AIMS African Institute for Mathematical Sciences RESEARCH

Imperial College London

MACHINE LEARNING & GLOBAL HEALTH NETWORK

Modern Statistics and Machine Learning for Population Health in Africa

A hands-on course for students and researchers at the intersection of probabilistic programming, statistics, AI and global health.

Dates: 24th - 28th March 2025

Organised by: Department of Mathematics, Imperial College London, AIMS South Africa, and the Machine Learning & Global Health Network

Course summary

One of the groundbreaking advances in machine learning research in the past decade is surrounding the emergence of increasingly sophisticated, robust, and easily usable probabilistic programming languages. These new tools, including Stan or numpyro, hide tedious calculations involving automatic differentiation and gradient-based optimization from the end-user, making modern statistical methods widely available to data scientists in Africa that wish to address some of the most urgent challenges on the continent, ranging from habitat degradation, air pollution, extreme weather events, disease outbreaks and population health in general.

This one-week course will cover how you can integrate modern statistical techniques with the Stan probabilistic programming language to effectively address a broad range of applications from epidemiological, genomic and spatial data. We hope this course will equip you with intelligence-driven statistical technologies to drive your own evidence-based discoveries in global health or other applications, and more broadly increase your fluency in artificial intelligence and modern statistics.

This short course will provide an introduction to modern methods at the intersection of epidemiology, statistical modelling and data science with applications in global health. Attendees can expect to learn probabilistic programming using Stan, computationally efficient non-parametric Bayesian inference, statistical techniques for infectious disease modelling and public health, and phylogenetics for the analysis of pathogen sequence data. This course will illustrate the application of these methods to epidemiological, genomic and spatial data, and demonstrate how to interpret results from statistical analyses, equipping attendees with the tools to carry out their own independent research in these areas.

Content covered/What attendees will learn

- Bayesian workflow with probabilistic programming in Stan
- Core regression models for hierarchical data
- Gaussian process regression with Stan
- State-of-the-art GP approximations for scalable inference
- Infectious disease modelling with probabilistic programming
- Pathogen phylogenetics with Stan

Practical real-world examples with applications in malaria modelling, HIV epidemiology, ecology, environmental health
Varied datasets including spatial, genomic and epidemiological data
Stan templates and **Python code** for implementing the methods and applications covered

Learning styles/course structure

- Lectures
- Individual labs
- Group project
- Presenting findings

Who should attend/pre-requisites

- Students and researchers interested in advanced statistical methods and probabilistic programming with applications in global health, including analysis of clinical trials and studies, infectious disease epidemiology and modelling outbreaks, and handling large genomic datasets for the surveillance of pathogens.
- Attendees should have good knowledge of Python and pandas to participate fully in the practical components. Previous experience with a probabilistic programming language (e.g. Stan, NumPyro, PyMc, Turing.jl) is advantageous but not essential.
- Attendees should be familiar with git for reproducible analyses and collaborative coding.

Full programme

Day 1 (9.00-17.00)
Introduction to probabilistic programming with Stan

1 Lecture 1: Welcome (09.00 - 09.30)

2 Review of the fundamentals of Bayesian inference (09.30 – 10:30)

3 Break (10.30-11.00)

4 Lecture 2: Statistical learning (11.00-12.00)

5 Lecture 3: Introduction to Stan (12.00 - 13:00)

6 Lunch (13.00 - 14.00)

7 Practical 1: Introduction to statistical learning with Stan in Python (14.00 - 16.00)

8 Break (16:00-16:30)

9 Lecture 4: Recap session (16:30-17:00)

Day 2 (9.00-17.00)
Scalable Gaussian process regression models in Stan

1 Q&A in small groups (09.00 - 09.30)

2 Lecture 5: Intro to Gaussian processes (09.30-10.30)

3 Break (10.30-11.00)

4 Lecture 6: Gaussian processes continued (11.00-12.00)

5 Practical 2: Gaussian processes in Stan (12.00-13.00)

6 Lunch (13.00-14:00)

7 Lecture 7: Scalable Gaussian process regression models (14.00-15:00)

8 Inspirational Lecture: Two research talks on real-world applications from the Machine Learning and Global Health Network (15:00 - 16.00)

