Bayesian nonparametric inference for stochastic infectious disease models

Theo Kypraios

School of Mathematical Sciences @ University of Nottingham

Athens University of Economics and Business, Greece – May, 2023

Acknowledgements

Rowland Seymour
University of Birmingham

Phil O'Neill
University of Nottingham

UKRI Engineering and Physical Sciences Research Council

Table of contents

1. Motivation
2. Roadmap – Homogeneously Mixing Populations
3. Heterogeneously Mixing Populations
4. Model Extensions
5. Challenges
6. Conclusions

Motivation

Epidemic Modelling

Mathematical and statistical modelling has become a valuable tool in the analysis of infectious disease dynamics:

- control strategies;
- informing policy–making at the highest levels;
- fundamental role in the fight against disease spread $\rightarrow$ see COVID-19!

Analysis of Outbreak Data on Infectious Diseases

1. Construct a model to understand the transmission mechanism
   [eg. taking into account individual's characteristics, location, household, ...]
2. Make inference for the model parameters.
   [Bayesian / Frequentist]
3. Use the fitted model to make predictions, evaluate control measures etc.
   [Inform policy makers.]

Modelling & Estimation

- Enormous attention has been given to the development of:
  - realistic (parametric) model of varying complexity, and
  - methods for efficient parameter estimation.
- Particular focus has been given to the construction of *computationally intensive methods*, for example
  - Markov Chain Monte Carlo (MCMC),
  - Sequential Monte Carlo (SMC),
  - Approximate Bayesian Computation (ABC),
  - Partially Observed Markov Processes (POMP).

Non-Parametric Methods

Non-parametric methods for stochastic infectious disease modesl had received relatively little attention in literature until recently; for example:

- Becker and Yip (1989) and Becker (1989) considered non-parametric estimation of a time-dependent infection rate in SIR models; [estimating equations, martingales, assumed infection times known].
- Lau and Yip (2008) assumed only removal times are observed, and used a kernel estimator to estimate the unobserved process of infectives; [assumed that the parameter of the infectious period distribution is known].
- Chen and others (2008) considered kernel estimation to estimate the infection rate in a large-scale epidemic model; [the depletion of susceptibles was ignored].
- More recent work during COVID-19 [primarily for ODE models].

Why Non-Parametric?

In this talk we will advocate a non–parametric approach. Why?

Because such an approach:

- helps to avoid erroneous conclusions . . .
- . . . and biased results arising from the use of parametric models with (perhaps) inappropriate assumptions.
- Offers great modelling flexibility.
- Allows the data to speak for themselves.

Why Non-Parametric?

In this talk we will adopt a non–parametric approach. Why?

Because such an approach:

- helps to avoid erroneous conclusions ...
- ... and biased results arising from the use of parametric models with (perhaps) inappropriate assumptions.
- Offers great modelling flexibility.
- Allows the data to speak for themselves.

Disclaimer

Parametric models are of great value!

Roadmap – Homogeneously Mixing Populations

Aspects of Epidemic Modelling

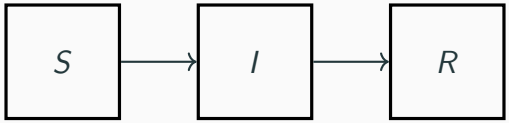
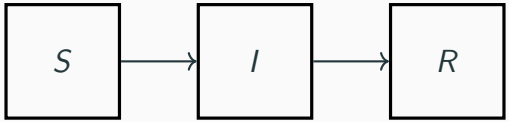


Figure 1: A compartmental SIR model.

- infection process;
- removal process;
- individual's infectiousness;
- population structure;
- ...

