Uncertainty in On-The-Fly Epidemic Fitting

Roxana Danila, Marily Nika, Thomas Wilding, and William J. Knottenbelt

Department of Computing, Imperial College London
South Kensington Campus, London SW7 2AZ
{id210,marily,tw610,wjk}@imperial.ac.uk

Abstract. The modern world features a plethora of social, technological and biological epidemic phenomena. These epidemics now spread at unprecedented rates thanks to advances in industrialisation, transport and telecommunications. Effective real-time decision making and management of modern epidemic outbreaks depends on the two factors: the ability to determine epidemic parameters as the epidemic unfolds, and the ability to characterise rigorously the uncertainties inherent in these parameters. This paper presents a generic maximum-likelihood-based methodology for online epidemic fitting of SIR models from a single trace which yields confidence intervals on parameter values. The method is fully automated and avoids the laborious manual efforts traditionally deployed in the modelling of biological epidemics. We present case studies based on both synthetic and real data.

Keywords: Epidemics, Compartmental disease models, SIR models, Maximum likelihood estimation

1 Introduction

> I have called the uncertainty that surrounds any response to a microbial outbreak the *Fog of Epidemics*, analogous to the *Fog of War* of which historians speak.
>
> Richard M. Krause

In this modern era, technological advances enable a deadly disease to spread across the globe in just a few days. If during the times of the Black Death people typically travelled less than 10 miles in a day, nowadays 14 000 miles can be covered in a day, resulting in unprecedented rates of infection spreading [19]. In addition to biological epidemics, phenomena such as social and technological epidemics have emerged due to the extensive coverage and penetration of the Internet and social media [3, 14, 23, 4]. Online phenomena are characterised by a rapid, exponential spread through the population and are often triggered by seemingly inconsequential causes compared to the magnitude of their effects. The ability to predict and control such events is a topic of increasing interest.

Several formal quantitative approaches are available for making predictions about infectious disease. Although widely used in contingency planning, predictive modelling is still "the art of possible". The key requirement for a good

model is to provide accurate predictions, although it is already well established that such predictions cannot achieve perfect accuracy. This uncertainty arises due to two main factors: (i) the transmission of infection is stochastic in nature, making it very unlikely that one observes identical dynamic disease trajectories, even when the underlying epidemic processes are parameterically identical; (ii) models are an approximation, and rare or unforeseen behavioural patterns cannot be captured, but can have a significant impact on the disease dynamics [16]. Uncertainty can also result from assumptions made about the infectious agent and the environment, or even the technical details of the model.

The main contribution of this paper is a generic methodology for on-the-fly epidemic fitting of a classical compartmental epidemiological model, namely Susceptible-Infected-Recovered (SIR) [17], with inbuilt characterisation of parameter uncertainty. Given a single data trace of an evolving outbreak, a technique is developed for the fitting of SIR model parameters using an optimization method that employs a maximum-likelihood-based objective function. The output is a set of confidence intervals on key parameter values. In contrast with traditional approaches deployed in biological epidemics, which require laborious manual work for index case identification, lab testing and contact tracing, this method is fully automated.

A novel aspect of this research is connected to one of the major challenges: not knowing or being able to estimate from past data the initial number of susceptible and infected individuals. These initial conditions cause potentially large uncertainties in the estimation procedure. Our previous attempts to address this challenge using a least-squares fitting procedure yielded point estimates for parameters without any characterisation of related uncertainty [21].

The rest of this paper is organised as follows. Section 2 presents background information regarding infectious disease modelling. Section 3 describes the main optimisation methods used for fitting the models, estimating the parameters and setting confidence intervals to capture their uncertainty. Section 4 presents some example analyses on synthetic and real disease data. Section 5 concludes.

2 Background

Improved sanitation, antibiotics and vaccination programs created a confidence in the 1960s that infectious disease spreading would be eliminated. However, infectious disease agents adapt and evolve over time, so that new infectious diseases have emerged and some existing diseases have re-emerged. Mathematical models have become important tools in planning, implementing, evaluating and optimizing various detection, prevention, therapy and control programs. Epidemiology modelling can contribute to the design and analysis of epidemiological surveys, suggest crucial data that should be collected, identify trends, make forecasts and estimate the uncertainty in forecasts [15, 13, 6, 9, 11].

An epidemic is defined as a widespread occurrence of an infectious disease in a community at a particular time. Real-time forecasts of epidemic spread using data-driven models have been hindered by technical challenges posed by param-

eter estimation and validation [21]. Furthermore, traditional approaches rely on laborious and often infeasible approaches to initial estimates for parameters, such as studying in detail the index cases of the outbreak to infer, for example, the recovery rate as the reciprocal of the average infectious period [7].

