b>
2.1 Background . . . . .	4
2.2 HIV surveillance . . . . .	9
<b>3 Bayesian spatio-temporal statistics</b>	<b>15</b>
3.1 Bayesian statistics . . . . .	15
3.2 Spatio-temporal statistics . . . . .	22
3.3 Model structure . . . . .	27
3.4 Model comparison . . . . .	31
3.5 Survey methods . . . . .	34
<b>4 Models for areal spatial structure</b>	<b>38</b>
4.1 Models based on adjacency . . . . .	39
4.2 Models using kernels . . . . .	48
4.3 Simulation study . . . . .	53
4.4 HIV prevalence study . . . . .	60
4.5 Discussion . . . . .	66

## *Contents*

<b>5 A model for risk group proportions</b>	<b>69</b>
5.1 Background . . . . .	69
5.2 Data . . . . .	71
5.3 Model for risk group proportions . . . . .	74
5.4 Prevalence and incidence by risk group . . . . .	86
5.5 Results . . . . .	88
5.6 Discussion . . . . .	94
<b>6 Fast approximate Bayesian inference</b>	<b>100</b>
6.1 Inference methods and software . . . . .	102
6.2 A universal INLA implementation . . . . .	120
6.3 The Naomi model . . . . .	138
6.4 AGHQ in moderate dimensions . . . . .	149
6.5 Malawi case-study . . . . .	152
6.6 Discussion . . . . .	162
<b>7 Conclusions</b>	<b>170</b>
7.1 Contributions . . . . .	170
7.2 Future work . . . . .	172
7.3 Broader reflections . . . . .	174
<b>Appendices</b>	
<b>A Models for areal spatial structure</b>	<b>178</b>
A.1 Comparison of AGHQ to NUTS . . . . .	178
A.2 Lengthscale prior sensitivity . . . . .	183
A.3 Simulation study . . . . .	184
A.4 HIV study . . . . .	186
<b>B A model for risk group proportions</b>	<b>223</b>
B.1 The Global AIDS Strategy . . . . .	223
B.2 Household survey data . . . . .	224
B.3 Spatial analysis levels . . . . .	226
B.4 Survey questions and risk group allocation . . . . .	227
B.5 Additional figures . . . . .	228

*Contents*

<b>C Fast approximate Bayesian inference</b>	<b>230</b>
C.1 Epilepsy example . . . . .	230
C.2 Loa loa example . . . . .	236
C.3 AGHQ with Laplace marginals algorithm . . . . .	237
C.4 Simplified Naomi model description . . . . .	239
C.5 NUTS convergence and suitability . . . . .	247
C.6 Use of PCA-AGHQ . . . . .	247
C.7 Inference comparison . . . . .	247
<b>Works cited</b>	<b>257</b>

# List of Figures

## *List of Figures*

## *List of Figures*

## *List of Figures*

4.1	Panel A shows the districts of Zimbabwe. Panel B shows the corresponding adjacency graph $\mathcal{G}$ with vertices positioned at the centre of the area they correspond to, and edges between adjacent areas. . . . .	40
4.2	Though they are quite different, the geometries shown in panels A, B, C, and D each have the same adjacency graph. Therefore, each geometries would have the same distribution under the Besag model. . . . .	43
4.3	A sequence of geometries where the number of neighbours of area one grows by one at each iteration, as the shaded area is split into more areas. In the limit, the precision of the spatial random effect in the first area tends to infinity. This is not reasonable behaviour if the amount of information being shared is not also increasing. . . . .	45
4.4	Each of the shaded areas in the geometry in Panel A are split into two in Panel B. . . . .	45
4.5	The $n = 33$ districts of Malawi. Panel A shows the centroids as in Section 4.2.1. Panel B shows $L_i = 10$ randomly chosen points, Panel C hexagonal points, and Panel D grid points in each area, each generated using the <code>sf::st_sample</code> function (E. Pebesma 2018). . . . .	51
4.6	Seven geometries were considered in the simulation study. These were the four geometries from Figure 4.2 shown in Panel A, B, C and D, and three more realistic geometries shown in Panel E, F and G. . . . .	54
4.7	The mean CRPS and its standard error for each inferential model and simulation model on the grid geometry (Panel 4.6E). . . . .	59
4.8	The mean CRPS and its standard error for each inferential model and simulation model on the Côte d’Ivoire geometry (Panel 4.6F). . . . .	60
4.9	The mean CRPS and its standard error for each inferential model and simulation model on the Texas geometry (Panel 4.6G). . . . .	61
4.10	Adult (15-49) HIV prevalence from the most recent PHIA survey conducted in Côte d’Ivoire (Panel A), Malawi (Panel B), Tanzania (Panel C), and Zimbabwe (Panel D). These estimates are survey weighted according to Equation (4.39). . . . .	62
4.11	In leave-one-out (LOO) cross-validation, one observation is left out of the training data and predicted upon in each fold. The spatial-leave-one-out (SLOO) cross-validation scheme considered here is similar, only differing in that observations corresponding to adjacent areas are also left out of the training data. . . . .	63

## List of Figures

4.12 The mean pointwise leave-one-out and spatial leave-one-out CRPS in estimating $\rho_i$ using each inferential model for the four PHIA surveys described in Table 4.3. The 95% credible intervals shown are generated using 1.96 times the standard error. . . . .	65
5.1 Risk of acquiring HIV depends on both individual-level risk behaviour and population-level HIV incidence. It is assumed here 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. . . . .	70
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. . . . .	72
5.3 Flowchart describing how respondents were classified to HIV risk groups based on their survey responses. . . . .	74
5.4 For the multinomial logistic regression model, under the conditional predictive ordinate (CPO) criterion, including Besag spatial random effects rather than IID spatial random effects improved model performance. On the other hand, under the deviance information criterion (DIC) and widely applicable information criterion (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. . . . .	82
5.5 For the logistic regression model, the CPO, DIC, and WAIC each agreed that the model containing Besag spatial random effects and the <code>cfswrecent</code> covariates was best. Inclusion of Besag spatial random effects consistently improved each criterion, whereas improvements from inclusion of any covariates were marginal. . . . .	84
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. . . . .	86
5.7 The posterior mean of the AGYW risk group proportions over space 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, though desirable for other reasons, makes it challenging to see spatial variation in the FSW risk group. . . . .	89

## List of Figures

5.8	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 5.7, this figure presents a clear view of within- and between-country variation in risk group proportions.	90
5.9	Probability integral transform (PIT) histograms (top row) and empirical cumulative distribution function (ECDF) difference plots (bottom row) for the final selected model.	92
5.10	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.	93
5.11	The modelled estimates display more plausible spatial smoothness than the direct estimates. In addition, missing values in the direct estimates are appropriately infilled by the model.	96
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 $\text{Gamma}(9, 4)$ . 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 $\log \tilde{p}_{\text{LA}}(\mathbf{y}) = -1.882458$ , resulting from the Gaussian approximation $p_{\mathcal{G}}(\phi   \mathbf{y}) = \mathcal{N}(\phi   \mu = 2, \tau = 2)$ .	105
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.	108
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 $\alpha \in \mathbb{R}$ .	112

## List of Figures

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 Håvard Rue, Martino, and Chopin (2009). Briefly, equally-weighted quadrature points are generated by starting at the mode and taking steps of size $\delta_z$ 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 $\delta_\pi$ . 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 Håvard Rue, Martino, and Chopin (2009). The CCD was generated using the <code>rsm</code> 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. . . . .	119
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. . . . .	123
6.6 A submatrix of the full parameter Hessian obtained from <code>TMB::sdreport</code> with <code>getJointPrecision = TRUE</code> on the log scale. Entries for the latent field parameters $\epsilon$ and $\nu$ 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. . . . .	127
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 <code>tmbstan</code> . Results from R-INLA and TMB are similar, especially for the posterior mean, but do differ in places. Differences could be attributable to bias corrections used in R-INLA. . . . .	132
6.8 The ECDF and ECDF difference for the $\beta_0$ latent field parameter. For this parameter, the Gaussian marginal results are inaccurate, and are corrected almost entirely by the Laplace marginal. An ECDF difference of zero corresponds to obtaining exactly the same results as NUTS, taken to be the gold-standard. Crucially, results obtained using R-INLA and TMB implementations are similar. . . . .	133

## List of Figures

6.9	The number of seconds taken to perform inference for the epilepsy GLMM using each method and software implementation given in Table 6.1. . . . .	134
6.10	Empirical prevalence of Loa loa in 190 sampled villages in Cameroon and Nigeria. The map in Panel A shows the village locations, empirical prevalences, presence of zeros, and sample sizes. The zeros are typically located in close proximity to each other. The histogram in Panel B shows the empirical prevalences, and high number of zeros. . . . .	134
6.11	Posterior mean of the suitability $\mathbb{E}[\phi_{\text{LA}}(s)]$ (Panel A) and prevalence $\mathbb{E}[\rho_{\text{LA}}(s)]$ (Panel B) random fields computed using Laplace marginals. Inferences over this fine spatial grid were using conditional Gaussian field simulation as implemented by <code>gstat::krige</code> . . . . .	136
6.12	Difference between the suitability posterior means with Gaussian marginals $\mathbb{E}[\phi_{\text{G}}(s)]$ and Laplace marginals $\mathbb{E}[\phi_{\text{LA}}(s)]$ to NUTS results. While the Gaussian approximation appears to systematically underestimate suitability, results from the Laplace approximation are substantially closer to results from NUTS. As $\beta_\phi$ was fixed this difference is as a result in differences in estimation of $u(s)$ . The diverging colour palette used in this figure is from Thyng et al. (2016). . . . .	137
6.13	Difference between the prevalence posterior means with Gaussian marginals $\mathbb{E}[\rho_{\text{G}}(s)]$ and Laplace marginals $\mathbb{E}[\rho_{\text{LA}}(s)]$ to NUTS results. Like the suitability in Figure 6.12, the error the the Gaussian approximation is higher than that of the Laplace approximation. As $\beta_\rho$ was fixed this difference is as a result in differences in estimation of $v(s)$ . The diverging colour palette used in this figure is from Thyng et al. (2016). . . . .	138
6.14	Absolute difference between the Gaussian and Laplace marginal posterior means and standard deviations to NUTS results at each $u(s_i), v(s_i) : i \in [190]$ . Relative differences are in Figure C.4. For close to every node, the Laplace approximation produced a more accurate posterior mean than the Gaussian approximation. For the posterior standard deviation (SD), the picture was more mixed. . . . .	139
6.15	The element of the latent field with maximum difference in absolute difference to NUTS for the posterior mean was $u_{184}$ . While the Gaussian approximation has substantial error as compared with NUTS, the Laplace approximation is a close match. . . . .	140
6.16	The number of minutes taken to perform inference for the Loa loa ELGM using each approach given in Table 6.2. . . . .	140

## *List of Figures*

6.17 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 usual AGHQ nodes with a spectral matrix decomposition. 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.	151
6.18 District-level HIV prevalence, ART coverage, and new HIV cases and HIV incidence for adults 15-49 in Malawi. Inference here was conducted using a Gaussian approximation and EB via TMB.	152
6.19 Under PCA, the proportion of total variation explained is given by the sum of the first $s$ eigenvalues over the sum of all eigenvalues. A typical rule-of-thumb is to include dimensions sufficient to explain 90% of total variation. In this case, for computational reasons, 87% was considered sufficient.	154
6.20 The full rank original covariance matrix (Panel A) was closely reproduced by its reduced rank ( $s = 8$ ) matrix approximation (Panel B).	155
6.21 Each principal component loading, obtained by the eigendecomposition of the inverse curvature, gives the direction of maximum variation conditional on inclusion of each previous principal component loading. For example, the first principal component loading is a sum of <code>log_sigma_alpha_as</code> and <code>logit_phi_alpha_as</code> .	155
6.22 The 6561 PCA-AGHQ nodes projected onto the 24 hyperparameter marginal distributions obtained with NUTS.	157
6.23 The number of hours taken to perform inference for the Naomi ELGM (Section 6.3.1) using each approach.	158
6.24 The latent field posterior mean and posterior standard deviation point estimates from each inference method as compared with those from NUTS. The root mean square error (RMSE) and mean absolute error (MAE) are displayed in the top left. For both the posterior mean and posterior standard deviation, GPCA-AGHQ reduced RMSE and MAE as compared with GEB.	159

## List of Figures

6.25 The average Kolmogorov-Smirnov (KS) test statistic for each latent field parameter of the Naomi model. Vectors of parameters were grouped together. For points above the dashed line at zero, performance of GEB was better. For points below the dashed line, performance of GPCA-AGHQ was better. Most notably, for the latent field parameters <code>ui_lambda_x</code> the test statistic for GEB was substantially higher than for GPCA-AGHQ. This parameter, of length 32, corresponds to $\mathbf{u}_x^\lambda$ and plays a key role in the ART attendance component of the Naomi (Section 6.3.1.4). . . . .	160
6.26 The parameter <code>ui_lambda_x</code> [26] had the greatest difference in KS test statistics between GEB and GPCA-AGHQ to NUTS. For this parameter, the potential scale reduction factor was 1 and effective sample size was 2100. . . . .	161
6.27 The probability each strata has met the second 90 (ART coverage above 81%) calculated using each inference method, as compared with NUTS. The root mean square error (RMSE) and mean absolute error (MAE) are displayed in the top left. . . . .	162
6.28 The probability each strata has high HIV incidence (above 1% per year) calculated using each inference method, as compared with NUTS. The root mean square error (RMSE) and mean absolute error (MAE) are displayed in the top left. . . . .	162
6.29 Monthly R package downloads from the Comprehensive R Archive Network (CRAN) for <code>brms</code> , <code>glmmTMB</code> , <code>nimble</code> , <code>rstan</code> and <code>TMB</code> , obtained using the <code>cranlogs</code> (Csárdi 2023) R package. Unfortunately, <code>R-INLA</code> is not available from CRAN, and so could not be included in this figure. The official <code>rstan</code> documentation recommends installation of a development version hosted outside CRAN. As such, this metric may underestimate the popularity of <code>rstan</code> . . . . .	168
7.1 Panel A shows the front page of UNAIDS (2023b). Panel B shows the page containing text and a figure based on the work done in Chapter 5. In this figure, 30 countries are included. . . . .	171
7.2 For the Loa loa ELGM (Section 6.2.2), increasing the number of quadrature nodes per hyperparameter dimension from $k = 3$ to $k = 7$ did little to improve accuracy. On the other hand, using Laplace marginals rather than Gaussian marginals did have a substantial effect (Figures 6.12 and 6.13). It would be valuable to better understand, and aspirationally have diagnostics for, the circumstances under which accuracy of INLA methods could be improved by additional computation. . . . .	173

## *List of Figures*

A.1 A comparison of time taken to fit AGHQ via <code>aghq</code> as compared with NUTS via <code>tmbstan</code> for each inferential model. For the models run using NUTS via <code>tmbstan</code> there was significant variation in time taken depending on initial random seed. As such, these timings and more broadly the inferences obtained from NUTS in Appendix A.1 should be interpreted with appropriate skepticism. . . . .	179
A.2 A comparison of the posterior means and standard deviations obtained with AGHQ via <code>aghq</code> as compared with NUTS via <code>tmbstan</code> fitting an IID inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 1686, and the maximum value of the potential scale reduction factor was 1.00.	179
A.3 A comparison of the posterior means and standard deviations obtained with AGHQ via <code>aghq</code> as compared with NUTS via <code>tmbstan</code> fitting a Besag inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 1056, and the maximum value of the potential scale reduction factor was 1.00.	180
A.4 A comparison of the posterior means and standard deviations obtained with AGHQ via <code>aghq</code> as compared with NUTS via <code>tmbstan</code> fitting a BYM2 inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 35, and the maximum value of the potential scale reduction factor was 1.06.	180
A.5 A comparison of the posterior means and standard deviations obtained with AGHQ via <code>aghq</code> as compared with NUTS via <code>tmbstan</code> fitting a FCK inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 355, and the maximum value of the potential scale reduction factor was 1.01.	181
A.6 A comparison of the posterior means and standard deviations obtained with AGHQ via <code>aghq</code> as compared with NUTS via <code>tmbstan</code> fitting a CK inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 1471, and the maximum value of the potential scale reduction factor was 1.00.	181
A.7 A comparison of the posterior means and standard deviations obtained with AGHQ via <code>aghq</code> as compared with NUTS via <code>tmbstan</code> fitting a FIK inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 289, and the maximum value of the potential scale reduction factor was 1.01.	182

## List of Figures

A.8 A comparison of the posterior means and standard deviations obtained with AGHQ via <code>aghq</code> as compared with NUTS via <code>tmbstan</code> fitting a IK inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 1623, and the maximum value of the potential scale reduction factor was 1.00.	182
A.9 The probability density for each lengthscale prior distribution as given in Table A.1. . . . .	183
A.10 Lengthscale posterior distributions obtained using NUTS to fit a centroid kernel model to integrated kernel data. The true value, 2.5, is shown as a dashed vertical line. Six different lengthscale prior distributions were considered as given in Table A.1. The geometry used was the grid (Panel 4.6E). . . . .	184
A.11 The lengthscale posterior mean and 95% credible interval obtained using the centroid kernel model on integrated kernel data for the first 40 simulation replicates on each geometry. The true lengthscale, and lengthscale obtained using the heuristic method of Best et al. (1999), are shown as dashed horizontal lines. . . . .	187
A.12 The BYM2 proportion parameter posterior mean and 95% credible interval obtained for the first 40 simulation replicates for the realistic geometries. When the simulated data is IID, the BYM2 proportion parameter is in the majority of cases below 0.5, corresponding to have inferred that the noise is mostly IID (spatially unstructured) When the simulated data is either Besag or IK, the BYM2 proportion parameter is in the majority of cases above 0.5, corresponding to have inferred that the noise is mostly Besag (spatially structured). .	188
A.13 The mean CRPS with 95% credible interval in estimating $\rho$ using each inferential model and simulation model on the first vignette geometry (Panel 4.6A). Credible intervals were generated using 1.96 times the standard error. . . . .	189
A.14 The mean CRPS with 95% credible interval in estimating $\rho$ using each inferential model and simulation model on the second vignette geometry (Panel 4.6B). Credible intervals were generated using 1.96 times the standard error. . . . .	190
A.15 The mean CRPS with 95% credible interval in estimating $\rho$ using each inferential model and simulation model on third vignette geometry (Panel 4.6C). Credible intervals were generated using 1.96 times the standard error. . . . .	191

## *List of Figures*

A.16 The mean CRPS with 95% credible interval in estimating $\rho$ using each inferential model and simulation model on the fourth vignette geometry (Panel 4.6D). Credible intervals were generated using 1.96 times the standard error. . . . .	192
A.17 Choropleths showing the mean value of the CRPS in estimating $\rho$ , under each inferential model and simulation model, at each area of the first vignette geometry (Panel 4.6A). . . . .	193
A.18 Choropleths showing the mean value of the CRPS in estimating $\rho$ , under each inferential model and simulation model, at each area of the second vignette geometry (Panel 4.6B). . . . .	194
A.19 Choropleths showing the mean value of the CRPS in estimating $\rho$ , under each inferential model and simulation model, at each area of the third vignette geometry (Panel 4.6C). . . . .	195
A.20 Choropleths showing the mean value of the CRPS in estimating $\rho$ , under each inferential model and simulation model, at each area of the fourth vignette geometry (Panel 4.6D). . . . .	196
A.21 Choropleths showing the mean value of the CRPS in estimating $\rho$ , under each inferential model and simulation model, at each area of the grid geometry (Panel 4.6E). . . . .	197
A.22 Choropleths showing the mean value of the CRPS in estimating $\rho$ , under each inferential model and simulation model, at each area of the Côte d'Ivoire geometry (Panel 4.6F). . . . .	198
A.23 Choropleths showing the mean value of the CRPS in estimating $\rho$ , under each inferential model and simulation model, at each area of the Texas geometry (Panel 4.6G). . . . .	199
A.24 Probability integral transform histograms and empirical cumulative distribution function difference plots for $\rho$ , under each inferential model and simulation model, for the first vignette geometry (Panel 4.6A). . . . .	200
A.25 Probability integral transform histograms and empirical cumulative distribution function difference plots for $\rho$ , under each inferential model and simulation model, for the second vignette geometry (Panel 4.6B). . . . .	201
A.26 Probability integral transform histograms and empirical cumulative distribution function difference plots for $\rho$ , under each inferential model and simulation model, for the third vignette geometry (Panel 4.6C). . . . .	202

## *List of Figures*

A.27 Probability integral transform histograms and empirical cumulative distribution function difference plots for $\rho$ , under each inferential model and simulation model, for the fourth vignette geometry (Panel 4.6D). . . . .	203
A.28 Probability integral transform histograms and empirical cumulative distribution function difference plots for $\rho$ , under each inferential model and simulation model, for the grid geometry (Panel 4.6E). . .	204
A.29 Probability integral transform histograms and empirical cumulative distribution function difference plots for $\rho$ , under each inferential model and simulation model, for the Côte d'Ivoire geometry (Panel 4.6F). . . . .	205
A.30 Probability integral transform histograms and empirical cumulative distribution function difference plots for $\rho$ , under each inferential model and simulation model, for the Texas geometry (Panel 4.6G). .	206
A.31 The lengthscale hyperparameter prior and posterior distributions for each of the four considered PHIA surveys (Table 4.3), using both the CK and IK inferential models. . . . .	207
A.32 The BYM2 proportion hyperparameter prior and posterior distributions for each of the four considered PHIA surveys (Table 4.3). A value of zero corresponds to IID noise. A value of one corresponds to Besag noise. For each survey, excluding the Côte d'Ivoire 2017 PHIA, the posterior distribution for the BYM2 proportion is concentrated towards a value of one. This result can be interpreted as suggesting that the variation in HIV prevalence from these surveys is spatially structured. . . . .	208
A.33 The HIV prevalence posterior mean and 95% credible interval for each area of Côte d'Ivoire, based on the 2017 PHIA survey. Direct estimates obtained from the survey are as shown in Panel 4.10A. . .	209
A.34 The HIV prevalence posterior mean and 95% credible interval for each area of Malawi, based on the 2016 PHIA survey. Direct estimates obtained from the survey are as shown in Panel 4.10B. . . . .	210
A.35 The HIV prevalence posterior mean and 95% credible interval for each area of Tanzania, based on the 2017 PHIA survey. Direct estimates obtained from the survey are as shown in Panel 4.10C. . . . .	211
A.36 The HIV prevalence posterior mean and 95% credible interval for each area of Zimbabwe, based on the 2016 PHIA survey. Direct estimates obtained from the survey are as shown in Panel 4.10D. . .	212

## List of Figures

A.37 The pointwise CRPS in estimating $\rho_i$ using either leave-one-out or spatial leave-one-out cross-validation, with mean and 95% credible interval for the Côte d'Ivoire 2017 PHIA survey (Panel 4.10A) . . . . .	213
A.38 The pointwise CRPS in estimating $\rho_i$ using either leave-one-out or spatial leave-one-out cross-validation, with mean and 95% credible interval, for the Malawi 2016 PHIA survey 4.10B. . . . .	214
A.39 The pointwise CRPS in estimating $\rho_i$ using either leave-one-out or spatial leave-one-out cross-validation, with mean and 95% credible interval, for the Tanzania 2017 PHIA survey 4.10C. . . . .	215
A.40 The pointwise CRPS in estimating $\rho_i$ using either leave-one-out or spatial leave-one-out cross-validation, with mean and 95% credible interval, for the Zimbabwe 2016 PHIA survey 4.10D. . . . .	216
A.41 Choropleth showing the pointwise CRPS in estimating $\rho_i$ using either leave-one-out or spatial leave-one-out cross-validation for the Côte d'Ivoire 2017 PHIA survey (Panel 4.10A). . . . .	217
A.42 Choropleth showing the pointwise CRPS in estimating $\rho_i$ using either leave-one-out or spatial leave-one-out cross-validation for the Malawi 2016 PHIA survey (Panel 4.10B). . . . .	217
A.43 Choropleth showing the pointwise CRPS in estimating $\rho_i$ using either leave-one-out or spatial leave-one-out cross-validation for the Tanzania 2017 PHIA survey (Panel 4.10C). . . . .	218
A.44 Choropleth showing the pointwise CRPS in estimating $\rho_i$ using either leave-one-out or spatial leave-one-out cross-validation for the Zimbabwe 2016 PHIA survey (Panel 4.10D). . . . .	218
A.45 Probability integral transform histograms and empirical cumulative distribution function difference plots in estimating $\rho$ for the Côte d'Ivoire 2017 PHIA survey (Panel 4.10A). . . . .	219
A.46 Probability integral transform histograms and empirical cumulative distribution function difference plots in estimating $\rho$ for the Malawi 2016 PHIA survey (Panel 4.10B). . . . .	220
A.47 Probability integral transform histograms and empirical cumulative distribution function difference plots in estimating $\rho$ for the Tanzania 2017 PHIA survey (Panel 4.10C). . . . .	221
A.48 Probability integral transform histograms and empirical cumulative distribution function difference plots in estimating $\rho$ for the Zimbabwe 2016 PHIA survey (Panel 4.10D). . . . .	222

## List of Figures

B.1	The proportion of posterior variance explained by each random effect, calculated as a ratio of the random effect variance posterior mean to the sum of all random effect variance posterior means. To allow calculation of this metric by country, the model was run for each country individually. . . . .	228
B.2	For the 20-24 and 25-29 age groups, the proportion of AGYW in the one cohabiting partner and non-regular or multiple partner(s) risk groups was bimodal. . . . .	229
C.1	Traceplots for the <code>tmbstan</code> parameters with the lowest ESS and highest potential scale reduction factor. These were <code>1_tau_nu</code> (an ESS of 377) and <code>beta[3]</code> (an $\hat{R}$ of 1.006). . . . .	235
C.2	Traceplots for the <code>rstan</code> parameters with the lowest ESS and highest potential scale reduction factor. These were <code>tau_nu</code> (an ESS of 437) and <code>tau_nu</code> (an $\hat{R}$ of 1.009). Rather than plotting the traceplot for <code>tau_nu</code> twice, the parameter <code>epsilon[18]</code> is included, which had the second highest $\hat{R}$ of 1.008. . . . .	236
C.3	Traceplots for the parameters with the lowest ESS and highest potential scale reduction factor for the Loa loa ELGM example. . . . .	236
C.4	Relative difference between the Gaussian and Laplace marginal posterior means and standard deviations to NUTS results at each $u(s_i), v(s_i) : i \in [190]$ . Absolute differences are in Figure 6.14. . . . .	237
C.5	For NUTS run on the Naomi ELGM, the maximum potential scale reduction factor was 1.021, below the value of 1.05 typically used as a cutoff for acceptable chain mixing, indicating that the results are acceptable to use. Additionally, the vast majority (93.7%) of $\hat{R}$ values were less than 1.1. . . . .	247
C.6	The efficiency of the NUTS, as measured by the ratio of effective sample size to total number of iterations run, was low for most parameters (Panel A). As a result, the number of iterations required for the the effective number of samples (mean 1265) to be satisfactory was high (Panel B). . . . .	248
C.7	Traceplots for the parameter with the lowest ESS which was <code>log_sigma_alpha_xs</code> (an ESS of 208, Panel A) and highest potential scale reduction factor which was <code>ui_lambda_x[10]</code> (an $\hat{R}$ of 1.021, Panel B). . . . .	248
C.8	Pairs plots for the parameters $\log(\sigma_A^\rho)$ and $\text{logit}(\phi_A^\rho)$ , or <code>log_sigma_rho_a</code> and <code>logit_phi_rho_a</code> as implemented in code. These parameters are the log standard deviation and logit lag-one correlation parameter of an AR1 process. In the posterior distribution obtained with NUTS, they have a high degree of correlation. . . . .	249

## *List of Figures*

C.9 Pairs plots for the parameters $\log(\sigma_X^\alpha)$ and $\text{logit}(\phi_X^\alpha)$ , or <code>log_sigma_alpha_x</code> and <code>logit_phi_alpha_x</code> as implemented in code. These parameters are the log standard deviation and logit BYM2 proportion parameter of a BYM2 process. In the posterior distribution obtained with NUTS, they are close to uncorrelated. . . . .	249
C.10 Prior standard deviations were calculated by using NUTS to simulate from the prior distribution. This approach is more convenient than simulating directly from the model, but can lead to inaccuracies. . . . .	250
C.11 The posterior contraction for each parameter in the model. Values are averaged for parameters of length greater than one. The posterior contraction is zero when the prior distribution and posterior distribution have the same standard deviation. This could indicate that the data is not informative about the parameter. The closer the posterior contraction is to one, the more than the marginal posterior distribution has concentrated about a single point. . . . .	251
C.12 The standard deviation of the quadrature nodes can be used as a measure of coverage of the posterior marginal distribution. Nodes spaced evenly within the marginal distribution would be expected to uniformly distributed quantile, corresponding to a standard deviation of 0.2871, shown as a dashed line. . . . .	252
C.13 The estimated posterior marginal standard deviation of each hyperparameter varied substantially based on its scale, either logarithmic or logistic. . . . .	252
C.14 The logarithm of the normalising constant estimated using PCA-AGHQ and a range of possible values of $k = 2, 3, 5$ and $s \leq 8$ . Using this range of settings, there was not convergence of the logarithm of the normalising constant estimate. The time taken by GPCA-AGHQ increases exponentially with number of PCA-AGHQ dimensions kept.	253
C.15 Differences in Naomi model output posterior means as estimated by GEB and GPCA-AGHQ compared to NUTS. Each point is an estimate of the indicator for a particular strata. In all cases, error is reduced by GPCA-AGHQ, most of all for ART coverage. . . . .	254
C.16 Differences in Naomi model output posterior standard deviations as estimated by GEB and GPCA-AGHQ compared to NUTS. Each point is an estimate of the indicator for a particular strata. Error is increased by GPCA-AGHQ for HIV prevalence and HIV indiccence, and reduced for ART coverage. . . . .	255

*List of Figures*

C.17 The Kolomogorov-Smirnov (KS) test statistic for each latent field parameter is correlated with the effective sample size (ESS) from NUTS, for both GEB and GPCA-AGHQ. This may be because parameters which are harder to estimate with INLA-like methods also have posterior distributions which are more difficult to sample from. Alternatively, it may be that high KS values are caused by inaccurate NUTS estimates generated by limited effective samples. . 256

# List of Tables

4.1	The three spatial random effect models used to generate synthetic data in the simulation study (Section 4.3). . . . .	53
4.2	The spatial random effect models used for inference. Each model is implemented in the <code>arealutils</code> package. The BYM2 model was implemented using the sparsity preserving parameterisation described in Section 3.2 of Riebler et al. (2016). . . . .	55
4.3	The four PHIA household surveys included in the HIV prevalence study (Section 4.4). . . . .	61
4.4	The mean pointwise leave-one-out and spatial leave-one-out CRPS in estimating $\rho_i$ , with standard errors, for each inferential model across the four considered PHIA surveys. The units used in this table are thousandths. . . . .	63
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. . . . .	71
5.2	Four multinomial regression models were considered. Observation random effects $\theta_{ita}$ , included in all models, are omitted from this table.	77
5.3	Applying sum-to-zero constraints to interaction effects ensured that the main effect was not interfered with. . . . .	81
5.4	Conditional predictive ordinate (CPO), deviance information criterion (DIC), and widely applicable information criterion (WAIC) values for the multinomial logistic regression model specifications with corresponding standard errors. . . . .	82

*List of Tables*

5.5	Six logistic regression models were considered. The covariate <code>cfswever</code> denotes the proportion of men who have ever paid for sex and <code>cfswrecent</code> denotes the proportion of men who have paid for sex in the past 12 months. . . . .	83
5.6	CPO, DIC, and WAIC values for the logistic regression model specifications with corresponding standard errors. . . . .	85
6.1	The inference methods and software considered to fit the epilepsy GLMM in Section 6.2.1. . . . .	122
6.2	The inference methods and software considered to fit the Loa loa ELGM in Section 6.2.2. . . . .	135
A.1	Six lengthscale prior distributions were considered for use in the simulation (Section 4.3) and HIV prevalence (Section 4.4) studies. .	183
A.2	The average mean squared error (MSE) of each inferential model in estimating $\rho$ , under different simulation and geometry settings. Entries for FCK and CK on geometry 2 are empty because model was undefined in that case. The units used in this table are thousandths.	184
A.3	The average continuous ranked probability score (CRPS) of each inferential model in estimating $\rho$ , under different simulation and geometry settings. Entries for FCK and CK on geometry 2 are empty because model was undefined in that case. The units used in this table are thousandths. . . . .	185
A.4	The mean pointwise leave-one-out and spatial leave-one-out MSE in estimating $\rho_i$ , with standard errors, for each inferential model across the four considered PHIA surveys. The units used in this table are thousandths. . . . .	187
B.1	Prioritisation strata for AGYW given by UNAIDS (2021b) based on to HIV incidence in the general population and behavioural risk. .	223
B.2	Commitments recommended by UNAIDS (2021b) to be met for each HIV intervention, given in terms of the proportion of the AGYW prioritisation strata reached. The symbol “-” represents no commitment.	223
B.3	The sample size by age group for each included survey in the analysis. The column “TS question” refers to whether or not the survey included a specific question about transactional sex (TS). . . . .	224
B.4	All of that household surveys that were excluded from the risk group model in Section 5.3. . . . .	226
B.5	The number of areas and analysis level for each country that was used in the analysis. . . . .	226

*List of Tables*

B.6	The behavioural survey questions included in AIDS Indicator Survey (AIS) and Demographic and Health Surveys (DHS) used to determine AGYW risk group membership. . . . .	227
B.7	The behavioural survey questions included in Population-Based HIV Impact Assessment (PHIA) surveys used to determine AGYW risk group membership. . . . .	227
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. .	240
C.2	Each term in Equation (C.14) together with, where applicable, its prior distribution and a written description of its role. . . . .	240
C.3	Each term in Equation (C.19) together with, where applicable, its prior distribution and a written description of its role. . . . .	242
C.4	Each term in Equations (C.21) and (C.22) together with (where applicable) its prior distribution and a written description of its role. The notation $\theta$ is used as stand in for $\theta \in \{\rho, \alpha\}$ . . . . .	243
C.5	Each term in Equation (C.24) together with, where applicable, its prior distribution and a written description of its role. As no terms include $x'$ , $\gamma_{x,x'}$ is only a function of $x$ . . . . .	244
C.6	Correspondence between the variable name used in the Naomi TMB template and the mathematical notation used in Appendix C.4. The parameter type, either a hyperparameter or element of the latent field, is also given. All of the parameters are defined on the real-scale in some dimension. In the final three columns ( $\rho$ , $\alpha$ , and $\lambda$ ) indication is given as to which component of the model the parameter is primarily used in. . . . .	246

## List of Abbreviations

<b>AIDS</b>	Acquired ImmunoDeficiency Syndrome.
<b>AIS</b>	AIDS Indicator Survey.
<b>ANC</b>	Antenatal Clinic.
<b>AGHQ</b>	Adaptive Gauss-Hermite Quadrature.
<b>ART</b>	Antiretroviral Therapy.
<b>BF</b>	Bayes Factor.
<b>BIC</b>	Bayesian Information Criterion.
<b>CCD</b>	Central Composite Design.
<b>CAR</b>	Conditionally Auto-regressive.
<b>CDC</b>	Centers for Disease Control and Prevention.
<b>CCP</b>	Conditional Predictive Ordinate.
<b>CRAN</b>	Comprehensive R Archive Network.
<b>CRPS</b>	Continuous Ranked Probability Score.
<b>DALY</b>	Disability Adjusted Life Year.
<b>DDC</b>	Data Defect Correlation.
<b>DHS</b>	Demographic and Health Surveys.
<b>DIC</b>	Deviance Information Criterion.
<b>EB</b>	Empirical Bayes.
<b>ECDF</b>	Empirical Cumulative Difference Function.
<b>ELGM</b>	Extended Latent Gaussian Model.
<b>ESS</b>	Effective Sample Size.
<b>FSW</b>	Female Sex Worker(s).
<b>GC</b>	Generalised Linear Model.
<b>GLM</b>	Generalised Linear Model.
<b>GLMM</b>	Generalised Linear Mixed effects Model.

*List of Abbreviations*

<b>GMRF</b>	.....	Gaussian Markov Random Field.
<b>GP</b>	.....	Gaussian Process.
<b>HIV</b>	.....	Human Immunodeficiency Virus.
<b>HMC</b>	.....	Hamiltonian Monte Carlo.
<b>ICAR</b>	.....	Intrinsic Conditionally Auto-regressive.
<b>IID</b>	.....	Independent and Identically Distributed.
<b>INLA</b>	.....	Integrated Nested Laplace Approximation.
<b>LM</b>	.....	Linear Model.
<b>LGM</b>	.....	Latent Gaussian Model.
<b>LS</b>	.....	Log Score.
<b>MCMC</b>	.....	Markov Chain Monte Carlo.
<b>MSM</b>	.....	Men who have Sex with Men.
<b>NUTS</b>	.....	No-U-Turn Sampler.
<b>PEP</b>	.....	Post-Exposure Prophylaxis.
<b>PEPFAR</b>	....	President's Emergency Plan for AIDS Relief.
<b>PHIA</b>	.....	Population-based HIV Impact Assessment.
<b>PIT</b>	.....	Probability Integral Transform.
<b>PLHIV</b>	....	People Living with HIV.
<b>PWID</b>	....	People Who Inject Drugs.
<b>PPL</b>	.....	Probabilistic Programming Language.
<b>PrEP</b>	.....	Pre-Exposure Prophylaxis.
<b>SAE</b>	.....	Small-Area Estimation.
<b>SR</b>	.....	Scoring Rule.
<b>SPSR</b>	....	Strictly Proper Scoring Rule.
<b>SSA</b>	....	Sub-Saharan Africa.
<b>STI</b>	.....	Sexually Transmitted Infection.
<b>TaSP</b>	.....	Treatment as Prevention.
<b>TGP</b>	.....	Transgender People.
<b>UNAIDS</b>	....	The Joint United Nations Programme on HIV/AIDS.
<b>VI</b>	.....	Variational Inference.
<b>VMMC</b>	....	Voluntary Medical Male Circumcision.
<b>WAIC</b>	....	Watanabe-Akaike Information Criterion.

# 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.
$\rho$	HIV prevalence.
$\lambda$	HIV incidence.
$\alpha$	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 $(y_1, \dots, y_n)$ .
$\boldsymbol{\phi}$	Parameters, a $d$ -vector $(\phi_1, \dots, \phi_d)$ .
$\mathbf{x}$	Latent field, a $N$ -vector $(x_1, \dots, x_N)$ .
$\boldsymbol{\theta}$	Hyperparameters, a $m$ -vector $(\theta_1, \dots, \theta_m)$ .
$x \sim p(x)$	$x$ has the probability distribution $p(x)$ .
$A_i$	Areal unit.
$A_i \sim A_j$	Adjacency between areal units.
$\mathbf{u}$	Random effects, often spatial.
$\mathbf{H}$	Hessian matrix.
$\mathbf{R}$	Structure matrix.
$\mathbf{Q}$	Precision matrix.
$\boldsymbol{\Sigma}$	Covariance matrix.
$\mathbf{M}^-$	The generalised inverse of a (potentially rank-deficient) matrix $\mathbf{M}$

*List of Notations*

$\mathcal{N}$	.....	Gaussian distribution.
$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$		Kernel function on the space $\mathcal{X}$ .
$A_i \sim A_j$	.....	Adjacency between areal units.
$\mathcal{Q}$	.....	A set of quadrature nodes.
$\omega : \mathcal{Q} \rightarrow \mathbb{R}$	....	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.
$\varphi$	.....	A standard (multivariate) Gaussian density.

# 1

## Introduction

This thesis is about applied and methodological Bayesian statistics. It is applied and methodological in that the primary concern is real world questions and the means to answer them. The statistical approach is Bayesian because probability theory is used 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 HIV (human immunodeficiency virus) epidemic in sub-Saharan Africa (SSA). Over 40 years since the beginning of the epidemic, among non-infants HIV is the largest annual cause of disability adjusted life years (DALYs) in SSA [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 used 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 observations were recorded. Spatio-temporal data have important recurring commonalities across a diverse range of application

## *Introduction*



**Figure 1.1:** HIV 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.

settings. The work conducted in this thesis uses, 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>. To facilitate reproducible research, the R package `orderly` (FitzJohn et al. 2023) was used to structure code repositories.

## 1.1 Chapter overview

This thesis is structured as follows:

- Chapter 2 provides an overview of the HIV/AIDS epidemic, and describes 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.

## *Introduction*

- 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, areas correspond to the administrative divisions of a country, which are typically not a grid. I used simulation and survey data studies to evaluate the practical consequences of this concern
- 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. This work (Howes et al. 2023) was included in the UNAIDS (Joint United Nations Programme on HIV/AIDS) Global AIDS Update 2022 and 2023.
- Chapter 6: The Naomi small-area estimation model (Jeffrey W Eaton et al. 2021) is used by countries to estimate district-level HIV indicators. First, to allow for compatibility with Naomi, I implemented the integrated nested Laplace approximations using automatic differentiation, opening the door to a new class of fast, flexible, and accurate Bayesian inference algorithms. The implementation was using models for a clinical trial of an epilepsy drug, and for the prevalence of the parasitic worm *Loa loa*. Second, I developed an approximate Bayesian inference method combining adaptive Gauss-Hermite quadrature with principal components analysis. 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 contributions of the research, avenues for future work, and some broader reflections.

Though chronological order is recommended, Chapters 4, 5 and 6 may be read in any order, or as stand-alone studies, if preferred.

# 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. 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 (UNAIDS 2023a).

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

## The HIV/AIDS epidemic



**Figure 2.1:** Globally, yearly new HIV infections peaked in 1995, and have since decreased by 59%. Yearly AIDS-related deaths peaked in 2004, and have since decreased by 68% (UNAIDS 2023a). Much of the global disease burden is concentrated in eastern and southern Africa, as well as western and central Africa. The unit “M” refers to millions. The colour palette used in this figure, and throughout the thesis, is that of Okabe and Ito (2008). It is designed to be colourblind friendly, and the default in Wilke (2019).

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). 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), indicating the scale of the response.

Implementation of HIV prevention and treatment has significantly reduced the number of new HIV infections and AIDS-related deaths per year since their respective peaks (Figure 2.1). The most significant evidence-based interventions, in approximately 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

## *The HIV/AIDS epidemic*

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 (M. S. 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 this way 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 (RCTs) (Auvert et al. 2005; Gray et al. 2007; R. C. 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 (WHO and UNAIDS 2007). 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 (McGillen et al. 2018; UNAIDS and WHO 2021).

## *The HIV/AIDS epidemic*

- Pre-exposure prophylaxis (PrEP) and post-exposure prophylaxis (PEP) are antiretroviral drugs which can be taken before and after exposure to prevent transmission. PrEP has been shown to be effective at an individual-level across a number of RCTs (Baeten et al. 2012; Thigpen et al. 2012), but there are few population-level studies. Though PEP cannot be studied with RCTs, observational studies indicate it is highly effective (Dominguez et al. 2016). These medical interventions are more costly than some other options, so are 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. By analogy to precision medicine, this paradigm has been termed precision public health (Khoury et al. 2016). While precision medicine tailors treatments to the individual, precision public health tailors treatments 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

## The HIV/AIDS epidemic



**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.

estimated that 66% of the 39 million PLHIV worldwide live in SSA. Adult HIV prevalence (ages 15-49) is above 10% in some countries in southern Africa, with some districts even exceeding 20% (Figure 2.2). Indeed, just as there is variation between countries, there is variation within countries. As an illustration, adult HIV prevalence at the district municipality level in South Africa ranges from 6% in Namakwa to 30% in uMkhanyakude. Accordingly, the work in this thesis is centred on measurement of HIV at the district level in SSA.

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

### *The HIV/AIDS epidemic*

people (TGP) (Stevens et al. 2023). 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, though concentrated subepidemics do occur (Tanser et al. 2014), risk is more diffuse across the population. 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. Chapter 5 focuses on measurement of HIV for AGYW and FSW.

There are a number of ways to practically implement differentiated HIV treatment and prevention services (Godfrey-Faussett et al. 2022). These include geographic and demographic prioritisation (Meyer-Rath et al. 2018), key population services (Organization et al. 2022), and risk screening based on individual-level risk characteristics (Jia et al. 2022). Each approach requires strategic information about HIV disease burden. This thesis focuses on using HIV surveillance to inform geographic and demographic prioritisation.

## **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 be used to track epidemic indicators, identify at-risk populations, uncover drivers of transmission, implement prevention and treatment programs, and assess their impact. Important indicators to measure 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, and the number of PLHIV, can be caused either by new HIV infections or more PLHIV remaining alive by taking treatment. For this reason it is important to take caution in directly interpreting changes in HIV prevalence. Nonetheless, as a primary measure of population disease

## *The HIV/AIDS epidemic*

burden, HIV prevalence is crucial in calculating all of the other indicators given below.

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

$$I = \int \lambda_t \cdot (1 - \rho_t) \cdot N_t dt.$$

Planning, delivery, and evaluation of prevention programming relies on estimates of HIV incidence and the number of new HIV infections. Knowing whether the rate of new infections is rising or declining within specific populations is crucial.

- **ART coverage** is the proportion  $\alpha \in [0, 1]$  of PLHIV who are on ART. The number of people taking ART is given by  $N \cdot \rho \cdot \alpha$ . Estimates of ART coverage play a direct role in planning provision of treatment services, and finding unmet treatment need.
- **Recent infection** is the proportion  $\kappa \in [0, 1]$  of PLHIV who have been recently infected. Recency assays use biomarkers to distinguish between recent and longstanding infection, with varying sensitivity and specificity. Estimates of recent infection are primarily used to help estimate HIV incidence (Kassanjee et al. 2012; UNAIDS, WHO, et al. 2022).
- **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 estimates of awareness of HIV status. HIV diagnosis allows for linkage to care and progression along the HIV treatment cascade and care continuum (CDC 2014).

### 2.2.1 Data

Measuring the above HIV indicators requires data. To give the most complete picture of the epidemic, it is important to use multiple sources of data. The most prominent categories are:

- **Household surveys** are large, national, cross-sectional studies. The surveys conducted in the most countries are Demographic and Health Surveys [DHS ;USAID (2012)], which include a wide range of health related questions, and more HIV-specific Population-based HIV Impact Assessment [PHIA; ICAP (2023)] and AIDS Indicator Surveys (AIS). Some countries also implement their own survey series, such as the South Africa Behavioural, Sero-status and Media Impact Survey (SABSSM). Household surveys provide high quality standardised data about HIV, typically designed to furnish nationally-representative estimates. Both DHS and PHIA surveys collecting demographic, behavioural, and clinical information. Additionally, HIV testing is conducted via home-based testing, with results returned immediately, or anonymous dried blood spot testing.
- **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. Due to their integration with regular service delivery, programmatic data are available at higher frequency than other data sources. However, in comparison with designed studies, less control can be exercised over collection of programmatic data. It is common to encounter issues of data quality and reliability, as well as bias, 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 have particular use in informing otherwise difficult to estimate epidemiological parameters. Such parameters include

## *The HIV/AIDS epidemic*

disease progression and mortality rates, transmission dynamics, and treatment outcomes. Examples of population-based cohort studies in SSA include the Manicaland Project Open Cohort Study in Zimbabwe (Gregson et al. 2006), the Rakai Community Cohort Study in Uganda (Grabowski et al. 2017), and the Karonga Demographic Surveillance Site in Malawi (A. C. Crampin et al. 2012).

### **2.2.2 Challenges**

Obtaining reliable, timely estimates of the HIV indicators at an appropriate spatial resolution using the available data sources is challenging. The most significant difficulties faced are enumerated below, providing important context for the work in this thesis:

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. Furthermore, the sample sizes in household surveys are typically designed to be representative at a national-level. As a result, data for subpopulations are usually sparse.
2. **Missing data:** The sampling frame of a survey may not correspond to the target population. For example, some KPs are difficult to reach, and may be omitted from sampling frames (Jin et al. 2021). Additionally, individuals included on the sampling frame may choose not to respond. Each of these issues can be characterised as being problems of missing data.
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 of HIV status, but still may be subject to measurement errors. Biases in behavioural data can be more difficult to disentangle.

## *The HIV/AIDS epidemic*

4. **Denominators and demography:** Many indicators are rates or proportions, which rely on estimates of the population at risk in the denominator. For example, HIV prevalence is a proportion of the population, and HIV incidence is a rate per person-years at risk. Accurately estimating population denominators over space, time, and demographics 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 sources of data that are collected might vary across space and time. Additionally, reporting protocols or definitions for the same data source can also change. Though household surveys tend to be more consistent than programmatic data, the questions included and design of the surveys do 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 mitigate these limitations as effectively as possible. The most important statistical approaches for estimating HIV indicators used in this thesis are:

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. Chapter 4 discusses models for borrowing information over space. These

## *The HIV/AIDS epidemic*

models, along with others for borrowing information in other dimensions, are applied in Chapters 5 and 6.

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. Chapter 6 develops methods suitable for the complex statistical models required to integrate data sources. Multiple data sources are used in Chapter 5 to overcome the limitations of household surveys for measuring KP population sizes.
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. UNAIDS process.
4. **Uncertainty quantification:** Conclusions drawn by synthesising multiple incomplete data sources are unlikely to be firm and unanimous. It is therefore especially important that the uncertainties inherent to any statistical analysis are accurately and transparently presented. The Bayesian statistical paradigm introduced in Chapter 3 and used throughout this thesis is particularly well suited to handling of uncertainty.

# 3

## Bayesian spatio-temporal statistics

### 3.1 Bayesian statistics

Bayesian statistics is a mathematical paradigm for learning from data. Two reasons stand out as to why it is especially well suited to facing the challenges presented in Section 2.2. First, it allows for principled and flexible integration of prior domain knowledge. Second, 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  $\mathbf{y} = (y_1, \dots, y_n)$ . Parameters  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_d)$  are used to describe features of the data. Both the data and parameters are assumed to be random variables, and their joint probability distribution is written as  $p(\mathbf{y}, \boldsymbol{\phi})$ . Subsequent calculations, and the conclusions which follow from them, are made based on manipulating the model using probability theory.

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

$$p(\mathbf{y}, \boldsymbol{\phi}) = p(\mathbf{y} | \boldsymbol{\phi})p(\boldsymbol{\phi}). \quad (3.1)$$

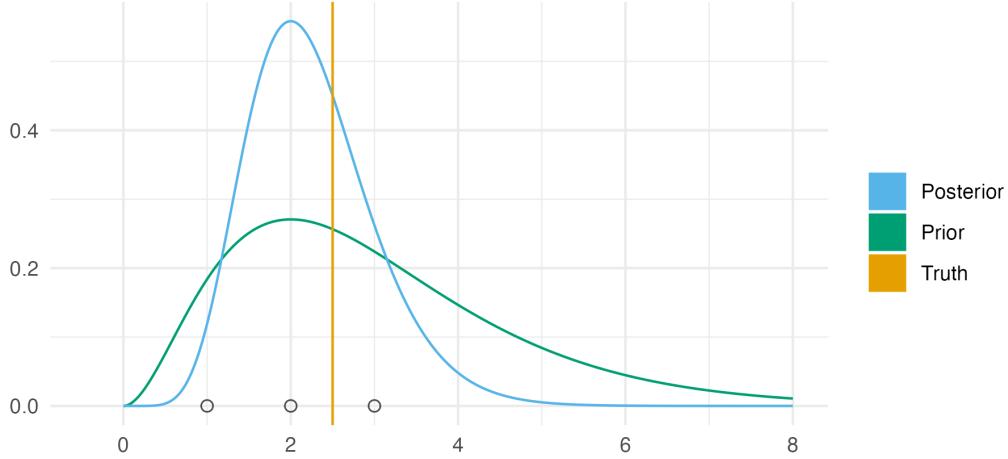
The likelihood, as a function of  $\boldsymbol{\phi}$  with  $\mathbf{y}$  fixed, reflects the probability of observing the data when the value of the parameters is  $\boldsymbol{\phi}$ . The prior distribution encapsulates beliefs about the parameters  $\boldsymbol{\phi}$  before the data are observed.

Recommendations for specifying prior distributions vary. The extent to which subjective information should be incorporated into the prior distribution is a central issue. 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, as we will in Section 3.3, 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 – that is,  $p(\mathbf{y}, \boldsymbol{\phi})$  – 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 samples of the data  $\mathbf{y}$  differ too greatly from what the analyst would expect to see in reality, then the model fails to 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  $p(\boldsymbol{\phi} | \mathbf{y})$ . 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 the 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  $\text{Poisson}(2.5)$ . As such, the true data generating process is within the space of models being considered. This situation is sometimes known (Bernardo and A. F. Smith 2001) as the  $\mathcal{M}$ -closed world, in contrast to the  $\mathcal{M}$ -open world where the model is said to be misspecified. Further, 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, Bayesian inference will typically be more computationally demanding than for conjugate models. The posterior distribution here is more tightly peaked than the prior distribution. Contraction of this kind is typical, 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

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

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

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

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

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

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. Broadly, computational strategies for approximating the posterior distribution (Martin et al. 2023) may be divided into Monte Carlo algorithms and deterministic approximations.

### 3.1.2.1 Monte Carlo algorithms

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

$$\boldsymbol{\phi}_s \sim p(\boldsymbol{\phi} | \mathbf{y}), \quad s \in 1, \dots, S. \quad (3.5)$$

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

$$\mathbb{E}(G | \mathbf{y}) = \int G(\boldsymbol{\phi})p(\boldsymbol{\phi} | \mathbf{y})d\boldsymbol{\phi} \approx \frac{1}{S} \sum_{s=1}^S G(\boldsymbol{\phi}_s), \quad (3.6)$$

using the samples from the posterior distribution in Equation (3.5). Most quantities of interest can be cast as posterior expectations, which may then be approximated empirically using samples in this way. Of course, it remains to discuss how the samples are obtained.

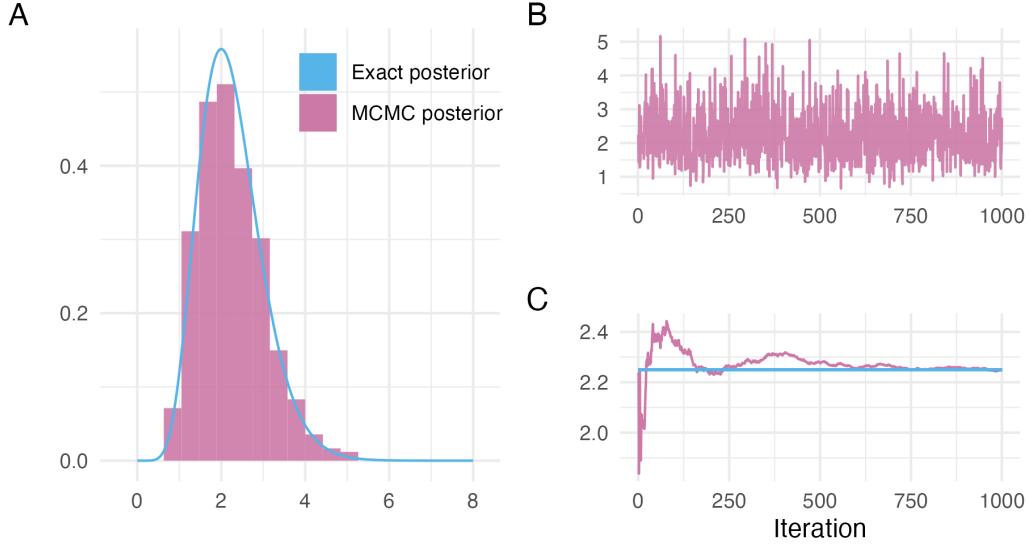
Markov chain Monte Carlo (MCMC) methods (Roberts and Rosenthal 2004) are the most popular class of sampling algorithms. Using MCMC, samples are generated

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  $q(\boldsymbol{\phi}_{s+1} | \boldsymbol{\phi}_s)$  to generate candidate parameters for the next step in the Markov chain. These candidate parameters are then accepted or rejected with some probability determined based on their log-posterior evaluation. Many MCMC algorithms, including the Gibbs sampler (S. Geman and D. Geman 1984), can be thought of as special cases of MH.

Other notable classes of sampling algorithms include importance sampling [IS; Tokdar and Kass (2010)] methods, which uses weighted samples, sequential Monte Carlo [SMC; Chopin, Papaspiliopoulos, et al. (2020)] methods, which are based on sampling from a sequence of distributions, and approximate Bayesian computation [ABC; Sisson et al. (2018)], which works by comparing simulated data to observed data, and does not require evaluation of the log-posterior. Though these methods have found applications in specific domains, MCMC is currently more widely used. The most important benefits of MCMC are its generality, theoretical reliability, and implementation in accessible software packages.

Illustrating the use of MCMC being supported by software, this thesis uses 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. Three tuning parameters control the behaviour of the HMC algorithm [Section 15.2; Stan Development Team (2023)]. NUTS automatically adapts these parameters 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 that diagnostic checks are used to evaluate convergence and assess whether the results of the Markov chain



**Figure 3.2:** NUTS can be used to sample from the posterior distribution described in 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, though it does not depend on tuning parameters like kernel density estimation. Other visualisations, such as empirical cumulative difference function plots, though less initially intuitive, are preferred for accurate distributional sample comparisons. Panel B is a traceplot showing the path of the Markov chain  $\{\phi_s\}_{s=1}^{1000}$  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 convergence of the empirical posterior mean  $\frac{1}{s} \sum_{l \leq s} \phi_l$  to the true value of  $\mathbb{E}(\phi)$  as more iterations of the Markov chain are included in the sum. In this case, the samples from NUTS are highly accurate in estimating this posterior expectation.

can be used to compute posterior quantities. It is never possible to be sure that results computed from MCMC will be accurate, though it is possible to know when they will be inaccurate. Panel 3.2B shows the traceplot for a Markov chain which has converged, and moves freely through the range of plausible parameter values. A range of convergence diagnostics have been developed for MCMC (Roy 2020; C. C. Margossian and Gelman 2023). Two widely used examples are the potential scale reduction factor  $\hat{R}$  (Gelman and Rubin 1992), which compares the variance between and within parallel Markov chains, and the effective sample size (ESS), which measures the efficiency of samples drawn from MCMC.

### 3.1.2.2 Deterministic approximations

The Monte Carlo methods discussed in Section 3.1.2.1 make use of stochasticity to generate samples from the posterior distribution. Deterministic approximations offer an alternative approach, often focused more directly on approximating the posterior distribution or posterior normalising constant. These approaches can be faster than Monte Carlo methods, especially for large datasets or models. That said, they lack strong theoretical guarantees of accuracy.

One prominent deterministic approximation is the Laplace approximation. It involves approximating the posterior normalising constant using Laplace’s method of integration. This is equivalent to approximating the posterior distribution by a Gaussian distribution. Numerical integration, or quadrature, is another deterministic approach in which the posterior normalising constant is approximated using a weighted sum of evaluations of the unnormalised posterior distribution. The integrated nested Laplace approximation [INLA; Håvard Rue, Martino, and Chopin (2009)] combines quadrature with the Laplace approximation. These methods are used throughout this thesis. In depth discussion is left to Chapter 6.

Variational inference [VI; Blei et al. (2017)] is another important deterministic approximation. The well-known expectation maximisation [EM; Dempster et al. (1977)] and expectation propagation [EP; Minka (2001)] algorithms are closely related to VI. In VI, 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 lacks theoretical guarantees and is known to often inaccurately estimate posterior variances (Giordano et al. 2018). As such, statisticians tend to approach VI with caution, despite its relative widespread acceptance within the machine learning community. Developing diagnostics to evaluate the accuracy of VI is an important area of ongoing research (Yao et al. 2018).

### 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, depending on the measure used, what can be argued to be the majority of cases. The analyst needs therefore not only to be concerned with choosing a model suitable for the data, but with choosing a model for which the posterior distribution may tractably be calculated in reasonable time. As such, there is an important interplay between modelling and computation, wherein models are bound by the limits of computation. As computational techniques and tools improve, the space of models available to the analyst expands. Exactly the focus of Chapter 6 is on expanding the space of models practically available to analysts.

## 3.2 Spatio-temporal statistics

Space and time are important features of infectious disease data, including those related to HIV. The field of spatio-temporal statistics (Cressie and Wikle 2015) is concerned with such observations, indexed by spatial and temporal location. It unifies the fields of spatial statistics (R. S. Bivand et al. 2008), concerned with observations indexed by space, and time series analysis (Shumway and Stoffer 2017), concerned with observations indexed by time. First, Section 3.2.1 characterises the shared properties of spatio-temporal data. Then, Section 3.2.2 describes how these properties facilitate the class of small-area estimation methods used in this thesis.

### 3.2.1 Properties of spatio-temporal data

Three important properties are discussed in this section: scale, correlation structure, and size.



**Figure 3.3:** In Panel A, the spatial location of Cape Town in South Africa can be considered a point, and the ZF Mgcau District Municipality (DM) can be considered as an area. In Panel B, World AIDS Day, designated on the 1st of December every year, can be considered a point in time, whereas 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, outside of mathematical abstraction, are rare.

### 3.2.1.1 Scale

The scale of spatio-temporal data refers to its extent and resolution. Its extent is the size of the spatial study region and length of time over which data was collected. Its resolution is how fine-grained those observations were.

In this thesis, the spatial study region  $\mathcal{S} \subseteq \mathbb{R}^2$  used is typically a country or collection of countries. It 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 spatial study region, illustrated in Panel A3.3. The temporal study period  $\mathcal{T} \subseteq \mathbb{R}$  can more generally be assumed to be one-dimensional. This feature, together with the fact that time only moves forward, is what distinguishes space and time. As with space, observations may be associated to a point  $t \in \mathcal{T}$  or period of time  $T \subseteq \mathcal{T}$ , illustrated in Panel B3.3.

The change-of-support problem (Gelfand et al. 2001) occurs when data is modelled at a scale different to the one it was observed at. For example, in this thesis, particularly Chapter 4, point data is modelled at an area-level. Special cases of the change-of-support problem include downscaling, upscaling, and dealing with

so-called misaligned data. It is also possible that spatio-temporal observations of the same process are made at multiple scales. Jointly modelling data at different scales simultaneously is another closely related challenge to the change-of-support problem.

### 3.2.1.2 Correlation structure

In “The Design of Experiments” Fisher (1936) observed that neighbouring crops were more likely to have similar yields than those far apart. This observation was later termed Tobler’s first law of geography: “everything is related to everything else, but near things are more related than distant things” (Tobler 1970). As well as space, Tobler’s first law applies to time, in that observations made close together in time tend to be similar.

This law can be formalised using space-time covariance functions, measuring the dependence of observations across their spatial and temporal dimensions. A 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). Spatial covariance functions are called isotropic when they apply equally in all directions, and stationary when they are invariant over space. Temporal covariance structures are often periodic, corresponding to daily, weekly, monthly, quarterly, or yearly cycles.

That spatio-temporal data are rarely IID is a statistically important point. The consequence is that it is rare to have true replicates available. Typically, only a single instance of a spatio-temporal can ever be realised.

### 3.2.1.3 Size

Data with both spatial and temporal dimensions are often large. For example, observations collected every week across a number of sites in a country can easily number in the thousands. Storage and mathematical operations with large spatio-temporal data can be challenging.

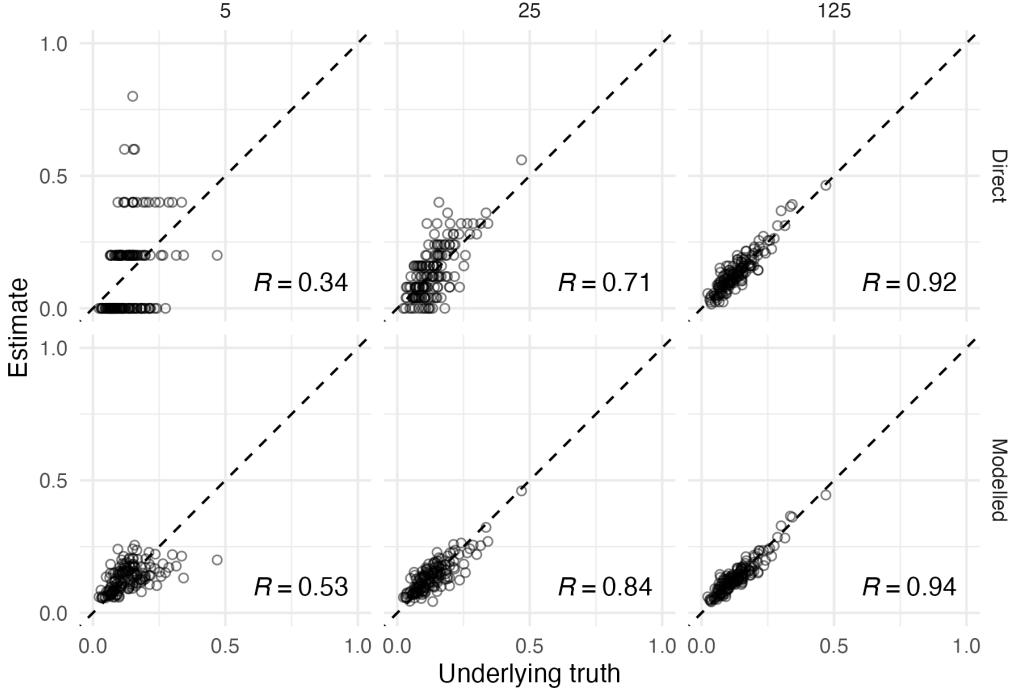


**Figure 3.4:** Simulation of a simple random sample  $y_i \sim \text{Bin}(m, p_i)$  with varying sample size  $m = 5, 25, 125$  in each of the  $i = 1, \dots, 156$  constituencies of Zambia. Direct estimates were obtained by the empirical ratio of data to sample size. Modelled estimates were obtained using a logistic regression with linear predictor given by an intercept and a spatial random effect. 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. The viridis colour palette, as implemented by the *viridis* R package (Garnier et al. 2023), was used in this figure. It is used often throughout this thesis because it is perceptually uniform and accessible to colourblind viewers (N. 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-un aids>.

Further, 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. In combination, large data (big  $n$ ) and models with a large number of parameters (big  $d$ ) make Bayesian inference, and other complex mathematical operations, challenging for spatio-temporal data.

### 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. In



**Figure 3.5:** The setting of this figure matches that of Figure 3.4. Estimates from surveys with higher sample size have higher sample 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.

spatio-temporal statistics, there are a large number of possible locations in space and time. Given the cost of data collection, 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. Prior knowledge that observations in one spatio-temporal location are correlated with those at another (Section 3.2.1.2) can be used to improve estimates.

Figures 3.4 and 3.5 illustrate the unreliability of direct estimates from small sample sizes, and the benefit of using a spatial model to overcome this limitation.

The effect is most pronounced for the sample size of 5, where the only possible direct estimates are 0, 0.2, 0.4, 0.6, 0.8 and 1. Using a spatial model to borrow information across space in this case results in improvement of the Pearson correlation coefficient between the estimates and the true underlying values from 0.34 to 0.53.

SAE methods are not only useful in the spatio-temporal setting. More generally, they apply in any situation where data are limited for subpopulations of interest. Just as these subpopulations can be generated by spatio-temporal variables, they can be generated by other variables. One such example is demographic variables. Analogous to spatio-temporal correlation structure, we also can often 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 socio-economic strata.

### 3.3 Model structure

The spatio-temporal data used in this thesis is not IID (Section 3.2.1.2). This section discusses ways to use statistical models to encode more complex relations between observations mathematically. Simple structures are discussed first, beginning with the linear model. Extensions are introduced one at a time, culminating in the model structures used throughout the thesis.

#### 3.3.1 Linear model

In a linear model, each observation  $y_i$  with  $i \in [n]$  is modelled using a Gaussian distribution

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

The conditional mean  $\mu_i$  is assumed to be linearly related to a collection of  $p$  covariates  $z_{1i}, \dots, z_{pi}$

$$\mu_i = \eta_i \quad (3.8)$$

$$\eta_i = \beta_0 + \sum_{l=1}^p \beta_l z_{li}. \quad (3.9)$$

Priors may be placed on the regression coefficients, as well as the observation standard deviation

$$\beta_l \sim p(\beta_l), \quad l = 0, \dots, p, \quad (3.10)$$

$$\sigma \sim p(\sigma). \quad (3.11)$$

While the linear model provides a useful foundation, its strong assumptions and limited flexibility call for careful use.

### 3.3.2 Generalised linear model

Generalised linear models (GLMs) extend the linear model by allowing the conditional mean  $\mu_i$  to be connected to the linear predictor  $\eta_i$  via a link function  $g$  as follows

$$y_i \sim p(y_i | \eta_i), \quad (3.12)$$

$$\mu_i = \mathbb{E}(y_i | \eta_i) = g(\eta_i). \quad (3.13)$$

The logistic function  $g(\eta) = \exp(\eta)/(1 + \exp(\eta))$  is commonly used as a link function to ensure that the conditional mean is in the range  $[0, 1]$ . Similarly, the exponential function  $g(\eta) = \exp(\eta)$  can be used to ensure the conditional mean is positive. The linear model is a special case of a GLM where the link function  $g$  is the identity. As well, GLMs admit a wider range of likelihoods  $p(y_i | \eta_i)$  than linear models, typically restricted to the so-called exponential family of distributions. The equation for the linear predictor is the same as the linear model case in Equation (3.9).

### 3.3.3 Generalised linear mixed effects model

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

$$\eta_i = \beta_0 + \sum_{l=1}^p \beta_l z_{li} + \sum_{k=1}^r u_k(w_{ki}). \quad (3.14)$$

The terms  $\beta_l$  are referred to as fixed effects. The terms  $u_k$  are called random effects, of additional covariates  $w_{ki}$ . The words fixed and random effects have

notoriously many different and incompatible definitions which unfortunately can cause confusion (Gelman 2005).