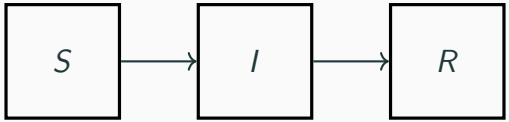
Homogeneously-mixing S-I-R Model



- Underlying assumptions:
  - **S** $\rightarrow$ **I**: New infections occur at the points of a time non–homogeneous Poisson process with rate, for example,
$$\beta S_t I_t$$
  - **I** $\rightarrow$ **R**: Infectives become removed after an infectious period which has an Exponential distribution with rate γ
$$R_i - I_i \sim \text{Exp}(\gamma)$$

[**GOAL**: Infer β and γ]

Homogeneously-mixing S-I-R Model



- Underlying assumptions:
  - **S** $\rightarrow$ **I**: New infections occur at the points of a time non–homogeneous Poisson process with rate, for example,

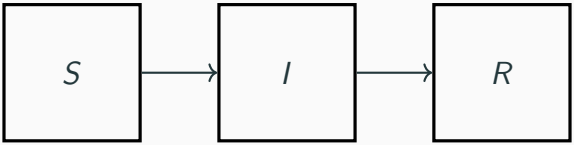
$$\beta S_t I_t \quad \text{or} \quad \beta S_t I_t^{\delta} \quad \text{or} \quad \beta S_t^{\delta_1} I_t^{\delta_2} \quad \text{or} \quad \ldots$$

  - **I** $\rightarrow$ **R**: Infectives become removed after an infectious period which has an arbitrary, but specified distribution, for example:

$$\text{Exp}(\gamma) \quad \text{or} \quad \text{Gamma}(\mu, \nu) \quad \text{or} \quad \text{Weibull}(\mu, \nu) \quad \text{or} \quad \ldots$$

[**GOAL**: Infer β, γ, μ, ν, δ, δ_1, δ_2]

Non-Parametric Estimation: Main Idea



- Underlying Assumptions:
  - **S** $\rightarrow$ **I**: New infections occur at the points of a time non–homogeneous Poisson process with rate
$$h(t) > 0 \quad (t \in \mathbb{R})$$
  - **I** $\rightarrow$ **R**: Infectives become removed after an infectious period which has an arbitrary, but specified distribution, for example:
$$\text{Exp}(\gamma) \quad \text{or} \quad \text{Gamma}(\mu, \nu) \quad \text{or} \quad \text{Weibull}(\mu, \nu) \quad \text{or} \quad \ldots$$

[**GOAL**: Infer $h(t)$, γ, μ, ν, δ, δ_1, δ_2]

Bayesian Inference

- We can infer $h(t)$ from data within Bayesian framework.
- We first place a prior distribution over the function $h(t)$;
  [Gaussian Process; B-Splines, Piecewise-constant];
- We develop an efficient Markov Chain Monte Carlo algorithm to sample from the target distribution:

$$\pi[h(t), \text{ set of infection times} \mid \text{observed data}];$$

- The algorithm is *exact* $--$ infer the whole *latent* history of the generative (time non-homogenous Poisson) process.

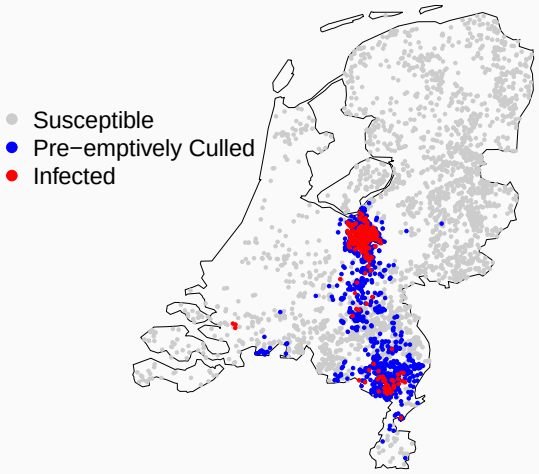
References for Homogeneously Mixing Models

- Xu, X., Kypraios, T. and O'Neill, P.D. (2016) Bayesian nonparametric inference for stochastic epidemic models using Gaussian Processes. *Biostatistics*, 17(4):619-633.
- Hensman, J. and Kypraios, T. (2016) Variational Bayesian Non-Parametric Inference for Infectious Disease Models in *Machine Learning for Healthcare*, IET.
- Knock, E. and Kypraios, T. (2016) Bayesian non-parametric inference for infectious disease data. `https://arxiv.org/abs/1411.2624`
- Kypraios, T. and O'Neill (2018) Bayesian nonparametrics for stochastic epidemic models. *Statistical Science*, 33(1): 44–56.