In 1927 Kermack and McKendrick proposed one of the classical compartmental models most widely used in epidemiology, namely SIR [17]. Using Ordinary Differential Equations (ODEs), this models the evolution of an epidemic over time in terms of the number of Susceptible, Infected and Recovered individuals. Given a closed population of individuals, it defines

- $S(t)$ = individuals not yet infected at time t, but susceptible to infection
- $I(t)$ = individuals infected at time t by contact with susceptibles at a rate β
- $R(t)$ = individuals recovered at time t at a constant rate γ

We assume that the size of each compartment is a differentiable function of time. We ignore intricacies related to the pattern of contact between individuals, considering the instantaneous rate of new infections to be βSI. The recovery rate γ is proportional to the number of infected individuals, as each individual is assumed to recover at a constant rate γ.

These assumptions lead to the set of differential equations:

$$\frac{dS}{dt} = -\beta SI \tag{1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I \tag{3}$$

The initial values of the SIR model must satisfy the following conditions:

$$S(0) = S_0 > 0 \tag{4}$$

$$I(0) = I_0 > 0 \tag{5}$$

$$R(0) = 0 \tag{6}$$

and at any time, t, $S(t) + I(t) + R(t) = N$, where N is the total population size.

Such compartmental models can forecast the disease spread between individuals, not only in one population but also in various subpopulations and across localities [20, 2]. An outbreak originating in a seed subpopulation could potentially lead to a global-scale epidemic. A computational model called the Global Epidemic and Mobility model (GLEAM) is capable of integrating high-resolution data on human demography and mobility on a global scale in a metapopulation stochastic epidemic framework. GLEAM can simulate the global spread of influenza in order to provide insights on intervention strategies including vaccinations, antiviral treatment and travel restrictions [22].

A compelling interdisciplinary analysis of methods through which model uncertainty can be negotiated is presented in [9]. The study shows that many

models provide only cursory reference to the uncertainties of the information and data, or the parameters used, concluding that a more careful consideration of the limitations and uncertainties present in modelling epidemic phenomena would drastically improve its value. It is therefore essential to implement a rigorous and transparent technique that can provide confidence intervals on parameters for a clear understanding of evolving scenarios.

3 Methodology

Given a data set, we estimate the parameters using a two-pass methodology that combines least squares (LS) and maximum likelihood (ML) based optimisation techniques. Uncertainty quantification is then performed using the profiles obtained from the ML estimates.

3.1 Model Fitting Procedure

Mathematical modelling of infectious disease dynamics relies on a series of assumptions regarding key parameters that cannot be measured directly. We discuss here the technique used to fit the parameters of our model as an outbreak unfolds over time. In particular, we consider the challenges of estimating the initial number of susceptible and infected individuals in the target population, when these values are unknown. Currently, there is no principled way of doing this, as traditionally they are either known or can be estimated from the context [21]. However, in an era of social and technological epidemics, we argue that time and speed of movement make it infeasible to obtain accurate manual estimates.

3.1.1 Online model fitting We attempt to account for uncertainty as each outbreak unfolds, over time. To achieve this, we apply our fitting methodology on truncated data sets. We initially consider the first 10 observations from the outbreak. We then create new truncated datasets by adding each subsequent observation as the outbreak unfolds.

Using the SIR model, we propose two methodologies, one for estimating the parameters β, γ, S_0, and another for estimating β, γ, S_0 and I_0. By definition, all these quantities are positive, allowing us to apply a log transformation, yielding an unconstrained optimisation landscape with no possibility to explore infeasible values. Similarly a scaled logistic transformation can be applied to the initial number of susceptibles S_0 and infecteds I_0 when these are known to be bounded above by some constant C. The transformation function is:

$$\text{trans}(x) = \log(\frac{x}{C-x}) \tag{7}$$

and its corresponding inverse is:

$$\text{trans}^{-1}(y) = \frac{C}{1+e^{-y}} \tag{8}$$

3.1.2 Parameter estimation using Maximum Likelihood The Maximum Likelihood method is an analytic procedure for finding parameter vectors which maximise the likelihood of a dataset of iid observations. The likelihood function is defined as:

$$\mathcal{L}(\boldsymbol{\theta} \,|\, x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) \tag{9}$$

where $f(x_1, x_2, \ldots, x_n|\boldsymbol{\theta})$ is the joint density function of the observations and $\boldsymbol{\theta}$ the vector of unknown parameters. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is then:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta} \,|\, x_1, \ldots, x_n) \tag{10}$$

Equivalently, one can minimise the negative log likelihood:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \big(-\log \mathcal{L}(\boldsymbol{\theta} \,|\, x_1, \ldots, x_n) \big) = \arg\min_{\boldsymbol{\theta}} \big(-\sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta}) \big) \tag{11}$$

In the present work we assume the observations to be Poisson distributed. Typically, epidemiologists model variability in disease occurrence using either Binomial, Poisson or Exponential distributions. [12] argues that the three distributions have common attributes and underlying assumptions that tend to yield similar results. They also state that the Poisson distribution is widely used by epidemiologists when the data involves summary counts of cases. Moreover, since we deal with discrete observations, the variance is expected to scale with the number of infected individuals [5, 10].