Random effects allow for more complex sharing of information between observations. To demonstrate this fact, first consider the model

$$\eta_i = \beta_0. \quad (3.15)$$

In this model all observations are assumed to be equivalent, and as such information is said to be completely pooled together. Second, consider the so-called no pooling model

$$\eta_i = \beta_0 + \beta_1 z_i, \quad (3.16)$$

with  $z_i \in \{0, 1\}$  a binary covariate. Now, there are two groups of observations, each of which with its own mean:  $\beta_0$  for the first group and  $\beta_0 + \beta_1$  for the second. No amount of information is shared between the two groups. Finally, consider an intermediate between these two extremes, known as the partial pooling model. In the partial pooling model, the extent to which information is shared between groups is learnt rather than fixed to either extreme at the outset, as with the complete or no pooling models. The parameter  $\beta_0$  applies to all groups, and each group is differentiated by a specific value of the random effects  $u_i$ .

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 encode smoothness in space or time. In contrast, unstructured random effects treat groups of observations as being exchangeable.

Generalised additive models [GAMs; Wood (2017); Hastie and Tibshirani (1987)] are another class of models which extend GLMs. Though GAMs place more of a focus on using  $u_k$  to model non-linear relationships between covariates and the response variable, they can also be cast to fit into the GLMM framework.

### 3.3.4 Latent Gaussian model

Latent Gaussian models [LGMs; Håvard Rue, Martino, and Chopin (2009)] are a type of GLMMs in which Gaussian priors are used for many of the models parameters. More specifically, these parameters are  $\beta_0$ ,  $\{\beta_j\}$ ,  $\{u_k(\cdot)\}$ , and can be collected into a vector  $\mathbf{x} \in \mathbb{R}^N$  called the latent field. The Gaussian prior distribution is then

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}(\boldsymbol{\theta}_2)^{-1}), \quad (3.17)$$

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  $p(\boldsymbol{\theta})$ . The posterior distribution under an LGM is then

$$p(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) \propto p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta}), \quad (3.18)$$

with the complete set of parameters  $\boldsymbol{\phi} = (\mathbf{x}, \boldsymbol{\theta})$ , and  $N + m = d$ .

### 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

$$\boldsymbol{\eta} = (\eta_1, \dots, \eta_{N_n}), \quad (3.19)$$

where  $N_n \in \mathbb{N}$  is a function of  $n$ . Unlike in the LGM case, it is possible that  $N_n \neq n$ . Each mean response  $\mu_i$  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  $g_i : \mathbb{R}^{|\mathcal{J}_i|} \rightarrow \mathbb{R}$ , such that  $\mu_i = g_i(\boldsymbol{\eta}_{\mathcal{J}_i})$ . Put together, ELGMs are of the form

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

The latent field and hyperparameter prior distributions are equivalent to the LGM case.

Though the ELGM model class was only introduced recently, it connects much of the work done in this 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 have ELGM-like features.

### 3.4 Model comparison

Many models can be fit to the same data during the course of an analysis. Model comparison methods are used to determine which is the most suitable for use. This section focuses on measuring suitability via the model's predictive performance (Vehtari and Ojanen 2012).

Ideally, new data  $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_n)$  drawn from the true data generating process would be available to test predictive performance. The log predictive density for new data (LPD) (Gelman, Hwang, et al. 2014) is one measure of out-of-sample predictive performance given by

$$\text{lpd} = \sum_{i=1}^n \log p(\tilde{y}_i | \mathbf{y}) = \sum_{i=1}^n \log p(\tilde{y}_i | \boldsymbol{\phi}) p(\boldsymbol{\phi} | \mathbf{y}) d\boldsymbol{\phi}. \quad (3.20)$$

The expected LPD (ELPD) integrates the LPD over the data generating process to give a measure of expected performance

$$\text{elpd} = \sum_{i=1}^n \log \int p(\tilde{y}_i | \mathbf{y}) p(\tilde{y}_i) d\tilde{y}_i. \quad (3.21)$$

In reality, such data are not usually available, and instead the ELPD must be approximated using the available data.

### 3.4.1 Information criteria

Information criteria can be constructed to approximate the ELPD using adjusted within-sample predictive performance. The Akaike [AIC; Akaike (1973)] and deviance [DIC; D. J. Spiegelhalter et al. (2002)] information criteria estimate ELPD by

$$\text{elpd}_{\text{IC}} = \log p(\mathbf{y} | \hat{\boldsymbol{\phi}}), \quad (3.22)$$

where  $\hat{\boldsymbol{\phi}}$  is a maximum likelihood estimate (AIC) or Bayesian point estimate (DIC). The widely applicable information criteria [WAIC; Watanabe (2013)] improves upon Equation (3.22) by instead using the predictive density of the data

$$\text{elpd}_{\text{WAIC}} = \sum_{i=1}^n \log p(y_i | \mathbf{y}). \quad (3.23)$$

As both Equations (3.22) and (3.23) are based on within-sample measures, they overestimate the ELPD. As such, they are adjusted downward by a complexity penalty  $p_{\text{IC}}$ . The particular penalty varies depending on the particular information criteria.

### 3.4.2 Cross-validation

Cross-validation is an alternative way to estimate the ELPD. Rather than use a complexity penalty, as in Section 3.4.1, to adjust a within-sample estimate, cross-validation (CV) partitions the data into training and held-out sets of data. For example, in a leave-out-out (LOO) CV there are  $n$  partitions, where each held-out set is a single observation. The LOO-CV estimate of ELPD is

$$\text{elpd}_{\text{LOO-CV}} = \sum_{i=1}^n \log p(y_i | \mathbf{y}_{-i}), \quad (3.24)$$

where the subscript  $-i$  refers to all elements of the vector excluding  $i$ . Naively, computing  $\text{elpd}_{\text{LOO-CV}}$  requires refitting the model  $n$  times. This can be computationally costly, and so approximation strategies have been developed. Importance sampling methods using the full posterior as a proposal are a notable example, including Pareto-smoothed importance sampling [PSIS; Vehtari, Gelman, et al. (2017)].

### 3.4.3 Scoring rules

Scoring rules [SR; Gneiting and Raftery (2007)] measure the quality of probabilistic forecasts. The log score, used above in the ELPD, is one example of a scoring rule. However, it is by no means the only possibility. Any information criterion (Section 3.4.1) or cross-validation strategy (Section 3.4.2) can be redefined using a different scoring rule, or utility function more broadly. Possible examples include the root mean square error (RMSE), variance explained ( $R^2$ ) or classification accuracy.

The log score (LS) is popular, in part because it is an example of a strictly proper scoring rule (SPSR). A scoring rule is strictly proper when the forecaster gains maximum expected reward by reporting their true probability distribution. Any scoring rule which does not admit this property is susceptible to manipulation, in some sense. The continuous ranked probability score [CRPS; Matheson and Winkler (1976)], which generalises the Brier score (Brier 1950) beyond binary classification, is another example of a SPSR. Ideally, the correct scoring rule to use in an analysis should be determined based upon the application setting.

### 3.4.4 Bayes factors

Finally, the evidence  $p(\mathbf{y})$ , given in Equation (3.3), can also be used as a measure of model performance. If  $\mathcal{M}_0$  and  $\mathcal{M}_1$  are two competing models, then the Bayes factor comparing  $\mathcal{M}_0$  to  $\mathcal{M}_\infty$  is

$$B_{01} = \frac{p(\mathbf{y} \mid \mathcal{M}_0)}{p(\mathbf{y} \mid \mathcal{M}_1)}, \quad (3.25)$$

where  $p(\mathbf{y} \mid \mathcal{M})$  denotes the evidence under model  $\mathcal{M}$ . The Bayes factor can be interpreted as supporting the maximum a posteriori model. If  $B_{01} > 1$  then support is provided for  $\mathcal{M}_0$  and if  $B_{01} < 1$  then support is provided for  $\mathcal{M}_1$ . Bayes factors can also be framed as predictive criteria according to the decomposition

$$p(\mathbf{y}) = p(y_1)p(y_2 \mid y_1) \cdots p(y_n \mid y_{n-1}, \dots, y_1). \quad (3.26)$$

## 3.5 Survey methods

Large national household surveys (Section 2.2.1) provide the highest quality population-level information about HIV indicators in SSA. Demographic and Health Surveys [DHS; USAID (2012)] 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.

Analysis of responses from survey methods can require specific methods. This section provides required background, before describing the survey design approach used by household surveys in SSA, and the methods used to analyse this data in this thesis.

### 3.5.1 Background

Consider a population of  $N$  individuals, indexed by  $i$ , with outcomes of interest  $y_i$ . If a census were run, with all responses recorded, then any population quantities of interest could be directly calculated. However, running a census is usually too expensive or otherwise impractical. As such, in a survey only a subset of individuals are sampled: let  $S_i$  be an indicator for whether or not individual  $i$  is sampled. Furthermore, only a subset of those sampled have their outcome recorded, due to nonresponse or otherwise: let  $R_i$  be an indicator for whether or not  $y_i$  is recorded. If  $S_i = 0$  then  $R_i = 0$ , and if  $S_i = 1$  then individual  $i$  may not respond such that  $R_i = 0$ . Consider a function  $G_i = G(y_i)$ . The population mean of  $G$  is

$$\bar{G} = \frac{1}{N} \sum_{i=1}^N G(y_i), \quad (3.27)$$

and a direct estimate of  $\bar{G}$  based on the recorded subset of the population is

$$\bar{G}_R = \frac{\sum_{i=1}^N R_i G(y_i)}{\sum_{i=1}^N R_i}, \quad (3.28)$$

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

In a probability sample, individuals are selected to be included in the survey at random. On the other hand, 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  $\mathbb{P}(S_i = 1) = 1/N$ . A survey design is called complex when the sampling probabilities for each individual vary, such that  $\mathbb{P}(S_i = 1) = \pi_i$ , 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, 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 under SRS, nonresponse analogous bias can be caused by non-response.

### 3.5.2 Survey design

The DHS employs a two-stage sampling procedure, outlined here following USAID (2012). In the first stage, enumeration areas from a recently conducted census are typically used as the primary sampling unit, or cluster. Each cluster is assigned to a strata  $h$  by region, as well as by urban-rural status. After appropriate strata sample sizes  $n_h$  are determined, EAs are sampled with probability proportional to number of households

$$\pi_{1hj} = n_h \times \frac{N_{hj}}{\sum_j N_{hj}}, \quad (3.29)$$

where  $N_{hj}$  is the number of households in strata  $h$  and cluster  $j$ . In the second stage, the secondary sampling units are households. All households in the selected cluster are listed, before being sampled systematically at a regular interval, with equal probability

$$\pi_{2hj} = \frac{n_{hj}}{N_{hj}}, \quad (3.30)$$

where  $n_{hj}$  is the number of households selected in cluster  $j$  and stratum  $h$ . All adults are interviewed in each selected household. As a result, the probability an individual is sampled is equal to the probability their household is sampled  $\pi_{hj} = \pi_{1hj} \times \pi_{2hj}$ .

### 3.5.3 Survey analysis

Suppose a survey is run with complex design, and sampling probabilities  $\pi_i$ . Some individuals are more likely to be included in the survey than others. By over-weighting the responses of those unlikely to be included, and under-weighting the responses of those likely to be included, this feature can be taken into account. Design weights  $\delta_i = 1/\pi_i$  can be thought of as the number of individuals in the population represented by the  $i$ th sampled individual. Let

$$\mathbb{P}(R_i = 1 \mid S_i = 1) = v_i \quad (3.31)$$

be the probability of response for sampled individual  $i$ . Nonresponse can be handled using nonresponse weights  $\gamma_i = 1/v_i$ , which analogously can be thought of as the number of sampled individuals represented by the  $i$ th recorded individual. Multiplying the design and nonresponse weights gives survey weights  $\omega_i = \delta_i \times \gamma_i$ .

Extending Equation (3.28), a weighted estimate (Hájek 1971) of the population mean using the survey weights  $\omega_i$  is

$$\bar{G}_\omega = \frac{\sum_{i=1}^N \omega_i R_i G(y_i)}{\sum_{i=1}^N \omega_i R_i}. \quad (3.32)$$

Following Meng (2018) and Bradley et al. (2021), decomposing the additive error  $\bar{G}_\omega - \bar{G}$  of Equation (3.32) provides useful intuition as to the benefits of survey weighting (M. A. Bailey 2023). Under SRS then, the error is a product of three terms

$$\bar{G}_\omega - \bar{G} = \frac{\mathbb{E}(\omega_i R_i G_i)}{\mathbb{E}(\omega_i R_i)} - \mathbb{E}(G_i) = \frac{\mathbb{C}(\omega_i R_i G_i)}{\mathbb{E}(\omega_i R_i)} \quad (3.33)$$

$$= \rho_{R_\omega, G} \times \sqrt{\frac{N - m_{R_\omega}}{m_{R_\omega}}} \times \sigma_G, \quad (3.34)$$

where  $R_\omega = \omega R$ . The first term is called the data defect correlation (DDC), and measures the correlation between the weighted recording mechanism and given function of the outcome of interest. The DDC is minimised when  $G \perp\!\!\!\perp R_\omega$ . The second term is the data scarcity, and measures the effective proportion of the population who have been recorded. Finally, the third term is the problem

## *Bayesian spatio-temporal statistics*

difficultly, and measures the intrinsic difficulty of the estimation problem. This term is independent of the sampling or analysis method used.

This thesis uses hierarchical Bayesian models defined using weighted direct survey estimates (Fay and Herriot 1979). Following C. Chen et al. (2014), the sampling distribution of these direct estimates is arrived at using by estimating the variance of Equation (3.32). Although this approach acknowledges the complex survey design, it has some important limitations. Importantly, it ignores clustering structure within the observations  $i$ . Furthermore, as a two-step procedure, it fails to fully propagate uncertainty from a Bayesian perspective. While progress has been made in dealing with survey data, the Gelman (2007) claim that “survey weighting is a mess” still holds some weight.

# 4

## Models for areal spatial structure

This chapter is about spatial random effect model specifications for areal data. A simple model based on the adjacency structure between areas is popular in HIV small-area estimation and beyond. The analysis aimed to determine if using a more complex model would result in more accurate predictions.

Modelling spatial correlation is particularly important for the small-area estimation of HIV. This is because the covariates which are most strongly associated with HIV, such as sexual behaviour and STI status (Benjamin K. Mayala et al. 2020), are themselves difficult to measure. As a result, previous small-area models of HIV have found including covariates only modestly improve predictive performance (Supplementary Figure 20, Dwyer-Lindgren, Cork, et al. 2019). The lack of predictive covariates emphasises 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 directly modelling spatial correlation is less important (Daniel J Weiss et al. 2015; Bhatt et al. 2015).

Spatial variation in areal data is 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;

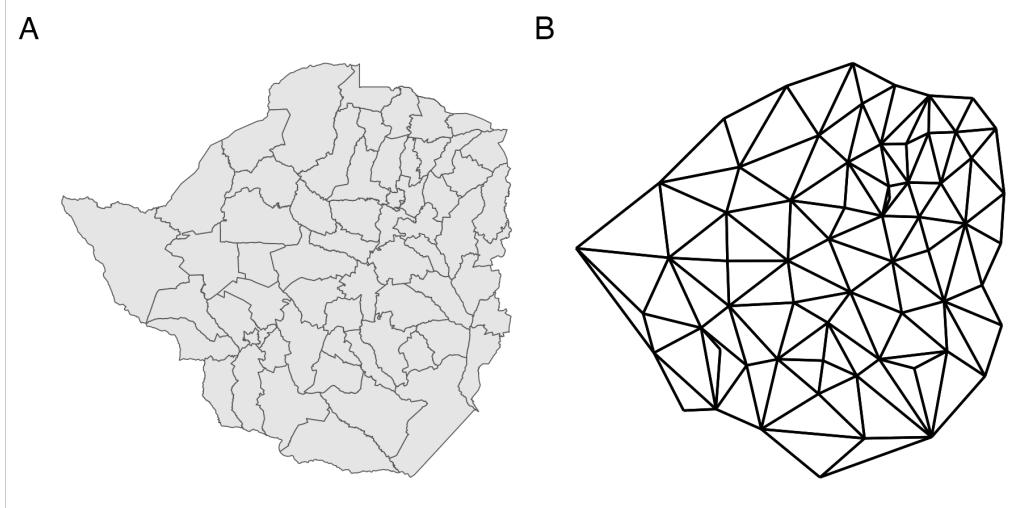
## *Models for spatial structure*

Havard Rue and Held (2005)]. These models combine a Gaussian distribution with Markov conditional independence assumptions between areas. Observations in areas close together are assumed to be related, with more distant relationships not directly accounted for. 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 in software such as **R-INLA** (Blangiardo et al. 2013), **rstan** (M. Morris et al. 2019; Donegan 2022), **NIMBLE** [Chapter 9; de Valpine et al. (2023)] and **PyMC** (Saunders 2023), among others. As a result, it has been widely used, including to model bird population dynamics from capture-recapture data (Saracco et al. 2010); for the analysis of magnetic resonance images (Gössl et al. 2001; Schmid et al. 2006); to map mortality from cancers (Rashid et al. 2023), injuries (Parks et al. 2020), and air pollution (Bennett et al. 2019); and to model alcohol use patterns (Dwyer-Lindgren, A. D. Flaxman, et al. 2015).

The Besag model was designed for 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 improves the performance of small-area estimation models. Performance in this context refers to accurate forecasts of parameters as measured by scoring rules. In doing so, practical recommendations for modelling areal spatial structure are offered. Code for the analysis in this chapter is available from <https://github.com/athowes/beyond-borders>, and supported by the **arealutils** R package (Howes 2023a).

## **4.1 Models based on adjacency**

This section discusses spatial random effect models based on a symmetric adjacency relation  $i \sim j$  between areas  $A_i$  and  $A_j$ . Adjacency is typically defined by a shared



**Figure 4.1:** Panel A shows the districts of Zimbabwe. Panel B shows the corresponding adjacency graph  $\mathcal{G}$  with vertices positioned at the centre of the area they correspond to, and edges between adjacent areas.

border, though other choices are possible (Christopher J Paciorek et al. 2013).

#### 4.1.1 The Besag model

The Besag model (Besag et al. 1991) is an improper conditional auto-regressive (ICAR) model where the conditional mean of the random effect  $u_i$  is the average of its neighbours  $\{u_j\}_{j \sim i}$  and the precision is proportional to the number of neighbours. The full conditional distribution of the  $i$ th spatial random effect is given by

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

where  $\delta i$  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  $i$ th entry removed. By Brook's lemma (Havard Rue and Held 2005) the set of full conditionals of the Besag model are equivalent to the Gaussian Markov random field (GMRF) given by

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

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

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

## Models for spatial structure

The Markov property arises due to the conditional independence structure  $p(u_i | \mathbf{u}_{-i}) = p(u_i | \mathbf{u}_{\delta i})$  whereby each area only depends on its neighbours. This is reflected in the sparsity of  $\mathbf{R}$  such that  $u_i \perp u_j | \mathbf{u}_{-ij}$  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 districts of Zimbabwe with corresponding adjacency graph.

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

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

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  $I_1, \dots, I_{n_c}$ , 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  $I = I_1, \dots, I_{n_c}$ .

### 4.1.2 Best practises for the Besag model

Directly implementing the Besag model as described in Section 4.1.1 is recommended against. Freni-Stabantino et al. (2018) provide three best practices:

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

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

### Models for spatial structure

The scaled structure matrix  $\mathbf{R}^*$  is given by

$$\mathbf{R}^* = \mathbf{R}/\sigma_{\text{GV}}^2(\mathbf{R}). \quad (4.6)$$

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 Sigrunn Holbek Sørbye and Håvard Rue (2014), ignores off-diagonal entries. More broadly, other measures of typical variance could be used.

Scaling mitigates the influence of the adjacency graph on the variance of  $\mathbf{u}$ . For consistent and interpretable prior distribution selection, it is important to allow the variance to be controlled by  $\tau_u$  alone.

When the adjacency graph is disconnected it is not appropriate to scale the structure matrix  $\mathbf{R}$  uniformly. This is because, given the precision  $\tau_u$ , local smoothing operates on each connected component independently. As such, each connected component  $I = I_1, \dots, I_{n_c}$  should be scaled independently to have generalised variance one

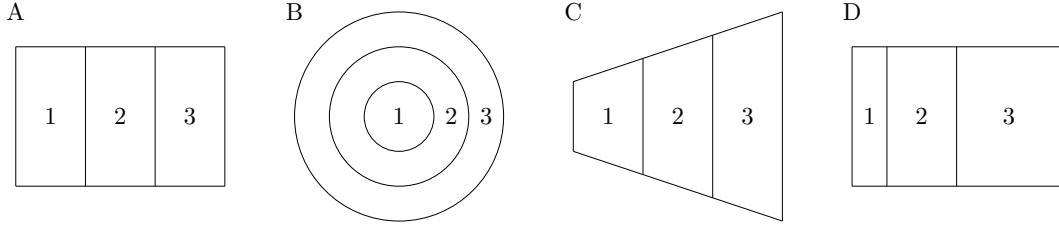
$$\mathbf{R}_I^* = \mathbf{R}_I/\sigma_{\text{GV}}^2(\mathbf{R}_I) \quad (4.7)$$

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

$$p(u_i) \propto \exp\left(-\frac{\tau_u}{2} \sum_{i \sim j} (u_i - u_j)^2\right) \propto 1 \quad (4.8)$$

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



**Figure 4.2:** Though they are quite different, the geometries shown in panels A, B, C, and D each have the same adjacency graph. Therefore, each geometries would have the same distribution under the Besag model.

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,

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

As such, in total the number of sum-to-zero constraints equals to the number of non-singleton connected components.

### 4.1.3 Concerns about the Besag model

The Besag model was originally proposed by Besag et al. (1991) for use in image analysis. In this setting, areas correspond to pixels arranged in a regular lattice structure. Since then, it has seen wider use. In some situations, like small-area estimation of HIV, the spatial structure is less regular than a lattice. This raises concerns about the Besag model's applicability to this broader setting. The discussion in this section is closely linked to:

- the modifiable areal unit problem (Openshaw and P. Taylor 1979), whereby statistical conclusions change as a result of seemingly arbitrary changes in data aggregation;
- the challenge of ecological inference and the ecological fallacy (Jonathan Wakefield and Lyons 2010).

#### 4.1.3.1 Compression to adjacency

A fundamental objection is that summarising a geometry by an adjacency graph represents a loss of information. Many geometries share the same adjacency graph, and as such are isomorphic under the Besag model (Figure 4.2). Though not in itself a problem, this fact prompts consideration as to whether the class of geometries with the same adjacency graph is sufficiently similar to merit identical models.

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

The regularity of realistic geometries may help to constrain each class to be similar. In other words, although pathological geometries can be constructed, they might be implausible in statistical practice and so of limited concern.

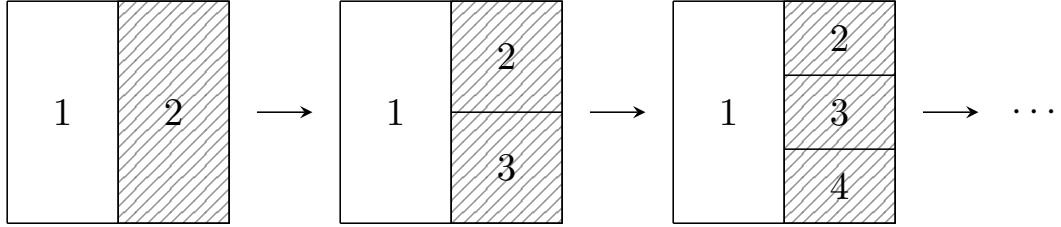
#### 4.1.3.2 Mean structure

In the Besag model all adjacent areas count equally in the equation for the conditional mean. This assumption is unsatisfying: for most geometries, we expect different amounts of correlation between neighbouring areas. Figure 4.2 illustrates a number of heuristic features for neighbour importance. In Panel 4.2C, the area with a longer shared border would be expected to be more highly correlated. In Panel 4.2D, the area with a closer centre would be expected to be more highly correlated.

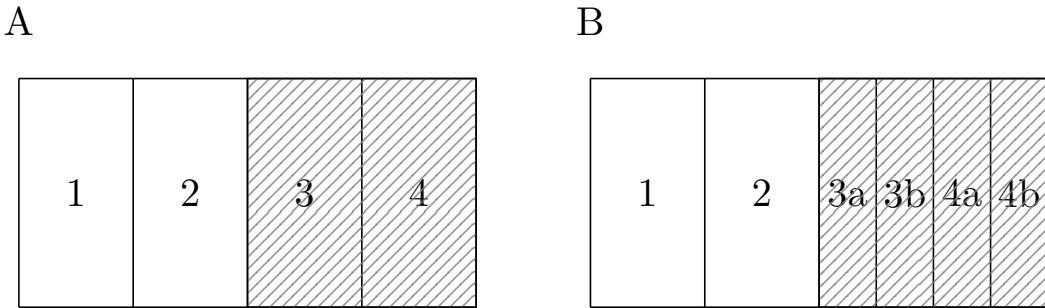
#### 4.1.3.3 Variance structure

In Equation (4.1) the precision of  $u_i$  is proportional to its number of neighbours  $n_{\delta i}$ . It follows that as  $n_{\delta i} \rightarrow \infty$  then  $\text{Var}(u_i) \rightarrow 0$ . This is illustrated by Figure 4.3 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

*Models for spatial structure*



**Figure 4.3:** A sequence of geometries where the number of neighbours of area one grows by one at each iteration, as the shaded area is split into more areas. In the limit, the precision of the spatial random effect in the first area tends to infinity. This is not reasonable behaviour if the amount of information being shared is not also increasing.



**Figure 4.4:** Each of the shaded areas in the geometry in Panel A are split into two in Panel B.

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.

As a second example of undesirable behaviour, suppose we fit a Besag model upon identical data using each of the two geometries in Figure 4.4. If the spatial variation is relatively smooth, dividing the shaded areas into two will result in a lower estimated variance  $\sigma_u^2$  in Panel 4.4B as compared with Panel 4.4A because there will appear to be less variation between neighbouring areas. This problem does not only apply locally: since the effect of  $\sigma_u^2$  applies everywhere, the smoothing will change even in unaltered parts of the study region.

#### 4.1.4 Weighted ICAR models

The Besag model is a special case of a more general class of (zero-mean) weighted ICAR models. These models can be specified in terms of scaled weights  $\{b_{ij}\}_{j \sim i}$  and a precision vector  $\boldsymbol{\kappa} = (\kappa_i)_{i \in [n]}$ . The full conditionals are then

$$u_i | \mathbf{u}_{-i} \sim \mathcal{N} \left( \sum_{j:j \sim i} b_{ij} u_j, \frac{1}{\kappa_i \tau_u} \right). \quad (4.10)$$

Setting  $b_{ij} = 1/n_{\delta i}$  and  $\kappa_i = n_{\delta i}$  recovers the Besag model in Equation (4.1). The structure matrix  $\mathbf{R}$  corresponding to the more general full conditionals in Equation (4.10) is

$$\mathbf{R} = \mathbf{D}_\kappa(\mathbf{I} - \mathbf{B}), \quad (4.11)$$

where the unscaled weights matrix  $\mathbf{B}$  has elements

$$\mathbf{B}_{ij} = \begin{cases} b_{ij}, & \text{for } i \sim j, \\ 0, & \text{for } i = j, i \not\sim j. \end{cases}, \quad (4.12)$$

and the matrix  $\mathbf{D}_\kappa$  is given by  $\text{diag}(\kappa_1, \dots, \kappa_n)$ . Ensuring that the structure matrix is symmetric requires that for all  $i, j \in [n]$

$$-b_{ij}\kappa_i = -b_{ji}\kappa_j. \quad (4.13)$$

To meet this condition, it can be simpler to directly consider symmetry of the unscaled weights matrix

$$\mathbf{W} = \mathbf{D}_\kappa \mathbf{B}, \quad (4.14)$$

such that  $\mathbf{R} = \mathbf{D}_\kappa - \mathbf{W}$ . For the Besag model the unscaled weights matrix  $\mathbf{W}$  corresponds to the adjacency matrix. Scaled weights can be recovered by  $b_{ij} = w_{ij}/\kappa_i$  where  $\kappa_i = \sum_{k:k \sim i} w_{ik}$ . Duncan et al. (2017) provide discussion of methods for specifying  $\mathbf{W}$ , including

$$w_{ij} = \left( \frac{1}{d_{ij}} \right), \quad (4.15)$$

$$w_{ij} = \exp(-d_{ij}). \quad (4.16)$$

Weighted ICAR models appear to overcome some of the limitations discussed in Section 4.1.3.

#### 4.1.5 The reparameterised Besag-York-Mollié model

Often, as well as spatial correlation, 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 component has its own respective precision parameter  $\tau_v$  and  $\tau_w$ . The resulting distribution is

$$\mathbf{u} \sim \mathcal{N}(0, \tau_v^{-1}\mathbf{I} + \tau_w^{-1}\mathbf{R}^-). \quad (4.17)$$

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  $\tau_v$  and  $\tau_w$ .

However, in the BYM model, 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 distribution

$$(\tau_v, \tau_w) \sim p(\tau_v, \tau_w) \quad (4.18)$$

which does not privilege either component is more easily accomplished if  $\mathbf{Q}$  and  $\mathbf{I}$  have the same scale. Additionally, supposing one has 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, without inadvertently altering the prior distribution on the amount of overall variation. This highlights identifiability issues of the parameters  $(\tau_v, \tau_w)$  resulting from them being non-orthogonal.

Building on the models of Leroux et al. (2000) and C. 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. This is known as the BYM2 model and given by

$$\mathbf{u} = \frac{1}{\tau_u} \left( \sqrt{1-\phi} \mathbf{v} + \sqrt{\phi} \mathbf{w}^* \right), \quad (4.19)$$

where  $\tau_u$  is the marginal precision of  $\mathbf{u}$ ,  $\phi \in [0, 1]$  gives the proportion of the marginal variance explained by each component, and  $\mathbf{w}^*$  is a scaled version of  $\mathbf{w}$  with precision matrix given by the scaled structure matrix  $\mathbf{R}^*$ . When  $\phi = 0$  the random effects are IID, and when  $\phi = 1$  the random effects follow the Besag model. To borrow an analogy (Håvard Rue 2020) the parameterisation  $(\tau_v, \tau_w)$  is like having one hot water and one cold water tap, whereas the parameterisation  $(\tau_u, \phi)$  is like a mixer tap where the amount of water and its temperature can be adjusted separately.

Although the BYM and BYM2 models were originally proposed using the Besag model as spatially structured component, this need not be the case. Indeed, more broadly it is reasonable to consider convolution random effects (of a form analogous to that in Equation (4.17) or (4.19)) with any model for spatially structured noise. Any limitations of the model for spatially structured random effects are inherited by the convolution random effects.

## 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. Such a function may be specified as

$$K : \mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S}) \rightarrow \mathbb{R}, \quad (4.20)$$

where  $\mathcal{P}$  denotes the power set such that  $\mathcal{P}(\mathcal{S})$  is the space of subsets of the study region. If the function  $K$  is positive semi-definite, then define areal kernel spatial random effects by

$$\mathbf{u} \sim \mathcal{N}\left(0, \frac{1}{\tau_u} \mathbf{K}\right), \quad (4.21)$$

where the  $n \times n$  Gram matrix  $\mathbf{K}$  with entries  $K_{ij} = K(A_i, A_j)$  is a valid covariance matrix. The precision parameter  $\tau_u$  is placed outside of the Gram matrix, analogous to the relation of the precision and structure matrices, but could be omitted. Areal kernels may be thought of as a type of kernels on sets (Gärtner et al. 2002).

## *Models for spatial structure*

It is challenging to think directly about the correlation structure between areas. Instead, most well-known spatial process models define the correlation structure between points using a kernel function

$$k : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}. \quad (4.22)$$

A simple method, and the one considered here henceforth, is to construct  $K$  (Equation (4.20)) from  $k$  (Equation (4.22)) by averaging the kernel  $k$  computed on some number of points representing each area. In Section 4.2.1 one point is used, and in Section 4.2.2 multiple points are used.

### 4.2.1 Centroid kernel

The simplest approach is to use a single point to represent each area such that

$$K(A_i, A_j) = k(p_i, p_j). \quad (4.23)$$

A natural choice is the centroid  $p_i = c_i$ , given by the arithmetic mean of the latitude and longitude. (Note that it is not guaranteed for the centroid to lie within the area i.e. it is possible  $c_i \notin A_i$ , and more generally points representing an area may not be contained by that area.) This choice results in the centroid kernel

$$K(A_i, A_j) = k(c_i, c_j). \quad (4.24)$$

The centroid kernel has been used in environmental epidemiology (Wakefield and S. Morris 1999), for US election modelling (S. R. Flaxman et al. 2015), and to model the reproduction number of COVID-19 (Teh et al. 2022). In a model comparison study Nicky Best et al. (2005) (Section 3) 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 unsurprising that a stationary covariance function would struggle to recover non-stationary structure.

### 4.2.2 Integrated kernel

Rather than choosing a single representative point, an alternative is to more completely represent the area by integrating the kernel over the areas of interest (Kelsall and Jonathan Wakefield 2002; Follestad and Håvard Rue 2003). This results in the integrated kernel

$$K(A_i, A_j) = \frac{1}{|A_i||A_j|} \int_{A_i} \int_{A_j} k(s, s') ds ds'. \quad (4.25)$$

Unlike for the centroid kernel where  $K_{ii} = 1$  for all  $i$ , the marginal variance of the  $i$ th 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.

This covariance structure is equivalent to that obtained by aggregating a spatially continuous Gaussian process with kernel  $k$  over the areal partition. In the machine learning literature, models of this kind have been studied under the name aggregated Gaussian processes (Law et al. 2018; Tanaka et al. 2019; Yousefi et al. 2019; Hamelijnck et al. 2019; Chau et al. 2021). Examples of use of this model in statistical practice are rare.

#### 4.2.2.1 Accounting for heterogeneity

Additional information accounting for heterogeneity over  $A_i$  may be incorporated into the integrated kernel. This can be accomplished using weighting distributions  $\{w_i\}$  which represent an unequal contribution of each point to the similarity measure. The weighted integrated kernel is given by

$$K(A_i, A_j) = \frac{1}{|A_i||A_j|} \int_{A_i} \int_{A_j} w_i(s) w_j(s') k(s, s') ds ds', \quad (4.26)$$

This areal kernel may be useful in disease mapping. For example, areas with populations who live close to a shared border are likely to be more strongly correlated than areas whose populations live far apart. This detail could be accounted for by weighting according to a high resolution measure of population density. Though e.g. weighted centroids may also be used in Equation (4.24),



**Figure 4.5:** The  $n = 33$  districts of Malawi. Panel A shows the centroids as in Section 4.2.1. Panel B shows  $L_i = 10$  randomly chosen points, Panel C hexagonal points, and Panel D grid points in each area, each generated using the `sf::st_sample` function (E. Pebesma 2018).

accounting for heterogeneity over an area is more natural within the integrated kernel than the centroid kernel.

#### 4.2.2.2 Computation

Most of the time it is not possible to calculate Equation (4.26) analytically. Instead, consider  $n$  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

$$K(A_i, A_j) \approx \frac{1}{L_i L_j} \sum_{l=1}^{L_i} \sum_{m=1}^{L_j} w_i(s_l^{(i)}) w_j(s_m^{(j)}) k(s_l^{(i)}, s_m^{(j)}). \quad (4.27)$$

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. Figure 4.5 shows three possible ways of choosing points  $s_l^{(i)}$ , together with the centroids approach.

Computing the  $n \times n$  Gram matrix  $K$  requires

$$\mathcal{O}\left(\sum_{i=1}^n \sum_{j=1}^n L_i L_j\right) \quad (4.28)$$

## Models for spatial structure

evaluations of the kernel  $k$ . This imposes a significant computational cost if the Gram matrix is often recomputed during inference. For example, during MCMC when the kernel has hyperparameters which are learnt then the Gram matrix is recomputed for each proposed set of hyperparameters. As such, there is a limit on the size of  $L_i$  which it is feasible to use. Kelsall and Jonathan Wakefield (2002) encounter this challenge, and take the approach of using a discrete hyperparameter prior to reduce the number of Gram matrix constructions and inversions required.

### 4.2.2.3 Connection to log-Gaussian Cox processes

The log-Gaussian Cox Process framework (Diggle, Moraga, et al. 2013) arrives naturally at the integrated kernel formulation (Li et al. 2012). 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

$$y_i | x(s) \sim \text{Poisson} \left( \int_{s \in A_i} x(s) ds \right). \quad (4.29)$$

In a LGCP the log intensity  $\log x(s) = \eta(s)$  is modelled using a Gaussian process prior  $\eta(s) \sim \mathcal{GP}(\mu(s), k(s, s'))$ . O. Johnson et al. (2019) obtain Equation (4.26) by considering a discrete Poisson log-linear mixed model approximation to a continuous LGCP, whereby  $\eta(s)$  is approximated by a piecewise constant  $\eta_i = \mu_i + u_i$  in each area  $A_i$ . The  $i$ th discrete spatial random effect is then  $u_i = \int_{A_i} w_i(s) u(s) ds$ , with covariance structure

$$\text{Cov} \left( \int_{A_i} w_i(s) u(s) ds, \int_{A_j} w_j(s') u(s') ds' \right) = \int_{A_i} \int_{A_j} w_i(s) w_j(s') k(s, s') ds ds', \quad (4.30)$$

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

#### 4.2.2.4 Connection to 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 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; Arambepola et al. 2022; Nandi et al. 2023). These two tasks are similar, and indeed it could be argued that accurate point-level estimates are a necessary intermediate step towards accurate area-level estimates. However, disaggregation regression is challenging without auxiliary covariate information, and therefore unlikely to be applicable to small-area estimation of HIV.

### 4.3 Simulation study

This simulation study tests 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 application in mind.

#### 4.3.1 Synthetic data

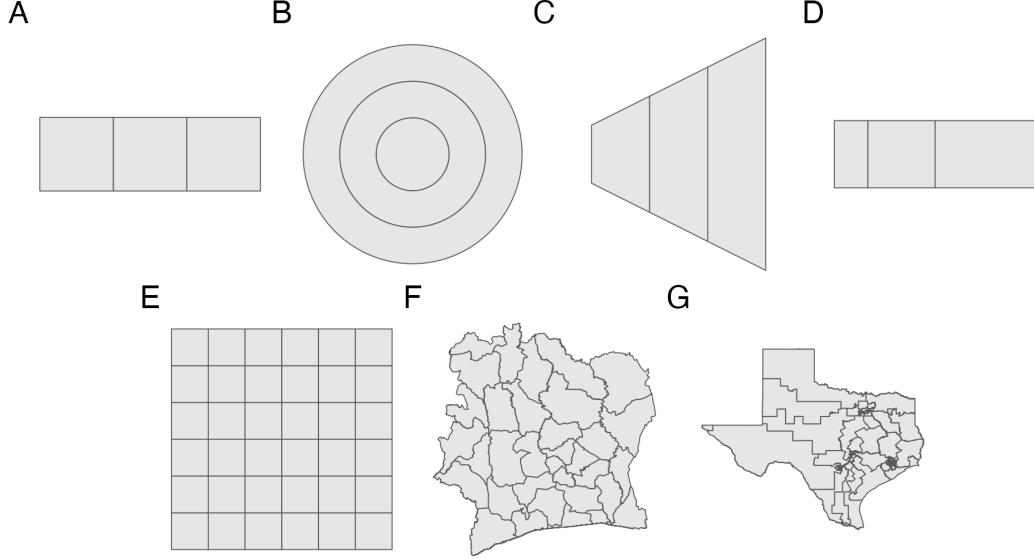
**Table 4.1:** The three spatial random effect models used to generate synthetic data in the simulation study (Section 4.3).

Model	Details
IID	$\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_n)$
Besag	$\mathbf{u} \sim \mathcal{N}(0, \mathbf{R}^{\star-})$ as in Section 4.1.1
Integrated kernel (IK)	$\mathbf{u} \sim \mathcal{N}(0, \mathbf{K}^{\star})$ as in Section 4.2.2 with Matérn kernel, $\nu = 3/2, l = 2.5$ and $L_i = 100$ points per area

Data  $\mathbf{y} = (y_i)_{i \in [n]}$  were simulated from a binomial likelihood  $y_i \sim \text{Bin}(m_i, \rho_i)$ . The probabilities  $\rho_i \in [0, 1]$  were linked to linear predictors  $\eta_i \in \mathbb{R}$  via

$$\log\left(\frac{\rho_i}{1 - \rho_i}\right) = \eta_i = \beta_0 + u_i, \quad i \in [n]. \quad (4.31)$$

## Models for spatial structure



**Figure 4.6:** Seven geometries were considered in the simulation study. These were the four geometries from Figure 4.2 shown in Panel A, B, C and D, and three more realistic geometries shown in Panel E, F and G.

Spatial random effects were generated according to three different models (Table 4.1). Sample sizes were fixed as  $m_i = 25$  for all  $i \in [n]$ , the intercept parameter as  $\beta_0 = -2$  and the spatial random effect precision parameter as  $\tau_u = 1$ .

Seven geometries were considered (Figure 4.6). These included the four vignette geometries from Figure 4.2 which share an adjacency graph. Three more realistic geometries were included to represent plausible variation over spatial regularity for the small-area estimation setting. From the most to the least spatially regular, these geometries were: a  $6 \times 6$  lattice grid; the 33 districts of Côte d'Ivoire; and the 36 congressional districts of Texas. For each of the three spatial random effect models and seven geometries 250 synthetic data were generated, resulting in a total of 5250 synthetic data.

### 4.3.2 Inferential models

## Models for spatial structure

**Table 4.2:** The spatial random effect models used for inference. Each model is implemented in the `arealutils` package. The BYM2 model was implemented using the sparsity preserving parameterisation described in Section 3.2 of Riebler et al. (2016).

Model	Details
IID	$\mathbf{u} \sim \mathcal{N}(0, \tau_u^{-1} \mathbf{I}_n)$
Besag	$\mathbf{u} \sim \mathcal{N}(0, \tau_u^{-1} \mathbf{R}^{\star-})$ as in Section 4.1.1
BYM2	$\mathbf{u} = \tau_u^{-1} (\sqrt{1-\pi} \mathbf{v} + \sqrt{\pi} \mathbf{w}^*)$ as in Section 4.1.5 with $\pi \sim \text{Beta}(0.5, 0.5)$
FCK	$\mathbf{u} \sim \mathcal{N}(0, \tau_u^{-1} \mathbf{K})$ with $K_{ij} = k(c_i, c_j)$ as in Section 4.2.1 with fixed length-scale $l$
CK	$\mathbf{u} \sim \mathcal{N}(0, \tau_u^{-1} \mathbf{K})$ with $K_{ij} = k(c_i, c_j)$ as in Section 4.2.1 with length-scale prior distribution $l \sim \text{Inv-Gamma}(a, b)$ with $a, b$ set based on the geometry
FIK	$\mathbf{u} \sim \mathcal{N}(0, \tau_u^{-1} \mathbf{K})$ with $K_{ij} = K(A_i, A_j)$ as in Section 4.2.2 with hexagonal points (Panel 4.5C), $L_i = 10$ , and fixed length-scale $l$
IK	$\mathbf{u} \sim \mathcal{N}(0, \tau_u^{-1} \mathbf{K})$ with $K_{ij} = K(A_i, A_j)$ as in Section 4.2.2 with hexagonal points (Panel 4.5C), $L_i = 10$ , and length-scale prior distribution $l \sim \text{Inv-Gamma}(a, b)$ with $a, b$ set based on the geometry

Eight inferential models were fit to the synthetic data (Table 4.2). Apart from the spatial random effect specification, each inferential model corresponded exactly to the simulation model.

### 4.3.2.1 Kernels

Gram matrices were computed using the Matérn kernel  $k : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$  (Stein 1999) given by

$$k(s, s') = \frac{1}{2^{\nu-1} \Gamma(\nu)} \left( \frac{\sqrt{2\nu} |s - s'|}{l} \right)^{\nu} B_{\nu} \left( \frac{\sqrt{2\nu} |s - s'|}{l} \right). \quad (4.32)$$

In Equation (4.32):

- $B_{\nu}$  is the modified Bessel function of the second kind;
- $|s - s'|$  is the Euclidean distance between the point locations  $s$  and  $s'$ ;
- $\nu$  is the smoothness hyperparameter;
- $l$  is the length-scale hyperparameter on the latitude-longitude scale.

## *Models for spatial structure*

The smoothness parameter  $\nu$  is difficult to identify from data and so was fixed at 3/2. This value matches that used to simulate data, and simplifies Equation (4.32) as follows

$$k(s, s') = \left(1 + \frac{\sqrt{3}|s - s'|}{l}\right) \exp\left(-\frac{\sqrt{3}|s - s'|}{l}\right). \quad (4.33)$$

The number of points per area  $L_i$  was set to 10 with a hexagonal spacing structure (Panel 4.5C). The actual values of  $L_i$  sometimes differed from 10 because `sf::st_sample` with `type = "hexagonal"` does not guarantee exactly the specified number of samples are returned (E. Pebesma 2018).

### 4.3.2.2 Prior distributions

A weakly informative half-Gaussian prior was placed on the standard deviation such that  $\sigma_u \sim \mathcal{N}_+(0, 2.5^2)$  (Gelman 2006). The value 2.5 avoids placing significant prior density on the region  $\sigma_u > 5$ , which after logistic transformation would facilitate undesirable variation on the probability scale very close to zero or one. A weakly informative  $\mathcal{N}(-2, 1)$  prior was placed on  $\beta_0$ , setting most of the prior probability density for  $\text{logit}^{-1}(\beta_0)$  within a range typical for a disease prevalence.

In cases where the length-scale  $l$  was fixed, it was set based on the geometry such that points an average distance apart had 1% correlation (Best et al. 1999). In cases where a prior distribution was set on the length-scale it was  $l \sim \text{Inv-Gamma}(a, b)$ , with  $a$  and  $b$  chosen for each geometry such that 5% of the prior mass was below the 5% quantile for distance between points and 5% of the prior mass was above the 95% quantile (Betancourt 2017). The sensitivity analysis in Appendix A.2 illustrates the extent to which six possible lengthscale prior distributions (Figure A.9) effect the lengthscale posterior distribution (Figure A.10).

### 4.3.2.3 Inference

Approximate Bayesian inference was conducted using adaptive Gauss-Hermite quadrature [AGHQ; Stringer et al. (2022)] with  $k = 3$  quadrature points over a marginal Laplace approximation via the `aghq` package (Stringer 2021). Models were

## Models for spatial structure

implemented using a Template Model Builder C++ template for the log-posterior via the **TMB** package (Kristensen et al. 2016). Appendix A.1 compares posterior mean and standard deviations from AGHQ to those obtained using the No-U-Turn Sampler (NUTS) Hamiltonian Monte Carlo (HMC) algorithm run using **Stan** (Carpenter et al. 2017) via the **tmbstan** package (Monnahan and Kristensen 2018).

### 4.3.3 Model assessment

Let the parameter  $\phi$  have posterior marginal  $f(\phi) = p(\phi | \mathbf{y})$  with cumulative distribution function  $F$ . Let  $\phi_s$  be samples  $s \in [S]$  from  $f$ . Here, the number of samples per posterior marginal was  $S = 200$ . Let  $\omega$  be the true value of  $\phi$  used in the simulation.

The accuracy of latent field parameter and hyperparameter posterior marginals from each model were assessed using three methods. These were the mean squared error (MSE), the continuous ranked probability score [CRPS; Matheson and Winkler (1976)], and the calibration.

The MSE is a simple and popular measure, calculated using samples as

$$\text{MSE}(f, \omega) \approx \frac{1}{S} \sum_{s=1}^S (\phi_s - \omega)^2. \quad (4.34)$$

The CRPS is a strictly proper scoring rule (SPSR) which has favourable properties and is often regarded as a default choice (Gneiting and Raftery 2007). Any scoring rule which is not strictly proper rewards a misrepresentation of beliefs. The CRPS is

$$\text{CRPS}(f, \omega) = \int_{-\infty}^{\infty} (F(\phi) - \mathbb{I}\{\phi \geq \omega\})^2 d\phi. \quad (4.35)$$

The CRPS may be estimated using samples by

$$\text{CRPS}(f, \omega) \approx \frac{1}{S} \sum_{s=1}^S |\phi_s - \omega| - \frac{1}{2S^2} \sum_{s=1}^S \sum_{l=1}^S |\phi_s - \phi_l|. \quad (4.36)$$

A posterior marginal is calibrated if over repeated simulations the quantile of the true value is uniformly distributed such that

$$F(\omega) \approx \frac{1}{S} \sum_{s=1}^S \mathbb{I}\{\phi_s \geq \omega\} = q \sim \mathcal{U}[0, 1]. \quad (4.37)$$

If Equation (4.37) holds then at any given nominal coverage  $1 - \alpha$  the proportion of quantile-based credible intervals containing  $\omega$  is also  $1 - \alpha$ . Uniformity was assessed using probability integral transform (PIT) histograms (Dawid 1984) and empirical cumulative distribution function (ECDF) difference plots (Aldor-Noiman et al. 2013) with simultaneous confidence bands as described in Säilynoja et al. (2022).

### 4.3.4 Results

#### 4.3.4.1 Vignette geometries

As each geometry only had three areas, the sample size of 250 synthetic data was insufficient to distinguish between inferential models for the vignette geometries. Figures A.13, A.14, A.15 and A.16 show that almost all 95% credible intervals for the mean CRPS overlap.

Additionally, for the vignette geometries, both the heuristic method for fixing a lengthscale, and lengthscale prior distribution, were misspecified. Three points was insufficient to learn the lengthscale, and as such misspecification of the prior distribution propagated to the posterior distribution (Figure A.11).

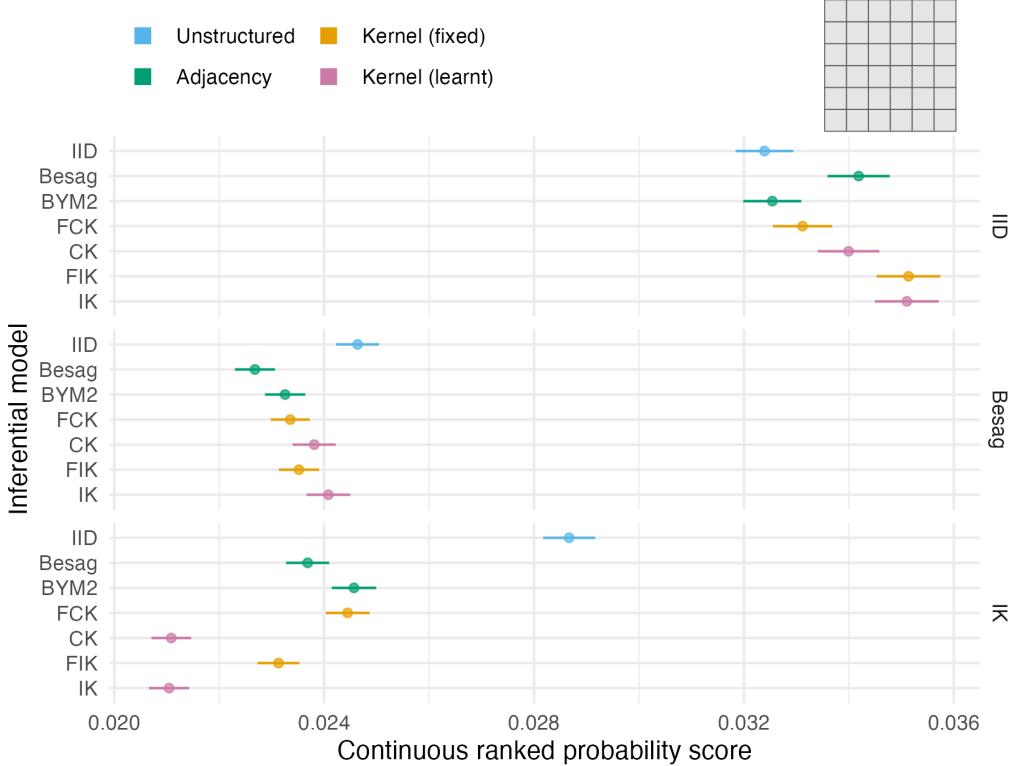
To produce higher resolution and more meaningful results, the simulation study for the vignette geometries should be rerun. Two changes should be made. First, an increase to the sample size. Second, more careful specification of study with regard to the lengthscale.

#### 4.3.4.2 Realistic geometries

The two problems with the vignette geometry study did not apply to the more realistic geometries. Figures 4.7, 4.8 and 4.9 show mean CRPS values with 95% credible intervals which rarely overlap, and hence provide meaningful findings. Mean MSE and CRPS values are provided in Tables ?? and ??.

The average CRPS varied substantially between the three models (Table 4.1) used to simulate synthetic data. IID structure is harder to predict than spatial structure, and to a lesser extent, Besag structure is harder to predict than IK. This observation is explained by correlation structure making forecasting easier.

## Models for spatial structure



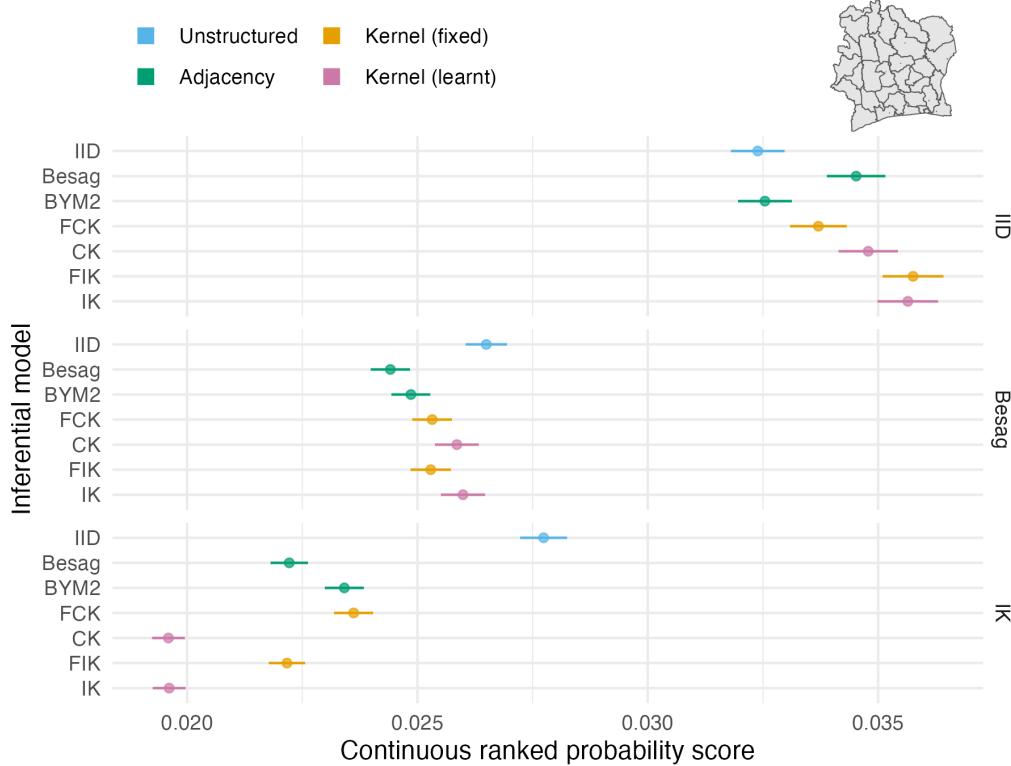
**Figure 4.7:** The mean CRPS and its standard error for each inferential model and simulation model on the grid geometry (Panel 4.6E).

For IID synthetic data, the IID and BYM2 models performed well. The BYM2 model also performed almost as well as the Besag model on the spatially structured synthetic data. Appendix A.3.2 shows that the BYM2 proportion parameter successfully recovers either IID or spatial structure. Meanwhile, the IID model performed poorly on spatially structured synthetic data.

The performance of kernel models on IID and Besag synthetic data diminished with increasingly spatially irregular geometry. For the most part, differences between the centroid and integrated kernel models were small, even for synthetic data generated from the IK model. Only for the IK simulated data there was a significant difference between the kernel models with a fixed lengthscale and prior distribution set on the lengthscale.

Interpretation of CRPS choropleths (Figures ??, ?? and ??) was challenging primarily due to two factors: varying scores by simulation model, and limited

## Models for spatial structure



**Figure 4.8:** The mean CRPS and its standard error for each inferential model and simulation model on the Côte d’Ivoire geometry (Panel 4.6F).

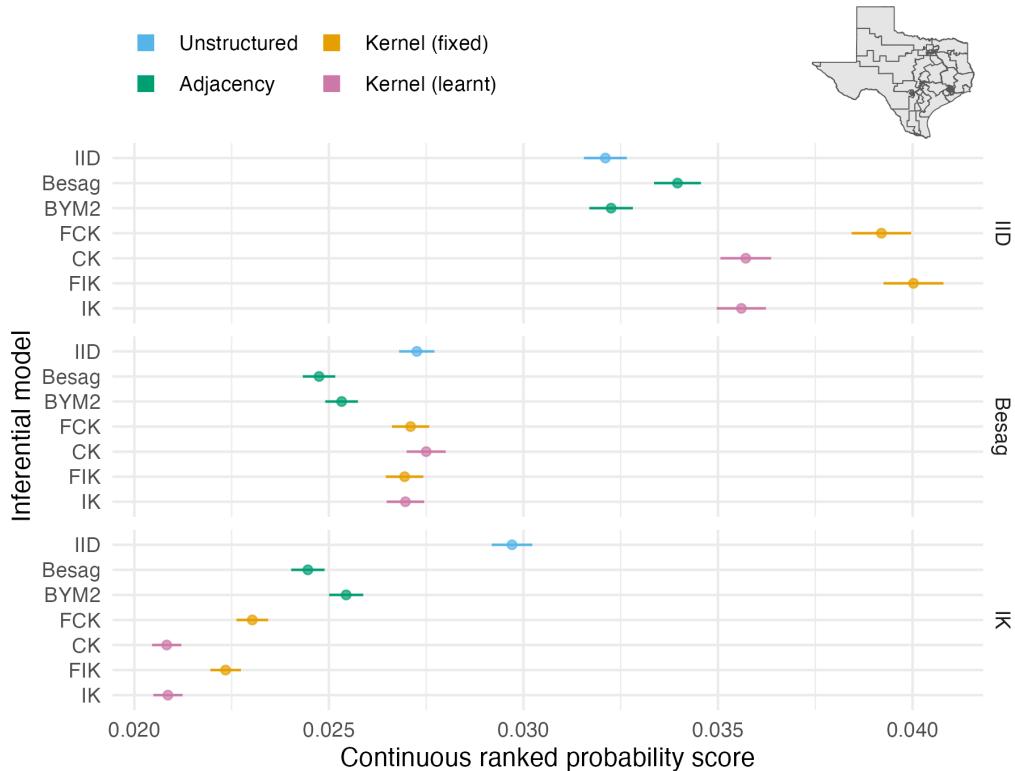
sample size at the area-level. It would be relatively simple to remedy these challenges, such that figures of this kind could help to uncover precise findings about spatial random effect models.

For IID synthetic data, spatial models tend to produce “U”-shaped ECDF difference plots (Figures ??, ?? and ??). In other words, the quantile of the true value is too often near zero or one. This pattern corresponds to over-smoothing.

## 4.4 HIV prevalence study

Simulation studies are a valuable tool for experimenting on models in controlled environments. However, it is difficult to capture the complexity of a realistic applied scenario using simulation. Therefore, it is important to complement simulation studies with studies conducted on real data. To this end, model performance was

## Models for spatial structure



**Figure 4.9:** The mean CRPS and its standard error for each inferential model and simulation model on the Texas geometry (Panel 4.6G).

compared in estimating district-level HIV prevalence  $\rho_i \in [0, 1]$  in adults aged 15-49. Household survey data was used from across four countries in sub-Saharan Africa (Table 4.3, Figure 4.10).

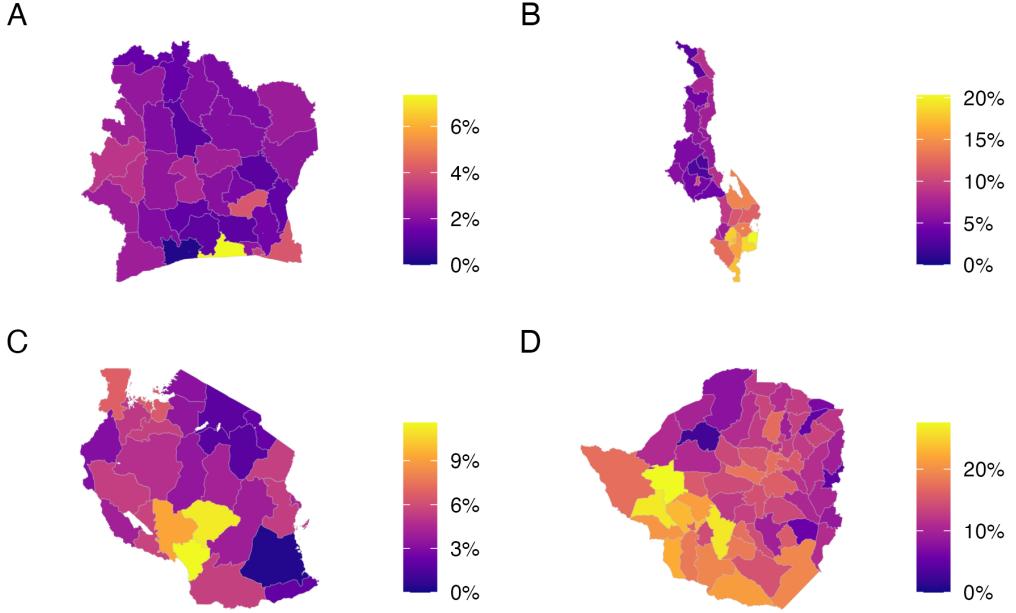
**Table 4.3:** The four PHIA household surveys included in the HIV prevalence study (Section 4.4).

Country	Survey	Number of areas	Analysis level
Côte d'Ivoire	PHIA 2017	33	Regions
Malawi	PHIA 2016	31	Health districts and cities, with islands removed
Tanzania	PHIA 2017	26	Regions, with islands removed
Zimbabwe	PHIA 2016	60	Districts

### 4.4.1 Household survey data

Data from the most recent publicly available Population Health Impact Assessment (PHIA) survey were used in each country. Let  $y_{ij} \in \{0, 1\}$  be the survey response for individual  $j$  in area  $i$ . The survey designs used were complex in that each individual had potentially unequal probabilities  $\pi_{ij}$  of being included in the survey.

## Models for spatial structure



**Figure 4.10:** Adult (15-49) HIV prevalence from the most recent PHIA survey conducted in Côte d'Ivoire (Panel A), Malawi (Panel B), Tanzania (Panel C), and Zimbabwe (Panel D). These estimates are survey weighted according to Equation (4.39).

### 4.4.2 Inferential models

The inferential models used correspond to those in Section 4.3 with a small modification. As before, prevalences  $\rho_i$  were modelled via  $\text{logit}(\rho_i) = \beta_0 + u_i$  with spatial random effect specification varied according to Table 4.2. Due to survey weighting, the effective number of cases  $y_i^* \in \mathbb{R}$  and effective sample size  $m_i^* \in \mathbb{R}$  may not be integers. Following C. Chen et al. (2014) a generalised binomial distribution  $y_i^* \sim \text{xBin}(m_i^*, \rho_i)$  was used, with working likelihood for  $m_i^* \geq y_i^*$  given by

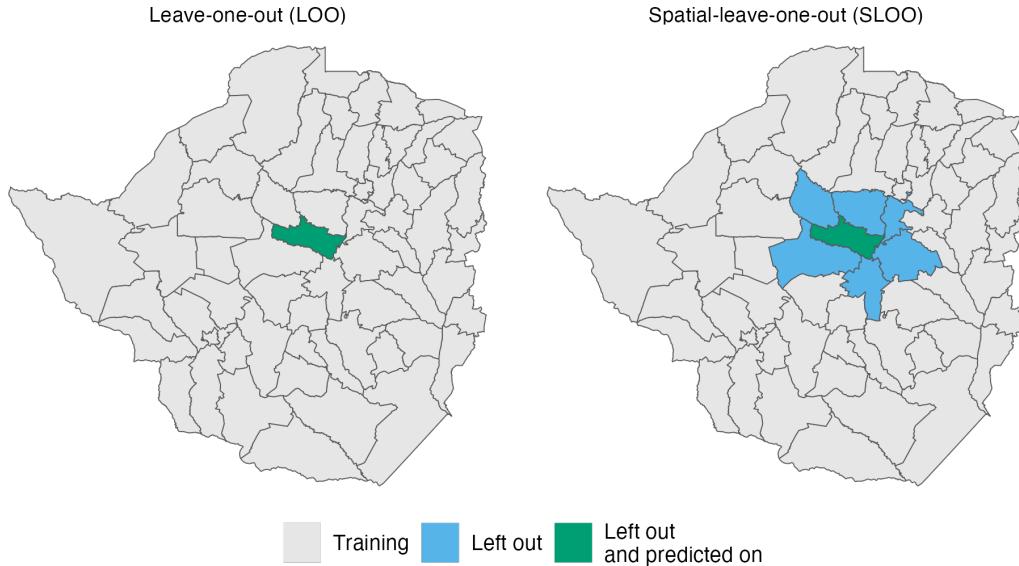
$$p(y_i^* | m_i^*, \rho_i) = \frac{\Gamma(m_i^* + 1)}{\Gamma(y_i^* + 1)\Gamma(m_i^* - y_i^* + 1)} \rho_i^{y_i^*} (1 - \rho_i)^{(m_i^* - y_i^*)}. \quad (4.41)$$

### 4.4.3 Model comparison

Each model was assessed using (Figure 4.11):

1. a regular leave-one-out cross-validation (LOO-CV);
2. a spatial leave-one-out cross-validation (SLOO-CV).

## Models for spatial structure



**Figure 4.11:** In leave-one-out (LOO) cross-validation, one observation is left out of the training data and predicted upon in each fold. The spatial-leave-one-out (SLOO) cross-validation scheme considered here is similar, only differing in that observations corresponding to adjacent areas are also left out of the training data.

At each fold the CRPS, MSE and quantile (as in Section 4.3.3) of posterior predictive samples as compared with the observed data were computed. In this section, the number of samples per posterior marginal was  $S = 1000$ .

### 4.4.4 Results

**Table 4.4:** The mean pointwise leave-one-out and spatial leave-one-out CRPS in estimating  $\rho_i$ , with standard errors, for each inferential model across the four considered PHIA surveys. The units used in this table are thousandths.

PHIA survey	Continuous ranked probability score (units: 1/1000)						
	IID	Besag	BYM2	FCK	CK	FIK	IK
<b>LOO</b>							
Côte d'Ivoire, 2017	6.6	6.6	6.7	6.7	6.9	6.9	6.9
Malawi, 2016	31.7	19.5	19.6	22.7	22.8	21.4	21.0
Tanzania, 2017	14.9	12.1	13.4	10.7	9.5	10.3	10.6
Zimbabwe, 2016	28.9	20.8	20.9	21.7	21.6	21.4	22.0
<b>SLOO</b>							

### *Models for spatial structure*

Côte d'Ivoire, 2017	6.5	6.6	6.6	6.4	6.9	6.4	6.8
Malawi, 2016	31.6	19.3	19.9	26.5	29.0	25.1	28.3
Tanzania, 2017	14.9	12.1	18.1	16.0	17.6	15.4	16.9
Zimbabwe, 2016	29.1	20.8	25.2	26.7	26.2	26.1	26.3

The results (Figure 4.12, Table 4.4, Table A.4) for each survey were as follows:

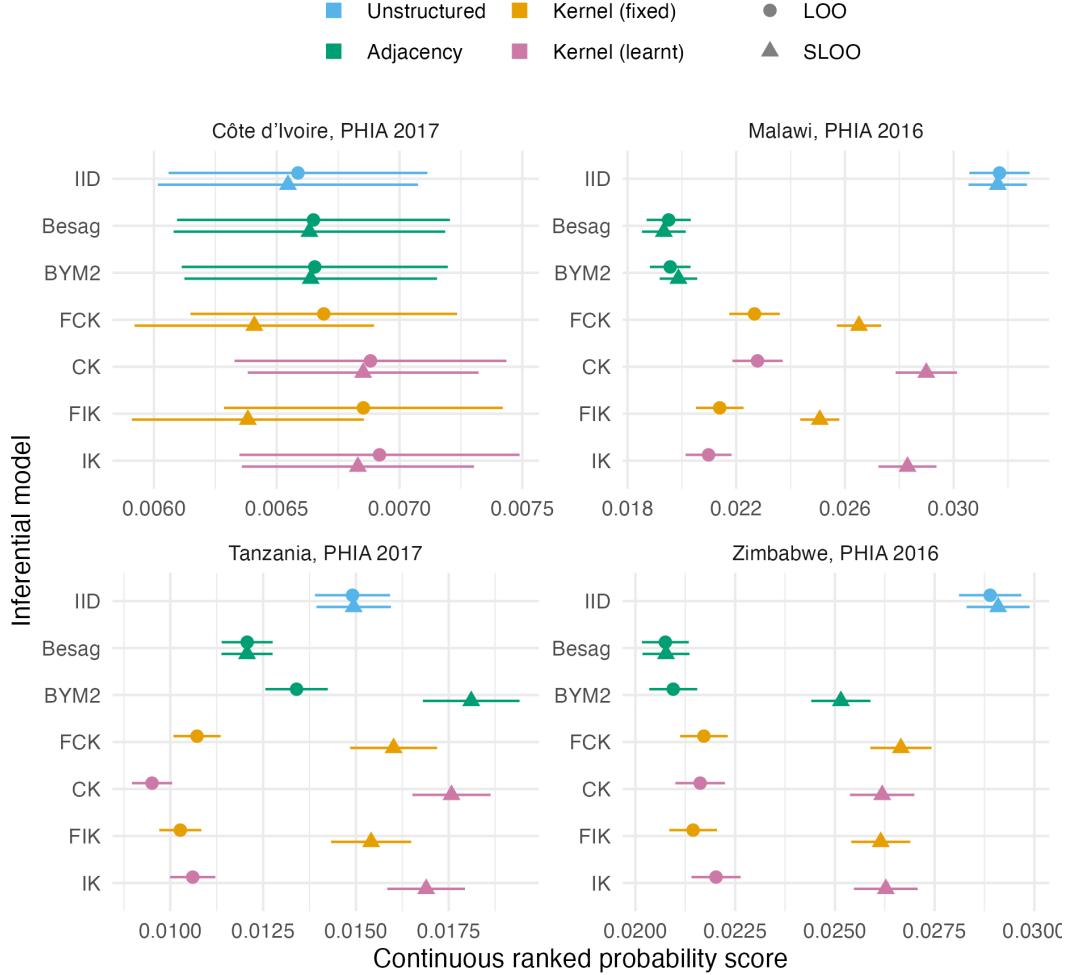
1. For the 2017 PHIA survey in Côte d'Ivoire, all of the models performed similarly, using both LOO- and SLOO-CV (Figure A.37) The pointwise CRPS for all models was high at one outlying district in the survey, Grand-Ponts. It is difficult to see how any spatial random model would perform well in this situation, without additional covariates or using a distribution with heavier tails than the Gaussian.

