# Bayesian Inference and Model Choice for Nonlinear Stochastic Processes:

## Applications to stochastic epidemic modelling

Theo Kypraios
http://www.maths.nott.ac.uk/∼tk

School of Mathematical Sciences, University of Nottingham

EPSRC Symposium Workshop on Markov Chain-Monte Carlo,
Warwick , March 2009

Joint work with:
Gareth O. Roberts @ University of Warwick
Phil O'Neill @ University of Nottingham
Ben Cooper @ Health Protection Agency

# Background

Understanding the spread of an infectious disease is an important issue in order to prevent major outbreaks of an epidemic.

In general, inference problems for disease outbreak data are complicated by the facts that

- the data are inherently dependent and
- the data are usually incomplete in the sense that the actual process of infection is not observed.

However, it is often possible to formulate simple stochastic models which describe the key features of epidemic spread.

# Outline

- Bayesian inference for partially observed stochastic epidemics
- Efficient Markov Chain Monte Carlo algorithms (including *Non-Centered* reparameterisations)
- Bayesian Model Choice using RJMCMC and other alternative methods.
- Illustration simulation study.

# The SIR Model

**Principles**

– Consider a *closed* population (no births/deaths) of size $\mathcal{N}$ individuals

– $\alpha$ initially infected individuals

– at any given time point, each individual *i* is in one of the three states:

- **S**usceptible
- **I**nfected
- **R**emoved

$$\mathbf{S} \longrightarrow \mathbf{I} \longrightarrow \mathbf{R}$$

# The Stochastic Process

- Each of the infected individuals remains infectious for some period which is distributed according to the distribution of random variable $D$.

- While infectious, makes contacts with each of the $\mathcal{N}$ individuals in the population at times given by the points of a Poisson process of rate $\beta > 0$.

- Any such contact with a susceptible individual immediately makes the susceptible an infective.

- At the end of its infectious period an infective no longer makes any contacts and is said to be removed.

# Infectious Period & Generalisations of the GSE

- The infectious periods of different infectives are assumed to be independent and identically distributed according to the distribution of a random variable $D$,...
- ... which can have any arbitrary but specified distribution.
- The special (*Markovian*) case where the infectious period follows an Exponential distribution is known as the *General Stochastic Epidemic* (GSE).

Although assuming an Exponential infectious period is mathematically convenient (limit results etc) it is not biologically motivated.

This lead to various generalisations of this particular model (infectious period, SEIR, $\beta_{ij}$ ...)

# A Heterogeneously Mixing Stochastic Epidemic Model

## Model Specification

Infection Rate:

$$\beta_{ij} := \beta_0 \cdot h(i,j) \tag{1}$$

Infectious Period:

$$D_i \sim Ga(\alpha, \gamma) \tag{2}$$

The parameters of interest are $\beta_0, \gamma > 0$ ($\alpha$ assumed to be known).

The (deterministic) function $h(i,j)$ refers to individual's characteristics which might involve some other parameters as well.

# Inference

**GOAL:** Draw inference for the parameters $(\beta_0, \gamma)$

**Likelihood:**

$$f\left(\mathbf{I}, \mathbf{R} | \beta_0, \gamma\right) \propto \prod_{i=1, i \neq k}^{n_I} \left( \sum_{j \in Y_i} \beta_{ji} \right) \times \exp\left\{ - \int_{I_k}^{T} \sum_{i \in \mathcal{I}_{t-}} \sum_{j \in \mathcal{S}_{t-}} \beta_{ij} \, \mathrm{d}t \right\} \times \prod_{i=1}^{n_I} f_D(R_i - I_i)$$

where

- $Y_i = \{j : I_j < I_i < R_j\}$
- $k$ denotes the initial infective.
- $\mathcal{I}_{t^-}$ and $\mathcal{S}_{t^-}$ denote the set of infectives and susceptibles just prior time $t$.
- $\mathbf{I} = (I_1, \ldots, I_{n_I})$ and $\mathbf{R} = (R_1, \ldots, R_{n_R})$
- $f_D(\cdot)$ is the density of the infectious distribution.