The estimates are computed using the *mle2* function in the *bbmle* R package, which requires a negative log-likelihood function and starting values for the initial parameters to be specified. A computational challenge arises through the calculation of confidence intervals within *mle2*. This requires calculating the covariance matrix for the parameters, which is done by inversion of the Hessian matrix at the optimum and can be unsuccessful depending on the initial parameters. To overcome this, we first applied a Least Square based fitting procedure and used the estimates provided as starting values in order to be able to successfully estimate the confidence intervals.

The set of parameters that gives the best Maximum Likelihood based fit to the data is found using the Nelder-Mead algorithm, a widely used gradient-free method for unconstrained multidimensional optimization [18]. The first-order ODEs are solved using the *lsoda* R package. For optimal results, it is important to specify a small threshold for the absolute error tolerance.

3.1.3 Confidence intervals We make use of profile confidence intervals to indicate how reliable the estimate for a parameter is. The level of confidence is taken to be the probability that the interval contains the true value of the parameter, given a distribution of samples.

Traditionally, Wald-type confidence intervals are used as an approximation to profile intervals. The standard procedure for computing such a confidence interval is:

$$\text{estimate} \pm (\text{percentile} \times \text{SE}(\text{estimate})) \tag{12}$$

where SE is the standard error and the percentile represents the desired confidence level with respect to some reference distribution. Although easier to compute for complex models, it performs poorly when the likelihood surface is not quadratic.

A more robust technique for constructing confidence regions can be derived from the asymptotic χ^2 distribution of the likelihood ratio test statistic. Given a maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of a parameter vector $\boldsymbol{\theta}_0$, an approximate $(1-\alpha)$ confidence interval for $\boldsymbol{\theta}_0$ is the set of values of satisfying:

$$\{\boldsymbol{\theta} : 2[l(\hat{\boldsymbol{\theta}}) - l(\boldsymbol{\theta}_0)] \leq c_{k;1-\alpha}\} \tag{13}$$

where $c_{k;1-\alpha}$ is the $(1-\alpha)$th quantile of the χ^2 distribution with k degrees of freedom. Confidence intervals for individual parameters can be obtained by treating the others as "nuisance parameters" and maximising over them [24].

We compute two-sided confidence intervals using the *confint* function in the *bbmle* R package, at various confidence levels: 99%, 95%, 90%, 80% and 50%. In addition, we provide a 3D visualisation of the confidence intervals for the case when the unknown parameters vector is β, γ and S_0. This representation takes the shape of an ellipsoid, with each of the axis corresponding to one of the parameters to estimate. Note that the semi-axes may be unequal due to their asymmetric confidence intervals.

4 Results

In order to illustrate key aspects of the proposed approach we use both synthetic and a real-world datasets. The synthetic datasets were generated based on Gillespie's Stochastic Simulation Algorithm, using the *ssa* function in the *GillespieSSA* R package. The real dataset represents positive laboratory tests for influenza summed over all subtypes of the flu virus, as reported to the Centre of Disease Control (CDC) during the 2012/2013 flu season (starting in September 2012). The data were obtained via the FluView Web Portal[1].

4.1 Synthetic Data

The synthetic data set used in this section was generated by simulating an SIR epidemic with known parameters $\beta = 0.001$, $\gamma = 0.1$ and initial conditions $S_0 = 500$, $I_0 = 10$, $R_0 = 0$.

[1] http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

Synthetic Data with β, γ, S_0 unknown We fitted truncated datasets obtained of 25%, 50%, 75% and 100% of the data in order to analyse the uncertainty in the parameters as more data becomes available. As time progresses, we observe that our fits become more and more stable as illustrated in Figure 1.