The CK and IK models had lengthscale posterior distributions largely unchanged from their prior distribution (Figure A.31). This uncertainty in lengthscale resulted wide prevalence 95% credible intervals for the CK and IK models in Figure A.33. This example shows the importance of being careful using kernel models, and the prior distributions set on their hyperparameters. It is surprisingly that this behaviour appears not to have resulted in poor LOO or SLOO performance.

For this survey the BYM2 proportion posterior distribution was also similar to its prior distribution, in contrast to each of the other surveys which had BYM2 proportion posteriors peaked at one, corresponding to spatially structured noise (Figure A.32).

2. For the 2017 PHIA survey in Malawi the Besag and BYM2 models performed the best, followed by the kernel models, and then the IID model (Figure A.38). While the LOO and SLOO CRPS values for IID, Besag and BYM2 models were similar, for the kernel models forecasting performance was substantially reduced by leaving out adjacent districts. This finding is surprising, as the kernel models make use of more distant correlations, and it is the adjacency-based models that one would intuitively expect to be hampered more by

## Models for spatial structure



**Figure 4.12:** The mean pointwise leave-one-out and spatial leave-one-out CRPS in estimating  $\rho_i$  using each inferential model for the four PHIA surveys described in Table 4.3. The 95% credible intervals shown are generated using 1.96 times the standard error.

the SLOO-CV. For the IID model, that LOO and SLOO performance are similar is no surprise as in all cases the IID model should be predicting the mean. Though less data is available in the SLOO case, this should be of little consequence.

3. For the 2017 PHIA survey in Tanzania (Figure A.39), under LOO-CV the kernel models performed better, but under SLOO-CV there was a significant drop in performance.
4. Finally, for the 2016 PHIA survey in Zimbabwe, performance for each of the spatially structured models was similar (Figure A.40). Again, under SLOO-

CV, performance of the BYM2 and kernel-based models dropped. Differences within the kernel-based models for this survey, and indeed across all four surveys, were limited.

## 4.5 Discussion

### 4.5.1 Modelling

Though there are situations where other models perform better, on the whole this study supports the use of adjacency-based spatial random effect models. For the study on HIV survey data, adjacency-based models performed well, if not the best, in all cases. That is not to say that under data truly generated from a kernel model, there isn't significant benefit to using the corresponding kernel model for inference. However, the transferability of this finding to applied settings is limited by the following factors. First and foremost, it is usually impossible to know that real data was generated from any particular process. Second, the synthetic data study used the same kernel, Matérn with  $\nu = 3/2$  (Equation (4.32)), for both simulation and inference, and as such represents a best-case. Third, specification of the lengthscale prior distribution is challenging, and easy to do badly. Finally, aggregation via the integrated kernel occurred at the level of the latent field, despite the fact that most of the time we expect aggregation to occur at the level of the data. If the link function  $g$  is the identity or linear then the two are equivalent, but for non-linear link functions there is a discrepancy, which was not addressed in this study.

This chapter did not consider use of the stochastic partial differential equation (SPDE) approximation of Lindgren et al. (2011) as a potentially more computationally efficient way to implement integrated kernel models (K. Wilson and Jon Wakefield 2018). Though the underlying models are ultimately similar, that is a continuous Matérn random field over space aggregated at an area-level, the findings from this work are likely to apply to use of the SPDE approximation. Nonetheless, it would be of value to confirm this empirically.

## *Models for spatial structure*

This chapter used of area-level models to for point-level data throughout. However, Konstantinoudis et al. (2020) found that using a point-level LGCP model rather than an area-level BYM model may have significant benefits. The work in this chapter does not address the broader question of under which circumstances use of an area or point-level model is sensible.

The adjacency-based models considered in this study were limited to the Besag and BYM2 model. Although these are perhaps the most widely used adjacency-based models, others could have been considered. Examples include the more general weighted ICAR model discussed in Section 4.1.4. Additionally, it would be of interest to implement the integrated kernel model with population-based weighting (Section 4.2.2.1).

The models used for spatial structure in this chapter were all stationary. Although stationarity assumptions may be violated by HIV survey data, it remains challenging to estimate non-stationary spatial structure (Christopher J. Paciorek and Schervish 2006).

### **4.5.2 Model comparison**

Previous spatial random effect comparison studies (Nicky Best et al. 2005; D. Lee 2011) were limited to the DIC measure of model performance. Use of the DIC is strongly discouraged by Vehtari, Gelman, et al. (2017). This study used less flawed measures of model performance, such as the cross-validated CRPS. It would be beneficial to compute the DIC and WAIC in Section 4.4 as a comparison. Additionally, the measures presented in this work are disaggregated by area. With refinements to the sample sizes used, these disaggregated measures could enable nuanced findings about spatial random effect models.

Cross-validation was performed using  $\rho$  as the forecasting target, rather than  $y$  as is typical. This decision was made because applied interest is in forecasting HIV prevalence at a district level, not forecasting the outcome of a household survey. It could be argued that a district does not become more important to forecast well by virtue of surveying a larger sample size in that district. That said, an

### *Models for spatial structure*

alternative viewpoint is that forecast accuracy should be incentivised in proportion to district population size, such that PLHIV is accurately estimated. If sample size is proportional to population size, then forecasting  $y$  could be a useful proxy. Choice of the particular parameter, or transformation of that parameter (Nikos I Bosse et al. 2023), to score is an ongoing topic of research.

The CRPS was used in preference to the log-score. Whereas the log-score requires a kernel density estimate of the posterior distribution, and is therefore sensitive to tuning parameters, the CRPS can be estimated from samples alone. A downside of use of the CRPS and MSE is their relative lack of interpretability. For example, it is difficult to determine whether a forecast is good, or suitable for practical use, on the basis of its CRPS or MSE. Measures such as the skill score have been used to contrast forecast performance with some baseline. A constant model, with no random effects, could be used as such a baseline.

#### **4.5.3 Inference**

A strength of this work is that all of the inferential models (Table 4.2) in this chapter were implemented in TMB. Inference was then conducted using AGHQ over the marginal Laplace approximation using the `aghq` package. The accuracy of inferences was compared to gold-standard results from NUTS obtained using the `tmbstan` package. An earlier version of this study used R-INLA. Not all of the inferential models were compatible with R-INLA, so `rstan` was used in some cases. However, due to the difference in inference algorithm, this study design conflated statistical models with inference algorithms. Consistent use of TMB, a fast and flexible tool for spatial modelling (Osgood-Zimmerman and Jon Wakefield 2023), overcame this limitation. Chapter 6 extends TMB to implement the INLA algorithm of R-INLA.

# 5

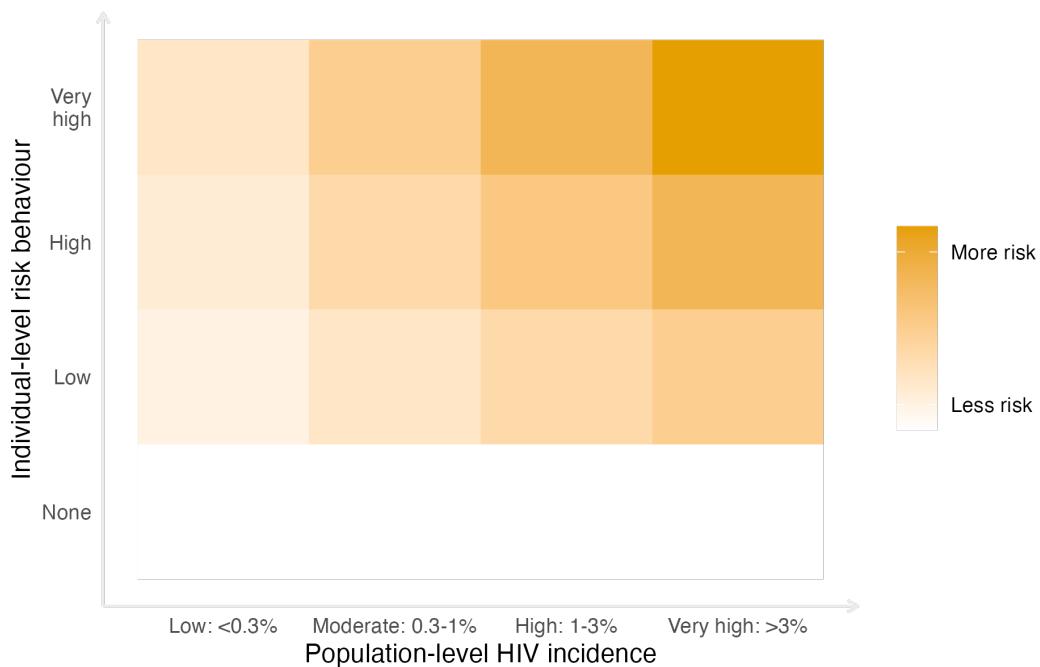
## A model for risk group proportions

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 et al. (2023). Outputs are 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> and supported by the `multi.utils` R package (Howes 2023b).

### 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

### *A model for risk group proportions*



**Figure 5.1:** Risk of acquiring HIV depends on both individual-level risk behaviour and population-level HIV incidence. It is assumed here 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.

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, including 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.

## *A model for risk group proportions*

Risk of acquiring HIV depends on both factors. As such, prioritisation of prevention services is more efficient if both factors are taken into account. Figure 5.1 illustrates this stylistically. The strategy encourages programmes to define targets for the proportion of AGYW to be reached with a range of interventions (Table B.2) based on prioritisation strata which incorporate behavioural risk (Table B.1). 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

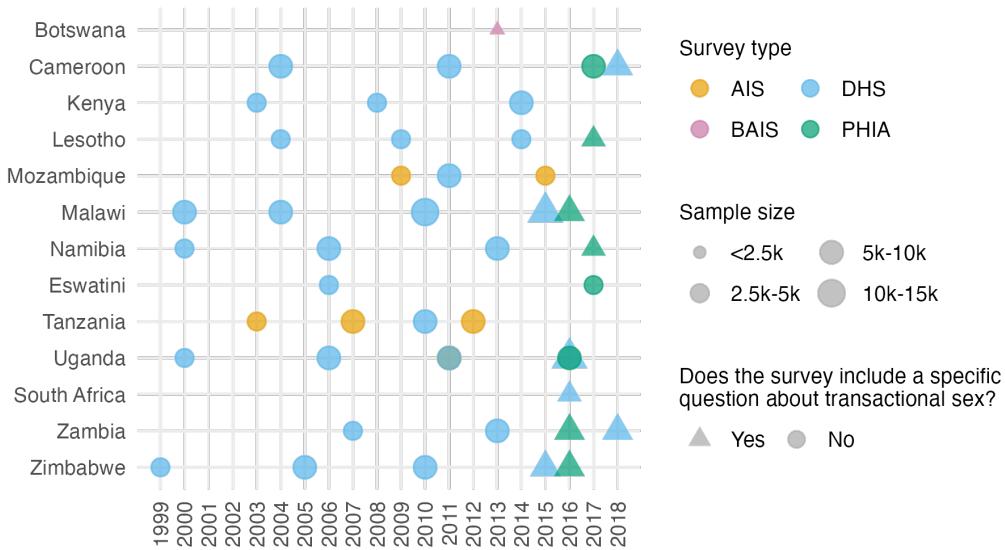
### 5.2.1 Behavioural data from household surveys

**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 (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

## A model for risk group proportions



**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.

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, respondents were classified 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.

### *A model for risk group proportions*

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.

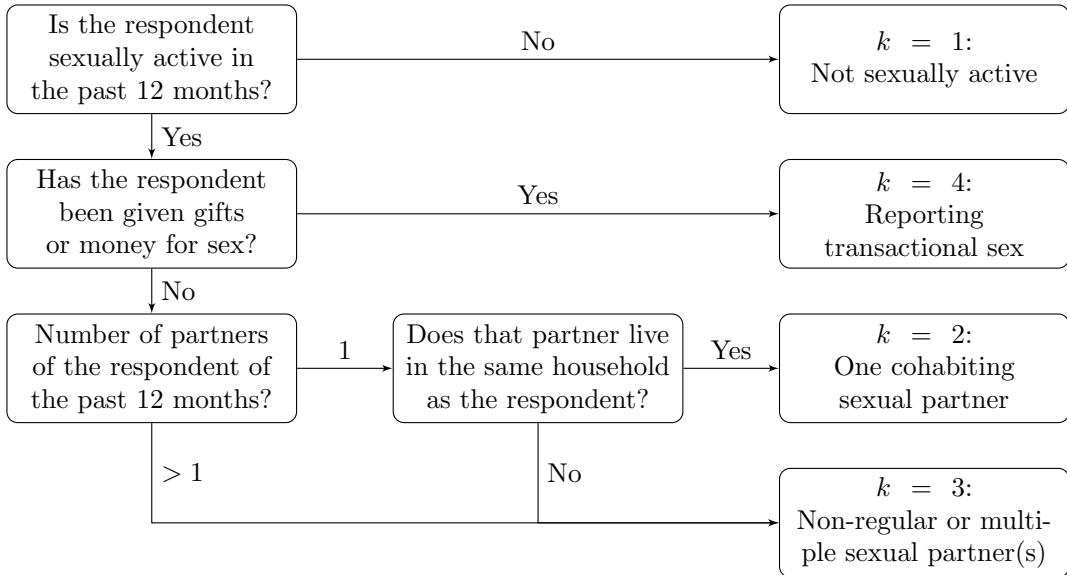
Exact survey questions varied slightly across survey types and between survey phases. Questions captured information about whether the respondent had been sexually active in the past twelve months, and if so with how many partners. For their three most recent partners, respondents were also asked about the type of partnership. Possible partnership types included spouse, cohabiting partner, partner not cohabiting with respondent, friend, sex worker, sex work client, and other. The 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. Table B.3 gives the sample size by age group for every survey included in the analysis.

### **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 estimates published by UNAIDS that were developed using the Naomi model (Jeffrey W 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

### A model for risk group proportions



**Figure 5.3:** Flowchart describing how respondents were classified to HIV risk groups based on their survey responses.

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 (Table B.5). 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 group 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  $i \in \{1, \dots, n\}$  denote districts partitioning the 13 studied AGYW priority countries  $c[i] \in \{1, \dots, 13\}$ . Consider the years 1999-2018 denoted as  $t \in \{1, \dots, T\}$ , and age groups  $a \in \{15-19, 20-24, 25-29\}$ . Let  $p_{itak} > 0$  with  $\sum_{k=1}^{3^+} p_{itak} = 1$ , be the probabilities of membership of risk group  $k$ .

#### 5.3.1.1 Multinomial logistic regression

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

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

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

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; Håvard 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 5.3.1.2). This modification allowed inference to be performed using the INLA (Håvard Rue, Martino, and Chopin 2009) algorithm via the R-INLA package (Martins et al. 2013).

#### 5.3.1.2 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

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

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

### A model for risk group proportions

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

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

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

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

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

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

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

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

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

$$= \prod_k \text{Poisson} (y_{itak} | \kappa_{itak}). \quad (5.10)$$

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 Poisson 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  $\theta_{ita}$  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  $\kappa_{ita}$  to recover

### A model for risk group proportions

$m_{ita}$  exactly. Although in theory an improper prior distribution  $\theta_{ita} \propto 1$  should be used, I found that in practice, by keeping  $\eta_{ita}$  otherwise small using appropriate constraints, so that arbitrarily large values of  $\theta_{ita}$  are not required, it is sufficient (and practically preferable for inference) to instead use a vague prior distribution.

#### 5.3.1.3 Model specifications

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

Category $\beta_k$	Country $\zeta_{ck}$	Age $\alpha_{ack}$	Spatial $\phi_{ik}$	Temporal $\gamma_{tk}$
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  $\eta_{ita}$  in the equivalent Poisson log-linear model of the form

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

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-category  $u_{ik}$  and time-category effects  $\gamma_{tk}$ , described in Sections 5.3.1.3 and 5.3.1.3.

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

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

The scaled structure matrix  $\mathbf{R}_u^* = \mathbf{R}_b^* \otimes \mathbf{I}$  is given by the Kronecker product of the scaled Besag structure matrix  $\mathbf{R}_b^*$  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  $\gamma_{tk}$  I considered two specifications:

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

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

## A model for risk group proportions

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

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

$$r_t = \rho r_{t-1} + \epsilon_t, \quad t = 2, \dots, T, \quad (5.13)$$

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 variable with the more complicated model was written first.

**Note on spatio-temporal interaction random effects** I also considered including separable space-time-category random effects  $\delta_{itk}$  in the model, using the specification

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

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

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

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.

## *A model for risk group proportions*

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

$$\tau \in \{\tau_\beta, \tau_\zeta, \tau_\alpha, \tau_u, \tau_\gamma, \tau_\delta\} \quad (5.15)$$

were given independent penalised complexity (PC) prior distributions (Simpson et al. 2017) with base model  $\sigma = 0$  given by

$$p(\tau) = 0.5\nu\tau^{-3/2} \exp(-\nu\tau^{-1/2}), \quad (5.16)$$

where  $\nu = -\ln(0.01)/2.5$  such that  $\mathbb{P}(\sigma > 2.5) = 0.01$ . For the lag-one correlation parameter  $\rho$ , I used the PC prior distribution, as derived by Sigrunn Holbek Sørbye and Håvard 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.

### **5.3.1.4 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 ensured that the main effect was 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	$\sum_i \delta_{itk} = 0, \forall t, k; \sum_t \delta_{itk} = 0, \forall i, k; \sum_k \delta_{itk} = 0, \forall i, t$

### 5.3.1.5 Survey weighted likelihood

I accounted for the survey design using a weighted pseudo-likelihood where the observed counts  $y$  are replaced by effective counts  $y^*$ , 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  $y^* \sim \text{xPoisson}(\kappa)$  given by

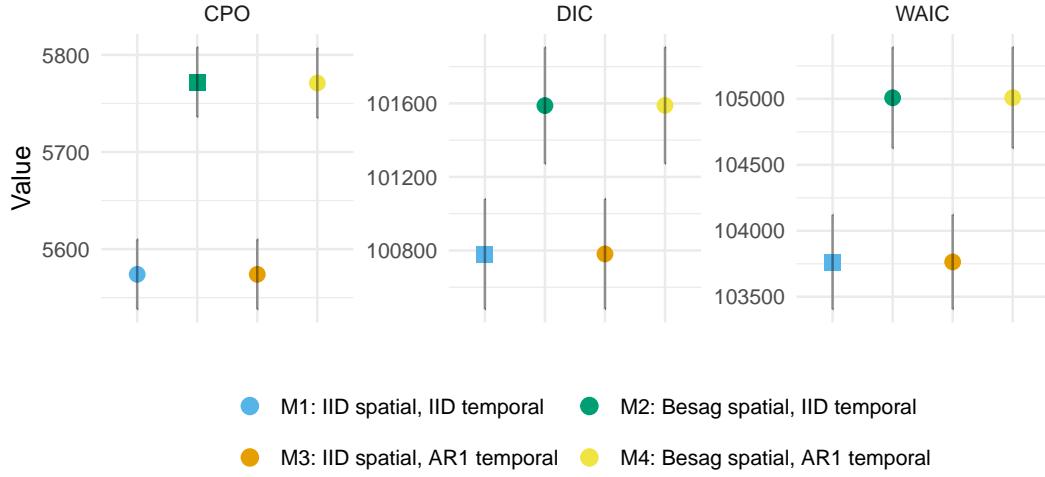
$$p(y^*) = \frac{\kappa^{y^*}}{[y^*!]} \exp(-\kappa), \quad (5.17)$$

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

### 5.3.1.6 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) (D. J. Spiegelhalter 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 Table 5.4 and Figure 5.4.

## A model for risk group proportions



**Figure 5.4:** For the multinomial logistic regression model, under the conditional predictive ordinate (CPO) criterion, including Besag spatial random effects rather than IID spatial random effects improved model performance. On the other hand, under the deviance information criterion (DIC) and widely applicable information criterion (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:** Conditional predictive ordinate (CPO), deviance information criterion (DIC), and widely applicable information criterion (WAIC) values for the multinomial logistic regression model specifications 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

$$y_{ia4} \sim \text{Binomial}(y_{ia3} + y_{ia4}, q_{ia}), \quad (5.18)$$

$$q_{ia} = \text{logit}^{-1}(\eta_{ia}), \quad (5.19)$$

where

$$q_{ia} = \frac{p_{ia4}}{p_{ia3} + p_{ia4}} = \frac{p_{ia4}}{p_{ia3^+}}. \quad (5.20)$$

### A model for risk group proportions

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.

#### 5.3.2.1 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.

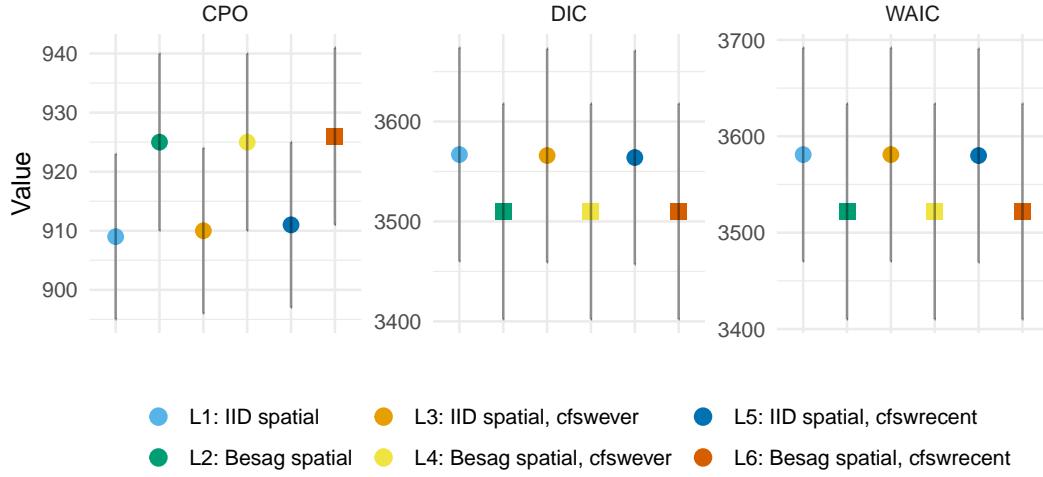
	Intercept $\beta_0$	Country $\zeta_c$	Age $\alpha_{ac}$	Spatial $u_i$	Covariates
L1	Constant	IID	IID	IID	None
L2	Constant	IID	IID	Besag	None
L3	Constant	IID	IID	IID	<code>cfswever</code>
L4	Constant	IID	IID	Besag	<code>cfswever</code>
L5	Constant	IID	IID	IID	<code>cfswrecent</code>
L6	Constant	IID	IID	Besag	<code>cfswrecent</code>

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  $\beta_0$  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  $\beta \in \{\beta_{\text{cfswever}}, \beta_{\text{cfswrecent}}\}$  I used the prior distribution  $\beta \sim \mathcal{N}(0, 2.5^2)$ .

#### 5.3.2.2 Survey weighted likelihood

As with the multinomial regression model, I used survey weighted counts  $y^*$  and sample sizes  $m^*$ . I used a generalised binomial pseudo-likelihood  $y^* \sim$

## A model for risk group proportions



**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.

`xBinomial( $m^*$ ,  $q$ )` given by

$$p(y^* | m^*, q) = \binom{\lfloor m^* \rfloor}{\lfloor y^* \rfloor} q^{y^*} (1 - q)^{m^* - y^*} \quad (5.21)$$

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

### 5.3.2.3 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 5.6 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.

### *A model for risk group proportions*

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

	L1	L2	L3	L4	L5	L6
CPO	950 (15)	969 (15)	951 (15)	970 (15)	950 (15)	970 (15)
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)

### **5.3.3 Female sex worker population size adjustment**

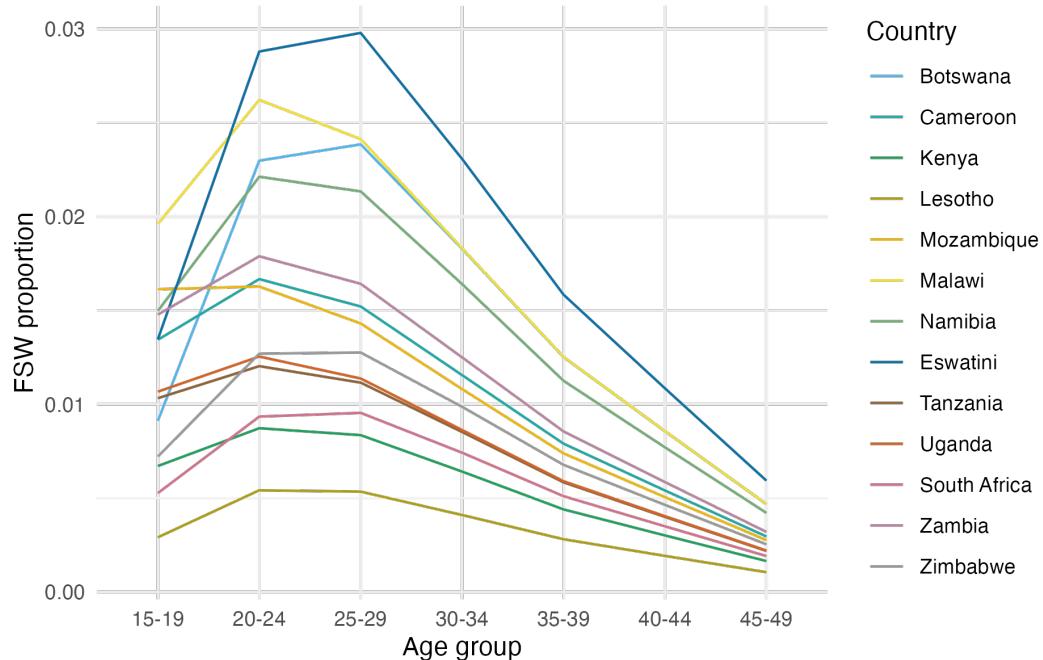
Having had sex “in return for gifts, cash or anything else in the past 12 months” is not considered sufficient to constitute sex work. As such, I adjusted the estimates obtained based on the transactional sex survey question to match FSW population size estimates obtained using an alternative method, 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 et al. 2023). To disaggregate these estimates by age, I took the following steps. First, I calculated the total sexually debited population in each age group, by country. To describe the distribution of age at first sex, I used skew logistic distributions (Nguyen and Jeffrey W. Eaton 2022) with cumulative distribution function given by

$$F(x) = (1 + \exp(\kappa_c(\mu_c - x)))^{-\gamma_c}, \quad (5.22)$$

where  $\kappa_c, \mu_c, \gamma_c > 0$  are country-specific shape, shape and skewness parameters respectively. Next, I used the assumed  $\text{Gamma}(\alpha = 10.4, \beta = 0.36)$  FSW age distribution in South Africa from the Thembisa model (L. Johnson and Dorrington 2020) to calculate the implied ratio between the number of FSW and the sexually debited 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.

## A model for risk group proportions



**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.

## 5.4 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  $\rho_{iak}$ ,
2. the number of people living with HIV (PLHIV)  $H_{iak}$ ,
3. HIV incidence  $\lambda_{iak}$ , 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

### A model for risk group proportions

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. To do so, I found the values of  $\theta_{ia}$  which minimised the equation

$$f(\theta_{ia}) = \sum_{k=1}^4 (\text{logistic}(\theta_{ia} + \log(\text{OR}_k)) \cdot N_{iak}) - H_{ia}, \quad (5.23)$$

where  $\text{logistic}(x) = \exp(x)/(1 + \exp(x))$  such that  $\text{logistic}(\hat{\theta}_{ia} + \log(\text{OR}_k)) = \rho_{iak}$ . These values were given by

$$\hat{\theta}_{ia} = \arg \min_{\theta_{ia} \in [-10, 10]} f(\theta_{ia})^2. \quad (5.24)$$

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 used linear disaggregation to calculate the number of new HIV infections by risk group

$$I_{ia} = \sum_k I_{iak} = \sum_k \lambda_{iak}(1 - \rho_{iak})N_{iak} \quad (5.25)$$

$$= 0 + \lambda_{ia2}(1 - \rho_{ia2})N_{ia2} + \lambda_{ia3}(1 - \rho_{ia3})N_{ia3} + \lambda_{ia4}(1 - \rho_{ia4})N_{ia4} \quad (5.26)$$

$$= \lambda_{ia2}((1 - \rho_{ia2})N_{ia2} + \text{RR}_3(1 - \rho_{ia3})N_{ia3} + \text{RR}_4(\lambda_{ia})(1 - \rho_{ia4})N_{ia4}), \quad (5.27)$$

where  $\text{RR}_2$ ,  $\text{RR}_3$  and  $\text{RR}_4(\cdot)$  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

### *A model for risk group proportions*

ratio for FSW was defined as a function of district-level incidence in the general population  $\lambda_{ia}$ . Risk group specific HIV incidence estimates were then given by

$$\lambda_{ia1} = 0, \quad (5.28)$$

$$\lambda_{ia2} = \frac{I_{ia}}{(1 - \rho_{ia2})N_{ia2} + RR_3(1 - \rho_{ia3})N_{ia3} + RR_4(\lambda_{ia})(1 - \rho_{ia4})N_{ia4}}, \quad (5.29)$$

$$\lambda_{ia3} = RR_3\lambda_{ia2}, \quad (5.30)$$

$$\lambda_{ia4} = RR_4(\lambda_{ia})\lambda_{ia2}. \quad (5.31)$$

These equations were evaluated 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**

The number of new infections that would be reached prioritising according to each possible stratification of the population were calculated. Each possible stratification refers to all  $2^3 = 8$  possible combinations of stratification by location, age, and risk group. As an illustration, consider stratification by age. I first aggregated the number of new HIV infections and HIV incidence such that

$$I_a = \sum_{ik} I_{iak}, \quad (5.32)$$

$$\lambda_a = I_a / \sum_{ik} (1 - \rho_{iak})N_{iak}. \quad (5.33)$$

I then considered prioritisation individuals by age group  $a$  according to the highest HIV incidence  $\lambda_a$ . 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.

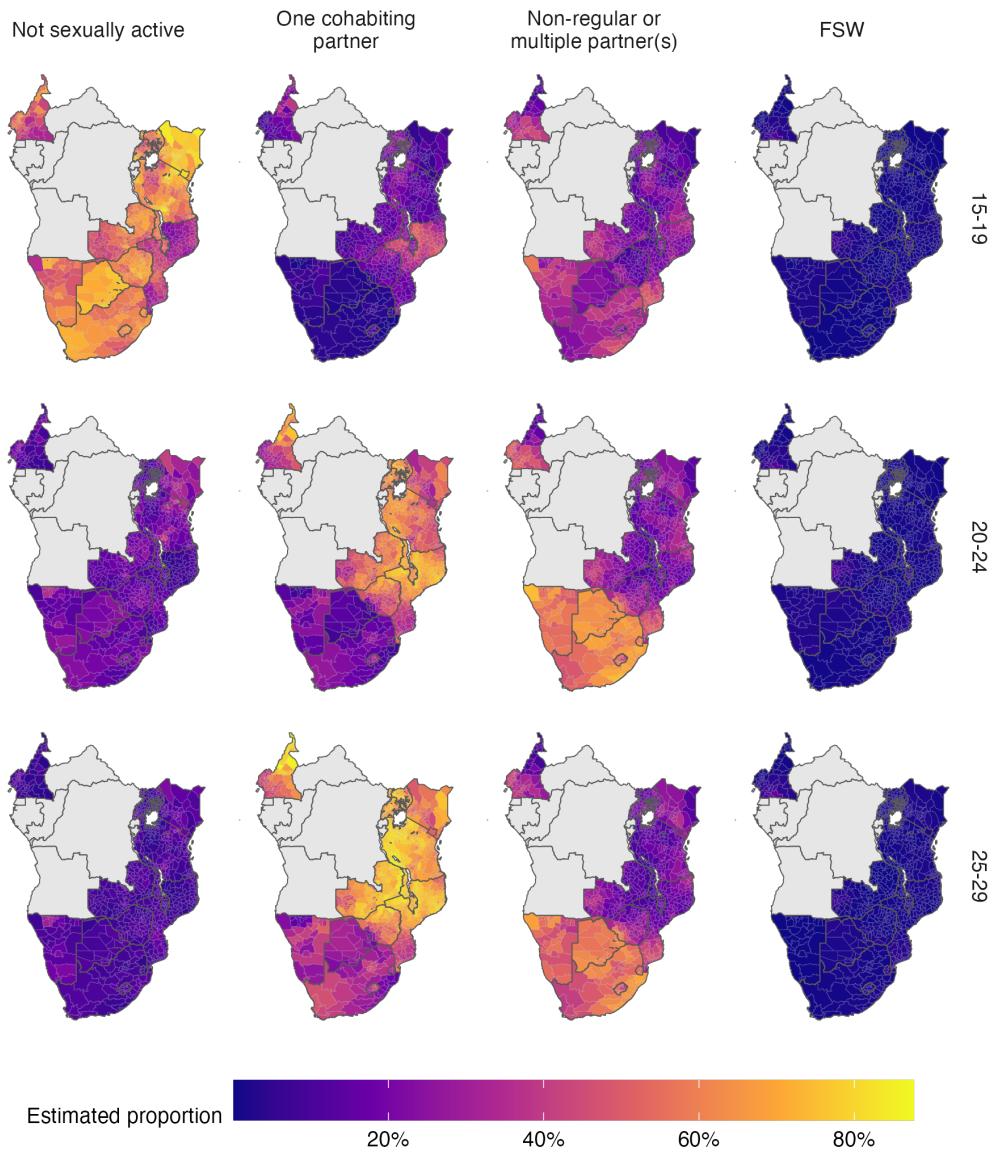
## **5.5 Results**

### **5.5.1 Model for risk group proportions**

#### **5.5.1.1 Estimates**

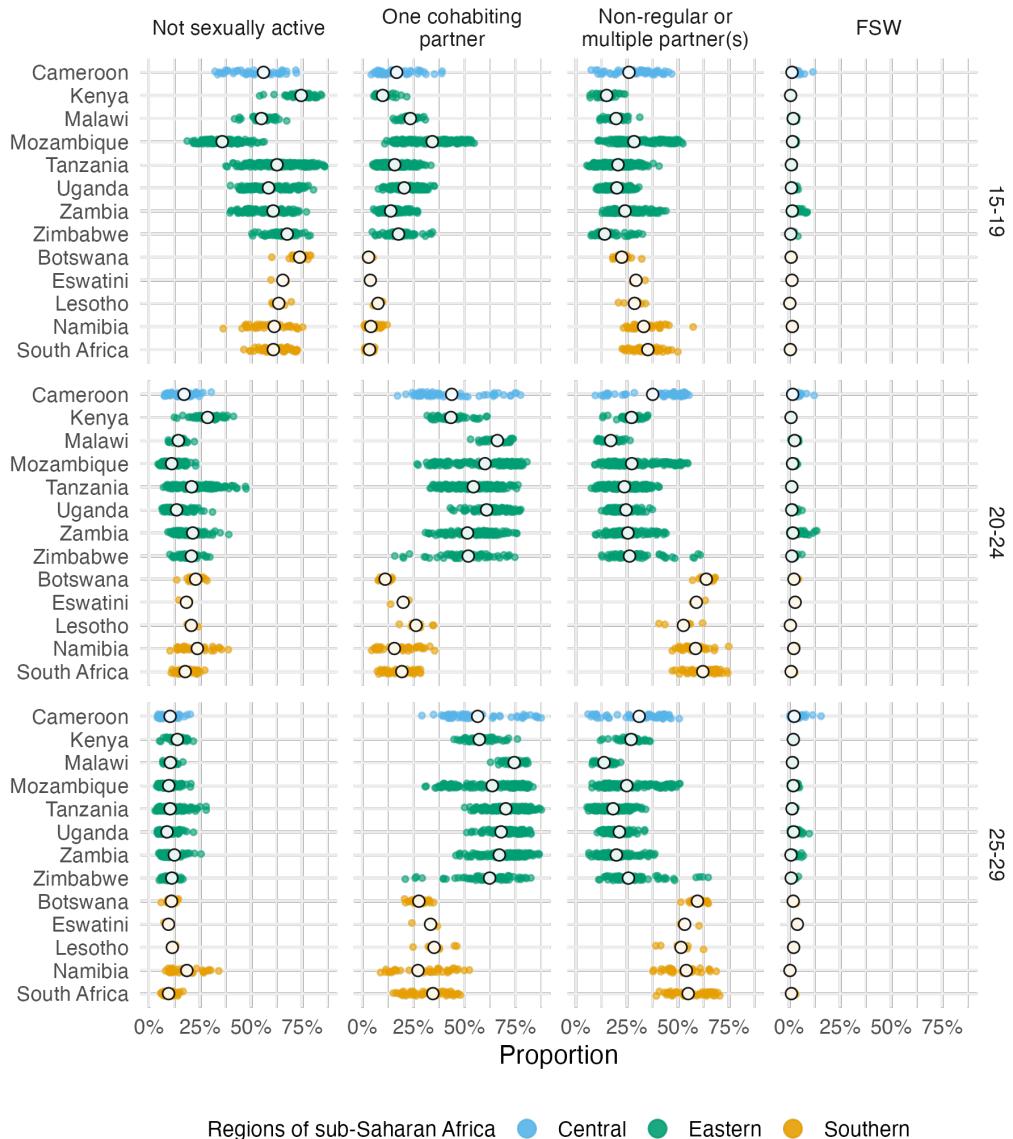
Figure B.1 and Figure 5.8 show posterior mean estimates for the proportion in each risk group for the final model in 2018, the most recent year included in

## *A model for risk group proportions*



**Figure 5.7:** The posterior mean of the AGYW risk group proportions over space 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, though desirable for other reasons, makes it challenging to see spatial variation in the FSW risk group.

*A model for risk group proportions*



**Figure 5.8:** 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 5.7, this figure presents a clear view of within- and between-country variation in risk group proportions.

### *A model for risk group proportions*

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.

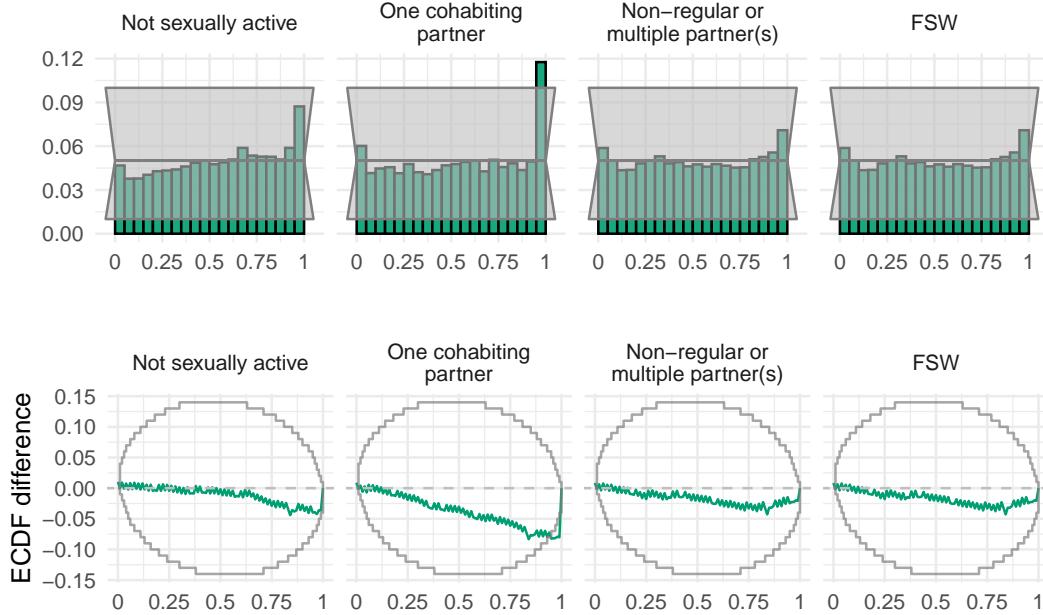
In the 20-24 and 25-29 year age groups, the majority of women were either cohabiting or had non-regular or multiple partner(s). Countries in eastern and central Africa (Cameroon, Kenya, Malawi, Mozambique, Tanzania, Uganda, Zambia and Zimbabwe) had a higher proportion of women in these age groups cohabiting (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 5.8. Figure 5.7 shows the geographic delineation to pass along the border of Mozambique, through the interior of Zimbabwe and along the border of Zambia. The bimodality of the 20-24 and 25-29 year age groups is shown in Figure B.2.

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.

#### **5.5.1.2 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; Nikos I. 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

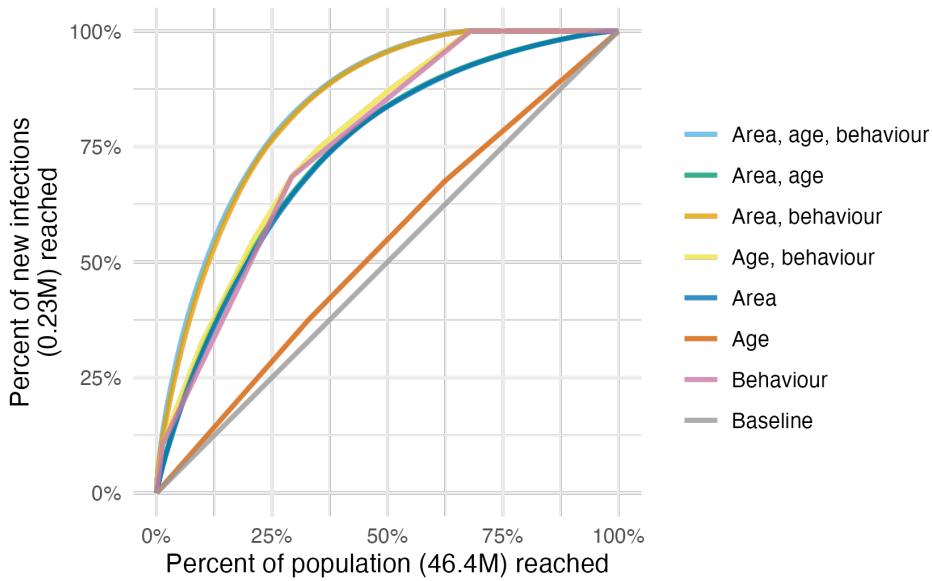
### A model for risk group proportions



**Figure 5.9:** Probability integral transform (PIT) histograms (top row) and empirical cumulative distribution function (ECDF) difference plots (bottom row) for the final selected model.

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. (2022). I found the only significant deviation from uniformity occurred in 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.

## *A model for risk group proportions*



**Figure 5.10:** 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.

### 5.5.1.3 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 primary change in risk group proportions by age group occurs between the 15-19 age group and 20-29 age group (Figure 5.7). 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 B.1) and using other model specifications.

### 5.5.2 Prevalence and incidence by risk group

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

### *A model for risk group proportions*

stratification was included (Figure 5.10). 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 5.5.1.3).

## 5.6 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

### *A model for risk group proportions*

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.

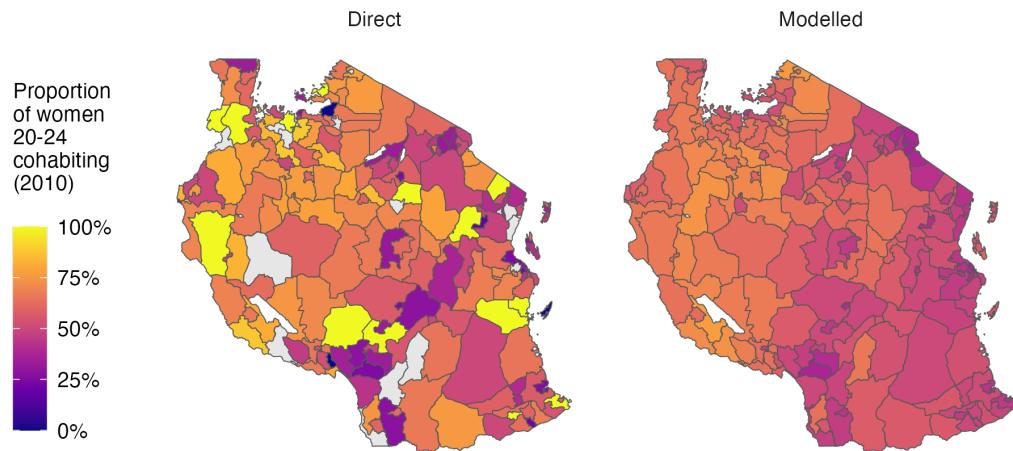
By considering a range of possible risk stratification strategies, I showed that successful implementation of a risk-stratified approach would allow substantially more of those at risk for infections to be identified before infection occurs. A considerable proportion of estimated new infections were 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

## *A model for risk group proportions*



**Figure 5.11:** The modelled estimates display more plausible spatial smoothness than the direct estimates. In addition, missing values in the direct estimates are appropriately infilled by the model.

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 5.11). Third, I provided estimates 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.

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

## *A model for risk group proportions*

Africa, and would allow incidence models to include stratification of individuals by sexual risk.

### **5.6.1 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 Jeffrey W. 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 et al. 2023) 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.

### *A model for risk group proportions*

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.

#### **5.6.2 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.

### *A model for risk group proportions*

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 deterministic Bayesian inference approach, motivated by the Naomi small-area estimation model (Jeffrey W Eaton et al. 2021). Development of the approach required meeting both methodological challenges and implementation difficulties. 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 Håvard 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 (Håvard 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 nearly ubiquitous in applied settings. However, the Naomi model is not compatible with **R-INLA**. The foremost reason is that Naomi is too complex to

be expressed using a formula interface (of the form  $y \sim \dots$ ). Additionally, Naomi has more hyperparameters (moderate-dimensional,  $>20$ ) than can typically be handled using INLA (low-dimensional, certainly below 10). As a result, inferences for the Naomi model have previously been obtained using an empirical Bayes [EB; Casella (1985)] approximation to full Bayesian inference, with Laplace approximation implemented by the more flexible Template Model Builder [TMB; Kristensen et al. (2016)] R package. Under the EB approximation, the hyperparameters are fixed by optimising an approximation to the marginal posterior. This is undesirable as fixing the hyperparameters underestimates their uncertainty. Ultimately, the resulting overconfidence may lead to worse HIV prevention policy decisions.

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 a fast Laplace approximation approach which does not rely on Markov structure of the latent field in the same way as Håvard Rue, Martino, and Chopin (2009). 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 Håvard 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 A. F. 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. 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

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

or some way to compute relevant functions of it. The posterior distribution encapsulates beliefs about the parameters  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_d)$  having observed data  $\mathbf{y} = (y_1, \dots, y_n)$ . 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

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

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

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

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

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

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

$$\int \exp(C h(\mathbf{z})) d\mathbf{z}, \quad (6.5)$$

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

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

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

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

be the posterior mode, and

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

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

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

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

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

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

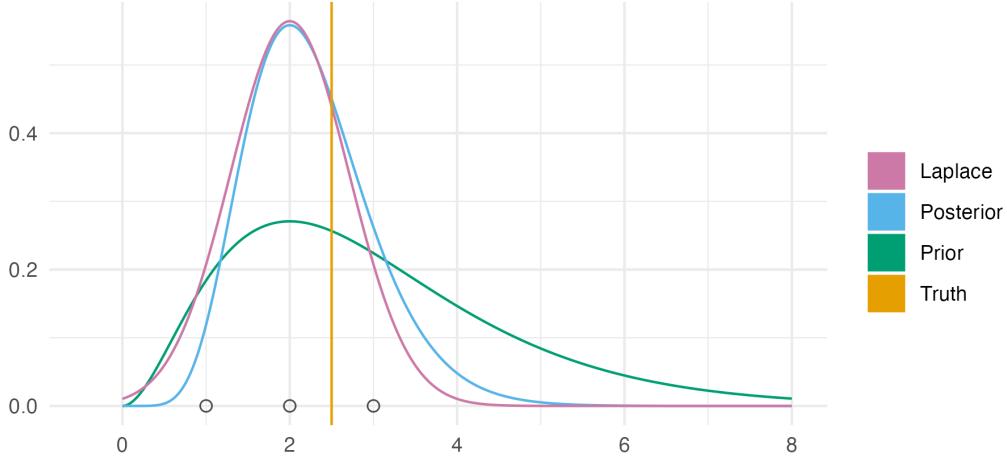
The Laplace approximation may be thought of as approximating the posterior distribution by a Gaussian distribution  $p(\boldsymbol{\phi} | \mathbf{y}) \approx p_{\text{G}}(\boldsymbol{\phi} | \mathbf{y})$  such that

$$\tilde{p}_{\text{LA}}(\mathbf{y}) = \frac{p(\boldsymbol{\phi}, \mathbf{y})}{p_{\text{G}}(\boldsymbol{\phi} | \mathbf{y})} \Big|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}. \quad (6.12)$$

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  $\boldsymbol{\phi}$ .

#### 6.1.1.1 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  $\text{Gamma}(9, 4)$ . 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  $\log \tilde{p}_{\text{LA}}(\mathbf{y}) = -1.882458$ , resulting from the Gaussian approximation  $p_{\mathcal{G}}(\phi | \mathbf{y}) = \mathcal{N}(\phi | \mu = 2, \tau = 2)$ .

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

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

where  $\mathbf{x} = (x_1, \dots, x_N)$  is the latent field, and  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$  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

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

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

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

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  $\tilde{p}_{\mathcal{G}}(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}(\mathbf{x} | \hat{\mathbf{x}}(\boldsymbol{\theta}), \hat{\mathbf{H}}(\boldsymbol{\theta})^{-1})$  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

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

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

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

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

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

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

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

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

is not exactly Gaussian. However, its deviation can be expected to be small if  $\log p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})$  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  $\omega : \mathbb{R}^d \rightarrow \mathbb{R}$  be a weighting function. Then, quadrature can be used to estimate the posterior normalising constant (Equation (6.3)) by

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

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  $\pi$ . A quadrature approximation of this integral is

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

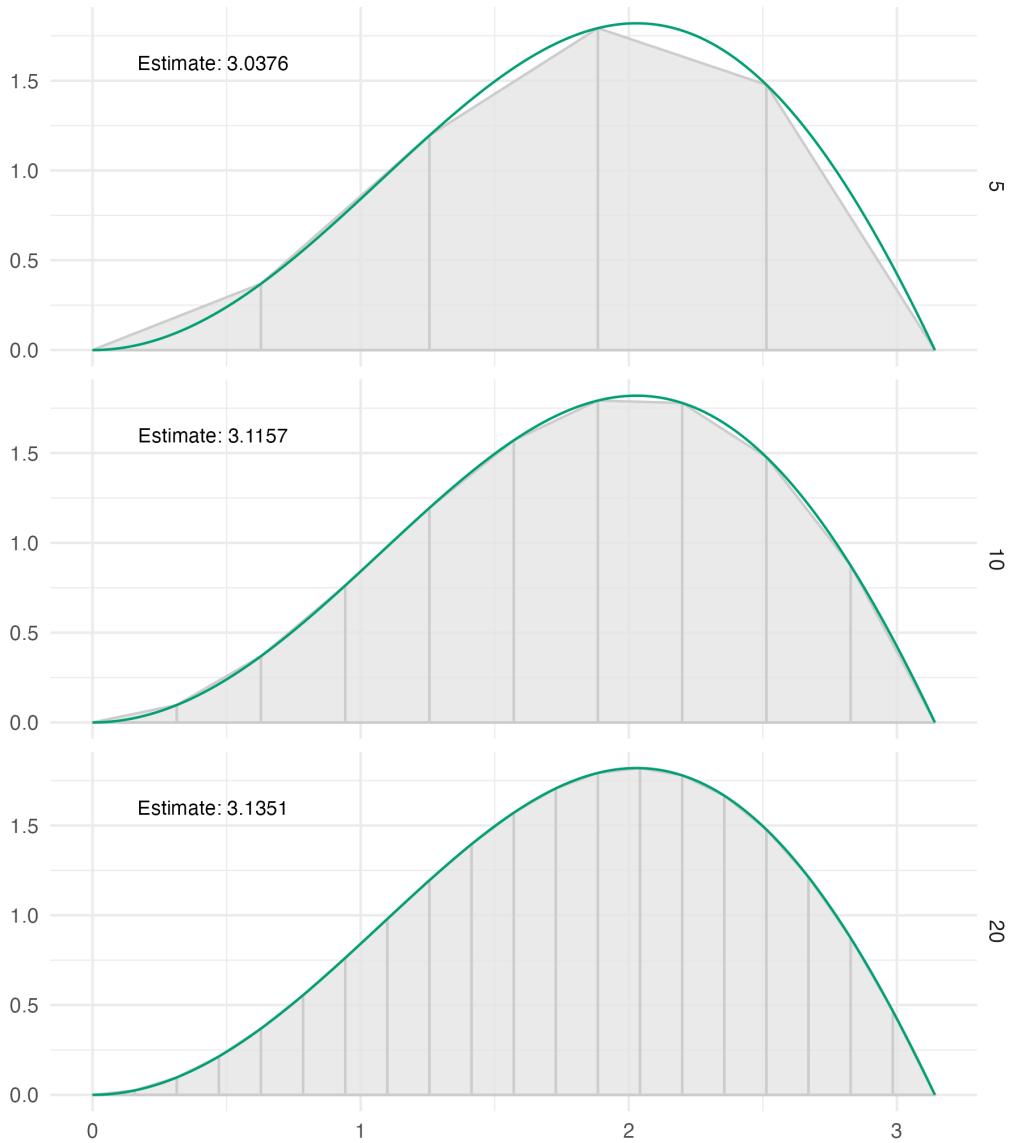
where  $\mathcal{Q} = \{z_1, \dots, z_k\}$  are a set of  $k$  quadrature nodes and  $\omega : \mathbb{R} \rightarrow \mathbb{R}$  is a weighting function.

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

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

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  $\epsilon \rightarrow 0$  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.

### 6.1.2.1 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

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

In this equation, the term  $\varphi(\cdot)$  is a standard multivariate normal density  $\mathcal{N}(\cdot | \mathbf{0}, \mathbf{I})$ , where  $\mathbf{0}$  and  $\mathbf{I}$  are the zero-vector and identity 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 converges 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  $k$ th (probabilist's) Hermite polynomial

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

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

$$\int H_k(z)H_l(z)\varphi(z)dz = \delta_{kl}, \quad (6.27)$$

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  $k$ th Hermite polynomial. For  $k = 1, 2, 3$  these zeros, up to three decimal places, are

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

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

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

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

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

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

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

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

### 6.1.2.2 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 A. F. 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

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

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

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

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

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

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

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

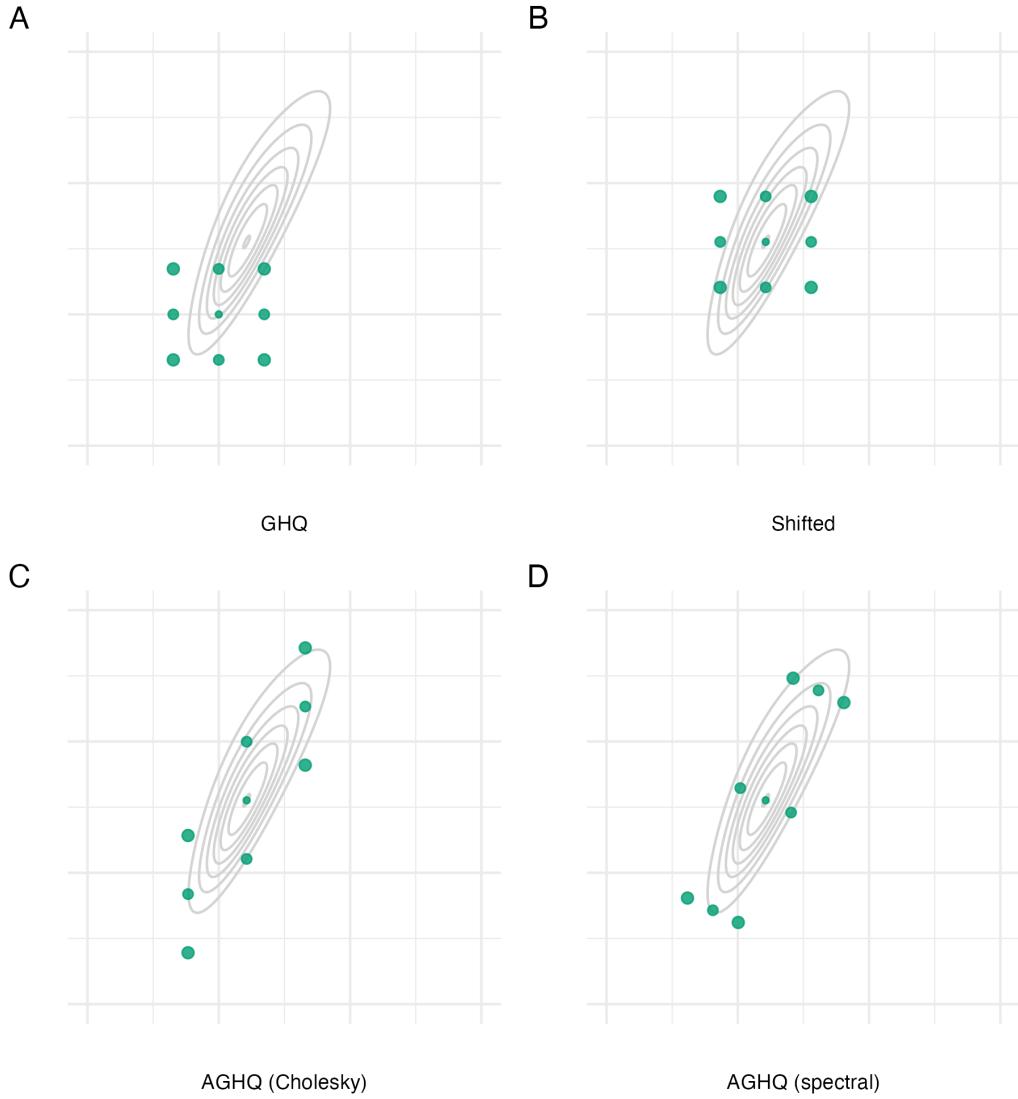
$$\tilde{p}_{\text{AQ}}(\mathbf{y}) = p(\mathbf{y}, \hat{\boldsymbol{\phi}}) \cdot |\hat{\mathbf{P}}| \cdot (2\pi)^{d/2} = p(\mathbf{y}, \hat{\boldsymbol{\phi}}) \cdot \frac{(2\pi)^{d/2}}{|\hat{\mathbf{H}}|^{1/2}}, \quad (6.37)$$

which corresponds to the Laplace approximation  $\tilde{p}_{\text{LA}}(\mathbf{y})$  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

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

is a lower triangular matrix. For the spectral decomposition  $\hat{\mathbf{P}} = \hat{\mathbf{E}}\hat{\Lambda}^{1/2}$ , where  $\hat{\mathbf{E}} = (\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_d)$  contains the eigenvectors of  $\hat{\mathbf{H}}^{-1}$  and  $\hat{\Lambda}$  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  $\alpha \in \mathbb{R}$ .

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 (Håvard Rue, Martino, and Chopin 2009) combines marginal Laplace approximations with quadrature to enable approximation of posterior marginal distributions.

Consider the marginal Laplace approximation (Section 6.1.1.1) for a three-stage hierarchical model given by

$$\tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{\tilde{p}_G(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\hat{\mathbf{x}}(\boldsymbol{\theta})}. \quad (6.39)$$

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)

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

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  $\omega : \mathbb{R}^m \rightarrow \mathbb{R}$  the corresponding weighting function. These nodes are adapted by  $\boldsymbol{\theta}(\mathbf{z}) = \hat{\mathbf{P}}_{\text{LA}} \mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}}$  where

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

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

$$\hat{\mathbf{H}}_{\text{LA}}^{-1} = \hat{\mathbf{P}}_{\text{LA}} \hat{\mathbf{P}}_{\text{LA}}^\top. \quad (6.43)$$

The nested AGHQ estimate of the posterior normalising constant is then

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

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

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

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

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

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.

#### 6.1.3.1 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 (Håvard Rue and Martino 2007)

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

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

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

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

each with multinomial probabilities

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

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 Håvard Rue (2001). The posterior marginals for any subset of the complete vector can simply be obtained by keeping the relevant entries of  $\mathbf{x}$ .

### 6.1.3.2 Laplace marginals

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

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

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

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

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

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

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

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

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

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

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

$$p(x_i | \mathbf{y}) = \int p(x_i, \boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta} \quad (6.56)$$

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

Equation (6.57) is a mixture of the normalised Laplace approximations  $\tilde{p}_{\text{LA}}(x_i, \boldsymbol{\theta} | \mathbf{y})$  over the hyperparameter quadrature nodes. However, unlike the Gaussian case (Section 6.1.3.1) 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. One approach is to chose these nodes based on a one-dimensional AGHQ rule, using the mode and standard deviation of the Gaussian approximation to avoid unnecessary computation of the Laplace marginal mode and standard deviation. The probability density function of the marginal posterior may then be recovered using a Lagrange polynomial or spline interpolant to the log probabilities.

### 6.1.3.3 Simplified Laplace marginals

When the latent field  $\mathbf{x}$  is a Gauss-Markov random fields [GMRF; Havard Rue and Held (2005)] it is possible to efficiently approximate the Laplace marginals in Section 6.1.3.2. 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 6.1.3.1 for location and skewness. Details are left to Section 3.2.3 of Håvard Rue, Martino, and Chopin (2009).

### 6.1.3.4 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 a setting in which, which similar to extended latent Gaussian models [ELGMs; Stringer et al. (2022)], precision matrices are not typically as sparse as GMRFs. Details are left to Wood (2020).

### 6.1.3.5 Augmenting a noisy structured additive predictor to the latent field

Discussion of INLA is concluded by briefly mentioning a difference in implementation between Håvard Rue, Martino, and Chopin (2009) and Stringer et al. (2022). Specifically, Håvard Rue, Martino, and Chopin (2009) augment the latent field to include a noisy structured additive predictor as follows

$$\boldsymbol{\eta}^* = \boldsymbol{\eta} + \boldsymbol{\epsilon}, \quad (6.58)$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \tau^{-1} \mathbf{I}_n), \quad (6.59)$$

$$\mathbf{x}^* = (\boldsymbol{\eta}^*, \mathbf{x}). \quad (6.60)$$

Stringer et al. (2022) (Section 3.2) omit this augmentation, highlighting several drawbacks including: fitting ELGMs, fitting LGMs to large datasets, and theoretical study of the approximation error. Similarly, in what Van Niekerk et al. (2023) (Section 2.1) refer to as the “modern” formula of INLA, the latent field is not augmented. The crux of the issue regards the dimensions and sparsity structure of the Hessian matrix  $\hat{\mathbf{H}}(\boldsymbol{\theta})$ . Details are left to Stringer et al. (2022). Based on these findings, this thesis does not augment the latent field.

## 6.1.4 Software

### 6.1.4.1 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 Håvard Rue 2009). Algorithms for sampling from GMRFs are used from the `GMRFlib` C library (Håvard Rue and Follstad 2001). First and second derivatives are either hard coded, or computed numerically using central finite differences (Fattah et al. 2022). 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>.

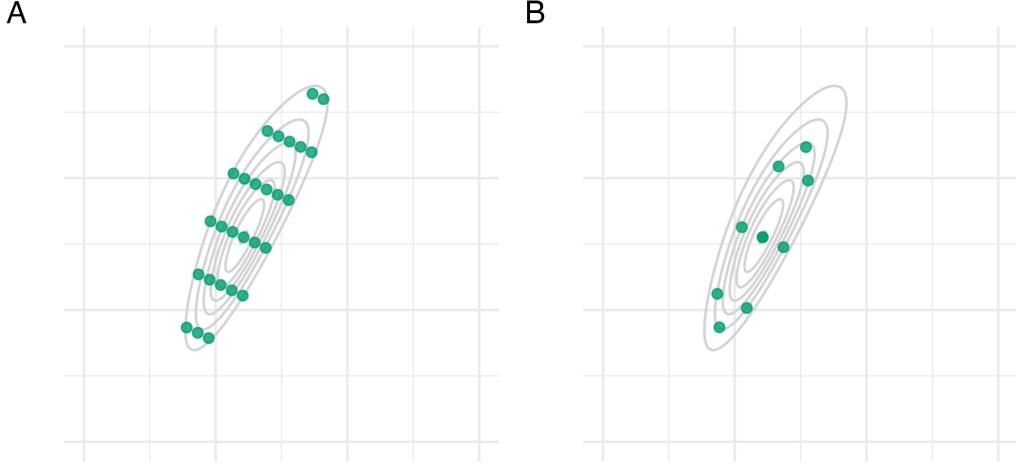
The connection between the latent field  $\mathbf{x}$  and structured additive predictor  $\boldsymbol{\eta}$  is specified in **R-INLA** using a formula interface of the form  $y \sim \dots$ . The interface is similar to that used in the **lm** function in the core **stats** R package. For example, a model with one fixed effect **a** and one IID random effect **b**, has the formula  $y \sim a + f(b, \text{model} = "iid")$ . 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  $\tilde{p}(x_i | y)$  can be chosen by setting **method** to "gaussian" (Section 6.1.3.1), "laplace" (Section 6.1.3.2) or **simplified.laplace** (Section 6.1.3.3). 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 K. B. 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 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. See the **inlabru** package vignettes for additional details. Further inference procedures which leverage **R-INLA** include INLA within MCMC (Gómez-Rubio and Håvard Rue 2018) and importance sampling with INLA (Berild et al. 2022).

#### 6.1.4.2 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



**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 Håvard Rue, Martino, and Chopin (2009). Briefly, equally-weighted quadrature points are generated by starting at the mode and taking steps of size  $\delta_z$  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  $\delta_\pi$ . 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 Håvard 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.

elementary operations may then be composed by repeat use of the chain rule to obtain the function's derivative. A review of AD and how it can be efficiently implemented is provided by C. C. Margossian (2019). TMB uses the C++ package `CppAD` (B. 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 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  $\log p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})$  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 2021) allows use of AGHQ, and AGHQ over the marginal Laplace approximation, via the `mvQuad` R package (Weiser 2016). The `glmmTMB` R package (M. E. 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 Jon Wakefield (2023).

#### 6.1.4.3 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) (Section 6.1.3.4).

## 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 (2020) 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 Jon 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 H. J. Skaug (2009) (a coauthor of **TMB**) in discussion of Håvard 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. The model was used in Section 5.2 of Håvard Rue, Martino, and Chopin (2009), and 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 demonstrates the correspondence between the Laplace marginal implementation developed in **TMB**, and that of **R-INLA** with `method` set to “`laplace`”.
2. Section 6.2.2 considers an extended latent Gaussian model (ELGM) of a tropical parasitic infection. The model was used in Section 5.2 of Bilodeau et al. (2022), and is not compatible with **R-INLA**. This example demonstrates the benefit of a more widely applicable INLA implementation.

### 6.2.1 Epilepsy GLMM

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

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

$$y_{ij} \sim \text{Poisson}(e^{\eta_{ij}}), \quad (6.61)$$

with structured additive predictor

$$\eta_{ij} = \beta_0 + \beta_{\text{Base}} \log(\text{Base}_i/4) + \beta_{\text{Trt}} \text{Trt}_i + \beta_{\text{Trt} \times \text{Base}} (\text{Trt}_i \times \log(\text{Base}_i/4)) \quad (6.62)$$

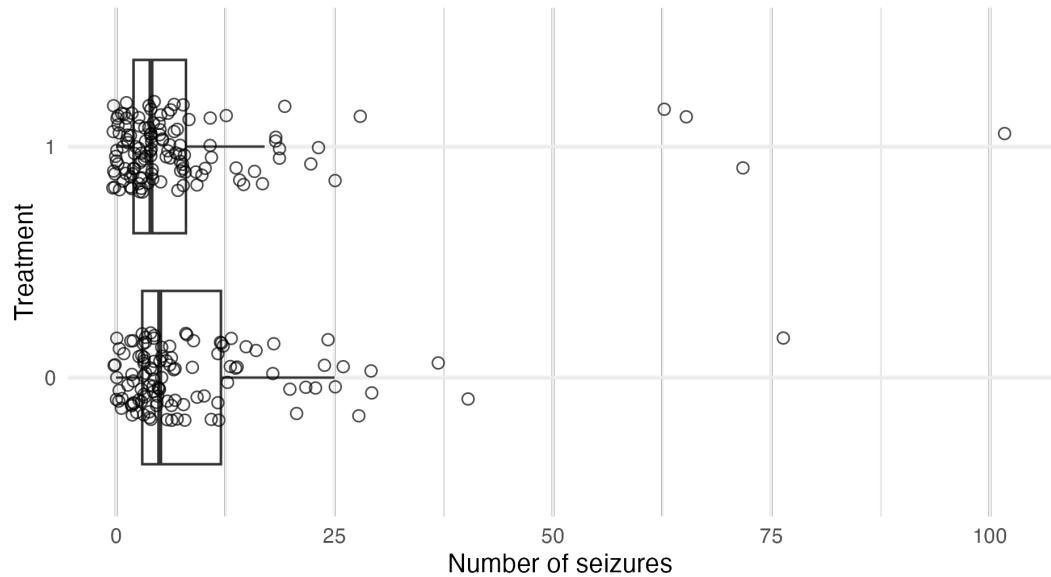
$$+ \beta_{\text{Age}} \log(\text{Age}_i) + \beta_{V_4} V_{4j} + \epsilon_i + \nu_{ij}, \quad i \in [59], \quad j \in [4]. \quad (6.63)$$

The prior distribution on each of the regression parameters, including the intercept  $\beta_0$ , 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 to fit the epilepsy GLMM in Section 6.2.1.

