Mathematical Biology

Limits of epidemic prediction using SIR models

Omar Melikechi[1] · Alexander L. Young[2] · Tao Tang[1] · Trevor Bowman[1] ·
David Dunson[1,3] · James Johndrow[4]

Abstract

The Susceptible-Infectious-Recovered (SIR) equations and their extensions comprise
a commonly utilized set of models for understanding and predicting the course of an
epidemic. In practice, it is of substantial interest to estimate the model parameters
based on noisy observations early in the outbreak, well before the epidemic reaches
its peak. This allows prediction of the subsequent course of the epidemic and design of
appropriate interventions. However, accurately inferring SIR model parameters in such
scenarios is problematic. This article provides novel, theoretical insight on this issue of
practical identifiability of the SIR model. Our theory provides new understanding of the
inferential limits of routinely used epidemic models and provides a valuable addition to
current simulate-and-check methods. We illustrate some practical implications through
application to a real-world epidemic data set.

Keywords SIR model · Epidemic prediction · Parameter inference · Identifiability ·
Nonlinear dynamics · Hypothesis testing

Mathematics Subject Classification 92-10 · 92D25 · 62M20

1 Introduction

The Susceptible-Infectious-Recovered (SIR) model, first introduced in the early twen-
tieth century, is a mathematical model describing the spread of a novel pathogen
through a population (Kermack and Mckendrick 1927; Ross 1916; Ross and Hudson
1917a, b). This model is governed by the ordinary differential equations

✉ Omar Melikechi
  omar.melikechi@duke.edu

[1] Department of Mathematics, Duke University, Durham, NC, USA

[2] Department of Statistics, Harvard University, Cambridge, MA, USA

[3] Department of Statistics, Duke University, Durham, NC, USA

[4] Department of Statistics, University of Pennsylvania, Philadelphia, PA, USA

$$\frac{ds}{dt} = -\beta i s, \quad \frac{di}{dt} = \beta i s - \gamma i, \quad \frac{dr}{dt} = \gamma i. \tag{1}$$

According to this model, the population is divided into three groups or "compartments," each of which represents a proportion of the population. The *susceptible compartment*, s, consists of the proportion of individuals who have never been infected with the pathogen. The *infected compartment*, i, consists of the proportion of individuals who are currently infected. The *removed compartment*, r, consists of the proportion of individuals who have either recovered from the pathogen and are immune or have died, and are therefore removed from the population. Since the SIR model assumes all recovered individuals are permanently immune to the pathogen, the value of r can be obtained from s and i via the identity $s + i + r = 1$.

During the last century the SIR equations have been modified and extended to model a diverse range of epidemics including Ebola, cholera, H1N1, tuberculosis, HIV/AIDS, influenza, malaria, Dengue fever, Zika, and most recently SARS-CoV-2 (Brauer et al. 2019; Coburn et al. 2009; Eisenberg et al. 2013; Khaleque and Sen 2017; Lee et al. 2020; Pasquali et al. 2021; Rachah and Torres 2015; Yang et al. 2020). In many of these examples, additional terms are added to account for pathogen specific characteristics of transmission. Additional compartments may also be added to model different subpopulations. One such example is the SEIR model, which includes a subpopulation of exposed (E) but non-infectious cases (Sauer et al. 2020). Collectively, the SIR model and its extensions and variations provide epidemiologists with a vast array of interpretable and highly expressive models to understand and predict the behavior of outbreaks. However, incorporating too many features can have subtle but important drawbacks including limited or unreliable inference of model parameters early in an epidemic. The main contribution of this article is insight on the inferential limits of epidemics (as captured by estimated parameters) which can be obtained from noisy, real-time observations of an outbreak.

Our work is motivated by the application of SIR and related compartmental models to real-time analysis of epidemics of human disease such as the 2014 Ebola outbreak in West Africa and the ongoing SARS-CoV-2 ("coronavirus") pandemic. In such outbreaks, the initial aim of the public health response is to extinguish the epidemic while the number of infected individuals is still small, or at least to significantly slow the rate of infection to allow time for the pathogen to be better understood and effective therapeutics or vaccines to be developed. The stay-at-home orders instituted by many countries due to SARS-CoV-2 are one recent example which has had profound global economic impacts. As such, mathematical models employed in the real-time analysis of epidemics must provide accurate inferences about properties of the epidemic – encapsulated by model parameters – *early in the epidemic*, when only a small fraction of the population has been infected. Hereafter, we refer to estimation of unknown model parameters from observations as the *inverse problem*.
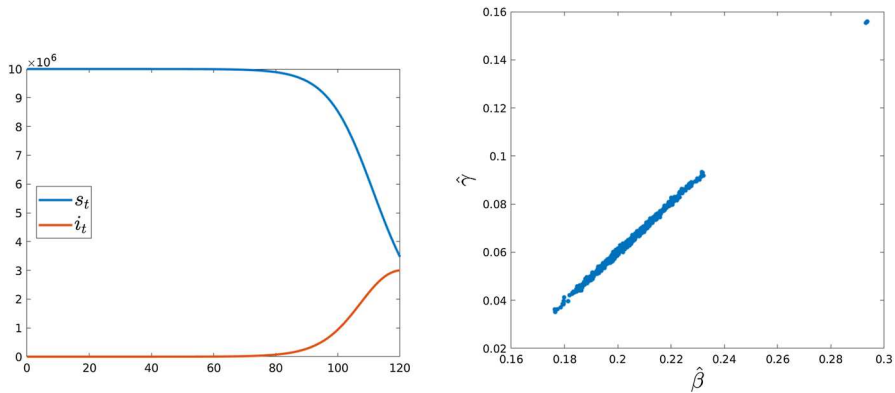
As noted in a review by Hamelin et al, many disease models proposed in the literature follow a similar structure: (1) a model is proposed, (2) a subset of model parameters are inferred from the literature, and (3) the remaining parameters are fit from data using least squares or maximum likelihood estimation (Hamelin et al. 2020). In order for these parameter estimates to be reliable the parameters must be statisti-

cally *identifiable*, ruling out settings in which multiple parameter values are equally consistent with observed data. Such issues were first considered in the context of compartmental models by Bellman and Aström in 1970 (Bellman and Åström 1970). Specific details relevant to the SIR model may be found in Hamelin et al. (2020). Here we provide a brief overview of the well-posedness of the inverse problem.

A model is structurally identifiable when there is a single value of the parameters consistent with noise-free data observations. A comprehensive review of analytic methods for assessing structural identifiability is given in Chis et al. (2011); alternatively, software packages such as DAISY can be used (Bellu et al. 2007). There are many examples in the literature (Brunel 2008; Chapman and Evans 2009; Daly et al. 2018; Eisenberg et al. 2013; Piazzola et al. 2020; Tuncer et al. 2016; Tuncer and Le 2018; Villaverde 2018). In particular, structural identifiability of the SIR parameters, β and γ, is well understood with strong theoretical support. See Hamelin et al. (2020) for specific cases based on different observations of the compartments. Similar considerations arise in literature related to branching process models which are also commonly used for modeling the dynamics of an outbreak. For example, Fok and Chou establish theoretical guarantees on ascertaining the progeny and lifetime distributions for Bellman-Harris processes when one knows the extinction time or population size distributions (Fok and Chou 2013). In practical applications, much less is typically known about the dynamics. Laredo et al. (2009) prove that when certain branching processes are observed only up to their nth generation, one can infer that the true model parameter belongs to a specific subset (which depends on n) of parameter space, but it is impossible to infer the exact true parameter for any finite n.

In practice, data observed during an epidemic tend to be very noisy, so we are far from the idealized noise-free case. *Practical identifiability* is the ability to discern different parameter values based on noisy observations. Despite considerable recent attention (Balsa-Canto et al. 2009, 2008; Chis et al. 2011; Srinath and Gunawan 2010), far less is known about practical identifiability. Present theoretical methods rely on sensitivity analysis and the computation of the Fisher information matrix, which is analytically intractable in the SIR model and its extensions. Instead, it is common to see Monte Carlo methods employed, wherein the model is simulated for a set value of the parameters, noise is added to the simulated observations, and a fitting procedure is conducted on the noisy data (Chis et al. 2011; Hamelin et al. 2020; Lee et al. 2020; Tuncer and Le 2018). The fidelity of parameter estimates relative to the known values is summarized using the average relative estimation error, which is then plotted as a function of the noise intensity.

Interestingly, the lack of practical identifiability manifests in a remarkably similar manner across multiple, different model formulations even in cases where the parameters are known to be structurally identifiable. As the magnitude of the noise is increased, Monte Carlo parameter estimates concentrate along a curve stretched throughout parameter space indicating a functional relationship between model parameters (Browning et al. 2020; Eisenberg et al. 2013; Piazzola et al. 2020; Tuncer et al. 2016; Tuncer and Le 2018). Importantly, there are often great disparities in parameter values along this curve and hence huge uncertainty in the parameters. See Fig. 1 for a representative example in the specific case considered herein.

Fig. 1 The trajectory of the SIR model used in the simulation (left). Plots of $\hat{\beta}$ vs $\hat{\gamma}$ from 1000 realizations from the sampling distribution of their MLE (right)

The goal of this article is to provide theoretical tools for understanding practical identifiability in the context of the SIR model. We propose a formulation based on realistic observations early in an outbreak. Then, using linearizations similar to those of Sauer et al. (2020), we construct analytically tractable approximations to the SIR dynamics from which theoretical guarantees of the performance of the inverse problem are developed. We begin by introducing the model under consideration, reemphasizing ideas discussed previously to provide overt examples of the challenges of practical identifiability.

2 Statistical model

The data available to infer the parameters of an SIR model are usually noisy, biased measurements of the rate of change in the size of the susceptible compartment, discretized to unit time intervals $\Delta_t = N(s_{t-1} - s_t)$. For simplicity, we take the time unit to be one day. Here, N represents the total population size in the jurisdiction under study and s_t is the size of the susceptible compartment at time t. The quantity Δ_t is the number of newly infected individuals between day $t - 1$ and day t. Data on daily confirmed cases, hospitalizations, or deaths are all examples of observable data that depend on the underlying value of Δ_t. Specifically, all are discrete convolutions of Δ_t of the form $p \sum_{s=0}^{t} \Delta_s \pi_{t-s}$, where p is the probability that an infected person goes on to be diagnosed, hospitalized, or die, and π_k is the conditional probability that a person tests positive, is hospitalized, or dies k days after becoming infected given that the corresponding outcome will eventually occur. It is likely that the parameters p and, to a lesser extent, π change over the course of an epidemic. However, changing values of these parameters can only make inference more difficult, and since our main focus is on studying limitations of inference, as a starting point we assume that p and π are fixed and known.

While the inverse problem with known initial conditions but unknown parameters $\theta = (\beta, \gamma)$ is well-posed when even a partial trajectory of Δ_t is observed, in reality

we observe Δ_t corrupted with noise, and we always have to work with finitely many discrete-time observations. In epidemic modeling, unlike some other inverse problems, we do not even have control of the sampling rate and are generally stuck with at best daily monitoring data. To simplify exposition, we focus on a simple but flexible noise model in which the observed data Y_t are realizations of a random variable satisfying $\mathbb{E}[Y_t] = p\Delta_t$ for some known $p \in (0, 1)$. In this case, $\pi_0 = 1$ and $\pi_k = 0$ for $k > 0$. While our results apply to many noise models, to fix ideas we begin with Gaussian noise

$$Y_t = p\Delta_t + \xi_t, \quad \xi_t \sim \mathcal{N}(0, \sigma_t^2). \tag{2}$$

In addition to simplifying exposition, our primary motivation for choosing Gaussian noise is to illustrate that the SIR model can, as we see shortly and explain later, be practically unidentifiable even for simple, idealized models like the one above. A secondary reason is that, despite its simplicity, (2) is not entirely unrealistic. For example, suppose any two people infected on day t have the same chance of eventually testing positive, that the chance any one such person tests positive is independent of whether any other such person does, and that the average number of people who became infected on day t who go on to test positive is roughly $p\Delta_t$. Then in any sufficiently large population the central limit theorem implies Y_t, which in this case is the number of people who become infected on day t and go on to get diagnosed, is approximately normally distributed with mean $p\Delta_t$ and some variance σ_t, i.e. Y_t satisfies (2).