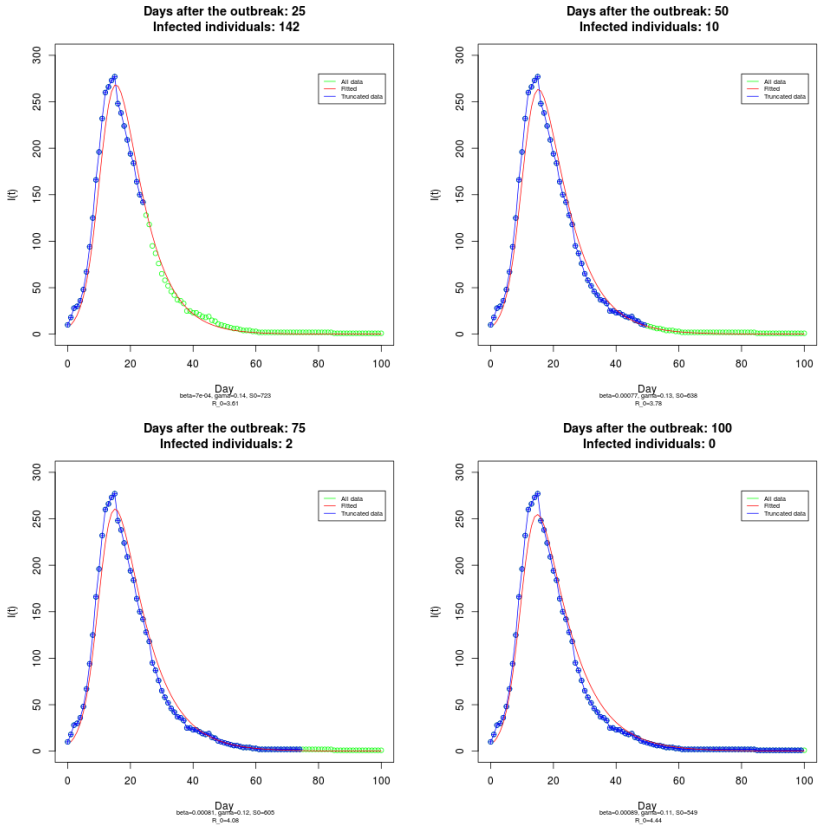

Fig. 1: Fitting of SIR model with β, γ, S_0 unknown to synthetic data

Figure 2 shows the profiles obtained from the ML estimate at various confidence levels for log-based transformations of each of the unknown parameters β, γ and S_0. For example we see that the 95% confidence interval for $\log(\beta)$ is (-7.083,-6.962), yielding a 95% confidence interval for β as (8.39e-04, 9.47e-04). As expected, the estimated range of possible values is wider as the confidence level increases. This is illustrated in the isosurface plot extended to three dimensions to visually represent the uncertainty inherent in the parameters.

Table 1 shows the lower and upper bounds on each parameter when the data is fitted over time. We observe the uncertainty of the parameters tends to decrease as more observations are considered.

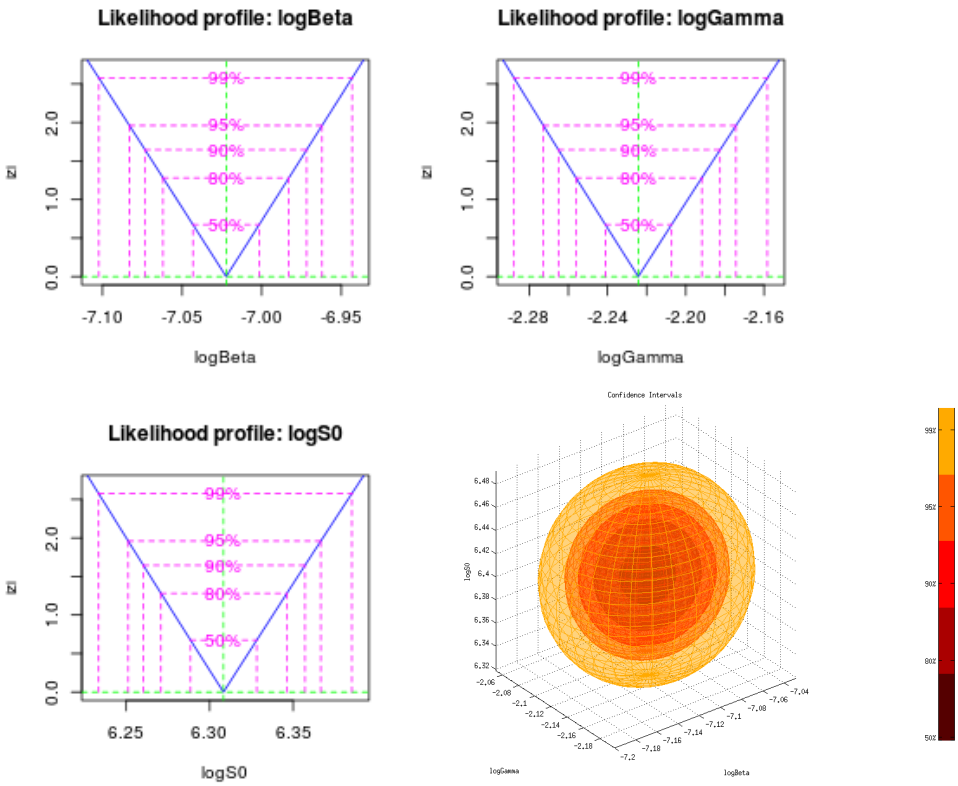

Fig. 2: Likelihood profile plots and corresponding isosurface plot for the estimated confidence intervals of transformed parameters when β, γ and S_0 are unknown (synthetic data)