Heterogeneously Mixing Populations

Avian Influenza A/H7N7 in the Netherlands

- $N = 5359$ bird farms; $n_I = 241$ were confirmed to be infected.
- One veterinarian died; non-fatal infection of 89 others; culling of over 300 million birds.
- Over 1200 farms with an unknown infection status were pre-emptively culled.



Avian Influenza A/H7N7 in the Netherlands

Data available: Culling dates; farm location.

ID	Coordinates	Status	Culling Date	Type
1	(5.32, 18.82)	Susceptible	*NA*	Turkey
2	(2.90, 15.67)	Susceptible	*NA*	Turkey
3	(2.86, 17.99)	Pre-Emptively Culled	3^{rd} May	Broiler
4	(4.56, 18.01)	Culled	30^{th} April	Duck
⋮	⋮	⋮	⋮	⋮

Questions of interest:

- Is there a spatial element to the spread of the disease?
- Can we include other information, such as farm type?

Homogeneously Mixing $\rightarrow$ Heterogeneously Mixing

It does not make sense assume that the rate of infection between a susceptible farm j and infective farm i is β, i.e. the same regardless where farms i and j are and/or their characteristics.

Instead, it makes sense to consider models where the infection rate between an infective farm i and susceptible farm j depends on the farm's characteristics:

$$\beta_{ij} = \textbf{function}\ (d_{ij}, \text{size}_i, \text{size}_j, \text{type}_i, \text{type}_j, \ldots)$$

Parametric infection rates

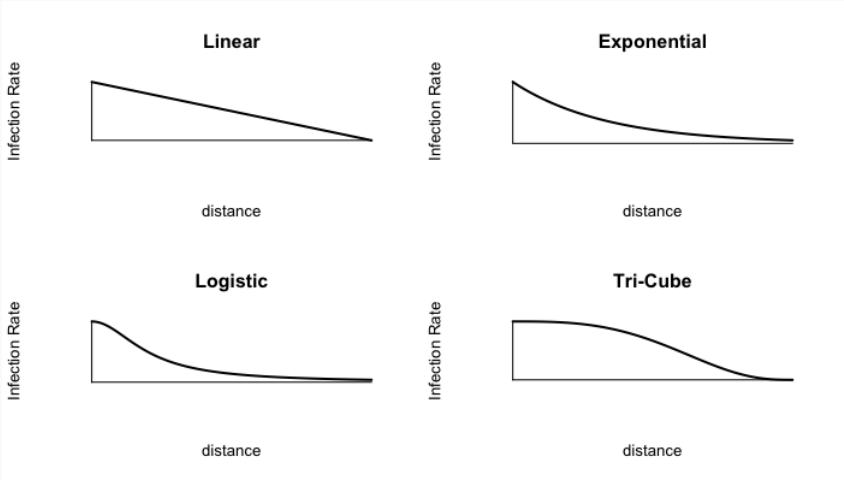


Figure 2: Four Possible Parametric Kernels.

Avian Influenza A/H7N7 in the Netherlands: Previous Work

Boender *et al.* (2007) considered the following transmission models (stochastic; discrete-time):

Model	Infection Rate
1	$\beta_{i,j} = \beta_0$
2	$\beta_{i,j} = \frac{\beta_0}{1+d_{i,j}}$
3	$\beta_{i,j} = \frac{\beta_0}{1+d_{i,j}^2}$
4	$\beta_{i,j} = \frac{\beta_0}{1+d_{i,j}^\alpha}$
5	$\beta_{i,j} = \frac{\beta_0}{1+(d_{i,j}/\beta_1)^\alpha}$

Approach Taken

Assume that each farm remained infectious for 7.5 days $\rightarrow$ Generalised Models $\rightarrow$ MLE $\rightarrow$ choose best model using AIC.