9 Break (16:00-16:30)

10 Lecture 8: Recap session (16:30-17:00)

Day 3 (9.00-17.00)
Gaussian processes continued

1 Q&A in small groups (09.00 - 09.30)

2 Practical 3: Scalable Gaussian process regression models (09:30 – 10:30)

3 Break (10.30-11.00)

4 Group project: Synthesising material analysing a real-world dataset (11.00 - 13:00)

5 Lunch (13:00-14:00)

6 Group project continued (14.00 - 15.30)

7 Break (15:30-16:00)

8 Groups present (16.00-16.30)

9 Quiz (16:30-16:50)

10 Introduce take-home assignment (16:50-17:00)

Day 4 (9.00-17.00)
Infectious Disease Modelling with Stan

1 Lecture 9: Introduction to Infectious Disease Modelling and Compartmental Modelling (09.00-10:30)

2 Break (10.30-11.00)

3 Practical 4: Deriving simple SIR type models (11:00 – 11:40)

4 Practical 5: SIR models in Stan (11.40-13:00)

5 Lunch (13:00-14:00)

7 Practical 5: SIR models in Stan (continued) (14:00-15.30)

8 Break (15:30-16:00)

9 Inspirational Lecture: Two talks on real-world applications from research and industry contacts based in Cape Town (16.00 - 17.00)

Day 5 (9.00-17.00)
Phylogenetics

1 Lecture 10: Recap session and Q&A in small groups (09.00 - 09.30)

2 Lecture 11: Introduction to phylogenetics (09:30-10:30)

3 Break (10.30-11.00)

4 Practical 6: Running a phylogenetic pipeline (11:00-13:00)

5 Lunch (13:00-14:00)

6 Practical 7: Phylogenetics continued (14:00 – 15:00)

7 Q&A (15:00 - 15:30)

8 Break (15:30-16:00)

9 Quiz (16:00 - 16:30)

10 Lecture 12: Discussion and wrap-up (16:30-17:00)

11 Social event (18:00 - 20:00)

Lecturers
Juliette Unwin

Dr Juliette Unwin is a lecturer in statistical science at the University of Bristol. She is interested in developing and applying novel methods for infectious disease outbreak analysis to help inform policy makers in real time. Her current research focuses on developing spatial temporal renewal-based transmission models alongside estimating the number of children affected by COVID-19 and crises. She has previously been involved in real-time analysis of Ebola in the Democratic Republic of Congo alongside the World Health Organisation and COVID-19 in New York State with the local government.

https://research-information.bris.ac.uk/en/persons/h-juliette-t-unwin

Alexandra Blenkinsop

Dr Alexandra Blenkinsop is a Research Associate in the Department of Mathematics at Imperial College London. Her research centres around developing and applying statistical methods to inform public health policy decisions. Research interests include Bayesian statistics, pathogen phylodynamics, and methods for partially observed data. She has collaborated with the HIV Transmission Elimination Amsterdam Initiative, The Botswana-Harvard Health Partnership and the United States Centers for Disease Control.

https://www.imperial.ac.uk/people/a.blenkinsop

Tristan Naidoo

Tristan Naidoo is a PhD student in the Department of Infectious Disease Epidemiology. His interests lie at the intersection of Natural Language Processing and Public Health. In line with these interests, his research focuses on using Large Language Models to investigate how Twitter data can be used to quantify adherence to protective behaviours during the COVID-19 pandemic.

https://mlgh.net/author/tristan-naidoo/

Shozen Dan

Shozen Dan, a PhD student at the StatML CDT, jointly hosted by Imperial College London and the University of Oxford, focuses his research on developing statistical methods with applications in public health, epidemiology, and infectious diseases. His current projects involve estimating social contact matrices with high precision, devising adjustments for reporting fatigue in longitudinal surveys, and analyzing the spatio-temporal variability of social contacts using established Bayesian statistical modeling techniques. Previously, he has collaborated with Center for Asian Research and Education (CARE) at Stanford University's School of Medicine on the analysis of health outcomes in Asian Americans. Additionally, he is an active member of the Machine Learning and Global Health Network (https://mlgh.net).

https://shozend.github.io/

Josh Corneck

Josh is a third year PhD student in the Department of Mathematics at Imperial College London. His research is primarily centered around network point processes, and more generally in the application of networks to aid modelling performance. He is currently working on developing models that incorporate Bayesian nonparametric approaches to help capture latent group structure among the nodes on the network.
https://profiles.imperial.ac.uk/josh.corneck-willcox20

Michael Whitehouse

Dr Michael Whitehouse is a Research Associate in the School of Public Health at Imperial College London. His research is around statistical inference in infectious disease models for heterogeneous populations and the modelling and control of infectious disease outbreaks on contact network structures.

https://michael-whitehouse.github.io/

Sahoko Ishida

Dr Sahoko Ishida is a Research Associate at the Department of Computer Science, University of Oxford. Her research focuses on Gaussian process regression, spatio-temporal analysis, small-area estimation, Bayesian inference, and statistical learning, with applications in food security, environmental studies, and epidemiology.
https://sahokoishida.github.io/