Table 1: 95% Confidence Intervals for synthetic data

Data%	β		γ		$S0$	
	Lower	Upper	Lower	Upper	Lower	Upper
25%	5.66e-04	8.47e-04	1.08e-01	1.93e-01	569	962
50%	7.17e-04	8.36e-04	1.17e-01	1.35e-01	590	692
75%	7.62e-04	8.68e-04	1.13e-01	1.26e-01	568	646
100%	8.39e-04	9.47e-04	1.03e-01	1.14e-01	519	582

Synthetic Data with β, γ, S_0, I_0 unknown Figure 3 captures the uncertainty characterised over the parameters β, γ, and the initial conditions S_0, I_0, where I_0 is bounded by S_0 using a logistic based transformation. The uncertainty ranges and estimated values are similar to the ones computed by the optimisation with known I_0, demonstrating the robustness of the optimisation.

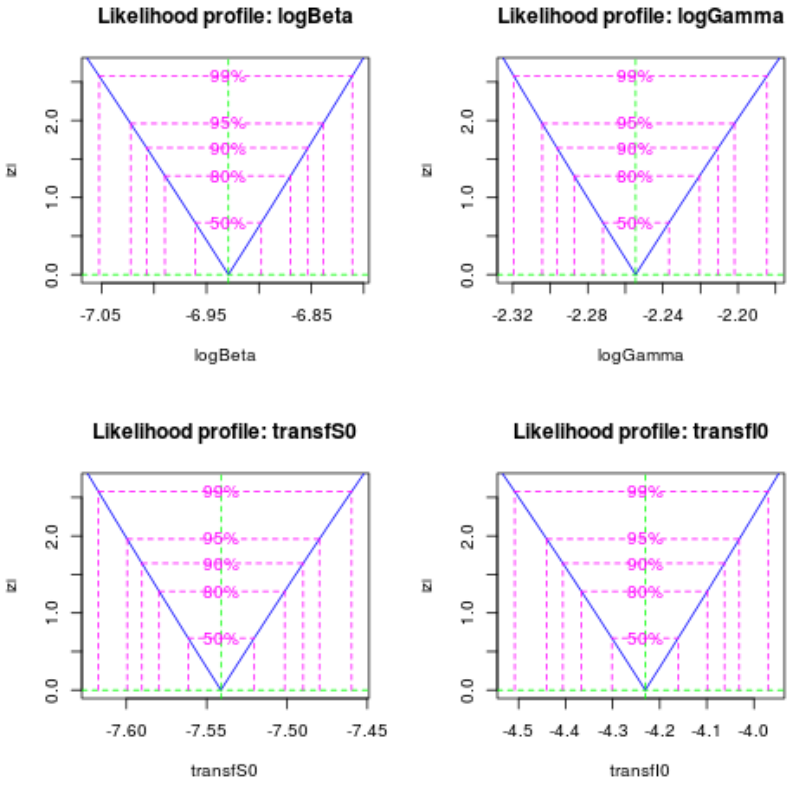

Fig. 3: Likelihood profile plots for the estimated confidence intervals of transformed parameters when β, γ, S_0 and I_0 are unknown (synthetic data)

True value recoverability rate for parameters For a known set of ground truth parameters, we use Gillespie's stochastic simulation algorithm to generate 1 000 sample trajectories of the number of infected individuals over time. For each trajectory, we apply our methodology to obtain 95% confidence intervals for each parameter. We might have expected that 95% of the time, the true values of the parameters should lie within the 95% confidence interval. However, Table 2 shows that this is not the case. This emphasises how difficult it is to obtain accurate

estimates of the uncertainty of the parameters from a single data trace. Such traces may be heavily affected by stochastic variation, especially in cases like our example where there are a relatively small number of initial susceptibles [1]. We also note the improvement in recovery rates for β and S_0 when I_0 is included as an unknown parameter, showing the benefits of maintaining flexibility with respect to this critical initial condition.

Table 2: True value recoverability rate for unknown parameters β, γ and S_0 (left) and for β, γ, S_0 and I_0 (right)

Parameter	Recoverability rate
β	26.59%
γ	26.28%
S_0	31.82%
β, γ, S_0	8.86%

Parameter	Recoverability rate
β	41.99%
γ	26.28%
S_0	34.44%
I_0	48.04%
β, γ, S_0, I_0	9.46%

4.2 CDC Influenza Data

We used data regarding positive lab-based influenza tests reported to the Center of Disease Control and Prevention (CDC) during the 2012/2013 influenza season.