Estimating the Infection Rate Non-Parametrically

- It can be difficult to propose parametric functions given the observed data.
- Specific functional forms are often based on strict assumptions about β_{ij}.
- Hence, we do not want to assume a specific functional form (eg. one of the forms shown in the previous Table).
- Instead, we want to assume that

$$\beta_{ij} = \text{a function of } d_{ij} = f(d_{ij})$$

and estimate $f(\cdot)$ non-parametrically within a Bayesian framework.

Estimating the Infection Rate Non-Parametrically: How?

Outline of the approach

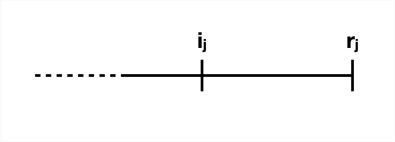
1. Assign a Gaussian Process prior to $\log f(\cdot)$.
2. The likelihood of the observed data (culling dates and farm locations) given the infection rate function $f(\cdot)$ and parameters associated with the infectious period distribution is intractable.
3. Augment the data with the unobserved infection times as well as the unknown status of the pre-emptively culled farms.
4. Develop efficient MCMC algorithms to sample from the posterior distribution.

Note

Step 4. is not a standard problem

Likelihood function

To construct the likelihood function, we first consider the contribution of one individual j.



It contributes to the likelihood is several ways:

- By avoiding infection up to time i_j ,
- By becoming infected at time i_j , and
- By being infectious until r_j .

(Augmented) Likelihood Function

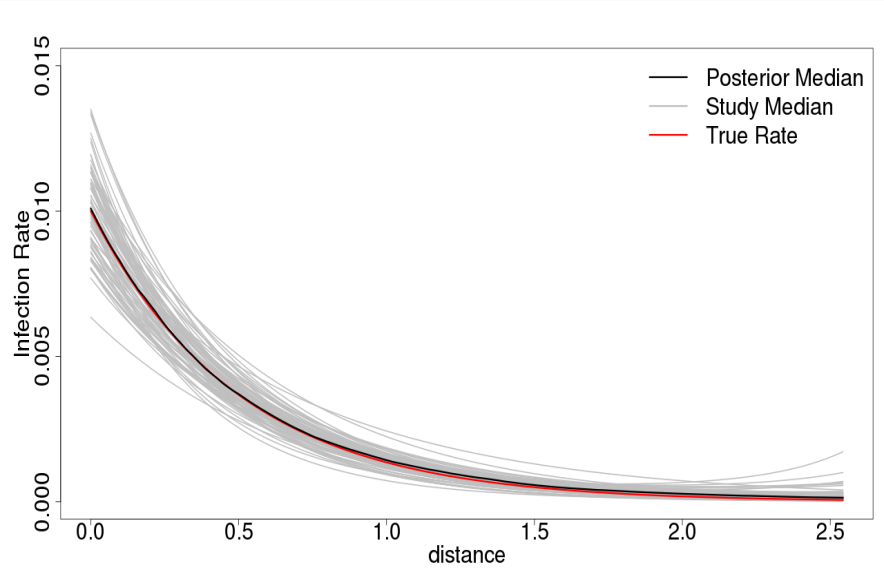
The augmented likelihood function for this model is given by

$$\pi(\mathbf{i}, \mathbf{r}|\boldsymbol{\beta}, \gamma) \propto \underbrace{\exp\Big(-\sum_{j=1}^{n}\sum_{k=1}^{N}\beta_{j,k}\big((r_j \wedge i_k) - (i_j \wedge i_k)\big)\Big)}_{\text{Avoiding infection}}$$
$$\times \underbrace{\prod_{\substack{j=1\\ j\neq\kappa}}^{n}\Big(\sum_{k\in\mathcal{Y}_j}\beta_{k,j}\Big)}_{\text{Becoming infectious}}$$
$$\times \underbrace{\prod_{j=1}^{n} g(r_j - i_j|\gamma)}_{\text{Remaining infected}}.$$

