Estimation of Single Locus Effects on Susceptibility, Infectivity and Recovery Rates in an Epidemic Using Temporal Data

C. M. Pooley[1], **S. C. Bishop**[1], **G. Marion**[2]
[1]The Roslin Institute & R(D)SVS, University of Edinburgh, Midlothian, UK, [2]Biomathematics and Statistics Scotland, Edinburgh, UK

ABSTRACT: Epidemic dynamics are modelled using a variant of the SIR (susceptible-infectious-recovered) model. We investigate scenarios in which a single (dominant) locus affects animal susceptibility, infectivity and recovery rates. In particular, we find that genetic differences in susceptibility and recovery can be readily inferred using data from a single epidemic, but that infectivity requires information from multiple or replicated epidemics. Inference from partially observed epidemics was conducted within a Bayesian framework using Markov chain Monte Carlo (MCMC). The method is tested using simulated data generated by applying the Doob-Gillespie algorithm to a suitable epidemic model. Limits in our ability to carry out inference were explored in the case when complete epidemic data is known, and theoretical expressions are presented for expectations of parameter accuracy. The practical utility of the approach is subsequently demonstrated using data for which infection times are uncertain or completely unknown.
Keywords: epidemic; infectivity; susceptibility; Markov chain Monte Carlo

Introduction

Reducing the impact of disease is of fundamental importance not only for the welfare of animals, but also in terms of the economic costs incurred by farmers. With the reduction in the cost of genetic testing, selective breeding based on the presence or absence of given alleles has become increasingly feasible. Identifying which alleles help to reduce disease liability is a challenging problem because data from epidemics can often be incomplete and the measurements taken from the system are often complex. The aim of this paper is to numerically and analytically investigate the circumstances under which epidemic data can be used to quantify single (dominant) locus effects, and to determine which parameters can be inferred.

Materials and Methods

Model. A common approach to modelling micro-parasitic epidemics is to use the stochastic SIR process (e.g. Keeling and Rohani (2007)) in which the state of the system is described by three integer valued variables S, I and R, i.e. the number of susceptible, infected and recovered individuals within a population. Two possible events can occur probabilistically: an infection (which causes a transition whereby S is reduced by one and I is increased by one) or a recovery (I is reduced and R is increased). In the Markovian implementation inter-event times are exponentially distributed. To incorporate genetic effects into the standard SIR model we consider that there are two subpopulations, labelled + and -, that differ in their phenotypic characteristics.

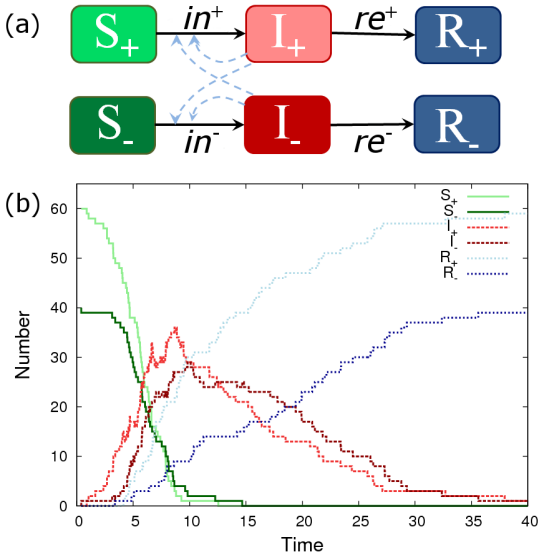

Figure 1: (a) The compartmental model. S, I and R refer to susceptible, infected and recovered, *in* and *re* refer to infection and recovery events, and + and – denote two subpopulations that differ in their phenotype. (b) An example of the system dynamics.