# Bayesian Inference for Stochastic Epidemics

In general, inference for disease outbreak data is hard and often requires problem-specific methodology:

- Incompleteness
  - Actual process not observed (eg. infection times),
  - Undetected colonisation (eg. imperfect sensitivity).
- Dependence
  - Correlation between unobserved data and model parameters can cause mixing problems for the MCMC.
- Dimensionality
  - $d_1 = 2$ (model parameters)
  - $d_2$ (missing data - infection times).

# MCMC Algorithms for Stochastic Epidemics

Gibson and Renshaw (1998) & Roberts and O'Neill (1999) were the first to apply MCMC methods for stochastic epidemics.

Outline of the MCMC (centered) algorithm:

1. Initialise the algorithm;
2. Choose uniformly one (or more) infection time(s) $I_j, j = 1, \ldots, n_I$ and update it (them) individually using a Metropolis Hastings step by proposing a replacement infection time;
3. Update the parameters $\beta_0$ and $\gamma$ via a Gibbs step.

# MCMC Implementation

## Terminology:

Random Scan (RS): At each MCMC iteration we update only one of the infection times (*chosen at random*).

$\delta$% Deterministic Scan ($\delta$% DS): At each MCMC iteration, we choose (at random) to update $\delta$% of the infection times.

## Computation:

The integral involved in the likelihood can be rewritten as follows:

$$\int_{I_k}^{T} \sum_{i \in \mathcal{I}_{t-}} \sum_{j \in \mathcal{S}_{t-}} \beta_{ij} \, \mathrm{d}t = \sum_{i=1}^{n_I} \sum_{j=1}^{\mathcal{N}} \beta_{ij} (R_i \wedge I_j - I_j \wedge I_i).$$

This reduces the computational cost of evaluating the likelihood significantly.

# Remarks

$-$ (Efficient) Block updating of the infection times is very hard especially when the size of the data $(n_I, \mathcal{N})$ is big.

$-$ Gibbs step for the infection times is possible, but difficult to construct.

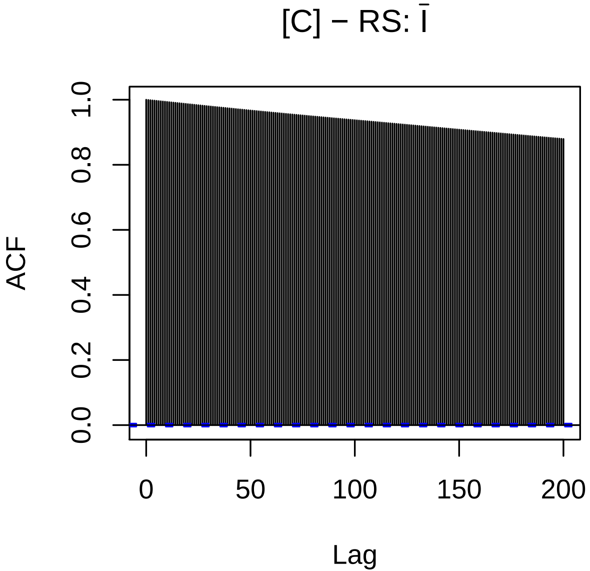
$-$ The independence sampler

$$R_i - I_i \sim Ga(\alpha, \gamma)$$

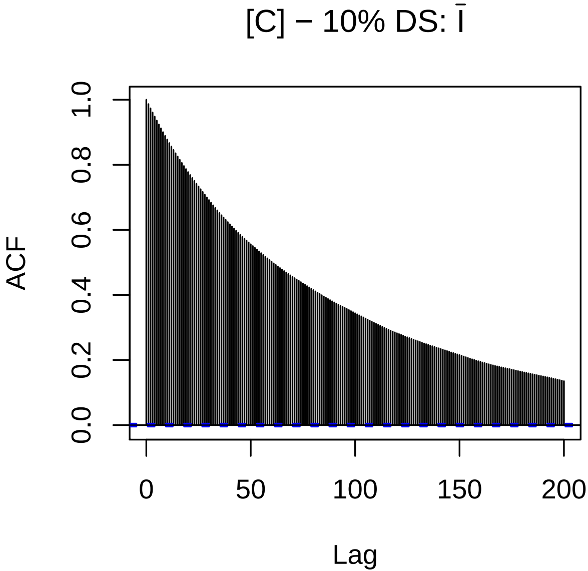
seems to be the most efficient proposal among the previously mentioned sampler.