Figure 4 shows the fitting over time of truncated datasets, illustrating that the algorithm is robust enough to be applied to real data.

Figure 5, Figure 6 and Table 3 characterise the uncertainty of the parameters for the real data set. The similar behaviour to the synthetic data reinforces our results and methodology.

Table 3: 95% Confidence intervals for influenza data (* - non convergence)

Data%	β		γ		$S0$	
	Lower	Upper	Lower	Upper	Lower	Upper
25%	*	*	*	*	*	*
50%	2.95e-05	3.22e-05	3.46e-01	3.81e-01	26769	30118
75%	3.50e-05	3.69e-05	2.90e-01	3.06e-01	22091	23515
100%	3.53e-05	3.70e-05	2.90e-01	3.03e-01	22031	23292

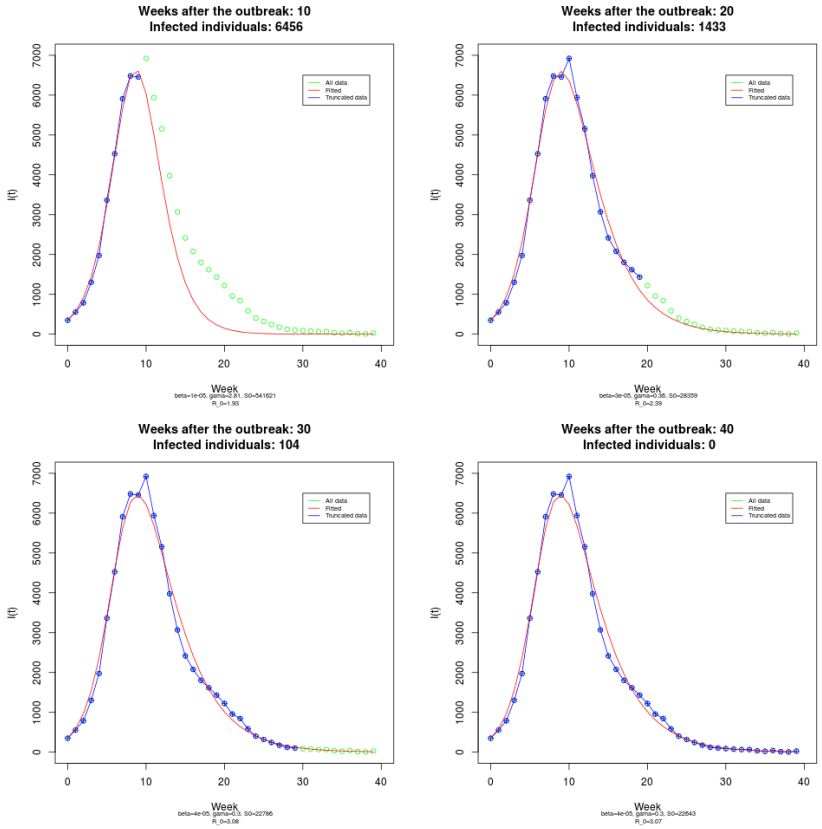

Fig. 4: Fitting of SIR model with β, γ, S_0 unknown to real influenza data

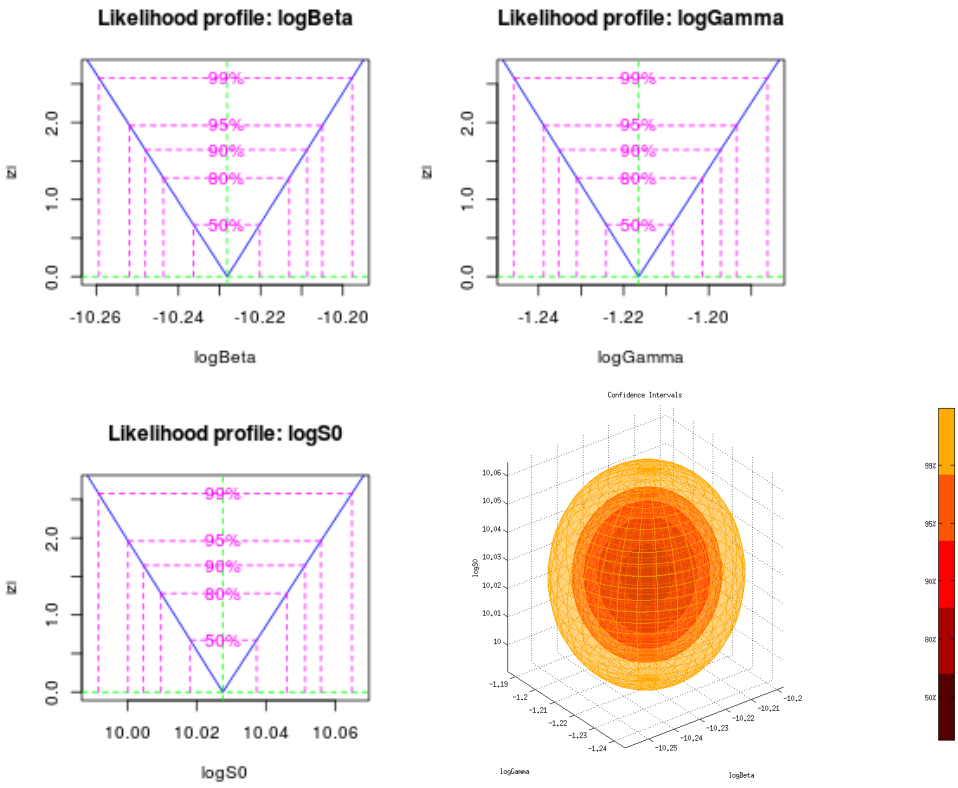

Fig. 5: Likelihood profile plots and corresponding isosurface plot for the estimated confidence intervals of transformed parameters when β, γ and S_0 are unknown (influenza data)

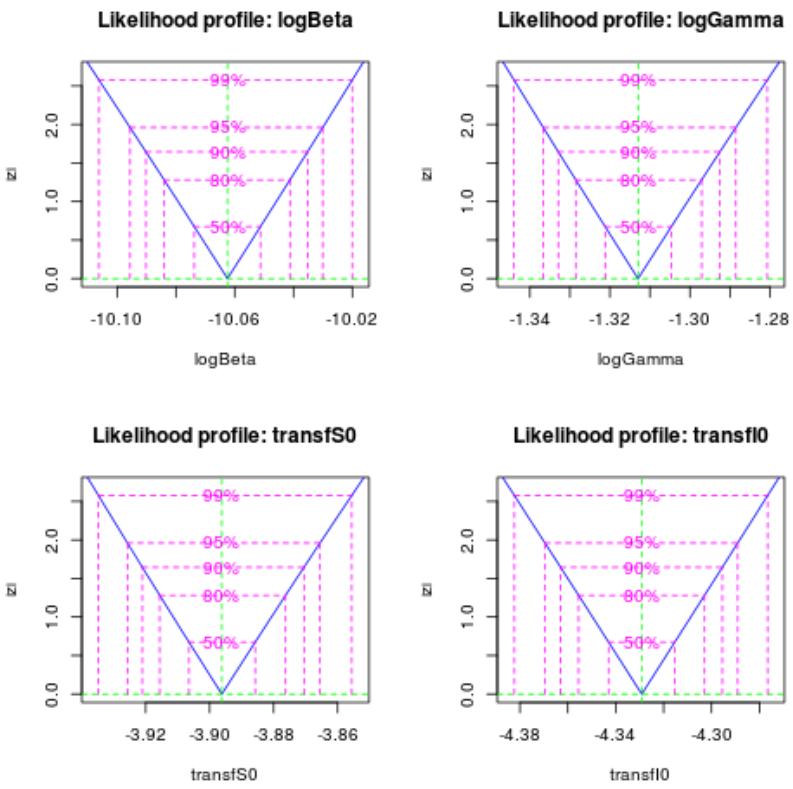

Fig. 6: Likelihood profile plots for the estimated confidence intervals of transformed parameters when β, γ, S_0 and I_0 are unknown (influenza data)

5 Conclusion

In this paper we provided a generic maximum-likelihood-based approach towards the on-the-fly epidemic fitting of SIR models from a single trace, which yields confidence intervals on parameter values. In contrast to traditional biological epidemiological modelling techniques, our approach is fully automated and the parameters to be estimated include the number of initial susceptibles and the initial number of infected in the population. Visualising the fitted parameters gives rise an isosurface plot of the feasible parameter ranges corresponding to each confidence level.

We generated multiple synthetic disease outbreak trajectories via stochastic simulation and fitted parameters to those trajectories. The "true" parameters were contained in the corresponding confidence bounds only for a relatively low proportion of the time, emphasising (a) the difficulty of obtaining accurate parameter estimations from a single epidemic trace and (b) the large potential impact of small random variations, especially those occurring early on in a trace.

It is expected that real systems are likely to exhibit different characteristics than the ideal ones assumed by the classical SIR model; for example real systems may feature time-varying parameters and the homogeneous mixing assumption may not apply. Nevertheless, the models may have utility in predicting the stochastic impact of candidate interventions in real systems with bounds [8]; a simulation-based methodology for this will be the focus of our future work.

