MATHEMATICAL BIOSCIENCES
AND ENGINEERING
Volume **9**, Number **3**, July **2012**

doi:10.3934/mbe.2012.9.553

pp. **553–576**

PARAMETER ESTIMATION AND UNCERTAINTY QUANTIFICATION FOR AN EPIDEMIC MODEL

Alex Capaldi

Center for Quantitative Sciences in Biomedicine and Department of Mathematics
North Carolina State University, Raleigh, NC 27695, USA

Current address:
Department of Mathematics & Computer Science, Valparaiso University
1900 Chapel Drive, Valparaiso, IN 46383, USA

Samuel Behrend

Department of Mathematics, University of North Carolina, Chapel Hill
CB #3250, Chapel Hill, NC 27599, USA

Benjamin Berman

Program in Applied Mathematics, University of Arizona
617 N. Santa Rita Ave., PO Box 210089, Tucson, AZ 85721-0089, USA

Jason Smith

Department of Mathematics, Morehouse College
830 Westview Drive SW Unit 142133, Atlanta, GA 30314, USA

Justin Wright

Department of Mathematics, North Carolina State University
Raleigh, NC 27695, USA

Alun L. Lloyd

Biomathematics Graduate Program and Department of Mathematics
North Carolina State University, Raleigh NC, 27695, USA

and

Fogarty International Center, National Institutes of Health
Bethesda, MD 20892, USA

(Communicated by Abba Gumel)

2000 *Mathematics Subject Classification.* Primary: 92D30; Secondary: 62F99, 62P10, 65L09.

Key words and phrases. Inverse problem, sampling methods, asymptotic statistical theory, sensitivity analysis, parameter identifiability.

Abstract. We examine estimation of the parameters of Susceptible-Infective-Recovered (SIR) models in the context of least squares. We review the use of asymptotic statistical theory and sensitivity analysis to obtain measures of uncertainty for estimates of the model parameters and the basic reproductive number (R_0)—an epidemiologically significant parameter grouping. We find that estimates of different parameters, such as the transmission parameter and recovery rate, are correlated, with the magnitude and sign of this correlation depending on the value of R_0. Situations are highlighted in which this correlation allows R_0 to be estimated with greater ease than its constituent parameters. Implications of correlation for parameter identifiability are discussed. Uncertainty estimates and sensitivity analysis are used to investigate how the frequency at which data is sampled affects the estimation process and how the accuracy and uncertainty of estimates improves as data is collected over the course of an outbreak. We assess the informativeness of individual data points in a given time series to determine when more frequent sampling (if possible) would prove to be most beneficial to the estimation process. This technique can be used to design data sampling schemes in more general contexts.

1. **Introduction.** The use of mathematical models to interpret disease outbreak data has provided much insight into the epidemiology and spread of many pathogens, particularly in the context of emerging infections. The basic reproductive number, R_0, which gives the average number of secondary infections that result from a single infective individual over the course of their infection in an otherwise entirely susceptible population (see, for example, [1] and [20]), is often of prime interest. In many situations, the value of R_0 governs the probability of the occurrence of a major outbreak, the typical size of the resulting outbreak and the stringency of control measures needed to contain the outbreak (see, for example [10, 26, 30]).

While it is often simple to construct an algebraic expression for R_0 in terms of epidemiological parameters, one or more of these values is typically not obtainable by direct methods. Instead, their values are usually estimated indirectly by fitting a mathematical model to incidence or prevalence data (see, for example, [3, 12, 32, 38, 41, 42]), obtaining a set of parameters that provides the best match, in some sense, between model output and data. It is, therefore, crucial that we have a good understanding of the properties of the process used to fit the model and its limitations when employed on a given data set. An appreciation of the uncertainty accompanying the parameter estimates, and indeed whether a given parameter is even individually identifiable based on the available data and model, is necessary for our understanding.

The simultaneous estimation of several parameters raises questions of parameter identifiability (see, for example, [2, 6, 17, 22, 24, 27, 36, 43, 44, 45, 46]), even if the model being fitted is simple. Oftentimes, parameter estimates are highly correlated: the values of two or more parameters cannot be estimated independently. For instance, it may be the case that, in the vicinity of the best fitting parameter set, a number of sets of parameters lead to effectively indistinguishable model fits, with changes in one estimated parameter value being able to be offset by changes in another.

Even if individual parameters cannot be reliably estimated due to identifiability issues, it might still be the case that a compound quantity of interest, such as the basic reproductive number, can be estimated with precision. This would occur, for instance, if the correlation between the estimates of indivdual parameters was such

that the value of R_0 varied little over the sets of parameters that provided equal quality fits.

Statistical theory is often used to guide data collection, with sampling theory providing an idea of the amount of data required in order to obtain parameter estimates whose uncertainty lies within a range deemed to be acceptable. In time-dependent settings, sampling theory can also provide insight into *when* to collect data in order to provide as much information as possible. Such analyses can be extremely helpful in biological settings where data collection is expensive, ensuring that sufficient data is collected for the enterprise to be informative, but in an efficient manner, avoiding excessive data collection or the collection of uninformative data from certain periods of the process.

In this paper we discuss the use of sensitivity analysis [21, 37] and asymptotic statistical theory [18, 39], to quantify the uncertainties associated with parameter estimates obtained by the use of least squares model fitting in an epidemiological context. The theory also quantifies the correlation between estimates of the different parameters, and we discuss the implications of correlations on the estimation of R_0. We investigate how the magnitude of uncertainty varies with both the number of data points collected and their collection times. We suggest an approach that can be used to identify the times at which more intensive sampling would be most informative in terms of reducing the uncertainties associated with parameter estimates.

In order to make our presentation as clear as possible, we throughout employ the simplest model for a single outbreak, the SIR model, and use synthetic data sets generated using the model. This idealized setting should be the easiest one for the estimation methodology to handle, so we imagine that any issues that arise (such as non-identifiability of parameters) would carry over to, and indeed be more delicate for, more realistic settings such as more complex models or real-world data sets. The use of synthetic data allows us to investigate the performance and behavior of the estimation for infections that have a range of transmission potentials, providing a broader view of the estimation process than would be obtained by focusing on a particular individual data set.

