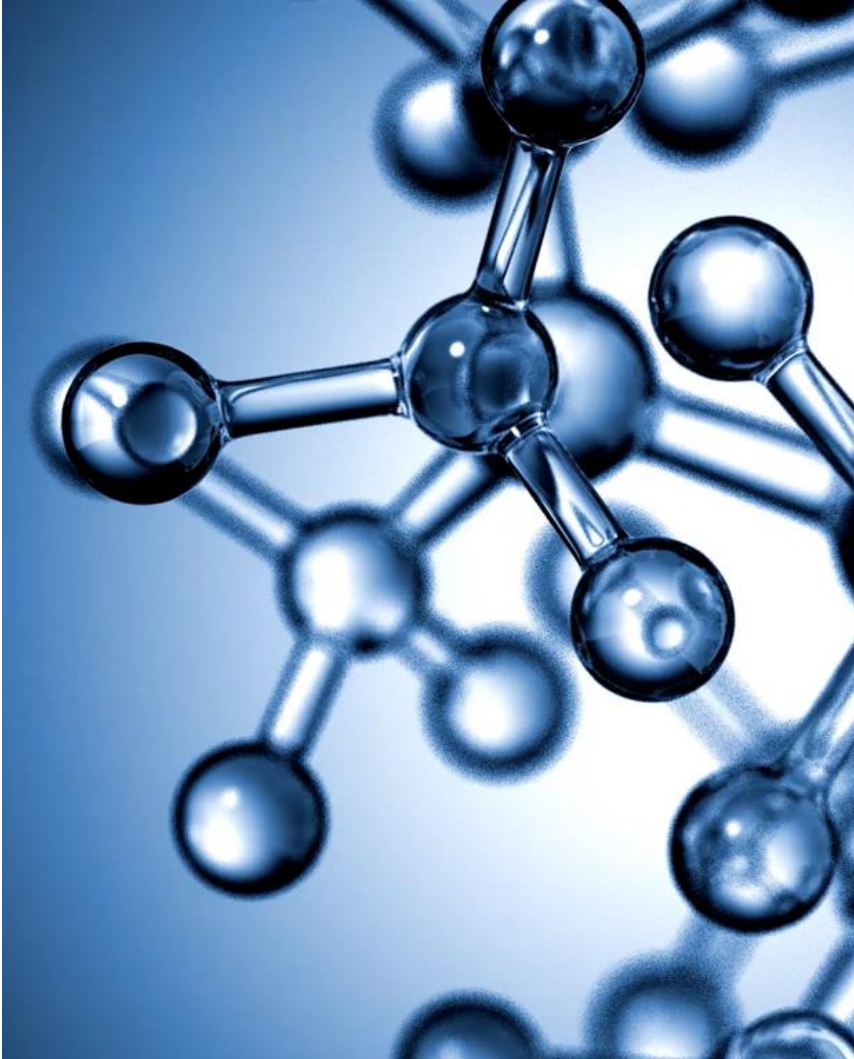




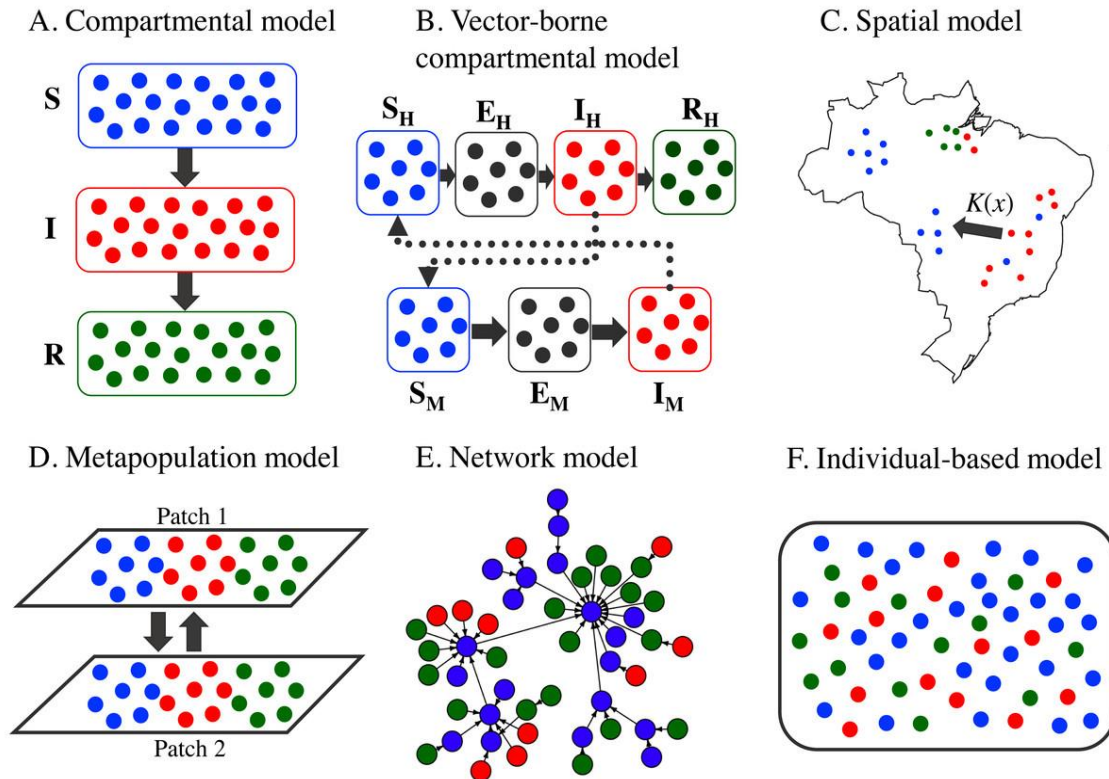
GEM: a domain-specific modelling language for disease outbreak models

Crystal Leong,

Lancaster University,

w.y.leong3@lancaster.ac.uk

What is Epidemiological Modelling?



(Wiratsudakul et al, 2018, Fig. 1)

Epidemiology is the study of how diseases are found, spread, and controlled

- Disease models come in many forms
- Inference is performed on the model
- Outbreak response require fast inference
- Existing probabilistic programming tools (BUGS, STAN)

How can we relieve the computational burden?