Figure 1(a) shows the proposed model, which consists of two coupled SIR processes. Infection and recovery events (corresponding to the four arrows in the diagram) have rates given by:

$$r_{in^+} = e^{\delta_S/2} F S_+, \quad r_{in^-} = e^{-\delta_S/2} F S_-,$$
$$r_{re^+} = e^{\delta_\gamma/2} \gamma I_+, \quad r_{re^-} = e^{-\delta_\gamma/2} \gamma I_-. \qquad (0.0)$$

Where the infection process is driven by a force of infection

$$F = \frac{\beta}{N}(e^{\delta_I/2} I_+ + e^{-\delta_I/2} I_-), \qquad (0.0)$$

which determines the average rate at which susceptible individuals acquire the disease. F depends on the number of infected individuals within the two subpopulations I_+ and I_- modified by parameter δ_I to differentiate the infectivity of the two genotypes. The exponential dependency on δ_I is used to ensure that F is strictly positive. Its form is chosen such that, for small values at least, δ_I represents the fractional change in infectivity between the subpopulations (e.g. δ_I=*0.1* implies that the + genotype is ca. 10% more infective than the - one). Differences in the susceptibility and recovery rates are similarly incorporated in Eq.(1.1) through parameters δ_S and δ_γ, respectively.

The data in this study was generated using the Doob-Gillespie algorithm, a standard stochastic simulation technique for generating event sequences from Markovian compartmental models. A typical output is shown in Fig. 1(b). Here the initial population is N=*100*, with a fraction

$p_+=0.6$ of the animals having the + genotype. At time $t=0$ one of the animals becomes infected and this animal starts to infect other individuals. In the case shown the process rapidly leads to an explosion in infection (i.e. an epidemic) until the susceptible population is exhausted, at which point the remaining infected animals recover.

Bayesian inference. We take the event data and attempt to estimate the values for the model parameters that could have plausibly generated it (i.e. parameter inference). The starting point for this is the complete likelihood (see Walker et al. (2006) for a derivation):

$$\mathrm{L}(\varepsilon \mid \theta) = \prod_{i=1}^{N_e} r_{\varepsilon_i} e^{-\mathrm{P}(t_{i+1}-t_i)}, \qquad (0.0)$$

which represents the probability that the model produces a particular event sequence ε given a set of system parameters $\theta=(\beta,\gamma,\delta_S,\delta_I,\delta_\gamma)$ that define the event rates in Eq.(1.1), with P being the sum of these rates. This form reflects the structure of the model and in particular the exponentially distributed nature of inter-event times. The product in Eq.(1.3) runs over all N_e events in the sequence ε. The quantities t_i and ε_i represent the time and type (i.e. in^+, re^+, in^-, re^- in Fig. 1(a)) of each event with the corresponding event rate given in Eq.(1.1).

In reality the complete event sequence is unlikely to be available, but within the Bayesian framework inference of both the event sequence and the parameters values from incomplete data y is possible based on the posterior distribution (e.g. Lee (2004)):

$$\pi(\theta,\varepsilon \mid y) \propto \pi(y \mid \varepsilon)\,\mathrm{L}(\varepsilon \mid \theta)\pi(\theta). \qquad (0.0)$$

Prior information about the system parameters is incorporated via $\pi(\theta)$, and $\pi(y|\varepsilon)$ is the probability that observation y is made given the actual system has event sequence ε. We considered three data scenarios for possible observations y made on the system:

1. All events are exactly known.
2. Only recovery events are known. Often, "recovery" can represent the death of animals, so this is pertinent when the only measurable quantity is the time at which animals die.
3. Animals are periodically checked for disease. Infection times, therefore, are not known exactly but are confined. Recovery events are assumed known.

MCMC is used to generate samples from the posterior. Specially designed transitions, which allow the system to explore all of the potential parameter values and event sequences consistent with the observations, are accepted or rejected based on their Metropolis-Hastings probability (Hastings (1970)). This procedure generates a Markov chain in the space of parameters and event sequences which converges to the posterior distribution as the number of steps increases. In this study, an initial 10^4 burn-in steps were discarded and the results presented are based on 10^6 MCMC iterations. The system parameters used were $\beta=1$, $\gamma=0.1$ and $p_+=0.5$, and $\pi(\theta)$ was taken to be an uninformative flat prior.