The paper is organized as follows: the simple SIR model employed in this study is outlined in Section 2. The statistical theory and sensitivity analysis of the model is presented in Section 3. Section 4 discusses the synthetic data sets that we use to demonstrate the approach. Section 5 presents the results of model fitting, and discusses the estimation of R_0. The impact of sampling frequency and sampling times are examined in Section 6. Section 7 explores parameter identifiability for the SIR model. We conclude with a discussion of the results.

2. **The model.**

Since our aim here is to present an examination of general issues surrounding parameter estimation, we choose to use a simple model containing a small number of parameters. We employ the standard deterministic Susceptible-Infective-Recovered compartmental model (see, for example, [1, 19, 25]) for an infection that leads to permanent immunity and that is spreading in a closed population (*i.e.*, we ignore demographic effects). The population is divided into three classes, susceptible, infectious and recovered, whose numbers are denoted by S, I, and R, respectively. The closed population assumption leads to the total population size, N, being constant and we have $S + I + R = N$.

We assume that transmission is described by the standard incidence term $\beta SI/N$, where β is the transmission parameter, which incorporates the contact rate and the probability that contact (between an infective and susceptible) leads to transmission. Individuals are assumed to recover at a constant rate, γ, which gives the average duration of infection as $1/\gamma$.

Because of the equation $S + I + R = N$, we can determine one of the state variables in terms of the other two, reducing the dimension of the system. Here, we choose to eliminate R, and we so focus our attention on the dynamics of S and I. The model can then be described by the following differential equations

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

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

together with the initial conditions $S(0) = S_0$, $I(0) = I_0$.

The behavior of this model is governed by the basic reproductive number. For this SIR model, $R_0 = \beta/\gamma$. The average number of secondary infections per individual at the beginning of an epidemic is given by the product of the rate at which new infections arise (β) and the average duration of infectiousness ($1/\gamma$). R_0 tells us whether an epidemic will take off ($R_0 > 1$) or not ($R_0 < 1$) in this deterministic framework.

This SIR model is formulated in terms of the number of infectious individuals, $I(t)$, *i.e.*, the prevalence of infection. Disease outbreak data, however, is typically reported in terms of the number of new cases that arise in some time interval, *i.e.*, the disease incidence. The incidence of infection over the time interval (t_{i-1}, t_i) is given by integrating the rate of infection over the time interval: $\int_{t_{i-1}}^{t_i} \beta S(t)I(t)/N\, dt$. Notice that, since the SIR model does not distinguish between infectious and symptomatic individuals—even though this is not the case for many infections—we equate the incidence of new infections and new cases. For the simple SIR model employed here, the incidence can be calculated by the simpler formula $S(t_{i-1}) - S(t_i)$, since the number of new infections is given by the decrease in the number of susceptibles over the interval of interest.

3. **Methodology.** Estimating the parameters of the model given a data set (solving the inverse problem) is here accomplished by using either ordinary least squares (OLS) or a weighted least squares method known as either iteratively reweighted least squares or generalized least squares (GLS) [18]. Uncertainty quantification is then performed using asymptotic statistical theory (see, for example, Seber and Wild [39]) applied to the statistical model that describes the epidemiological data set. Although the application of this theory to epidemiological settings has been developed and explained in a number of previous works (see, for example, [3, 15, 16]), to aid the reader we provide a brief general summary of this theory. In order to facilitate comparison with previous papers cowritten by us, we largely follow the development and notation laid out in [3, 15, 16], albeit with a few notational deviations and changes in emphasis.

The statistical model assumes that the epidemiological system is exactly described by some underlying dynamic model (for us, the deterministic SIR model) together with some set of parameters, known as the true parameters, but that the observed data arises from some corruption of the output of this system by noise (*e.g.*, observational errors). We write the true parameter set as the p-element vector

θ_0, noting that some of these parameters may be initial conditions of the dynamic model if one or more of these are unknown. The n observations of the system, $Y_1, Y_2, \ldots, Y_n$, are made at times $t_1, t_2, \ldots, t_n$. We assume the statistical model can be written as

$$Y_i = M(t_i; \theta_0) + E_i, \tag{3}$$

where $M(t_i; \theta_0)$ is our deterministic model (either for prevalence or incidence, as appropriate) evaluated at the true value of the parameter, θ_0, and the E_i depict the errors. We write $Y = (Y_1, \ldots, Y_n)^T$.

The appropriate estimation procedure depends on the properties of the errors E_i. We assume that the errors have the following form

$$E_i = M(t_i; \theta_0)^{\xi} \epsilon_i, \tag{4}$$

where $\xi \geq 0$. The ϵ_i are assumed to be independent, identically distributed random variables with zero mean and (finite) variance σ_0^2. The random variables Y_i have means given by $E(Y_i) = M(t_i; \theta_0)$ and variances $\mathrm{Var}(Y_i) = M(t_i; \theta_0)^{2\xi} \sigma_0^2$.

If ξ is taken to equal 0 then $E_i = \epsilon_i$, and the error variance is assumed to be independent of the magnitude of the predicted value of the observed quantity. This noise structure is often termed absolute noise in the literature. Positive values of ξ correspond to the assumption that the error variance scales with the predicted value of the quantity being measured. If $\xi = 1$, the standard deviation of the noise is assumed to scale linearly with M: the average magnitude of the noise is a constant fraction of the true value of the quantity being measured. This situation is often referred to as relative noise. If, instead, $\xi = 1/2$, the variance of the error scales linearly with M: we refer to this as Poisson noise.