Aim of GEM project: Epidemiologist describes the disease, and the library builds the model.

Data: Daily Cases over Time

Consider that we have an outbreak on our hands...

We might be interested in how quickly the rate of disease spread is increasing, or how many initial cases there may be.

Assume that infection rate grows exponentially.

$$C(t) = ae^{rt} \quad \text{Cases grow exponentially}$$

$$\log C(t) = \log(a) + rt \quad \text{Linear model in log space}$$

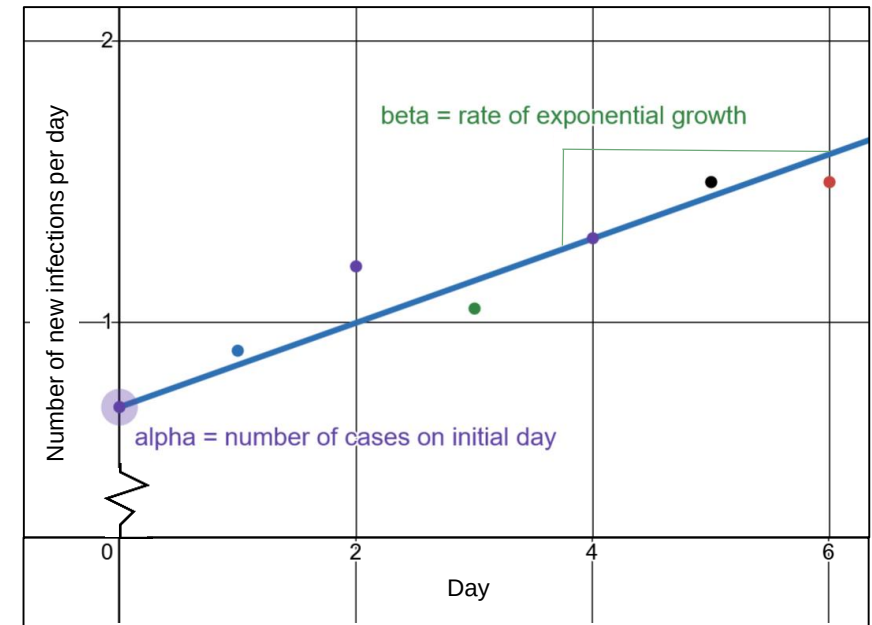
Initial cases = 5

Daily exponential growth rate = 0.15 (all hidden)