	Method	Software
Section 6.2.1.1	Gaussian, EB	R-INLA
Section 6.2.1.1	Gaussian, grid	R-INLA
Section 6.2.1.1	Laplace, EB	R-INLA
Section 6.2.1.1	Laplace, grid	R-INLA
Section 6.2.1.2	Gaussian, EB	TMB
Section 6.2.1.3	Gaussian, AGHQ	TMB and aghq
Section 6.2.1.4	Laplace, EB	TMB
Section 6.2.1.5	Laplace, AGHQ	TMB and aghq
Section 6.2.1.6	NUTS	tmbstan
Section 6.2.1.7	NUTS	rstan

Inference for the epilepsy GLMM was conducted using a range of approaches (Table 6.1). Section 6.2.1.8 compares the results. The foremost objective of this



**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.

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.

#### 6.2.1.1 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 (6.63) 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 6.1.4.1, `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"`.

### 6.2.1.2 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 than 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 6.2.1.1. 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. For example, the data `y` are passed via:

```
DATA_VECTOR(y);
```

Lines with a `PARAMETER` prefix specify the parameters  $\phi = (\mathbf{x}, \boldsymbol{\theta})$  to be estimated. For example, the regression coefficients  $\boldsymbol{\beta}$  are specified by:

```
PARAMETER_VECTOR(beta);
```

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 `nll == ddist(...)` increment the negative log-posterior, where

`dist` is the name of a distribution. For example, the Gaussian prior distributions on  $\beta$  are implemented by:

```
nll -= dnorm(beta, Type(0), Type(100), true).sum();
```

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

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

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

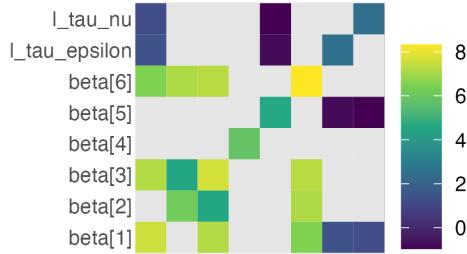
```
obj <- TMB::MakeADFun(
  data = dat,
  parameters = param,
  random = c("beta", "epsilon", "nu"),
  DLL = "epil"
)
```

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

$$(\beta_0, \beta_{\text{Base}}, \beta_{\text{Trt}}, \beta_{\text{Trt} \times \text{Base}}, \beta_{\text{Age}}, \beta_{V_4}, \epsilon_1, \dots, \epsilon_{59}, \nu_{1,1}, \dots, \nu_{59,4}) = (\boldsymbol{\beta}, \boldsymbol{\epsilon}, \boldsymbol{\nu}) = \mathbf{x}. \quad (6.64)$$

The objective function `obj` may then be optimised using a gradient based optimiser to obtain  $\hat{\theta}_{\text{LA}}$ . 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,
```



**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  $\epsilon$  and  $\nu$  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.

```
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  $\hat{\mathbf{H}}_{\text{LA}}$ . 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,
  getJointPrecision = TRUE
)
```

Note that 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")
)
```