The least squares estimator $\hat{\theta}_{\mathrm{LS}}$ is a random variable obtained by consideration of the cost functional

$$J(\theta | Y) = \sum_{i=1}^{n} w_i (Y_i - M(t_i; \theta))^2, \tag{5}$$

in which the weights w_i are given by

$$w_i = \frac{1}{M(t_i; \theta)^{2\xi}}. \tag{6}$$

If $\xi = 0$, then $w_i = 1$ for all i, and in this case the estimator is obtained by minimizing $J(\theta|Y)$, that is

$$\hat{\theta}_{\mathrm{LS}} = \arg\min_{\theta} J(\theta | Y). \tag{7}$$

In this case, known as ordinary least squares (OLS), all data points are of equal importance in the fitting process.

When $\xi > 0$, the weights lead to more importance being given to data points that have a lower variability (*i.e.*, those corresponding to smaller values of the model). If the values of the weights were known ahead of time, estimation could proceed by a weighted least squares minimization of the cost functional (5). The weights, however, depend on θ and so an iterative process is instead used, employing estimated weights. An initial ordinary (unweighted) least squares is carried out and the resulting model is used to provide an initial set of weights. Weighted least squares is then carried out using these weights, providing a new model and hence a new

set of weights. The weighted least squares step is repeated with successively updated weights until some termination criterion, such as the convergence of successive estimates to within some specified tolerance, is achieved [18].

The asymptotic statistical theory, as detailed in [18, 39], describes the distribution of the estimator $\hat{\theta}_{\text{LS}} = \hat{\theta}_{\text{LS}}^{(n)}$ as the sample size $n \to \infty$. (In this paragraph we include the superscript n to emphasize sample size dependence.) Provided that a number of regularity and sampling conditions are satisfied (discussed in detail in [39]), this estimator has a p-dimensional multivariate normal distribution with mean θ_0 and variance-covariance matrix Σ_0 given by

$$\Sigma_0 = \lim_{n\to\infty} \Sigma_0^{(n)} = \lim_{n\to\infty} \sigma_0^2 \left(n\Omega_0^{(n)}\right)^{-1}, \tag{8}$$

where

$$\Omega_0^{(n)} = \frac{1}{n}\chi^{(n)}(\theta_0)^T W^{(n)}(\theta_0)\chi^{(n)}(\theta_0). \tag{9}$$

So, $\hat{\theta}_{\text{LS}} \sim N(\theta_0, \Sigma_0)$.

We note that existence and invertibility of the limiting matrix $\Omega_0 = \lim_{n\to\infty} \Omega_0^{(n)}$ is required for the theory to hold. In Equation (9), $W^{(n)}(\theta)$ is the diagonal weight matrix, with entries w_i, and $\chi^{(n)}(\theta)$ is the $n \times p$ sensitivity matrix, whose entries are given by

$$\chi^{(n)}(\theta)_{ij} = \frac{\partial M(t_i;\theta)}{\partial \theta_j}. \tag{10}$$

Because we do not have an explicit formula for $M(t_i;\theta)$, the sensitivities must be calculated using the so-called sensitivity equations. As outlined in [21, 37], for the general m-dimensional system

$$\dot{x} = F(x, t; \theta), \tag{11}$$

with state variable $x \in \mathbb{R}^m$ and parameter $\theta \in \mathbb{R}^p$, the matrix of sensitivities, $\partial x/\partial\theta$, satisfies

$$\frac{d}{dt}\frac{\partial x}{\partial \theta} = \frac{\partial F}{\partial x}\frac{\partial x}{\partial \theta} + \frac{\partial F}{\partial \theta}, \tag{12}$$

with initial conditions

$$\frac{\partial x(0)}{\partial \theta} = 0_{m\times p}. \tag{13}$$

Here, $\partial F/\partial x$ is the Jacobian matrix of the system. This initial value problem must be solved simultaneously with the original system (11).

Sensitivity equations for the state variables with respect to initial conditions can be derived in a similar way, except that the second term on the right side of Equation (12) is absent and the appropriate matrix of initial conditions is $I_{m\times m}$. The sensitivity equations for the specific case of the SIR model of interest here are presented in the appendix.

Because the true parameter θ_0 is usually not known, we use the estimate of θ in its place in the estimation formulae. The value of σ_0^2 is approximated by

$$\sigma^2 = \frac{1}{n-p}\sum_{i=1}^{n} w_i(M(t_i;\theta) - y_i)^2, \tag{14}$$

where the factor $1/(n-p)$ ensures that the estimate is unbiased. The matrix

$$\Sigma = \sigma^2[\chi^T(\theta)W(\theta)\chi(\theta)]^{-1} \tag{15}$$

provides an approximation to the covariance matrix Σ_0.

Standard errors for the components of the estimator $\hat{\theta}_{\mathrm{LS}}$ are approximated by taking square roots of the diagonal entries of Σ, while the off-diagonal entries provide approximations for the covariances between pairs of these components. The uncertainty of an estimate of an individual parameter is conveniently discussed in terms of the coefficient of variation (CV), that is the standard error of an estimate divided by the estimate itself. The dimensionless property of the CV allows for easier comparison between uncertainties of different parameters. In a related fashion, the covariances can be conveniently normalized to give correlation coefficients, defined by

$$\rho_{\hat{\theta}_i,\hat{\theta}_j} = \frac{\mathrm{cov}(\hat{\theta}_i,\hat{\theta}_j)}{\sqrt{\mathrm{Var}(\hat{\theta}_i)\mathrm{Var}(\hat{\theta}_j)}}. \tag{16}$$