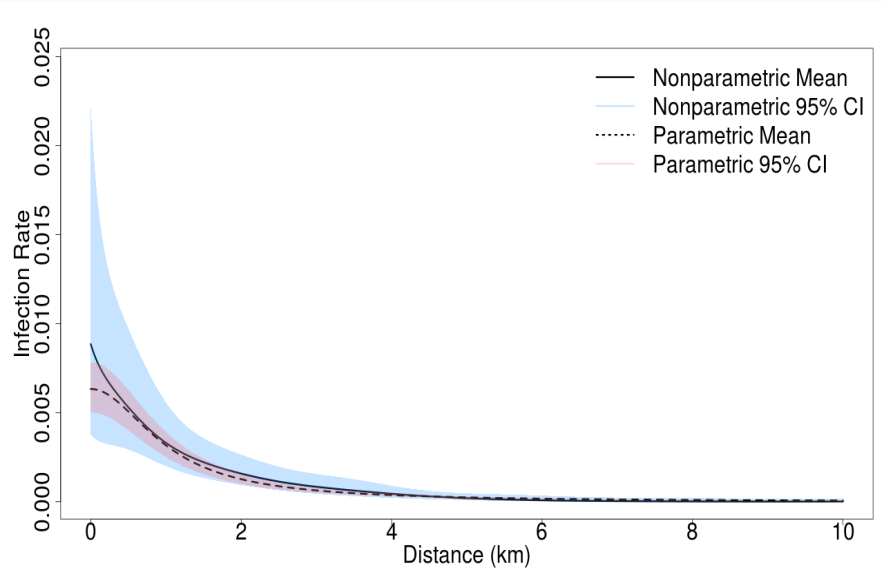
where $g(\cdot)$ is the pdf of the infectious period distribution.

Does this really work?

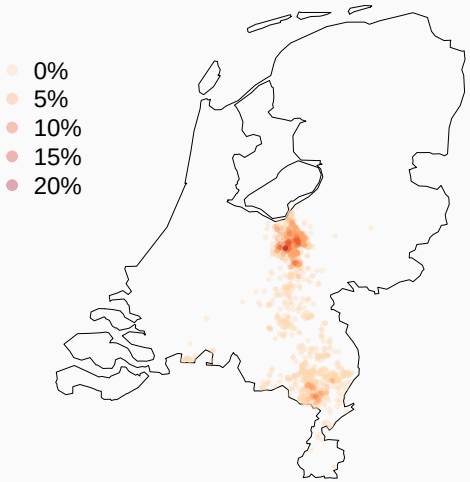
Simulation Study: $N = 1000$, $\beta_{ij} = 0.01 \cdot \exp\{-2.2 d_{ij}\}$



Avian Influenza: Results (1)



Avian Influenza: Results (2)



Estimate posterior probability of pre-emptively culled farms being infected.

Avian Influenza Results: Culling Strategies

Table 1: Posterior predictive medians (95% probability intervals) for the number of infected and culled farms and the amount of compensation paid.

Radius (km)	Infected Farms	Culled Farms	Compensation (€mil)
0	443 (151, 644)	443 (151, 644)	24.8 (8.62, 35.9)
1	297 (110, 535)	489 (215, 709)	27.2 (12.2, 38.9)
2	283 (108, 608)	488 (217, 740)	27.5 (12.2, 41.7)
3	283 (112, 582)	517 (242, 775)	29.0 (13.2, 43.1)
4	274 (105, 564)	512 (228, 793)	28.5 (12.3, 43.9)
5	280 (109, 549)	527 (226, 797)	39.2 (12.4, 41.9)

Posterior Predictive Distribution for Inform Policy Making

True model: $\beta_{ij} = 0.7 \exp\{-0.7 d_{ij}\}$

Table 2: The results of a culling strategy

Model	Infection function (β_{ij})	Mean final size	Probability of a severe outbreak
M_1 (Exponential)	$\theta_1 \exp\{-\theta_2 d_{ij}\}$	327	0.790
M_2 (Logistic)	$\lambda_1/(1 + d_{ij})$	555	0.658
M_3 (BNP)	$\exp(f(d_{ij}))$	303	0.796

Posterior Predictive Distribution for Inform Policy Making

Table 3: Assessing disease control strategies: results of the ring-culling strategy and time taken to run the MCMC algorithm.