References

1. H. Anderson and T. Britton. *Stochastic Epidemic Models and their Statistical Analysis*. Springer, 2000.
2. J. Angulo, H. Yu, A. Langousis, A. Kolovos, J. Wang, A. Madrid, and G. Christakos. Spatiotemporal Infectious Disease Modeling: A BME-SIR Approach. *PLoS ONE*, 2013.
3. E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The Role of Social Networks in Information Diffusion. In *Proc. 21st International Conference on the World Wide Web (WWW 2012)*, 2012.
4. F. Bauer and J. Lizier. Identifying Influential Spreaders. *CoRR*, abs/1203.0502, 2012.
5. B. Bolker and S. Ellner. Likelihood and all that, for Disease Ecologists, 2011. Available online at `http://kinglab.eeb.lsa.umich.edu/EEID/eeid/2011_eco/mle_2011.pdf`.
6. A. Briggs, M. Weinstein, E. Fenwick, J. Karnon, M. Sculpher, and A. Paltiel. Model Parameter Estimation and Uncertainty: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. *Value in Health*, 15(6):835 – 842, 2012.
7. E. Brooks-Pollock and K. Eames. Pigs didn't Fly, but Swine Flu. *Mathematics Today*, 47:36–40, 2011.
8. T. Burr and G. Chowell. Observation and Model Error Effects on Parameter Estimates in Susceptible-Infected-Recovered Epidemiological Models. *Far East Journal of Theoretical Statistics*, 19(2):163–183, 2013.

9. R. Christley, M. Mort, B. Wynne, J. Wastling, A. Heathwaite, R. Pickup, Z. Austin, and S. Latham. "Wrong, but Useful": Negotiating Uncertainty in Infectious Disease Modelling. *PLoS ONE*, 8(10):e76277, 10 2013.
10. R. Dolgoarshinnykh. Epidemic Modelling Graduate Topics Course. Lecture Notes. Available at `http://www.stat.columbia.edu/~regina/research/`.
11. B. Elderd, M. Vanja, V. Dukic, and G. Dwyer. Uncertainty in Predictions of Disease Spread and Public Health Responses to Bioterrorism and Emerging Diseases. *Proceedings of the National Academy of Sciences*, 103(42):15693–15697, 2006.
12. W. Flanders and D. Kleinbaum. Basic Models for Disease Occurrence in Epidemiology. *International Journal of Epidemiology*, 24(1):1–7, 1995.
13. J. Gilbert, L. Meyers, A. Galvani, and J. Townsend. Probabilistic Uncertainty Analysis of Epidemiological Modeling to Guide Public Health Intervention Policy. *Epidemics*, 6:37 – 45, 2014.
14. W. Hartmann, P. Manchanda, H. Nair, M. Bothner, P. Dodds, D. Godes, K. Hosanagar, and C. Tucker. Modeling Social Interactions: Identification, Empirical Methods and Policy Implications. *Marketing Letters*, 19(3):287–304, 2008.
15. H. Hethcote. The Mathematics of Infectious Diseases. *SIAM Review*, 42(4):599–653, 2000.
16. M. Keeling. State-Of-Science Review: Predictive and Real-time Epidemiological Modellig., 2006. Available online at `http://www.dti.gov.uk/assets/foresight/docs/infectious-diseases/s9.pdf`.
17. W. Kermack and A. McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772):700–721, 1927.
18. J. Lagarias, J. Reeds, M. Wright, and P. Wright. Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. *SIAM Journal of Optimization*, 9:112–147, 1998.
19. R. Lerner. The Black Death and Western European Eschatological Mentalities. *The American Historical Review*, 86:533–552, 1981.
20. M. Nika, D. Fiems, K. Turck, and W. J. Knottenbelt. Modelling Interacting Epidemics in Overlapping Populations. In *Proc. 21st International Conference on Analytical & Stochastic Modelling Techniques & Applications (ASMTA 2014)*, Budapest, Hungary, 2014.
21. M. Nika, G. Ivanova, and W. J. Knottenbelt. On Celebrity, Epidemiology and the Internet. In *Proc. 7th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS)*, Turin, Italy, December 2013.
22. M. Tizzoni, P. Bajardi, C. Poletto, J. Ramasco, D. Balcan, B. Goncalves, N. Perra, V. Colizza, and A. Vespignani. Real-time Numerical Forecast of Global Epidemic Spreading: Case study of 2009 A/H1N1pdm. *BMC Medicine*, 10(1):165, 2012.
23. V. Tweedle and R. Smith. A Mathematical Model of Bieber Fever: The most infectious disease of our time. In S. Mushayabasa and C. P. Bhunu, editors, *Understanding the Dynamics of Emerging and Re-Emerging Infectious Diseases using Mathematical Models*, chapter 7, pages 157–177. Transworld Research Network, 2012.
24. D. Venzon and S. Moolgavkar. A Method for Computing Profile-Likelihood-Based Confidence Intervals. *Applied Statistics*, 37(1):87–94, 1988.