The asymptotic statistical theory provides uncertainties for individual parameters, but not for compound quantities—such as the basic reproductive number—that are often of interest. For instance, if we had the estimator $\hat{\theta}_{\mathrm{LS}} = (\hat{\beta},\hat{\gamma})^T$, a simple point estimate for R_0 would be β/γ, where β and γ are the realized values of $\hat{\beta}$ and $\hat{\gamma}$. To understand the properties of the corresponding estimator we examine the expected value and variance of the estimator $\hat{\beta}/\hat{\gamma}$. Because this quantity is the ratio of two random variables, there is no simple exact form for its expected value or variance in terms of the expected values and variances of the estimators $\hat{\beta}$ and $\hat{\gamma}$. Instead, we have to use approximation formulas derived using the method of statistical differentials (effectively a second order Taylor series expansion, see [29]), and obtain

$$\mathrm{E}\left(\frac{\hat{\beta}}{\hat{\gamma}}\right) \approx \frac{\beta_0}{\gamma_0}\left(1 - \frac{\mathrm{cov}(\hat{\beta},\hat{\gamma})}{\beta_0\gamma_0} + \frac{\mathrm{Var}(\hat{\gamma})}{\gamma_0^2}\right), \tag{17}$$

and

$$\mathrm{Var}\left(\frac{\hat{\beta}}{\hat{\gamma}}\right) \approx \left(\frac{\beta_0}{\gamma_0}\right)^2\left(\frac{\mathrm{Var}(\hat{\beta})}{{\beta_0}^2} + \frac{\mathrm{Var}(\hat{\gamma})}{{\gamma_0}^2} - \frac{2\mathrm{cov}(\hat{\beta},\hat{\gamma})}{\beta_0\gamma_0}\right). \tag{18}$$

Here we have made use of the fact that $\mathrm{E}(\hat{\beta}) = \beta_0$, the true value of the parameter, and $\mathrm{E}(\hat{\gamma}) = \gamma_0$.

The variance equation has previously been used in an epidemiological setting by Chowell et al [13]. Equation (17), however, shows us that estimation of R_0 by dividing point estimates of β and γ provides a biased estimate of R_0. The bias factor can be written in terms of the correlation coefficient and coefficients of variation giving

$$\left(1 - \frac{\mathrm{cov}(\hat{\beta},\hat{\gamma})}{\beta_0\gamma_0} + \frac{\mathrm{Var}(\hat{\gamma})}{\gamma_0^2}\right) = \left(1 - \rho_{\hat{\beta},\hat{\gamma}}CV_{\hat{\beta}}CV_{\hat{\gamma}} + CV_{\hat{\gamma}}^2\right). \tag{19}$$

This factor only becomes important when the CVs are on the order of one. In such a case, however, the estimability of the parameters is already in question. Thus, under most useful circumstances, estimating R_0 by the ratio of point estimates of β and γ suffices.

4. **Generation of synthetic data, model fitting and estimation.** In order to facilitate our exploration of the parameter estimation problem, we choose to use simulated data. This 'data' is generated using a known model, a known parameter set and a known noise structure, putting us in an idealized situation in which we know that we are fitting the correct epidemiological model to the data, that the correct statistical model is being employed and where we can compare the estimated

parameters with their true values. Furthermore, since we know the noise process, we can generate multiple realizations of the data set and hence directly assess the uncertainty in parameter estimates by fitting the model to each of the replicate data sets. As a consequence, we can more completely evaluate the performance of the estimation process than would be possible using a single real-world data set.

The use of synthetic data also allows us to investigate parameter estimation for diseases that have differing levels of transmissibility. We considered three hypothetical infections, with low, medium and high transmissibility, using R_0 values of 1.2, 3 and 10, respectively. In each case we took the recovery rate γ to equal 1, which corresponds to measuring time in units of the average infectious period. The value of β was then chosen to provide the desired value of R_0. (In terms of the "true values" of our statistical model, we have $\gamma_0 = 1$ and $\beta_0 = R_0$). We took a population size of $10,000$, of which 100 people were initially infectious, with the remainder being susceptible. (Altering the initial number of infectives makes no qualitative difference to the results that follow.)

The model was solved for S and I using the `MATLAB ode45` routine, starting from $t = 0$, giving output at $n + 1$ evenly spaced time points $(0, t_1, ..., t_n)$. The duration of the outbreak depends on R_0 and so, in order to properly capture the time scale of the epidemic, we choose t_n to be the time at which $I(t)$ falls back to its initial value. A data set for prevalence was then obtained by adding noise generated by multiplying independent draws, e_i, from a normal distribution with mean zero and variance σ_0^2 by $I(t_i, \theta_0)^\xi$. Thus, our data,

$$y(t_i, \theta_0) \equiv I(t_i, \theta_0) + I(t_i, \theta_0)^\xi e_i, \qquad i = 1, 2, \ldots, n, \tag{20}$$

satisfies the assumptions made in Section 3 and allows us to apply the asymptotic statistical theory. Notice that, for convenience, we have chosen normally distributed e_i, but we re-emphasize that the asymptotic statistical theory does not require this. Data sets depicting incidence of infection can be created in a similar way, replacing $I(t_i)$ by $S(t_i) - S(t_{i-1})$, as discussed above, for $i = 1, \ldots, n$.

Three different values of ξ, namely $\xi = 0$ (absolute noise), $\xi = 1/2$ (Poisson noise) and $\xi = 1$ (relative noise), were used to generate synthetic data sets. Given that prevalence (or incidence) increases with R_0, the use of absolute noise, with the same value of σ_0^2 across the three transmissibility scenarios, leads to noise being much more noticeable for the low transmissibility situation. This complicates comparisons of the success of the estimation process between differing R_0 values. Visual inspection of real-world data sets, however, indicates that variability increases with either prevalence or incidence [23]. If this variability reflected reporting errors, with individual cases being reported independently with some fixed probability, the variance of the resulting binomial random variable would be proportional to its mean value. As a result, we direct most of our attention to data generated using $\xi = 1/2$.

Because we know the true values of the parameters and the variance of the noise, we can calculate the variance-covariance matrix Σ_0 (Equation 8) exactly, without having to use estimated parameter values or error variance. This provides a more reliable value than that obtained using the estimate Σ, allowing us to more easily detect small changes in standard errors, such as those that occur when a single data point is removed from or added to a data set as we do in Section 6. This approach was employed to obtain many of the results that follow (in each instance, it will be stated whether Σ_0 or Σ was used to provide uncertainty estimates).

5. **Results: Parameter estimation.** We could attempt to fit any combination of the parameters and initial conditions of the SIR model, *i.e.*, β, γ, N, S_0 and I_0. We shall concentrate, however, on the simpler situation in which we just fit β and γ, imagining that the other values are known. This might be the case if a new pathogen were introduced into a population at a known time, so that the population was known to be entirely susceptible apart from the initial infective. Importantly, the estimation of β and γ allows us to estimate the value of R_0. We shall return to consider estimation of three or more parameters in a later section.