### 6.2.1.3 Gaussian marginals and AGHQ with TMB

The objective function `obj` created in Section 6.2.1.2 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 = 1000)$samps
```

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

### 6.2.1.4 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 `x` via the `data` argument of `MakeADFun`. The  $j$ th 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  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\text{LA}}$  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  $(x_i, \hat{\boldsymbol{\theta}}_{\text{LA}})$

$$\tilde{p}(x_i | \mathbf{y}) \approx \tilde{p}_{\text{LA}}(x_i | \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y}) \tilde{p}_{\text{LA}}(\hat{\boldsymbol{\theta}}_{\text{LA}} | \mathbf{y}) \quad (6.65)$$

$$\propto \tilde{p}_{\text{LA}}(x_i, \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y}). \quad (6.66)$$

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  $p_{\mathbf{G}}(\mathbf{x} \mid \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y})$  with the  $i$ th 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  $i$ th 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  $i = 1, \dots, N$ . 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 data items which can be updated (as used in `epil_modified.cpp`) are not supported, so this is not yet possible.

#### 6.2.1.5 Laplace marginals and AGHQ with TMB

The approach taken in Section 6.2.1.4 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

$$\tilde{p}(x_i \mid \mathbf{y}) \propto \sum_{\mathbf{z} \in \mathcal{Q}(m,k)} \tilde{p}_{\text{LA}}(x_i, \boldsymbol{\theta}(\mathbf{z}), \mathbf{y}) \omega(\mathbf{z}). \quad (6.67)$$

As in Section 6.2.1.4 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  $p_{\mathbf{G}}(\mathbf{x} \mid \boldsymbol{\theta}(\mathbf{z}), \mathbf{y})$ . 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. The complete algorithm is given in Appendix C.3.

### 6.2.1.6 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`. Four chains of 2000 iterations, with the first 1000 iterations from each chain discarded as warm-up, were run. Convergence diagnostics are in Appendix C.1.4.1.

### 6.2.1.7 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 being equivalent to the TMB template up to a constant of proportionality. Inferences from `Stan` may be obtained by

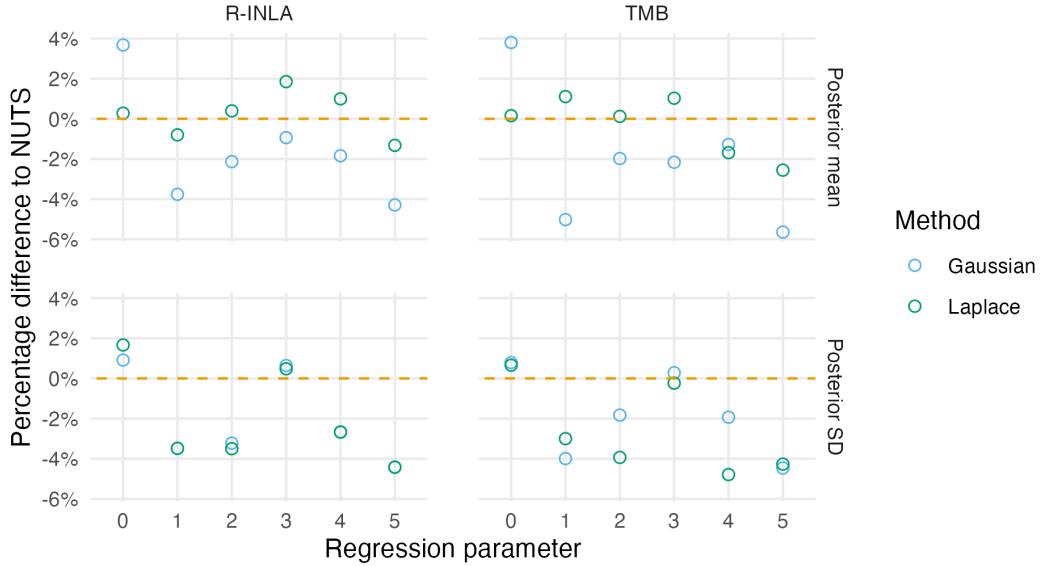
```
fit <- rstan::stan(file = "epil.stan", data = dat, chains = 4)
```

Like for `tmbstan`, four chains of 2000 iterations, with 1000 iterations of burn-in, were run. Convergence diagnostics are in Appendix C.1.4.2.

### 6.2.1.8 Comparison

Posterior means and standard deviations for the six regression parameters  $\beta$  from the inference methods implemented in TMB (Section 6.2.1.2, 6.2.1.3, 6.2.1.3, 6.2.1.5) were highly similar to their R-INLA analogues in Section 6.2.1.1 (Figure 6.7). Posterior distributions obtained were also similar. Figure 6.8 shows ECDF difference plots for Gaussian or Laplace marginals from TMB and R-INLA (as compared with results from NUTS implemented in `tmbstan`) for  $\beta_0$ . These results provide evidence that the implementation of INLA in TMB is correct.

Figures 6.9 shows the number of seconds taken to fit the epilepsy GLMM model for each approach. Gaussian marginals with either EB or AGHQ via TMB were the



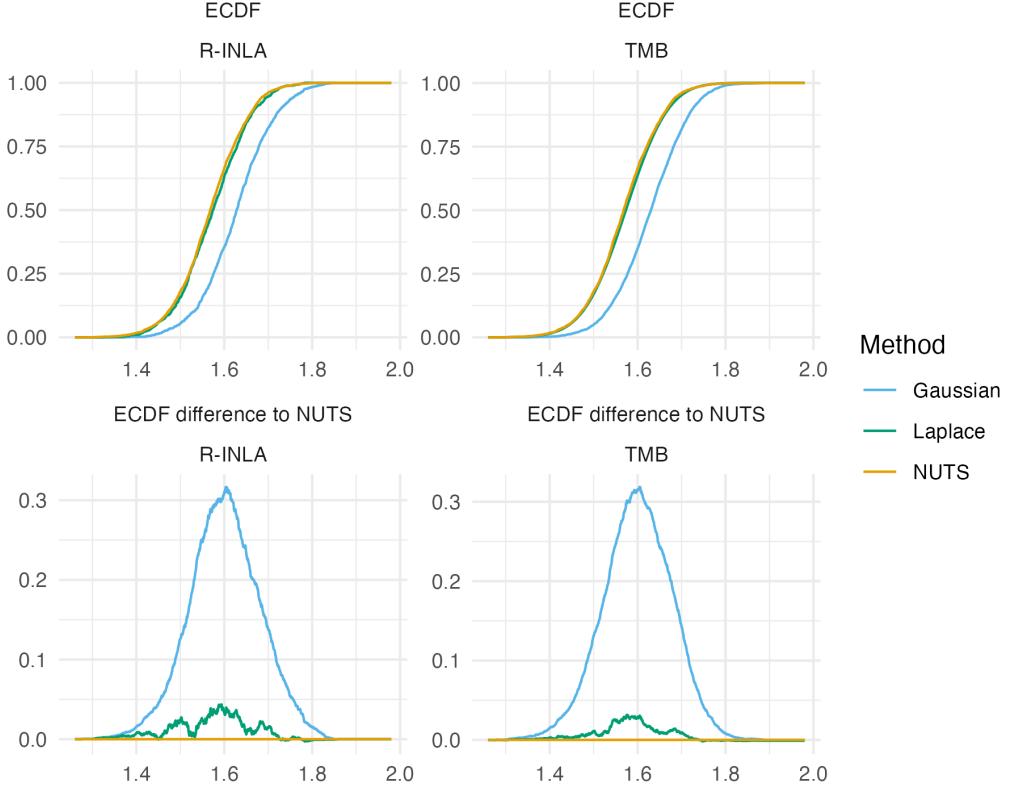
**Figure 6.7:** Percentage difference in posterior summary estimate obtained from NUTS as compared to that obtained from a Gaussian or Laplace marginal with quadrature over the hyperparameters. NUTS results were obtained with `tmbstan`. Results from R-INLA and TMB are similar, especially for the posterior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

fastest approach. All of the approaches using R-INLA took a similar amount of time. The approaches using TMB to implement Laplace marginals were slower than their equivalent in R-INLA. The TMB implementation is relatively naive, based on a simple for loop, and does not use the more advanced approximations of R-INLA. Laplace marginals in TMB with AGHQ ( $k^2 = 3^2 = 9$  quadrature nodes) took 3.4 times as long as Laplace marginals in TMB with EB ( $k^2 = 1^2 = 1$  quadrature node).

For this problem, the `tmbstan` implementation of NUTS took 38.9% of time of the `rstan` implementation. Diagnostics (Figures C.1 and C.2) show that both implementations converged. Monnahan and Kristensen (2018) (Supporting information) found runtime with `rstan` and `tmbstan` to be comparable, so the relatively large difference in this case is surprising.

### 6.2.2 Loa loa ELGM

Bilodeau et al. (2022) considered a ELGM for the prevalence of the parasitic worm Loa loa. Counts of cases  $y_i \in \mathbb{N}^+$  from a sample of size  $n_i \in \mathbb{N}^+$  were obtained from

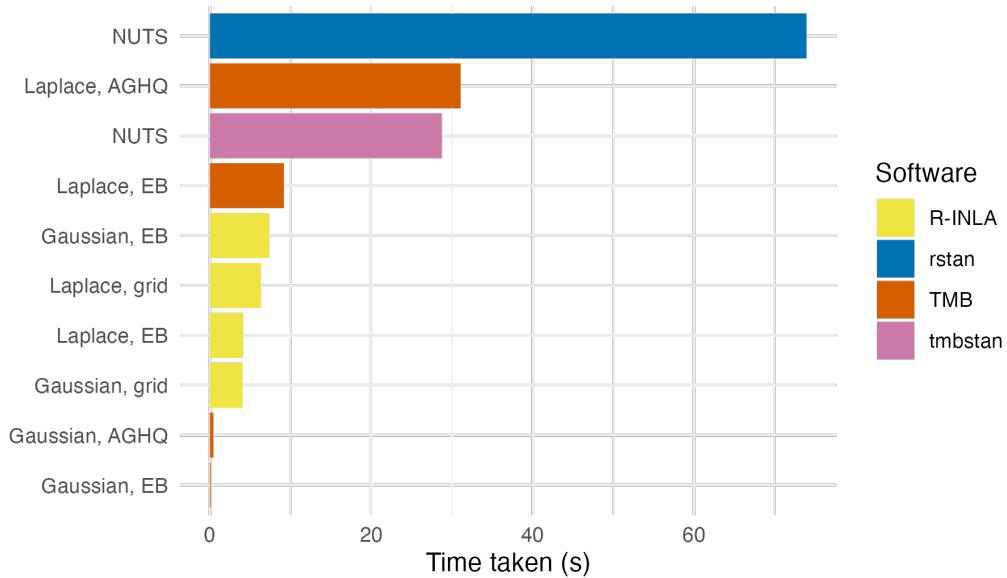


**Figure 6.8:** The ECDF and ECDF difference for the  $\beta_0$  latent field parameter. For this parameter, the Gaussian marginal results are inaccurate, and are corrected almost entirely by the Laplace marginal. An ECDF difference of zero corresponds to obtaining exactly the same results as NUTS, taken to be the gold-standard. Crucially, results obtained using R-INLA and TMB implementations are similar.

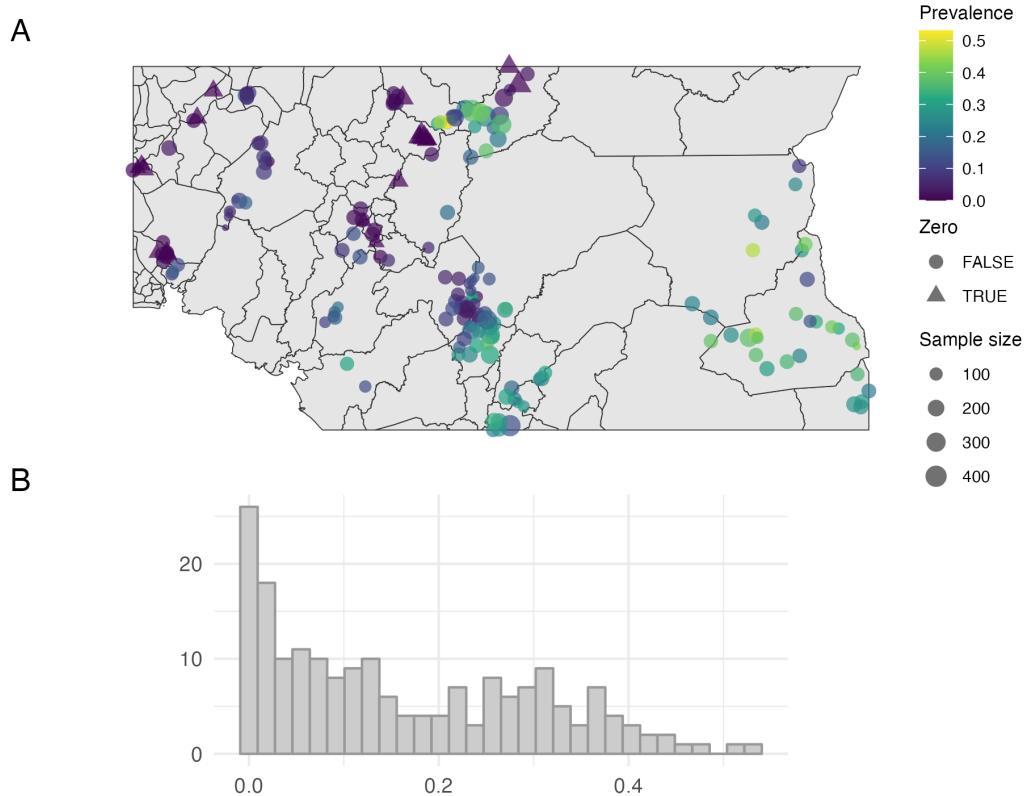
field studies in  $n = 190$  villages in Cameroon and Nigeria [Schlüter et al. (2016); Figure 6.10]. Some areas are thought to be unsuitable for disease transmission, and possibly as a result there are relatively high number of villages with zero prevalence. To account for the possibility of structural zeros, following Diggle and Giorgi (2016), a zero-inflated binomial likelihood was used

$$p(y_i) = (1 - \phi(s_i))\mathbb{I}(y_i = 0) + \phi(s_i)\text{Bin}(y_i \mid n_i, \rho(s_i)) \quad (6.68)$$

where  $s_i \in \mathbb{R}^2$  is the village location,  $\phi(s_i) \in [0, 1]$  is the suitability probability, and  $\rho(s_i) \in [0, 1]$  is the disease prevalence. The prevalence and suitability were



**Figure 6.9:** The number of seconds taken to perform inference for the epilepsy GLMM using each method and software implementation given in Table 6.1.



**Figure 6.10:** Empirical prevalence of Loa loa in 190 sampled villages in Cameroon and Nigeria. The map in Panel A shows the village locations, empirical prevalences, presence of zeros, and sample sizes. The zeros are typically located in close proximity to each other. The histogram in Panel B shows the empirical prevalences, and high number of zeros.

modelled jointly using logistic regressions

$$\text{logit}[\phi(s)] = \beta_\phi + u(s), \quad (6.69)$$

$$\text{logit}[\rho(s)] = \beta_\rho + v(s). \quad (6.70)$$

The two regression coefficients  $\beta_\phi$  and  $\beta_\rho$  were given diffuse Gaussian prior distributions

$$\beta_\phi, \beta_\rho \sim \mathcal{N}(0, 1000). \quad (6.71)$$

Independent Gaussian processes  $u(s)$  and  $v(s)$  were specified by a Matérn kernel (Stein 1999) with shared hyperparameters. Gamma penalised complexity (Simpson et al. 2017; Fuglstad et al. 2019) prior distributions were used for the standard deviation  $\sigma$  and range  $\rho$  hyperparameters such that (Brown 2015)

$$\mathbb{P}(\sigma < 4) = 0.975, \quad (6.72)$$

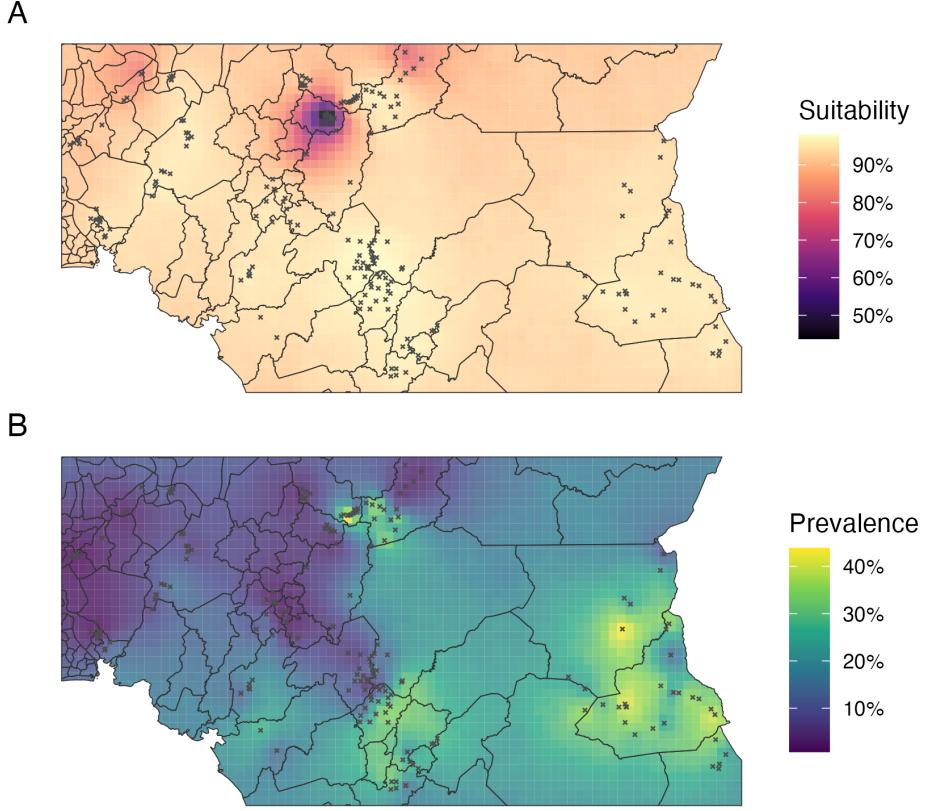
$$\mathbb{P}(\rho < 200\text{km}) = 0.975. \quad (6.73)$$

The smoothness parameter  $\nu$  was fixed to 1.

The zero-inflated likelihood in Equation (6.68) is not compatible with **R-INLA**. Section 2.2 of Brown (2015) demonstrates use of **R-INLA** to fit a simpler LGM model which includes covariates. Instead, Bilodeau et al. (2022) implemented this model in **TMB**. Inference was then performed using Gaussian marginals and AGHQ via **aghq** and NUTS via **tmbstan**. This section considers inference using three approaches (Table 6.2), extending Bilodeau et al. (2022) by including AGHQ with Laplace marginals.

**Table 6.2:** The inference methods and software considered to fit the Loa loa ELGM in Section 6.2.2.

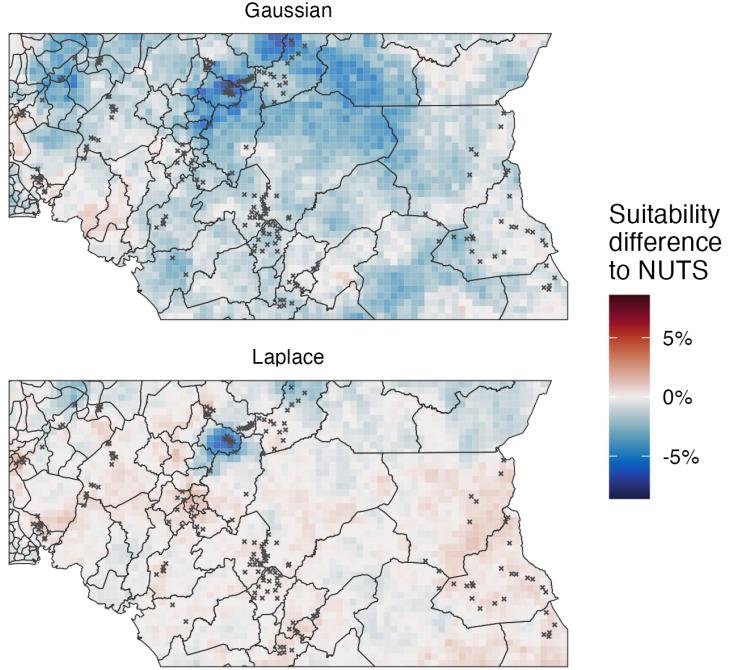
Method	Software	Details
Gaussian, AGHQ	TMB and <b>aghq</b>	$k = 3$
Laplace, AGHQ	TMB and <b>aghq</b>	$k = 3$
NUTS	<b>tmbstan</b>	4 chains of 5000 iterations, with default NUTS settings as implemented in <b>rstan</b> (Carpenter et al. 2017)



**Figure 6.11:** Posterior mean of the suitability  $\mathbb{E}[\phi_{\text{LA}}(s)]$  (Panel A) and prevalence  $\mathbb{E}[\rho_{\text{LA}}(s)]$  (Panel B) random fields computed using Laplace marginals. Inferences over this fine spatial grid were using conditional Gaussian field simulation as implemented by `gstat::krige`.

Bilodeau et al. (2022) found that NUTS did not converge for the full model, but did converge when the values of  $\beta_\phi$  and  $\beta_\rho$  were fixed at their posterior mode (obtained using AGHQ with Gaussian marginals). To allow for comparison between Gaussian and Laplace marginals, the same approach was taken here.

After obtaining posterior inferences at each  $s_i$ , the `gstat::krige` function (E. J. Pebesma 2004) was used to implement conditional Gaussian field simulation [E. Pebesma and R. Bivand (2023); Chapter 12] over a fine spatial grid. Independent latent field and hyperparameter samples were used in each conditional simulation. For each method (Table 6.2) 500 conditional Gaussian field simulations were obtained.



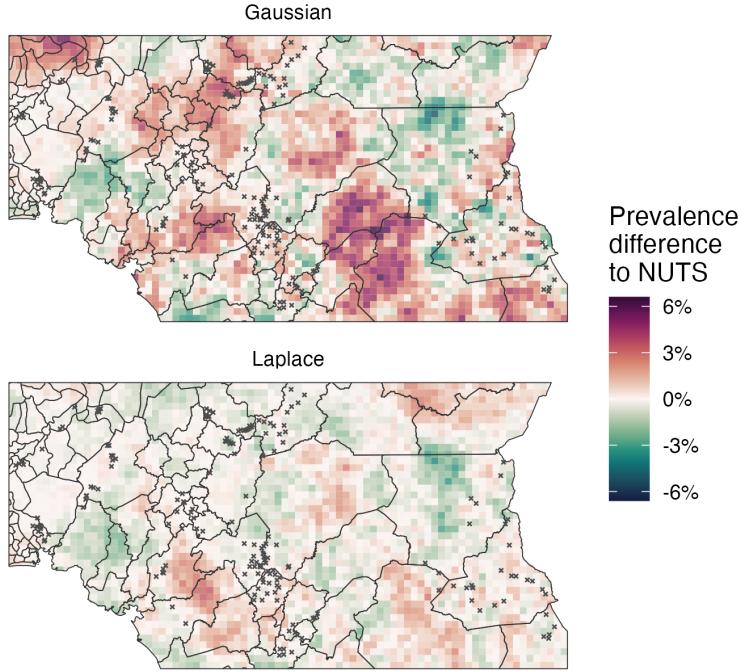
**Figure 6.12:** Difference between the suitability posterior means with Gaussian marginals  $\mathbb{E}[\phi_G(s)]$  and Laplace marginals  $\mathbb{E}[\phi_{LA}(s)]$  to NUTS results. While the Gaussian approximation appears to systematically underestimate suitability, results from the Laplace approximation are substantially closer to results from NUTS. As  $\beta_\phi$  was fixed this difference is as a result in differences in estimation of  $u(s)$ . The diverging colour palette used in this figure is from Thyng et al. (2016).

### 6.2.2.1 Results

Figure 6.11 shows the suitability and prevalence posterior means across the fine grid obtained using AGHQ with Laplace marginals.

For both the suitability and prevalence posterior mean, using Laplace marginals rather than Gaussian marginals substantially reduced error compared to NUTS (Figures 6.12 and 6.13). As the hyperparameter posteriors for each approach were the same, differences in Gaussian field simulation results were due to differences in latent field posterior marginals at each of the 190 sites, shown in Figure 6.14. At some sites, the differences in ECDF were substantial (Figure 6.15). Figure C.3 shows that the results from NUTS were suitable for use, and therefore that this comparison is valid.

Laplace marginals with AGHQ took 12% of time taken (23.1 hours) by NUTS (Figure 6.16). That said, Gaussian marginals with AGHQ took less than a minute to run: substantially less than the 2.77 hours taken by the Laplace marginals.

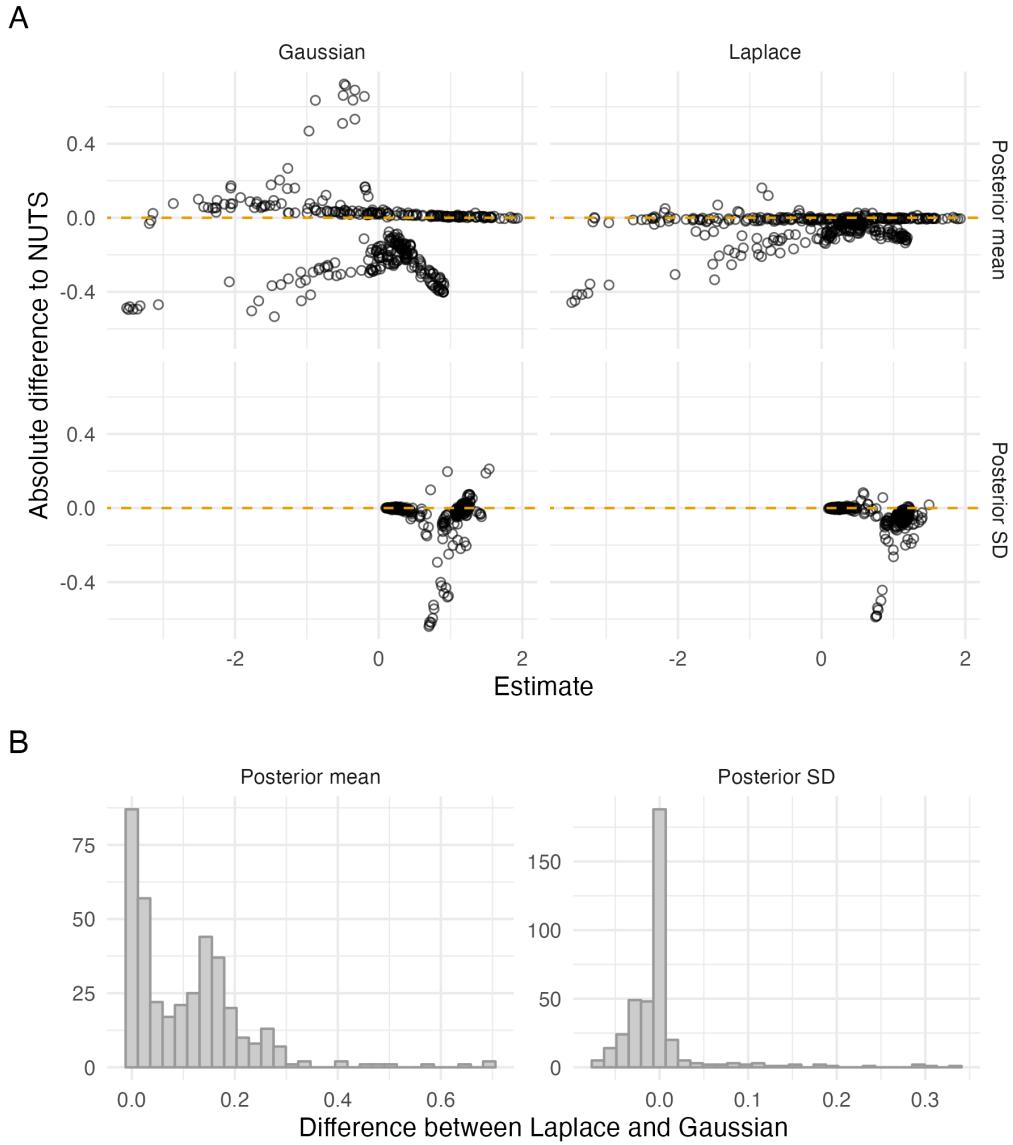


**Figure 6.13:** Difference between the prevalence posterior means with Gaussian marginals  $\mathbb{E}[\rho_G(s)]$  and Laplace marginals  $\mathbb{E}[\rho_{LA}(s)]$  to NUTS results. Like the suitability in Figure 6.12, the error the the Gaussian approximation is higher than that of the Laplace approximation. As  $\beta_\rho$  was fixed this difference is as a result in differences in estimation of  $v(s)$ . The diverging colour palette used in this figure is from Thyng et al. (2016).

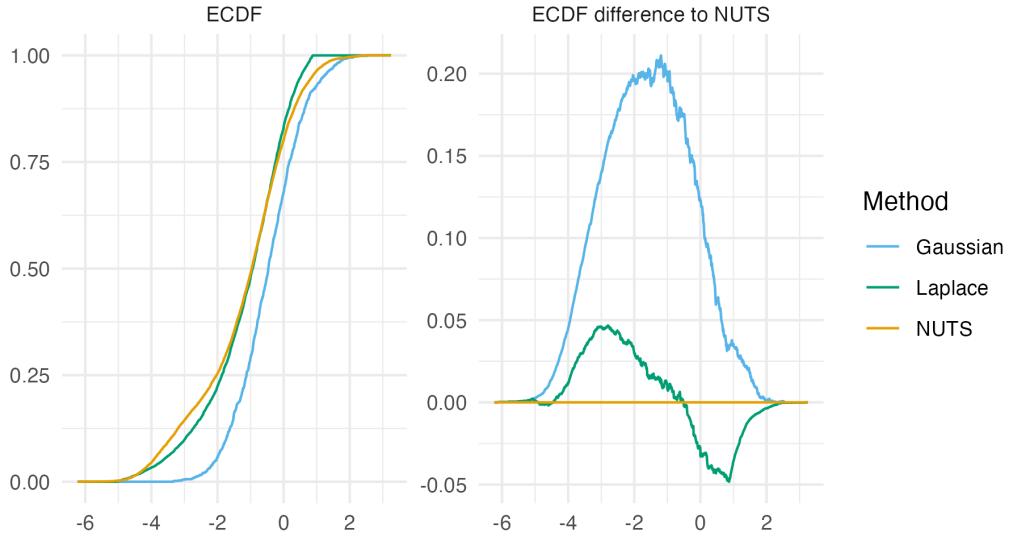
A less naive Laplace implementation may achieve a runtime more competitive to the Gaussian.

### 6.3 The Naomi model

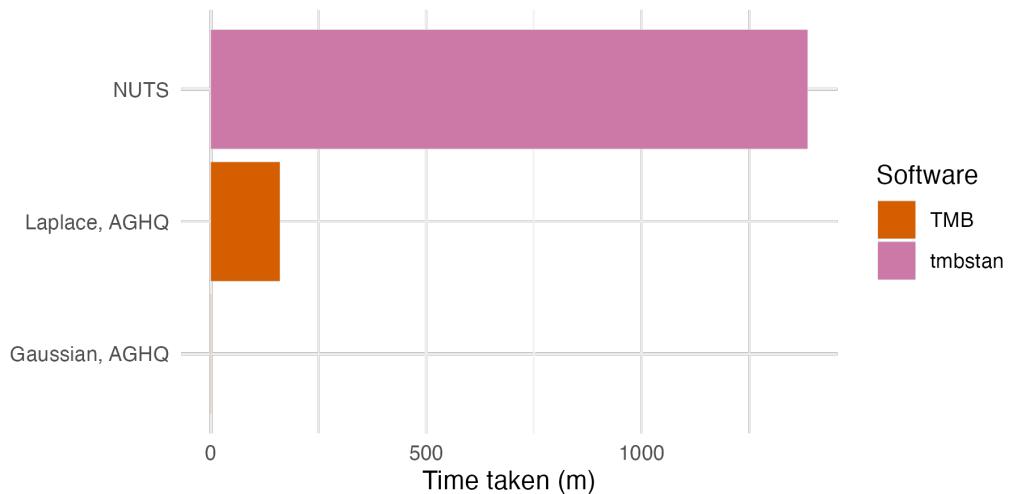
The work in this chapter was conducted in search of a fast and accurate Bayesian inference method for the Naomi model (Jeffrey W Eaton et al. 2021). Software has been developed for Naomi to allow countries to input their data and interactively generate estimates during week long workshops as a part of a yearly process supported by UNAIDS. Generation of estimates by country teams, rather than external agencies or researchers, is an important and distinctive feature of the HIV response. Drawing on expertise closest to the data being modelled improves the accuracy of the process, as well as strengthening trust in the resulting estimates, creating a virtuous cycle of data quality, use and ownership (Noor 2022). To



**Figure 6.14:** Absolute difference between the Gaussian and Laplace marginal posterior means and standard deviations to NUTS results at each  $u(s_i), v(s_i) : i \in [190]$ . Relative differences are in Figure C.4. For close to every node, the Laplace approximation produced a more accurate posterior mean than the Gaussian approximation. For the posterior standard deviation (SD), the picture was more mixed.



**Figure 6.15:** The element of the latent field with maximum difference in absolute difference to NUTS for the posterior mean was  $u_{184}$ . While the Gaussian approximation has substantial error as compared with NUTS, the Laplace approximation is a close match.



**Figure 6.16:** The number of minutes taken to perform inference for the Loa loa ELGM using each approach given in Table 6.2.

allow interactive review and iteration of model results by workshop participants, any inference procedure for Naomi should ideally be fast and have low memory usage. Additionally, it should be reliable and automatic, across a range of country settings. Naomi is a complex model, comprised of multiple linked generalized linear mixed models (GLMMs), and as such these requirements present a challenging Bayesian inference problem.

This section begins (Section 6.3.1) by describing a simplified version of Naomi. 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. 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 likely transferable to the complete model. Description of some features of the simplified model is left to the more exhaustive Appendix C.4. After outlining the model, Section 6.3.2 explains why it is an ELGM (Stringer et al. 2022) rather than an LGM (Håvard 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 6.3.1.2);
- the antenatal care (ANC) clinic testing component (Section 6.3.1.4);
- the antiretroviral therapy (ART) attendance component (Section 6.3.1.4).

After specifying common notation used throughout the model (Section 6.3.1.1) each of these components is described in turn.

### 6.3.1.1 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  $I \in \mathcal{I}$ .

### 6.3.1.2 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

$$\text{logit}(\rho_i) = \eta_i^\rho, \quad (6.74)$$

and

$$\text{logit}(\alpha_i) = \eta_i^\alpha. \quad (6.75)$$

HIV incidence rate is modelled on the log scale as

$$\log(\lambda_i) = \eta_i^\lambda. \quad (6.76)$$

The structured additive predictor  $\eta_i^\lambda$  includes terms for adult HIV prevalence and adult ART coverage. The proportion recently infected among HIV positive persons is linked to HIV incidence via

$$\kappa_i = 1 - \exp\left(-\lambda_i \cdot \frac{1 - \rho_i}{\rho_i} \cdot (\Omega_T - \beta_T) - \beta_T\right), \quad (6.77)$$

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

The three processes in Equations (6.74), (6.75), and (6.76) are each primarily informed by household survey data. Let  $j$  denote a surveyed individual, in district-age-sex strata  $i(j)$ . Weighted aggregate survey observations are calculated based on individual responses  $\theta_j \in \{0, 1\}$  as

$$\hat{\theta}_I = \frac{\sum_{i(j) \in I} w_j \cdot \theta_j}{\sum_{i(j) \in I} w_j}, \quad (6.78)$$

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 (6.78) by the Kish effective sample size [ESS; Kish (1965)]

$$y_I^\theta = m_I^\theta \hat{\theta}_I, \quad (6.79)$$

where

$$m_I^\theta = \frac{\left(\sum_{i(j) \in I} w_j\right)^2}{\sum_{i(j) \in I} w_j^2}. \quad (6.80)$$

As the Kish ESS is maximised by constant survey weights, in exchange for reducing bias, survey weighting increases variance. Equations (6.78) and (6.80) 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  $\alpha$  and the proportion recently infected among HIV positive persons  $\kappa$ , only those individuals who are HIV positive are included in the set. The denominator for HIV prevalence  $\rho$  includes all individuals.

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

$$y_I^\theta \sim \text{xBin}(m_I^\theta, \theta_I). \quad (6.81)$$

The terms  $\theta_I$  are the following weighted aggregates

$$\rho_I = \frac{\sum_{i \in I} N_i \rho_i}{\sum_{i \in I} N_i}, \quad \alpha_I = \frac{\sum_{i \in I} N_i \rho_i \alpha_i}{\sum_{i \in I} N_i \rho_i}, \quad \kappa_I = \frac{\sum_{i \in I} N_i \rho_i \kappa_i}{\sum_{i \in I} N_i \rho_i}, \quad (6.82)$$

where the denominators of  $\alpha_I$  and  $\kappa_I$  reflect their restriction to HIV positive persons.

### 6.3.1.3 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

$$\text{logit}(\rho_i^{\text{ANC}}) = \text{logit}(\rho_i) + \eta_i^{\rho^{\text{ANC}}}, \quad (6.83)$$

$$\text{logit}(\alpha_i^{\text{ANC}}) = \text{logit}(\alpha_i) + \eta_i^{\alpha^{\text{ANC}}}. \quad (6.84)$$

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 (6.82) 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  $\Psi_i$  are the number of pregnant women, which are assumed to be fixed.

#### 6.3.1.4 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 et al. 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  $x$

$$\tilde{\gamma}_{x,x'} = \text{logit}(\gamma_{x,x'}) = \eta_x^{\tilde{\gamma}}. \quad (6.85)$$

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

$$y_I^A = \sum_{i \in I} y_i^A. \quad (6.86)$$

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

$$\pi_{i,x(i)=x,x'} = \rho_i \alpha_i \gamma_{x(i)=x,x'}. \quad (6.87)$$

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

1. the probability of having HIV  $\rho_i$ ,
2. the probability of taking ART  $\alpha_i$ ,
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

$$A_i = \sum_{x' \sim x, x' = x} A_{i,x(i)=x',x}. \quad (6.88)$$

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

$$A_{i,x(i)=x',x} \sim \text{Bin}(N_{i,x(i)=x'}, \pi_{i,x(i)=x',x}). \quad (6.89)$$

Each aggregate attendance observation (Equation (6.86)) 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

$$y_I^A \sim \mathcal{N}(\mu_I^A, \sigma_I^{A^2}), \quad (6.90)$$

where the mean is

$$\mu_I^A = \sum_{i \in I} \sum_{x' \sim x, x' = x} N_{i,x(i)=x'} \cdot \pi_{i,x(i)=x',x}, \quad (6.91)$$

and the variance is

$$\sigma_I^A = \sum_{i \in I} \sum_{x' \sim x, x' = x} N_{i,x(i)=x'} \cdot \pi_{i,x(i)=x',x} \cdot (1 - \pi_{i,x(i)=x',x}). \quad (6.92)$$

Equations (6.91) and (6.92) 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

$$\mathbf{y} = (y_I^\theta), \quad \theta \in \{\rho, \alpha, \kappa, \rho^{\text{ANC}}, \alpha^{\text{ANC}}, A\}, \quad I \in \mathcal{I}. \quad (6.93)$$

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.4.

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 6.3.1.2, HIV incidence depends on district-level adult HIV prevalence and ART coverage (Equation (C.19))). 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 ( $\{15\text{-}19, \dots, 45\text{-}49\}$ ), 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 6.3.1.2, the proportion recently infected  $\kappa_i$  is given by a non-linear function (Equation (6.77)) of HIV incidence  $\lambda_i$ , HIV prevalence  $\rho_i$ , mean duration of recent infection  $\Omega_T$  and proportion of long-term HIV infections misclassified as recent  $\beta_T$ . Though arguably a contorting of the ELGM framework, by considering  $\Omega_T$  and  $\beta_T$  as (Gaussian) linear predictors, then each  $\kappa_i$  depends on four structured additive predictors.
5. In Section 6.3.1.3, 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 6.3.1.3, nested binomial likelihoods are used.
7. In Section 6.3.1.4 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 6.3.1.4 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 was previously conducted using a marginal Laplace approximation, and optimisation over the hyperparameters, implemented using **TMB**. This approach was illustrated for the epilepsy example in Section 6.2.1.2 and is analogous for Naomi. It would be desirable to instead integrate with respect to the hyperparameters, taking an INLA-like approach as described in Section 6.1.3.

Section 6.2 attends to part of the challenge, by developing INLA methods which compatible with the Naomi model log-posterior as implemented in **TMB**. However, naive quadrature methods are not directly applicable to Naomi. This is because Naomi has  $m = 24$  hyperparameters. Although  $m = 24$  cannot be described as high-dimensional, it is certainly more than the  $m < 4$  or so hyperparameters typical for use of INLA. Hence here the term moderate-dimensional is used. 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  $\theta$ , it is possible to use a variable number of nodes per dimension. In line with the terminology used in the `mvQuad` package, the number of nodes per dimension are referred to as levels. Let  $\mathbf{k} = (k_1, \dots, k_m)$  be a vector of levels, where each  $k_j \in \mathbb{Z}^+$ . A GHQ grid with (potentially) variable levels is then given by

$$\mathcal{Q}(m, \mathbf{k}) = \mathcal{Q}(1, k_1) \times \cdots \times \mathcal{Q}(1, k_m). \quad (6.94)$$

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

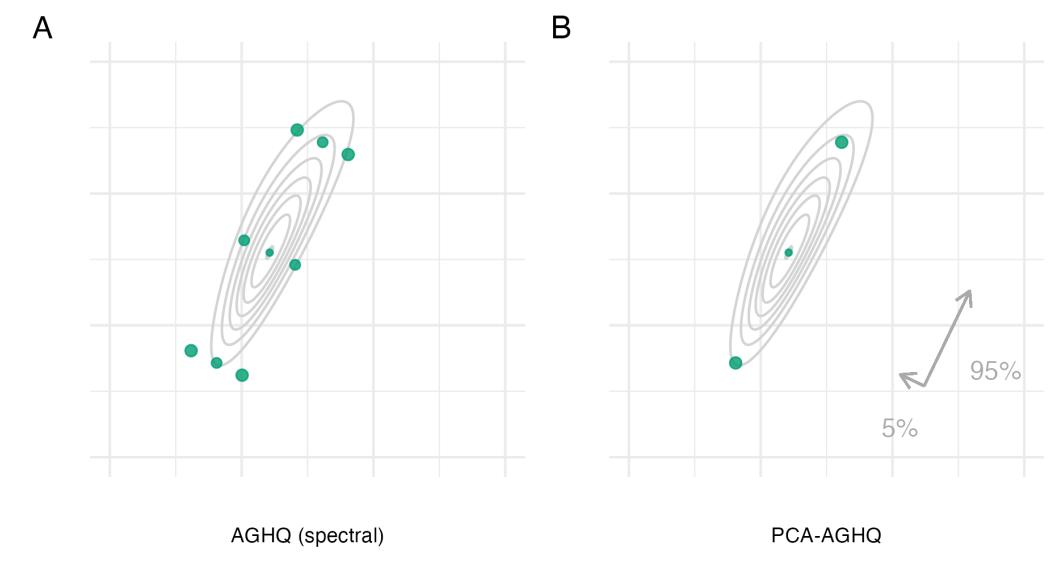
$$\omega(\mathbf{z}) = \prod_{j=1}^m \omega_{k_j}(z_j). \quad (6.95)$$

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 6.17 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,



**Figure 6.17:** 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 usual AGHQ nodes with a spectral matrix decomposition. 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.

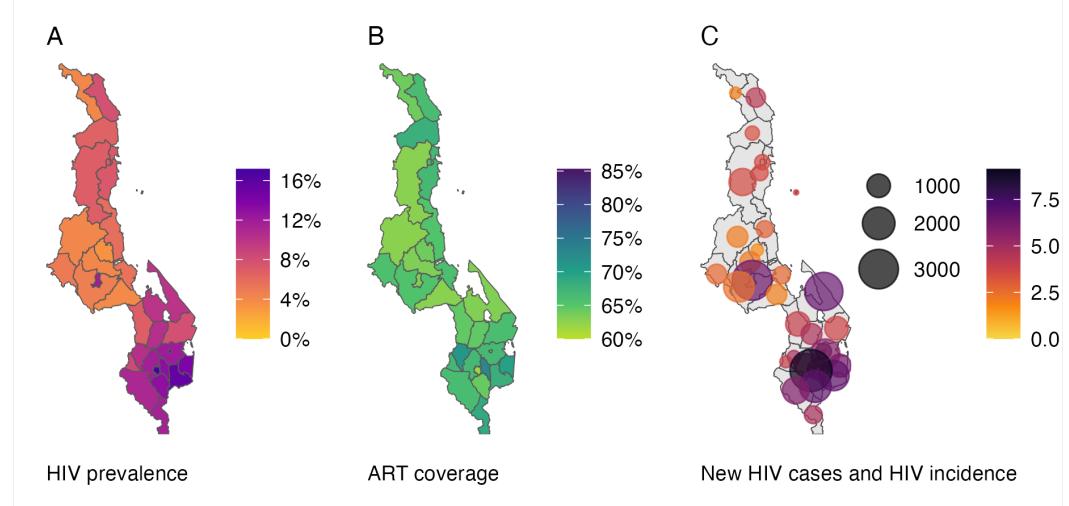
and a Laplace approximation on the remaining  $m - s$  principal components. As such, it may be argued that PCA-AGHQ provides a natural compromise between the EB and AGHQ integration strategies.

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

$$\tilde{p}_{\text{PCA}}(\mathbf{y}) = |\hat{\mathbf{E}}_{\text{LA}} \hat{\Lambda}_{\text{LA}}^{1/2}| \sum_{\mathbf{z} \in \mathcal{Q}(m, s, k)} \tilde{p}_{\text{LA}}(\hat{\mathbf{E}}_{\text{LA}, s} \hat{\Lambda}_{\text{LA}, s}^{1/2} \mathbf{z} + \hat{\boldsymbol{\theta}}_{\text{LA}}, \mathbf{y}) \omega(\mathbf{z}), \quad (6.96)$$

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

$$\omega(\mathbf{z}) = \prod_{j=1}^s \omega_s(z_j) \times \prod_{j=s+1}^d \omega_1(z_j). \quad (6.97)$$



**Figure 6.18:** District-level HIV prevalence, ART coverage, and new HIV cases and HIV incidence for adults 15-49 in Malawi. Inference here was conducted using a Gaussian approximation and EB via TMB.

## 6.5 Malawi case-study

This section presents a case-study of approximate Bayesian inference methods applied to the Naomi model in Malawi. Data from Malawi has previously been used to demonstrate the Naomi model, including as a part of the `naomi` R package vignette available from <https://github.com/mrc-ide/naomi>. Malawi was chosen for the vignette and this case-study in part because it has a small number of districts,  $n = 30$ , limiting the computational demand of the model.

Three Bayesian inference approaches were considered:

1. Gaussian marginals and EB with TMB. This approach was previously used in production for Naomi. As short-hand, this approach is referred to as GEB.
2. Gaussian marginals and PCA-AGHQ with TMB. This is a novel approach, enabled by the methodological work of Section 6.4. As short-hand, this approach is referred to as GPCA-AGHQ.
3. NUTS with `tmbstan`. Conditional on assessing chain convergence and suitability, to be discussed in Section 6.5.1, inferences from NUTS represent a gold-standard.

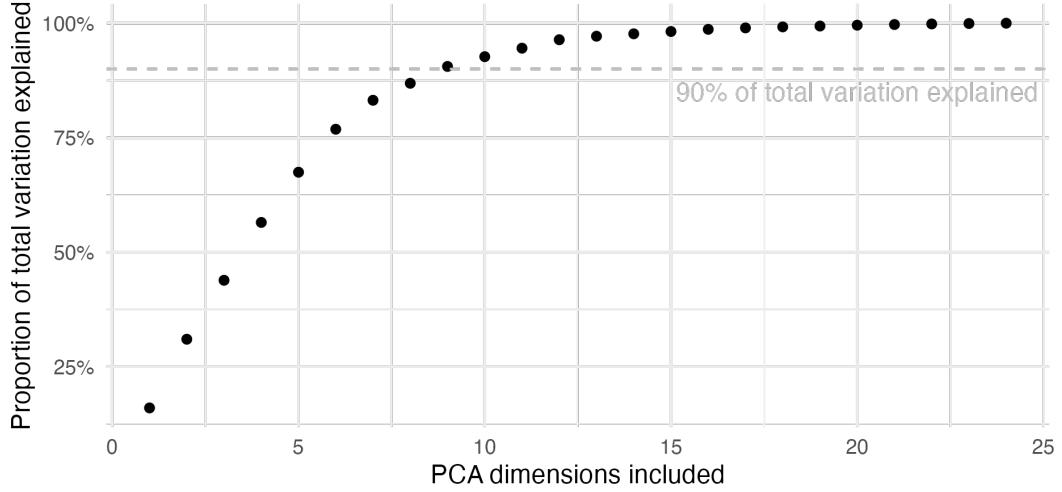
The TMB C++ user-template used to specify the log-posterior, described in Appendix C.4.4, was the same for each approach. The dimension of the latent field was  $N = 467$  and the dimension of the hyperparameters was  $m = 24$ . For GEB and GPCA-AGHQ, hyperparameter and latent field samples were simulated 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 from GEB are illustrated in Figure 6.18.

### 6.5.1 NUTS convergence and suitability

The Naomi model was difficult to efficiently sample from using NUTS via `tmbstan`. Four chains run in parallel for 100 thousand iterations each were required to obtain acceptable NUTS diagnostics. For ease-of-storage, the samples were thinned by a factor of 20, resulting in 5000 iterations kept per chain, with the first 2500 removed as burn-in. The effective sample size ratios were typically low (Figure C.6). The lowest effective sample size was 208 (2.5% quantile 318, 50% quantile 1231, and 97.5% quantile 2776; Panel C.7A). The largest potential scale reduction factor was 1.021 (2.5% quantile 1, 50% quantile 1.003, and 97.5% quantile 1.017; Panel C.7B). Though inaccuracies remain possible, these diagnostics are sufficient to treat inferences obtained from NUTS as a gold-standard.

Correlation structure in the posterior can result in sampler inefficiency. Each of the four pairs of AR1 log standard deviation  $\log(\sigma)$  and logit lag-one autocorrelation parameter  $\text{logit}(\phi)$  posteriors were positively correlated (mean absolute correlation 0.81, Figure C.8). These parameters are partially identifiable as variation can either be explained by high standard deviation and high autocorrelation or low standard deviation and low autocorrelation. On the other hand, the BYM2 log standard deviation  $\log(\sigma)$  and logit proportion parameter  $\text{logit}(\phi)$  were, as designed, more orthogonal (mean absolute correlation 0.17, Figure C.9).

The informativeness of data about a parameter can be summarised by the posterior contraction (Schad et al. 2021) which compares the prior variance  $\mathbb{V}_{\text{prior}}(\phi)$



**Figure 6.19:** Under PCA, the proportion of total variation explained is given by the sum of the first  $s$  eigenvalues over the sum of all eigenvalues. A typical rule-of-thumb is to include dimensions sufficient to explain 90% of total variation. In this case, for computational reasons, 87% was considered sufficient.

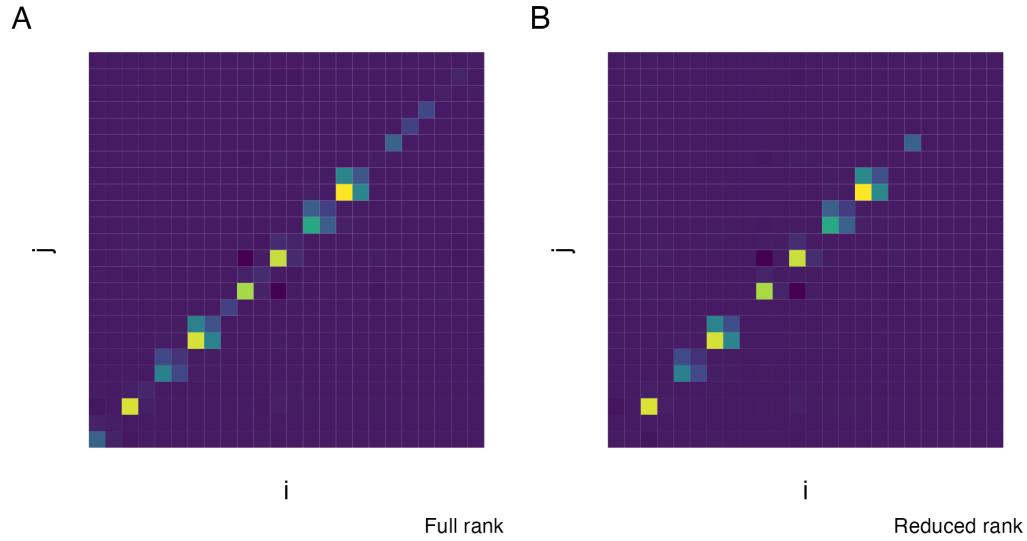
to posterior variance  $\mathbb{V}_{\text{post}}(\phi)$  via

$$c(\phi) = 1 - \frac{\mathbb{V}_{\text{prior}}}{\mathbb{V}_{\text{post}}(\phi)}. \quad (6.98)$$

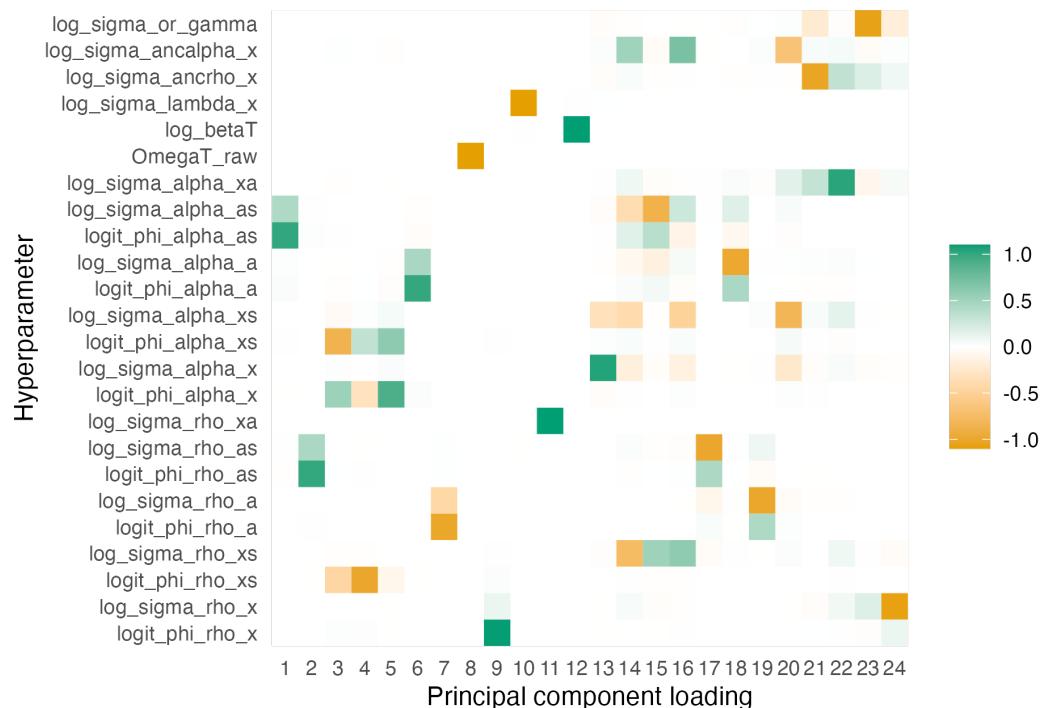
Posterior variances were extracted from NUTS results, and prior variances obtained by simulating from a model with the likelihood component removed (Figure C.10). The average posterior contraction was positive for all latent field parameter vectors, and for the majority of hyperparameters (Figure C.11). However, for seven hyperparameters the posterior contraction was very close to zero. Furthermore, for some latent field parameter vectors, the average contraction was small. Based on this finding, these parameters may not be identifiable.

### 6.5.2 Use of PCA-AGHQ

For the PCA-AGHQ quadrature grid, a Scree plot based on the spectral decomposition of  $\hat{\mathbf{H}}_{\text{LA}}(\theta_{\text{LA}})^{-1}$  was used to select the number of principal components to keep (Figure 6.19). Keeping  $s = 8$  principal components was sufficient to explain 87% of total variation. The reduced rank approximation to the inverse curvature with this choice of  $s$  was visually similar to the full rank matrix (Figure 6.20).



**Figure 6.20:** The full rank original covariance matrix (Panel A) was closely reproduced by its reduced rank ( $s = 8$ ) matrix approximation (Panel B).



**Figure 6.21:** Each principal component loading, obtained by the eigendecomposition of the inverse curvature, gives the direction of maximum variation conditional on inclusion of each previous principal component loading. For example, the first principal component loading is a sum of `log_sigma_alpha_as` and `logit_phi_alpha_as`.

The principal component (PC) loadings (Figure ??) provide interpretable information about which directions had the greatest variation. Many of the first PC loadings are sums of two hyperparameters. As such, there is some redundancy in the hyperparameter parameterisation, supporting the findings of Section 6.5.1 regarding correlation structure in the hyperparameter posterior. It is exactly this correlation structure that PCA, and PCA-AGHQ, looks to utilise.

Projecting the  $3^8 = 6561$  PCA-AGHQ quadrature nodes onto each hyperparameter dimension, there was substantial variation in coverage by hyperparameter (Figure 6.22). Approximately 12 hyperparameters had well covered marginals: greater than the 8 naively obtained with a dense grid, but nonetheless far fewer than the full 24. Coverage was higher among hyperparameters on the logistic scale, and lower among hyperparameters on the logarithmic scale. This discrepancy occurred due to logistic hyperparameters naturally having higher posterior marginal standard deviation than logarithmic hyperparameters (Figure C.13).

### 6.5.3 Time taken

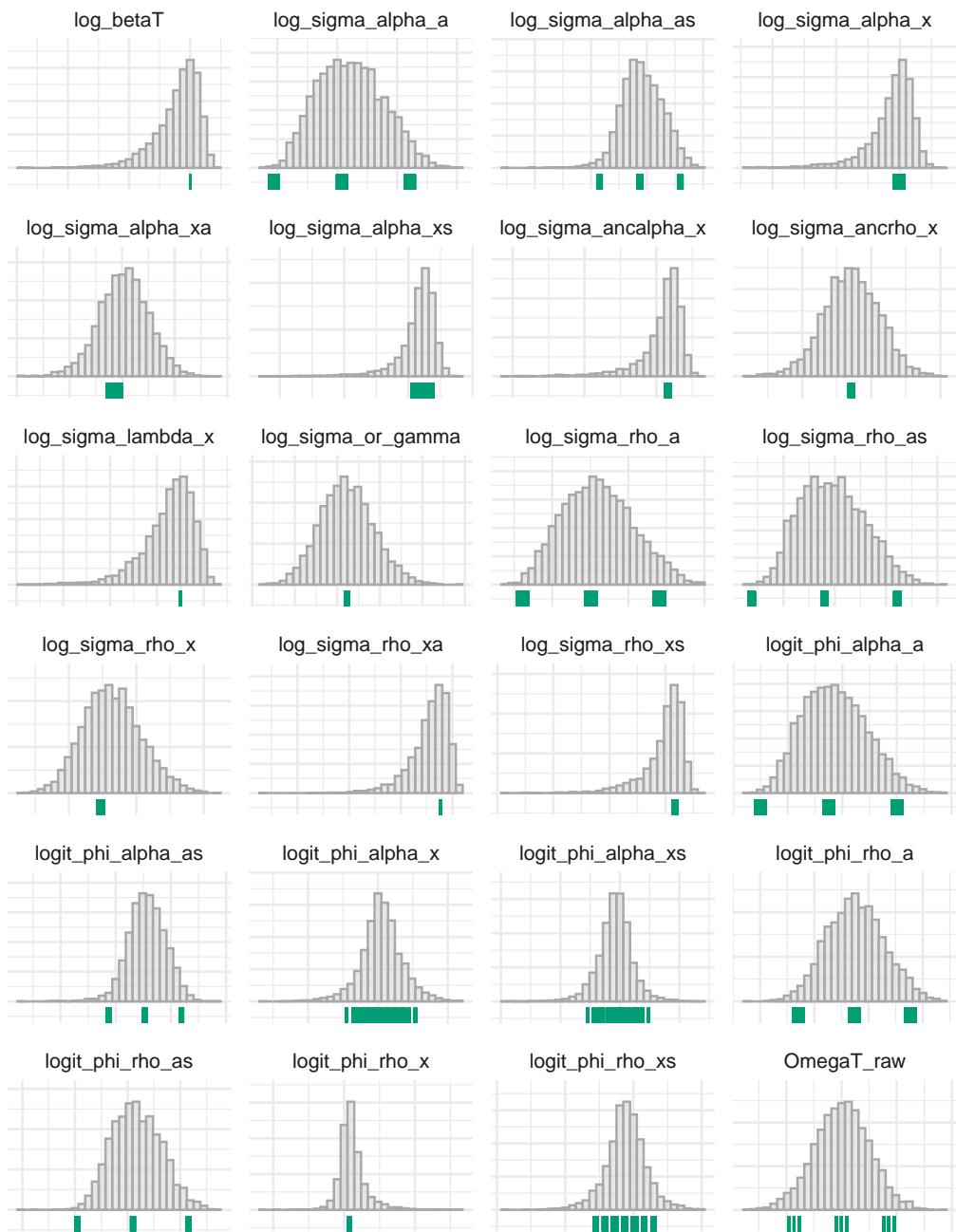
Inference with NUTS took 79 hours, while inference with GPCA-AGHQ took 1.2 hours and GEB just 0.9 minutes (Figure 6.23). Both the NUTS and GPCA-AGHQ algorithms can be run under a range of settings, trading off accuracy and runtime.

### 6.5.4 Inference comparison

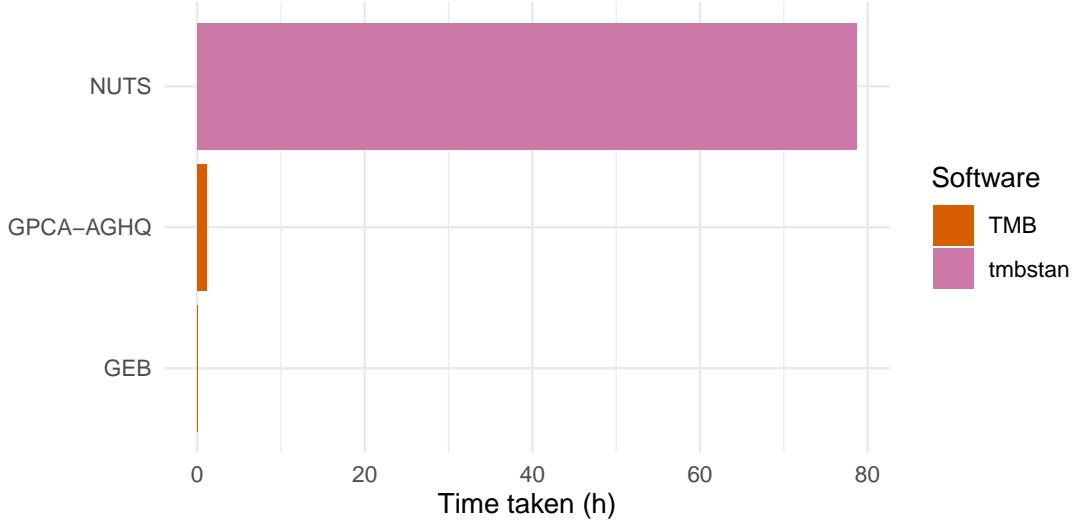
Posterior inferences from GEB, GPCA-AGHQ and NUTS were compared using point estimates (Section 6.5.4.1) distributional quantities (Section 6.5.4.2).

#### 6.5.4.1 Point estimates

Latent field point estimates obtained from GPCA-AGHQ were closer to the gold-standard results from NUTS than those obtained from GEB (Figure 6.24). The root mean square error (RMSE) between posterior mean estimates from GPCA-AGHQ and NUTS (0.063) was 20% lower than that between GEB and NUTS (0.078). For the posterior standard deviation estimates, there was a substantial 60%



**Figure 6.22:** The 6561 PCA-AGHQ nodes projected onto the 24 hyperparameter marginal distributions obtained with NUTS.



**Figure 6.23:** The number of hours taken to perform inference for the Naomi ELGM (Section 6.3.1) using each approach.

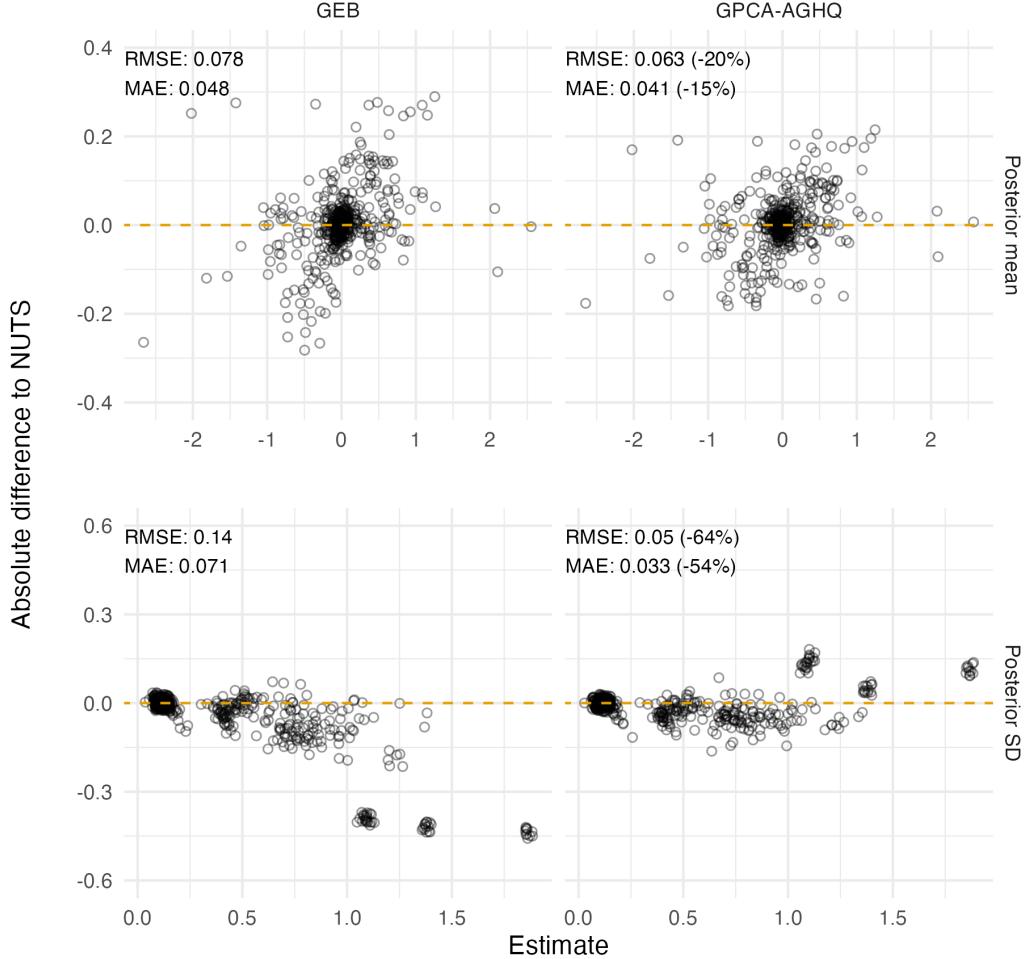
reduction in RMSE: from 0.14 (TMB) to 0.05 (PCA-AGHQ). However, puzzlingly, improvements in latent field estimate accuracy only transferred to model outputs to a limited extent (Figures C.15 and C.16).

#### 6.5.4.2 Distributional quantities

**Kolmogorov-Smirnov** The two-sample Kolmogorov-Smirnov (KS) test statistic (Smirnov 1948) is the maximum absolute difference between two ECDFs  $F(\omega) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\phi_i \leq \omega}$ . It is a relatively stringent, worst case, measure of distance between empirical distributions. The average KS test statistic for GPCA-AGHQ (0.077) was 8.6% less than the average KS test statistic for GEB (0.084).

For both GEB and GPCA-AGHQ the KS test statistic for a parameter was correlated with low NUTS ESS (Figure C.17). This may be due to by difficulties estimating particular parameters for all inference methods, or high KS values caused by NUTS inaccuracies.

**Maximum mean discrepancy** Let  $\Phi^1 = \{\phi_i^1\}_{i=1}^n$  and  $\Phi^2 = \{\phi_i^2\}_{i=1}^n$  be two sets of joint posterior samples, and  $k$  be a kernel. The maximum mean discrepancy [MMD; Gretton et al. (2006)] is a measure of distance between joint distributions,

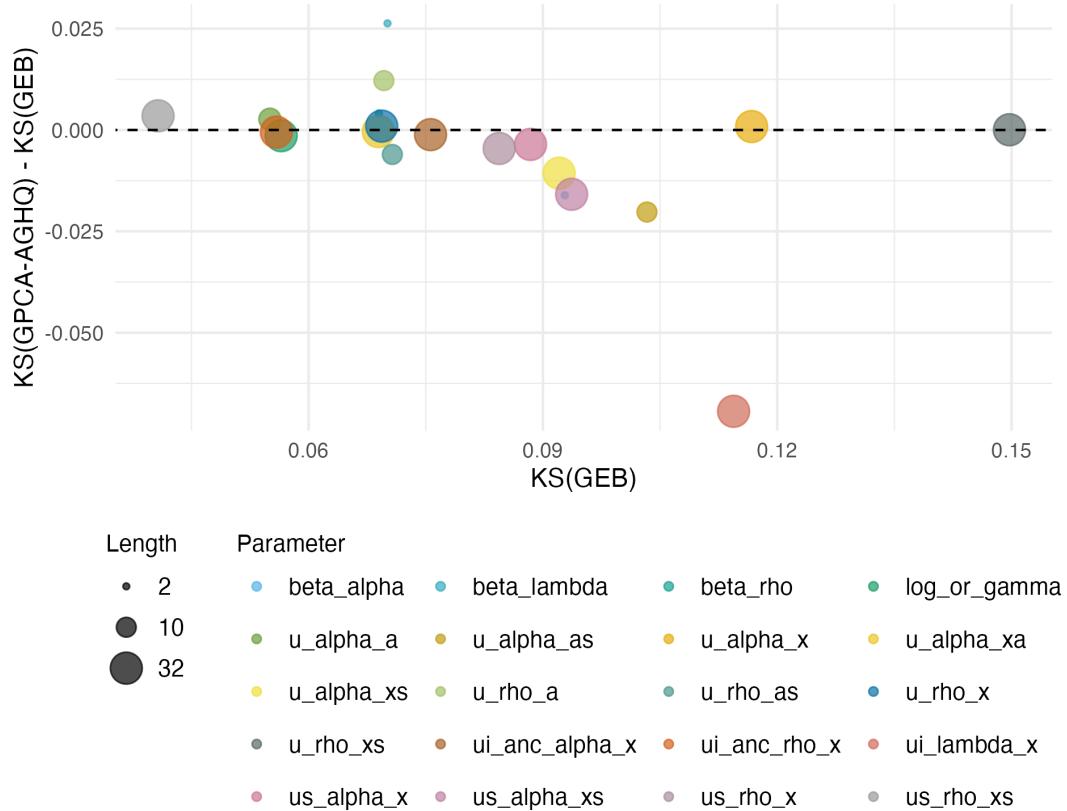


**Figure 6.24:** The latent field posterior mean and posterior standard deviation point estimates from each inference method as compared with those from NUTS. The root mean square error (RMSE) and mean absolute error (MAE) are displayed in the top left. For both the posterior mean and posterior standard deviation, GPCA-AGHQ reduced RMSE and MAE as compared with GEB.

and can be estimated empirically by samples

$$\text{MMD}(\Phi^1, \Phi^2) = \sqrt{\frac{1}{n^2} \sum_{i,j=1}^n k(\phi_i^1, \phi_j^1) - \frac{2}{n^2} \sum_{i,j=1}^n k(\phi_i^1, \phi_j^2) + \frac{1}{n^2} \sum_{i,j=1}^n k(\phi_i^2, \phi_j^2)}. \quad (6.99)$$

The kernel was set to  $k(\phi^1, \phi^2) = \exp(-\sigma \|\phi^1 - \phi^2\|^2)$  with  $\sigma$  estimated from data using the **kernlab** R package (Karatzoglou et al. 2019). The first and third order MMD statistics for GEB were 0.08 and 0.0048. Those of GPCA-AGHQ (0.078 and 0.0044) were just 3% and 7% lower.



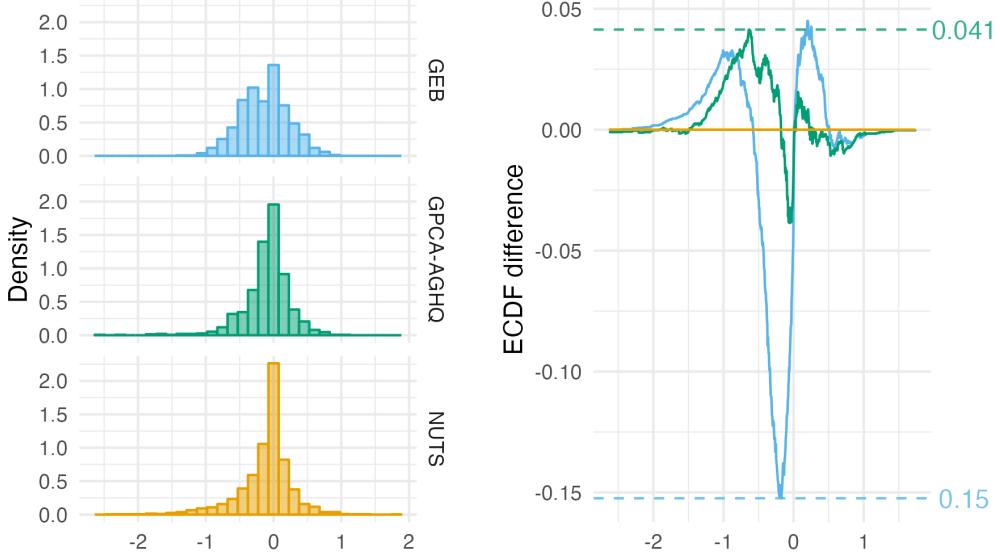
**Figure 6.25:** The average Kolmogorov-Smirnov (KS) test statistic for each latent field parameter of the Naomi model. Vectors of parameters were grouped together. For points above the dashed line at zero, performance of GEB was better. For points below the dashed line, performance of GPCA-AGHQ was better. Most notably, for the latent field parameters  $ui\_lambda\_x$  the test statistic for GEB was substantially higher than for GPCA-AGHQ. This parameter, of length 32, corresponds to  $\mathbf{u}_x^\lambda$  and plays a key role in the ART attendance component of the Naomi (Section 6.3.1.4).

### 6.5.5 Exceedance probabilities

As a more realistic use case for the Naomi model outputs, consider the two following case-studies based on exceedance probabilities.

#### 6.5.5.1 Meeting the second 90

Ambitious targets for scaling up ART treatment have been developed by UNAIDS, with the goal of ending the AIDS epidemic by 2030 (UNAIDS 2014). Meeting the 90-90-90 fast-track target requires that 90% of people living with HIV know their status, 90% of those are on ART, and 90% of those have a suppressed viral load. Inferences from Naomi can be used to identify treatment



**Figure 6.26:** The parameter `ui_lambda_x[26]` had the greatest difference in KS test statistics between GEB and GPCA-AGHQ to NUTS. For this parameter, the potential scale reduction factor was 1 and effective sample size was 2100.

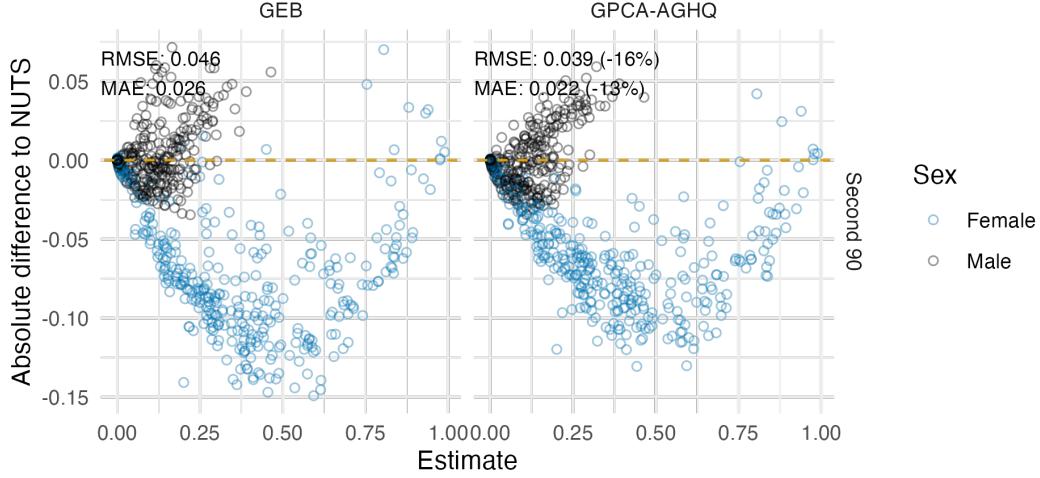
gaps by calculating the probability that the second 90 target has been met, that is  $\mathbb{P}(\alpha_i > 0.9^2 = 0.81)$  for each strata  $i$ .

Strata probabilities of having met the second 90 target were more accurately estimated by GPCA-AGHQ than GEB (Figure 6.27). Both GPCA-AGHQ and GEB had substantial error as compared to results from NUTS, however, particularly for girls and women. This discrepancy in accuracy by sex may be caused by interactions between the household survey and ANC components of the model creating a more challenging posterior geometry.

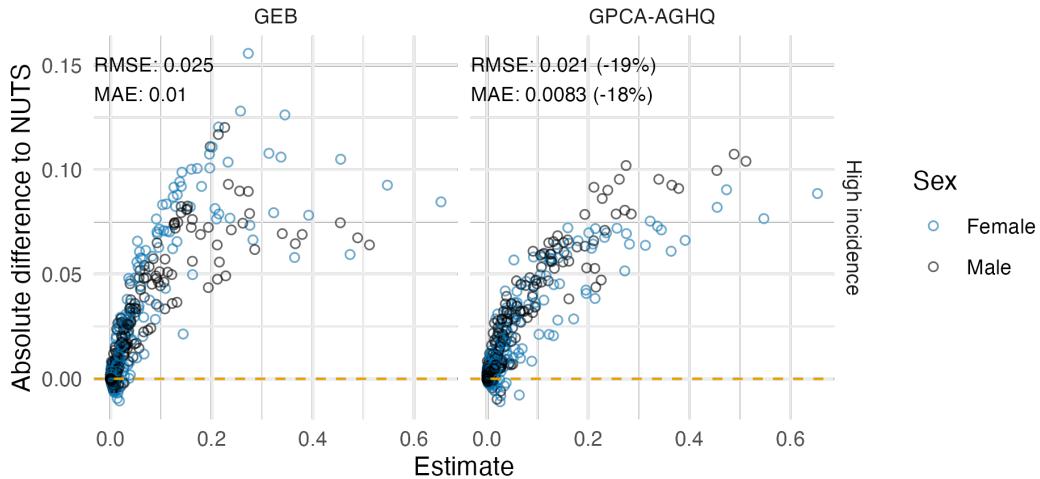
#### 6.5.5.2 Finding strata with high incidence

Some HIV interventions are cost-effective only within high HIV incidence settings, typically defined as higher than 1% incidence per year. Inferences from Naomi can be used to calculate the probability of a strata having high incidence by evaluating  $\mathbb{P}(\lambda_i > 0.01)$ .

GPCA-AGHQ gave more accurate estimates of the probability that a strata has



**Figure 6.27:** The probability each strata has met the second 90 (ART coverage above 81%) calculated using each inference method, as compared with NUTS. The root mean square error (RMSE) and mean absolute error (MAE) are displayed in the top left.



**Figure 6.28:** The probability each strata has high HIV incidence (above 1% per year) calculated using each inference method, as compared with NUTS. The root mean square error (RMSE) and mean absolute error (MAE) are displayed in the top left.

high HIV incidence than GEB (Figure 6.28). Again, both methods had significant error. Unlike in Section 6.5.5.1, there was little difference in error by sex.

## 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). Both were applied

to the Naomi model in Malawi in Section 6.5. These contributions are discussed in turn, before outlining suggestions for future work.

### 6.6.1 A universal INLA implementation

Monnahan and Kristensen (2018) write that “to our knowledge, TMB is the only software platform capable of toggling between integration tools [the Laplace approximation and NUTS] so effortlessly”. Section 6.2 made important progress towards adding INLA to the integration tools easily accessible using TMB. Reaching this milestone would be of significant value to both applied and methodological researchers.

The implementation is not intended to replace R-INLA, and indeed for the majority of users a formula-based interface is preferred. Both formula-based and universal statistical tools have value, as they inhabit different use-cases. For the NUTS algorithm, a universal interface is available via Stan, and packages such as `brms` (Bürkner 2017) and `rstanarm` (Goodrich et al. 2020) enable researchers to fit common models using a formula interface. Furthermore, developers of formula-based tools do have incentives to engage with the needs of their users, and indeed do so. For example, after requesting for the generalised binomial distribution used in Equation (6.81) to be included in R-INLA, a prototype version was shortly made available. That said, it is ultimately more sustainable for advanced users to have capacity to implement their own distributions and models.

### 6.6.2 PCA-AGHQ with application to INLA for Naomi

For the simplified Naomi model applied to data from Malawi, GPCA-AGHQ more accurately inferred latent field posterior marginal distributions than GEB. However, model output posterior marginals did not see the same improvements. Approximate posterior exceedance probabilities from both GEB and GPCA-AGHQ had systematic inaccuracies as compared with NUTS. GEB and GPCA-AGHQ were substantially faster than NUTS, which took over two days to reach convergence.

Inaccuracies in model outputs from GEB and GPCA-AGHQ do have potential to meaningfully mislead policy (Sections 6.5.5.1 and 6.5.5.2). As such, where

possible, gold-standard NUTS results should be computed. Though NUTS is too slow to run during a workshop, it could be run afterwards. As the UNAIDS HIV estimates process occurs annually, requiring days to compute more accurate estimates is viable. 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 be impossible to run NUTS to convergence, and approximate methods may be required.

To empower users, GPCA-AGHQ and NUTS could be added to the Naomi web interface (<https://naomi.unaids.org>) as alternatives to GEB. 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.

#### **6.6.3.1 Further comparisons**

Comparison to further Bayesian inference methods could be included in Section 6.5. Four possibilities 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. Rather than use quadrature to integrate the marginal Laplace approximation, an alternative approach is to run HMC (Monnahan and Kristensen 2018; C. 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. Preliminary testing of this approach, using `tmbstan` and setting `laplace = TRUE`, did not show immediate success but likely could be worked on.
3. NUTS is not especially well suited to sampling from Gaussian latent field models like Naomi. Other MCMC algorithms, such as blocked Gibbs sampling (S. Geman and D. Geman 1984) or slice sampling (Neal 2003), could be considered. It may be difficult to implement such algorithms using TMB. Many MCMC algorithms are implemented and customisable (including, for example, the choice of block structure) within the NIMBLE probabilistic programming language (de Valpine et al. 2017). Requiring rewriting the Naomi model log-posterior outside of TMB would be a substantial downside.
4. Finally, it would be of substantial interest to implement the Naomi model using the iterative INLA method via `inlabru`. However, as `inlabru`, like R-INLA, is based on a formula interface, it may not be possible to do so directly.