Noise size = 0.25

Day	0	1	2	...	30
New infections	5	6	7	...	309
Log (New infections)	1.807	1.779	1.956	...	5.773

NOTE: log here refers to the natural log, a.k.a. ln

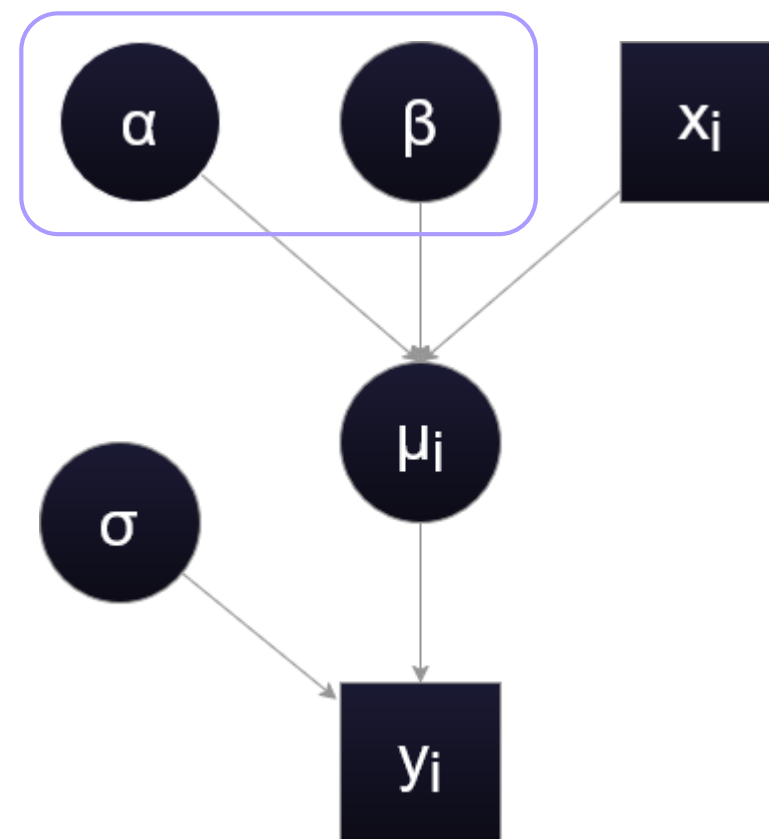


Infections per day in **log** space

The Directed Acyclic Graph

Working in Python, Gemlib, numpy and JAX libraries.
We define classes for the nodes (stochastic, data or logical) that make up the DAG.

- 1) Model is parsed into a DAG
- 2) Tell it what to do (parameters of interest, tuning)
- 3) Choose appropriate sampler for each node
- 4) Chain together steps in order
- 5) Execution (JAX jit)
 - i. Start with a guess
 - ii. Nudge one node (single/multi-site)
 - iii. Check if the nudge improved the fit
 - iv. Repeat for every node, many times over



$\alpha \sim \text{Normal}(0, 1)$ intercept, initial infections
 $\beta \sim \text{Normal}(0, 1)$ slope, daily exp. growth rate
 $\sigma \sim \text{Gamma}(1, 1)$ noise size (must be positive)
* $\mu_i = \alpha + \beta x_i$ the line, growth trend, log
 $y_i \sim \text{Normal}(\mu_i, \sigma)$ observed daily cases, log