The least squares estimation procedure works well for synthetic data sets generated using the three different values of R_0 (results not shown). Diagnostic plots of the residuals were used to examine potential departures from the assumptions of the statistical model: unsurprisingly, none were seen when the value of ξ used in the fitting process matched that used to generate the data, and clear deviations were seen when the incorrect value of ξ was used in the fitting process (results not shown).

TABLE 1. Coefficients of variation (CV) for parameter estimates of β, γ, R_0, and the correlation coefficient between β and γ, $\rho_{\beta,\gamma}$. The coefficients of variation and correlation coefficient were obtained from the asymptotic stastical theory where the variance-covariance matrix Σ_0 was calculated exactly (*i.e.*, no curve-fitting was carried out). Calculations were done under a Poisson noise structure, $\xi = 1/2$, with $\sigma_0^2 = 1$, and $n = 50$ data points. Parameter values and initial conditions used were $\beta = R_0$, $\gamma = 1$, $N = 10,000$, $S_0 = 9900$, and $I_0 = 100$.

Parameter	**Value**	**CV**
β	1.2	0.0121
γ	1	0.0110
R_0	1.2	0.0023
$\rho_{\hat{\beta},\hat{\gamma}}$	0.9837	-

Parameter	**Value**	**CV**
β	3	0.0019
γ	1	0.0034
R_0	3	0.0037
$\rho_{\hat{\beta},\hat{\gamma}}$	0.1132	-

Parameter	**Value**	**CV**
β	10	0.0035
γ	1	0.0027
R_0	10	0.0050
$\rho_{\hat{\beta},\hat{\gamma}}$	-0.3122	-

A Monte Carlo approach can be used to verify the distributional results of the asymptotic statistical theory. A set of point estimates of the parameter (β, γ) was generated by applying the estimation process to a large number of replicate data sets generated using different realizations of the noise process, allowing estimates of variances and covariances of parameter estimates to be directly obtained. Unsurprisingly, good agreement was seen when the correct value of ξ was employed in the estimation process and the distribution of (β, γ) estimates appears to be consistent with the appropriate bivariate normal distribution predicted by the theory.

Table 1 and Figure 1a demonstrate that estimates of β and γ are correlated, with the sign and magnitude of the correlation coefficient depending strongly on the value of R_0. Standard errors for the estimates also depend strongly on the value of R_0 (Figure 1b).

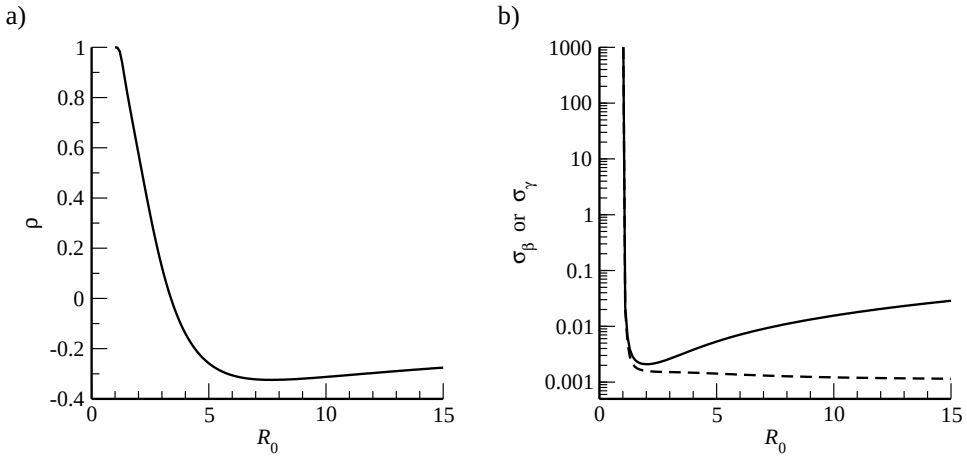

FIGURE 1. Dependence of the correlation coefficient and standard errors for estimates of β and γ on the value of R_0. Panel (a) displays the correlation coefficient, ρ, between estimates of β and γ for a range of R_0 values. Panel (b) shows, on a log scale, standard errors for estimates of β (solid curve) and γ (dashed curve). The variance-covariance matrix Σ_0 was calculated exactly (*i.e.*, no curve-fitting was carried out) under a Poisson noise structure, $\xi = 1/2$, with $\sigma_0^2 = 1$, and $n = 250$ data points. Parameter values and initial conditions used were $\beta = R_0$, $\gamma = 1$, $N = 10,000$, $S_0 = 9900$, and $I_0 = 100$.

As R_0 approaches 1, the correlation coefficient approaches 1 and the standard errors become extremely large. It is, therefore, difficult to obtain good estimates of the individual parameters in this case. Examination of the cost functional J in the (γ, β) plane reveals the origin of the strong correlation and large standard errors (Figure 2a). Near its minimum value, the contours of J are well approximated by long thin ellipses whose major axes are oriented along the line $\beta = R_0\gamma$. Thus there is a considerable range of β and γ values that give almost identical model fits, but for which the ratio β/γ varies relatively little. In a later section we shall see that these long thin elliptical contours arise as a consequence of sensitivities of the model to changes in β and γ being almost equal in magnitude but of opposite signs. (The derivation of these contour curves can be found in [9].)

For values of R_0 that lead to lower correlation between estimates of β and γ, the contours of J near its minimum point are closer to being circular and are less tilted (Figure 2b), allowing for easier identification of the two individual parameters. The standard error for the estimate of γ is seen to decrease with R_0, while that of β exhibits non-monotonic behavior. For a fixed value of γ, increasing R_0 leads to more rapid spread of the infection and hence an earlier and higher peak in prevalence (Figure 3). For large values of R_0, the majority of the transmission events occur over the timespan of the first few data points, meaning that fewer points within the data set are informative regarding the spread of the infection. Consequently, it becomes increasingly difficult to estimate β as R_0 is increased beyond some critical value.