#### 6.6.3.2 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 indices, 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.
4. Finally, it may be possible to achieve better performance by pruning and prerotation, as discussed by Jäckel (2005).

#### 6.6.3.3 Computational improvements

1. Approximation: The most significant improvement likely could come by using approximations to the Laplace marginals. In particular, he simplified Laplace marginals of Wood (2020) (Section 6.1.3.4) should be implemented, as the ELGM setting has relatively dense precision matrices.
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.

#### **6.6.3.4 Statistical theory**

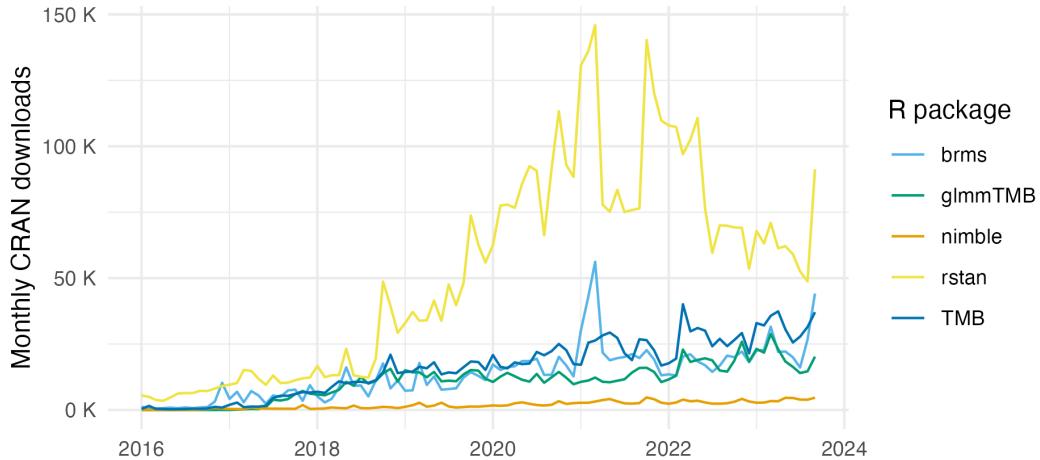
The class of functions which are integrated exactly by PCA-AGHQ remains to be shown. Theorem 1 of Stringer et al. (2022) bounds the total variation error of AGHQ, establishing convergence in probability of coverage probabilities under the approximate posterior distribution to those under the true posterior 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.

#### **6.6.3.5 Testing quadrature assumptions**

It may be possible to test the assumptions made by use of AGHQ grids, allowing their suitability for a particular integral to be assessed. Specifically, AGHQ assumes that the integrand is closely approximated by a polynomial multiplied by a Gaussian density. Given NUTS hyperparameter samples (or better yet, hyperparameter samples from the Laplace NUTS hybrid discussed in Section 6.6.3.1) this assumption could be tested by fitting a model using a polynomial times Gaussian kernel. This approach could be generalised to also test the suitability of PCA-AGHQ grids.

#### **6.6.3.6 Exploration of the accuracy of INLA for complex models**

The universal INLA implementation can be used to measure the accuracy of INLA for a wider range of models than were previously possible. An important benefit of using TMB is that comparisons to NUTS can easily be made using exactly the same model template. Among the ELGM-type structures of particular interest for spatial epidemiology are aggregated likelihood models and evidence synthesis models.



**Figure 6.29:** Monthly R package downloads from the Comprehensive R Archive Network (CRAN) for `brms`, `glmmTMB`, `nimble`, `rstan` and `TMB`, obtained using the `cranlogs` (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. As such, this metric may underestimate the popularity of `rstan`.

#### 6.6.3.7 Methods dissemination

The approach used to implement Laplace marginals with `TMB` was relatively ad-hoc, and involved modification of the `TMB` C++ template (Section 6.2.1.4). For wider dissemination of this method, it is important that the user is not burdened with making these modifications. One possibility would be to change the `random` argument in `TMB::MakeADFun` to allow for indexing. Another (less desirable) option would be to algorithmically 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 6.29). 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 C. 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

This chapter concludes the thesis by discussing its most important contributions, some promising avenues for future work, and broader reflections about the work.

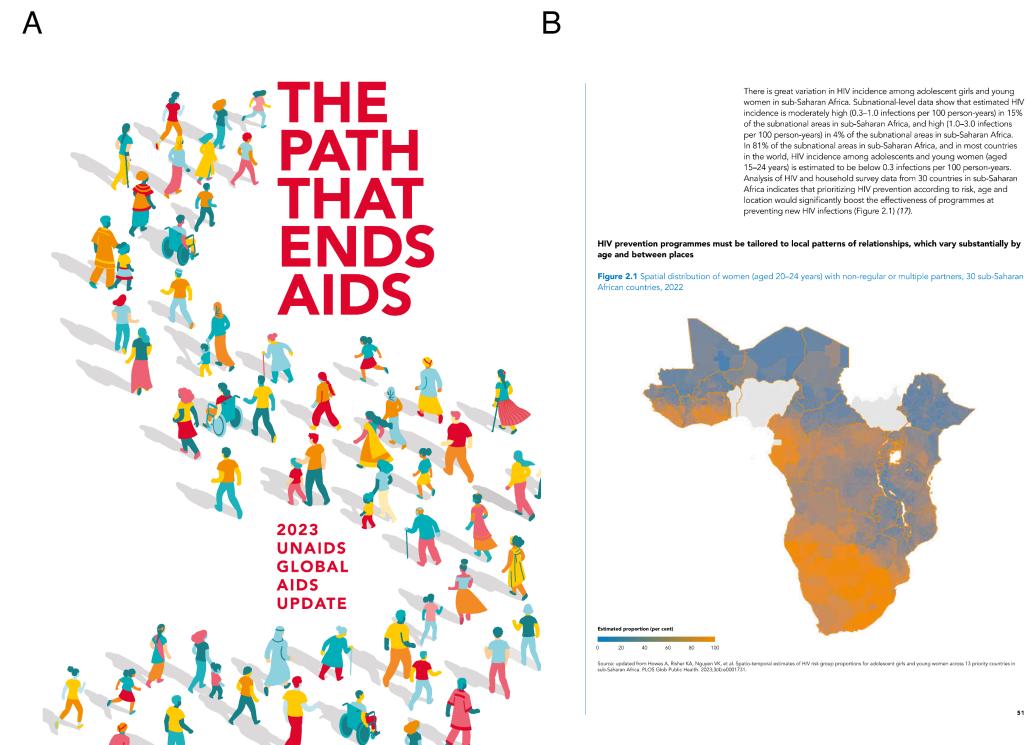
### 7.1 Contributions

Effective response to the HIV epidemic depends on strategic information provided by models of data. The contributions of this thesis are both in generating this information, and in understanding and advancing statistical methods.

Chapter 4 found that spatially structured random effects should be used in small-area models for HIV. Kernel models performed better for data simulated from an adjacency-based spatial process than adjacency-based models did for data simulated from a kernel model. However, adjacency-based models performed better under cross-validation of real HIV survey data. Model comparison was conducted using strictly proper scoring rules, with checks for calibration.

Chapter 5 estimated HIV risk group proportions for AGYW to enable implementation of the Global AIDS strategy (UNAIDS 2021b). Risk group proportion estimates were used to behaviourally disaggregate HIV prevalence and incidence, and assess the benefits of a variety of risk stratification strategies. This work is the basis for a tool used to prioritise delivery of HIV prevention services by

## Conclusions



**Figure 7.1:** Panel A shows the front page of UNAIDS (2023b). Panel B shows the page containing text and a figure based on the work done in Chapter 5. In this figure, 30 countries are included.

countries in SSA. The tool now encompasses at least 30 countries, expanding from the initial 13 included [Figure 7.1; UNAIDS (2023b)]. Models will be rerun each year to populate the tool with updated information as a part of the UNAIDS annual HIV estimates process. Alongside these applied contributions, Chapter 5 exemplified specification of complex multinomial spatio-temporal models in **R-INLA** using the Poisson-multinomial transformation, including using two- and three-way Kronecker product interactions.

The Naomi model has been used in over 35 countries in SSA to produce district-level estimates of HIV indicators by synthesising evidence from multiple sources. Chapter 6 developed deterministic Bayesian inference methods, motivated by the aim of providing more accurate inferences for this challenging and practically important model. Its most important methodological contributions are two-fold. First, an implementation of INLA which is compatible with models specified using a **TMB C++** template. For the first time, practitioners can now fit essentially any model using

## *Conclusions*

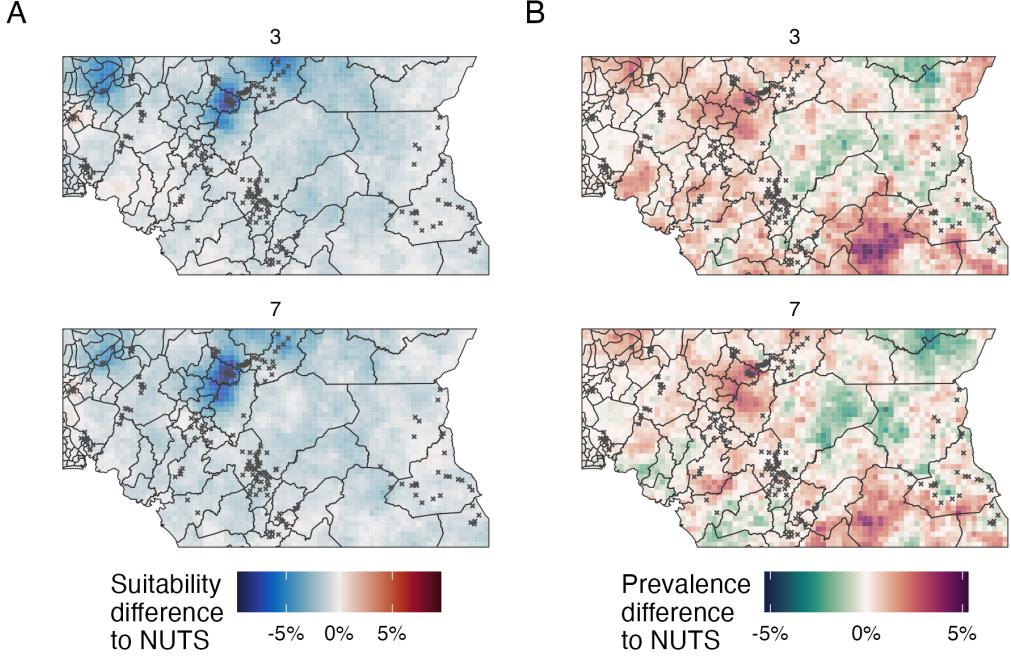
the INLA method. Second, a quadrature rule which combines PCA and AGHQ to naturally extend the applicability of INLA methods to moderate hyperparameter dimension, allowing more complex models to be fit. Additionally, Chapter 6 provides detailed description and analysis of the Naomi model. Indeed, Esra et al. (2023) used tables and text from Appendix C in an update to Jeffrey W Eaton et al. (2021).

## 7.2 Future work

Promising avenues for future work, that I might prioritise, include:

1. It would be valuable to extend the risk group model developed in Chapter 5, and the resulting tool, to include all adults 15-49. Although AGYW are disproportionately at risk of HIV infection, 56% of new infections in SSA occur in other demographic groups. Modelling of age-stratified sexual partnerships (Wolock et al. 2021) may help to overcome reporting biases by harmonising male and female reporting. This model would likely fall out of the scope of **R-INLA**, but would be possible to write with **TMB** and therefore amenable to the inference methods advanced in Chapter 6.
2. Although suitable for early stage research, wider adoption of the INLA implementation developed in Chapter 6 would be enhanced greatly by improvements to its speed and usability. The most important speed enhancement would come from using the simplified approximation to the Laplace marginals developed by Wood (2020). Although the naive implementation used in this thesis is viable for integrating Laplace marginals over a small number of hyperparameter quadrature nodes, such as the  $3^2 = 9$  nodes used Sections 6.2.2 and 6.2.1, it becomes prohibitively slow for larger numbers. Usability would be improved by providing the method as a part of statistical software, likely via the **aghq** package. The primary difficulty which would have to be overcome to do so is that the **random** argument of **TMB**:**MakeADFun** does not allow indexing.

## Conclusions



**Figure 7.2:** For the Loa loa ELGM (Section 6.2.2), increasing the number of quadrature nodes per hyperparameter dimension from  $k = 3$  to  $k = 7$  did little to improve accuracy. On the other hand, using Laplace marginals rather than Gaussian marginals did have a substantial effect (Figures 6.12 and 6.13). It would be valuable to better understand, and aspirationally have diagnostics for, the circumstances under which accuracy of INLA methods could be improved by additional computation.

3. The universal INLA implementation developed in Chapter 6 enables empirical and methodological research that was previously not possible, or prohibitively difficult.

INLA-like methods can now be tested for a broader class of models, such as the Loa loa and Naomi ELGMs (Sections 6.2.2 and 6.5). That a single TMB C++ template for the log-posterior supports inference using multiple methods, including gold-standard NUTS via `tmbstan`, is a substantial asset in conducting this type of research.

As an example research question, within this class of models, what is the best way to obtain accurate inferences within a fixed computational budget. Is it better to use additional hyperparameter grid points, or more accurate latent field approximations? For the Loa loa ELGM in Section 6.2.2, the benefit of Laplace marginals was greater than a denser AGHQ grid (Figure 7.2). It

## *Conclusions*

would also be of interest to find methods to obtain accurate inferences for particular parameters, or functions of parameters, using INLA-like methods. For example, in Section 6.5, although the PCA-AGHQ grid improved latent field parameter inferences, it did little to improve model output accuracy. Is there a way in which computational could be focused on obtaining accurate estimates of Naomi model outputs?

Additionally, it is relatively easy to make alterations to the implementation, facilitating possible innovation in the design of INLA-like algorithms. Previously, it has been difficult for researchers not involved in development of **R-INLA** to engage in methodological work about the INLA method.

Theoretical research could be conducted to complement the work described above, extending the findings of Bilodeau et al. (2022). This work is benefited by the complete specification (Appendix C.3) of the INLA-like algorithm used in this thesis.

## **7.3 Broader reflections**

Conducting the work in this thesis involved testing the boundaries of available statistical software. For example, I found it challenging, if not impossible, to implement a common model using different inferential software. As the frequently asked questions section of the **R-INLA** website (Havard Rue 2023) notes: “the devil is in the details”. Similarly, I encountered issues implementing a desired collection of different models in a common inferential software. From personal experience, my colleagues have also encountered similar problems. Needless to say, conflation of statistical models and inference methodologies limits the validity of any findings. To avoid this issue, I implemented all models in Chapters 4 and 6 using **TMB** model templates. (Additionally, I would recommend implementing the model used in Chapter 5 in **TMB** for future development.) Alongside being sufficiently flexible to meet my model specification requirements, **TMB** is compatible with a range of inference methodologies, including those advanced in this thesis.

## *Conclusions*

As such, TMB remains (Osgood-Zimmerman and Jon Wakefield 2023) an under-rated statistical tool. In demonstrating some of its capabilities, I hope this thesis contributes to its wider adoption.

The work done in this thesis, particularly Chapters 4 and 6, focused on producing experimental, empirical evidence. This approach reflects the complexity of the models and methods used in this thesis. Understanding complex systems from a theoretical perspective can be challenging. That said, in my opinion the work in this thesis could benefit from closer integration with statistical theory. Though comprehensive theoretical understanding of models or algorithms may be too optimistic, better understanding simplified examples, limiting cases, or constituent parts may provide value.

Working with the data in Chapter 5 deepened my appreciation for the realistic challenges faced in applied work, and data quality being linchpin for any successful statistical analysis. While from the real world, the data in Chapters 4 and 6 underwent substantial cleaning, processing, and vetting before I handled them, as is typical in methodological research. It is important that methodological and theoretical statisticians appreciate the real challenges of applied work, by doing it themselves, or working in close collaboration with those who do.

There are both direct and indirect paths to impact for the work in this thesis. Directly, the methodological contributions of Chapters 4 and 6 may eventually lead to marginally more accurate indicator estimates, contributing to a broadly more effective response. However, these improvements in accuracy seem of minor consequence within the broader context of the HIV response, and factors limiting its effectiveness. The applied contributions of Chapter 5 have a more promising case for direct impact. Indeed, I have seen evidence of engagement with this work by decision makers.

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 having done so facilitates my work to be scrutinised, and more optimistically, built upon. In part this hope has already been realised, as with limited input from me, Dr. Kathryn Risher

### *Conclusions*

was able to extend my code for Chapter 5 to include additional countries (Panel 7.1B). This would not have been possible without tools from the R ecosystem such as **rmarkdown** and **rticles** for reporting, **devtools** for R package development, as well as those written by software engineers within the MRC Centre for Global Infectious Disease Analysis such as **orderly** and **didehpc**. It is crucial that academia adjusts to appropriately incentivises software contributions, and encourages adaption of open science best practices. Work done to inform public health decision making should be held to high standards of transparency, reproducibility and collaboration. Especially so in an outbreak response scenario (Grieve et al. 2023), where time is limited and decisions may be of significant consequence.

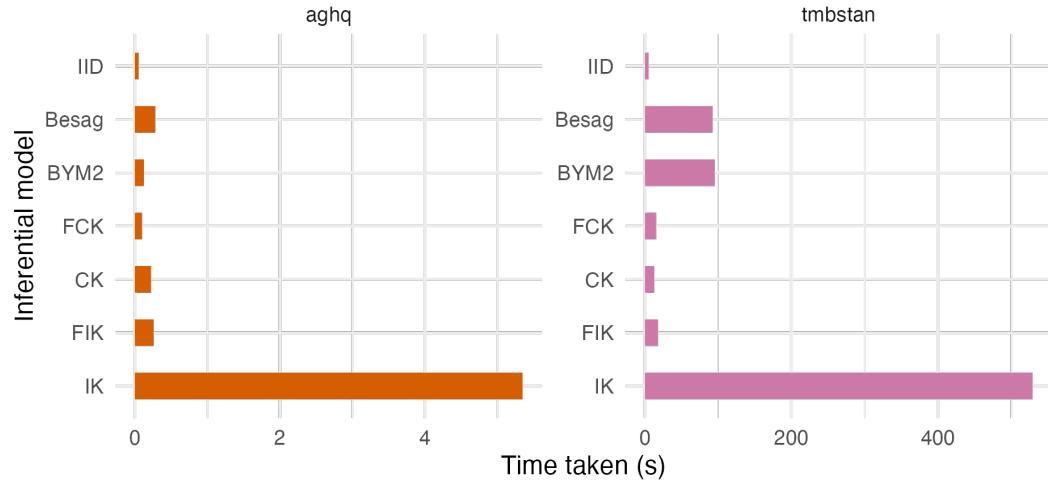
# Appendices

# A

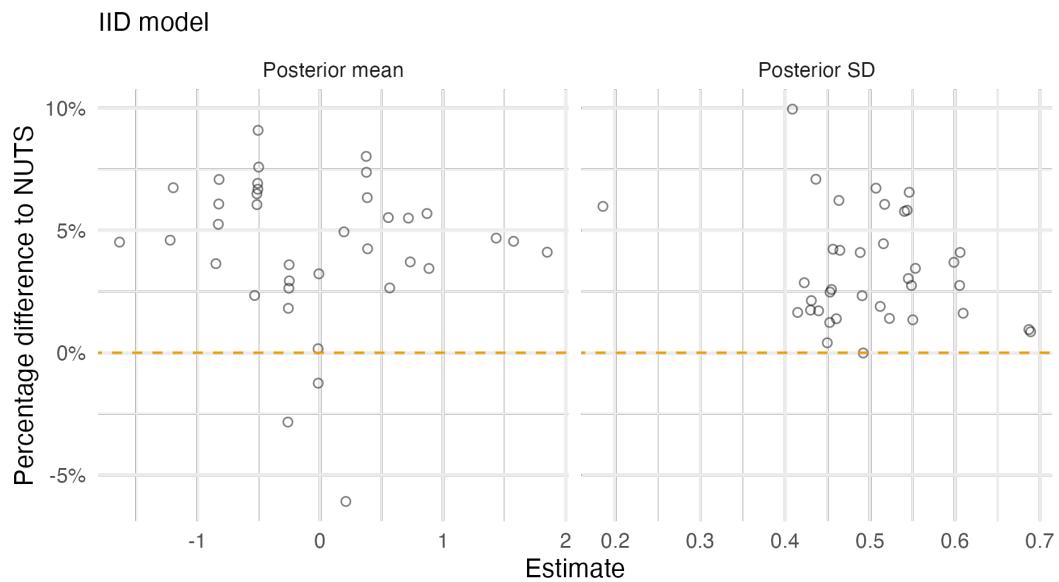
Models for areal spatial structure

## A.1 Comparison of AGHQ to NUTS

### A. Models for areal spatial structure

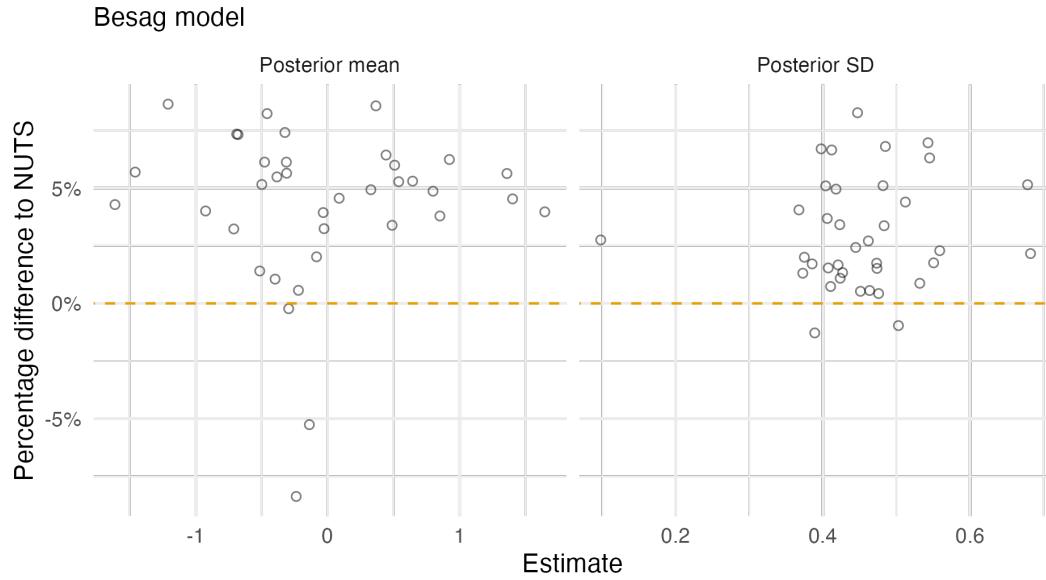


**Figure A.1:** A comparison of time taken to fit AGHQ via `aghq` as compared with NUTS via `tmbstan` for each inferential model. For the models run using NUTS via `tmbstan` there was significant variation in time taken depending on initial random seed. As such, these timings and more broadly the inferences obtained from NUTS in Appendix A.1 should be interpreted with appropriate skepticism.

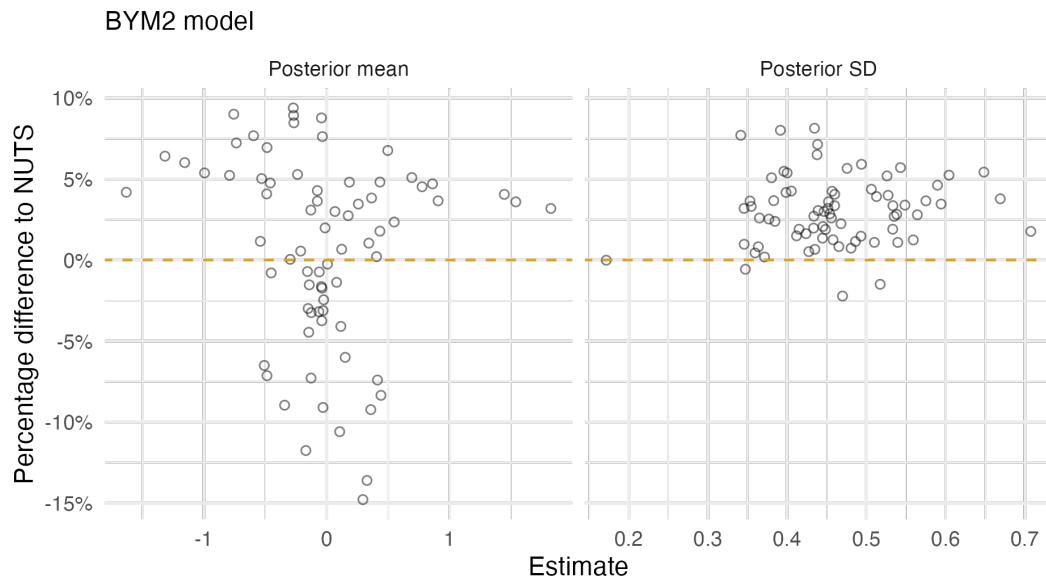


**Figure A.2:** A comparison of the posterior means and standard deviations obtained with AGHQ via `aghq` as compared with NUTS via `tmbstan` fitting an IID inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 1686, and the maximum value of the potential scale reduction factor was 1.00.

*A. Models for areal spatial structure*

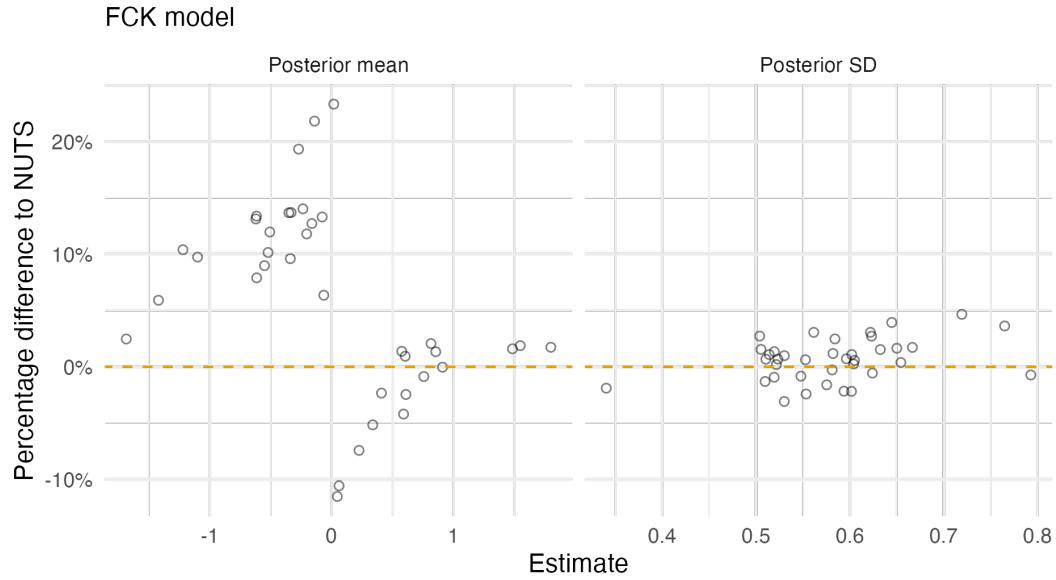


**Figure A.3:** A comparison of the posterior means and standard deviations obtained with AGHQ via `aghq` as compared with NUTS via `tmbstan` fitting a Besag inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 1056, and the maximum value of the potential scale reduction factor was 1.00.

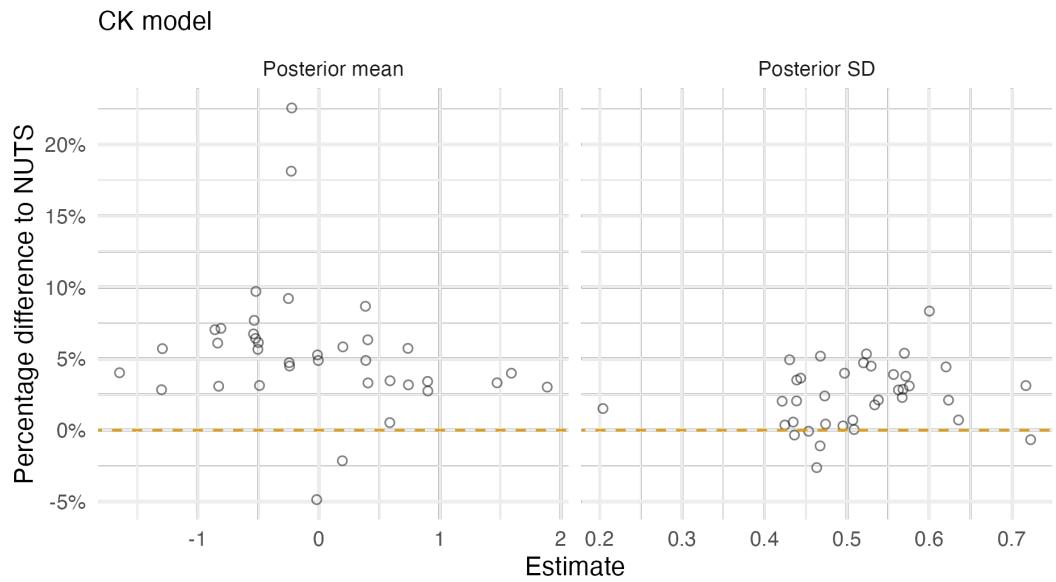


**Figure A.4:** A comparison of the posterior means and standard deviations obtained with AGHQ via `aghq` as compared with NUTS via `tmbstan` fitting a BYM2 inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 35, and the maximum value of the potential scale reduction factor was 1.06.

### A. Models for areal spatial structure

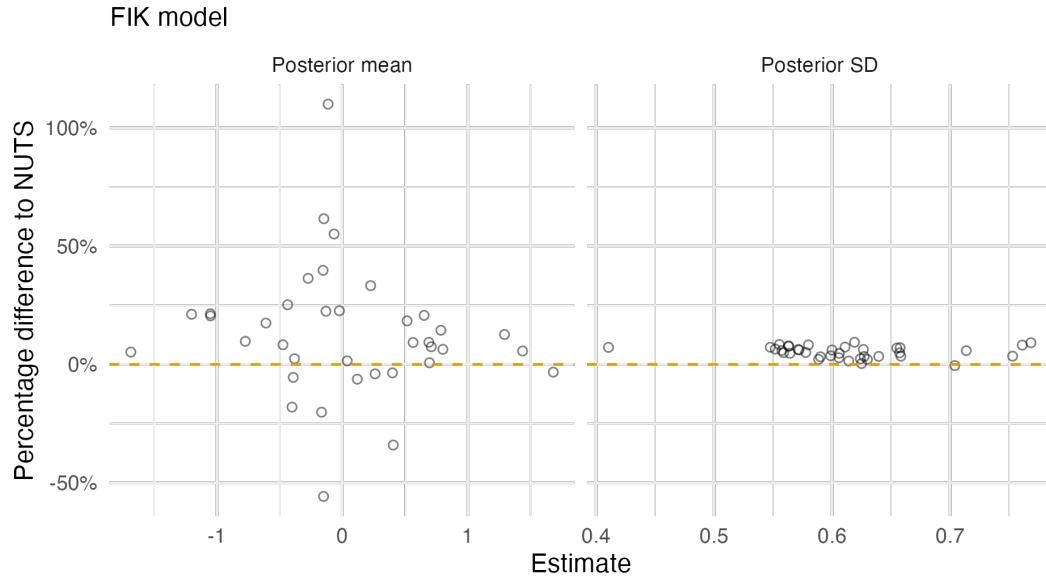


**Figure A.5:** A comparison of the posterior means and standard deviations obtained with AGHQ via `aghq` as compared with NUTS via `tmbstan` fitting a FCK inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 355, and the maximum value of the potential scale reduction factor was 1.01.

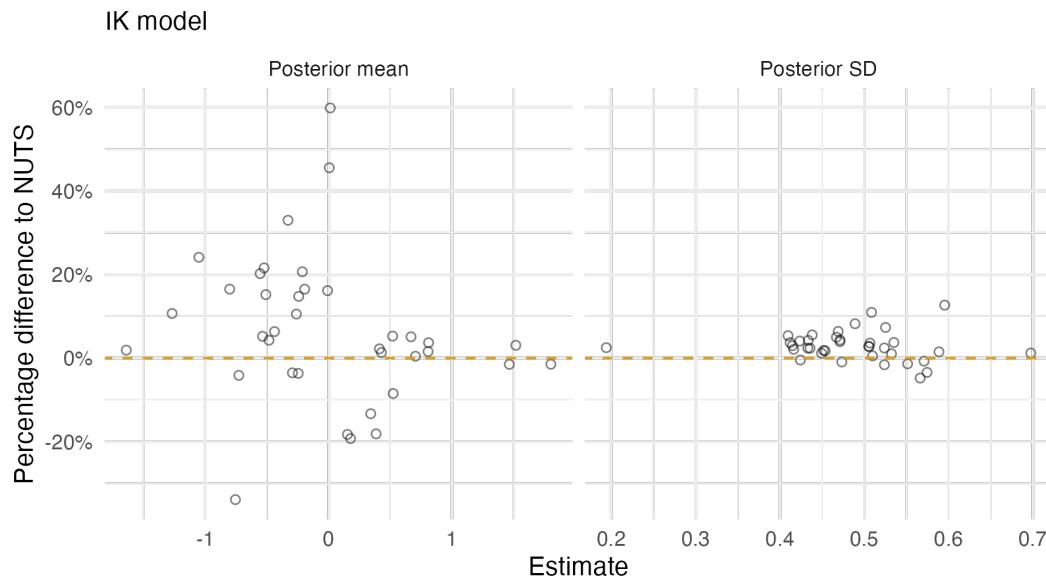


**Figure A.6:** A comparison of the posterior means and standard deviations obtained with AGHQ via `aghq` as compared with NUTS via `tmbstan` fitting a CK inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 1471, and the maximum value of the potential scale reduction factor was 1.00.

### A. Models for areal spatial structure

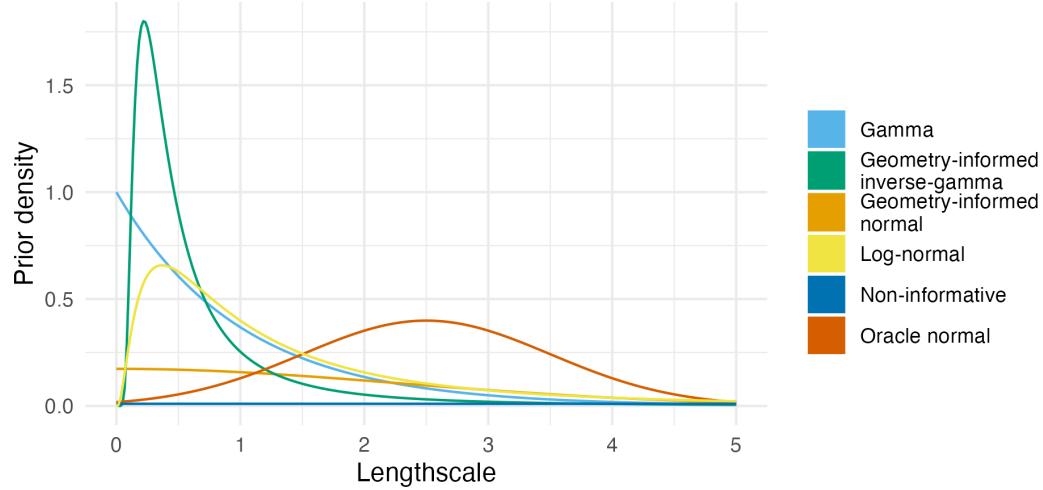


**Figure A.7:** A comparison of the posterior means and standard deviations obtained with AGHQ via `aghq` as compared with NUTS via `tmbstan` fitting a FIK inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 289, and the maximum value of the potential scale reduction factor was 1.01.



**Figure A.8:** A comparison of the posterior means and standard deviations obtained with AGHQ via `aghq` as compared with NUTS via `tmbstan` fitting a IK inferential model to IID synthetic data on the grid geometry (Panel 4.6E). For NUTS, the minimum ESS was 1623, and the maximum value of the potential scale reduction factor was 1.00.

### A. Models for areal spatial structure



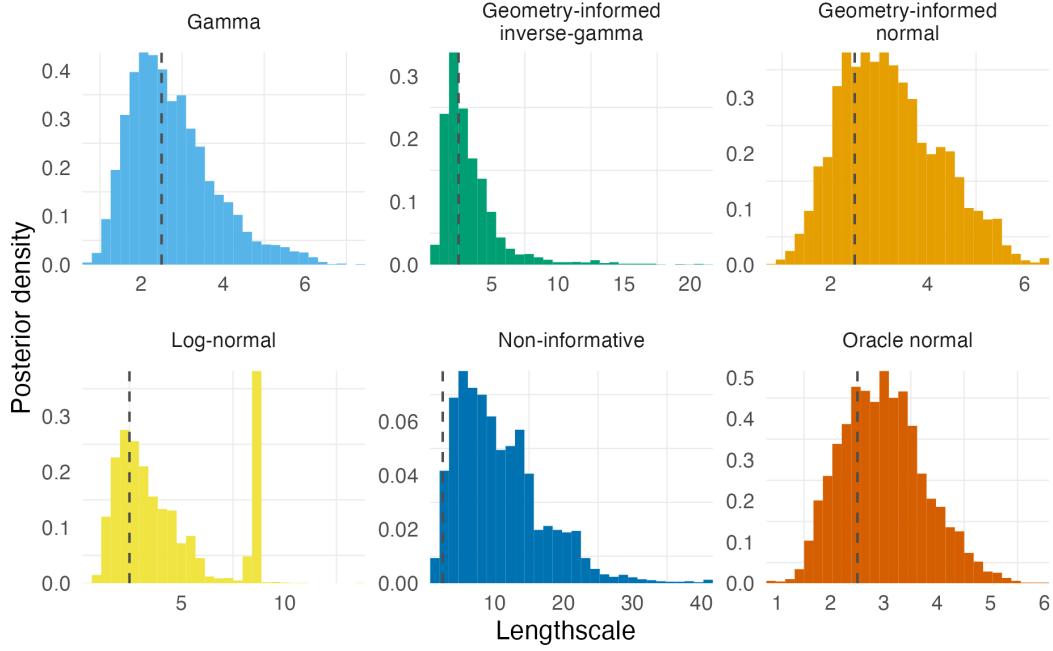
**Figure A.9:** The probability density for each lengthscale prior distribution as given in Table A.1.

## A.2 Lengthscale prior sensitivity

**Table A.1:** Six lengthscale prior distributions were considered for use in the simulation (Section 4.3) and HIV prevalence (Section 4.4) studies.

Description	Prior	Additional details
Gamma	$l \sim \text{Gamma}(1, 1)$	—
Geometry-informed inverse-gamma	$l \sim \text{IG}(a, b)$	The parameters $a$ and $b$ chosen such that 5% of the prior mass was below and above the 5% and 95% quantile for distance between points (Betancourt 2017)
Geometry-informed normal	$l \sim \mathcal{N}^+(0, \sigma)$	The parameter $\sigma$ set as one third the difference between the minimum and maximum distance between points (Betancourt 2017)
Log-normal	$l \sim \text{Log-normal}(0, 1)$	—
Non-informative	$p(l) = 1$	This is an improper prior in that it does not integrate to one
Oracle normal	$l \sim \mathcal{N}^+(2.5, 1)$	The mean of this prior was set to the true value of the lengthscale

### A. Models for areal spatial structure



**Figure A.10:** Lengthscale posterior distributions obtained using NUTS to fit a centroid kernel model to integrated kernel data. The true value, 2.5, is shown as a dashed vertical line. Six different lengthscale prior distributions were considered as given in Table A.1. The geometry used was the grid (Panel 4.6E).

## A.3 Simulation study

### A.3.1 Lengthscale recovery

### A.3.2 BYM2 proportion

### A.3.3 Mean squared error

**Table A.2:** The average mean squared error (MSE) of each inferential model in estimating  $\rho$ , under different simulation and geometry settings. Entries for FCK and CK on geometry 2 are empty because model was undefined in that case. The units used in this table are thousandths.

Simulation model	Inferential model						
	IID	Besag	BYM2	FCK	CK	FIK	IK
1							
IID	8.20	7.56	7.99	7.84	7.67	7.90	7.61
Besag	7.31	6.39	7.15	7.31	6.76	7.27	6.63
IK	7.44	6.30	7.27	7.74	6.83	7.58	6.62

### A. Models for areal spatial structure

2							
IID	8.43	7.62	8.23	-	-	7.99	8.32
Besag	7.56	6.58	7.39	-	-	7.25	6.42
IK	7.16	5.91	6.95	-	-	6.91	4.95
3							
IID	8.23	7.72	8.19	8.09	7.85	8.05	7.75
Besag	7.73	6.71	7.63	7.78	7.01	7.55	6.67
IK	7.56	6.24	7.30	7.75	6.78	7.53	6.18
4							
IID	8.71	8.03	8.49	8.53	8.31	8.35	8.12
Besag	7.48	6.65	7.34	7.55	7.08	7.44	6.89
IK	7.38	6.11	7.12	7.60	6.71	7.45	6.36
Grid							
IID	7.63	7.65	7.66	7.72	7.79	7.89	7.84
Besag	4.06	3.29	3.77	3.94	3.36	3.71	3.32
IK	5.97	4.30	4.81	4.98	3.50	4.47	3.41
Cote d'Ivoire							
IID	7.72	7.78	7.74	7.89	7.99	8.08	7.96
Besag	4.88	3.96	4.45	4.62	4.07	4.36	4.00
IK	5.61	3.96	4.50	4.73	3.18	4.19	3.10
Texas							
IID	7.63	7.71	7.65	8.59	8.05	8.60	7.80
Besag	5.13	4.05	4.62	4.60	4.36	4.34	4.26
IK	6.29	4.51	5.06	4.44	3.45	4.04	3.37

### A.3.4 Continuous ranked probability score

**Table A.3:** The average continuous ranked probability score (CRPS) of each inferential model in estimating  $\rho$ , under different simulation and geometry settings. Entries for FCK and CK on geometry 2 are empty because model was undefined in that case. The units used in this table are thousandths.

Simulation model	Inferential model						
	IID	Besag	BYM2	FCK	CK	FIK	IK
1							
IID	32.6	33.9	32.7	32.1	33.4	32.3	33.5
Besag	30.7	29.5	30.6	30.7	30.0	30.7	29.9
IK	31.2	29.1	31.1	32.1	30.1	31.7	29.7

### A. Models for areal spatial structure

2							
IID	33.1	33.4	32.8	-	-	32.7	39.9
Besag	32.0	30.6	31.6	-	-	31.2	33.2
IK	28.9	26.2	28.6	-	-	28.4	24.2
3							
IID	32.9	33.8	33.1	32.4	33.5	32.6	35.0
Besag	32.9	31.1	32.4	33.0	31.5	32.2	31.6
IK	30.7	28.1	30.3	31.4	29.0	30.8	27.9
4							
IID	34.3	34.9	34.2	34.2	34.8	33.8	34.7
Besag	32.3	31.2	31.9	32.1	31.8	31.9	31.7
IK	29.8	27.3	29.3	30.5	28.3	29.9	27.7
Grid							
IID	32.4	34.2	32.5	33.1	34.0	35.1	35.1
Besag	24.6	22.7	23.3	23.4	23.8	23.5	24.1
IK	28.7	23.7	24.6	24.4	21.1	23.1	21.0
Cote d'Ivoire							
IID	32.4	34.5	32.5	33.7	34.8	35.8	35.6
Besag	26.5	24.4	24.9	25.3	25.9	25.3	26.0
IK	27.7	22.2	23.4	23.6	19.6	22.2	19.6
Texas							
IID	32.1	34.0	32.3	39.2	35.7	40.0	35.6
Besag	27.3	24.7	25.3	27.1	27.5	26.9	27.0
IK	29.7	24.5	25.4	23.0	20.8	22.3	20.9

### A.3.5 Calibration

## A.4 HIV study

### A.4.1 Lengthscale

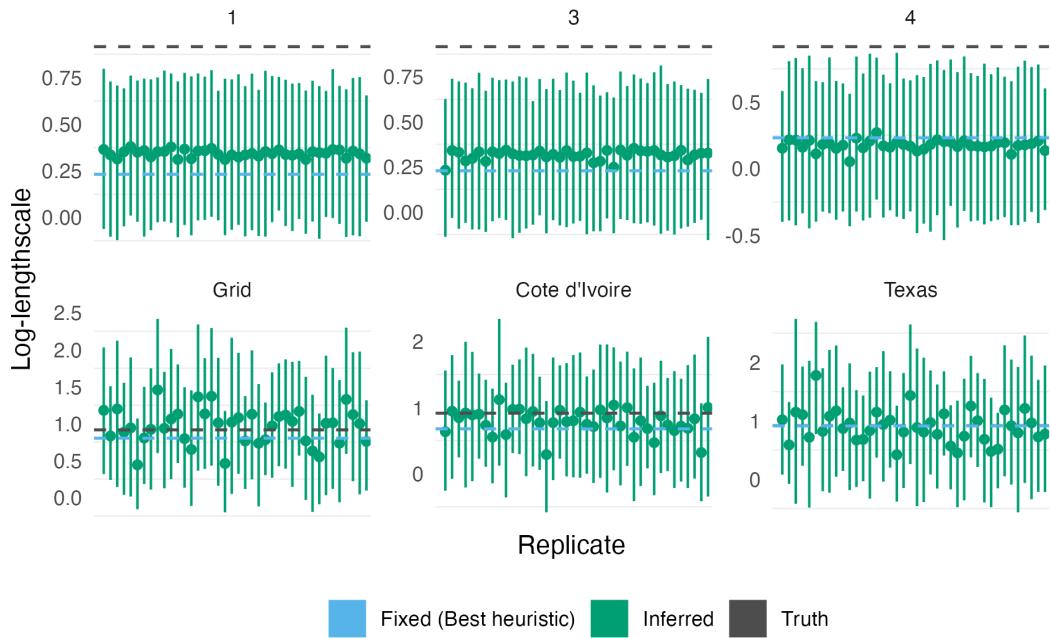
### A.4.2 BYM2 proportion

### A.4.3 Estimates

### A.4.4 Cross-validation

#### A.4.4.1 Mean squared error

### A. Models for areal spatial structure



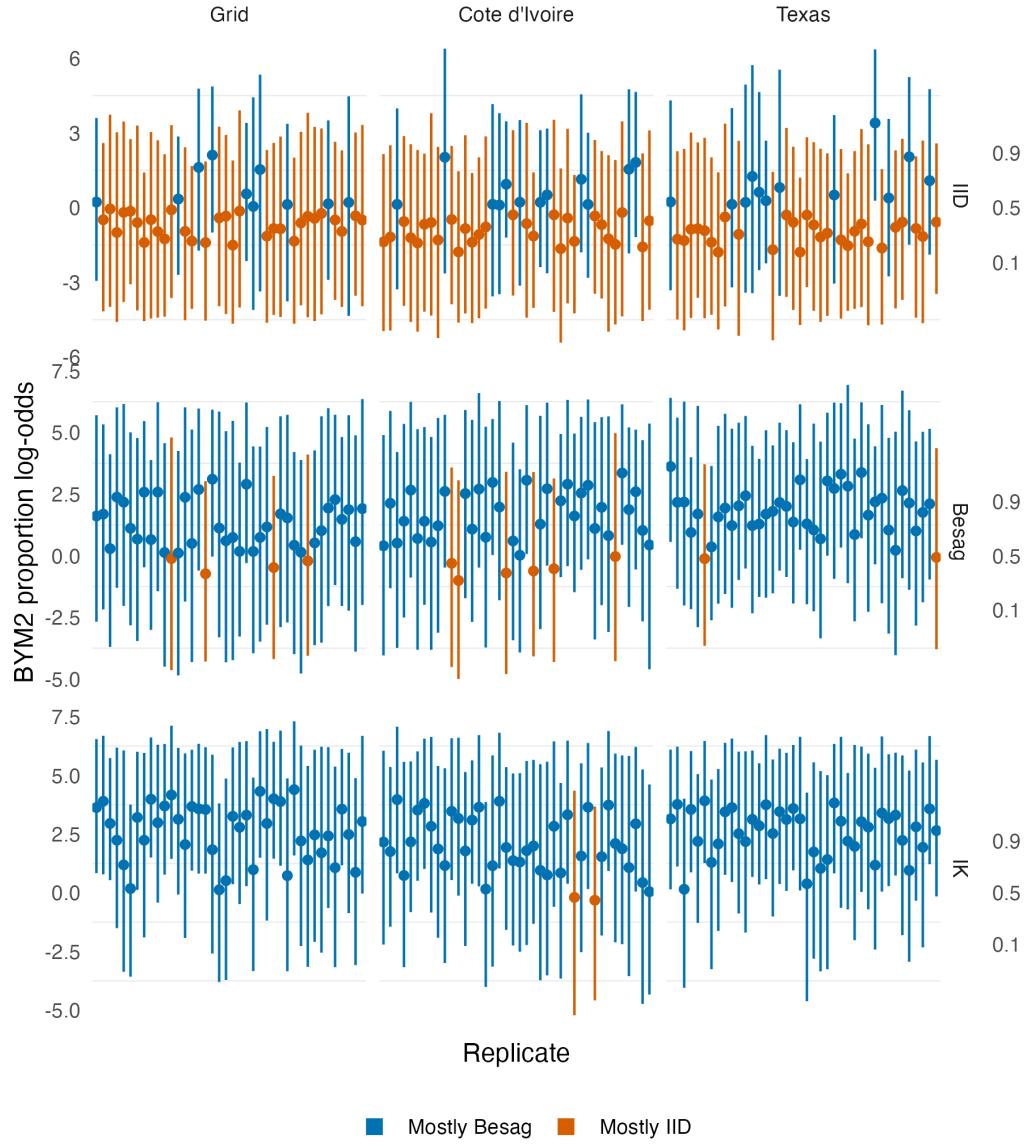
**Figure A.11:** The lengthscale posterior mean and 95% credible interval obtained using the centroid kernel model on integrated kernel data for the first 40 simulation replicates on each geometry. The true lengthscale, and lengthscale obtained using the heuristic method of Best et al. (1999), are shown as dashed horizontal lines.

**Table A.4:** The mean pointwise leave-one-out and spatial leave-one-out MSE in estimating  $\rho_i$ , with standard errors, for each inferential model across the four considered PHIA surveys. The units used in this table are thousandths.

PHIA survey	Mean squared error (units: 1/1000)						
	IID	Besag	BYM2	FCK	CK	FIK	IK
<b>LOO</b>							
Côte d'Ivoire, 2017	0.21	0.22	0.20	0.21	0.19	0.21	0.20
Malawi, 2016	7.10	2.39	2.59	3.59	3.70	2.43	2.54
Tanzania, 2017	1.66	1.14	1.43	0.95	0.65	0.78	0.66
Zimbabwe, 2016	4.76	2.51	2.54	2.51	1.88	2.15	1.83
<b>SLOO</b>							
Côte d'Ivoire, 2017	0.20	0.22	0.21	0.24	0.25	0.26	0.25
Malawi, 2016	7.13	2.41	3.32	8.22	7.95	7.05	6.70
Tanzania, 2017	1.65	1.09	2.46	1.86	2.80	1.86	2.59
Zimbabwe, 2016	4.73	2.49	3.44	3.95	3.36	3.93	3.42

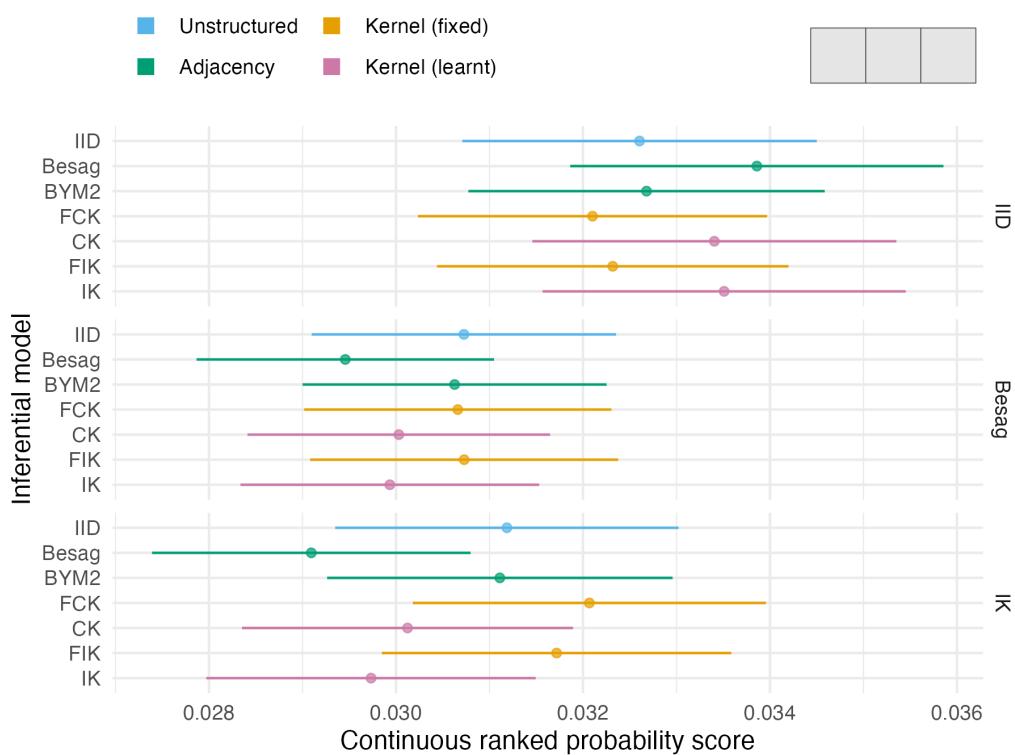
#### A.4.4.2 Continuous ranked probability score

*A. Models for areal spatial structure*



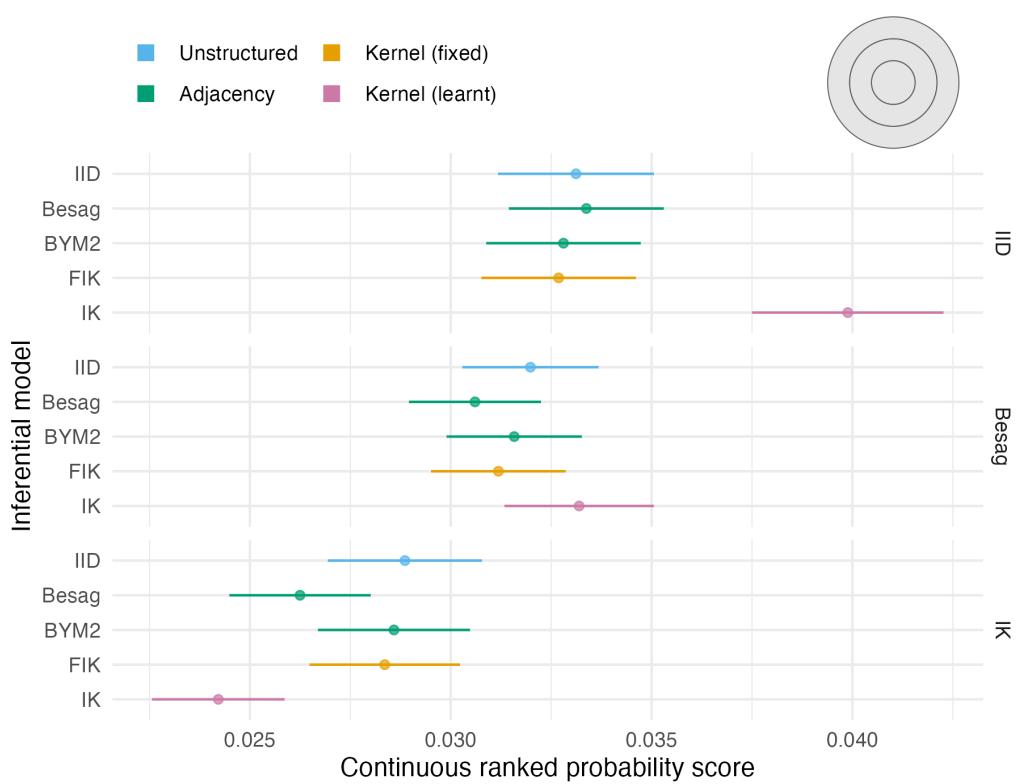
**Figure A.12:** The BYM2 proportion parameter posterior mean and 95% credible interval obtained for the first 40 simulation replicates for the realistic geometries. When the simulated data is IID, the BYM2 proportion parameter is in the majority of cases below 0.5, corresponding to have inferred that the noise is mostly IID (spatially unstructured). When the simulated data is either Besag or IK, the BYM2 proportion parameter is in the majority of cases above 0.5, corresponding to have inferred that the noise is mostly Besag (spatially structured).

### A. Models for areal spatial structure



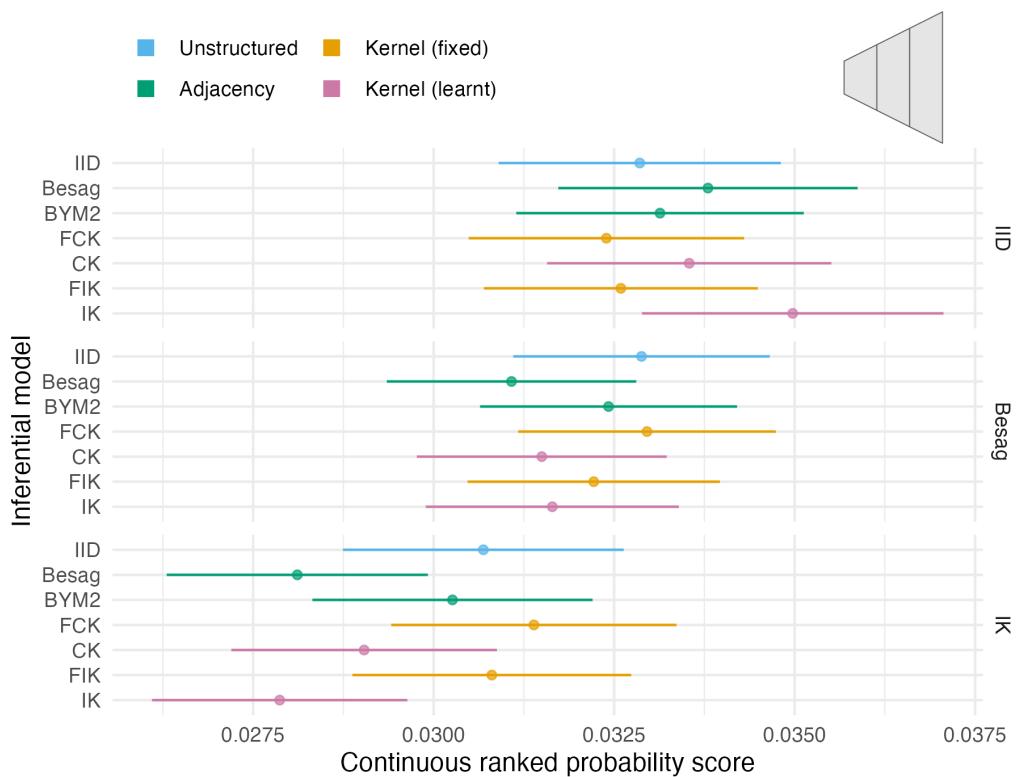
**Figure A.13:** The mean CRPS with 95% credible interval in estimating  $\rho$  using each inferential model and simulation model on the first vignette geometry (Panel 4.6A). Credible intervals were generated using 1.96 times the standard error.

*A. Models for areal spatial structure*



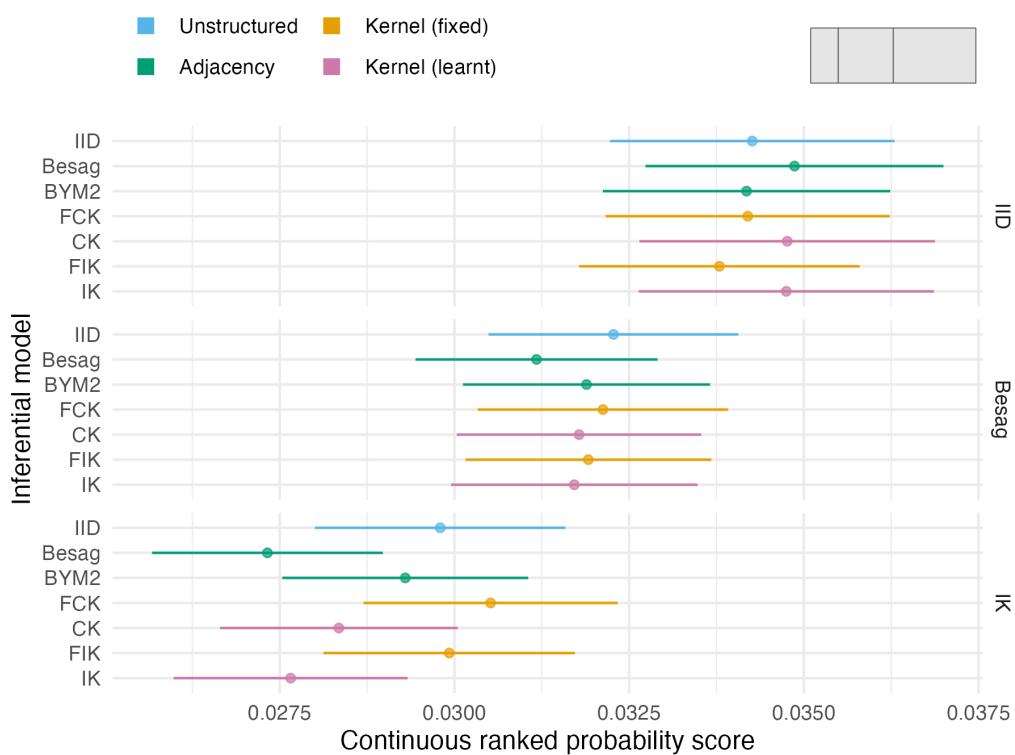
**Figure A.14:** The mean CRPS with 95% credible interval in estimating  $\rho$  using each inferential model and simulation model on the second vignette geometry (Panel 4.6B). Credible intervals were generated using 1.96 times the standard error.

### A. Models for areal spatial structure



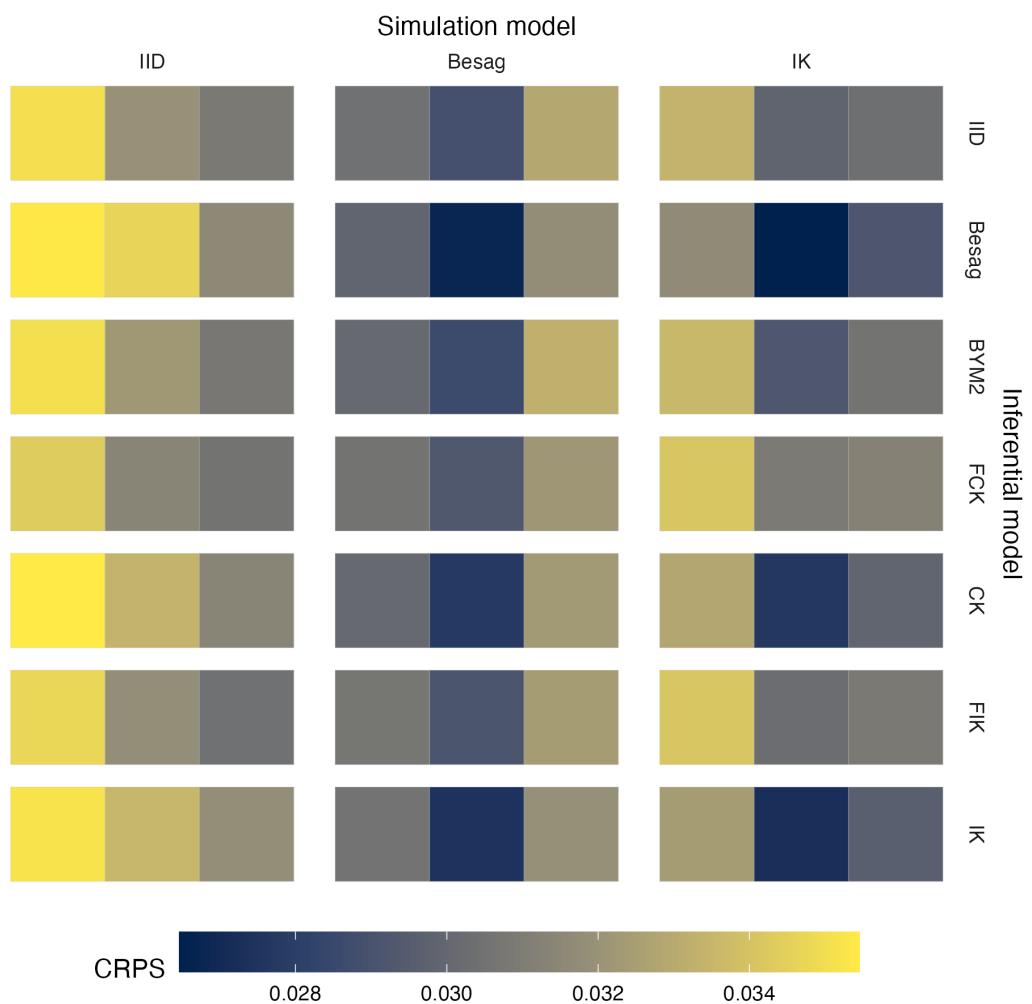
**Figure A.15:** The mean CRPS with 95% credible interval in estimating  $\rho$  using each inferential model and simulation model on third vignette geometry (Panel 4.6C). Credible intervals were generated using 1.96 times the standard error.

### A. Models for areal spatial structure



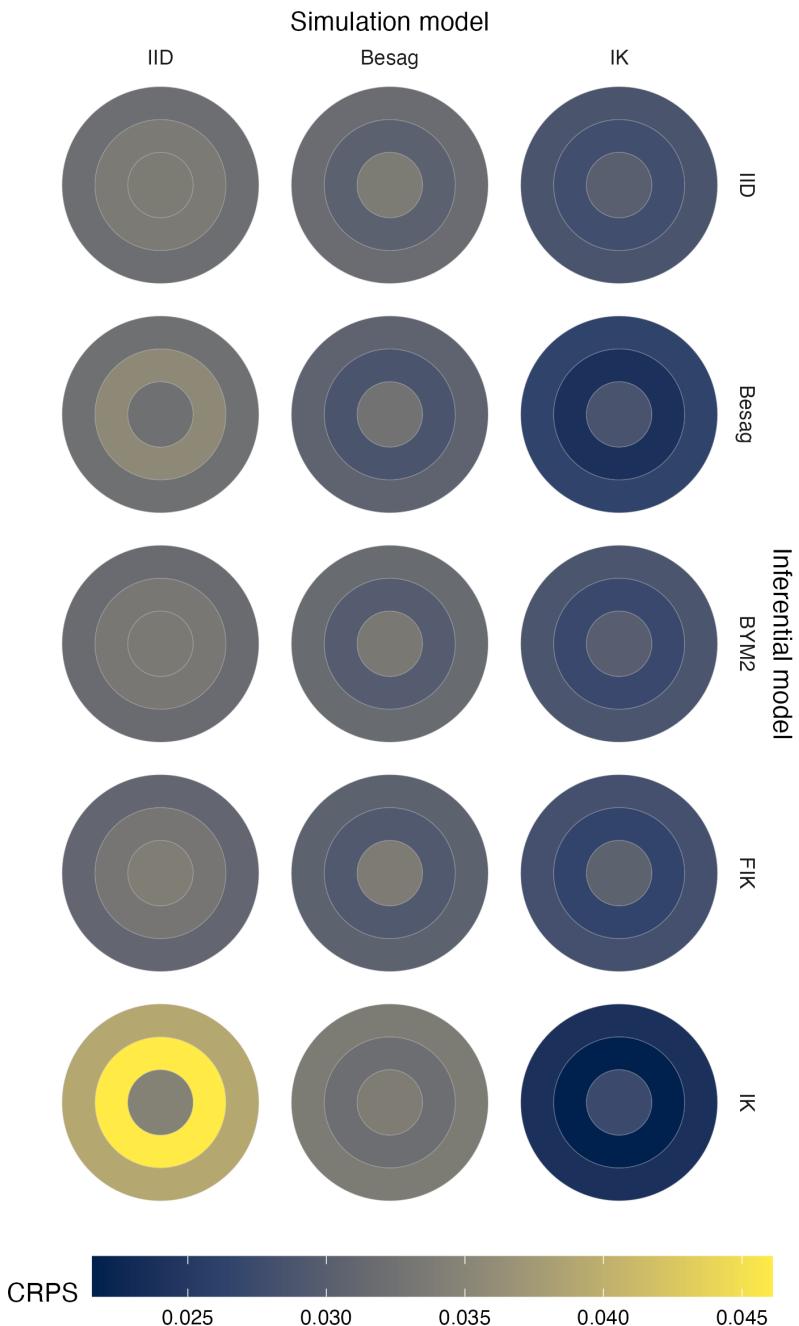
**Figure A.16:** The mean CRPS with 95% credible interval in estimating  $\rho$  using each inferential model and simulation model on the fourth vignette geometry (Panel 4.6D). Credible intervals were generated using 1.96 times the standard error.

*A. Models for areal spatial structure*



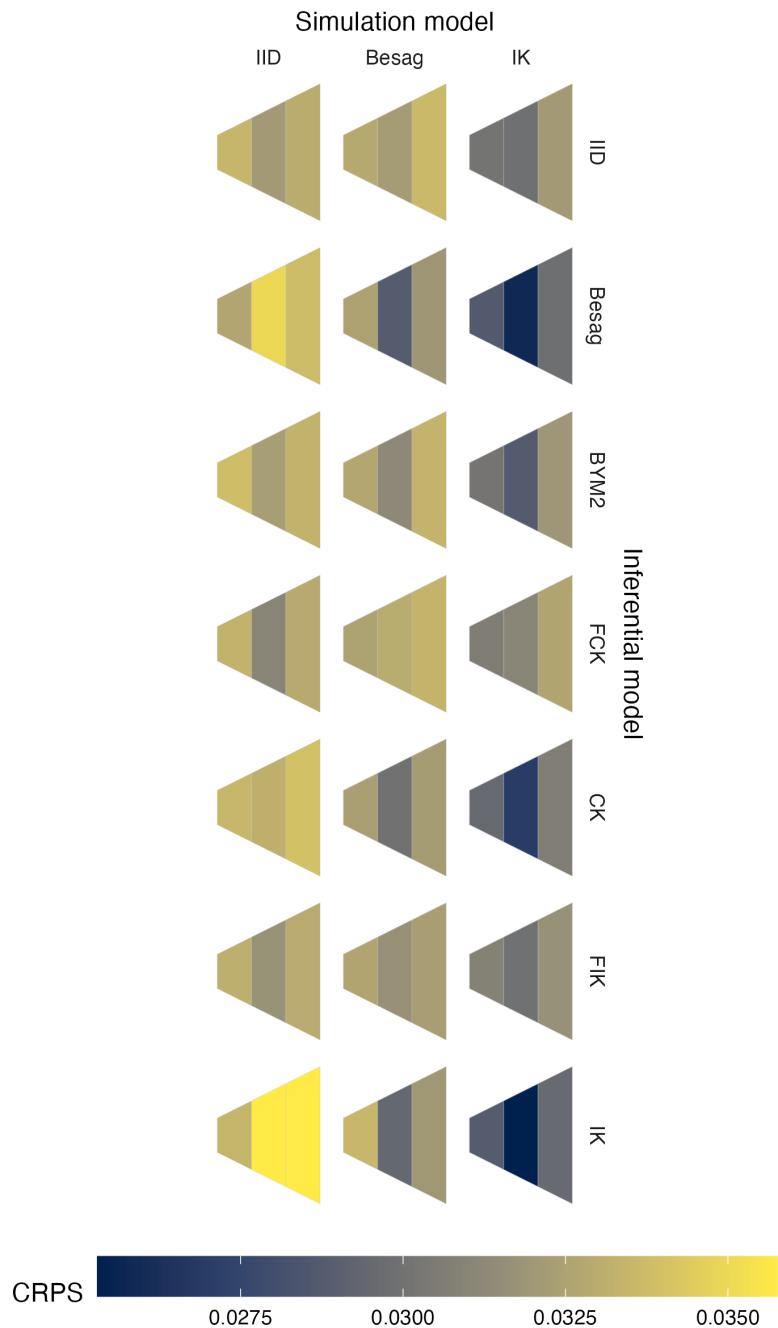
**Figure A.17:** Choropleths showing the mean value of the CRPS in estimating  $\rho$ , under each inferential model and simulation model, at each area of the first vignette geometry (Panel 4.6A).

*A. Models for areal spatial structure*



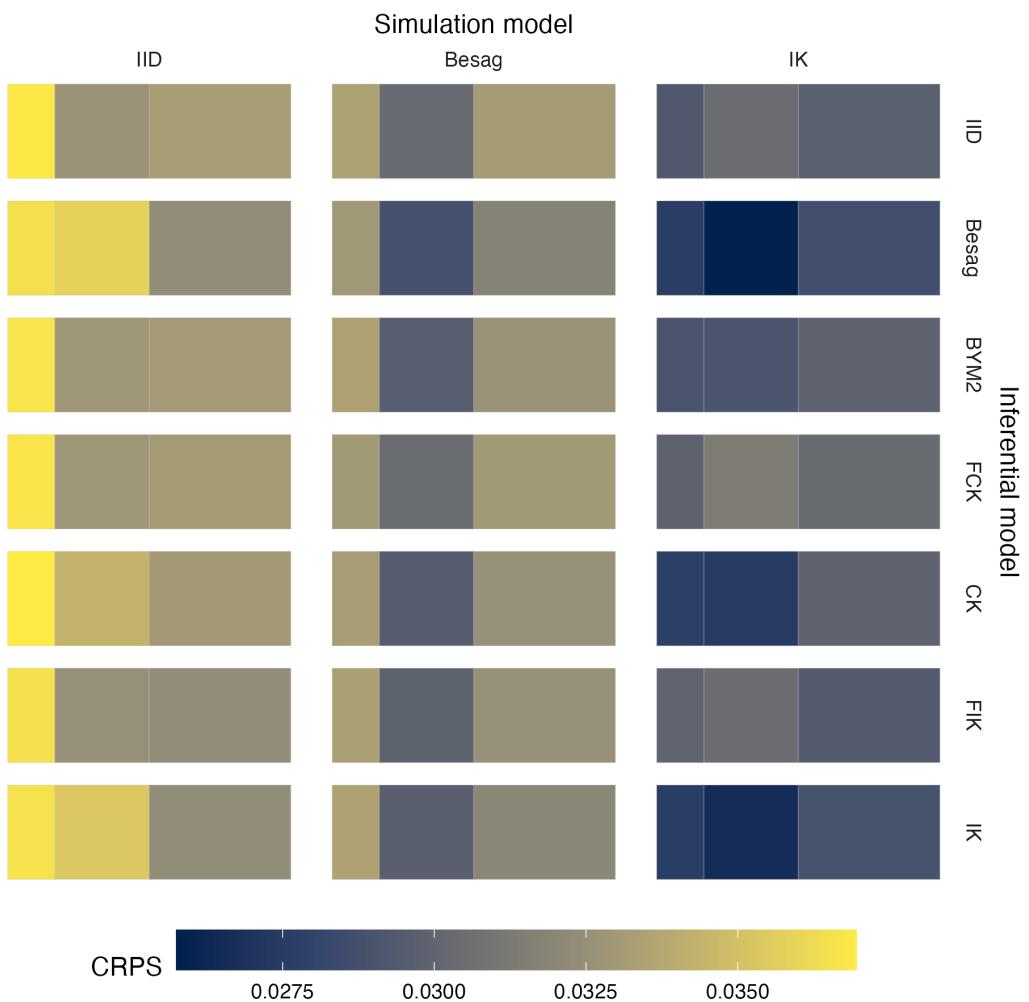
**Figure A.18:** Choropleths showing the mean value of the CRPS in estimating  $\rho$ , under each inferential model and simulation model, at each area of the second vignette geometry (Panel 4.6B).

*A. Models for areal spatial structure*



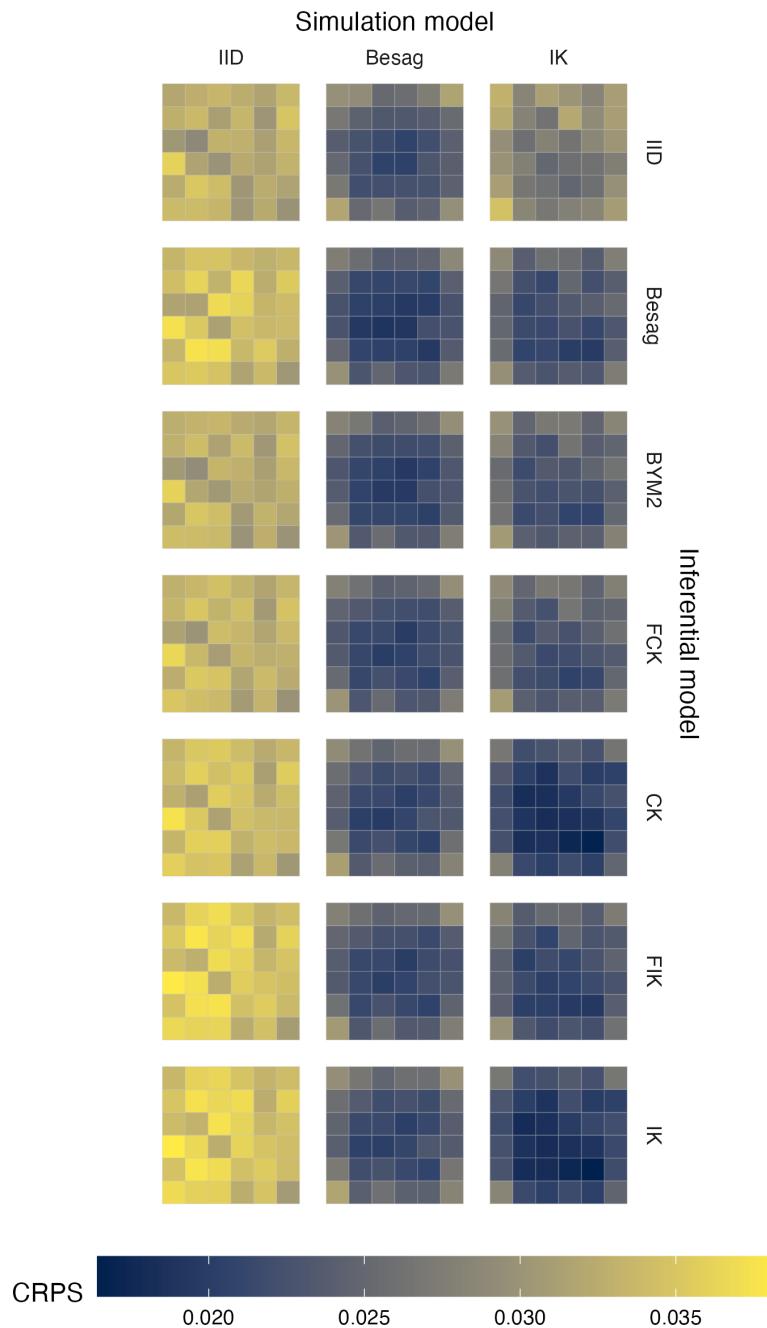
**Figure A.19:** Choropleths showing the mean value of the CRPS in estimating  $\rho$ , under each inferential model and simulation model, at each area of the third vignette geometry (Panel 4.6C).

*A. Models for areal spatial structure*



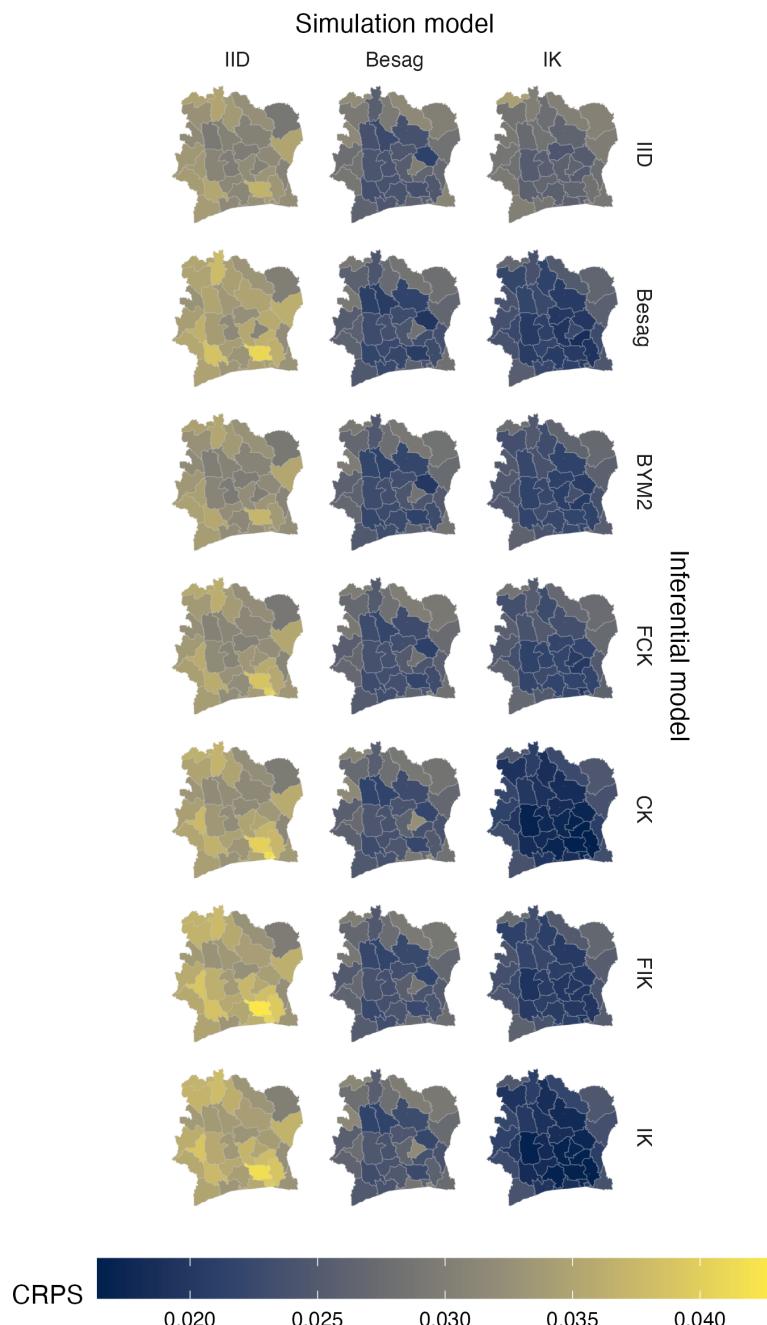
**Figure A.20:** Choropleths showing the mean value of the CRPS in estimating  $\rho$ , under each inferential model and simulation model, at each area of the fourth vignette geometry (Panel 4.6D).

*A. Models for areal spatial structure*



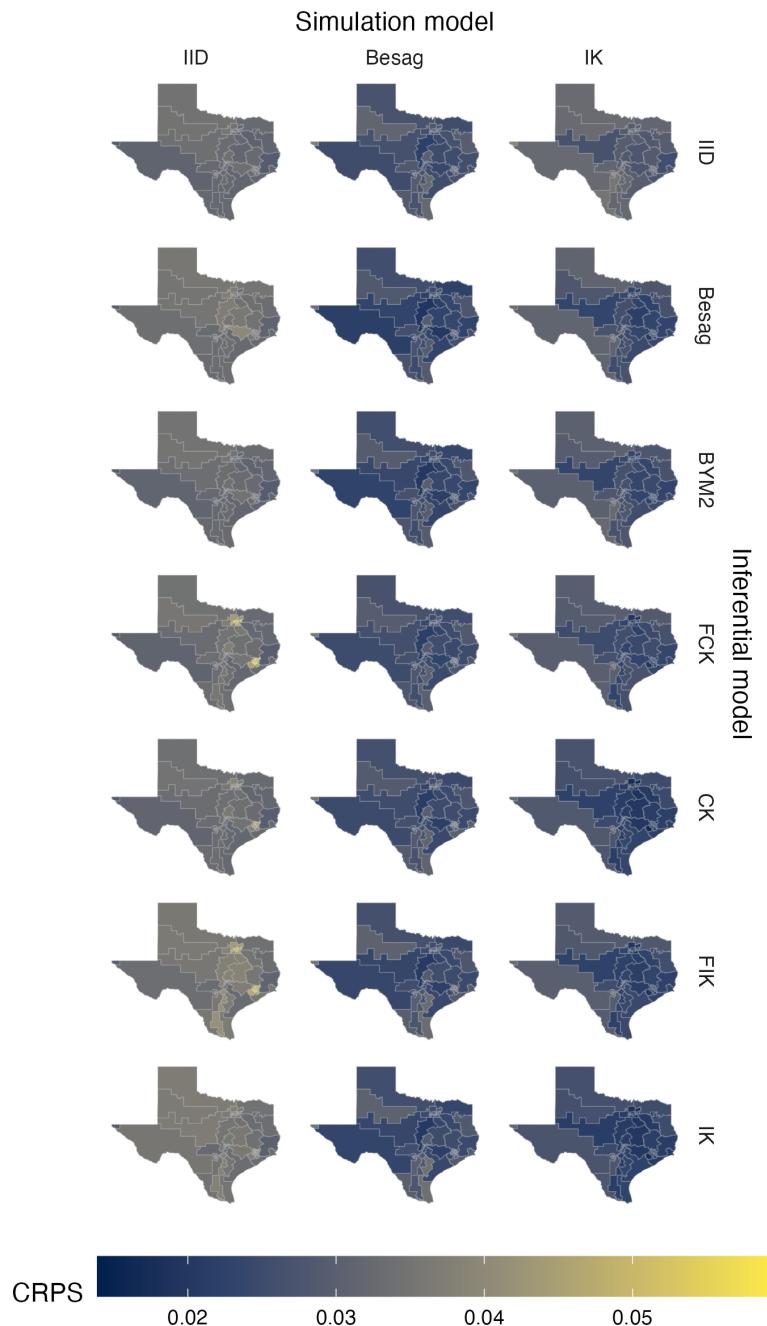
**Figure A.21:** Choropleths showing the mean value of the CRPS in estimating  $\rho$ , under each inferential model and simulation model, at each area of the grid geometry (Panel 4.6E).

*A. Models for areal spatial structure*



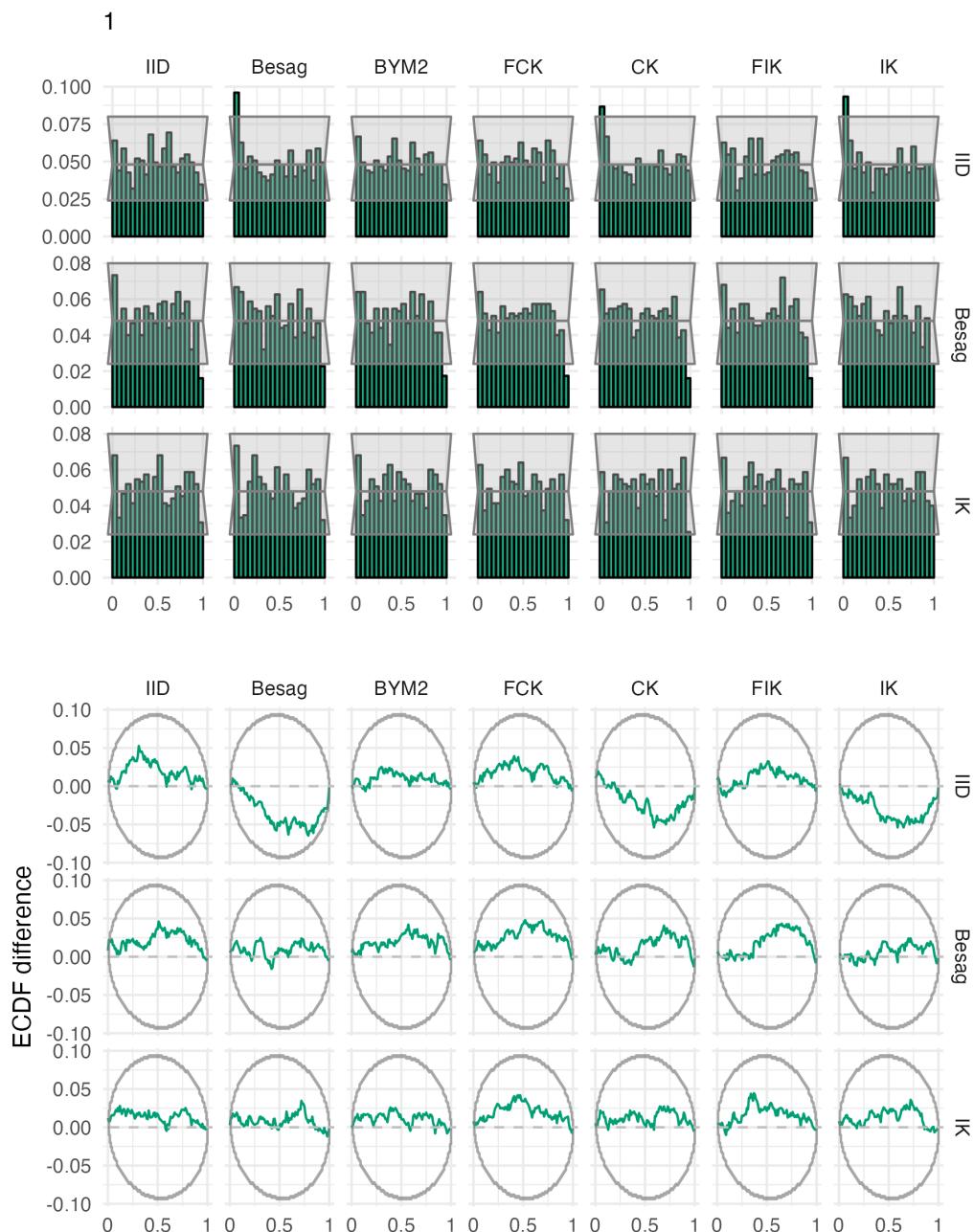
**Figure A.22:** Choropleths showing the mean value of the CRPS in estimating  $\rho$ , under each inferential model and simulation model, at each area of the Côte d'Ivoire geometry (Panel 4.6F).

*A. Models for areal spatial structure*



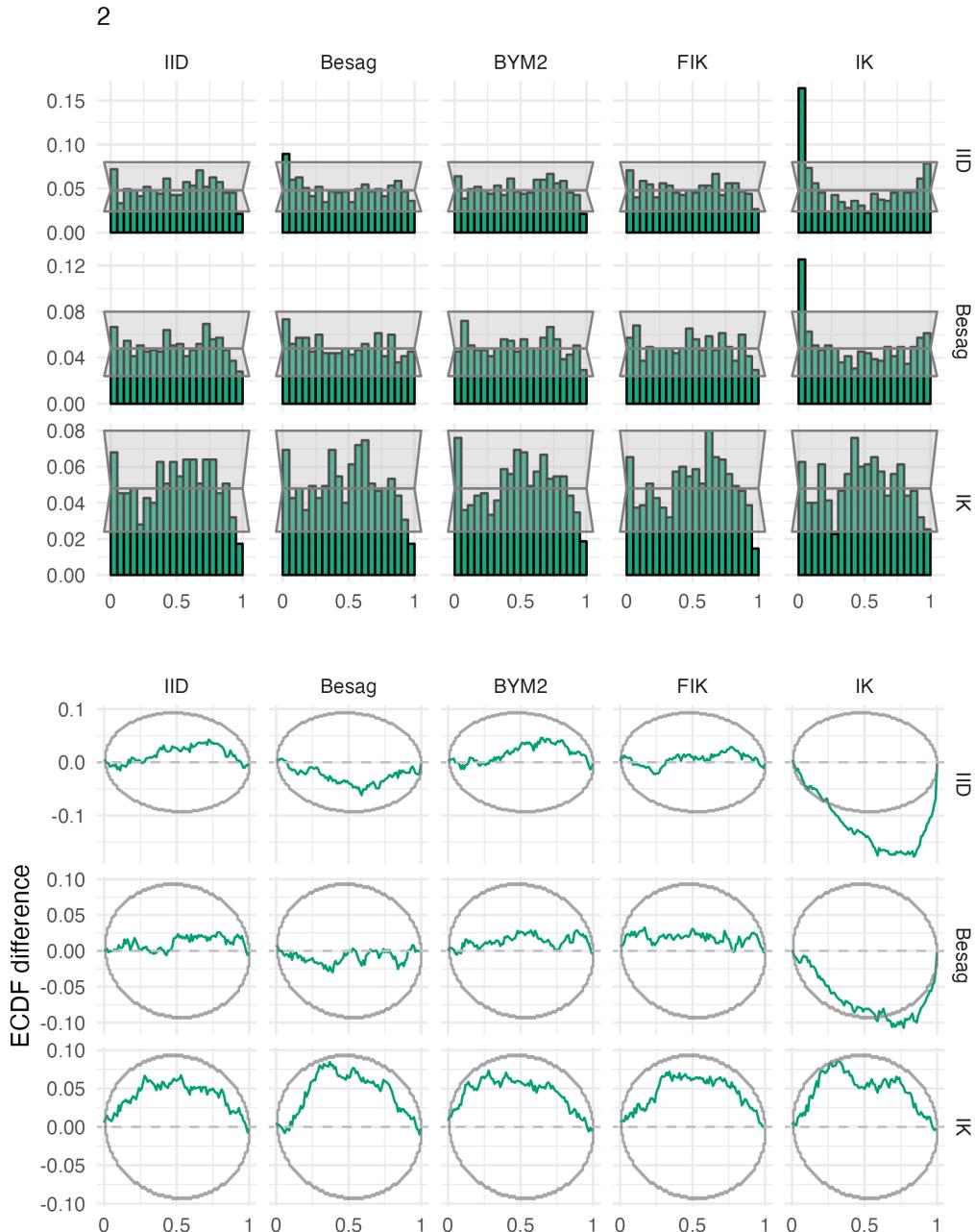
**Figure A.23:** Choropleths showing the mean value of the CRPS in estimating  $\rho$ , under each inferential model and simulation model, at each area of the Texas geometry (Panel 4.6G).

*A. Models for areal spatial structure*



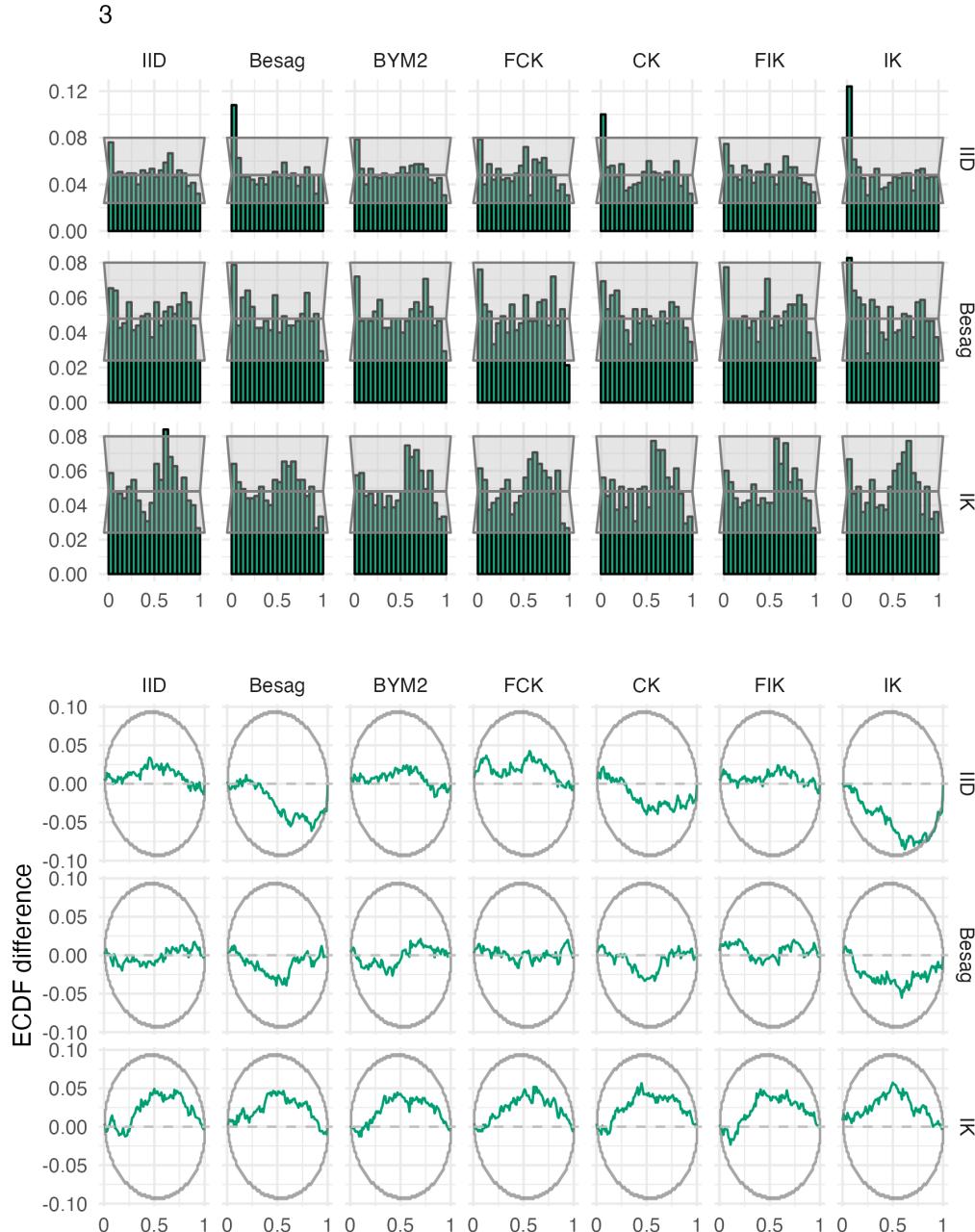
**Figure A.24:** Probability integral transform histograms and empirical cumulative distribution function difference plots for  $\rho$ , under each inferential model and simulation model, for the first vignette geometry (Panel 4.6A).

*A. Models for areal spatial structure*



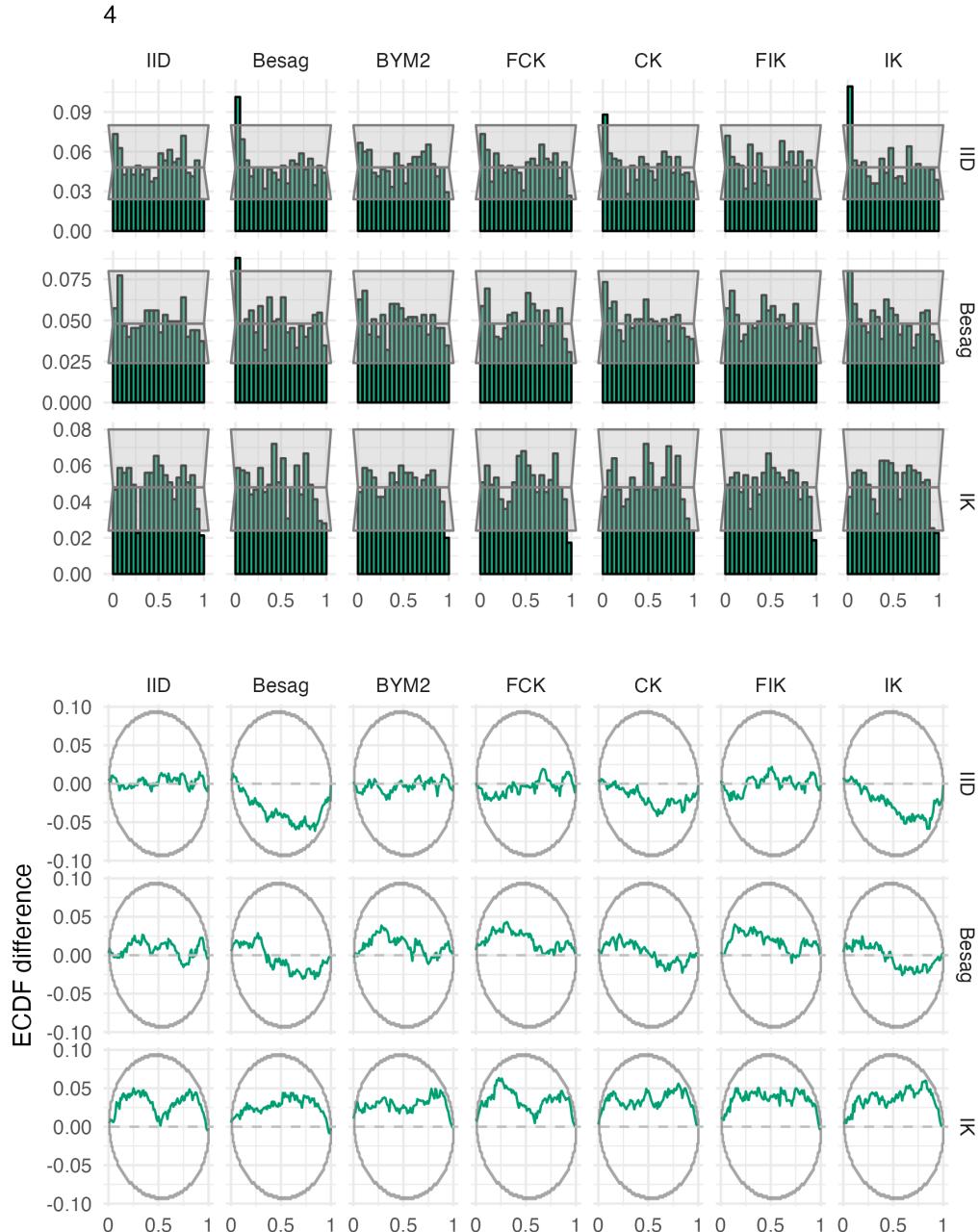
**Figure A.25:** Probability integral transform histograms and empirical cumulative distribution function difference plots for  $\rho$ , under each inferential model and simulation model, for the second vignette geometry (Panel 4.6B).

*A. Models for areal spatial structure*



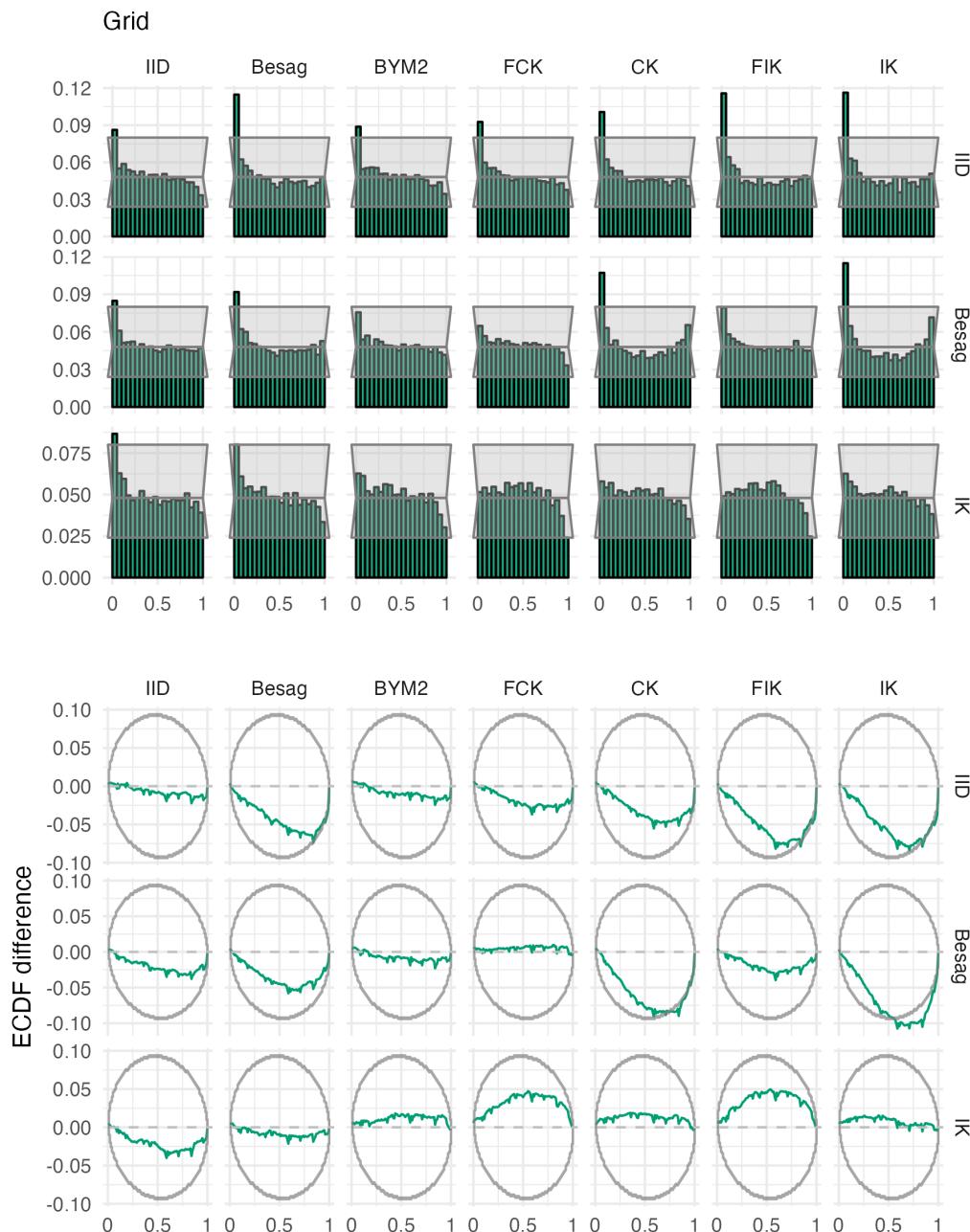
**Figure A.26:** Probability integral transform histograms and empirical cumulative distribution function difference plots for  $\rho$ , under each inferential model and simulation model, for the third vignette geometry (Panel 4.6C).

*A. Models for areal spatial structure*



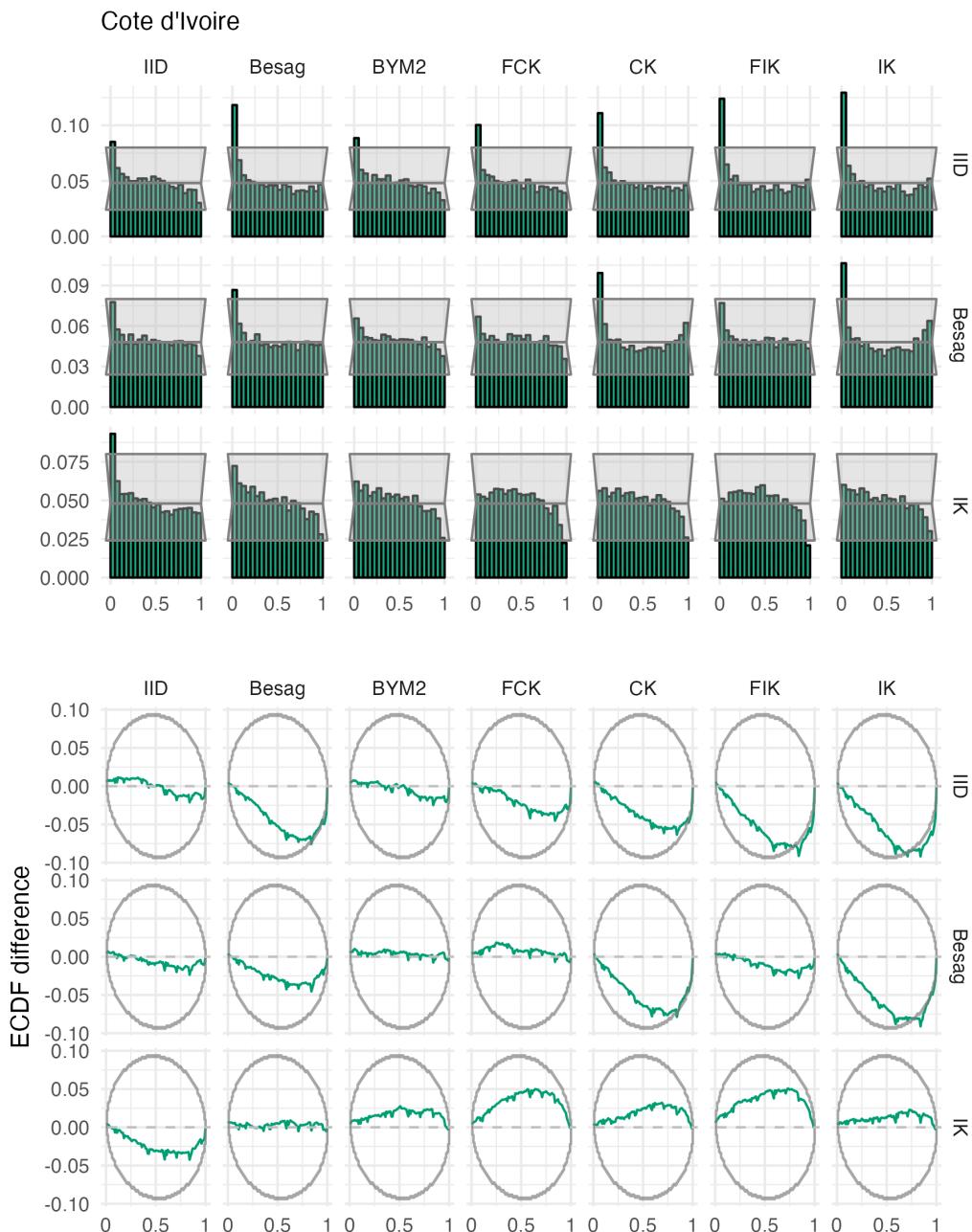
**Figure A.27:** Probability integral transform histograms and empirical cumulative distribution function difference plots for  $\rho$ , under each inferential model and simulation model, for the fourth vignette geometry (Panel 4.6D).

*A. Models for areal spatial structure*



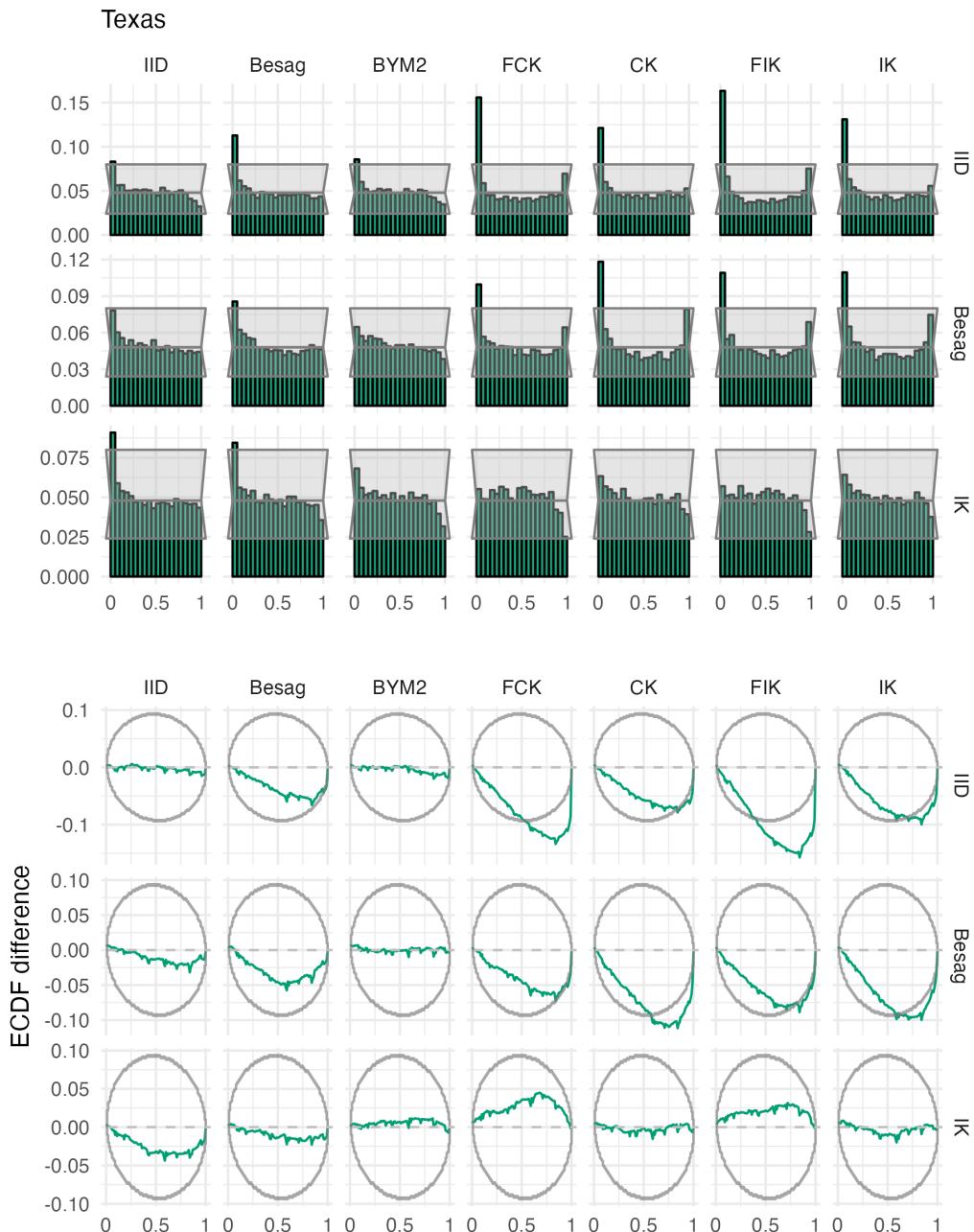
**Figure A.28:** Probability integral transform histograms and empirical cumulative distribution function difference plots for  $\rho$ , under each inferential model and simulation model, for the grid geometry (Panel 4.6E).

*A. Models for areal spatial structure*



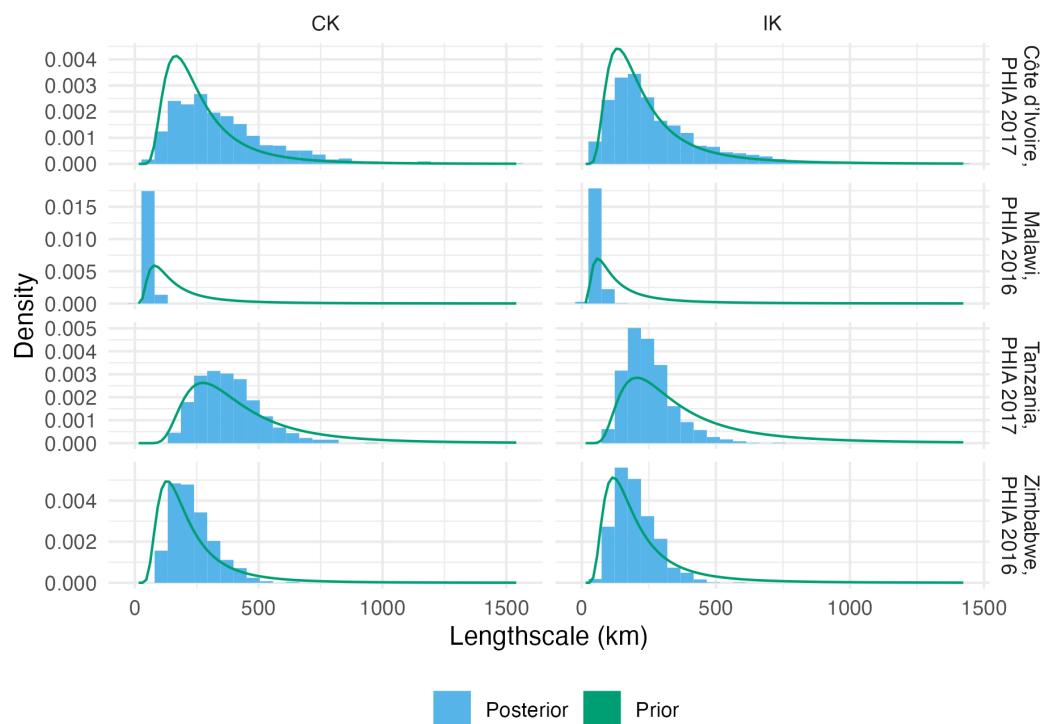
**Figure A.29:** Probability integral transform histograms and empirical cumulative distribution function difference plots for  $\rho$ , under each inferential model and simulation model, for the Côte d'Ivoire geometry (Panel 4.6F).

*A. Models for areal spatial structure*



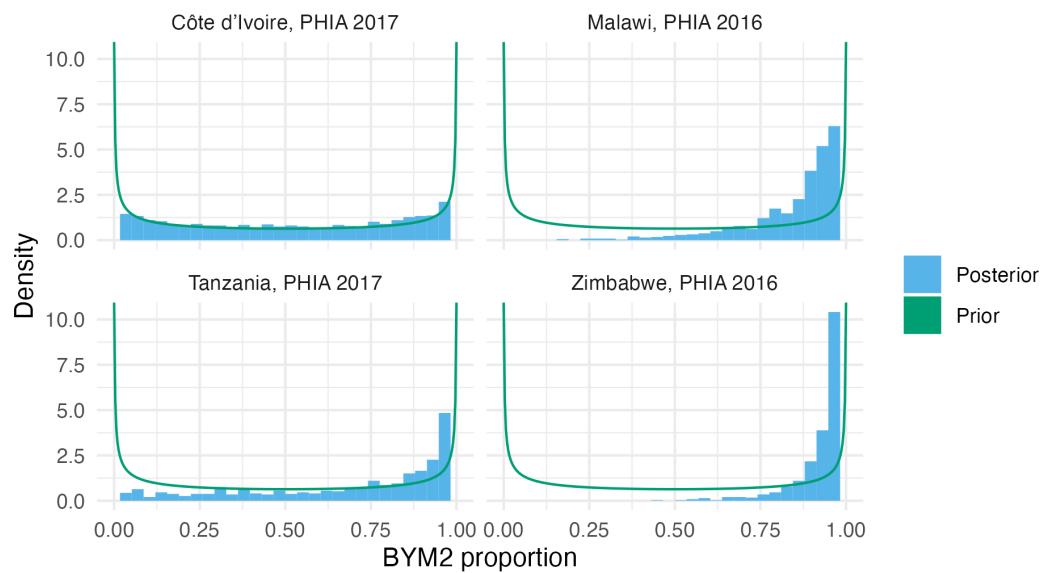
**Figure A.30:** Probability integral transform histograms and empirical cumulative distribution function difference plots for  $\rho$ , under each inferential model and simulation model, for the Texas geometry (Panel 4.6G).

*A. Models for areal spatial structure*



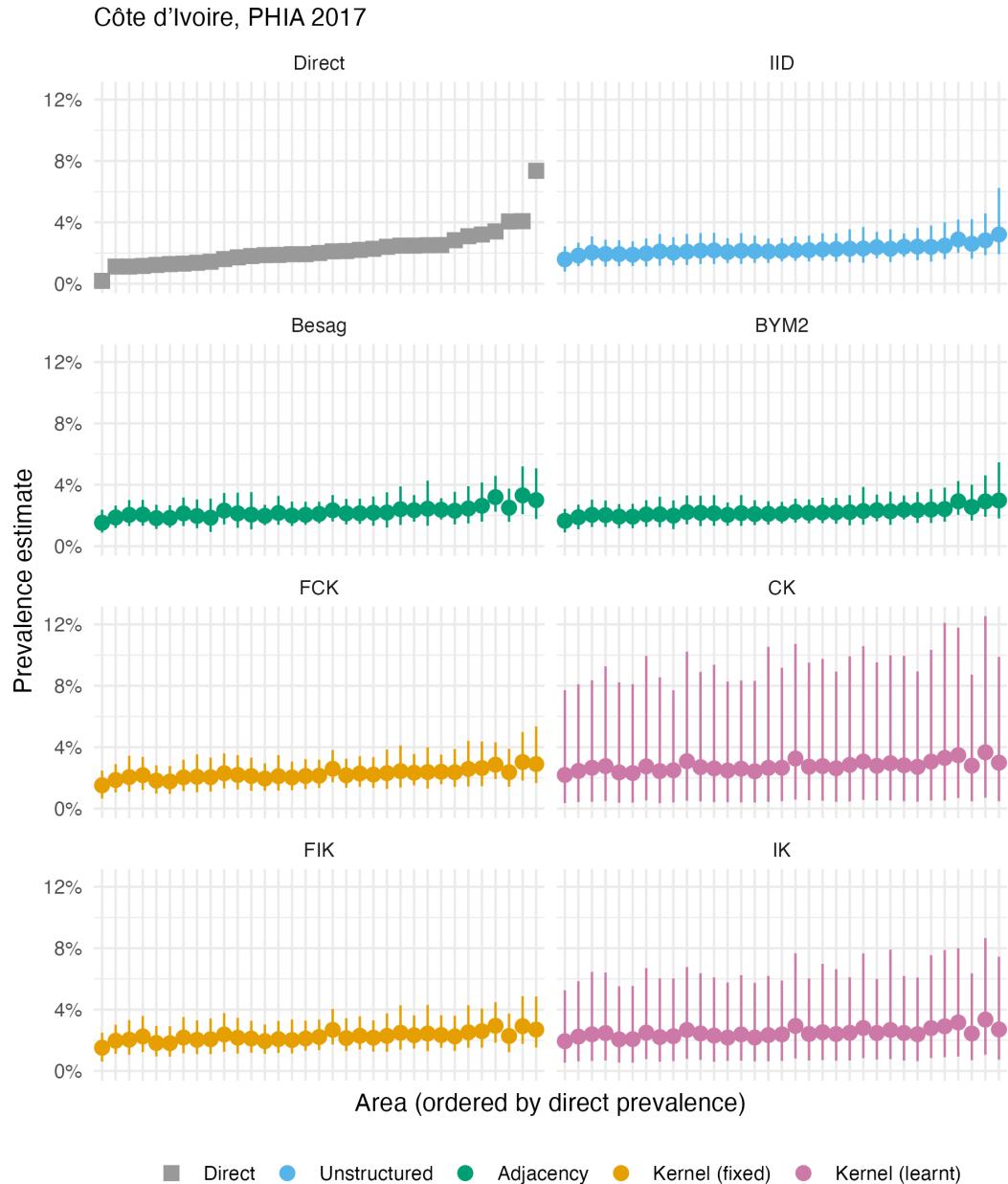
**Figure A.31:** The lengthscale hyperparameter prior and posterior distributions for each of the four considered PHIA surveys (Table 4.3), using both the CK and IK inferential models.

*A. Models for areal spatial structure*



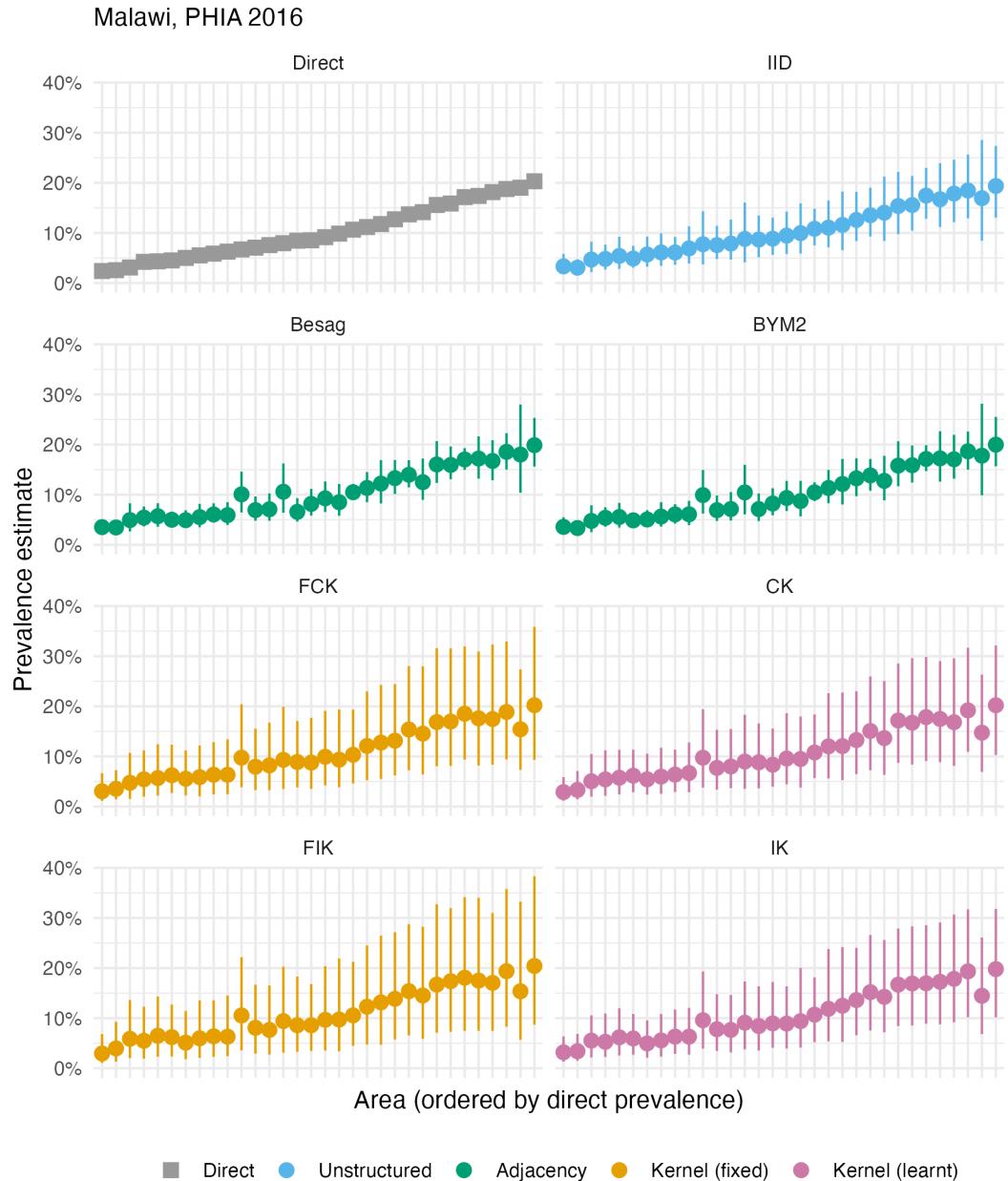
**Figure A.32:** The BYM2 proportion hyperparameter prior and posterior distributions for each of the four considered PHIA surveys (Table 4.3). A value of zero corresponds to IID noise. A value of one corresponds to Besag noise. For each survey, excluding the Côte d'Ivoire 2017 PHIA, the posterior distribution for the BYM2 proportion is concentrated towards a value of one. This result can be interpreted as suggesting that the variation in HIV prevalence from these surveys is spatially structured.

*A. Models for areal spatial structure*



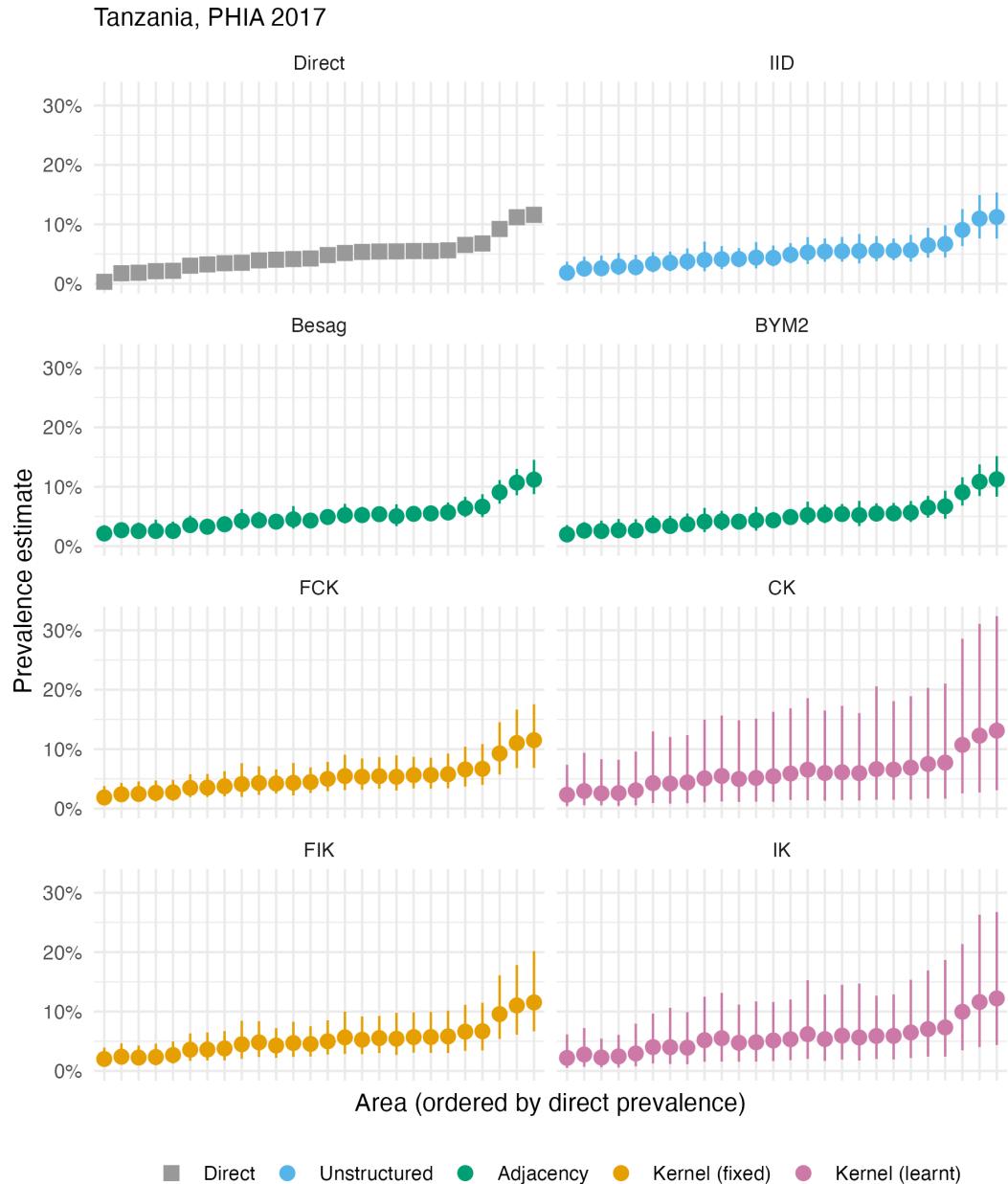
**Figure A.33:** The HIV prevalence posterior mean and 95% credible interval for each area of Côte d'Ivoire, based on the 2017 PHIA survey. Direct estimates obtained from the survey are as shown in Panel 4.10A.

*A. Models for areal spatial structure*



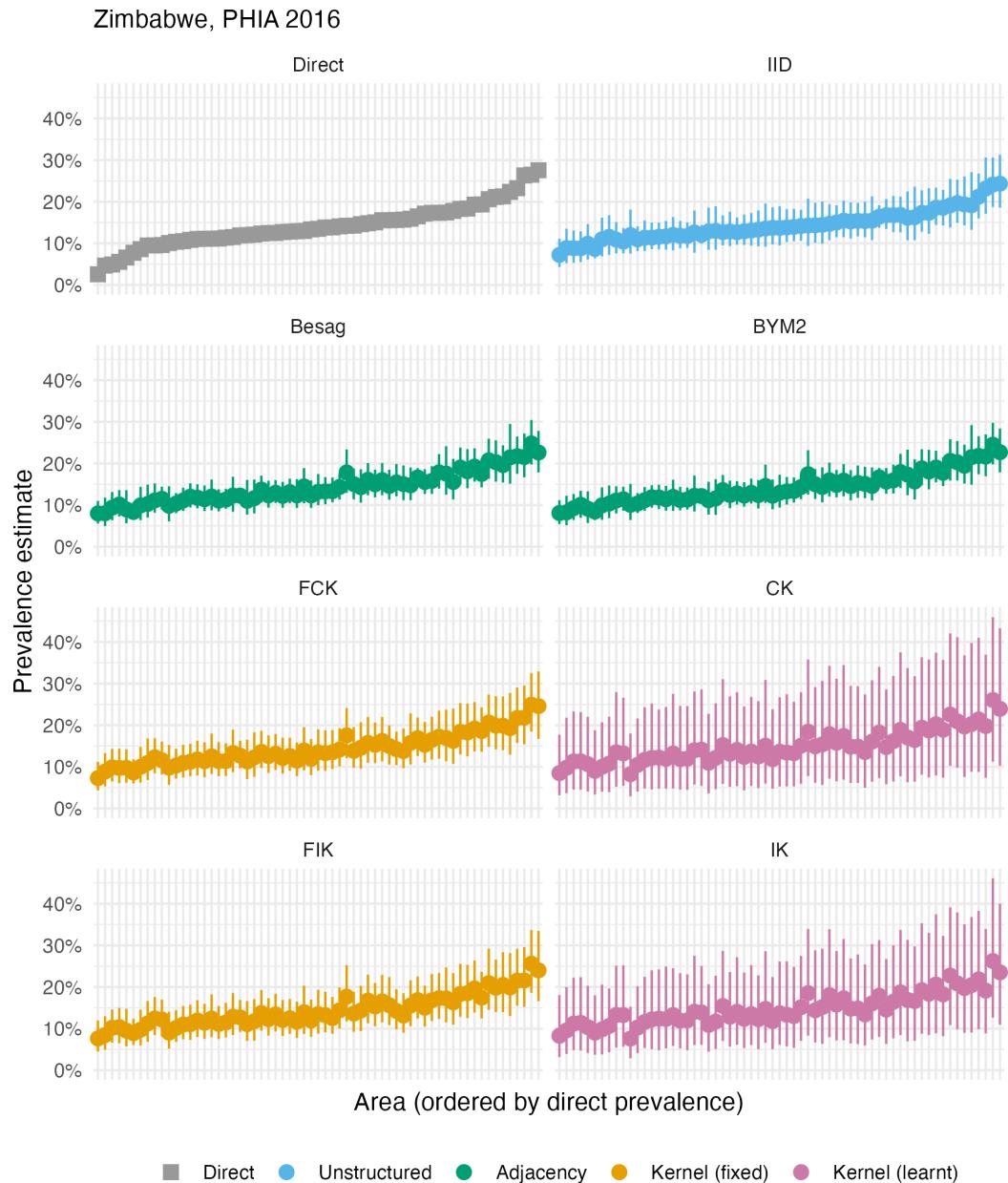
**Figure A.34:** The HIV prevalence posterior mean and 95% credible interval for each area of Malawi, based on the 2016 PHIA survey. Direct estimates obtained from the survey are as shown in Panel 4.10B.

*A. Models for areal spatial structure*



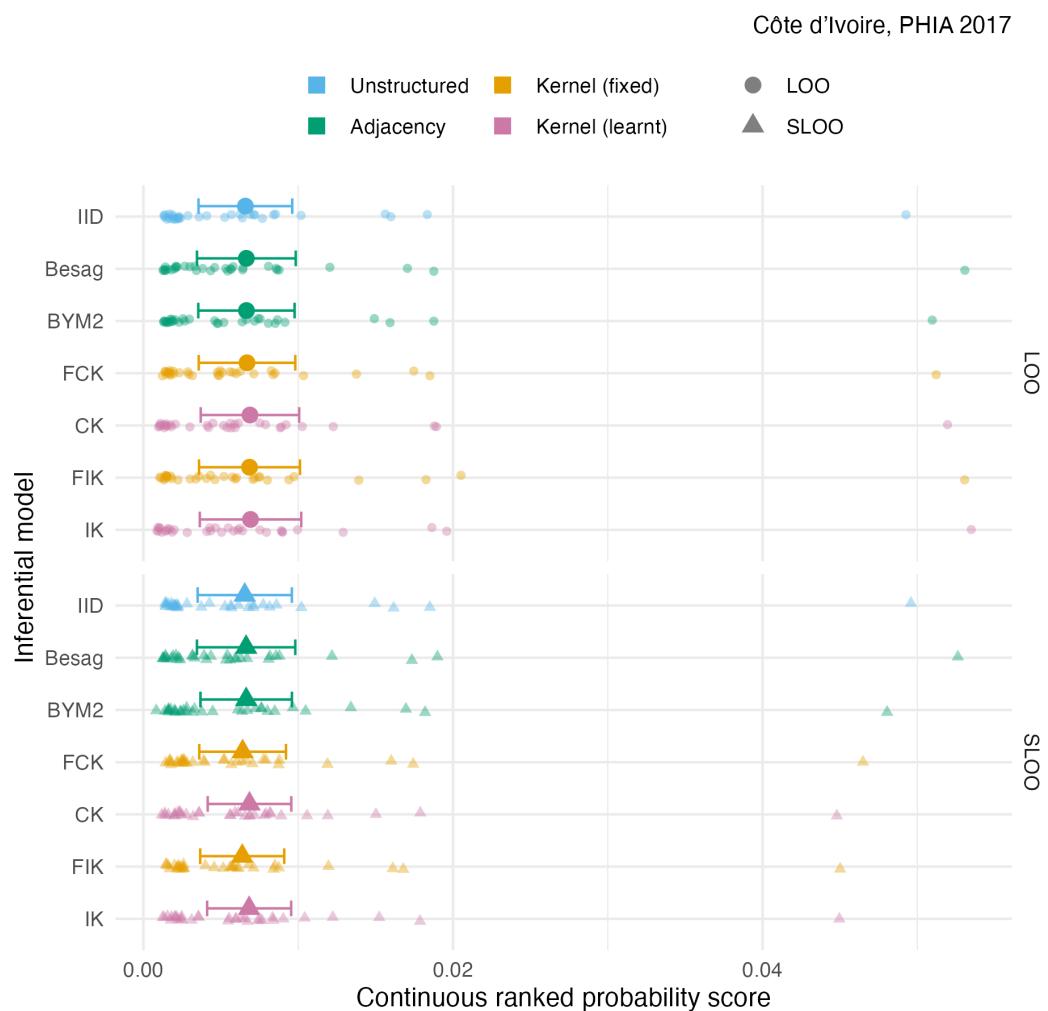
**Figure A.35:** The HIV prevalence posterior mean and 95% credible interval for each area of Tanzania, based on the 2017 PHIA survey. Direct estimates obtained from the survey are as shown in Panel 4.10C.

*A. Models for areal spatial structure*



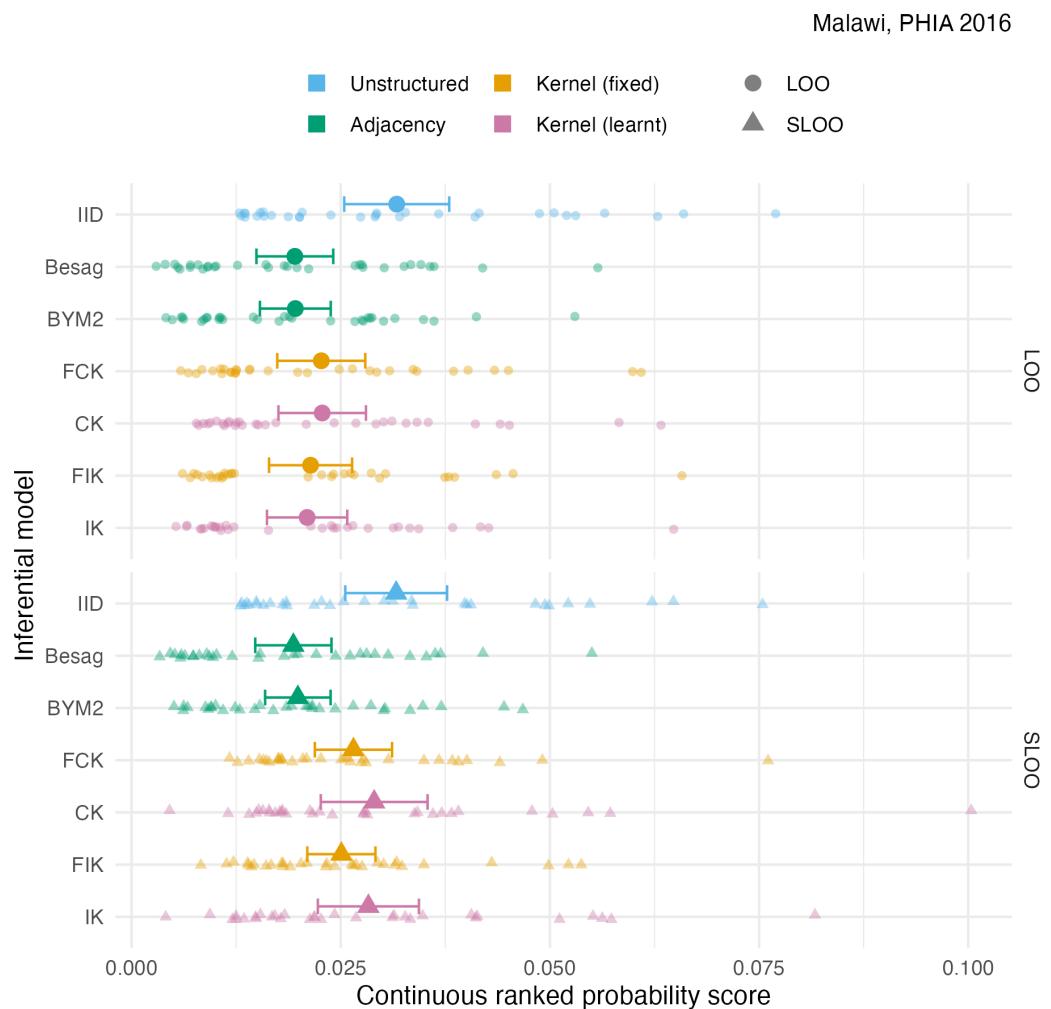
**Figure A.36:** The HIV prevalence posterior mean and 95% credible interval for each area of Zimbabwe, based on the 2016 PHIA survey. Direct estimates obtained from the survey are as shown in Panel 4.10D.

*A. Models for areal spatial structure*



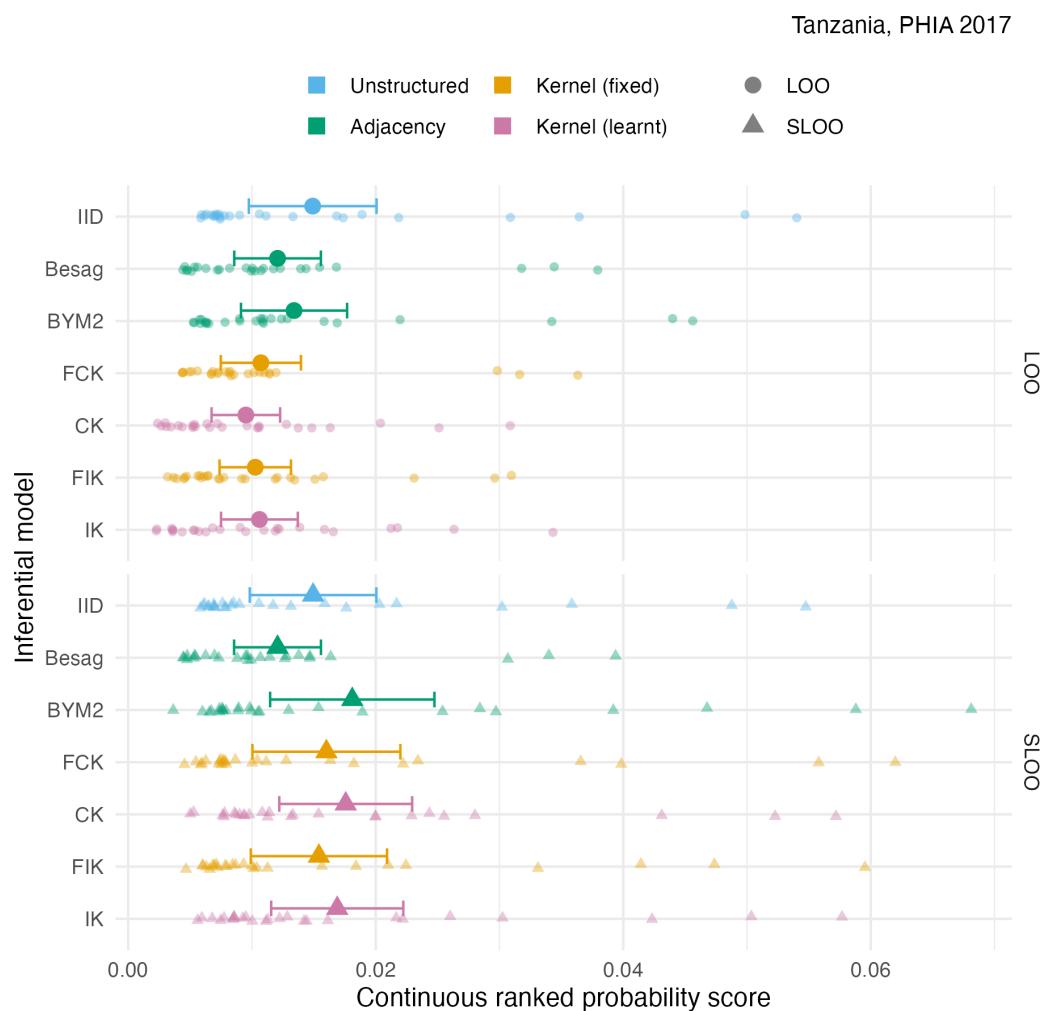
**Figure A.37:** The pointwise CRPS in estimating  $\rho_i$  using either leave-one-out or spatial leave-one-out cross-validation, with mean and 95% credible interval for the Côte d'Ivoire 2017 PHIA survey (Panel 4.10A).

*A. Models for areal spatial structure*



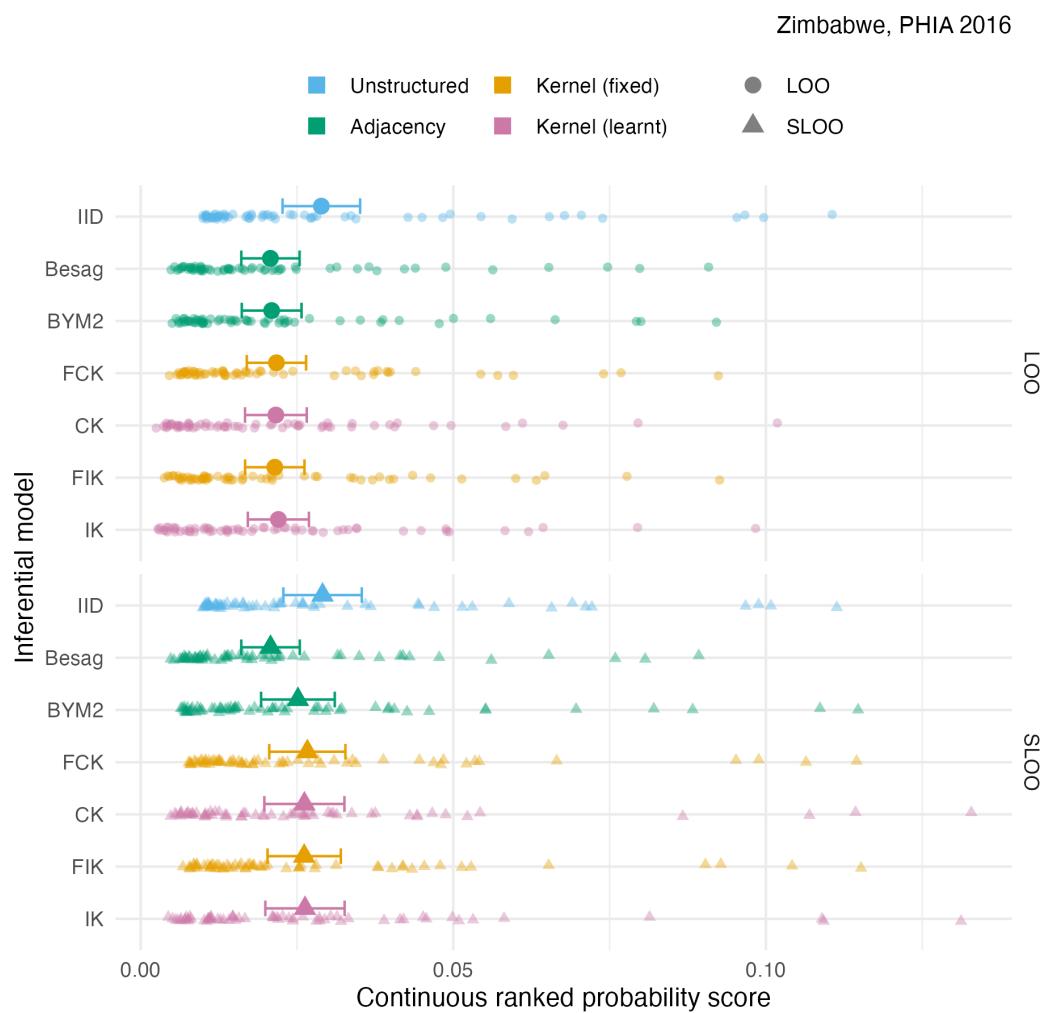
**Figure A.38:** The pointwise CRPS in estimating  $\rho_i$  using either leave-one-out or spatial leave-one-out cross-validation, with mean and 95% credible interval, for the Malawi 2016 PHIA survey 4.10B.

*A. Models for areal spatial structure*



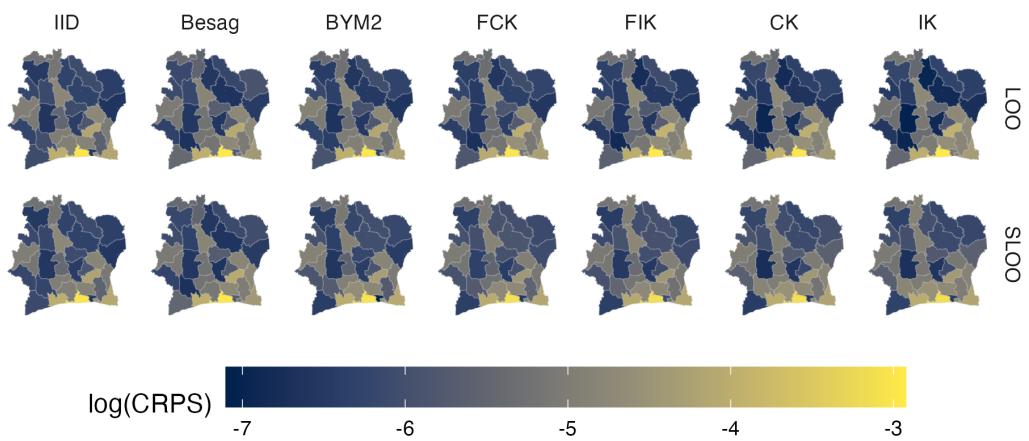
**Figure A.39:** The pointwise CRPS in estimating  $\rho_i$  using either leave-one-out or spatial leave-one-out cross-validation, with mean and 95% credible interval, for the Tanzania 2017 PHIA survey 4.10C.

*A. Models for areal spatial structure*

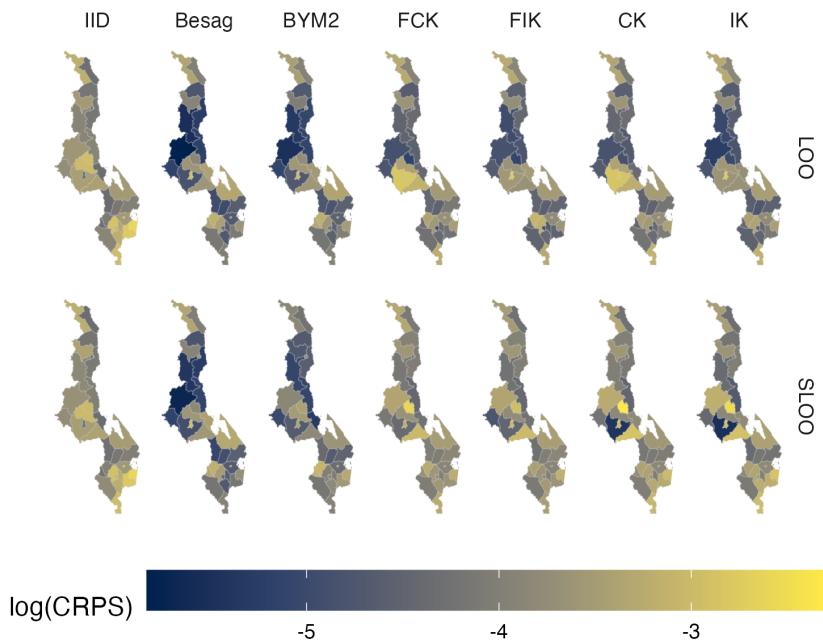


**Figure A.40:** The pointwise CRPS in estimating  $\rho_i$  using either leave-one-out or spatial leave-one-out cross-validation, with mean and 95% credible interval, for the Zimbabwe 2016 PHIA survey 4.10D.

*A. Models for areal spatial structure*

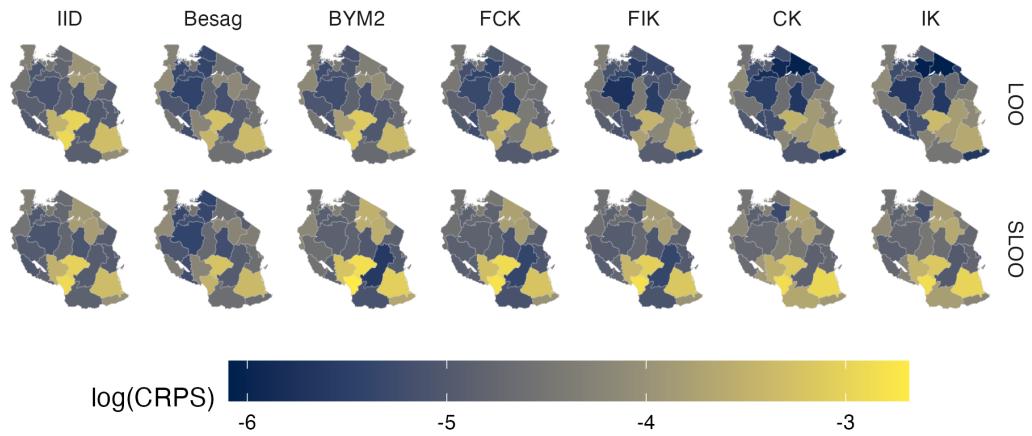


**Figure A.41:** Choropleth showing the pointwise CRPS in estimating  $\rho_i$  using either leave-one-out or spatial leave-one-out cross-validation for the Côte d'Ivoire 2017 PHIA survey (Panel 4.10A).

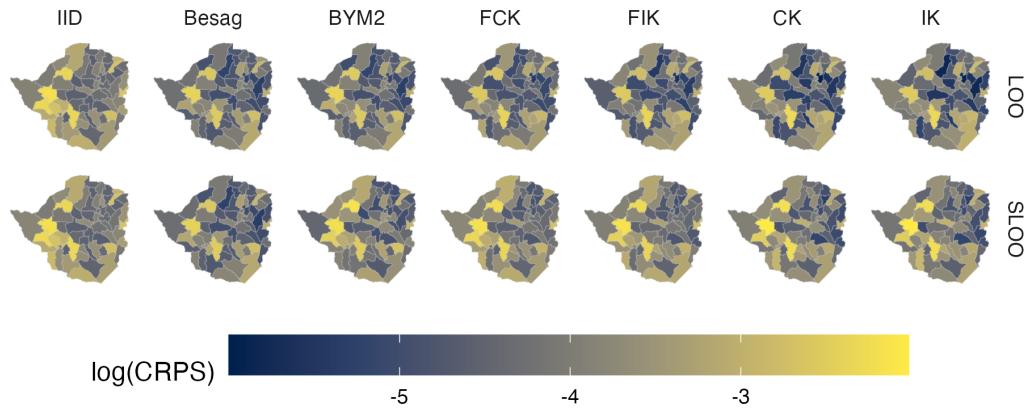


**Figure A.42:** Choropleth showing the pointwise CRPS in estimating  $\rho_i$  using either leave-one-out or spatial leave-one-out cross-validation for the Malawi 2016 PHIA survey (Panel 4.10B).

*A. Models for areal spatial structure*

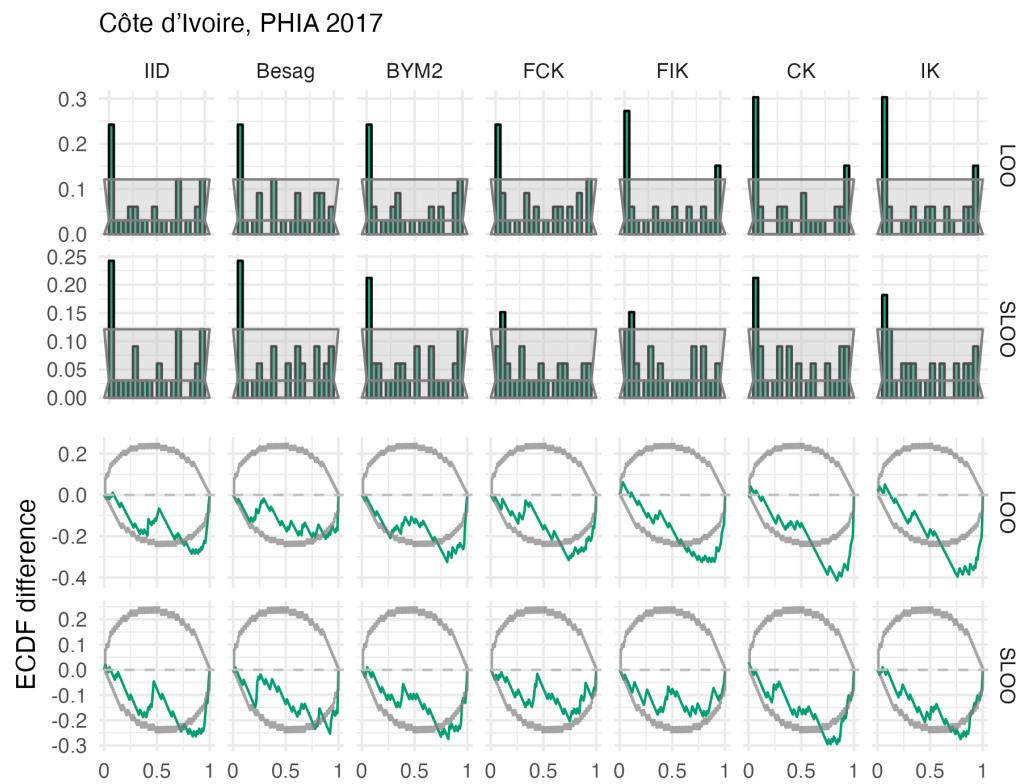


**Figure A.43:** Choropleth showing the pointwise CRPS in estimating  $\rho_i$  using either leave-one-out or spatial leave-one-out cross-validation for the Tanzania 2017 PHIA survey (Panel 4.10C).



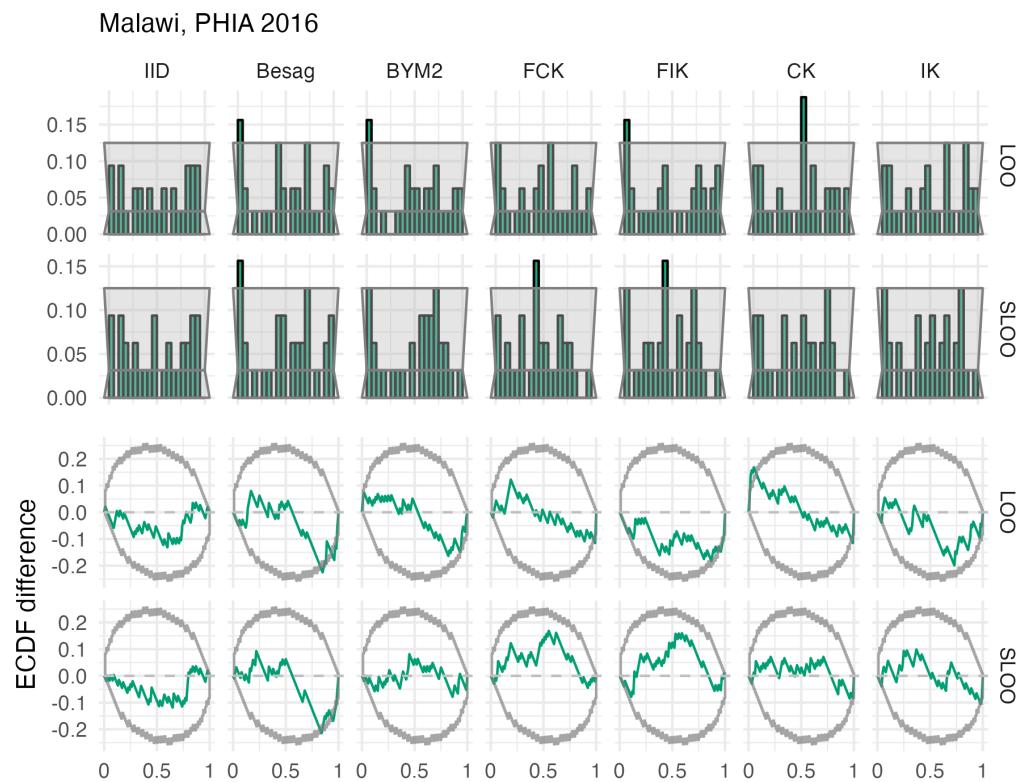
**Figure A.44:** Choropleth showing the pointwise CRPS in estimating  $\rho_i$  using either leave-one-out or spatial leave-one-out cross-validation for the Zimbabwe 2016 PHIA survey (Panel 4.10D).

*A. Models for areal spatial structure*



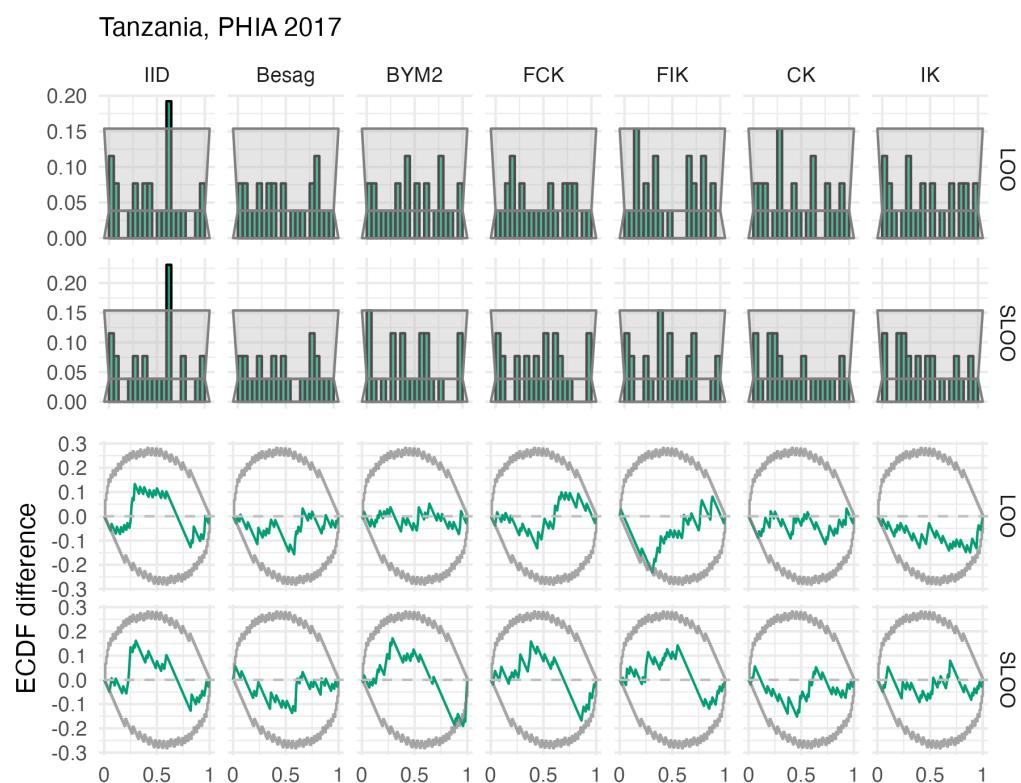
**Figure A.45:** Probability integral transform histograms and empirical cumulative distribution function difference plots in estimating  $\rho$  for the Côte d'Ivoire 2017 PHIA survey (Panel 4.10A).

*A. Models for areal spatial structure*



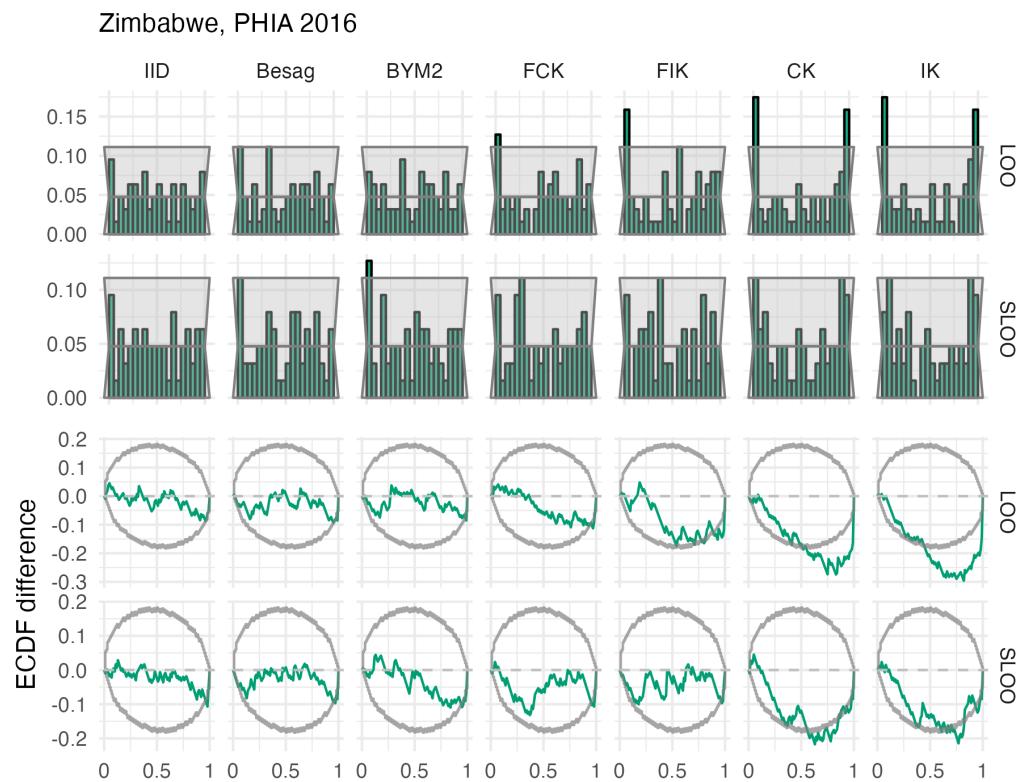
**Figure A.46:** Probability integral transform histograms and empirical cumulative distribution function difference plots in estimating  $\rho$  for the Malawi 2016 PHIA survey (Panel 4.10B).

*A. Models for areal spatial structure*



**Figure A.47:** Probability integral transform histograms and empirical cumulative distribution function difference plots in estimating  $\rho$  for the Tanzania 2017 PHIA survey (Panel 4.10C).

*A. Models for areal spatial structure*



**Figure A.48:** Probability integral transform histograms and empirical cumulative distribution function difference plots in estimating  $\rho$  for the Zimbabwe 2016 PHIA survey (Panel 4.10D).

# B

## A model for risk group proportions

### B.1 The Global AIDS Strategy

**Table B.1:** Prioritisation strata for AGYW given by UNAIDS (2021b) based on to HIV incidence in the general population and behavioural risk.

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.2:** Commitments recommended by UNAIDS (2021b) to be met for each HIV intervention, given in terms of the proportion of the AGYW prioritisation strata reached. The symbol “-” represents no commitment.

Intervention	Low	Moderate	High	Very High
Condoms and lube for those with non-regular partners(s), unknown STI status, 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%

*B. A model for risk group proportions*

Intervention	Very			
	Low	Moderate	High	High

## B.2 Household survey data

**Table B.3:** The sample size by age group for each included survey in the analysis. The column “TS question” refers to whether or not the survey included a specific question about transactional sex (TS).

Type	Year	TS question	Sample size				
			15-19	20-24	25-29	Total	
<b>Botswana</b>							
	BAIS	2013	Yes	557	588	649	1794
	Total			557	588	649	1794
<b>Cameroon</b>							
	DHS	2004	No	2678	2210	1732	6620
	DHS	2011	No	3588	3115	2656	9359
	PHIA	2017	No	2140	1923	1851	5914
	DHS	2018	Yes	3349	2463	2345	8157
	Total			11755	9711	8584	30050
<b>Kenya</b>							
	DHS	2003	No	1819	1709	1391	4919
	DHS	2008	No	1767	1743	1420	4930
	DHS	2014	No	2861	2534	2858	8253
	Total			6447	5986	5669	18102
<b>Lesotho</b>							
	DHS	2004	No	1761	1456	1026	4243
	DHS	2009	No	1834	1545	1195	4574
	DHS	2014	No	1537	1293	1069	3899
	PHIA	2017	Yes	1156	1202	1054	3412
	Total			6288	5496	4344	16128
<b>Mozambique</b>							
	AIS	2009	No	1031	1106	987	3124
	DHS	2011	No	3065	2468	2340	7873
	AIS	2015	No	1554	1390	1080	4024
	Total			5650	4964	4407	15021

*B. A model for risk group proportions*

Malawi

	DHS	2000	No	2914	2998	2358	8270
	DHS	2004	No	2407	2823	2135	7365
	DHS	2010	No	5032	4387	4309	13728
	DHS	2015	Yes	5273	5094	3976	14343
	PHIA	2016	Yes	1646	1934	1511	5091
Total				17272	17236	14289	48797

Namibia

	DHS	2000	No	1428	1313	1099	3840
	DHS	2006	No	2203	1870	1544	5617
	DHS	2013	No	1852	1709	1482	5043
	PHIA	2017	Yes	1491	1525	1370	4386
Total				6974	6417	5495	18886

Eswatini

	DHS	2006	No	1265	1027	731	3023
	PHIA	2017	No	1031	895	811	2737
Total				2296	1922	1542	5760

Tanzania

	AIS	2003	No	1466	1377	1270	4113
	AIS	2007	No	2137	1676	1509	5322
	DHS	2010	No	2221	1860	1613	5694
	AIS	2012	No	2474	1923	1815	6212
Total				8298	6836	6207	21341

Uganda

	DHS	2000	No	1687	1541	1326	4554
	DHS	2006	No	1948	1661	1406	5015
	AIS	2011	No	2451	2164	1921	6536
	DHS	2011	No	2025	1664	1614	5303
	DHS	2016	Yes	4276	3782	3014	11072
	PHIA	2016	No	3289	3059	2574	8922
Total				15676	13871	11855	41402

South Africa

	DHS	2016	Yes	1505	1408	1397	4310
Total				1505	1408	1397	4310

Zambia

	DHS	2007	No	1598	1405	1373	4376
	DHS	2013	No	3685	3036	2789	9510

*B. A model for risk group proportions*

	PHIA	2016	Yes	2120	2045	1619	5784
	DHS	2018	Yes	3112	2687	2166	7965
Total				10515	9173	7947	27635
Zimbabwe							
	DHS	1999	No	1468	1230	1011	3709
	DHS	2005	No	2128	1943	1438	5509
	DHS	2010	No	1966	1796	1680	5442
	DHS	2015	Yes	2154	1779	1647	5580
	PHIA	2016	Yes	2114	1817	1573	5504
Total				9830	8565	7349	25744
Total				103063	92173	79734	274970

**Table B.4:** All of that household surveys that were excluded from the risk group model in Section 5.3.

Survey	Reason for exclusion
Mozambique 2003 DHS	No GPS coordinates available to place survey clusters within districts.
Tanzania 2015 DHS	Insufficient sexual behaviour questions.
Uganda 2004 AIS	Unable to download region boundaries.
Zambia 2002 DHS	No GPS coordinates available to place survey clusters within districts.

### B.3 Spatial analysis levels

**Table B.5:** The number of areas and analysis level for each country that was used in the analysis.

Country	Number of areas	Analysis level
Botswana	27	Health district
Cameroon	58	Department
Kenya	47	County
Lesotho	10	District
Mozambique	161	District
Malawi	33	Health district and cities
Namibia	38	District
Eswatini	4	Region
Tanzania	195	District
Uganda	136	District

*B. A model for risk group proportions*

Country	Number of areas	Analysis level
South Africa	52	District
Zambia	116	District
Zimbabwe	63	District

## B.4 Survey questions and risk group allocation

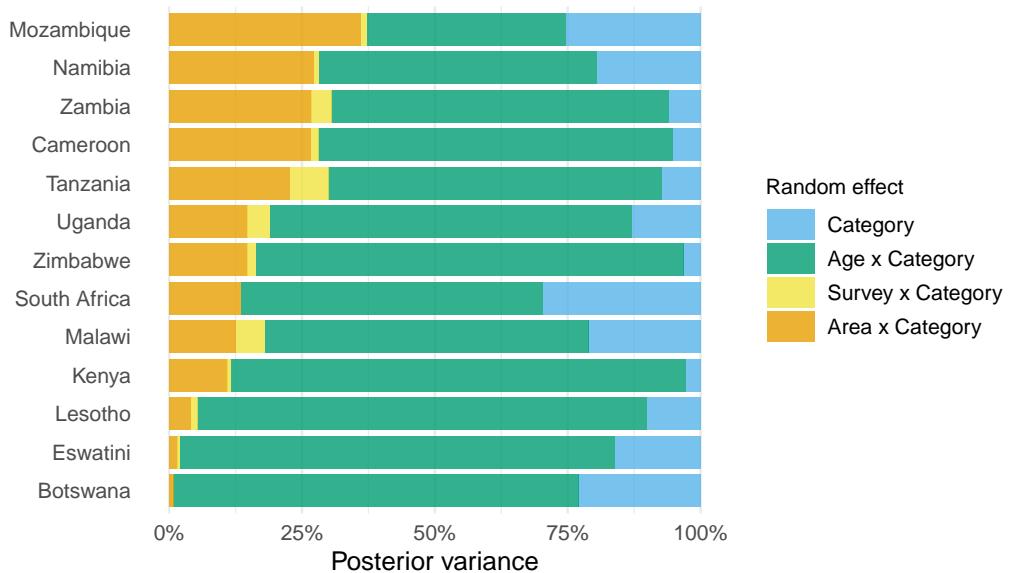
**Table B.6:** The behavioural survey questions included in AIDS Indicator Survey (AIS) and Demographic and Health Surveys (DHS) used to determine AGYW risk group membership.

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.7:** The behavioural survey questions included in Population-Based HIV Impact Assessment (PHIA) surveys used to determine AGYW risk group membership.

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.

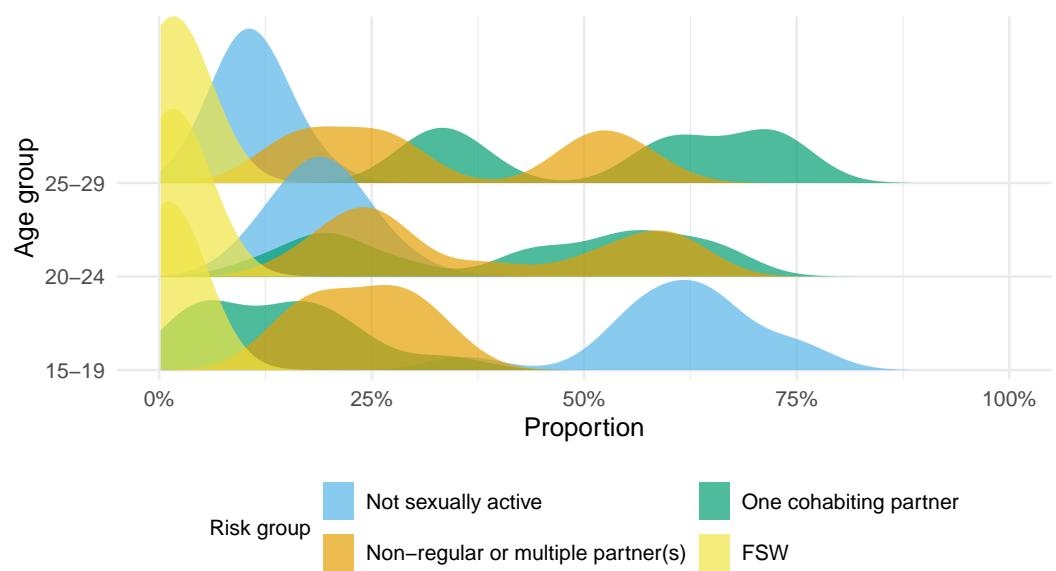
## B. A model for risk group proportions



**Figure B.1:** The proportion of posterior variance explained by each random effect, calculated as a ratio of the random effect variance posterior mean to the sum of all random effect variance posterior means. To allow calculation of this metric by country, the model was run for each country individually.

## B.5 Additional figures

*B. A model for risk group proportions*



**Figure B.2:** For the 20-24 and 25-29 age groups, the proportion of AGYW in the one cohabiting partner and non-regular or multiple partner(s) risk groups was bimodal.

# C

## Fast approximate Bayesian inference

### C.1 Epilepsy example

#### C.1.1 TMB C++ template