# Random Scan vs Deterministic Scan

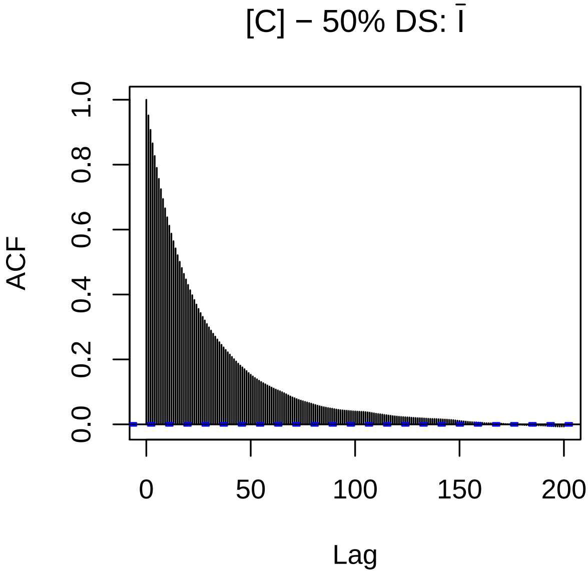
# Updating the infection times I



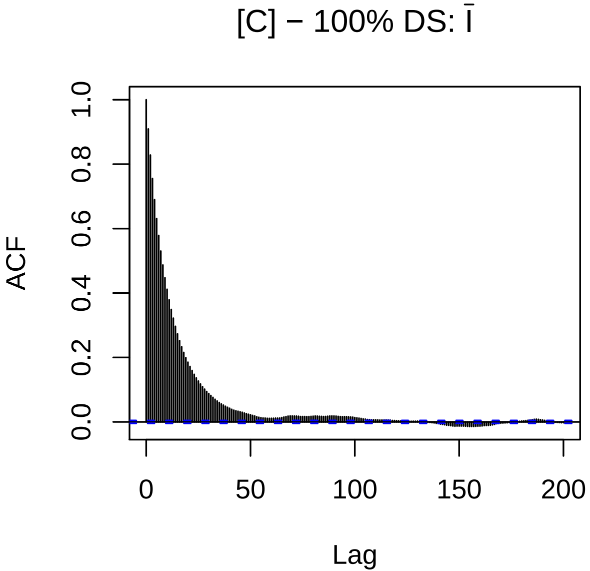
# Updating the infection times $\mathbf{I}$ (cont.)



# Updating the infection times $\mathbf{I}$(cont.)



# Updating the infection times **I**(cont.)



# Comments

- Ignoring the cpu time needed to run an algorithm, 100% Deterministic scan performs better than any other chosen algorithm (ie. $\delta < 1$).
- In general, increasing $\delta$ improves the mixing of the MCMC algorithm, nevertheless is more computationally costly (slower in cpu time).
- It is of interest to quantify the *optimal* percentage of deterministic scan taking into account mixing and cpu time.
- The optimal $\delta$ will depend on the size of the epidemic $(n_I, \mathcal{N})$. Easy to do it computationally $-$ hard to derive a theoretical .

# Alternative Target Distributions – Any better?

- Due to the conditional independence of $\beta_0$ and $\gamma$ (given the infection and removal times), $\beta_0$ or $\gamma$ (or even both) could be integrated out.
- It turns out that if $\delta$ is chosen to be relatively small, then pretty similar results (regarding efficiency) are obtained.
- However ...