   Initially, suppose that the variances σ_t^2 in (2) are known. A simple procedure for solving the inverse problem from data Y_t is maximum likelihood. The gradient of the log-likelihood can be obtained by numerically solving an extended ODE system (Gronwall 1919) which allows for easy fitting via gradient-based optimization methods. It can be shown that, even when the trajectory $p\Delta_t$ is observed only at discrete time intervals and the peak of infections has not yet occurred, the maximum likelihood estimator (MLE) exists and is unique, and so the model is structurally identifiable (Hamelin et al. 2020). Problems become apparent however when one seeks to study uncertainty in the estimated parameters. Figure 1 gives a stark indication of the challenges. We simulate data from an SIR model with parameters $\theta = (\beta, \gamma) = (0.21, 0.07)$ and initial conditions $s_0 = 1 - 1/N, i_0 = 1/N$ for $N = 10^7$. These parameters were selected to roughly approximate the dynamics of the coronavirus epidemic in New York City prior to the lockdown of March 16, 2020. The trajectories s_t, i_t for $0 \leq t \leq 120$ are shown in the left panel. By $t = 80$, about 1 percent of the population has been infected, and the peak size of the infected compartment occurs around $t = 120$. The right panel of Fig. 1 is obtained by repeatedly simulating data from (2) using the trajectory in the left panel, with $p = 1$ and $\sigma_t^2 = 100N$ chosen for illustrative purposes. Other potentially more realistic values of p and σ_t are considered later in the text; see for example Table 3 in Sect. 3.4 and Cases 1 and 2 in Sect. 3.2. For each replicate simulation, the model is fit by maximum likelihood. The resulting estimates of $\hat{\theta}$ are shown in Fig. 1, which plots $\hat{\beta}$ against $\hat{\gamma}$. These are samples from the sampling distribution of the maximum likelihood estimator for these parameters. The estimates exhibit very tight concentration along a line of slope 1. The variation in $\hat{R}_0 = \hat{\beta}/\hat{\gamma}$ observed for

these values is large, ranging from 1.88 to 5.01. This high degree of uncertainty occurs despite the fact that we have observed data up through the time when over half the population has been infected.

The linear shape of the plot in Fig. 1 suggests a practical identifiability problem in this model. That is, while the MLE exists and is unique, the curvature of the log-likelihood in the neighborhood of the MLE is very small in the direction where $\hat{\beta}$, $\hat{\gamma}$ lie along a line. We are not the first to notice this phenomenon. Previous works include (Chis et al. 2011; Hamelin et al. 2020; Lee et al. 2020; Tuncer and Le 2018), which experience qualitatively similar issues despite notable differences in the formulation of the likelihood in those settings.

While various empirical studies exist, our main contribution is a theoretical analysis of this phenomenon and the resulting limitations for solving the inverse problem from noisy observations. We take a two-step approach to the analysis. First, we characterize sensitivity of trajectories Δ_t to perturbations of the parameters θ, and show that perturbations of θ in the directions $\pi/4$ and $5\pi/4$ (equivalently, along the line of slope 1 through θ), closely approximate the smallest variation in the trajectory s_t among all perturbations for which $\|\theta_\epsilon - \theta\| = \epsilon$. We then give a computable approximate lower bound on $\inf_{\theta_\epsilon : \|\theta - \theta_\epsilon\| = \epsilon} |s_t(\theta_\epsilon) - s_t(\theta)|$ for times t prior to the peak infection time. Taken together, these results provide an explanation for the phenomenon in Fig. 1.

In the second part of the analysis, we relate the problem of uncertainty quantification to hypothesis tests of the form

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_\epsilon$$

for $\|\theta_\epsilon - \theta_0\| = \epsilon$. We use the result of the first part of our analysis to approximate the type II error of the test, which in turn allows for both theoretical and empirical analysis of the limits of epidemic prediction using SIR models.

3 Results

3.1 Perturbation bound for SIR trajectories

Informally, the phenomenon in Fig. 1 is a manifestation of the fact that very different values of θ can lead to SIR model trajectories that are very close. To formalize this, let $\varphi_t(x_0, \theta)$ be the (s, i)-trajectory of the SIR model starting from $x_0 = (s_0, i_0)$ with parameters $\theta = (\beta, \gamma)$. To aid the reader, all relevant notation is summarized in Table 1. We also remark that the analysis in this subsection and its associated appendices, "Appendices A and B", applies directly to the deterministic SIR system (1). In particular, it is independent of our choice of statistical model, which will not become relevant until our discussion of hypothesis testing in Sect. 3.2.

For $\epsilon > 0$, let $S_\epsilon(\theta)$ denote the circle of radius ϵ about θ. That is,

$$S_\epsilon(\theta) = \{\theta_\epsilon(\omega) : \omega \in [0, 2\pi)\}$$

Table 1 Summary of notation used throughout this article

Notation	Description
$x_0 = (s_0, i_0)$	Shorthand for initial conditions with $s_0 + i_0 = 1$
$\theta = (\beta, \gamma)$	Shorthand for the parameters of the SIR model
$R_0 = \beta/\gamma$	The reproductive number
$\delta = \beta - \gamma$	An important combination of the model parameters appearing in later analysis
θ_ϵ	Perturbation of θ such that $\|\theta_\epsilon - \theta\| = \epsilon$
$\theta_\epsilon(\omega)$	Perturbation of θ in the direction $\omega \in [0, 2\pi)$ such that $\|\theta_\epsilon(\omega) - \theta\| = \epsilon$
$\varphi_t(x_0, \theta) = (s_t(x_0, \theta), i_t(x_0, \theta))$	Solution of the SIR equation with initial condition x_0 and parameter θ
$Y_{1:T}$	Observed data on days 1 through T
$L(Y_{1:T}|\theta)$	Likelihood of θ given observed data

where $\theta_\epsilon(\omega) = \theta + \epsilon(\cos(\omega), \sin(\omega))$. We set $\delta = \beta - \gamma$ and assume throughout that $\delta > 0$; if not, then the *reproductive number* $R_0 = \beta/\gamma$ is at most 1 and the epidemic does not grow even at time 0. Similarly, we assume $\epsilon < \delta$. This ensures R_0 values of the perturbed parameters $\theta_\epsilon(\omega) = (\beta + \epsilon\cos(\omega), \gamma + \epsilon\sin(\omega))$ are also strictly greater than 1

$$\beta + \epsilon\cos(\omega) - \gamma - \epsilon\sin(\omega) = \delta + \epsilon(\cos(\omega) - \sin(\omega)) \geq \delta - \epsilon > 0$$

and so $R_0(\epsilon, \omega) = (\beta + \epsilon\cos(\omega))/(\gamma + \epsilon\sin(\omega)) > 1$ for every ω. Finally, for any fixed initial condition x_0 and parameter θ we define the *peak time*, denoted t_*, to be the deterministic time at which the number of infected individuals $i_t(x_0, \theta)$ is greatest; that is, $t_* = \text{argmax}\{i_t(x_0, \theta) : t \geq 0\}$. Since $di/dt = 0$ if and only if $i = 0$ or $s = 1/R_0$, it is follows that t_* exists and is unique whenever $R_0 > 1$. With this notation, the main result of this subsection is the following proposition.

Approximation 1 *Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^2$ and let t_* be the time of peak infection corresponding to θ. Then for all $t \in [0, 0.8t_*)$,*

$$\frac{\epsilon}{\delta\sqrt{2}}\left(e^{\delta t} - 1\right)i_0 \approx \inf_{\omega\in[0,2\pi)} \left\|\varphi_t\big(x_0, \theta_\epsilon(\omega)\big) - \varphi_t(x_0, \theta)\right\|. \tag{3}$$

Furthermore the infimum is approximately achieved when $\omega = \pi/4$ or $5\pi/4$.

The derivation of (3) is in "Appendix A". Approximation 1 says for any perturbation $\theta_\epsilon(\omega)$ of θ, the distance between the perturbed trajectory $\varphi_t(x_0, \theta_\epsilon(\omega))$ and true trajectory $\varphi_t(x_0, \theta)$ is approximately bounded below by the left side of (3) for all times t up to roughly 80% of t_*. The "$\approx$" in (3) indicates the bound is subject to error. Specifically, our derivation of Approximation 1 involves two approximations: First, we approximate the SIR model by a differential Eq. (A1) whose solution $\widetilde{\varphi}_t$ is given by (A2). Second, we use first-order Taylor expansions to approximate perturbations of $\widetilde{\varphi}_t$

resulting from perturbations in parameter space. Despite these approximations, numerical analysis of the error given in "Appendix B" indicates (3) holds for a wide range of parameter values and population sizes; see Fig. 2 below and Fig. 10 in "Appendix B". This numerical analysis also motivates our choice of 80% of the peak time as a cutoff, though this cutoff can be extended to 85% or even 90% for larger populations and certain parameter values; see Table 5. To complement the numerical results of "Appendix B", we give a theoretical upper bound on the error in "Appendix C". The theoretical result is more mathematically rigorous than the numerical one; however, it is significantly less precise than the control on error obtained in "Appendix B". We therefore use results derived from the numerical analysis, e.g. the 80% threshold, of "Appendix B" rather than the theoretical analysis of "Appendix C" for the remainder of this paper.

Approximation 1 successfully predicts the directions in parameter space, namely $\omega = \pi/4$ and $5\pi/4$ (equivalently, along the line of slope 1 through θ), corresponding to the most uncertainty about parameters even when data are observed up to the peak time, as in Fig. 1. In other words, the inverse problem of determining θ from data is least practically identifiable when distinguishing between θ and parameter values lying approximately on the line of slope 1 through θ. Furthermore, the approximate lower bound (3) quantifies the extent to which the inverse problem will not be practically identifiable which, as we discuss in the next subsection, is necessary for meaningful hypothesis testing. Finally, we find that the lower bound in (3) approximately holds for the s trajectory alone. That is, if $s_t(x_0, \theta_\epsilon(\omega))$ and $s_t(x_0, \theta)$ are the s trajectories corresponding to $\theta_\epsilon(\omega)$ and θ, respectively, then
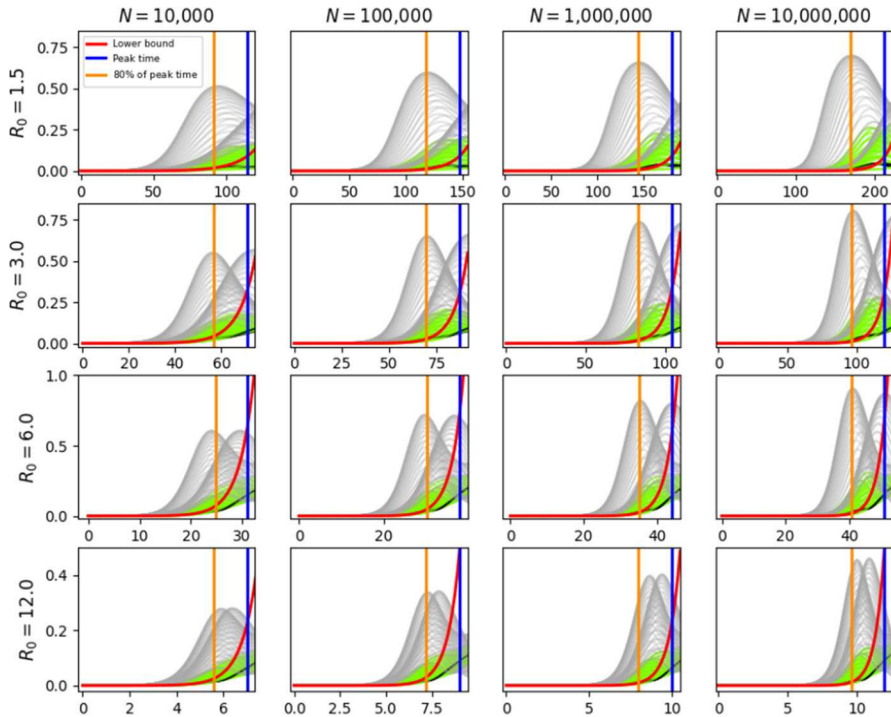
$$\frac{\epsilon}{\delta\sqrt{2}}\left(e^{\delta t} - 1\right)i_0 \approx \inf_{\omega \in [0, 2\pi)} \left| s_t\left(x_0, \theta_\epsilon(\omega)\right) - s_t(x_0, \theta) \right|, \qquad (4)$$