Model	Infection function (β_{ij})	Mean final size	Severe outbreak prob.
M_1	$0.3 \times \frac{\theta_1}{\theta_2+(d_{ij}-\theta_3)^2} + 0.7 \times \frac{\theta_1}{\theta_4+d_{ij}}$	370	0.634
M_2	$\frac{\lambda_1}{\lambda_2+d_{ij}}$	575	0.609
M_3	$\nu_1 \exp(-\nu_2 d_{ij})$	402	0.450
M_4	$\frac{\sigma_1}{\sigma_2+d_{ij}^2}$	274	0.645
M_5	$\frac{\psi_1}{\psi_2+(d_{ij}-\psi_3)^2}$	391	0.511
M_6	BNP	362	0.590

Model Extensions

Foot-and-Mouth Disease (FMD) in the UK

There was a large outbreak of Foot and Mouth Disease (FMD) in the UK in 2001.

In Cumbria, which was the most affected rea, there were 5,436 farms consisting of $N_1 = 1,061$ sheep farms, $N_2 = 1,064$ cattle farms, and $N_3 = 3,253$ farms with both sheep and cattle.

Out of 5,436 initially susceptible farms, $1,021 (= n)$ were infected.

Foot-and-Mouth in Cumbria

Incorporating More Information: Farm Type

We consider farms of different type, i.e. assume that each farm is of type k, where $k = 1, \ldots, m$.

Multitype susceptibility models

Farms have varying susceptibility to the disease, but are assumed to be equally infectious if infected.

$$\beta_{ij} = \begin{cases} f_1(d_{ij}) & \text{if } j \text{ is of type 1} \\ f_2(d_{ij}) & \text{if } j \text{ is of type 2} \\ \quad \vdots & \\ f_k(d_{ij}) & \text{if } j \text{ is of type } k \end{cases}$$

Multi-output Covariance (MOC) Model

We place a joint prior distribution on the functions $f_1(\cdot), f_2(\cdot), \ldots, f_p(\cdot)$:

$$\beta^{(j)} = \exp\left(f^{(j)}\right), \quad j = 1, \ldots, p$$

$$\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_p \end{pmatrix} \sim \mathcal{GP}\left(0, \begin{pmatrix} \Sigma^{(1,1)} & \ldots & \rho_{1,p}\Sigma^{(1,p)} \\ \rho_{2,1}\Sigma^{(2,1)} & \ldots & \rho_{2,p}\Sigma^{(2,p)} \\ \vdots & & \vdots \\ \rho_{p,1}\Sigma^{(p,1)} & \ldots & \Sigma^{(p,p)} \end{pmatrix}\right),$$

[We assume all covariance functions have the same length scale hyper-parameter.]

Independent GP (IGP) model

- Setting $\rho_{j,k} = 0$ for all j and k gives rise to an *independent GP* model.
- It is assumed that there is no relationship between the infection rate acting on different types of individuals *a priori*.
- An advantage of this model is its simplicity, no need to specify the relationship between f_j and f_k.

[We allow the p independent GPs to have their own length scales.]