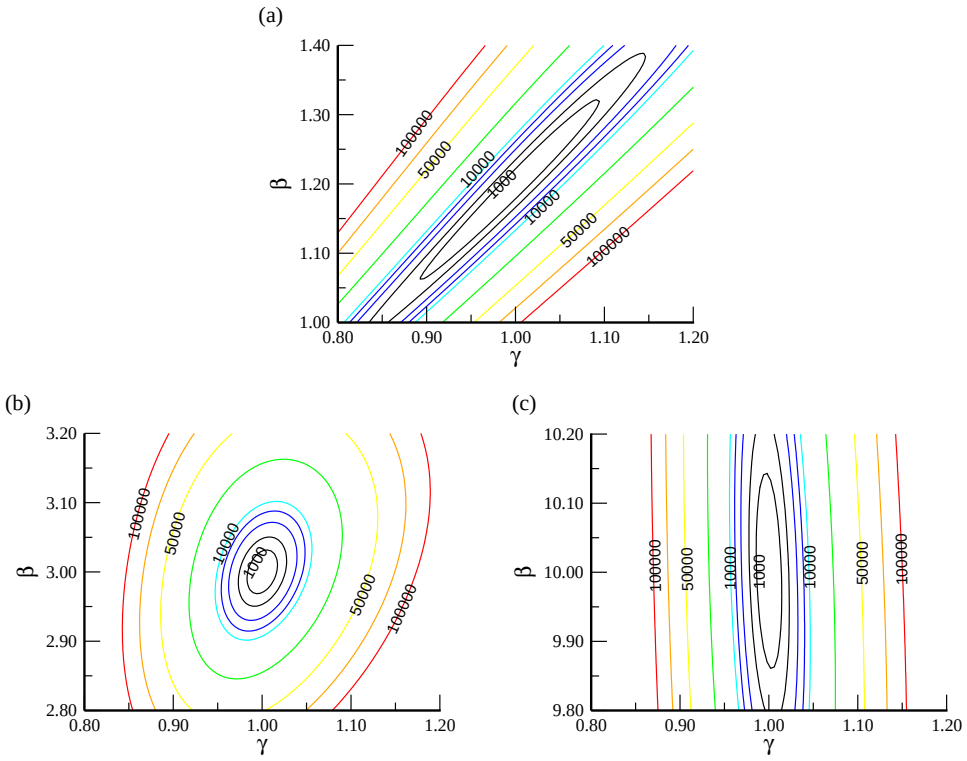

FIGURE 2. Contours of the cost functional J in the (γ, β)-plane (solid curves) for R_0 equal to (a) 1.2, (b) 3, and (c) 10. A Poisson noise structure was assumed ($\xi = 1/2$), with $\sigma_0^2 = 1$ and $n = 50$ data points. Parameter values and initial conditions used were $\beta = R_0$, $\gamma = 1$, $N = 10,000$, $S_0 = 9900$, and $I_0 = 100$. Contours are at heights 1000, 2500, 5000, 7500, 10000, 25000, 50000, 75000 and 100000. For clarity, not all contours are labeled with their height.

As seen in Table 1, estimates of β and γ have relatively large uncertainties when R_0 is small. It would, for instance, be difficult to accurately estimate the average duration of infection, $1/\gamma$, for an infection such as seasonal influenza—which is typically found to have R_0 about 1.3 (ranging from 0.9 to 2.1) [14]—using the least squares approach. Importantly, however, the estimate of R_0 has a much lower variation (as measured by the CV) than the estimates of β and γ. The strong positive correlation between the estimates of β and γ reduces the variance of the R_0 estimate, as can be seen in Equation (18), and reflecting the earlier observation concerning the orientation of the contours of the cost functional along lines of the form $\beta = R_0\gamma$.

6. **Results: Sampling schemes and uncertainty of estimates.** Biological data is often difficult or costly to collect, so it is desirable to collect data in such a way to maximize its informativeness. Consequently it is important to understand how parameter estimation depends on the number of sampled data points and the times at which the data are collected. This information can then be used to guide

future data collection. In this section we examine two approaches to address this question: sensitivity analysis and data sampling.

6.1. **Sensitivity.** The sensitivities of a system provide temporal information on how states of the system respond to changes in the parameters [21, 37]. They can, therefore, be used to identify time intervals where the system is most sensitive to such changes. Noting that the sensitivities are used to calculate the standard errors in estimates of parameters, direct observation of the sensitivity function provides an indication of time intervals in which data points carry more or less information for the estimation process [4, 5]. For instance, if the sensitivity to some parameter is close to zero in some time interval, changes in the value of the parameter would have little impact on the state variable. Conversely, more accurate knowledge of the state variable at that time could not cause the estimated parameter value to change by much.

For low values of R_0, for example $R_0 = 1.2$, we see that the sensitivity functions of $I(t)$ with respect to β and γ are near mirror images of each other (Figure 3a). This mirror image phenomenon allows a change in one parameter to be easily compensated by a corresponding change in the other parameter, giving rise to the strong correlation between the estimates of the two parameters. Early in the epidemic, we see a similar phenomenon for all values of R_0. We comment further on this observation in the next section.

As R_0 increases, the two sensitivity functions take on quite different shapes. Prevalence is much less sensitive to changes in β than to changes in γ. The sensitivity of prevalence to β is greatest right before the epidemic peak, before becoming negative, but small, during the late stages of the outbreak. The sensitivity becomes negative because an increase in β would cause the peak of the outbreak to occur earlier, reducing the prevalence at a fixed, later time. I remains sensitive with respect to γ throughout much of the epidemic, reaching its largest absolute value slightly later than the time at which the outbreak peaks.