Results and Discussion

Inference was performed under data scenario 1 using a simulation of $N=500$ animals with parameter values $\delta_S=0.5$ and $\delta_I=\delta_\gamma=0$. The solid line in Fig. 2 shows a typical probability distribution for the inferred values of δ_S. The true value, denoted by the vertical line, falls within this normally shaped profile, indicating that inference has successfully estimated the true parameter's value. Of particular importance in this graph is the width of the posterior distribution, σ_S. If our null hypothesis is that there is no difference in the susceptibility between the two genotypes, this can only be rejected provided δ_S is significantly larger than σ_S. An estimate for σ_S, therefore, provides a useful guide as to how small a phenotypic affect can be observed given a particular set of data.

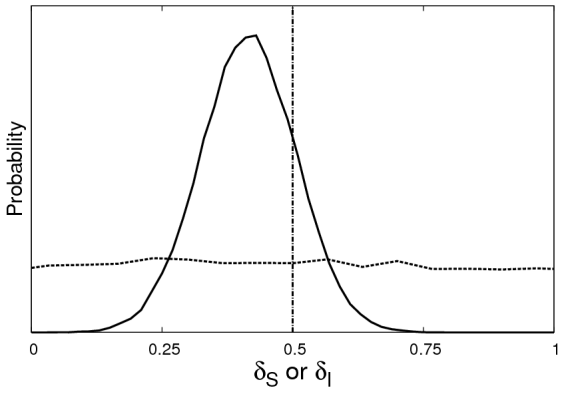

Figure 2: The marginalized posterior distribution for the susceptibility difference δ_S (solid line) and infectivity difference δ_I (dotted line), for a typical epidemic. The vertical dot-dashed line indicates the true parameter value.

This point is emphasized in Fig. 3, in which the crosses denote how σ_S varies with the number of animals. We find that to be able to detect a difference in susceptibility of 10% between the genotypes, it is necessary to have significantly more than 400 animals in the epidemic. Although we do not provide details of the derivation, when the complete epidemic history is known the likelihood Eq.(1.3) can be simplified using the Laplace approximation, and the following approximate analytical expression can be obtained:

$$\sigma_S \cong \frac{1}{\sqrt{Np_+(1-p_+)}}. \qquad (0.0)$$

where p+ represent the proportion of + phenotype in the population. This is represented by the dashed line in Fig. 3, which shows excellent agreement with the MCMC results. Additional analysis yields the following approximations:

$$\sigma_\beta \cong \frac{\beta}{\sqrt{N}}, \quad \sigma_\gamma \cong \frac{\gamma}{\sqrt{N}}, \quad \sigma_I \cong \infty, \quad \sigma_{\delta_\gamma} \cong \sigma_S, \quad (0.0)$$

which are the posterior standard deviations for β, γ, δ_I and δ_γ, respectively.

Numerical results based on MCMC show that our Bayesian framework can reliably infer the parameters β, γ, δ_γ and δ_S, as well as accurately reproduce the posterior standard deviations in Eqs.(1.5) and (1.6). The only parameter that cannot be estimated is the infectivity, as illustrated by the very broad dashed curve in Fig. 2 obtained

using numerical simulation (δ_I=0.5 and δ_S=δ_γ=0) and the fact that σ_I in Eq.(1.6) is infinite. Since the rates of infection for both genotype populations in Eq.(1.1) depends on F, and not on δ_I directly, any change in this force of infection affects both equally. Thus, the only way to measure differences in infectivity is to take data from multiple (or replicated) epidemics. Those epidemics in which there are more of the infective genotype will tend to proceed faster. MCMC simulations show that Bayesian inference can extract information from these correlations. Following the derivation of Eq.(1.5), for the case of multiple epidemics, it can be shown that

$$\sigma_I \cong \frac{1}{\sqrt{NM}\sigma_\chi}. \qquad (0.0)$$

Here, M is the number of epidemics and σ_χ is the standard deviation (between epidemics) in their composition, as defined by χ=½-p_+. Equation (1.7) is supported by inference based on simulated data (not shown).

MCMC results from scenario 2, with only recovery times known, are shown by the circles in Fig. 3. We find that the number of animals needed to achieve the same accuracy from inference is significantly larger. For example, with perfect data 20 animals were needed to give σ_S =0.44, but this now increases by factor k=10 to over 200 animals. The situation slightly improves (triangles) when system parameters other than δ_S are assumed known.