and the infimum is again achieved when $\omega = \pi/4$ and $5\pi/4$. The intuition behind (4) is that the s compartment is substantially larger than the i compartment early in an epidemic and therefore contributes significantly more to $\|\varphi_t^\epsilon - \varphi_t\|$ than i. This observation will be used for the hypothesis testing in Sect. 3.2 since our statistical model depends crucially on $\Delta_t = N(s_{t-1} - s_t)$, which in turn depends only on s rather than on s and i together. The approximation error implicit in the $\approx$ symbol in (3) and (4) is the one quantity we do not have rigorous control over; see "Appendix B" for details.
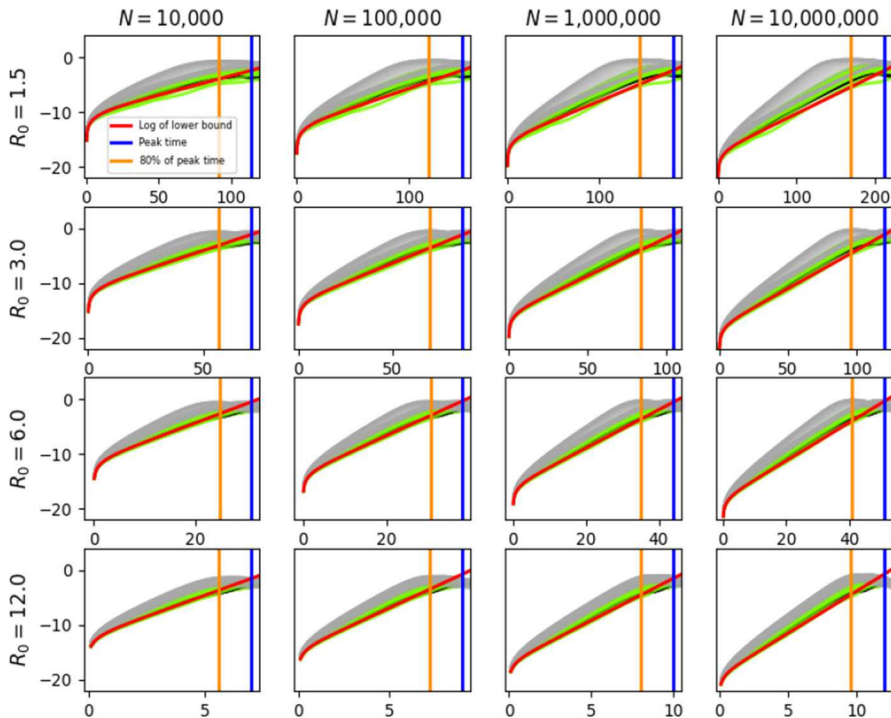
3.2 Hypothesis testing for the inverse problem

In this subsection we revisit the inverse problem in light of the perturbation bounds (3) and (4). To give context and motivate the main result of this subsection, namely Approximation 2 and its subsequent discussion, we first give a brief overview of simple hypothesis testing and the Neyman-Pearson Lemma.

Suppose we observe data Y taking values in a space $\mathcal{Y}$ and that these data are drawn from an unknown probability distribution belonging to a parametrized family of probability distributions $\{\mathbb{P}_\theta\}$. Given two parameters θ_0 and θ_1, a natural question

Fig. 2 Distance between perturbed and true trajectories for different parameter values and population sizes. In each graph the horizontal axis is the number of days since the start of the epidemic and the vertical axis is the distance $\|\varphi_t(x, \theta_\epsilon(\omega)) - \varphi_t(x, \theta)\|$ between a perturbed trajectory and the true trajectory at time t. The gray, green, and black curves correspond to 90 perturbed trajectories, one for each of 90 equally spaced angles ω in $[0, 2\pi)$. The black curves correspond to the angles $\pi/4$ and $5\pi/4$. The green curves correspond to the remaining angles in the intervals $[\pi/4 - \pi/12, \pi/4 + \pi/12)$ and $[5\pi/4 - \pi/12, 5\pi/4 + \pi/12)$, i.e. in intervals of width $\pi/6$ centered at $\pi/4$ and $5\pi/4$, respectively. The gray curves correspond to those angles in $[0, 2\pi)$ outside these two intervals. Note the distances corresponding to angles close to $\pi/4$ and $5\pi/4$ (the green and black curves) are smaller than those distances corresponding to angles farther away from $\pi/4$ and $5\pi/4$ (the gray curves), which supports the claim that the inverse problem is least practically identifiable for parameter perturbations approximately along a line of slope 1. The approximate lower bound of Approximation 1 is in red. The peak time of the trajectory corresponding to θ is indicated by the vertical blue line, and 80% of it by the vertical orange line. The first through fourth columns have population sizes 10^4, 10^5, 10^6, and 10^7, respectively, with only one initial infection in each case. The perturbation sizes for the first through fourth rows are $\epsilon = .03, .03, .06$, and $.1$, respectively. The SIR paramaters for the first through fourth rows are $(\beta, \gamma) = (.21, .14), (.21, .07), (.42, .07)$, and $(1.68, .14)$, which give respective R_0 values of 1.5, 3, 6, and 12. Note the approximate lower bound holds roughly up to 80% of the peak time in all cases despite the wide range of parameters. Finally, we remark that the two seemingly "distinct" classes of gray curves in each plot correspond to different subsets of the 90 distinct angles. This as well as the multimodality of certain curves (which becomes more apparent when our graphs are extended further beyond the peak time) are consequences of the nonlinearity of the SIR model and are not directly relevant to our analysis

Fig. 3 Logarithm of distance between perturbed and true trajectories for different parameter values and population sizes. Everything is the same as in Fig. 2 except now we plot $\log\|\varphi_t(x, \theta_\epsilon(\omega)) - \varphi_t(x, \theta)\|$ instead of $\|\varphi_t(x, \theta_\epsilon(\omega)) - \varphi_t(x, \theta)\|$. This gives a better view of the approximate lower bound early in the epidemic. Note the vertical axis is now a log scale

is whether the observed data came from $\mathbb{P}_{\theta_0}$ or $\mathbb{P}_{\theta_1}$. This is a *simple hypothesis test*, denoted by

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1, \tag{5}$$

where H_0 and H_1 are the *null* and *alternative hypotheses*, respectively. *Simple* here refers to the fact that both H_0 and H_1 correspond to single θ values which completely determine the distributions $\mathbb{P}_{\theta_0}$ and $\mathbb{P}_{\theta_1}$. The aim is to decide whether to reject H_0 in favor of H_1, which is done by choosing a subset $\mathcal{R}$ of $\mathcal{Y}$ called the *rejection region*. This choice of $\mathcal{R}$ completely determines the test: If $Y \in \mathcal{R}$, then reject H_0 in favor of H_1; if $Y \notin \mathcal{R}$, then do not reject H_0. *Type I error* occurs when H_0 is true but is rejected, and *type II error* occurs when H_0 is false but not rejected; see Table 2. This is quantified[1] as

$$\mathcal{E}_1(\mathcal{R}) = Type\ I\ error\ rate = \mathbb{P}_{\theta_0}(Y \in \mathcal{R}) = \mathbb{P}_{\theta_0}(\text{Reject } H_0),$$
$$\mathcal{E}_2(\mathcal{R}) = Type\ II\ error\ rate = \mathbb{P}_{\theta_1}(Y \notin \mathcal{R}) = \mathbb{P}_{\theta_1}(\text{Do not reject } H_0).$$

---

[1] Type I and II error rates are commonly denoted by α and β, but since β is already used as an SIR parameter we adopt the unconventional notation $\mathcal{E}_1$ and $\mathcal{E}_2$.

Table 2 Simple hypothesis test

	$H_0 : \theta = \theta_0$	$H_1 : \theta = \theta_\epsilon$
Reject H_0	Type I error	Success
Do not reject H_0	Success	Type II error

Ideally one would find a rejection region $\mathcal{R}$ that simultaneously minimizes type I and type II error rates, but this is generally impossible. Instead, a common statistical paradigm is to fix a *significance level $\alpha > 0$* and minimize $\mathcal{E}_2(\mathcal{R})$ subject to the constraint $\mathcal{E}_1(\mathcal{R}) = \alpha$. For such an α, a region $\mathcal{R}$ is called *a most powerful level-α rejection region* if $\mathcal{E}_1(\mathcal{R}) = \alpha$ and $\mathcal{E}_2(\mathcal{R}) \leq \mathcal{E}_2(\mathcal{R}')$ for all $\mathcal{R}'$ satisfying $\mathcal{E}_2(\mathcal{R}') = \alpha$. That is, $\mathcal{R}$ minimizes type II error over all rejection regions with type I error equal to α. The Neyman-Pearson Lemma gives the most powerful rejection region in the case of a simple hypothesis test.

Lemma 1 *(Neyman-Pearson) Let $L(Y|\theta)$ denote the likelihood function for data Y and a parameter θ, and fix $\alpha > 0$. Then there exists an $\eta \in \mathbb{R}$ such that*

$$\mathcal{R}_{LR} = \left\{ Y : \frac{L(Y|\theta_1)}{L(Y|\theta_0)} \geq \eta \right\} \tag{6}$$

is a most powerful level-α rejection region for the hypothesis test (5).

$L(Y|\theta_1)/L(Y|\theta_0)$ is called the *likelihood ratio* and the decision to reject or not reject H_0 based on the rejection region $\mathcal{R}_{LR}$ is called the *likelihood ratio test*. Since the Neyman-Pearson Lemma guarantees the likelihood ratio test is most powerful in our setting, we henceforth consider only the rejection region $\mathcal{R}_{LR}$ and set $\mathcal{E}_1 = \mathcal{E}_1(\mathcal{R}_{LR})$ and $\mathcal{E}_2 = \mathcal{E}_2(\mathcal{R}_{LR})$.

Returning now to the SIR model, our hypothesis test of interest is

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_\epsilon(\omega) \tag{7}$$

where, as before, $\theta_\epsilon(\omega)$ is a perturbation of θ of size ϵ in the direction ω. The observed data are $Y_{1:T} = (Y_1, \ldots, Y_T)$ for any time T before the time of peak infection, with each $Y_t = p\Delta_t + \xi_t$ as in (2). For the rest of this paper $\mathcal{E}_2(\omega)$ will denote the type II error rate of (7) for the likelihood ratio test with angle ω. As such, the likelihood ratio test (6) minimizes $\mathcal{E}_2(\omega)$ thereby providing the most powerful technique for detecting differences of order ϵ in the SIR model parameters. We also set $\Delta_t^\epsilon(\omega) = \Delta_t(\theta_\epsilon(\omega))$ where, recall, $\Delta_t(\theta) = N(s_{t-1}(\theta) - s_t(\theta))$, and let Φ denote the standard normal cumulative distribution function. With this notation we now present the main result of this subsection.