Results and Quality

Initial cases = $e^{1.6263} = 5.0850$ cases

Single-site Metropolis Hastings

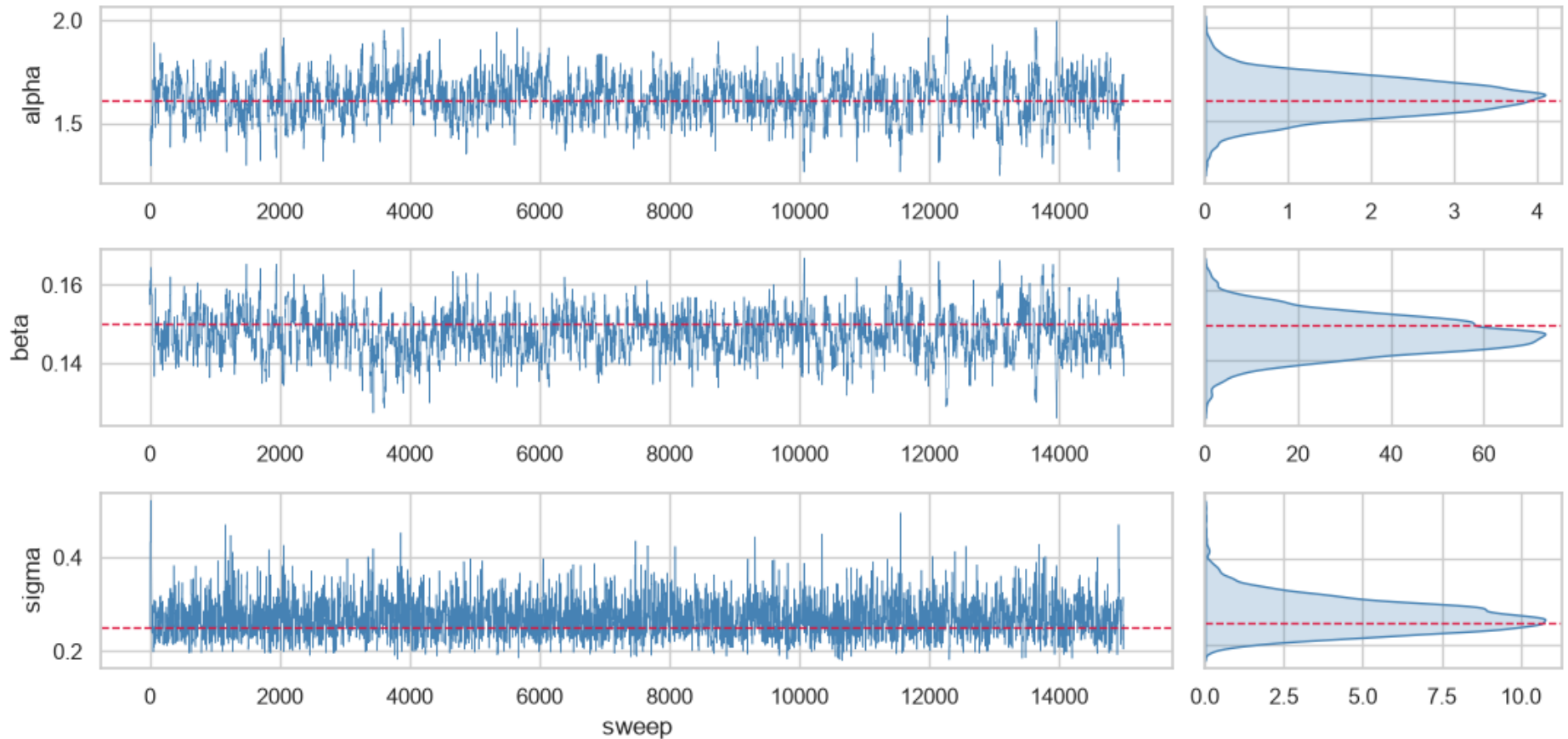
20000 sweeps, 5000 discarded as burn-in

Parameter	Mean	SD	Truth
(α) Initial cases, log	1.6263	0.0977	1.609
(β) Rate of growth	0.1474	0.0055	0.150
(σ) Noise	0.2689	0.0385	0.250

“Walking the posterior”

α and β are compensating

We like fuzzy caterpillars!
This is good mixing.