```
// epi_l.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);
```

### C. Fast approximate Bayesian inference

```

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);
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);

```

### C. Fast approximate Bayesian inference

```
    return(nll);
}
```

#### C.1.2 Modified TMB C++ template

```
// 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
```

### C. Fast approximate Bayesian inference

```

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);
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();

```

### C. Fast approximate Bayesian inference

```

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

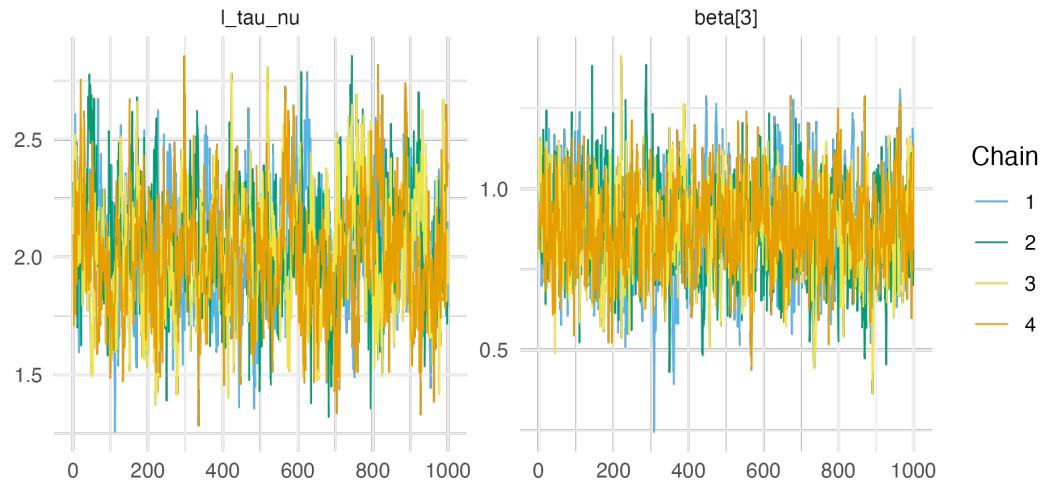
```

// 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
}
```

### C. Fast approximate Bayesian inference



**Figure C.1:** Traceplots for the `tmbstan` parameters with the lowest ESS and highest potential scale reduction factor. These were `l_tau_nu` (an ESS of 377) and `beta[3]` (an  $\hat{R}$  of 1.006).

```

}

transformed parameters {
  vector[N * J] eta = X * beta + nu + E * epsilon;
}

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);
}

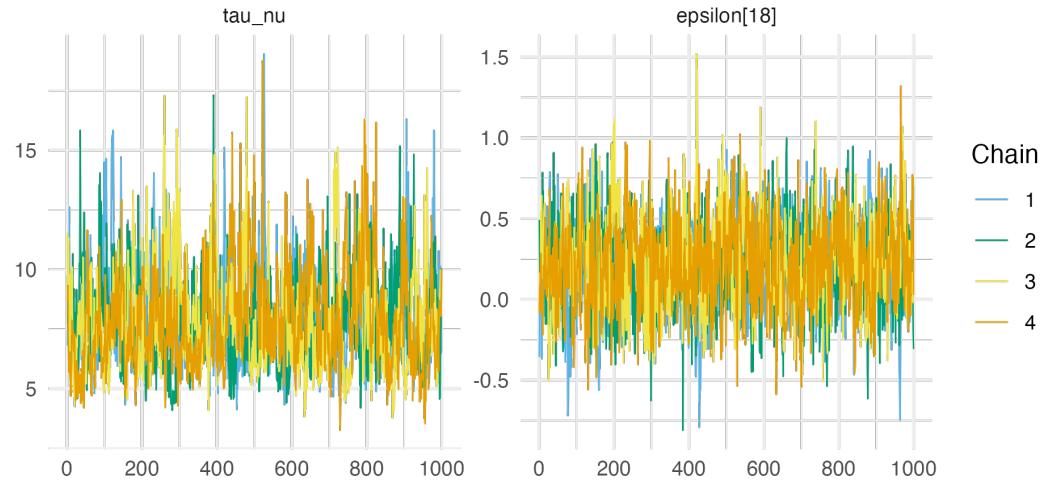
```

#### C.1.4 NUTS convergence and suitability

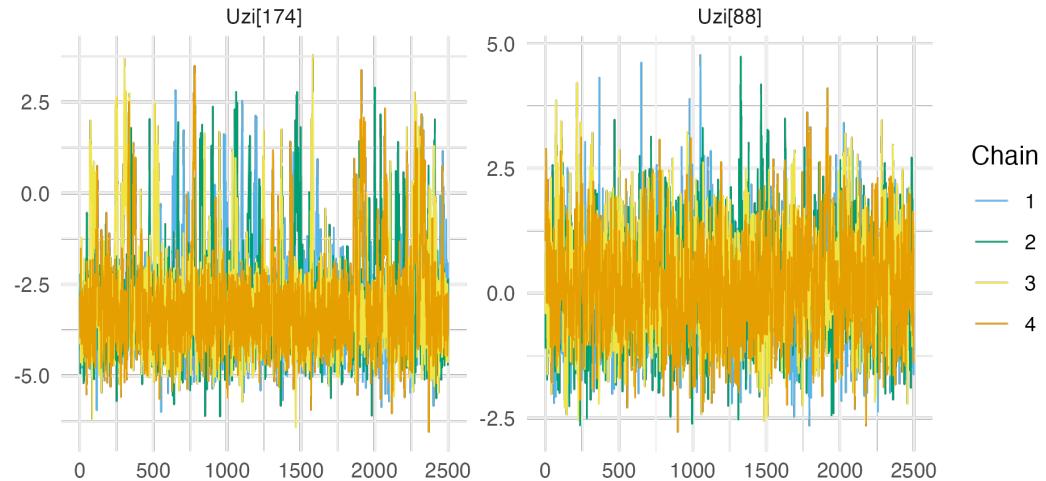
##### C.1.4.1 `tmbstan`

##### C.1.4.2 `rstan`

### C. Fast approximate Bayesian inference



**Figure C.2:** Traceplots for the `rstan` parameters with the lowest ESS and highest potential scale reduction factor. These were `tau_nu` (an ESS of 437) and `tau_nu` (an  $\hat{R}$  of 1.009). Rather than plotting the traceplot for `tau_nu` twice, the parameter `epsilon[18]` is included, which had the second highest  $\hat{R}$  of 1.008.



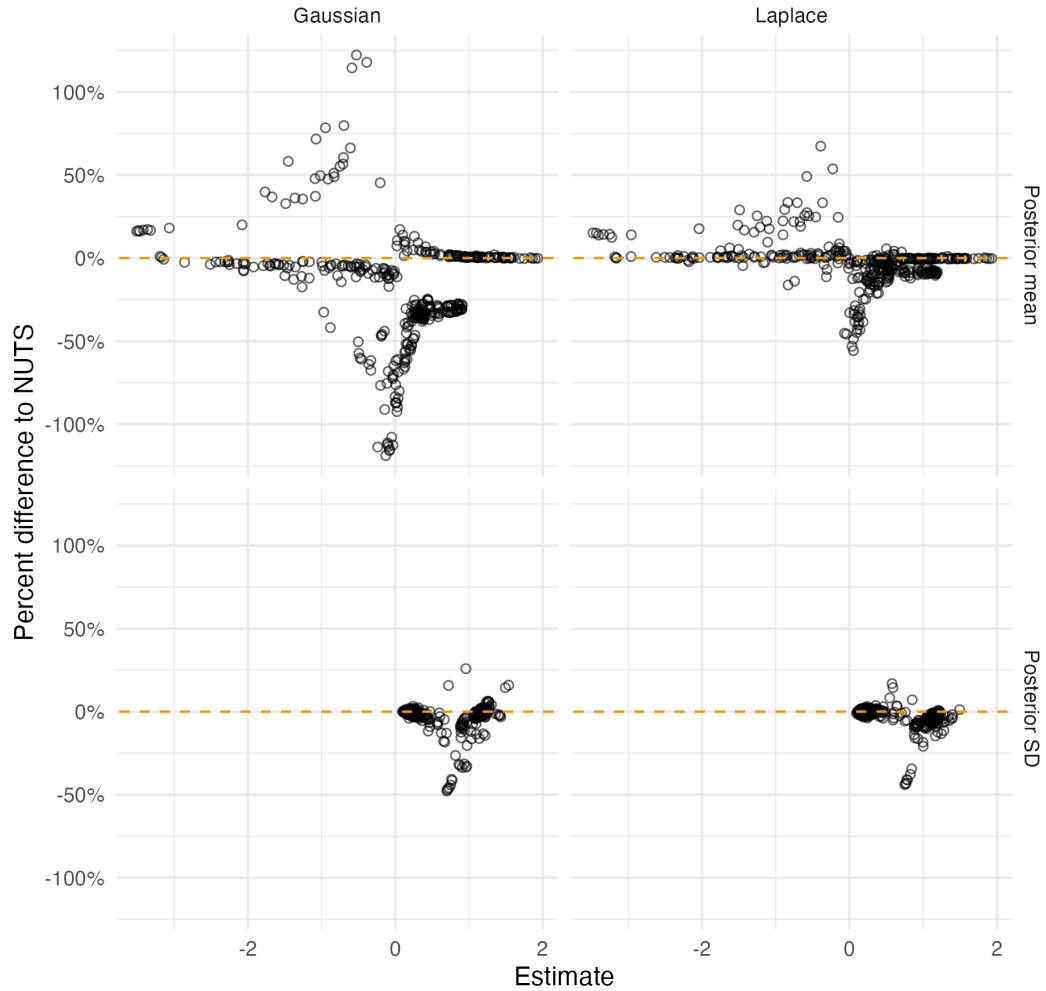
**Figure C.3:** Traceplots for the parameters with the lowest ESS and highest potential scale reduction factor for the Loa loa ELGM example.

## C.2 Loa loa example

### C.2.1 NUTS convergence and suitability

### C.2.2 Inference comparison

### C. Fast approximate Bayesian inference



**Figure C.4:** Relative difference between the Gaussian and Laplace marginal posterior means and standard deviations to NUTS results at each  $u(s_i), v(s_i) : i \in [190]$ . Absolute differences are in Figure 6.14.

### C.3 AGHQ with Laplace marginals algorithm

This section provides the INLA-like algorithm for AGHQ with Laplace marginals used in this thesis. The algorithm for AGHQ with Gaussian marginals used in this thesis is as given in Stringer et al. (2022), and implemented in the `aghq` package.

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

### C. Fast approximate Bayesian inference

approximation

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y}), \quad (\text{C.1})$$

$$\hat{\mathbf{H}} = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \log \tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad (\text{C.2})$$

$$\hat{\mathbf{H}}^{-1} = \hat{\mathbf{L}} \hat{\mathbf{L}}^\top, \quad (\text{C.3})$$

$$\tilde{p}_{\text{LA}}(\boldsymbol{\theta}, \mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{\tilde{p}_{\text{G}}(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\hat{\mathbf{x}}(\boldsymbol{\theta})}, \quad (\text{C.4})$$

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

$$\hat{\mathbf{x}}(\boldsymbol{\theta}) = \arg \max_{\mathbf{x}} \log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}), \quad (\text{C.5})$$

$$\hat{\mathbf{H}}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}^\top} \log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})|_{\mathbf{x}=\hat{\mathbf{x}}(\boldsymbol{\theta})}. \quad (\text{C.6})$$

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

$$\tilde{p}_{\text{AQ}}(\mathbf{y}) = \sum_{\mathbf{u} \in \mathcal{Q}(m, k)} \tilde{p}_{\text{LA}}(\boldsymbol{\theta}(\mathbf{u}), \mathbf{y}) \omega(\mathbf{u}). \quad (\text{C.7})$$

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

$$\hat{\mathbf{x}}_{-i}(x_i, \boldsymbol{\theta}) = \arg \max_{\mathbf{x}_{-i}} \log p(\mathbf{y}, x_i, \mathbf{x}_{-i}, \boldsymbol{\theta}), \quad (\text{C.8})$$

$$\hat{\mathbf{H}}_{-i,-i}(x_i, \boldsymbol{\theta}) = -\frac{\partial^2}{\partial \mathbf{x}_{-i} \partial \mathbf{x}_{-i}^\top} \log p(\mathbf{y}, x_i, \mathbf{x}_{-i}, \boldsymbol{\theta})|_{\mathbf{x}_{-i}=\hat{\mathbf{x}}_{-i}(x_i, \boldsymbol{\theta})}, \quad (\text{C.9})$$

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

### C. Fast approximate Bayesian inference

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

$$p_{\text{AQ}}(x_i | \mathbf{y}) = \frac{\tilde{p}_{\text{LA}}(x_i, \mathbf{y})}{\tilde{p}_{\text{AQ}}(\mathbf{y})}, \quad (\text{C.10})$$

where

$$\tilde{p}_{\text{LA}}(x_i, \mathbf{y}) = \sum_{\mathbf{u} \in \mathcal{Q}(m,k)} \tilde{p}_{\text{LA}}(x_i, \boldsymbol{\theta}(\mathbf{u}), \mathbf{y}) \omega(\mathbf{u}). \quad (\text{C.11})$$

and

$$\tilde{p}_{\text{LA}}(x_i, \boldsymbol{\theta}, \mathbf{y}) = \frac{p(x_i, \mathbf{x}_{-i}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_{\text{G}}(\mathbf{x}_{-i} | x_i, \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}_{-i}=\hat{\mathbf{x}}_{-i}(x_i, \boldsymbol{\theta})}. \quad (\text{C.12})$$

Equation (C.10) can be calculated using the estimate of the evidence given in Equation (C.7), but it is more numerically accurate, and requires little extra computation, to use the estimate

$$\tilde{p}_{\text{AQ}}(\mathbf{y}) = \sum_{\mathbf{v} \in \mathcal{Q}(1,l)} \tilde{p}_{\text{LA}}(x_i(\mathbf{v}), \mathbf{y}) \omega(\mathbf{v}) \quad (\text{C.13})$$

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

## C.4 Simplified Naomi model description

This section describes the simplified version of the Naomi model (Jeffrey W 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.4.1 gives the process specifications, giving the terms in each structured additive predictor, along with their distributions.
2. Section C.4.2 gives additional details about the likelihood terms not provided in Section 6.3.
3. Section C.4.3 gives identifiability constraints used in circumstances where incomplete data is available for the country.
4. Section C.4.4 provides details of the TMB implementation.

### C. Fast approximate Bayesian inference

#### C.4.1 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.4.1.1	HIV prevalence	$22 + 5n$	9
Section C.4.1.2	ART coverage	$25 + 5n$	9
Section C.4.1.3	HIV incidence rate	$2 + n$	3
Section C.4.1.4	ANC testing	$2 + 2n$	2
Section C.4.1.5	ART attendance	$n$	1
	Total	$51 + 14n$	24

##### C.4.1.1 HIV prevalence

HIV prevalence  $\rho_{x,s,a} \in [0, 1]$  was modelled on the logit scale using the structured additive predictor

$$\text{logit}(\rho_{x,s,a}) = \beta_0^\rho + \beta_S^{\rho,s=M} + \mathbf{u}_a^\rho + \mathbf{u}_a^{\rho,s=M} + \mathbf{u}_x^\rho + \mathbf{u}_x^{\rho,s=M} + \mathbf{u}_x^{\rho,a<15} + \boldsymbol{\eta}_{R_x,s,a}^\rho. \quad (\text{C.14})$$

Table C.2 provides a description of the terms included in Equation (C.14). Independent half-normal prior distributions were chosen for the five standard deviation terms

$$\{\sigma_A^\rho, \sigma_{AS}^\rho, \sigma_X^\rho, \sigma_{XS}^\rho, \sigma_{XA}^\rho\} \sim \mathcal{N}^+(0, 2.5), \quad (\text{C.15})$$

independent uniform prior distributions for the two AR1 correlation parameters

$$\{\phi_A^\rho, \phi_{AS}^\rho\} \sim \mathcal{U}(-1, 1), \quad (\text{C.16})$$

and independent beta prior distributions for the two BYM2 proportion parameters

$$\{\phi_X^\rho, \phi_{XS}^\rho\} \sim \text{Beta}(0.5, 0.5). \quad (\text{C.17})$$

**Table C.2:** Each term in Equation (C.14) together with, where applicable, its prior distribution and a written description of its role.