  ... If we increase $\delta$, and in particular we choose a 100% DS, integrating both model parameters out (i.e. $\beta_0$ and $\gamma$) and choose as target distribution

  $$\pi(\mathbf{I}|\mathbf{R})$$

  seems to increase (slightly) the algorithm's efficiency compared to the standard target distribution ($\pi(\mathbf{I}, \beta_0, \gamma|\mathbf{R})$).

# A Working Example

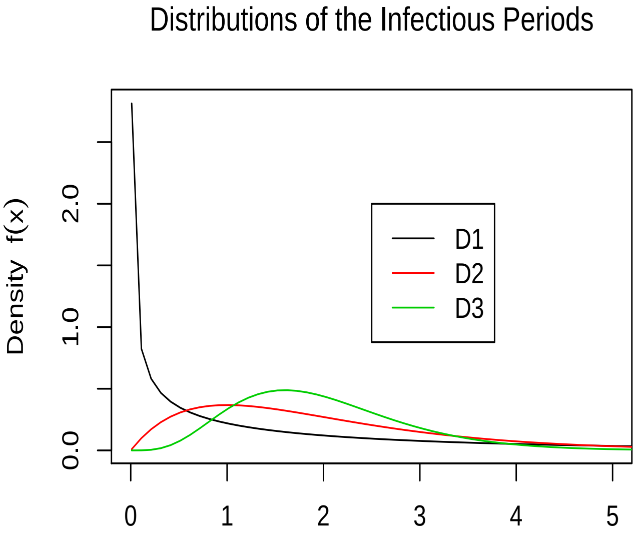
## Model's Equations

$$\begin{aligned} \beta_{ij} &= \beta_0 \exp\{-\rho(i,j)\} \\ R_i - I_i &\sim \text{Gamma}(\alpha,\gamma) \\ \mathcal{N} &= 501 \text{ (1 initially infective)} \end{aligned}$$

where $\rho(i,j)$ denotes the distance between the two individuals.

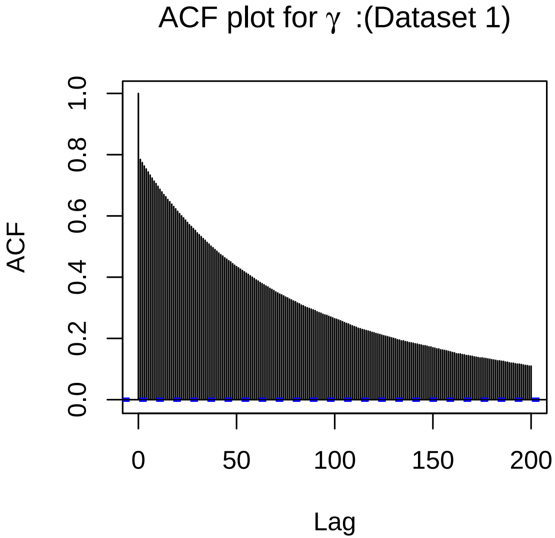
| **Simulated Dataset** | **D1** | **D2** | **D3** |
|---|---|---|---|
| True $\beta_0$ | 0.0025 | 0.0025 | 0.0025 |
| True $\alpha$ | 0.5 | 2 | 5 |
| True $\gamma$ | 0.25 | 1 | 2.5 |
| $n_I = n_R$ | 284 | 275 | 286 |