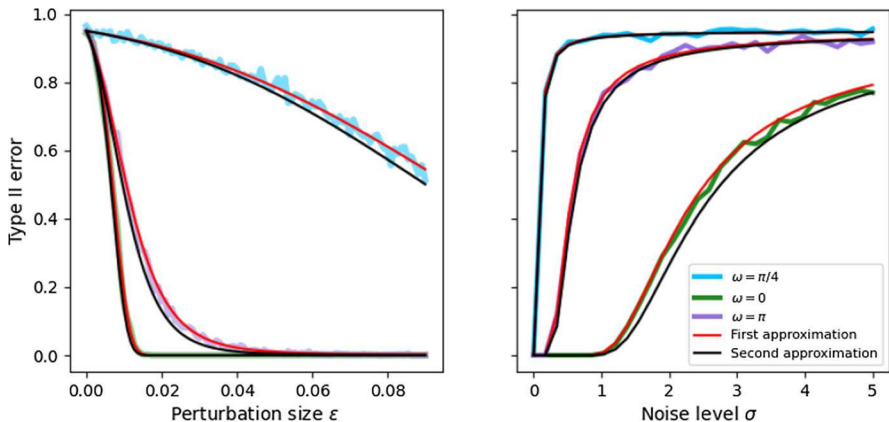**Approximation 2** *For any $\epsilon > 0$, $\omega \in [0, 2\pi)$, and significance level $\alpha > 0$,*

$$\mathcal{E}_2(\omega) \approx 1 - \Phi\left(\Phi^{-1}(\alpha) + pNi_0 \sqrt{\sum_{t=1}^{T} \frac{e^{2\delta t}}{\sigma_t^2} \left[\beta_\epsilon \left(\frac{e^{-\delta_\epsilon} - 1}{-\delta_\epsilon} \right) e^{\epsilon t f(\omega)} - \beta \left(\frac{e^{-\delta} - 1}{-\delta} \right) \right]^2} \right) \tag{8}$$

$$\approx 1 - \Phi \left(\Phi^{-1}(\alpha) + pNi_0 \sqrt{\sum_{t=1}^{T} \frac{e^{2\delta t}}{\sigma_t^2} \left[(\beta + \epsilon \cos \omega) e^{\epsilon t f(\omega)} - \beta \right]^2} \right), \tag{9}$$

where $f(\omega) = \cos(\omega) - \sin(\omega)$, $\beta_\epsilon = \beta + \epsilon \cos \omega$, and $\delta_\epsilon = \delta + \epsilon f(\omega)$. Moreover,

$$\mathcal{E}_2(\pi/4) \approx \mathcal{E}_2(5\pi/4) \approx \sup_{\omega \in [0, 2\pi)} \mathcal{E}_2(\omega).$$

The derivation of Approximation 2 is in the "Appendix". The first and second approximations of $\mathcal{E}_2(\omega)$ correspond to the red and black curves in Fig. 4, respectively. Comparing these to the empirical type II error rates (the blue, green, and purple curves) we see these approximations are sound. In particular, the last part of Approximation 2 indicates the angles $\pi/4$ and $5\pi/4$ give rise to the largest type II error rate for the hypothesis test (7) with perturbation size ϵ and significance level α. To quantify the

Fig. 4 Type II error as a function of perturbation size and noise level. The left panel shows the empirical and theoretical type II errors for the angles $\omega = 0, \pi/4$, and π as a function of perturbation size ϵ with fixed noise level $\sigma = 0.3$. The right panel shows the empirical and theoretical type II errors for the same angles as a function of noise level σ with fixed perturbation size $\epsilon = .03$. In each case the SIR parameters are those from Sect. 2, namely $(\beta, \gamma) = (.21, .07)$, $N = 10^7$, and initial condition $i_0 = 1/N$. The time horizon T is 60 days into the epidemic, which in this case is 60 days prior to the peak time. The significance level is $\alpha = .05$. Here, *theoretical* refers to the first (red) and second (black) approximations of type II error $\mathcal{E}_2(\omega)$ in Approximation 2, i.e. Eqs. (8) and (9), respectively. *Empirical* refers to the type II error obtained by performing 1000 simulations of the noisy SIR model (2) followed by a likelihood ratio test of the hypothesis in (7) for each set of parameters. More specifically, the red and black curves lying over the blue line are the type II error approximations (8) and (9) when $\omega = \pi/4$, those lying over the purple line are when $\omega = \pi$, and those lying over the green line are when $\omega = 0$, with the blue, green, and purple curves corresponding to the empirically computed type II error rates when $\omega = \pi/4, 0$, and π, respectively. In each case both theoretical results closely align with the empirical ones, with the first approximation being slightly better than the second as expected. Also as predicted, the empirical type II errors all approach $1 - \alpha = .95$ both as perturbation size goes to 0 and as the noise level gets large, and this approach is most rapid when $\omega = \pi/4$. In each case the noise model is Case 2, $\sigma_t = N\sigma i_t$. For the simulated blue, green, and purple curves, we used a numerical integrator to obtain the i_t values, while for the red and black curves we used the pre-peak approximation $i_t \approx e^{\delta t} i_0$

magnitude of type II error in these cases, we substitute into the second approximation to get

$$\mathcal{E}_2(\tfrac{\pi}{4}) \approx \mathcal{E}_2(\tfrac{5\pi}{4}) \approx 1 - \Phi\left(\Phi^{-1}(\alpha) + \frac{pNi_0\epsilon}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\frac{e^{2\delta t}}{\sigma_t^2}}\right). \tag{10}$$

Note that as the noise level σ_t^2 goes to 0, the sum under the square root goes to infinity and the entire expression goes to $1 - \Phi(\infty) = 0$. That is, if there is no noise then the type II error rate of the likelihood ratio test will vanish. If there is any noise at all however, the sum is finite and (10) becomes arbitrarily close to $1 - \Phi(\Phi^{-1}(\alpha)) = 1 - \alpha$ as either p, the probability of detecting an infected individual, or ϵ, the perturbation size, go to 0. For example, if we set the type I error rate to $\alpha = 0.1$ then as either p or ϵ go to 0, the probability of making a type II error will approach 0.9. Similarly, type II error will go to $1 - \alpha$ as σ_t^2 goes to infinity. This limit is unrealistic though since σ_t^2 is the variance of observed data and as such should be less than the population size. This leads us to consider two cases for noise.

Case 1. Noise proportional to population size, i.e. $\sigma_t = N\sigma$ for σ in (0, 1).
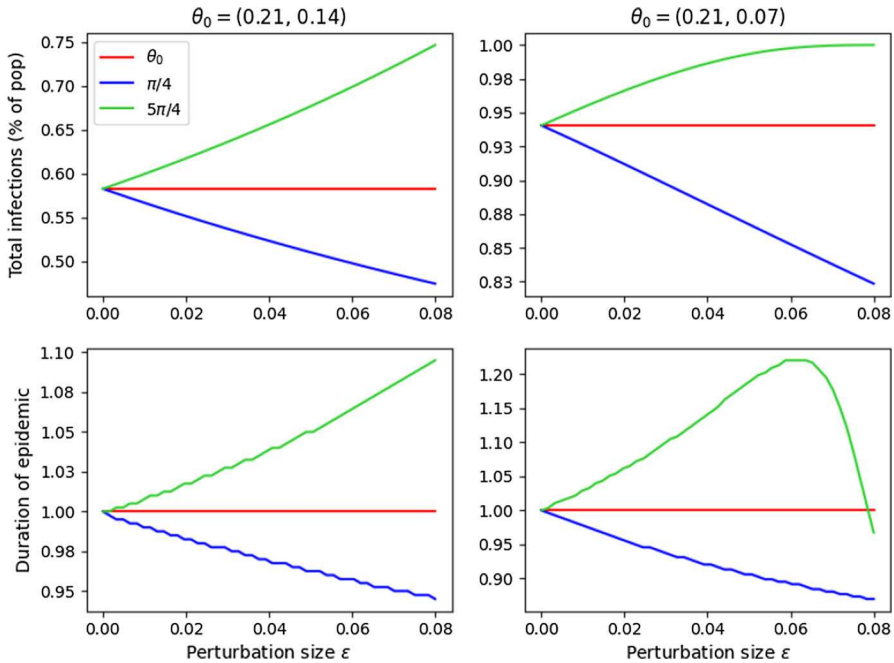*Case 2. Noise proportional to number of infections, i.e. $\sigma_t = N\sigma i_t$ for $\sigma > 0$.*

In both cases σ is constant and independent of t. Case 2 involves i_t which is not expressible in closed-form. However, we can use Approximation 1 and its derivation, specifically the approximate solution (A2), to circumvent this issue by replacing i_t with $e^{\delta t}i_0$. As discussed in Sect. 3.1, this approximation is appropriate early in the epidemic. In Case 1, Eq. (10) becomes

$$\mathcal{E}_2 \approx 1 - \Phi\left(\Phi^{-1}(\alpha) + \frac{pi_0\epsilon}{\sigma\sqrt{2}}\sqrt{\sum_{t=1}^{T}e^{2\delta t}}\right).$$

In addition to the aforementioned limits, we see in this case that the expression, and hence the type II error, approaches $1 - \alpha$ as the population N goes to infinity (so that $i_0 = 1/N$ goes to 0). In Case 2, Eq. (10) becomes

$$\mathcal{E}_2(\tfrac{\pi}{4}) \approx \mathcal{E}_2(\tfrac{5\pi}{4}) \approx 1 - \Phi\left(\Phi^{-1}(\alpha) + \frac{p\epsilon\sqrt{T}}{\sigma\sqrt{2}}\right). \tag{11}$$

The above expression does not depend on population size, N, nor on the SIR parameters β and γ, while the asymptotic results for p, ϵ, and σ still apply. Since Case 1 has noise proportional only to N, it implicitly assumes relative noise is larger earlier in the outbreak which may not be realistic. Case 2 avoids this since relative noise will be small whenever the reported number of infected individuals is small, e.g. early in an epidemic. For this reason and its invariance under different model parameters and population sizes, we consider only Case 2 moving forward.

Fig. 5 Consequences of type II error. The top panels show the total number of infections 10 days past the time of peak infection as a percentage of the total population. The bottom panels show the duration of the epidemic, which is defined to be the first day past the peak when less than 10 individuals are infectious. The left panels correspond to the parameter $\theta_0 = (0.21, 0.14)$ and the right panels to $\theta_0 = (.21, .07)$. The red lines give the total percent infected or duration of the epidemic for the true parameter θ_0 in each of their respective plots, while the blue and green curves give these values for $\theta_\epsilon(\pi/4)$ and $\theta_\epsilon(5\pi/4)$ over a range of ϵ values, respectively. In all cases $N = 10^7$

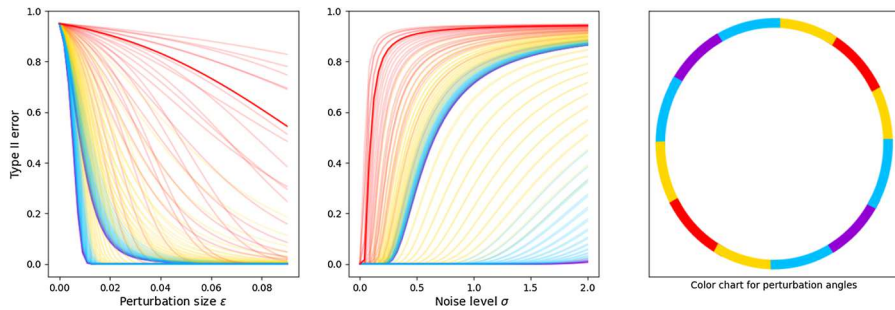### 3.3 Simple illustration: implications of Approximation 2

Approximation 2 says that given an SIR parameter θ_0, the probability of failing to reject the hypothesis $H_0 : \theta = \theta_0$ when the alternative $H_1 : \theta = \theta_\epsilon$ is true can be very high, especially when the angle of perturbation is $\pi/4$ or $5\pi/4$. In this subsection we take a closer look at what this means for epidemic prediction.

Consider for concreteness the familiar setting $\theta_0 = (.21, .07)$, $N = 10^7$, and $i_0 = 1/N$. Figure 5 shows the total number of infections 10 days past the peak time as well as the duration[2] of the epidemic for θ_0, $\theta_\epsilon(\pi/4)$, and $\theta_\epsilon(5\pi/4)$ and varying perturbation sizes ϵ.

Setting $\omega = \pi/4$ or $5\pi/4$ and letting noise be as in Case 2, the first approximation in Approximation 2 can be rearranged to obtain

$$\epsilon \approx \frac{\left[\Phi^{-1}(1 - \mathcal{E}_2) - \Phi^{-1}(\alpha)\right]\sigma\delta e^\delta\sqrt{2}}{(e^\delta - 1)p\sqrt{T}}.$$

---

[2] We define the duration of the epidemic to be first day after the peak time such that less than 10 individuals are infected.