Term	Distribution	Description
$\beta_0^\rho$	$\mathcal{N}(0, 5)$	Intercept

### C. Fast approximate Bayesian inference

Term	Distribution	Description
$\beta_s^{\rho,s=M}$	$\mathcal{N}(0, 5)$	The difference in logit prevalence for men compared to women
$\mathbf{u}_a^\rho$	AR1( $\sigma_A^\rho, \phi_A^\rho$ )	Age random effects for women
$\mathbf{u}_a^{\rho,s=M}$	AR1( $\sigma_{AS}^\rho, \phi_{AS}^\rho$ )	Age random effects for the difference in logit prevalence for men compared to women age $a$
$\mathbf{u}_x^\rho$	BYM2( $\sigma_X^\rho, \phi_X^\rho$ )	Spatial random effects for women
$\mathbf{u}_x^{\rho,s=M}$	BYM2( $\sigma_{XS}^\rho, \phi_{XS}^\rho$ )	Spatial random effects for the difference in logit prevalence for men compared to women in district $x$
$\mathbf{u}_x^{\rho,a<15}$	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 $R_x$

#### C.4.1.2 ART coverage

ART coverage  $\alpha_{x,s,a} \in [0, 1]$  was modelled on the logit scale using the structured additive predictor

$$\text{logit}(\alpha_{x,s,a}) = \beta_0^\alpha + \beta_S^{\alpha,s=M} + \mathbf{u}_a^\alpha + \mathbf{u}_a^{\alpha,s=M} + \mathbf{u}_x^\alpha + \mathbf{u}_x^{\alpha,s=M} + \mathbf{u}_x^{\alpha,a<15} + \boldsymbol{\eta}_{R_x,s,a}^\alpha \quad (\text{C.18})$$

with terms and prior distributions analogous to the HIV prevalence process model in Section C.4.1.1 above.

#### C.4.1.3 HIV incidence rate

HIV incidence rate  $\lambda_{x,s,a} > 0$  was modelled on the log scale using the structured additive predictor

$$\log(\lambda_{x,s,a}) = \beta_0^\lambda + \beta_S^{\lambda,s=M} + \log(\rho_x^{15-49}) + \log(1 - \omega \cdot \alpha_x^{15-49}) + \mathbf{u}_x^\lambda + \boldsymbol{\eta}_{R_x,s,a}^\lambda. \quad (\text{C.19})$$

Table C.3 provides a description of the terms included in Equation (C.19).

### C. Fast approximate Bayesian inference

**Table C.3:** Each term in Equation (C.19) together with, where applicable, its prior distribution and a written description of its role.

Term	Distribution	Description
$\beta_0^\lambda$	$\mathcal{N}(0, 5)$	Intercept term proportional to the average HIV transmission rate for untreated HIV positive adults
$\beta_S^{\lambda, s=M}$	$\mathcal{N}(0, 5)$	The log incidence rate ratio for men compared to women
$\rho_x^{15-49}$	—	The HIV prevalence among adults 15-49 in district $x$ calculated by aggregating age-specific HIV prevalences
$\alpha_x^{15-49}$	—	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

$$\kappa_{x,s,a} = 1 - \exp\left(-\lambda_{x,s,a} \cdot \frac{1 - \rho_{x,s,a}}{\rho_{x,s,a}} \cdot (\Omega_T - \beta_T) - \beta_T\right), \quad (\text{C.20})$$

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  $\Omega_T$  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  $\beta_T = 0$ .

#### C.4.1.4 ANC testing

HIV prevalence  $\rho_{x,a}^{\text{ANC}}$  and ART coverage  $\alpha_{x,a}^{\text{ANC}}$  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

$$\text{logit}(\rho_{x,a}^{\text{ANC}}) = \text{logit}(\rho_{x,F,a}) + \beta^{\rho^{\text{ANC}}} + \mathbf{u}_x^{\rho^{\text{ANC}}} + \boldsymbol{\eta}_{R_x,a}^{\rho^{\text{ANC}}}, \quad (\text{C.21})$$

$$\text{logit}(\alpha_{x,a}^{\text{ANC}}) = \text{logit}(\alpha_{x,F,a}) + \beta^{\alpha^{\text{ANC}}} + \mathbf{u}_x^{\alpha^{\text{ANC}}} + \boldsymbol{\eta}_{R_x,a}^{\alpha^{\text{ANC}}}. \quad (\text{C.22})$$

### C. Fast approximate Bayesian inference

Table C.4 provides a description of the terms included in Equation (C.21) and Equation (C.22).

**Table C.4:** Each term in Equations (C.21) and (C.22) together with (where applicable) its prior distribution and a written description of its role. The notation  $\theta$  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
$\mathbf{u}_x^{\theta^{\text{ANC}}}$	$\mathcal{N}(0, \sigma_X^{\theta^{\text{ANC}}})$	IID district random effects with $\sigma_X^{\theta^{\text{ANC}}} \sim \mathcal{N}^+(0, 1)$
$\boldsymbol{\eta}_{R_x,a}^{\theta^{\text{ANC}}}$	—	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

$$\log(\Psi_{x,a}) = \log(N_{x,F,a}) + \psi_{R_x,a} + \beta^\psi + \mathbf{u}_x^\psi, \quad (\text{C.23})$$

where  $N_{x,F,a}$  are the female population sizes,  $\psi_{R_x,a}$  are fixed age-sex fertility ratios in Spectrum region  $R_x$ ,  $\beta^\psi$  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.

#### C.4.1.5 ART attendance

Let  $\gamma_{x,x'} \in [0, 1]$  be the probability that a person on ART residing in district  $x$  receives ART in district  $x'$ . 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

$$\tilde{\gamma}_{x,x'} = \log \left( \frac{\gamma_{x,x'}}{1 - \gamma_{x,x'}} \right) = \tilde{\gamma}_0 + \mathbf{u}_x^{\tilde{\gamma}}. \quad (\text{C.24})$$

### C. Fast approximate Bayesian inference

Table C.5 provides a description of the terms included in Equation (C.24). Fixing  $\tilde{\gamma}_{x,x} = 0$  then the multinomial probabilities may be recovered using the softmax

$$\gamma_{x,x'} = \frac{\exp(\tilde{\gamma}_{x,x'})}{\sum_{x^* \in \{x, \text{ne}(x)\}} \exp(\tilde{\gamma}_{x,x^*})}. \quad (\text{C.25})$$

**Table C.5:** Each term in Equation (C.24) together with, where applicable, its prior distribution and a written description of its role. As no terms include  $x'$ ,  $\gamma_{x,x'}$  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.4.2 Additional likelihood specification

Though Section 6.3 provides a complete description of Naomi’s likelihood specification, any additional useful details are provided here.

### C.4.2.1 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

$$\log p(y) = \log \Gamma(m + 1) - \log \Gamma(y + 1) \quad (\text{C.26})$$

$$- \log \Gamma(m - y + 1) + y \log p + (m - y) \log(1 - p), \quad (\text{C.27})$$

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

## C.4.3 Identifiability constraints

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

### C. Fast approximate Bayesian inference

1. If survey data on HIV prevalence or ART coverage by age and sex are not available then  $\mathbf{u}_a^\theta = \mathbf{0}$  and  $\mathbf{u}_{a,s=M}^\theta = \mathbf{0}$ . 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  $|\{0-4, \dots, 50-54\}| = 13$  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

$$\text{logit}(\alpha_{x,s,a}) = \mathbf{u}_x^\alpha + \boldsymbol{\eta}_{R_x,s,a}^\alpha. \quad (\text{C.28})$$

3. If survey data on recent HIV infection are not included in the model, then  $\beta_0^\lambda = \beta_S^{\lambda,s=M} = \mathbf{0}$  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.4.4 Implementation

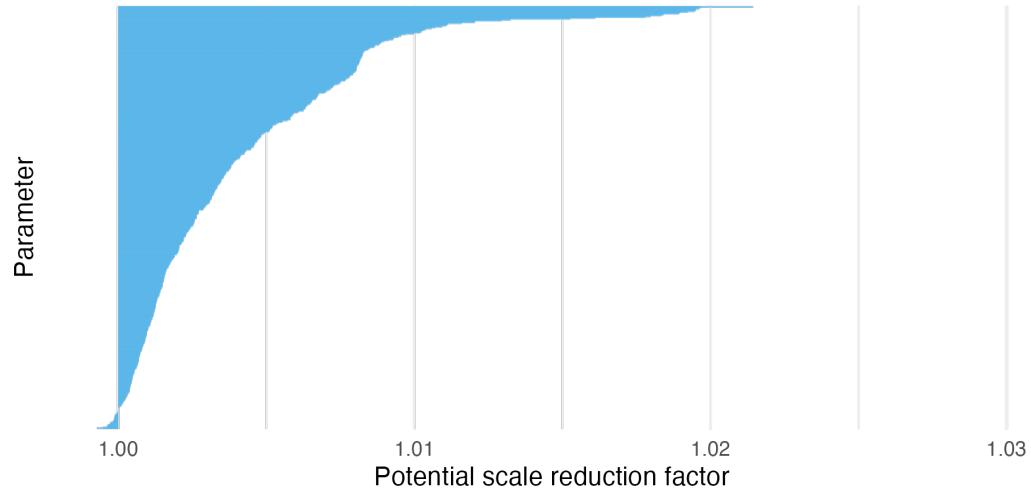
The TMB C++ code for the negative log-posterior of the simplified Naomi model is available from <https://github.com/athowes/naomi-aghq>. For ease of understanding, Table C.6 provides correspondence between the mathematical notation used in Section C.4 and the variable names used in the TMB code, for all hyperparameters and latent field parameters. For further reference on the TMB software see Kristensen (2021).

### C. Fast approximate Bayesian inference

**Table C.6:** Correspondence between the variable name used in the Naomi TMB template and the mathematical notation used in Appendix C.4. The parameter type, either a hyperparameter or element of the latent field, is also given. All of the parameters are defined on the real-scale in some dimension. In the final three columns ( $\rho$ ,  $\alpha$ , and  $\lambda$ ) indication is given as to which component of the model the parameter is primarily used in.

Variable name	Notation	Type	Domain	$\rho$	$\alpha$	$\lambda$
logit_phi_rho_x	$\text{logit}(\phi_X^\rho)$	Hyper	$\mathbb{R}$	Yes		
log_sigma_rho_x	$\log(\sigma_X^\rho)$	Hyper	$\mathbb{R}$	Yes		
logit_phi_rho_xs	$\text{logit}(\phi_{XS}^\rho)$	Hyper	$\mathbb{R}$	Yes		
log_sigma_rho_xs	$\log(\sigma_{XS}^\rho)$	Hyper	$\mathbb{R}$	Yes		
logit_phi_rho_a	$\text{logit}(\phi_A^\rho)$	Hyper	$\mathbb{R}$	Yes		
log_sigma_rho_a	$\log(\sigma_A^\rho)$	Hyper	$\mathbb{R}$	Yes		
logit_phi_rho_as	$\text{logit}(\phi_{AS}^\rho)$	Hyper	$\mathbb{R}$	Yes		
log_sigma_rho_as	$\log(\sigma_{AS}^\rho)$	Hyper	$\mathbb{R}$	Yes		
log_sigma_rho_xa	$\log(\sigma_{XA}^\rho)$	Hyper	$\mathbb{R}$	Yes		
logit_phi_alpha_x	$\text{logit}(\phi_X^\alpha)$	Hyper	$\mathbb{R}$		Yes	
log_sigma_alpha_x	$\log(\sigma_X^\alpha)$	Hyper	$\mathbb{R}$		Yes	
logit_phi_alpha_xs	$\text{logit}(\phi_{XS}^\alpha)$	Hyper	$\mathbb{R}$		Yes	
log_sigma_alpha_xs	$\log(\sigma_{XS}^\alpha)$	Hyper	$\mathbb{R}$		Yes	
logit_phi_alpha_a	$\text{logit}(\phi_A^\alpha)$	Hyper	$\mathbb{R}$		Yes	
log_sigma_alpha_a	$\log(\sigma_A^\alpha)$	Hyper	$\mathbb{R}$		Yes	
logit_phi_alpha_as	$\text{logit}(\phi_{AS}^\alpha)$	Hyper	$\mathbb{R}$		Yes	
log_sigma_alpha_as	$\log(\sigma_{AS}^\alpha)$	Hyper	$\mathbb{R}$		Yes	
log_sigma_alpha_xa	$\log(\sigma_{XA}^\alpha)$	Hyper	$\mathbb{R}$		Yes	
OmegaT_raw	$\Omega_T$	Hyper	$\mathbb{R}$			Yes
log_betaT	$\log(\beta_T)$	Hyper	$\mathbb{R}$			Yes
log_sigma_lambda_x	$\log(\sigma^\lambda)$	Hyper	$\mathbb{R}$			Yes
log_sigma_ancrho_x	$\log(\sigma_X^{\rho_{\text{ANC}}})$	Hyper	$\mathbb{R}$		Yes	
log_sigma_ancalpha_x	$\log(\sigma_X^{\alpha_{\text{ANC}}})$	Hyper	$\mathbb{R}$		Yes	
log_sigma_or_gamma	$\log(\sigma_X^{\tilde{\gamma}})$	Hyper	$\mathbb{R}$			
beta_rho	$(\beta_0^\rho, \beta_s^{\rho, s=M})$	Latent	$\mathbb{R}^2$	Yes		
beta_alpha	$(\beta_0^\alpha, \beta_S^{\alpha, s=M})$	Latent	$\mathbb{R}^2$		Yes	
beta_lambda	$(\beta_0^\lambda, \beta_S^{\lambda, s=M})$	Latent	$\mathbb{R}^2$			Yes
beta_anc_rho	$\beta^{\rho_{\text{ANC}}}$	Latent	$\mathbb{R}$		Yes	
beta_anc_alpha	$\beta^{\alpha_{\text{ANC}}}$	Latent	$\mathbb{R}$		Yes	
u_rho_x	$\mathbf{w}_x^\rho$	Latent	$\mathbb{R}^n$	Yes		
us_rho_x	$\mathbf{v}_x^\rho$	Latent	$\mathbb{R}^n$	Yes		
u_rho_xs	$\mathbf{w}_x^{\rho, s=M}$	Latent	$\mathbb{R}^n$	Yes		
us_rho_xs	$\mathbf{v}_x^{\rho, s=M}$	Latent	$\mathbb{R}^n$	Yes		
u_rho_a	$\mathbf{u}_a^\rho$	Latent	$\mathbb{R}^{10}$	Yes		
u_rho_as	$\mathbf{u}_a^{\rho, s=M}$	Latent	$\mathbb{R}^{10}$	Yes		
u_rho_xa	$\mathbf{u}_x^{\rho, a < 15}$	Latent	$\mathbb{R}^n$	Yes		
u_alpha_x	$\mathbf{w}_x^\alpha$	Latent	$\mathbb{R}^n$			Yes

### C. Fast approximate Bayesian inference



**Figure C.5:** For NUTS run on the Naomi ELGM, the maximum potential scale reduction factor was 1.021, below the value of 1.05 typically used as a cutoff for acceptable chain mixing, indicating that the results are acceptable to use. Additionally, the vast majority (93.7%) of  $\hat{R}$  values were less than 1.1.

Variable name	Notation	Type	Domain	$\rho$	$\alpha$	$\lambda$
us_alpha_x	$\mathbf{v}_x^\alpha$	Latent	$\mathbb{R}^n$		Yes	
u_alpha_xs	$\mathbf{w}_x^{\alpha,s=M}$	Latent	$\mathbb{R}^n$		Yes	
us_alpha_xs	$\mathbf{v}_x^{\alpha,s=M}$	Latent	$\mathbb{R}^n$		Yes	
u_alpha_a	$\mathbf{u}_a^\alpha$	Latent	$\mathbb{R}^{13}$		Yes	
u_alpha_as	$\mathbf{u}_a^{\alpha,s=M}$	Latent	$\mathbb{R}^{10}$		Yes	
u_alpha_xa	$\mathbf{u}_x^{\alpha,a<15}$	Latent	$\mathbb{R}^n$		Yes	
ui_lambda_x	$\mathbf{u}_x^\lambda$	Latent	$\mathbb{R}^n$			Yes
ui_anc_rho_x	$\mathbf{u}_x^{\rho_{ANC}}$	Latent	$\mathbb{R}^n$		Yes	
ui_anc_alpha_x	$\mathbf{u}_x^{\alpha_{ANC}}$	Latent	$\mathbb{R}^n$		Yes	
log_or_gamma	$\mathbf{u}_x^\gamma$	Latent	$\mathbb{R}^n$			

## C.5 NUTS convergence and suitability

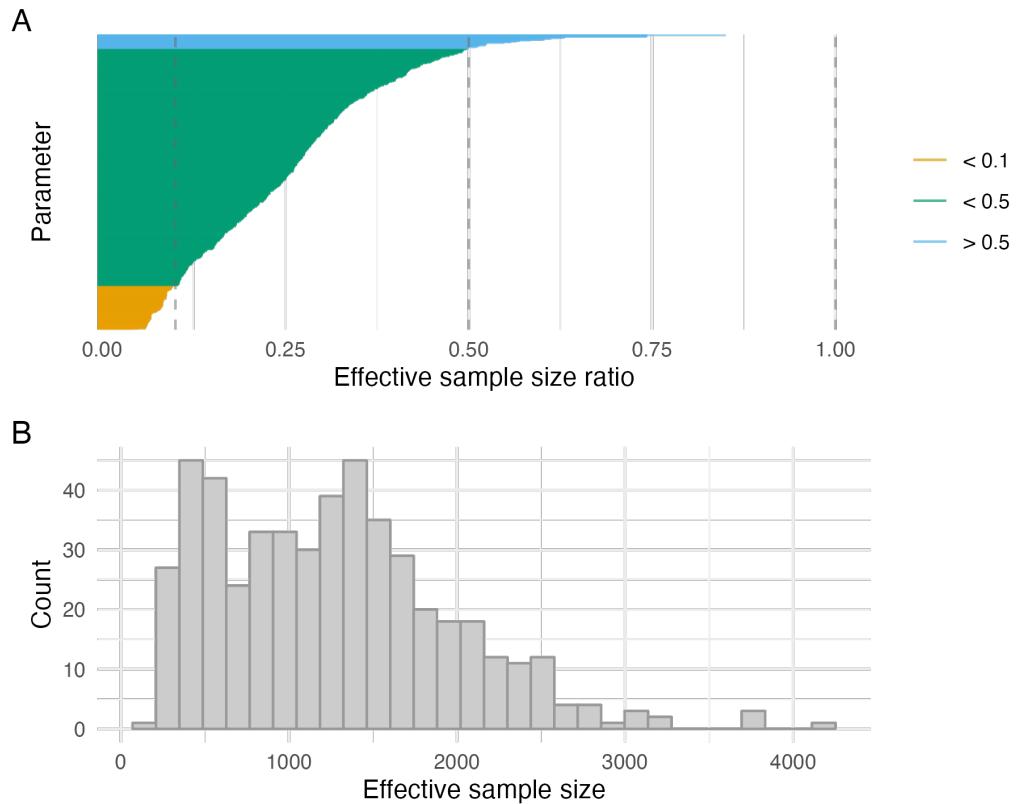
## C.6 Use of PCA-AGHQ

## C.7 Inference comparison

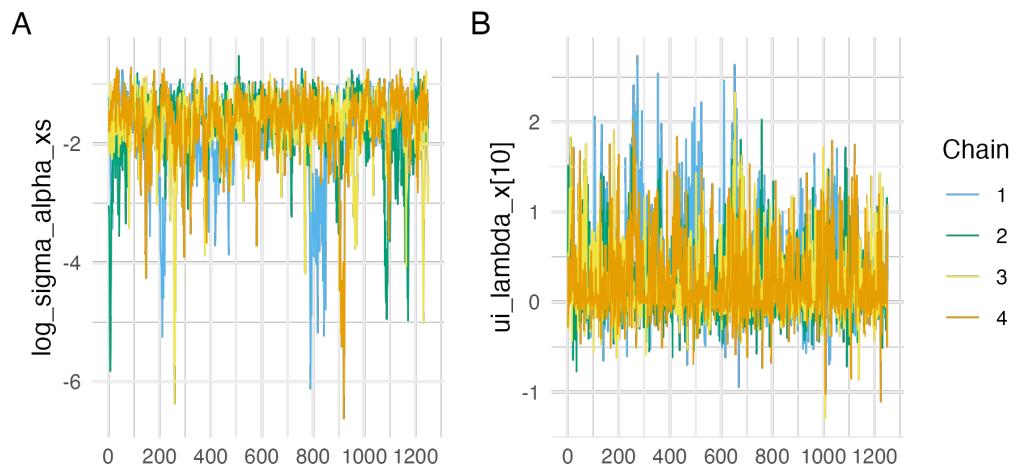
### C.7.1 Point estimates

### C.7.2 Distributional quantities

C. Fast approximate Bayesian inference

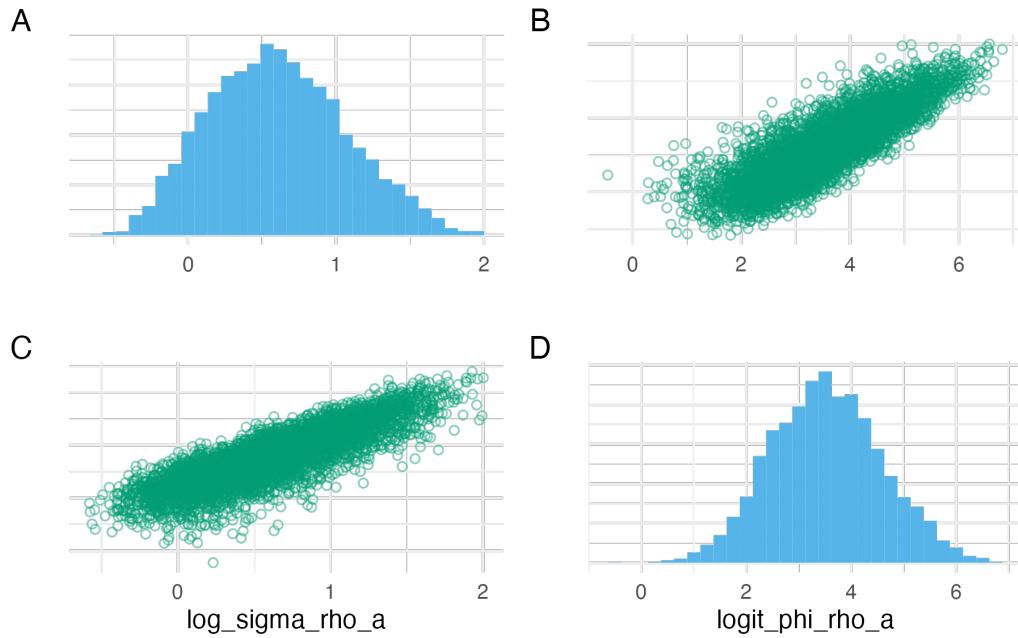


**Figure C.6:** The efficiency of the NUTS, as measured by the ratio of effective sample size to total number of iterations run, was low for most parameters (Panel A). As a result, the number of iterations required for the effective number of samples (mean 1265) to be satisfactory was high (Panel B).

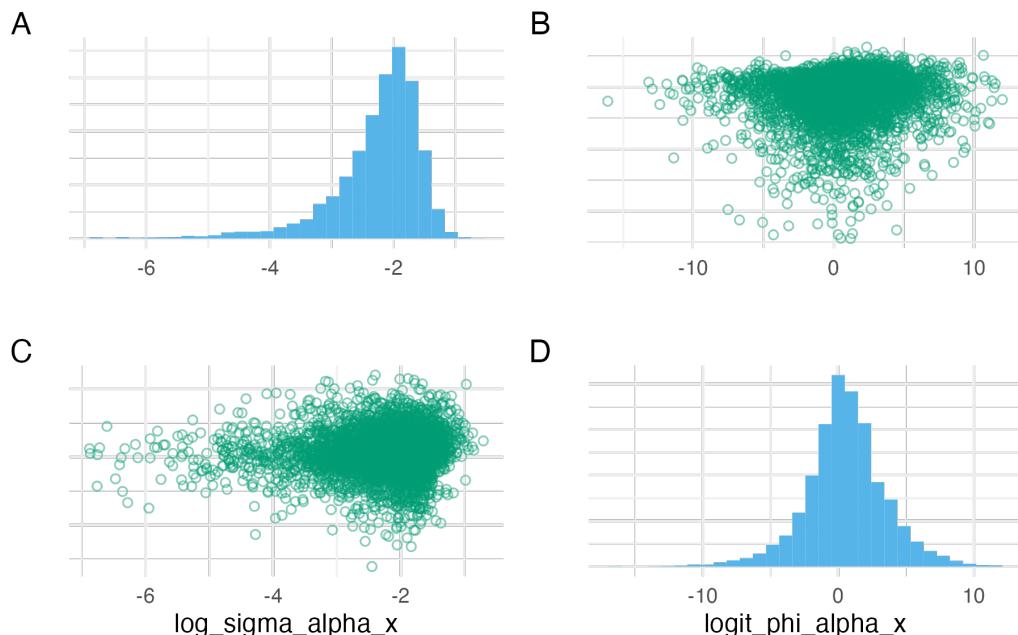


**Figure C.7:** Traceplots for the parameter with the lowest ESS which was `log_sigma_alpha_xs` (an ESS of 208, Panel A) and highest potential scale reduction factor which was `ui_lambda_x[10]` (an  $\hat{R}$  of 1.021, Panel B).

C. Fast approximate Bayesian inference

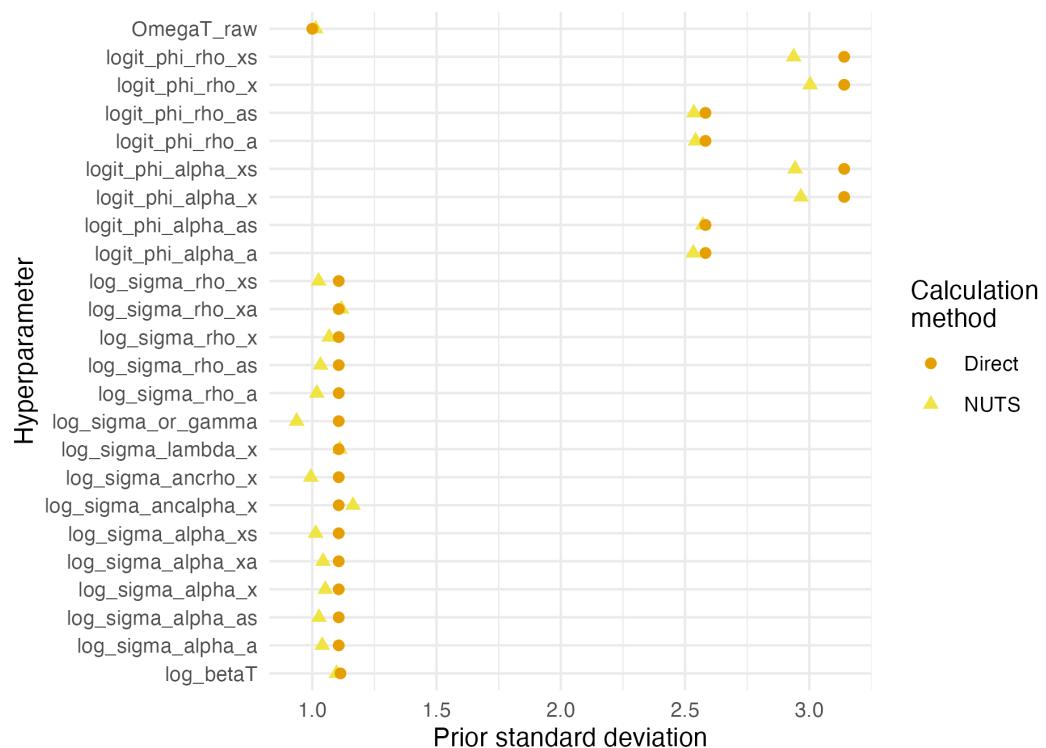


**Figure C.8:** Pairs plots for the parameters  $\log(\sigma_A^\rho)$  and  $\log(\phi_A^\rho)$ , or  $\log\_sigma\_rho\_a$  and  $\logit\_phi\_rho\_a$  as implemented in code. These parameters are the log standard deviation and logit lag-one correlation parameter of an AR1 process. In the posterior distribution obtained with NUTS, they have a high degree of correlation.



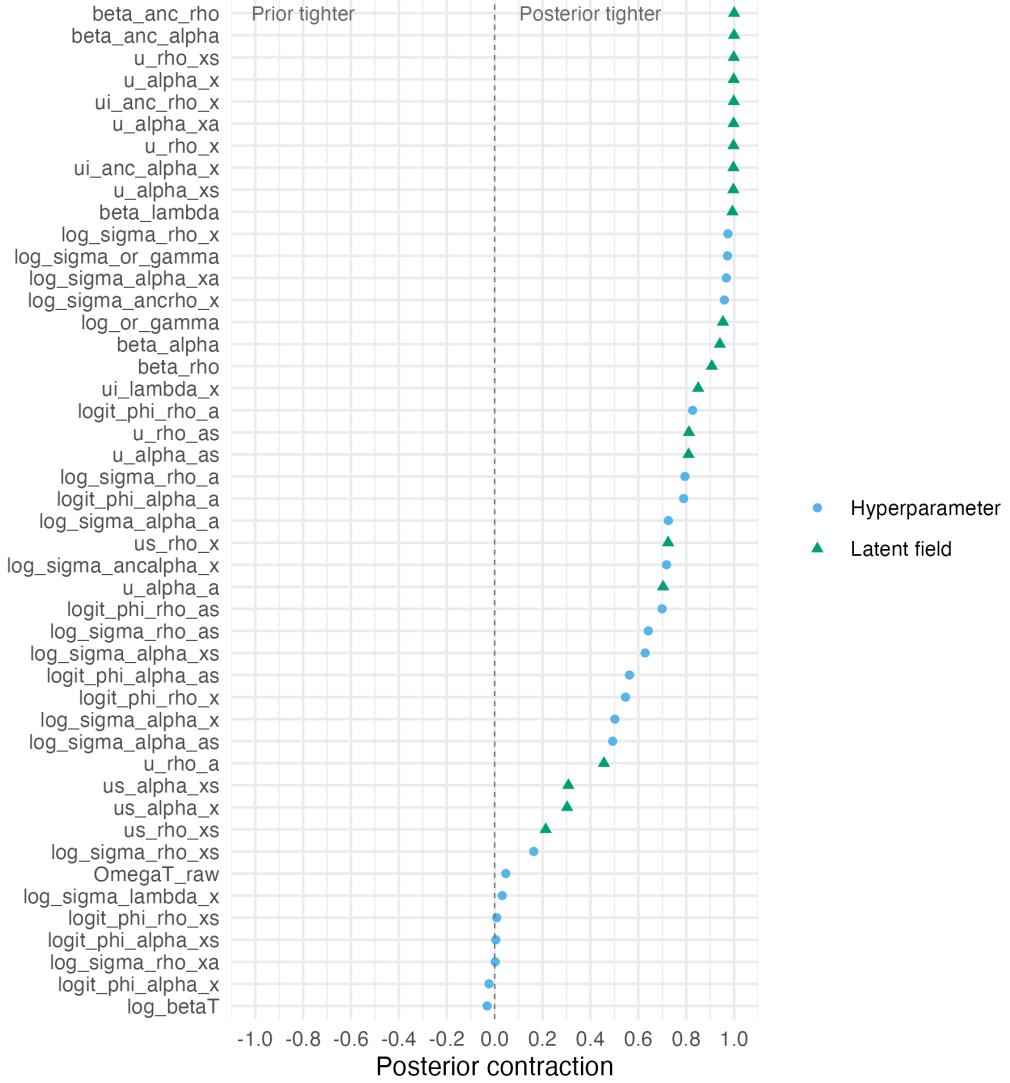
**Figure C.9:** Pairs plots for the parameters  $\log(\sigma_X^\alpha)$  and  $\log(\phi_X^\alpha)$ , or  $\log\_sigma\_alpha\_x$  and  $\logit\_phi\_alpha\_x$  as implemented in code. These parameters are the log standard deviation and logit BYM2 proportion parameter of a BYM2 process. In the posterior distribution obtained with NUTS, they are close to uncorrelated.

### C. Fast approximate Bayesian inference



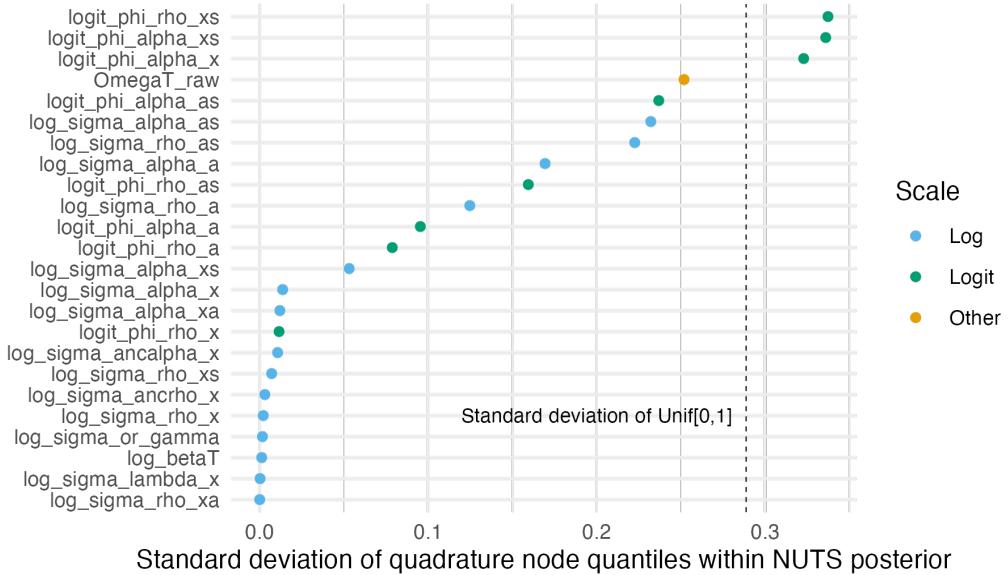
**Figure C.10:** Prior standard deviations were calculated by using NUTS to simulate from the prior distribution. This approach is more convenient than simulating directly from the model, but can lead to inaccuracies.

### C. Fast approximate Bayesian inference

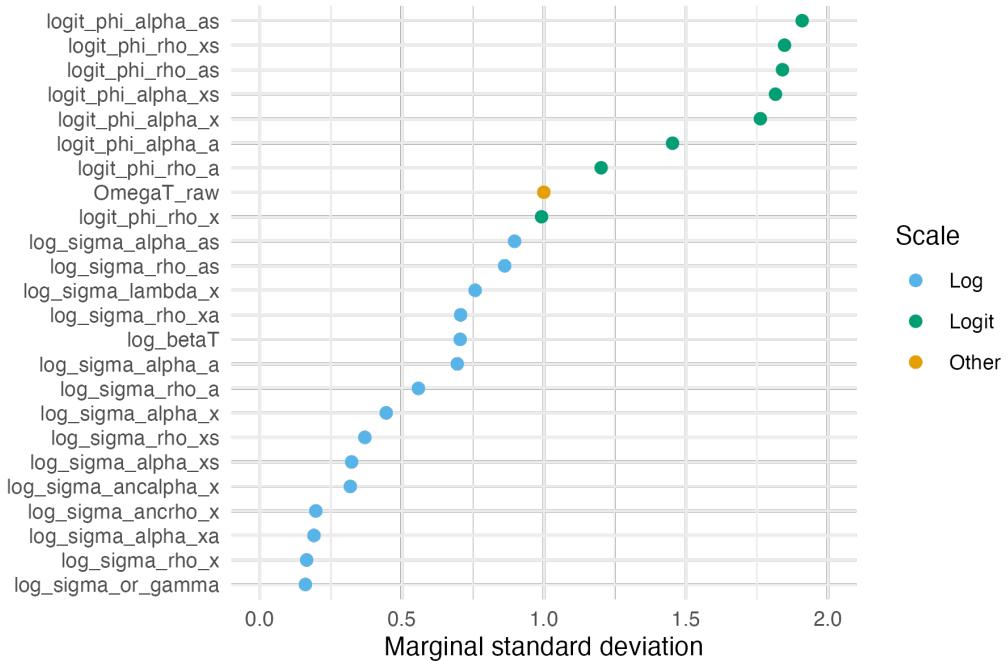


**Figure C.11:** The posterior contraction for each parameter in the model. Values are averaged for parameters of length greater than one. The posterior contraction is zero when the prior distribution and posterior distribution have the same standard deviation. This could indicate that the data is not informative about the parameter. The closer the posterior contraction is to one, the more than the marginal posterior distribution has concentrated about a single point.

### C. Fast approximate Bayesian inference

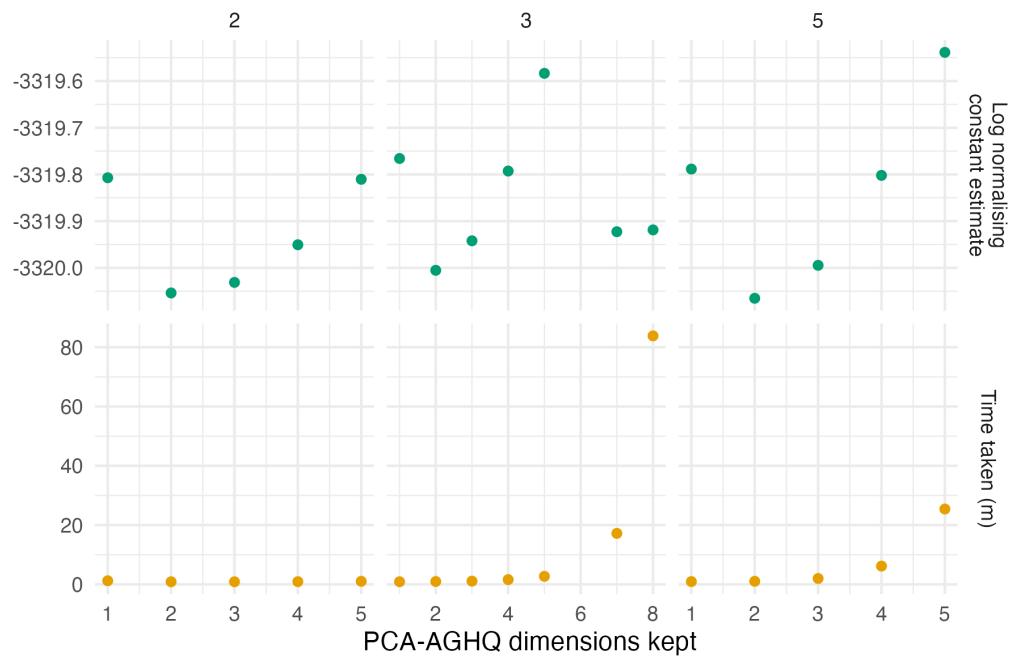


**Figure C.12:** The standard deviation of the quadrature nodes can be used as a measure of coverage of the posterior marginal distribution. Nodes spaced evenly within the marginal distribution would be expected to uniformly distributed quantile, corresponding to a standard deviation of 0.2871, shown as a dashed line.



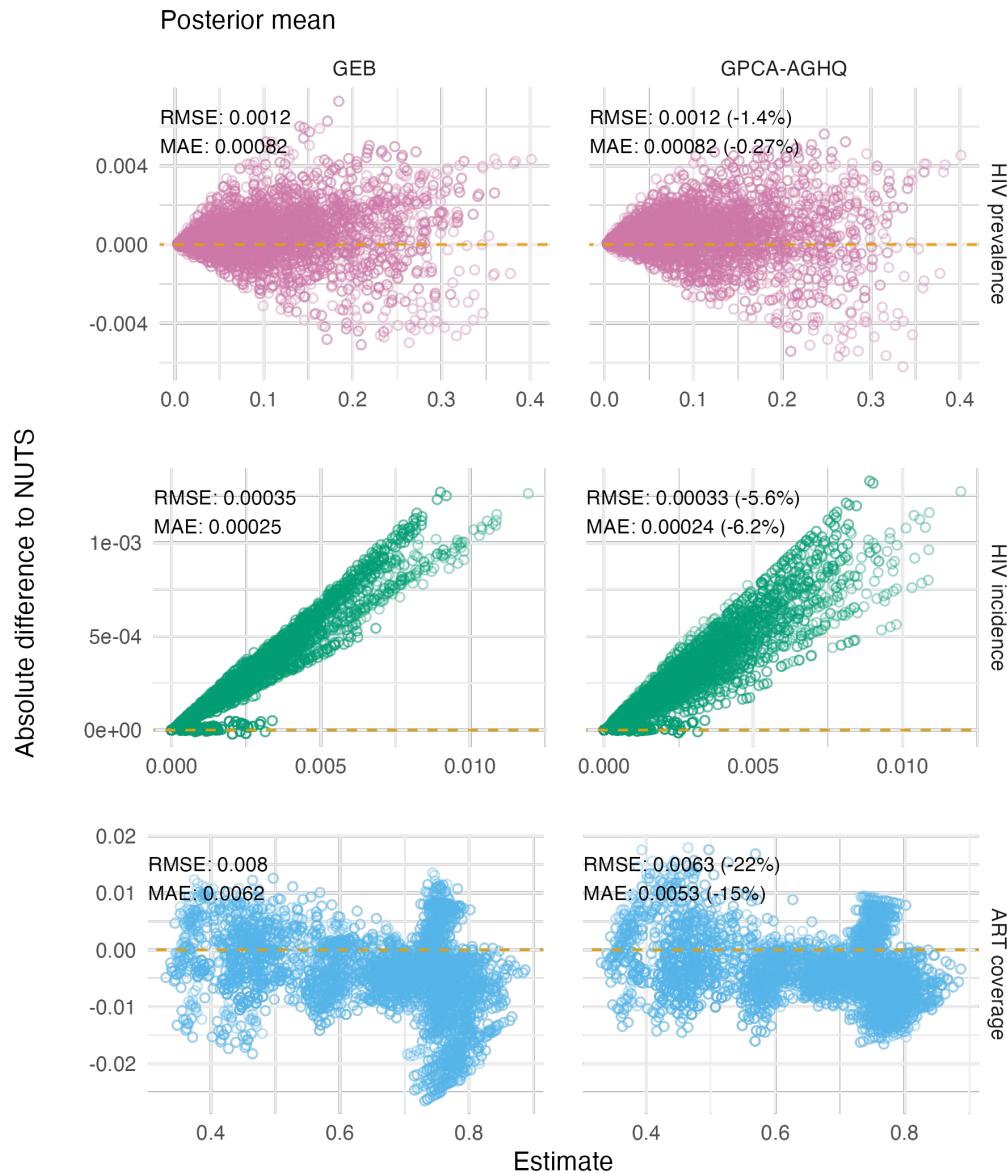
**Figure C.13:** The estimated posterior marginal standard deviation of each hyperparameter varied substantially based on its scale, either logarithmic or logistic.

C. Fast approximate Bayesian inference



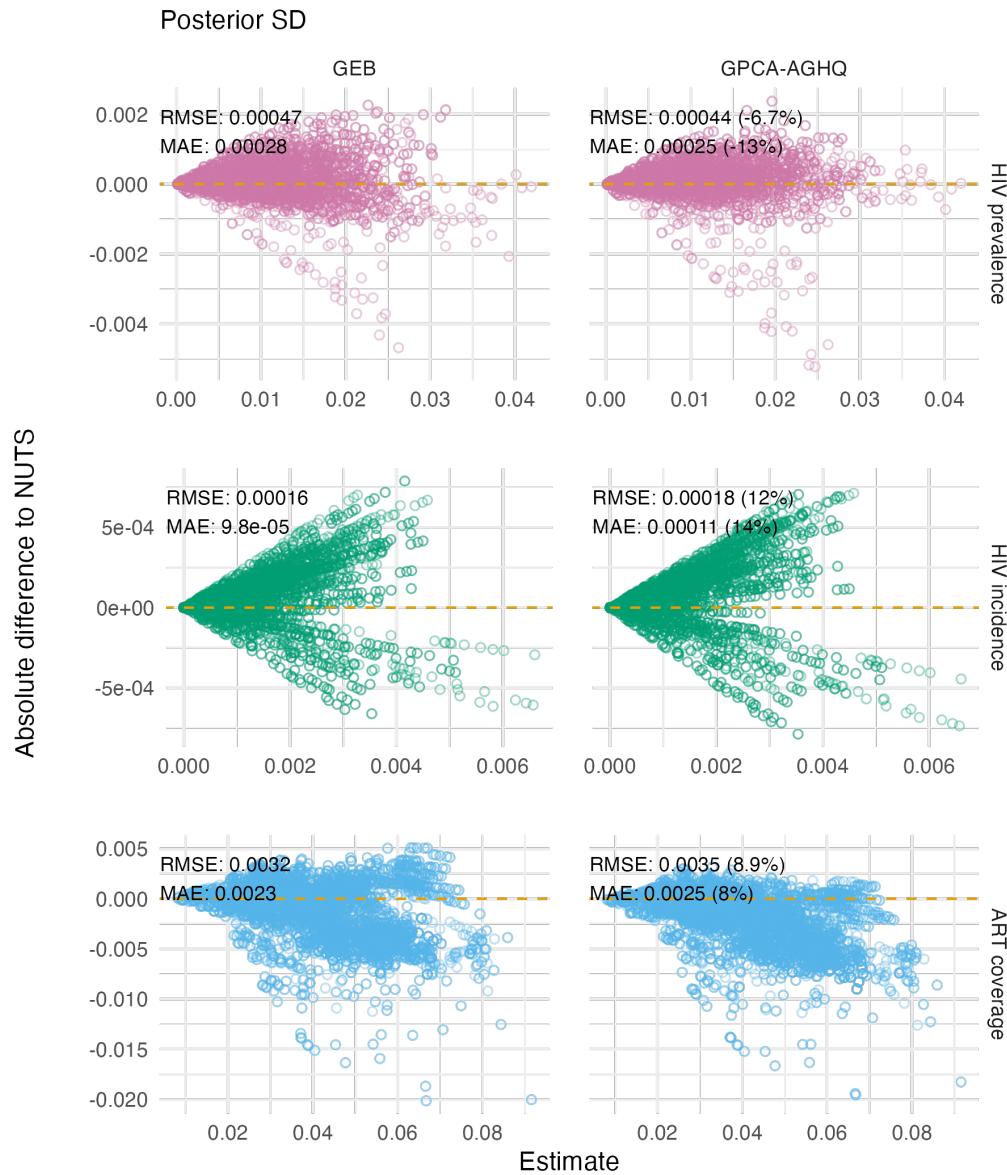
**Figure C.14:** The logarithm of the normalising constant estimated using PCA-AGHQ and a range of possible values of  $k = 2, 3, 5$  and  $s \leq 8$ . Using this range of settings, there was not convergence of the logarithm of the normalising constant estimate. The time taken by GPCA-AGHQ increases exponentially with number of PCA-AGHQ dimensions kept.

### C. Fast approximate Bayesian inference



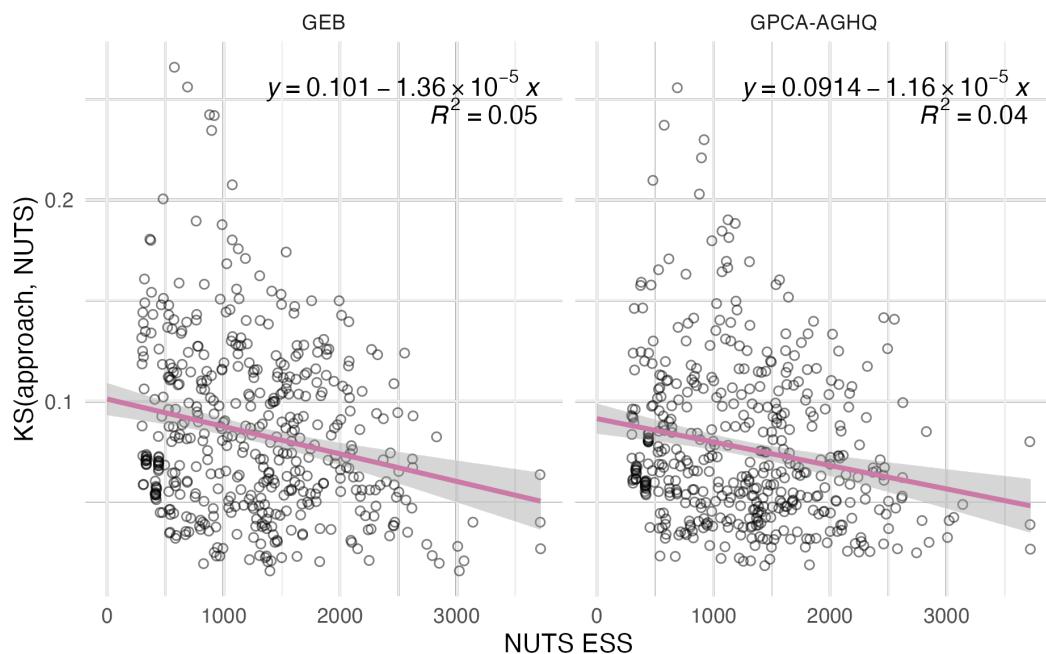
**Figure C.15:** Differences in Naomi model output posterior means as estimated by GEB and GPCA-AGHQ compared to NUTS. Each point is an estimate of the indicator for a particular strata. In all cases, error is reduced by GPCA-AGHQ, most of all for ART coverage.

C. Fast approximate Bayesian inference



**Figure C.16:** Differences in Naomi model output posterior standard deviations as estimated by GEB and GPCA-AGHQ compared to NUTS. Each point is an estimate of the indicator for a particular strata. Error is increased by GPCA-AGHQ for HIV prevalence and HIV incidence, and reduced for ART coverage.

*C. Fast approximate Bayesian inference*



**Figure C.17:** The Kolmogorov-Smirnov (KS) test statistic for each latent field parameter is correlated with the effective sample size (ESS) from NUTS, for both GEB and GPCA-AGHQ. This may be because parameters which are harder to estimate with INLA-like methods also have posterior distributions which are more difficult to sample from. Alternatively, it may be that high KS values are caused by inaccurate NUTS estimates generated by limited effective samples.

# Works cited

- Akaike, Hirotugu (1973). “Information theory as an extension of the maximum likelihood principle—In: Second International Symposium on Information Theory (Eds) BN Petrov, F”. In: *Csaki. BNPBF Csaki Budapest: Academiai Kiado*.
- Aldor-Noiman, Sivan et al. (2013). “The power to see: A new graphical test of normality”. In: *The American Statistician* 67.4, pp. 249–260.
- Arambepola, Rohan et al. (2022). “A simulation study of disaggregation regression for spatial disease mapping”. In: *Statistics in Medicine* 41.1, pp. 1–16.
- 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.
- Baeten, Jared M et al. (2012). “Antiretroviral prophylaxis for HIV prevention in heterosexual men and women”. In: *New England Journal of Medicine* 367.5, pp. 399–410.
- Bailey, Michael A (2023). “A new paradigm for polling”. In: *Harvard Data Science Review* 5.3.
- 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, Atilim Güneş 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.
- Bennett, James E et al. (2019). “Particulate matter air pollution and national and county life expectancy loss in the USA: A spatiotemporal analysis”. In: *PLoS medicine* 16.7, e1002856.
- Berger, James (2006). “The Case for objective Bayesian analysis”. In: *Bayesian Analysis* 1.3, pp. 385–402.
- Berild, Martin Outzen et al. (2022). “Importance sampling with the integrated nested Laplace approximation”. In: *Journal of Computational and Graphical Statistics* 31.4, pp. 1225–1237.
- Bernardo, José M and Adrian FM Smith (2001). *Bayesian theory*. John Wiley & Sons.

## Works cited

- 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, N et al. (1999). “Bayesian models for spatially correlated disease and exposure data”. In: *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*. Vol. 6. Oxford University Press, p. 131.
- 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.
- Betancourt, Michael (2017). *Robust Gaussian processes in Stan*. URL: <https://betanalpha.github.io/assets/case%5C%5Fstudies/gp%5C%5Fpart3/part3.html>.
- Bhatt, Samir et al. (2015). “The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015”. In: *Nature* 526.7572, pp. 207–211.
- 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*. Springer.
- Blangiardo, Marta et al. (2013). “Spatial and spatio-temporal models with R-INLA”. In: *Spatial and Spatio-temporal Epidemiology* 4, pp. 33–49.
- 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 Model Builder, 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. (2023). “Scoring epidemiological forecasts on transformed scales”. In: *PLoS Computational Biology* 19.8, e1011393.
- Bosse, Nikos I. et al. (2022). *Evaluating Forecasts with scoringutils in R*. 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.
- Bradley, Valerie C et al. (2021). “Unrepresentative big surveys significantly overestimated US vaccine uptake”. In: *Nature* 600.7890, pp. 695–700.
- 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.
- Brier, Glenn W (1950). “Verification of forecasts expressed in terms of probability”. In: *Monthly Weather Review* 78.1, pp. 1–3.
- 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.
- Brown, Patrick E (2015). “Model-based geostatistics the easy way”. In: *Journal of Statistical Software* 63, pp. 1–24.

## Works cited

- 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.
- Bürkner, Paul-Christian (2017). “brms: An R Package for Bayesian Multilevel Models Using Stan”. In: *Journal of Statistical Software* 80.1, pp. 1–28. DOI: [10.18637/jss.v080.i01](https://doi.org/10.18637/jss.v080.i01).
- 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.
- CDC (2014). *Understanding the HIV Care Continuum*. CDC. URL: [http://www.cdc.gov/hiv/pdf/dhap\\_continuum.pdf](http://www.cdc.gov/hiv/pdf/dhap_continuum.pdf).
- Chau, Siu Lun, Shahine Bouabid, and Dino Sejdinovic (2021). “Deconditional downscaling with Gaussian processes”. In: *Advances in Neural Information Processing Systems* 34, pp. 17813–17825.
- 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*.
- Crampin, Amelia C et al. (2012). “Profile: the Karonga health and demographic surveillance system”. In: *International Journal of Epidemiology* 41.3, pp. 676–685.
- 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](https://doi.org/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.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). “Maximum Likelihood from Incomplete Data Via the EM Algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22.

## Works cited

- 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.
- 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 and Emanuele Giorgi (2016). “Model-based geostatistics for prevalence mapping in low-resource settings”. In: *Journal of the American Statistical Association* 111.515, pp. 1096–1120.
- Diggle, Peter J, Paula Moraga, et al. (2013). “Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm”. In: *Statistical Science* 28.4, pp. 542–563.
- Dominguez, Kenneth L. et al. (Apr. 2016). “Updated guidelines for antiretroviral postexposure prophylaxis after sexual, injection drug use, or other nonoccupational exposure to HIV—United States, 2016”. In: URL: <https://stacks.cdc.gov/view/cdc/38856>.
- Donegan, Connor (2022). “geostan: An R package for Bayesian spatial analysis”. In: *The Journal of Open Source Software* 7.79, p. 4716. DOI: 10.21105/joss.04716.
- Duane, Simon et al. (1987). “Hybrid Monte Carlo”. In: *Physics letters B* 195.2, pp. 216–222.
- Duncan, Earl W, Nicole M White, and Kerrie Mengersen (2017). “Spatial smoothing in Bayesian models: a comparison of weights matrix specifications and their impact on inference”. In: *International Journal of Health Geographics* 16.1, pp. 1–16.
- 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: *Journal of the International AIDS Society* 24, e25788.
- Economist Impact (2023). *A triple dividend: the health, social and economic gains from financing the HIV response in Africa*.
- Esra, Rachel et al. (2023). *Improved indicators for subnational unmet antiretroviral therapy need in the health system: updates to the Naomi model in 2023*.
- Fattah, EA, JV Niekerk, and H Rue (2022). *Smart gradient—an adaptive technique for improving gradient estimation*.
- Fay, Robert E and Roger A Herriot (1979). “Estimates of income for small places: an application of James-Stein procedures to census data”. In: *Journal of the American Statistical Association* 74.366a, pp. 269–277.
- Fisher, Ronald Aylmer (1936). “Design of experiments”. In: *British Medical Journal* 1.3923, p. 554.
- FitzJohn, Rich et al. (2023). *orderly: Lightweight Reproducible Reporting*. <https://www.vaccineimpact.org/orderly/>, <https://github.com/vimc/orderly>.

## Works cited

- Flaxman, Seth R, Yu-Xiang Wang, and Alexander J Smola (2015). “Who supported Obama in 2012? Ecological inference through distribution regression”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 289–298.
- Follestad, Turid and Håvard Rue (2003). *Modelling spatial variation in disease risk using Gaussian Markov random field proxies for Gaussian random fields*.
- 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-Storti, 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.
- Fuglstad, Geir-Arne et al. (2019). “Constructing priors that penalize the complexity of Gaussian random fields”. In: *Journal of the American Statistical Association* 114.525, pp. 445–452.
- 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/>.
- Gärtner, Thomas et al. (2002). “Multi-instance kernels”. In: *ICML*. Vol. 2. 3, p. 7.
- 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. — (2006). “Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)”. In: *Bayesian Analysis* 1.3, pp. 515–534.
- (2007). “Struggles with survey weighting and regression modeling”. In.
- Gelman, Andrew, John B Carlin, et al. (2013). *Bayesian data analysis*. CRC press.
- Gelman, Andrew, Jessica Hwang, and Aki Vehtari (2014). “Understanding predictive information criteria for Bayesian models”. In: *Statistics and Computing* 24.6, pp. 997–1016.
- 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.

## Works cited

- Gneiting, Tilmann and Adrian E Raftery (2007). “Strictly proper scoring rules, prediction, and estimation”. In: *Journal of the American Statistical Association* 102.477, pp. 359–378.
- Godfrey-Faussett, Peter et al. (2022). “HIV prevention for the next decade: Appropriate, person-centred, prioritised, effective, combination prevention”. In: *PLoS Medicine* 19.9, e1004102.
- Goldstein, Michael (2006). “Subjective Bayesian analysis: principles and practice”. In: Gómez-Rubio, Virgilio (2020). *Bayesian inference with INLA*. CRC Press.
- Gómez-Rubio, Virgilio and Håvard Rue (2018). “Markov chain Monte Carlo with the integrated nested Laplace approximation”. In: *Statistics and Computing* 28, pp. 1033–1051.
- Goodrich, Ben et al. (2020). *rstanarm: Bayesian applied regression modeling via Stan*. R package version 2.21.1. URL: <https://mc-stan.org/rstanarm>.
- 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: *Morbidity and Mortality Weekly Report* 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.
- Gretton, Arthur et al. (2006). “A kernel method for the two-sample-problem”. In: *Advances in Neural Information Processing Systems* 19.
- Grieve, Richard et al. (2023). “The importance of investing in data, models, experiments, team science, and public trust to help policymakers prepare for the next pandemic”. In: *PLOS Global Public Health* 3.11, e0002601.
- 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. K. (1970). “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. In: *Biometrika* 57.1, pp. 97–109. URL: <http://www.jstor.org/stable/2334940> (visited on 12/29/2023).
- 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.

## Works cited

- 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 (2023a). *arealutils: Utility functions for beyond-borders*. R package version 0.0.1.
- (2023b). *multi.utils: Utility functions for multi-agyw*. R package version 0.1.0.
- Howes, Adam 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>.
- ICAP (2023). *Population-based HIV impact assessment: guiding the global HIV response*. Accessed 13/12/2023. URL: <https://phia.icap.columbia.edu>.
- 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.
- Jin, Harry, Arjee Restar, and Chris Beyrer (2021). “Overview of the epidemiological conditions of HIV among key populations in Africa”. In: *Journal of the International AIDS Society* 24, e25716.
- 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.
- Karatzoglou, Alexandros et al. (2019). “Package ‘kernlab’”. In: *CRAN R Project*.
- Kassanjee, Reshma et al. (2012). “A New General Biomarker-based Incidence Estimator”. In: *Epidemiology* 23.5.
- 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.
- Konstantoudis, Garyfallos et al. (2020). “Discrete versus continuous domain models for disease mapping”. In: *Spatial and Spatio-temporal Epidemiology* 32, p. 100319.
- Kristensen, Kasper (2021). *The comprehensive TMB documentation*. [https://kaskr.github.io/adcomp/\\_book/Introduction.html](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*.

## Works cited

- Law, Ho Chung et al. (2018). “Variational learning on aggregate outputs with Gaussian processes”. In: *Advances in Neural Information Processing Systems* 31.
- Lee, Duncan (2011). “A comparison of conditional autoregressive models used in Bayesian disease mapping”. In: *Spatial and Spatio-temporal Epidemiology* 2.2, pp. 79–89.
- 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.
- Li, Ye et al. (2012). “Log Gaussian Cox processes and spatially aggregated disease incidence data”. In: *Statistical Methods in Medical Research* 21.5. PMID: 22544855, pp. 479–507. DOI: 10.1177/0962280212446326. URL: <https://doi.org/10.1177/0962280212446326>.
- 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.
- Margossian, Charles C (2019). “A review of automatic differentiation and its efficient implementation”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.4, e1305.
- Margossian, Charles C and Andrew Gelman (2023). “For how many iterations should we run Markov chain Monte Carlo?” In: *arXiv preprint arXiv:2311.02726*.
- 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 (2020). “Integrated Nested Laplace Approximations (INLA)”. In: *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, pp. 1–19. DOI: <https://doi.org/10.1002/9781118445112.stat08212>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08212>.
- 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.
- Matheson, James E and Robert L Winkler (1976). “Scoring rules for continuous probability distributions”. In: *Management science* 22.10, pp. 1087–1096.
- Mayala, Benjamin K., Samir Bhatt, and Peter Gething (2020). *Predicting HIV/AIDS at Subnational Levels using DHS Covariates related to HIV*. DHS Spatial Analysis Reports 18. Rockville, Maryland, USA: ICF.
- 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.

## Works cited

- McGillen, Jessica B et al. (2018). "The emerging health impact of voluntary medical male circumcision in Zimbabwe: an evaluation using three epidemiological models". In: *PloS one* 13.7, e0199453.
- 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: *J. Chem. Phys* 21, p. 1087.
- Meyer-Rath, Gesine et al. (2018). "Targeting the right interventions to the right people and places: the role of geospatial analysis in HIV program planning". In: *AIDS (London, England)* 32.8, p. 957.
- 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). "Longitudinal population-level HIV epidemiologic and genomic surveillance highlights growing gender disparity of HIV transmission in Uganda". In: *Nature Microbiology*.
- Morris, Mitzi et al. (2019). "Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan". In: *Spatial and Spatio-temporal Epidemiology* 31, p. 100301. DOI: <https://doi.org/10.1016/j.sste.2019.100301>. URL: <https://www.sciencedirect.com/science/article/pii/S1877584518301175>.
- 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](https://doi.org/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.
- Noor, Abdisalan Mohamed (2022). "Country ownership in global health". In: *PLOS Global Public Health* 2.2, e0000113.
- Okabe, Masataka and Kei Ito (2008). *Color Universal Design (CUD): How to Make Figures and Presentations That Are Friendly to Colorblind People*. URL: <http://jfly.iam.u-tokyo.ac.jp/color/>.
- 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.

## *Works cited*

- Ord, Toby (2013). "The moral imperative toward cost-effectiveness in global health". In: *Center for Global Development* 12.
- Organization, World Health et al. (2022). *Consolidated guidelines on HIV, viral hepatitis and STI prevention, diagnosis, treatment and care for key populations*. World Health Organization.
- 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.
- Paciorek, Christopher J. and Mark J. Schervish (2006). "Spatial modelling using a new class of nonstationary covariance functions". In: *Environmetrics* 17.5, pp. 483–506. DOI: <https://doi.org/10.1002/env.785>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/env.785>.
- Parks, Robbie M et al. (2020). "Anomalously warm temperatures are associated with increased injury deaths". In: *Nature medicine* 26.1, pp. 65–70.
- Pebesma, Edzer (2018). "Simple Features for R: Standardized Support for Spatial Vector Data". In: *The R Journal* 10.1, pp. 439–446. DOI: 10.32614/RJ-2018-009. URL: <https://doi.org/10.32614/RJ-2018-009>.
- Pebesma, Edzer and Roger Bivand (2023). *Spatial Data Science: With Applications in R*. Chapman and Hall/CRC. DOI: 10.1201/9780429459016. URL: <https://doi.org/10.1201/9780429459016>.
- Pebesma, Edzer J. (2004). "Multivariable geostatistics in S: the gstat package". In: *Computers & Geosciences* 30, pp. 683–691.
- 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>.
- Rashid, T et al. (2023). "Inequalities in mortality from leading cancers in districts of England from 2002 to 2019: population-based high-resolution spatiotemporal analysis of vital registration data". In: *The Lancet Oncology*. URL: <http://hdl.handle.net/10044/1/107364>.
- Riebler, Andrea et al. (2016). "An intuitive Bayesian spatial model for disease mapping that accounts for scaling". In: *Statistical methods in medical research* 25.4, pp. 1145–1165.

## Works cited

- 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: *Probability Surveys* 1.none, pp. 20–71. DOI: 10.1214/154957804100000024. URL: <https://doi.org/10.1214/154957804100000024>.
- 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 (2022). "Graphical test for discrete uniformity and its applications in goodness-of-fit evaluation and multiple sample comparison". In: *Statistics and Computing* 32.2, p. 32.
- 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.
- Saunders, Daniel (2023). "The Besag-York-Mollie Model for Spatial Data". In: *PyMC examples*. Ed. by PyMC Team. DOI: 10.5281/zenodo.5654871.
- Schad, Daniel J, Michael Betancourt, and Shravan Vasishth (2021). "Toward a principled Bayesian workflow in cognitive science." In: *Psychological methods* 26.1, p. 103.
- Schlüter, Daniela K et al. (2016). "Using community-level prevalence of Loa loa infection to predict the proportion of highly-infected individuals: statistical modelling to support lymphatic filariasis and onchocerciasis elimination programs". In: *PLoS neglected tropical diseases* 10.12, e0005157.
- 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.

## Works cited

- Shapley, Lloyd S et al. (1953). *A value for n-person games*.
- Shumway, Robert H and David S Stoffer (2017). *Time Series Analysis and Its Applications With R Examples*. Springer.
- 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*.
- Smirnov, N (1948). "Table for estimating the goodness of fit of empirical distributions". In: *Annals of Mathematical Statistics* 19.2, pp. 279–281.
- 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 et al. (1996). "BUGS 0.5 Examples". In: *MRC Biostatistics Unit, Institute of Public health, Cambridge, UK* 256.
- Spiegelhalter, David J 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.
- Stan Development Team (2023). *Stan Reference Manual*. URL: <https://mc-stan.org/docs/reference-manual/index.html>.
- Stein, Michael L (1999). "Interpolation of spatial data: some theory for kriging". In: *Stevens, Oliver et al. (2023). "Population size, HIV prevalence, and antiretroviral therapy coverage among key populations in sub-Saharan Africa: collation and synthesis of survey data 2010–2023". In: medRxiv. URL: https://www.medrxiv.org/content/early/2023/11/22/2022.07.27.22278071.*
- 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.Supp1 3, S227.
- Stover, John and Yu Teng (2021). "The impact of condom use on the HIV epidemic". In: *Gates Open Research* 5.
- Stringer, Alex (2021). "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.
- Tanaka, Yusuke et al. (2019). "Spatially aggregated Gaussian processes with multivariate areal outputs". In: *Advances in Neural Information Processing Systems*, pp. 3005–3015.

## Works cited

- 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. (Dec. 2022). "Efficient Bayesian inference of Instantaneous Reproduction Numbers at Fine Spatial Scales, with an Application to Mapping and Nowcasting the Covid-19 Epidemic in British Local Authorities". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 185.1, S65–S85. doi: 10.1111/rssa.12971. URL: <https://doi.org/10.1111/rssa.12971>.
- 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%5C%5Fadolescentsgirlsandyoungwomenprograms%5C%5Fframeworkmeasurement%5C%5Fen.pdf>.
- Thigpen, Michael C et al. (2012). "Antiretroviral preexposure prophylaxis for heterosexual HIV transmission in Botswana". In: *New England Journal of Medicine* 367.5, pp. 423–434.
- Thyng, Kristen M et al. (2016). "True colors of oceanography: Guidelines for effective and accurate colormap selection". In: *Oceanography* 29.3, pp. 9–13.
- 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](https://www.state.gov/wp-content/uploads/2022/11/PEPFAR-Latest-Global-Results_December-2022.pdf). Accessed: 10/08/2023.
- UNAIDS (2014). "90-90-90. An ambitious treatment target to help end the AIDS epidemic". In: Accessed: December 2023.
- (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.
- UNAIDS, WHO, et al. (2022). *Using recency assays for HIV surveillance: 2022 technical guidance*. World Health Organization.

## Works cited

- UNAIDS and WHO (2021). *Voluntary Medical Male Circumcision Progress Brief*. UNAIDS. URL: [https://hivpreventioncoalition.unaids.org/wp-content/uploads/2021/04/JC3022\\_VMMC\\_4-pager\\_En\\_v3.pdf](https://hivpreventioncoalition.unaids.org/wp-content/uploads/2021/04/JC3022_VMMC_4-pager_En_v3.pdf).
- UNICEF (2019). *Adolescent & social norms situation in Mozambique*. Accessed 25/03/2022. URL: <https://www.unicef.org/mozambique/en/adolescent-social-norms>.
- USAID (2012). *Sampling and Household Listing Manual: Demographic and Health Surveys Methodology*. URL: [https://dhsprogram.com/pubs/pdf/DHSM4/DHS6\\_Sampling\\_Manual\\_Sept2012\\_DHSM4.pdf](https://dhsprogram.com/pubs/pdf/DHSM4/DHS6_Sampling_Manual_Sept2012_DHSM4.pdf).
- 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, Andrew Gelman, and Jonah Gabry (2017). “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. In: *Statistics and computing* 27, pp. 1413–1432.
- Vehtari, Aki and Janne Ojanen (2012). “A survey of Bayesian predictive methods for model assessment, selection and comparison”. In.
- 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.
- WHO and UNAIDS (2007). “New data on male circumcision and HIV prevention: policy and programme implications”. In: *Geneva: World Health Organization*.
- Wilke, Claus O (2019). *Fundamentals of Data Visualization: A Primer on making informative and compelling figures*. O'Reilly Media.
- Wilson, Katie and Jon Wakefield (Sept. 2018). “Pointless spatial modeling”. In: *Biostatistics* 21.2, e17–e32. DOI: 10.1093/biostatistics/kxy041. URL: <https://doi.org/10.1093/biostatistics/kxy041>.
- 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.

*Works cited*

- 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.