# A Working Example (cont.)



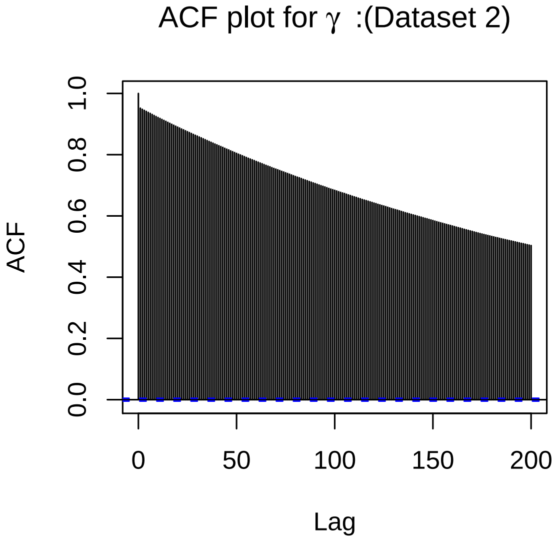
# Mixing of the MCMC Algorithm

$(\alpha, \gamma) = (0.5, 0.25)$



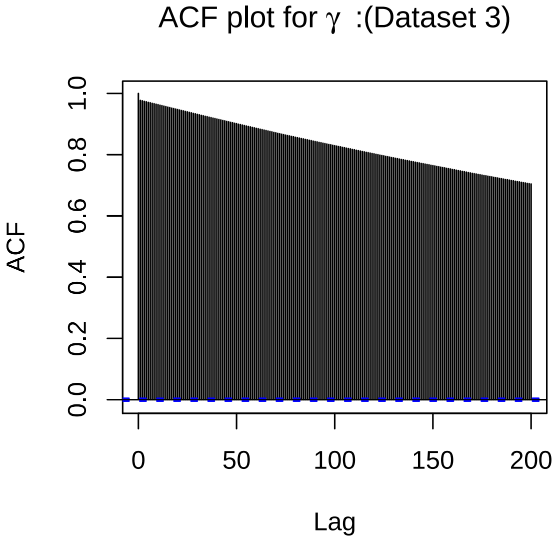
# Mixing of the MCMC Algorithms (cont.)

$(\alpha, \gamma) = (2, 1)$



# Mixing of the MCMC Algorithms (cont.)

$(\alpha, \gamma) = (5, 2.5)$



# Reasons for Poor Mixing

If we carefully look at one of the likelihood equations which describes the infectious period of each individual:

$$R_i - I_i \sim Ga(\alpha, \gamma).$$

This reveals that *a-priori* $\gamma$ and $\mathbf{I}$ are dependent and this causes problems to the MCMC mixing.

$$\begin{aligned} R_i - I_i &\sim Ga(\alpha, \gamma), \text{ for } i = 1, \ldots, n_I \\ \sum_{i=1}^{n_I} (R_i - I_i) &\sim Ga(\alpha n_I, \gamma) \end{aligned}$$

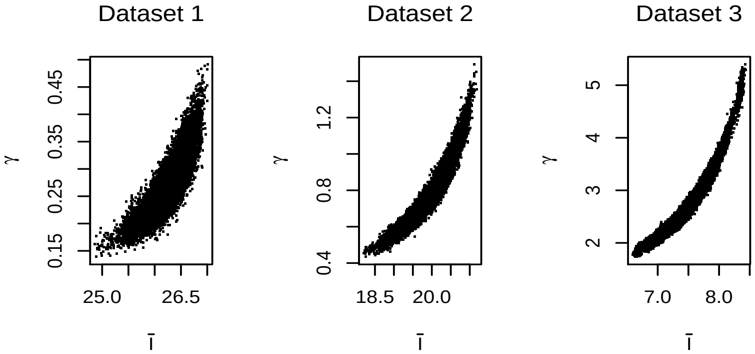
# Reasons for Poor Mixing (cont.)

Thus for large $n_I$ or $\alpha$, the parameter $\gamma$ and the sum of the infectious periods $\sum_{i=1}^{n_I} (R_i - I_i)$ are *a-priori* heavily dependent.

If these two were the parameters of interest, then this *a-priori* correlation would have caused mixing problems in the case of a two-state Gibbs sampler [see Amit(1991), Roberts and Sahu (1997)].

Things are more complicated in practice since the MCMC schemes used so far involved deterministic scan update of the each of infection times $I_i$, $i = 1, \ldots, n_I$.

# Correlation between Infection times and $\gamma$



# Non−Centered Parameterisations (NCP)

The non-centered methodology presented in Papaspiliopoulos *et. al* (2003) suggests that appropriate *non−centered* parameterisations out perform the centered existing centered algorithms over a range of examples.