Discrepancy-Based (DB) Model

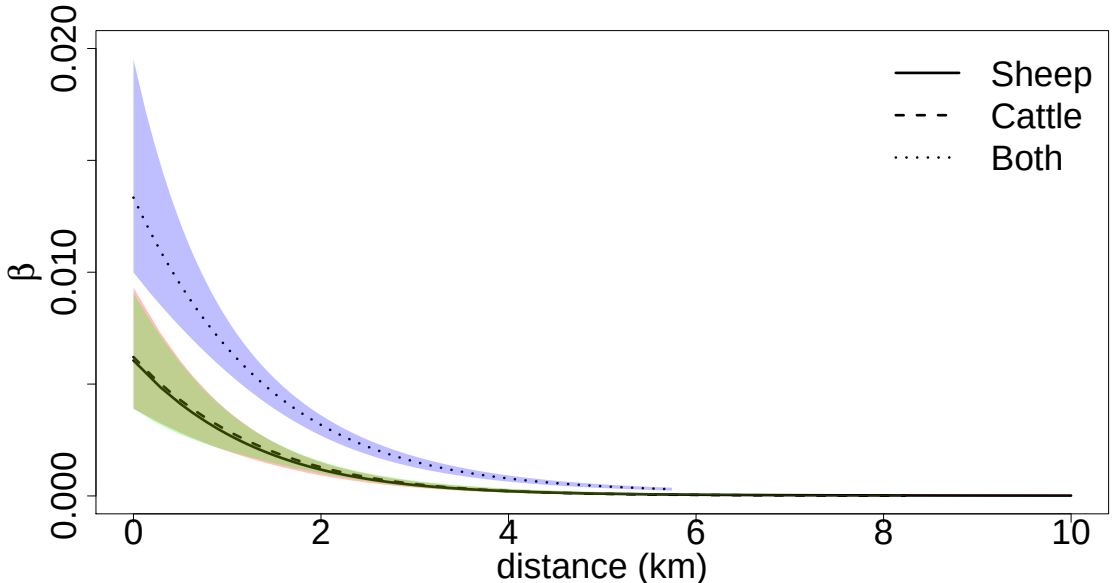
- In the *Discrepancy-Based model* we first set f_1 as a baseline, to which we assign a GP prior with mean zero and covariance matrix $\Sigma^{(1)}_{j,k}$.
- For $j = 2, \ldots p$ we then assume that

$$f_j = f_1 + u^{(j)}, \quad u^{(j)} \sim \mathcal{GP}\left(0, \Sigma^{(j)}_{j,k}\right),$$

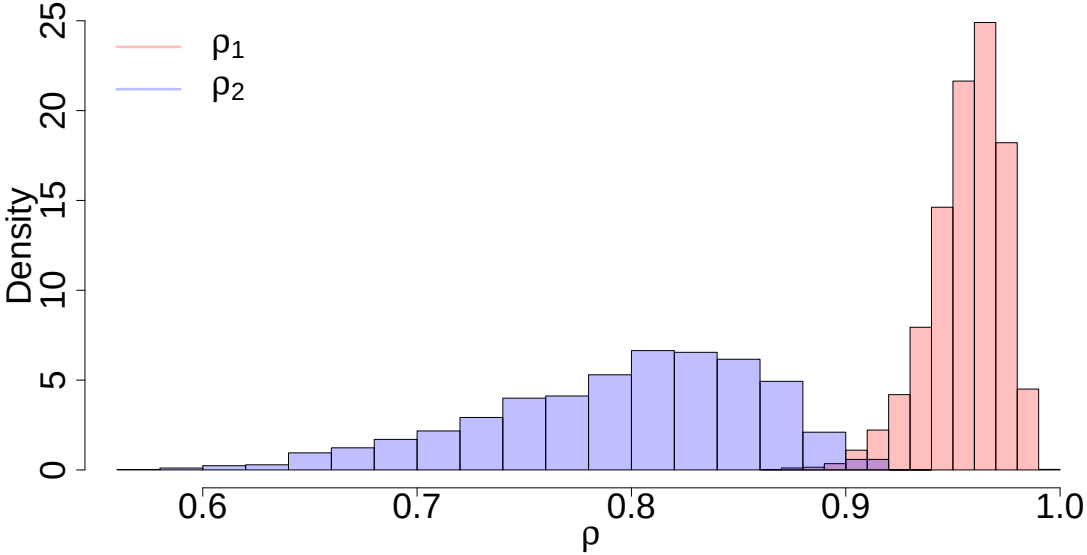
  where $u^{(j)}$ represents the discrepancy between f_j and f_1, with $f_1, u^{(2)}, \ldots, u^{(p)}$ assumed to be mutually independent.
- When fitted to data, this model enables a direct comparison between infection rates of different types of farms to be made, which can be useful for policy makers.

[The discrepancies have their own length scales.]