Fig. 6 Practical identifiability of δ. The left and center panels use the first type II error approximation in Approximation 2 to graph type II error as a function of perturbation size, ϵ, and noise level, σ, respectively. Each of the rainbow colored curves in both panels correspond to one of 150 different values of ω spread uniformly across $[0, 2\pi)$. The color chart in the right panel indicates the colors corresponding to different angles ω: The light red curves correspond to the ω closest to $\pi/4$ and $5\pi/4$, the yellow are those a bit farther away, the blue still farther, and the purple are those farthest from $\pi/4$ and $5\pi/4$, i.e. closest to $3\pi/4$ and $7\pi/4$. Finally, the dark red curve in each of the two panels corresponds to $\omega = \pi/4$ and $5\pi/4$, which have the same type II error. Note the rapid fall off in type II error as angles get farther from $\pi/4$ and $5\pi/4$, especially as a function of ϵ. This agrees with the empirical observation in Fig. 1 that MLE favors parameters lying along a line of slope 1. In particular, the inverse problem for δ is practically identifiable
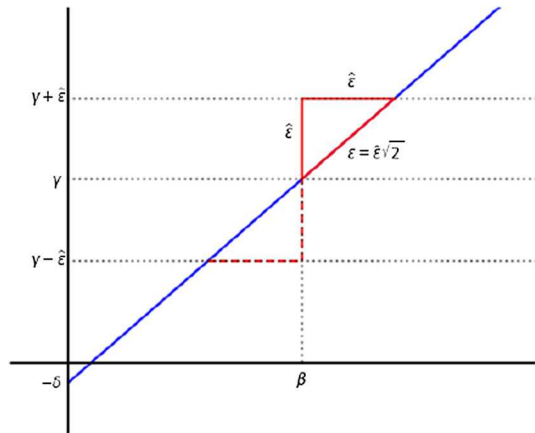
From this we can compute the consequences of type II error. For example, suppose we are 60 days into an epidemic ($T = 60$) and wish to test the hypothesis $\theta_0 = (0.21, 0.07)$ versus $\theta_\epsilon(5\pi/4)$ as above. Moreover, suppose $p = 1$ (perfect diagnostics), $\sigma = 0.2$ (infection standard deviation of $\pm 20\%$ of new cases), and we set $\alpha = .05$ and $\mathcal{E}_2 = 0.5$. Then the above equation gives $\epsilon \approx .064$. Thus, reading off the right panels in Fig. 5, we see that under these fairly generous conditions a type II error—which has a 50% chance of occurring—will result in underestimating the total number of infections of an epidemic by over 5% of the total population and the duration of an epidemic by over 20% of the predicted duration. For $\pi/4$ a type II error in this setting will result in overestimating the total infections by nearly 10% of the total population and the duration by approximately 10% of the predicted one. A similar computation shows $\epsilon \approx .062$ when $\theta_0 = (0.21, 0.14)$ with all other parameters the same, and again from the left panels in Fig. 5 we observe significantly different predicted outcomes depending on whether or not the null hypothesis $H_0 : \theta = \theta_0$ is rejected.

While our empirical and theoretical results indicate the inverse problem of finding $\theta = (\beta, \gamma)$ is prone to error, they also show inference of δ is robust and reliable (see for instance Figs. 1 and 6). In particular, since $\theta = (\beta, \gamma)$ is completely determined by[3] γ and δ, knowledge of δ reduces the inverse problem to finding γ, the reciprocal of the average number of days an individual is infectious. In this case our new hypothesis test becomes

$$H_0 : \gamma = \gamma_0 \quad \text{vs.} \quad H_1 : \gamma = \gamma_0 + \hat{\epsilon}. \tag{12}$$

---

[3] We choose to focus on γ because, unlike β which depends on the average number of people an infectious person will come in contact with, γ depends only on the pathogen, not on human social behavior, and therefore tends to be more stable and better approximated in practice.

for some real number $\hat{\epsilon}$. Furthermore, knowing δ implies θ lies on the line of slope 1 with vertical intercept $-\delta$. So by a simple geometric argument (see Fig. 7), the above hypothesis test is equivalent to the hypothesis test (7) with $\epsilon = |\hat{\epsilon}|\sqrt{2}$ and angle $\pi/4$ if $\hat{\epsilon} > 0$ or $5\pi/4$ if $\hat{\epsilon} < 0$. So by (11) the type II error of (12) is
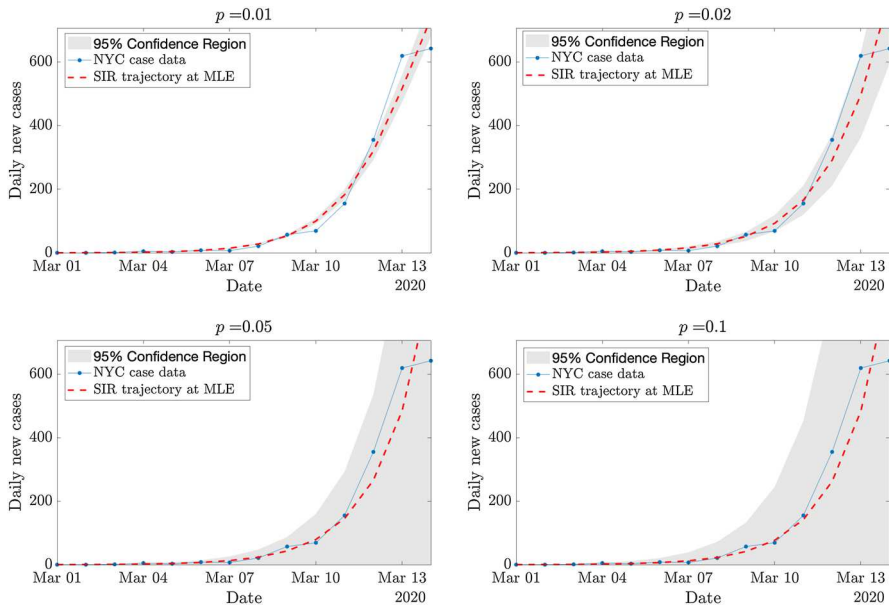
$$\mathcal{E}_2 = 1 - \Phi\left(\Phi^{-1}(\alpha) + \frac{p\epsilon\sqrt{T}}{\sigma\sqrt{2}}\right) = 1 - \Phi\left(\Phi^{-1}(\alpha) + \frac{p|\hat{\epsilon}|\sqrt{T}}{\sigma}\right).$$

Thus rather than consider the original hypothesis test, one can first infer δ, then consider the hypothesis test (12) with type II error rate as above.

3.4 Empirical analysis: NYC Covid cases, March 2020

In this section we discuss the extension of the theoretical results on parametric non-identifiability to a real world dataset. Consider the Spring 2020 COVID-19 outbreak in New York City. The New York City Health Department keeps a repository of all public COVID-19 data online [32]. Using their daily case data as a proxy for new infections, we directly apply Eq. (2) to the noisy data. To focus on estimation early in the pandemic, we focus on reported daily cases from February 29, 2020 through March 14, 2020, which approximately represent the first two weeks of the pandemic in New York City. This period precedes the statewide lockdown including the closing of schools on March 15th. However, the increasing awareness of COVID-19 and increased testing capacity strongly suggest that the contact rate β and reporting rate p were likely non-constant during this time. These parameters are also not jointly identifiable. Thus, we make the simplifying assumption that they are constant. Below, we show estimates of β, γ, and σ for fixed values of p ranging from 0.01 to 0.25 consistent with the current literature on the underreporting rates of COVID-19 infection (de Oliveira et al. 2020; Richterich 2020; Lau et al. 2021).

   To connect with the earlier analysis, we are following Case 2 as discussed in Sect. 3.2, in which the noise is proportional to number of infections, i.e. $\sigma_t = N\sigma i_t$ for

Fig. 8 New York City public testing results for COVID-19 from the first known case on February 29, 2020 to March 15, 2020. We have used maximum likelihood estimation to generate an SIR trajectory through the noisy data for each reporting rate

some $\sigma > 0$ which is also inferred via maximum likelihood. We thus model daily infections by

$$y_t = pN(s_{t-1} - s_t) + \sqrt{Ni_t}\epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2)$$

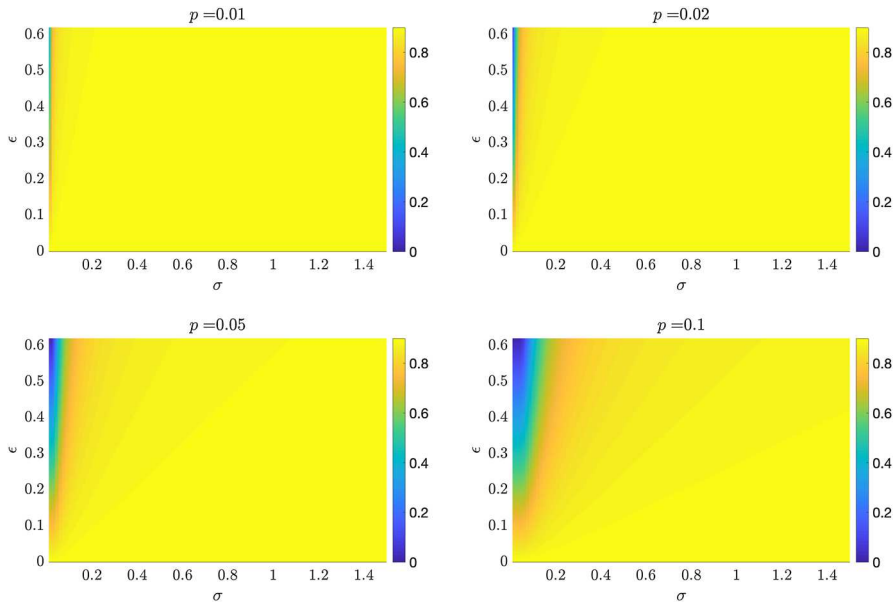from which we obtain the log-likelihood function

$$\ell(y_t | \beta, \gamma, \sigma) \approx -\frac{1}{2} \sum_{k=1}^{t} \frac{(y_k - pN(s_{k-1} - s_k))^2}{Ni_t \sigma^2}.$$

For a fixed value of $p = 0.05$, maximizing the above likelihood gives estimates

$$(\hat{\beta}, \hat{\gamma}, \hat{\sigma}) = (4.82, 4.22, 1.37)$$

and a corresponding estimate of $\hat{R}_0 = 1.14$. The SIR curve generated by the maximum likelihood estimates of β and γ is shown in Fig. 8 with corresponding 95% confidence regions based on the maximum likelihood estimate of σ. Additional results for $p = 0.01, 0.02,$ and 0.1 are also shown in the figure.

Returning to the testing framework, Fig. 9 provides type II error estimates based on the approximation of Eq. (11) with significance level $\alpha = 0.1$, reporting rates $p = 0.01, 0.02, 0.05,$ and 0.1, and $T = 14$ days of new infection counts. For all values of p considered herein, the MLE of $\hat{\sigma}$ is greater than 0.75 and corresponds with

Fig. 9 Type II error rate as a function of ϵ and σ at significance level $\alpha = 0.1$, reporting rate $p = 0.15$, and $T = 14$ days of new infection observations

Type II error greater than 80% for all values of ϵ such that $\hat{\theta}_\epsilon(\pi/4)$ or $\hat{\theta}_\epsilon(5\pi/4)$ with corresponding $R_0 > 1$.

Thus, while there may be a large disparity between the true SIR parameters and our maximum likelihood estimates—hence large differences in the estimate of R_0—the hypothesis testing framework has very low power to detect such differences. This result is based on the most difficult to detect perturbations in θ. However, it provides pessimistic but important lower bounds on the extent to which one can rely on parameter estimates from noisy, early pandemic data.

One final note on the preceding example. The MLEs of β and γ in the previous analysis are quite sensitive to the reporting rate. For reference, Table 3 provides corresponding MLE estimates for β, γ, and σ as a function of p.

However, the type II error plot in Fig. 9 is largely unchanged for the range of p in the preceding table. Since σ is fairly robust to different choices of p, our conclusion about the limited power of testing holds true for the range of p considered. Therefore, one has limited statistical power to detect large differences in SIR model parameters in the worst case scenario, regardless of the choice of reporting rate. As such, we believe this article serves as a cautionary tale to those fitting SIR-type models in the early days of an epidemic.

There is an important distinction to make. Having low power to detect a difference is not equivalent to being unable to tell that there is a difference. Certain parameter values are essentially impossible given natural assumptions about the dynamics of a pandemic. For example, $1/\gamma$ is the average time an infected individual can spread the disease before they are either recovered or removed from the population by quarantine.

Table 3 Maximum likelihood estimates of β, γ, and σ are shown for different choices of reporting rate p

p	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\sigma}$	$\hat{R}_0$
0.01	19.69	19.03	0.83	1.03
0.02	15.59	14.99	1.29	1.04
0.03	9.54	8.95	1.38	1.07
0.04	6.43	5.82	1.38	1.10
0.05	4.82	4.22	1.37	1.14
0.1	2.20	1.58	1.36	1.39
0.15	1.44	0.82	1.35	1.7
0.2	1.07	0.46	1.35	2.35
0.25	0.86	0.24	1.36	3.58

The corresponding estimates of R_0 using the MLEs is also provided

Extremely large values of γ and hence small values of $1/\gamma$, such as those attained in Table 3, are likely unrealistic. Thus, the inclusion of side or prior information on γ and/or β akin to the analysis in Sect. 3.3 can greatly improve one's ability to disambiguate different SIR parameters.