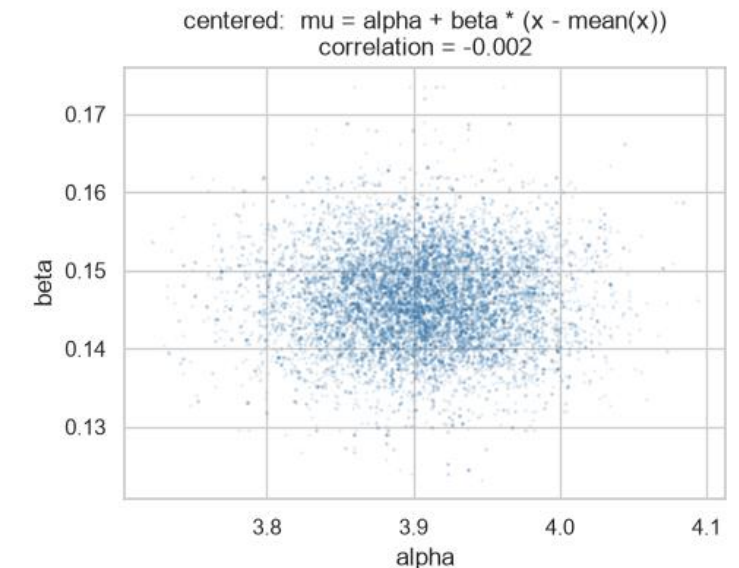
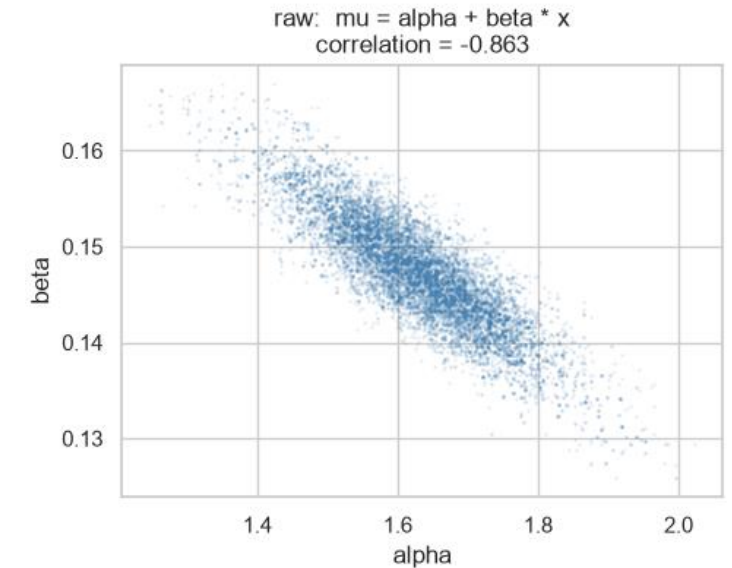


Single Site vs. Multi Site

- Normal posterior (top), centred posterior (bottom)
- Raw correlation = -0.863, Centred correlation = -0.002
- ESS = Effective Sample Size
- High ESS → Posterior estimates (mean, SD) are more precise → Good

Posterior	Sampler	ESS alpha	ESS beta	ESS sigma
Raw	Single	474	452	2114
Raw	Grouped	908	961	1843
Centred	Single	2906	2073	2381
Centred	Grouped	2272	793	1994

Compared to single-site sampling, multi-site sampling is...
2.0x better in raw data, and 1.6x worse in centred data



Discussion

- A look into epidemiological modelling
- Log-linear regression model as daily new infections over time
- Model description → automatic graph construction → automatic inference
- Challenges around designing a probabilistic modelling interface
- Multi-site sampling and centering of data can be effective mitigation against correlated parameters

Future Direction

- Automation of multi-site choices through analysis of correlation
- Mixed models that require multiple different MCMC algorithms

With thanks to Chris Jewell (supervisor and RSE mentor) for their support and guidance throughout.