Their findings can be summarized as follows:

> *When the observed (missing) data are much more informative about the parameter of interest than the missing (observed) then is better to choose a centered (non−centered) parameterisation.*

# NCP for Stochastic Epidemics

We introduce a non−centered parameterisation:

$$U_i = \gamma(R_i - I_i), i = 1, \ldots, n_I$$

It easy to see that since $U_i \sim Ga(\alpha, 1)$,$i = 1, \ldots, n_I$, the random variables $\mathbf{U} = (U_1, \ldots, U_{n_I})^T$ and $\gamma$ are *apriori* independent.
*Change in variables*:

$$(\mathbf{I}, \beta_0, \gamma, \mathbf{R}) \rightarrow (\mathbf{U}, \beta_0, \gamma, \mathbf{R}).$$

**Implementation**

1. Get a sample of $(\gamma, \mathbf{I})$ via a centered algorithm;
2. Transform the $\mathbf{I}$'s to $\mathbf{U}$'s via

$$U_i = \gamma(R_i - I_i)$$

3. Update $\gamma$ using Metropolis Hastings algorithm using $\pi(\gamma|\mathbf{U}, \beta_0, \mathbf{R})$ ;

# (P)NCP for Stochastic Epidemics (2)

1. Draw samples of $(\gamma, \mathbf{I})$ via MCMC on $\pi(\gamma, \mathbf{I}|\mathbf{R})$ → update $\gamma$ using RwM (Neal and Roberts, 2005)
2. Draw samples of $(\gamma, \mathbf{I})$ via MCMC on the most efficient standard algorithm (e.g. on $\pi(\mathbf{I}|\mathbf{R})$) → update $\gamma$ using various independence MH sampler (Kypraios 2007)

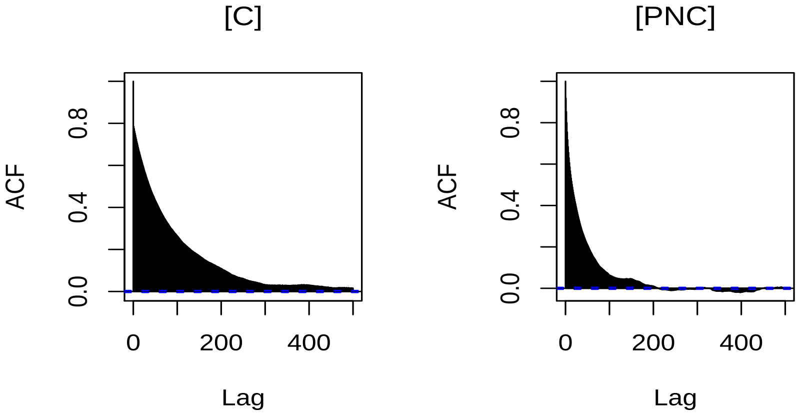
## Partially Non−Centered Parameterisations

The set of the infected individuals in the epidemic, is partitioned into two groups, $\mathcal{C}$ and $\mathcal{U}$ (assign with probability $\mu$).

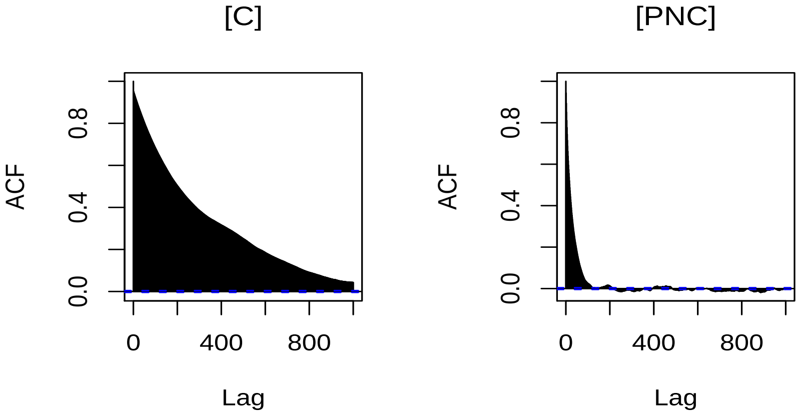
For those individuals in the $\mathcal{U}$, let

