Methods and applications of Bayesian spatio-temporal statistics for prioritised HIV prevention

Imperial College London

Adam Howes

Imperial College London

A thesis submitted for the degree of

Doctor of Philosophy

2023

For $\sum_i u_i$

Acknowledgements

Thanks to Jeff Eaton and Seth Flaxman for supervision of this research; staff and students of the StatML CDT at Imperial and Oxford; members of the HIV inference group at Imperial; Mike McLaren, Kevin Esvelt, the Nucleic Acid Observatory team, and the Sculpting Evolution lab for hosting my visit to MIT; Alex Stringer for hosting my visit to Waterloo; the Effective Altruism community; the Bill & Melinda Gates Foundation and EPSRC for funding this PhD; my friends and family for their support.

Adam Howes
Imperial College London
2023

Abstract

HIV remains a large problem. Disease burden is unevenly distributed. Effective public health response and prioritised prevention requires accurate, timely, high-resolution estimates of epidemic and demographic indicators. Thoughtful statistical modelling is required to overcome significant data challenges. In this thesis, I develop and apply Bayesian spatio-temporal methods for HIV surveillance.

Contents

Appendices

List of Figures

List of Tables

List of Abbreviations

HIV	Human Immunodeficiency Virus.
AIDS	Acquired Immune Deficiency Syndrome.
PEPFAR	President's Emergency Plan for AIDS Relief.
HIV	Demographic and Health Surveys.
AIS	AIDS Indicator Survey.
MCMC	Markov Chain Monte Carlo.
INLA	Integrated Nested Laplace Approximation.
GP	Gaussian Process.
CAR	Conditionally Auto-regressive.
ANC	Antenatal Clinic.
ART	Antiretroviral Therapy.
UNAIDS	United Nations Joint Programme on HIV/AIDS.
CDC	Centers for Disease Control and Prevention.
UAT	Unlinked Anonymous Testing.
PMTCT	Prevention of Mother-to-Child Transmission.
PLHIV	People Living with HIV.
MPES	Multi-parameter Evidence Synthesis.
VI	Variational Inference.
SAE	Small Area Estimation.
GMRF	Gaussian Markov Random Field.
HMC	Hamiltonian Monte Carlo.

List of Notations

ρ	HIV prevalence.
α	ART coverage.
$\mathcal{S}$	Spatial study region $\mathcal{S} \subseteq \mathbb{R}^2$.
$s \in \mathcal{S}$	Point location.
$\mathcal{T}$	Temporal study period $\mathcal{T} \subseteq \mathbb{R}$.
$t \in \mathcal{T}$	Time.

1 Background

1.1 Disease surveillance and small-area estimation

- Disease surveillance is a central application of statistics
- Small-area estimation in health, epidemiology and environment
- The Small-Area Health Statistics Unit at Imperial was set-up to monitor health around point sources of environmental pollution in response to the Sellafield enquiry into the increased incidence of childhood leukemia leukaemia near a nuclear reprocessing plant (Elliott et al. 1992). This research has a focus on ratios of observed events to expected events, and testing hypothesis about hot-spots.

1.2 HIV/AIDS

- HIV/AIDS has a large disease burden
- The disease burden is unevenly distributed in space and across communities and individuals
- Surveillance techniques and statistical models have been used to respond to the epidemic

Background

- Key HIV indicators are HIV prevalence, HIV incidence, ART coverage and coverage of other interventions such as PrEP, PEP
- Data difficulties including sparsity in space and time, survey bias, conflicting information sources, hard to reach populations, demography
- Aims for HIV response going forward, and surveillance capabilities are needed to meet them
- Phasing out of nationally-representative household surveys for HIV
  - Bayesian survey design
- Importance of relying on multiple sources of information Creates requirement for for complex models e.g. evidence synthesis, Naomi, multivariate models
- Why isn't case-based surveillance included yet?
  - There aren't individual linked databases and patient records have to be consolidated
  - Passive case-based surveillance
  - Post-hoc matching and create a case-based surveillance record
- Drivers of transmission
- Possible interventions are ART, condoms, PrEP and PEP, education, economic empowerment, VMMC
- Geographic priorisation versus demographic priorisation: hotspots, key populations, screening and individual level risk characteristics
- Adolescent girls and young women identified as a key demographic, stratification by sexual risk
- Interventions more likely to be demographic specific rather than geographic specific so if majority of difference in effectiveness depends on intervention type then demographic targeting may be more priority
- The population strategy of Geoffrey Rose