4 Discussion

The preceding analysis was based on a simple implementation of the SIR model. Practitioners studying future outbreaks may consider a multitude of modifications to our model construction which result in different likelihood functions. Thus, we have decided to conclude this article with a short discussion of how one may adapt our techniques to these different settings to better understand issues of practical identifiability with *noisy* or sparse observations.

To construct an analytically tractable approximation to the type II error, we assumed the proportion of susceptible individuals remains essentially 1 and thus obtained a linear system, namely (A1), that approximates the SIR equations. Such approximations are suitable locally in time and are therefore appropriate when one is focused on the early stages of an outbreak. Importantly, a similar approach can be used to construct analytic approximations to the dynamics of any epidemic model. Such expressions will depend on unknown SIR parameters, fixed parameters such as population size, and other parameters such as reporting rate or behavioral factors, as in Cori et al. (2020). For example, Britton and Scalia Tomba (2019) assume the proportion of susceptible individuals remains 1 early in an epidemic to study the problem of inferring infection rate from observations of generation and serial times, which are often available via contact tracing. In all cases one can investigate the use of this and other realistic simplifications of the dynamics to approximate type II error and better understand potential limitations of their particular model. We believe this approach remains an interesting, potentially fruitful avenue toward understanding identifiability in a wide array of epidemic models.

On the theoretical side, the upper bound for the error term in Proposition 1 of "Appendix C" gives some rigorous justification for the approximate dynamics used throughout this work. However, the numerical results of the main text and "Appendix B" indicate the approximation is more accurate than the theoretical bound suggests. It is therefore an open question whether our theoretical bound on the approximation error can be improved upon, perhaps via other approximate solutions of the SIR model found in, for example, (Turkyilmazoglu 2021; Barlow and Weinstein 2020; Schlickeiser and Kröger 2021) and references therein. Finally, since the approximation of s by 1 is used in other models and to investigate other questions about epidemics (Sauer et al. 2020; Britton and Scalia Tomba 2019), it is also of interest whether estimates of the error in our setting can be used to control error for similar approximations in related settings.

Data availability Data used in this paper are freely available online at https://github.com/nychealth/coronavirus-data. See [32] in the References section below.

Declarations

Conflict of interest The authors declare no competing interests.

Appendix A: Derivation of Approximation 1

The main observation leading to (3) is that s remains close to 1 early in the epidemic. Motivated by this, we replace s with 1 in the SIR model to obtain

$$\frac{ds}{dt} = -\beta i, \quad \frac{di}{dt} = (\beta - \gamma)i = \delta i. \tag{A1}$$

The corresponding solution starting from $x_0 = (s_0, i_0)$ is

$$\widetilde{\varphi}_t(x_0, \theta) = \left(s_0 - \frac{\beta}{\delta}\left(e^{\delta t} - 1\right)i_0, \ e^{\delta t}i_0\right). \tag{A2}$$

Fix $\omega \in [0, 2\pi)$ and set $\varphi_t = \varphi_t(x_0, \theta)$, $\varphi_t^\epsilon = \varphi_t(x_0, \theta_\epsilon(\omega))$, $\widetilde{\varphi}_t = \widetilde{\varphi}_t(x_0, \theta)$, and $\widetilde{\varphi}_t^\epsilon = \widetilde{\varphi}_t(x_0, \theta_\epsilon(\omega))$. The expression of interest, $\|\varphi_t^\epsilon - \varphi_t\|$, can be written

$$\|\varphi_t^\epsilon - \varphi_t\| = \|\widetilde{\varphi}_t^\epsilon - \widetilde{\varphi}_t\| + E_t^\epsilon \tag{A3}$$

where E_t^ϵ is the error incurred from approximating $\|\varphi_t^\epsilon - \varphi_t\|$ by $\|\widetilde{\varphi}_t^\epsilon - \widetilde{\varphi}_t\|$. As mentioned in the main text, we do not have explicit analytic control over E_t^ϵ but the numerical analysis in "Appendix B" indicates it is negligible compared to $\|\widetilde{\varphi}_t^\epsilon - \widetilde{\varphi}_t\|$ early in the epidemic. In particular, the approximation

$$\|\widetilde{\varphi}_t^\epsilon - \widetilde{\varphi}_t\| \approx \|\varphi_t^\epsilon - \varphi_t\| \tag{A4}$$

is valid up to roughly 80% of the time of peak infection. Thus we turn attention to $\widetilde{\varphi}_t^\epsilon - \widetilde{\varphi}_t$. Fixing t and Taylor expanding $\widetilde{\varphi}_t$ to first order in $\eta = (\beta, \delta)$ gives

$$\widetilde{\varphi}_t^\epsilon - \widetilde{\varphi}_t = D_\eta \widetilde{\varphi}_t(x, \eta)\big(\eta_\epsilon(\omega) - \eta\big) + o\big(\|\eta\|^2\big)$$

$$\approx \epsilon \begin{pmatrix} -\frac{1}{\delta}(e^{\delta t} - 1)i_0 & -\frac{\beta}{\delta}\Big(te^{\delta t} - \frac{1}{\delta}(e^{\delta t} - 1)\Big)i_0 \\ 0 & te^{\delta t} i_0 \end{pmatrix} \begin{pmatrix} \cos(\omega) \\ \cos(\omega) - \sin(\omega) \end{pmatrix}$$

$$\approx \epsilon i_0 \begin{pmatrix} -t & -\beta t^2 \\ 0 & te^{\delta t} \end{pmatrix} \begin{pmatrix} \cos(\omega) \\ \cos(\omega) - \sin(\omega) \end{pmatrix}$$

where D_η is the derivative in η and $\eta_\epsilon(\omega) = \eta + \epsilon(\cos(\omega) - \sin(\omega))$. The first approximation is a result of simply dropping the $o(\|\eta\|^2)$ term and the second is obtained by substituting the first order Taylor approximation $\exp(\delta t) - 1 = \delta t + o(\delta^2)$ about $\delta = 0$ into the expression above it. We see from the latter expression that any vector (x_1, x_2) in $\mathbb{R}^2$ with $x_2 \neq 0$ will grow exponentially in time under the above matrix due to the $te^{\delta t}$ term in the bottom right. On the other hand, the first component x_1 will only grow linearly in time provided $x_2 = 0$. So the magnitude of growth is minimized for vectors of the form $(x_1, 0)$. This implies that, under the above approximations which come at a cost of $o(\|\eta\|^2)$ and $o(\delta^2)$, respectively, the difference $\tilde{\varphi}_t^\epsilon - \tilde{\varphi}_t$ will grow the least when $(\cos(\omega), \cos(\omega) - \sin(\omega)) = \pm(1, 0)$. Therefore the perturbations $\eta_\epsilon(\omega)$ that yield the smallest separation between $\widetilde{\varphi}_t^\epsilon$ and $\widetilde{\varphi}_t$ are those corresponding to the directions $\pm(1, 0)$ or, equivalently, the angles $\omega = \pi/4$ and $5\pi/4$. Now

$$\widetilde{\varphi}_t^\epsilon(x_0, \theta_\epsilon(\omega)) - \widetilde{\varphi}_t(x_0, \theta) = \begin{pmatrix} \frac{\beta_\epsilon}{\delta_\epsilon} - \frac{\beta}{\delta} + \big(\frac{\beta}{\delta} - \frac{\beta_\epsilon}{\delta_\epsilon} e^{\epsilon t(\cos\omega - \sin\omega)}\big)e^{\delta t} \\ \big(e^{\epsilon t(\cos\omega - \sin\omega)} - 1\big)e^{\delta t} \end{pmatrix} i_0 \quad \text{(A5)}$$

where $\beta_\epsilon = \beta + \epsilon\cos\omega$ and $\delta_\epsilon = \delta + \epsilon(\cos\omega - \sin\omega)$. Plugging in $\pi/4$ gives

$$\widetilde{\varphi}_t^\epsilon(x_0, \theta_\epsilon(\omega_1)) - \widetilde{\varphi}_t(x_0, \theta) = \begin{pmatrix} 1 - e^{\delta t} \\ 0 \end{pmatrix} \frac{\epsilon i_0}{\delta\sqrt{2}} \quad \text{(A6)}$$

and similarly for $5\pi/4$, only negative. Thus, in combination with (A4),

$$\frac{\epsilon}{\delta\sqrt{2}}(e^{\delta t} - 1)i_0 \lesssim \|\widetilde{\varphi}_t^\epsilon - \widetilde{\varphi}_t\| \approx \|\varphi_t^\epsilon - \varphi_t\|$$

for all ω, which is precisely (3).

Appendix B: Numerical analysis of error

In this section we revisit the expression

$$\|\varphi - \varphi_t\| = \|\widetilde{\varphi}_t^\epsilon - \widetilde{\varphi}_t\| + E_t^\epsilon$$

which is (A3) in "Appendix A"; see the sentences preceding (A3) for notation. A key requirement in the derivation of Approximation 1 is that $\|\varphi_t^\epsilon - \varphi_t\|$ is well-approximated by $\|\widetilde{\varphi}_t^\epsilon - \widetilde{\varphi}_t\|$. By (A3) this is the case whenever

$$\frac{|E_t^\epsilon|}{\|\varphi_t^\epsilon - \varphi_t\|} < \tau \tag{B7}$$

for some prescribed tolerance τ, which we take to be 1 both for simplicity and because it agrees with the numerical observations from Figs. 2, 3. As mentioned before, we do not have sufficient control to verify (B7) analytically. However, we can verify (B7) numerically – in our case with the odeint function from the scipy.integrate package in Python. Specifically, we numerically solve both the SIR and approximate SIR equations and compute

$$\log|E_t^\epsilon| - \log\|\varphi_t^\epsilon - \varphi_t\| \tag{B8}$$

for 25 evenly spaced angles ω in the interval $[\pi/4 - \pi/12, \pi/4 + \pi/12)$ and 25 more such angles in $[5\pi/4 - \pi/12, 5\pi/4 + \pi/12)$. We restrict attention to these intervals because, as observed in Figs. 2, 3, these are the angles for which the inverse problem is least practically identifiable. The results, plotted in Fig. 10, indicate (B8) grows approximately linearly in time. For each of the 16 subplots in Fig. 10 (which correspond to 16 different β, γ, ϵ, and N combinations) we average the linear approximations for each of the 50 curves to obtain a single "average" linear approximation indicated by the red line in each subplot. Table 4 gives the equations for each of these averaged lines. So, for example, when $(\beta, \gamma) = (.21, .07)$, $\epsilon = .03$, and $N = 10^6$ the table gives
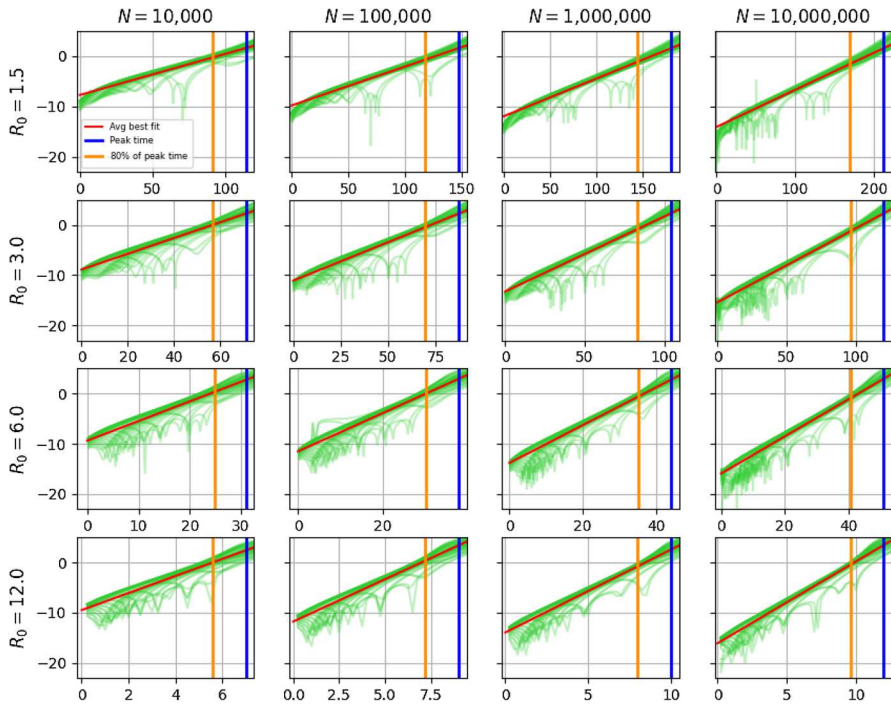
$$\log|E_t^\epsilon| - \log\|\varphi_t^\epsilon - \varphi_t\| \approx 0.15t - 13.3.$$

Setting this expression equal to 0 (which corresponds to $\tau = 1$) and solving for t gives $t = 13.3/0.15 \approx 89$ days, i.e. the error becomes intolerable at roughly 89 days. We can then divide 89 by the peak time to estimate the percentage of the peak time at which the error becomes sufficiently large that the approximation, and hence the lower bound from Approximation 1, no longer holds. The resulting percentages are given in Table 5 and motivate our choice of 80% in Approximation 1.