While the sensitivity functions provide an indication of when additional, or more accurate data, is likely to be informative, they have clear limitations, not least because they do not provide a quantitative measure of how uncertainty estimates, such as standard errors, are impacted. Being a univariate approach they cannot account for any impact of correlation between parameter estimates, as we shall see below, although they can indicate instances in which parameter estimates are likely to be correlated. Furthermore, they do not account for the different weighting accorded to different data points on account of the error structure of the model, such as the relationship between error variance and the magnitude of the observation being made. Another type of sensitivity function, the generalized sensitivity function (GSF) introduced by Thomaseth and Cobelli [40], which is based on the Fisher information matrix, does account for these two factors. While the GSF does provide qualitative information that can guide data collection, its interpretation is not without its own complications [4] and, given that we found that it provided little additional insight in the current setting, we shall not discuss it further here.

6.2. **Data sampling.** In order to gain quantitative information about sampling schemes on parameter estimation, as opposed to the qualitative information provided by inspection of the sensitivity functions, we carried out three numerical experiments in which different sampling schemes were implemented. The first approach involves altering the frequency at which data are sampled within a fixed

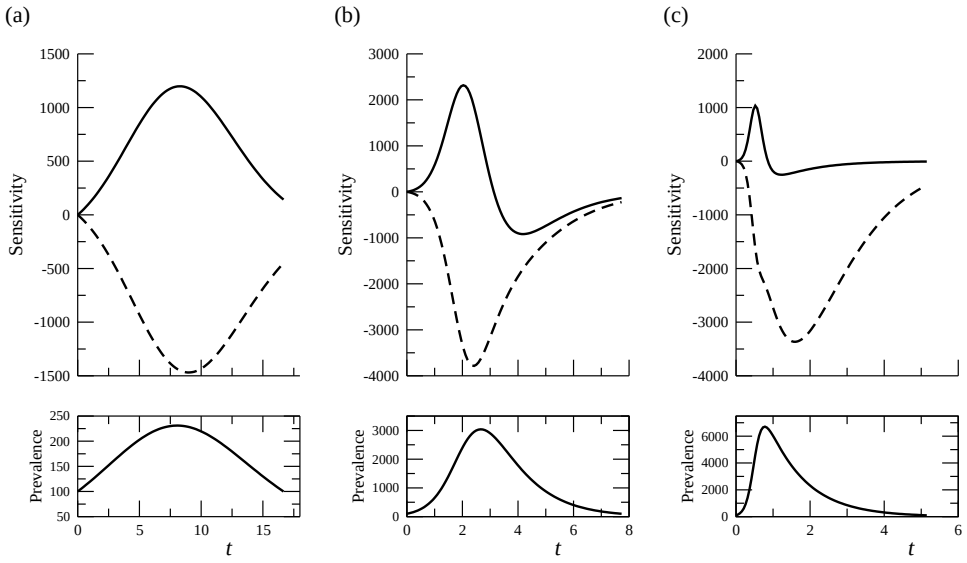

FIGURE 3. Sensitivities of $I(t)$ (*i.e.*, prevalence) with respect to the model parameters β (solid curves) and γ (dashed curves) are shown on the upper panels of the graphs for a) $R_0 = 1.2$, b) $R_0 = 3$ and c) $R_0 = 10$. The lower panel of each graph displays the corresponding prevalence-time curve. The initial conditions of the SIR model were $S_0 = 9900$, $I_0 = 100$, with $N = 10,000$ and γ was taken equal to one, so $\beta = R_0$.

observation window that covers the duration of the outbreak. The second approach considers sampling at a fixed frequency but over observation windows of differing durations. The third approach examines increasing the sampling frequency within specified sub-intervals of a fixed observation window.

In the first sampling method we alter the frequency at which observations are taken while keeping the observation window fixed. In other words, we increase n while fixing $t_0 = 0$ and $t_n = t_{\text{end}}$. For incidence data, increasing the observation frequency—*i.e.*, reducing the period over which each observation is made—has the important effect of reducing the values of the observed data and the corresponding model values. Under relative observational error ($\xi = 1$) there is a corresponding change in the error variance, keeping a constant signal to noise ratio. If $\xi < 1$, increasing n decreases the signal to noise ratio of the data.

Adding additional data points in this way increases the accuracy of parameter estimates, with standard errors eventually decreasing as $n^{-1/2}$ (Figure 4, in which prevalence data is used), in accordance with the asymptotic theory [39]. This is still the case for incidence data even when $\xi < 1$ where the signal to noise ratio decreases in n. We point out that changing the sampling frequency will typically not be an option in epidemiological settings because data will be collected at some fixed frequency, such as once each day or week, although, conceivably, a weekly sampling frequency could be replaced by daily sampling.