1.3 Bayesian spatio-temporal statistics

- The practice of doing Bayesian statistics primarily concerns construction of a generative model for the data we observe
- In spatio-temporal statistics, the data is indexed by spatial and or temporal location
- The independent and identically distributed (IID) assumptions commonly used for observations are rarely suitable in the spatio-temporal setting
- We expect there to be spatio-temporal structure
- Given a generative model, computation of the posterior distribution proceeds using approximate Bayesian inference methods
- Markov chain Monte Carlo (MCMC) is the most popular approach and works by simulating samples from a Markov chain which by construction has stationary distribution equal to the distribution of interest
- Variational Bayes approaches assume the posterior distribution belongs to some class and use optimisation to choose the best member of that class
- Laplace approximation and integrated nested Laplace approximation
- Empirical Bayes
- Definition of a latent Gaussian model (Rue et al. 2009)

$$\text{(Observations)} \qquad y_i \sim p(y_i \,|\, x_i, \boldsymbol{\theta}), \quad i = 1, \ldots, n, \tag{1.1}$$

$$\text{(Latent field)} \qquad \mathbf{x} \sim \mathcal{N}(\mathbf{x} \,|\, \mathbf{0}, \mathbf{Q}(\boldsymbol{\theta})^{-1}), \tag{1.2}$$

$$\text{(Parameters)} \qquad \boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \tag{1.3}$$

  - Common examples

- Examples of models used in HIV inference which are close to being latent Gaussian models, but aren't, and hence can't be fit using INLA
  - Disaggregation models
  - Evidence synthesis models like Naomi (Eaton, Dwyer-Lindgren, et al. 2021; Eaton, Bajaj, et al. 2019)

  - Compartmental models
  - ART attendance models
  - Multinomial models like for district-level risk factors
    - Multinomial logistic regression

- Other complex models from ecology that can't currently be fit using INLA
- Definition of extended latent Gaussian models (Stringer et al. 2021)
  - Many-to-one is not an issue for `R-INLA`, the latent field is implemented as a concatenation of many vectors already. For example, for $\eta_i = \beta_0 + \phi_i$ with $i = 1, \ldots, n$ the latent field is $(\eta_1, \ldots, \eta_n, \beta_0, \phi_1, \ldots, \phi_n)^\top$ of dimension $2n + 1$
  - For additive models, the only non-linearity is in the link function
- Particular properties of spatio-temporal models (and LGMs) which make INLA, if feasible, often the best option
- The increasing popularity of empirical Bayes approaches, like Template Model Builder (Osgood-Zimmerman and Wakefield 2021)
- Adaptive Gauss Hermite quadrature (AGHQ), like the central composite design (CCD) and grid strategies, is one way to choose the hyper-parameter integration points in the integrated nested Laplace approximation (INLA)
- Finn Lindgren is working on a method for non-linear predictors, called the iterative INLA method
  - More slides here
- Thesis work of Follestad that stayed as a preprint
- How does the ecological fallacy relate to aggregated output models

2
Understanding models for spatial structure

Code for the analysis in this chapter is available from `athowes/areal-comparison` and supported by the R package `arealutils`. Include an edited version of the corresponding paper here.

3

A multinomial spatio-temporal model for risk group proportions

In this chapter I describe an application of Bayesian spatio-temporal statistics to small-area estimation of HIV risk group proportions. This project was initially worked on by Kathryn Risher, who continues to lead dissemination of the results as a tool to be used by countries, as well as model development and data extensions, and is described in Howes, Risher, et al. (2022), which I draw from here. Code for the analysis in this chapter is available from `athowes/multi-agyw` and supported by the R package `multi.utils`.

3.1 Background

3.2 Data

3.3 Statistical model

3.4 Results

3.5 Conclusions

4

Fast, approximate inference for the Naomi model

Code for the analysis in this chapter is available from `athowes/elgm-inf` and supported by the R package `inf.utils`. Include an edited version of the corresponding paper here.

5 Future work and conclusions

5.1 Future work

Avenues for future work include:

1. Extending the risk group model described in Chapter 3 to include all adults 15-49. This may involve modelling of age-stratified sexual partnerships (Wolock et al. 2021). Such a model would likely fall out of the scope of `R-INLA`, but may be possible using `aghq` with Laplace marginals as described in Chapter 4.
2. Evaluating the accuracy of `aghq` with Laplace marginals for a greater variety of extended latent Gaussian models.

5.2 Conclusions

The spatial structure chapter is interesting because:

- I designed experiments to thoroughly compare models for spatial structure using tools for model assessment such as proper scoring rules and posterior predictive checks.