$$U_i = \gamma(R_i - I_i) \quad (i \in \mathcal{U})$$

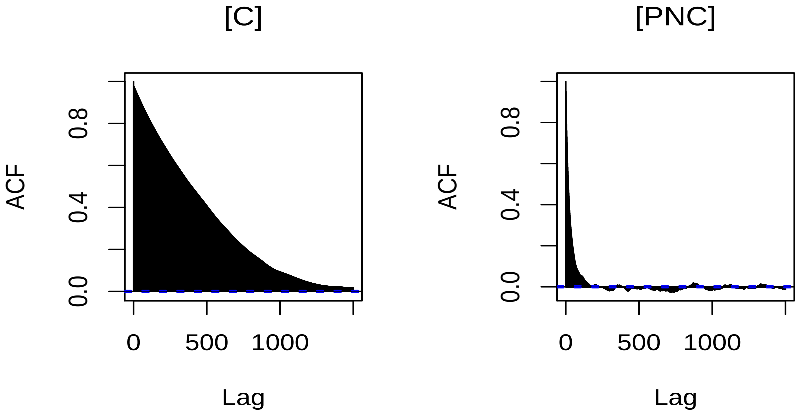
# Results - Dataset D1



# Results - Dataset D2



# Results - Dataset D3



# Bayesian Model Choice in Stochastic Epidemics

# Model Choice (1)

Motivated by applications in healthcare associated infections (MRSA etc) various models have been proposed (Kypraios *et al* 2009):

1. $M_1 : \lambda_j = \beta_0$
2. $M_2 : \lambda_j = \beta_0 + \beta_1 n_C + \beta_2 n_Q$
3. $M_3 : \lambda_j = \beta_0 + \beta_1 \mathbb{I}(n_C > 0) + \beta_2 \mathbb{I}(n_Q > 0)$

Question: Is there evidence to support the assumption of linear colonisation pressure?

- Bayesian Model Choice.
- Posterior Model Probabilities - Bayes Factors
- Within-Model prior distributions and Lindley's paradox : Prior's Matching & Prior Senstivity
- RJMCMC

# Matched Prior Distributions

The idea is to match the prior distributions for the two models by trying to make the pressure experienced by a susceptible individual similar in both models.

Assign a prior distribution to $\beta_1^F$ say, and then derive the prior distribution for $\beta_1^S$ by matching the moments of the prior distributions:

$$E[\lambda^F] = E[\lambda^S]$$
$$V[\lambda^F] = V[\lambda^S]$$

Easy to derive the "matched priors" when assuming Exponential priors.

# Computing Marginal Likelihoods?

{Work in progress with Tony Pettit and Nial Friel}

When the number of models to explore is not too big, it is worth seeking for alternative methods to RJMCMC.

The idea of power posteriors (Friel and Pettit, 2008) seems very promising when infection times are assumed to be known.

$$\log\{\pi(\mathbf{Y})\} = \int_0^1 E_{\boldsymbol{\theta}|\mathbf{y},t}[\log\{L(\mathbf{y}|\theta)\}]\ \mathrm{d}t$$

where $\pi(\boldsymbol{\theta}|y,t) \propto L(\mathbf{y}|\boldsymbol{\theta})^t \pi(\boldsymbol{\theta})$, i.e. the power posterior.

Significant improvements have been made to the original idea of power posterior but there is still more work required when infection times are treated as unknown.

# Conclusions

- Although "easy" to implement, standard (centered) MCMC algorithm for SIR-type epidemic models perform poorly as the number of infected individuals $(n_I)$ increases.
- Non−Centered Parameterisations often offer much more robust algorithms, there is relatively little extra cost in implementing such an algorithm.
- "Matched-Priors" in the model choice context offer an easy way towards avoiding Lindleys paradox.
- Power posteriors could be used to numerically evaluate marginal likelihood in the epidemic modelling context.