Appendix C: Theoretical analysis of error

In this section we again revisit (A3) and prove the following.

Proposition 1 *For every initial condition $x_0 = (s_0, i_0)$, parameter pair (β, γ), $\epsilon > 0$, ω in $[0, 2\pi)$, and $t \geq 0$,*

Fig. 10 Error analysis corresponding to Figs. 2, 3. As in Figs. 2, 3, the horizontal axis in each graph is the number of days since the start of the epidemic, the vertical blue line is the peak time, and the orange line is 80% of the peak time. The vertical axis is the logarithm of the relative error, $\log|E_t^\epsilon| - \log\|\varphi_t^\epsilon - \varphi_t\|$, at time t. There are 50 green curves in each plot which are the relative errors corresponding to 50 angles in the intervals $[\pi/4 - \pi/12, \pi/4 + \pi/12]$ and $[5\pi/4 - \pi/12, 5\pi/4 + \pi/12]$. The red line in each plot is the average linear approximation of the 50 green curves. Also as in Figs. 2, 3, the first through fourth columns have population sizes 10^4, 10^5, 10^6, and 10^7, respectively with one initial infection in each case, and the first through fourth rows have parameters $(\beta, \gamma, \epsilon) = (.21, .14, .03)$, $(.21, .07, .03)$, $(.42, .07, .06)$, and $(1.68, .14, .1)$ which give R_0 values of 1.5, 3, 6, and 12. The main point is that for every combination of parameters the relative error is exponentially small until about 80% of the peak time at which point it becomes $\mathcal{O}(1)$ and subsequently blows up exponentially

$$|E_t^\epsilon| \leq \left(\frac{\sqrt{2\beta_\epsilon^2 + \gamma_\epsilon^2}}{\delta_\epsilon} \left(e^{\delta_\epsilon t} - 1 \right) + \frac{\sqrt{2\beta^2 + \gamma^2}}{\delta} \left(e^{\delta t} - 1 \right) \right) i_0, \qquad (C9)$$

where, as before, $\delta = \beta - \gamma$, $\beta_\epsilon = \beta + \epsilon \cos(\omega)$, $\gamma_\epsilon = \gamma + \epsilon \sin(\omega)$, and $\delta_\epsilon = \beta_\epsilon - \gamma_\epsilon$.

Proof By (A3), the reverse triangle inequality, and the triangle inequality,

$$\begin{aligned} |E_t^\epsilon| = \left| \|\varphi_t^\epsilon - \varphi_t\| - \|\widetilde{\varphi}_t^\epsilon - \widetilde{\varphi}_t\| \right| &\leq \|\varphi_t^\epsilon - \varphi_t - \widetilde{\varphi}_t^\epsilon - \widetilde{\varphi}_t\| \\ &\leq \|\widetilde{\varphi}_t^\epsilon - \varphi_t^\epsilon\| + \|\widetilde{\varphi}_t - \varphi_t\|. \end{aligned} \qquad (C10)$$

Table 4 Linear approximations of log relative error

$(\beta, \gamma, \epsilon)$	$N = 10^4$	$N = 10^5$	$N = 10^6$	$N = 10^7$
$(.21, .14, .03)$	$.08t - 7.7$	$.08t - 9.8$	$.08t - 11.9$	$.07t - 14.1$
$(.21, .07, .03)$	$.16t - 8.9$	$.15t - 11.1$	$.15t - 13.3$	$.15t - 15.4$
$(.42, .07, .06)$	$.39t - 9.4$	$.38t - 11.5$	$.38t - 13.8$	$.37t - 15.9$
$(1.68, .14, .1)$	$1.7t - 9.5$	$1.68t - 11.8$	$1.65t - 14$	$1.63t - 16.1$

Equations for the average linear approximations to the log relative error $\log|E_t^\epsilon| - \log\|\varphi_t^\epsilon - \varphi_t\|$ as functions of time for each combination of parameter values appearing in Figs. 2, 3 and 10

Table 5 *Percent of peak time at which error becomes intolerable*

$(\beta, \gamma, \epsilon)$	$N = 10^4$	$N = 10^5$	$N = 10^6$	$N = 10^7$
$(.21, .14, .03)$	83%	86%	88%	90%
$(.21, .07, .03)$	80%	83%	85%	87%
$(.42, .07, .06)$	77%	80%	83%	84%
$(1.68, .14, .1)$	80%	78%	85%	83%

Approximate percent of peak time when the log relative error $\log|E_t^\epsilon| - \log\|\varphi_t^\epsilon - \varphi_t\|$ becomes intolerable, i.e. $\mathcal{O}(1)$

Next, note that $\widetilde{\varphi}_t - \varphi_t = \int_0^t \frac{d}{d\tau}(\widetilde{\varphi}_\tau - \varphi_\tau)d\tau$ and hence

$$
\begin{aligned}
\|\widetilde{\varphi}_t - \varphi_t\| &\leq \int_0^t \left\| \frac{d}{d\tau}(\widetilde{\varphi}_\tau - \varphi_\tau) \right\| d\tau \\
&= \int_0^t \|(-\beta(\widetilde{\iota}_\tau - i_\tau s_\tau), \beta(\widetilde{\iota}_\tau - i_\tau s_\tau) - \gamma(\widetilde{\iota}_\tau - i_\tau))\| d\tau \\
&= \int_0^t \sqrt{2\beta^2(\widetilde{\iota}_\tau - i_\tau s_\tau)^2 - 2\beta\gamma(\widetilde{\iota}_\tau - i_\tau s_\tau)(\widetilde{\iota}_\tau - i\tau) + \gamma^2(\widetilde{\iota}_\tau - i_\tau)^2}\, d\tau \\
&\leq \int_0^t \widetilde{\iota}_\tau\sqrt{2\beta^2 + \gamma^2}\, d\tau = \frac{\sqrt{2\beta^2 + \gamma^2}}{\delta}\left(e^{\delta t} - 1\right)i_0.
\end{aligned}
$$

The first equality holds because $\widetilde{\varphi}$ and φ are solutions of (A1) and (1), respectively, and the second inequality holds because $\widetilde{\iota}_t \geq i_t \geq i_t s_t$ for all $t \geq 0$ and hence

$$
2\beta^2(\widetilde{\iota}_\tau - i_\tau s_\tau)^2 - 2\beta\gamma(\widetilde{\iota}_\tau - i_\tau s_\tau)(\widetilde{\iota}_\tau - i\tau) + \gamma^2(\widetilde{\iota}_\tau - i_\tau)^2 \leq (2\beta^2 + \gamma^2)\widetilde{\iota}_\tau^2 \tag{C11}
$$

for every $\tau \geq 0$. Finally, the last equality follows from $\widetilde{\iota}_\tau = e^{\delta\tau}i_0$. Since the above holds for arbitrary β and γ, we also have

$$
\|\widetilde{\varphi}_t^\epsilon - \varphi_t^\epsilon\| \leq \frac{\sqrt{2\beta_\epsilon^2 + \gamma_\epsilon^2}}{\delta_\epsilon}\left(e^{\delta_\epsilon t} - 1\right)i_0.
$$

Combining with (C10) therefore gives

$$|E_t^\epsilon| \le \left(\frac{\sqrt{2\beta_\epsilon^2 + \gamma_\epsilon^2}}{\delta_\epsilon} \left(e^{\delta_\epsilon t} - 1 \right) + \frac{\sqrt{2\beta^2 + \gamma^2}}{\delta} \left(e^{\delta t} - 1 \right) \right) i_0,$$

as claimed.   □

A few remarks are in order. First, if $i_0 = c/N$ for some constant c, then i_0, and hence E_t^ϵ for any finite t, go to 0 as the population size N goes to infinity. Thus Approximation 1 improves with increasing N, which agrees with the numerical results of the last section. Second, the bound (C9) is not sharp. This is due to the use of the triangle and reverse triangle inequalities, as well as the other inequalities implemented in the above proof. In particular, inequality (C11) is quite coarse because we have essentially thrown away all i_τ and s_τ terms. This is rather necessary since the SIR model has no known exact analytic solution, so gaining explicit control over terms involving i_τ and s_τ is, to the best of our knowledge, largely intractable (Barlow and Weinstein 2020). Finally, the upper bound (C9) grows exponentially in time. This is a byproduct of the fact that the proportion of infected individuals $\tilde{\iota}$ in the approximate dynamics (A1) grows exponentially without bound. On the other hand, growth of the i compartment in the true SIR dynamics (1) is offset by the decreasing number of susceptible individuals as the epidemic progresses, i.e. fewer susceptible individuals means there are fewer people to infect. This is the primary reason why the approximate dynamics are only valid early in the epidemic, and certainly do not hold beyond the time of peak infection.

Appendix D: Derivation of Approximation 2

Fix $\omega \in [0, 2\pi)$ and set $\theta_\epsilon = \theta_\epsilon(\omega)$, $\Delta_t^\epsilon = \Delta_t(\theta_\epsilon)$, and $\Delta_t^0 = \Delta_t(\theta_0)$. The likelihood for observed data $Y_{1:T}$ is

$$L(Y_{1:T}|\theta_0) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{1}{2\sigma_t^2}(Y_t - p\Delta_t^0)^2 \right)$$

and similarly for $L(Y_{1:T}|\theta_\epsilon)$. So the log-likelihood ratio between θ_ϵ and θ_0 is

$$\log\left(\frac{L(Y_{1:T}|\theta_\epsilon)}{L(Y_{1:T}|\theta_0)} \right) = \sum_{t=1}^{T} \frac{1}{2\sigma_t^2}\left[(Y_t - p\Delta_t^0)^2 - (Y_t - p\Delta_t^\epsilon)^2 \right]$$

$$= \sum_{t=1}^{T} \frac{1}{2\sigma_t^2}\left[2pY_t(\Delta_t^\epsilon - \Delta_t^0) - p^2((\Delta_t^\epsilon)^2 - (\Delta_t^0)^2) \right].$$

If the Y_t satisfy (2) with parameter θ_ϵ, i.e. $Y_t = p\Delta_t^\epsilon + \xi_t$, then

$$2pY_t(\Delta_t^\epsilon - \Delta_t^0) - p^2((\Delta_t^\epsilon)^2 - (\Delta_t^0)^2) = p^2(\Delta_t^\epsilon - \Delta_t^0)^2 + 2p(\Delta_t^\epsilon - \Delta_t^0)\xi_t.$$

So, letting η be the value in (6) for the likelihood ratio test,

$$
\begin{aligned}
\mathcal{E}_2(\omega) &= \mathbb{P}_{\theta_\epsilon}\left(\frac{L(Y_{1:T}|\theta_\epsilon)}{L(Y_{1:T}|\theta_0)} < \eta\right) = \mathbb{P}_{\theta_\epsilon}\left(\log\left(\frac{L(Y_{1:T}|\theta_\epsilon)}{L(Y_{1:T}|\theta_0)}\right) < \log\eta\right) \\
&= \mathbb{P}_{\theta_\epsilon}\left(\sum_{t=1}^{T}\frac{1}{2\sigma_t^2}\left[p^2(\Delta_t^\epsilon - \Delta_t^0)^2 + 2p(\Delta_t^\epsilon - \Delta_t^0)\xi_t\right] < \log\eta\right) \\
&= \mathbb{P}_{\theta_\epsilon}\left(\sum_{t=1}^{T}\frac{p}{\sigma_t^2}(\Delta_t^\epsilon - \Delta_t^0)\xi_t < \log\eta - \frac{1}{2}V_T^\epsilon(\omega)\right).
\end{aligned}
$$

where

$$
V_T^\epsilon(\omega) = \sum_{t=1}^{T}\frac{p^2}{\sigma_t^2}(\Delta_t^\epsilon - \Delta_t^0)^2. \tag{D12}
$$