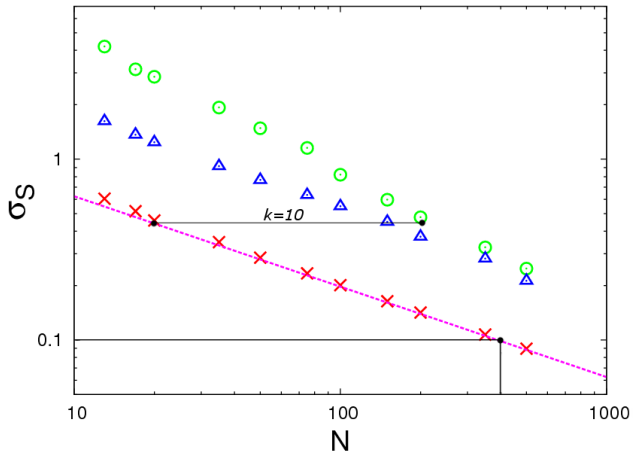

Figure 3: The standard deviation σ_S in the posterior distribution for the inferred susceptibility difference δ_S as a function of the population size N. 1) Event data is known precisely (crosses), 2) Only recovery data is known (circles), and 3) Recovery data and other model parameters are known (triangles).

Figure 4 shows results from scenario 3, in which the infective status of animals is checked at regular intervals. If an animal is found to be infected, it can be concluded that the actual infection time was at some point since the last check. Thus, partial information about infection is known in addition to the recovery times. In Fig. 4 the time between checks Δt is scaled by t_{av}, which is the average time of infection for animals over the entire epidemic. The y-axis shows the factor k by which the population size must be increased under scenario 3 in order to achieve the same σ_S as under scenario 1 where infection events are known exactly. The crosses show MCMC results, and when Δt becomes small k approaches 1, as would be expected. Perhaps surprising from this graph, however, is that even when measurements are made on the same timescale as the complete epidemic (i.e. $\Delta t/t_{av}\approx 1$), the accuracy of the inference is only marginally degraded.

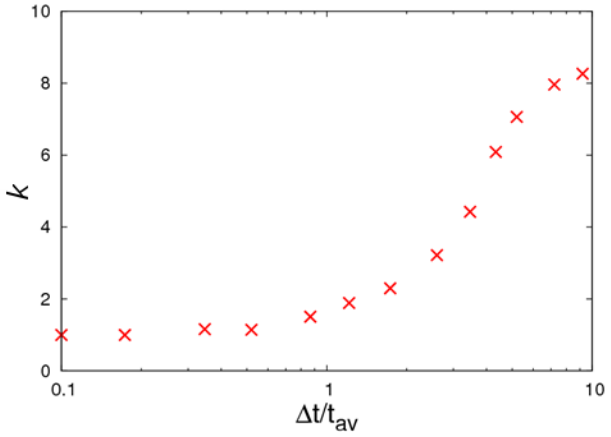

Figure 4: The infectious status of animals is periodically checked with time interval Δt. This graph shows the factor k by which the population number needs to be scaled to give an equivalent accuracy as when exact infection times are known. t_{av} is the average infection time.

The work here focuses on dominant alleles impacting on phenotypic epidemiological traits. The general procedure, however, is readily extendible to additive or partial dominance cases. The two coupled SIR models in Fig 1(a) would then expand to three representing two homozygote and one heterozygote genetic states.

Conclusion

Using MCMC within a Bayesian framework we have shown it is possible to infer parameters from temporal epidemiological data, even when the times of infection are not directly observed. To be able to measure small differences in epidemiological traits it will be necessary to have data from many hundreds, if not thousands, of animals. Furthermore, to estimate infectivity accurately, data from many epidemics will be needed, or alternatively existing data needs to be supplemented with information about which animals infect which. Thus, many challenges remain.

Acknowledgments

We wish to thank the Scottish Government for funding through SPASE and the BBSRC.

Literature Cited

Hastings, W. K. (1970). Biometrika 57:97-109

Keeling, M. J. and Rohani, P. (2007). Princeton University Press, NJ.

Lee, P. M. (2004). Arnold, London.

Walker, D. M., Pérez-Barbería, F. J., Marion, G. (2006). Ecol. Modell. 198:40-52