The risk group chapter is interesting because:

- I estimated HIV risk group proportions for AGYW, enabling countries to prioritise their delivery of HIV prevention services.
- I analysed the number of new infections that might be reached under a variety of risk stratification strategies.
- I used `R-INLA` to specify multinomial spatio-temporal models via the Poisson-multinomial transformation. This includes complex two- and three-way Kronecker product interactions defined using the `group` and `replicate` options.

The fast, approximate inference chapter is interesting because:

- I developed a novel Bayesian inference method, motivated by a challenging and practically important problem in HIV inference.
- The method enables integrated nested Laplace approximations to be fit to and studied on a wider class of models than was previously possible.
- My implementation of the method was straightforward, building on the `TMB` and `aghq` packages, and described completely and accessibly in Howes, Stringer, et al. (2023).

My final conclusions are:

- Modelling complex data, more often than not, pushes the boundaries of the statistical toolkit available
- One challenge I encountered was that of trying to implementing identical models across multiple frameworks with the aim of studying the inference method. Or, of a similarly fraught nature, comparing different models implemented in different frameworks with the aim of studying model differences. The frequently asked questions section of the `R-INLA` website (Rue 2023) notes that, "the devil is in the details". I have resolved this challenge by using a given `TMB` model template to fit models using multiple inference methodologies: empirical Bayes with Gaussian marginals (Kristensen et al. 2016), AGHQ with Gaussian marginals (Stringer 2021b), AGHQ with Laplace marginals (Howes,

Stringer, et al. 2023), and HMC using NUTS (Monnahan and Kristensen 2018). The benefits of such a ecosystem of packages are noted by Stringer (2021a). I would particularly highlight the benefit of enabling analysts to easily vary their choice of inference method based on the stage of model development that they are in.

- I have aimed to write this thesis, and the work described within it, in keeping with the principles of open science. I hope that doing so allows my work to be scrutinised, and, optimistically, built upon. This would not have been possible without a range of tools from the R ecosystem such as `rmarkdown` and `rticles`, as well as those developed within the MRC Centre for Global Infectious Disease Analysis like `orderly` and `didehpc`.

Appendices

A
The First Appendix

Works Cited

Eaton, Jeffrey W, Sumali Bajaj, et al. (2019). "Joint small-area estimation of HIV prevalence, ART coverage and HIV incidence". In: *Working paper*.

Eaton, Jeffrey W, Laura Dwyer-Lindgren, et al. (2021). "Naomi: A New Modelling Tool for Estimating HIV Epidemic Indicators at the District Level in Sub-Saharan Africa". In.

Elliott, Paul et al. (1992). "The Small Area Health Statistics Unit: a national facility for investigating health around point sources of environmental pollution in the United Kingdom." In: *Journal of Epidemiology & Community Health* 46.4, pp. 345–349.

Howes, Adam, Kathryn A Risher, et al. (2022). "Spatio-temporal estimates of HIV risk group proportions for adolescent girls and young women across 13 priority countries in sub-Saharan Africa". In: *medRxiv*.

Howes, Adam, Alex Stringer, et al. (2023). "Integrated nested Laplace approximations for extended latent Gaussian models with application to the Naomi HIV model". In: *arXiv*.

Kristensen, Kasper et al. (2016). "TMB: Automatic Differentiation and Laplace Approximation". In: *Journal of Statistical Software* 70.i05.

Monnahan, Cole C and Kasper Kristensen (2018). "No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages". In: *PloS one* 13.5, e0197954.

Osgood-Zimmerman, Aaron and Jon Wakefield (2021). *A Statistical Introduction to Template Model Builder: A Flexible Tool for Spatial Modeling*. arXiv: `2103.09929 [stat.ME]`.

Rue, Havard (2023). "'R-INLA' Project - FAQ". Accessed 23/01/2023. URL: `https://www.r-inla.org/faq`.

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.

Stringer, Alex (2021a). "Implementing Approximate Bayesian Inference Using Adaptive Quadrature". Statistics Graduate Student Research Day 2021, The Fields Institute for Research in Mathematical Sciences. URL: `http://www.fields.utoronto.ca/talks/Implementing-Approximate-Bayesian-Inference-Using-Adaptive-Quadrature`.

— (2021b). "Implementing Approximate Bayesian Inference using Adaptive Quadrature: the aghq Package". In: *arXiv preprint arXiv:2101.04468*.

Stringer, Alex, Patrick Brown, and Jamie Stafford (2021). "Fast, Scalable Approximations to Posterior Distributions in Extended Latent Gaussian Models". In: *arXiv preprint arXiv:2103.07425*.

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.