Now $\xi_t \sim \mathcal{N}(0, \sigma_t^2)$ implies $\sum_{t=1}^{T}(p/\sigma_t^2)(\Delta_t^\epsilon - \Delta_t^0)\xi_t \sim \mathcal{N}(0, V_T^\epsilon)$ and hence

$$
\frac{1}{\sqrt{V_T^\epsilon(\omega)}}\sum_{t=1}^{T}\frac{p}{\sigma_t^2}(\Delta_t^\epsilon - \Delta_t^0)\xi_t \sim \mathcal{N}(0, 1).
$$

So, letting Z denote a standard normal random variable,

$$
\mathcal{E}_2(\omega) = \mathbb{P}_{\theta_\epsilon}\left(Z < \frac{\log\eta}{\sqrt{V_T^\epsilon(\omega)}} - \frac{1}{2}\sqrt{V_T^\epsilon(\omega)}\right) = \Phi\left(\frac{\log\eta}{\sqrt{V_T^\epsilon(\omega)}} - \frac{1}{2}\sqrt{V_T^\epsilon(\omega)}\right).
$$

By an entirely similar computation (note the symmetry between θ_ϵ and θ_0),

$$
\alpha = \mathbb{P}_{\theta_0}\left(\frac{L(Y_{1:T}|\theta_\epsilon)}{L(Y_{1:T}|\theta_0)} \geq \eta\right) = \Phi\left(-\frac{\log\eta}{\sqrt{V_T^\epsilon(\omega)}} - \frac{1}{2}\sqrt{V_T^\epsilon(\omega)}\right)
$$

and so $\frac{\log\eta}{\sqrt{V_T^\epsilon(\omega)}} = -\Phi^{-1}(\alpha) - \frac{1}{2}\sqrt{V_T^\epsilon(\omega)}$. Therefore, since $\Phi(-x) = 1 - \Phi(x)$,

$$
\mathcal{E}_2(\omega) = \Phi\left(-\Phi^{-1}(\alpha) - \sqrt{V_T^\epsilon(\omega)}\right) = 1 - \Phi\left(\Phi^{-1}(\alpha) + \sqrt{V_T^\epsilon(\omega)}\right). \tag{D13}
$$

To obtain the approximations, first note that from (A2) we have

$$
\Delta_t^\epsilon(\omega) - \Delta_t^0 \approx \left[\beta_\epsilon\left(\frac{e^{-\delta_\epsilon} - 1}{-\delta_\epsilon}\right)e^{\epsilon(\cos\omega - \sin\omega)t} - \beta\left(\frac{e^{-\delta} - 1}{-\delta}\right)\right]Ni_0 e^{\delta t}, \tag{D14}
$$

where $\beta_\epsilon = \beta + \epsilon \cos \omega$ and $\delta_\epsilon = \delta + \epsilon(\cos \omega - \sin \omega)$. This can be further simplified by using the Taylor approximation $(e^x - 1)/x \approx 1$ to obtain

$$\Delta_t^\epsilon(\omega) - \Delta_t^0 \approx \left[\beta_\epsilon e^{\epsilon(\cos \omega - \sin \omega)t} - \beta \right] Ni_0 e^{\delta t}. \tag{D15}$$

Plugging (D14) and subsequently (D15) into (D13) then gives

$$\mathcal{E}_2(\omega) \approx 1 - \Phi \left(\Phi^{-1}(\alpha) + pNi_0 \sqrt{\sum_{t=1}^{T} \frac{e^{2\delta t}}{\sigma_t^2} \left[\beta_\epsilon \left(\frac{e^{-\delta_\epsilon} - 1}{-\delta_\epsilon} \right) e^{\epsilon(\cos \omega - \sin \omega)t} - \beta \left(\frac{e^{-\delta} - 1}{-\delta} \right) \right]^2} \right)$$

$$\approx 1 - \Phi \left(\Phi^{-1}(\alpha) + pNi_0 \sqrt{\sum_{t=1}^{T} \frac{e^{2\delta t}}{\sigma_t^2} \left[(\beta + \epsilon \cos \omega) e^{\epsilon(\cos \omega - \sin \omega)t} - \beta \right]^2} \right),$$

as claimed. Finally, from Approximation 1 we know the directions of least separation between the trajectories corresponding to θ_0 and $\theta_\epsilon(\omega)$ are approximately $\omega = \pi/4$ and $5\pi/4$. This suggests it will be most difficult to distinguish the null and alternative hypotheses of (7) when $\omega = \pi/4$ or $5\pi/4$. So when constrained to a significance level α, one of these two angles will approximately maximize the type II error rate $\mathcal{E}_2(\omega)$ over all $\omega \in [0, 2\pi)$.

References

Balsa-Canto E, Alonso AA, Banga JR (2008) Computational procedures for optimal experimental design in biological systems. IET Syst Biol 2(4):163–172. https://doi.org/10.1049/iet-syb:20070069

Balsa-Canto E, Alonso A, Banga J (2009) An iterative identification procedure for dynamic modeling of biochemical networks. BMC Syst Biol 4:11–11

Barlow NS, Weinstein SJ (2020) Accurate closed-form solution of the sir epidemic model. Phys D 408:132540

Bellman R, Åström KJ (1970) On structural identifiability. Math Biosci 7(3):329–339. https://doi.org/10.1016/0025-5564(70)90132-X

Bellu G, Saccomani MP, Audoly S, D'Angiò L (2007) DAISY: a new software tool to test global identifiability of biological and physiological systems. Comput Methods Programs Biomed 88(1):52–61. https://doi.org/10.1016/j.cmpb.2007.07.002

Brauer F, Castillo-Chavez C, Feng Z (2019) Mathematical models in epidemiology. Texts in applied mathematics, vol 69. Springer, New York, NY. https://doi.org/10.1007/978-1-4939-9828-9

Britton T, Scalia Tomba G (2019) Estimation in emerging epidemics: biases and remedies. J R Soc Interface 16(150):20180670

Browning AP, Warne DJ, Burrage K, Baker RE, Simpson MJ (2020) Listen to the noise: identifiability analysis for stochastic differential equation models in systems biology. bioRxiv. https://doi.org/10.1101/2020.08.10.245233

Brunel NJB (2008) Parameter estimation of ODE's via nonparametric estimators. Electron J Stat 2(March 2007):1242–1267. https://doi.org/10.1214/07-EJS132

Chapman JD, Evans ND (2009) The structural identifiability of susceptible-infective-recovered type epidemic models with incomplete immunity and birth targeted vaccination. Biomed Signal Process Control 4(4):278–284. https://doi.org/10.1016/j.bspc.2009.02.003

Chis O-T, Banga JR, Balsa-Canto E (2011) Structural identifiability of systems biology models: a critical comparison of methods. PLoS ONE 6(11):27755. https://doi.org/10.1371/journal.pone.0027755

Coburn BJ, Wagner BG, Blower S (2009) Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1). BMC Med 7:30. https://doi.org/10.1186/1741-7015-7-30

Cori L, Bianchi F, Cadum E, Anthonj C (2020) Risk perception and covid-19. Int J Environ Res Public Health 17(9):3114

Daly AC, Gavaghan D, Cooper J, Tavener S (2018) Inference-based assessment of parameter identifiability in nonlinear biological models. J R Soc Interface. https://doi.org/10.1098/rsif.2018.0318

de Oliveira ACS, Morita LHM, da Silva EB, Zardo LAR, Fontes CJF, Granzotto DCT (2020) Bayesian modeling of covid-19 cases with a correction to account for under-reported cases. Infect Dis Model 5:699–713. https://doi.org/10.1016/j.idm.2020.09.005

Eisenberg MC, Robertson SL, Tien JH (2013) Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. J Theor Biol 324:84–102. https://doi.org/10.1016/j.jtbi.2012.12.021

Fok P-W, Chou T (2013) Identifiability of age-dependent branching processes from extinction probabilities and number distributions. J Stat Phys 152(4):769–786. https://doi.org/10.1007/s10955-013-0781-1

Gronwall TH (1919) Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. Ann Math 20(4):292. https://doi.org/10.2307/1967124

Hamelin F, Iggidr A, Rapaport A, Sallet G, Sallet Observability G, Hamelin F, Iggidr A, Rapaport A, Sallet G (2020) Observability, Identifiability and Epidemiology A survey. Technical report . https://hal.archives-ouvertes.fr/hal-02995562

Kermack WO, Mckendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Character 115(772):700–721. https://doi.org/10.1098/rspa.1927.0118

Khaleque A, Sen P (2017) An empirical analysis of the Ebola outbreak in West Africa. Sci Rep. https://doi.org/10.1038/srep42594

Laredo C, David O, Garnier A (2009) Inference for partially observed multitype branching processes and ecological applications. https://doi.org/10.48550/ARXIV.0902.4520

Lau H, Khosrawipour T, Kocbach P, Ichii H, Bania J, Khosrawipour V (2021) Evaluating the massive underreporting and undertesting of covid-19 cases in multiple global epicenters. Pulmonology 27(2):110–115. https://doi.org/10.1016/j.pulmoe.2020.05.015

Lee C, Li Y, Kim J (2020) The susceptible-unidentified infected-confirmed (SUC) epidemic model for estimating unidentified infected population for COVID-19. Chaos, Solitons Fractals 139:110090. https://doi.org/10.1016/j.chaos.2020.110090

NYC Health Department: NYC Coronavirus Disease 2019 (COVID-19) Data. GitHub (2021). https://github.com/nychealth/coronavirus-data

Pasquali S, Pievatolo A, Bodini A, Ruggeri F (2021) A stochastic SIR model for the analysis of the COVID-19 Italian epidemic arXiv:2102.0756

Piazzola C, Tamellini L, Tempone R (2020) A note on tools for prediction under uncertainty and identifiability of SIR-like dynamical systems for epidemiology. Math Biosci. https://doi.org/10.1016/j.mbs.2020.108514

Rachah A, Torres DFM (2015) Mathematical modelling, simulation, and optimal control of the 2014 ebola outbreak in West Africa. Discret. Dyn. Nat. Soc. **2015**https://doi.org/10.1155/2015/842792, arXiv:1503.07396

Richterich P (2020) Severe underestimation of covid-19 case numbers: effect of epidemic growth rate and test restrictions. medRxiv. https://doi.org/10.1101/2020.04.13.20064220

Ross L-CSR (1916) An application of the theory of probabilities to the study of a priori pathometry.-Part I. Proc. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Character 92(638):204–230. https://doi.org/10.1098/rspa.1916.0007

Ross L-CSR, Hudson HP (1917) An application of the theory of probabilities to the study of a priori pathometry.-Part III. Proc. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Character 93(650):225–240. https://doi.org/10.1098/rspa.1917.0015

Ross R, Hudson HP (1917) An application of the theory of probabilities to the study of a priori pathometry.-Part II. Proc. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Character 93(650):212–225. https://doi.org/10.1098/rspa.1917.0014

Sauer T, Berry T, Ebeigbe D, Norton M, Whalen A, Schiff S (2020) Identifiability of infection model parameters early in an epidemic. medRxiv. https://doi.org/10.1101/2020.06.15.20132217

Schlickeiser R, Kröger M (2021) Analytical solution of the SIR-model for the temporal evolution of epidemics: part b. Semi-time case. J Phys A Math Theor 54(17):175601. https://doi.org/10.1088/1751-8121/abed66

Srinath S, Gunawan R (2010) Parameter identifiability of power-law biochemical system models. J Biotechnol 149(3):132–140. https://doi.org/10.1016/j.jbiotec.2010.02.019

Tuncer N, Le TT (2018) Structural and practical identifiability analysis of outbreak models. Math Biosci 299(February):1–18. https://doi.org/10.1016/j.mbs.2018.02.004

Tuncer N, Gulbudak H, Cannataro VL, Martcheva M (2016) Structural and practical identifiability issues of immuno-epidemiological vector-host models with application to rift valley fever. Bull Math Biol 78(9):1796–1827. https://doi.org/10.1007/s11538-016-0200-2

Turkyilmazoglu M (2021) Explicit formulae for the peak time of an epidemic from the SIR model. Phys D Nonlinear Phenom 422:132902. https://doi.org/10.1016/j.physd.2021.132902

Villaverde AF (2018) Observability and structural identifiability of nonlinear biological systems. arXiv **2019**arXiv:1812.04525

Yang X, Wang S, Xing Y, Li L, Da Xu RY, Friston KJ, Guo Y (2020) Revealing the transmission dynamics of COVID-19: a bayesian framework for R_t estimation arXiv:2101.01532