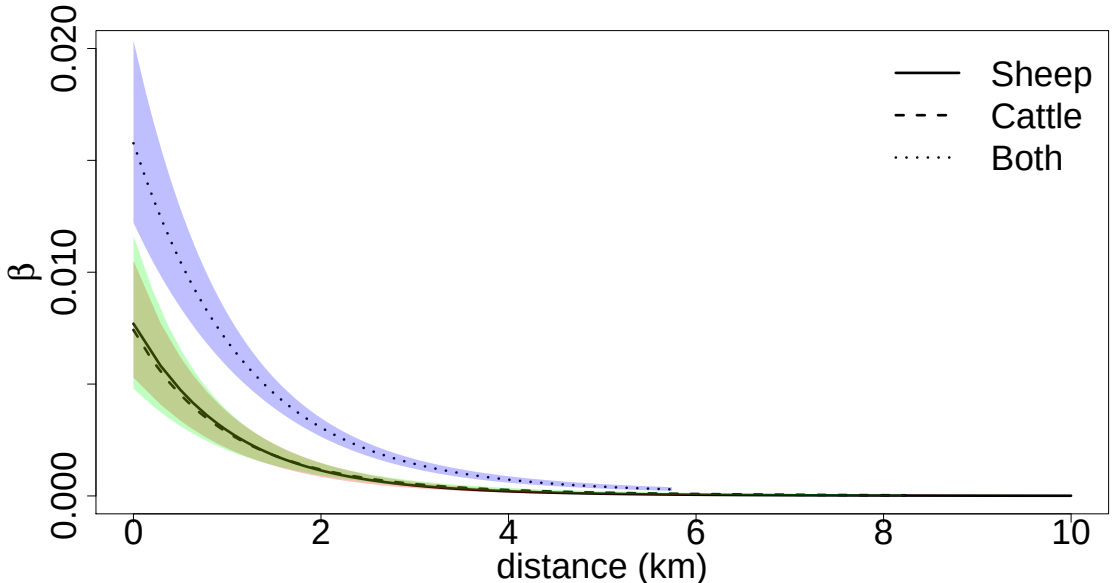
FMD: Multioutput-Covariance Model: MOC (1)



FMD: Multioutput-Covariance Model: MOC(2)



FMD: Discrepancy-Based Model



Challenges

Challenges

- Fitting stochastic epidemic models to data, parametric or non-parametric, is far from trivial.
- Off-the-shelf data augmentation MCMC algorithms struggle in large populations due to the non-linear dependence between infection times and model parameters.
- Computing the target posterior density for a given set of infection and removal times and a function $f(\cdot)$ is computationally expensive (for large N).
    1. There is double sum of order $O(n\,N)$ and any proposed set of infections must be consistent with the observed data.
    2. The posterior density requires evaluation of a MVN pdf $\rightarrow$ computing the inverse of its covariance matrix $\rightarrow$ problematic for $N > 300$.

Mean Projection Approximation

- Mean Projection Approximation works by using a subset ofthe original dataset (set of distances);
- this subset is suitably representative of the original (e.g. its size is sufficiently large and its elements are suitably placed across the entire domain to capture the features of the infection rate function);
- we infer the infection rate function given this subset;
- we then project the result onto the full data set.

Conclusions

Conclusions

- We have presented a general framework for Bayesian nonparametric inference for infection rate functions in individual-level stochastic epidemic models.
- The methodology is applicable to a wide class of epidemic models, including household models, network models and age-structured models.
- Our approach removes the need to make specific parametric assumptions about infection rate functions.
- We have also demonstrated that our approach can be used successfully for large data sets by employing MPA methods, but there is more work to be done in this direction.

References

Seymour, R. G., Kypraios, T., O'Neill, P. D. (2022). Bayesian nonparametric inference for heterogeneously mixing infectious disease models. *Proceedings of the National Academy of Sciences*, 119(10), e2118425119.

Seymour, R.G., Kypraios, T., O'Neill, P.D (2021) A Bayesian Nonparametric Analysis of the 2003 Outbreak of Highly Pathogenic Avian Influenza in the Netherlands. *JRSS, Series C*, 70(5):1323–1343.

Seymour, R.G (2020) Bayesian nonparametric methods for individual-level stochastic epidemic mod els. Phd thesis. University of Nottingham.

Code

`github.com/rowlandseymour/BNP_4_HMSEM`