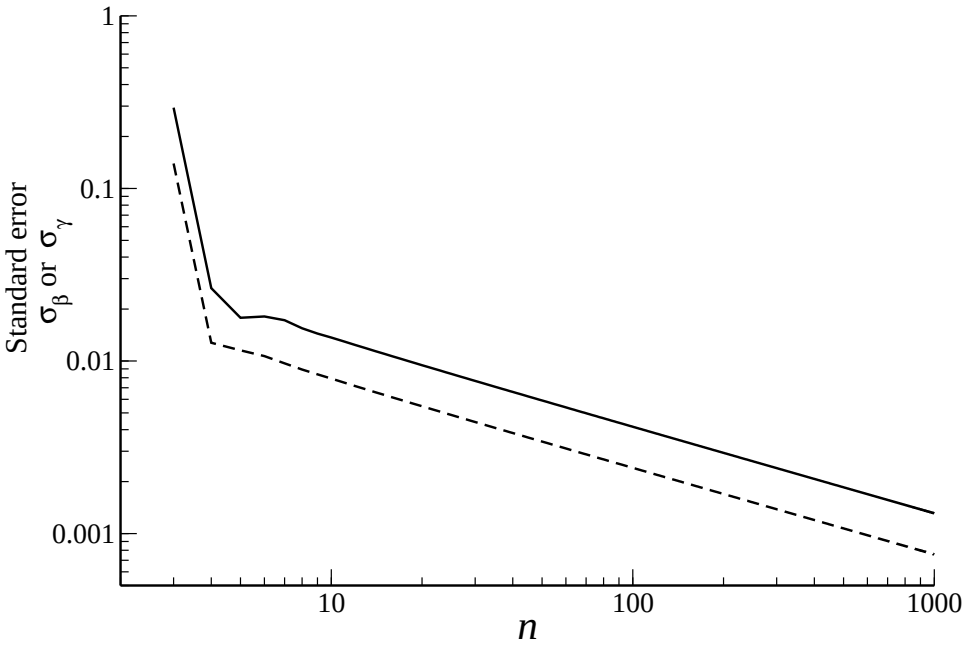

FIGURE 4. Standard errors of β (solid curve) and γ (dashed curve) as the number of observations, n, changes while maintaining a constant window of observation (fixed t_{end}). Apart from the smallest few values of n, the points fall on a line of slope $-\frac{1}{2}$ on this log-log plot. Standard errors are calculated using Equation (8), using the true values of the parameters. The variance-covariance matrix Σ_0 was calculated exactly (*i.e.*, no curve-fitting was carried out) with the disease prevalence under a Poisson noise structure, $\xi = 1/2$, with $\sigma_0^2 = 1$. Parameter values and initial conditions used were $\beta = 3$, $\gamma = 1$, $N = 10,000$, $S_0 = 9900$, and $I_0 = 100$.

For real-time outbreak analysis, the amount of available data will increase over time as the epidemic unfolds. Consequently, it is of practical importance to understand how much data—and hence observation time—is required to obtain reliable estimates and the extent to which estimates will improve with additional data points. Using Equation (8) and the known values of the parameters, we calculated standard errors for parameter estimates based on the first n_{used} data points, where $p + 1 \leq n_{\text{used}} \leq n$. As seen in Figures 5a and 5b, when only one parameter is fitted, the standard error decreases rapidly at first, but its decrease slows significantly just before the peak of the epidemic. Once this point in time has been reached, subsequent data points provide considerably less additional information than did earlier data points. In this setting, the most important time interval extends from the initial infection to just before the peak of the outbreak. However, when both β and γ are fitted, the interval of steep descent extends slightly beyond the peak of the epidemic, as seen in Figure 6a. This indicates that it would be useful to collect data over a longer interval in this case. Notice the log scale on the vertical axis for each of the aforementioned plots. These figures suggest that the amount of

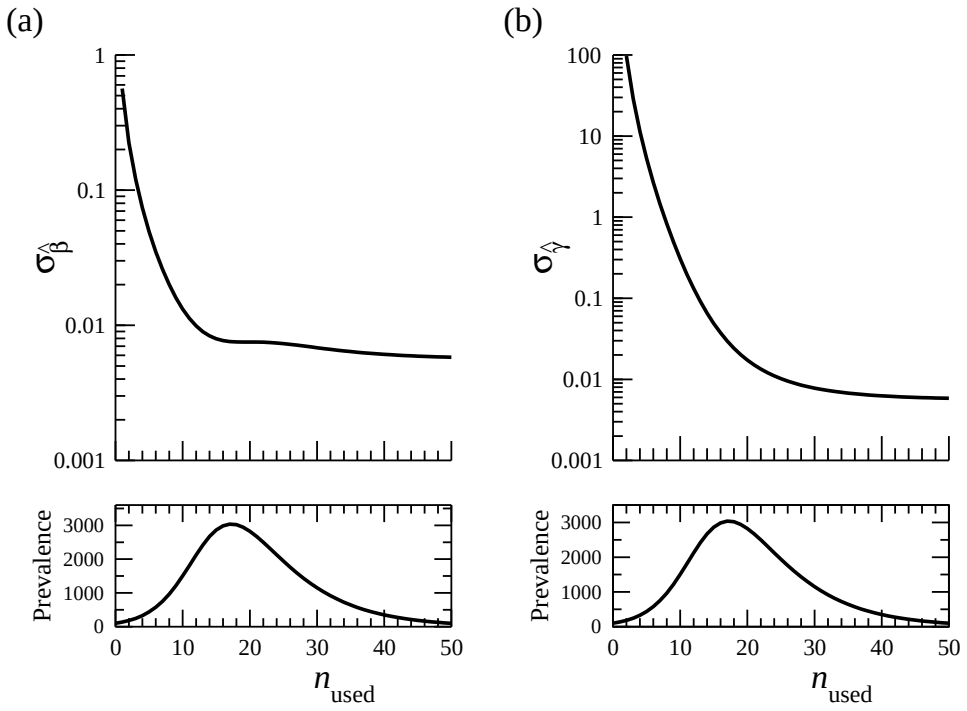

FIGURE 5. Impact of increasing the length of the observation window on standard errors of estimates of (a) β and (b) γ when each is estimated separately from prevalence data. The observation window is $[0, t_{n_{\text{used}}}]$, *i.e.*, estimation was carried out using n_{used} data points. Because data points are equally spaced, the horizontal axis depicts both the number of data points used and time since the start of the outbreak. For reference, the prevalence curve, $I(t)$, is shown in the lower panel of each graph. Standard errors are plotted on a logarithmic scale. The exact formula for Σ_0 was used, with $\sigma_0^2 = 1$, $S_0 = 9900$, $I_0 = 100$, $N = 10,000$, $\beta = 3$ and $\gamma = 1$. The Poisson noise structure, $\xi = 1/2$, was employed.

information contained in the earliest portion of an outbreak is orders of magnitude higher than that contained in later portions.

Figure 6b shows the correlation coefficient between estimates of β and γ as the epidemic progresses. It can be seen that estimates of β and γ are highly correlated until the first inflection point of the epidemic curve, causing the significantly higher standard errors as seen in Figure 6a. This behavior is not unexpected due to the two sensitivity curves for prevalence being near mirror images early in the outbreak, during the exponential growth phase.

Our final sampling method investigated the impact of removing a single data point as a means of identifying the data points which provide the most information for the estimation of the parameters. A baseline data set consisting of fifty evenly-spaced points taken over the course of the outbreak was generated using absolute noise ($\xi = 0$). Fifty reduced data sets were created by removing, in turn, a single data point from the baseline data set. Standard errors were then computed for

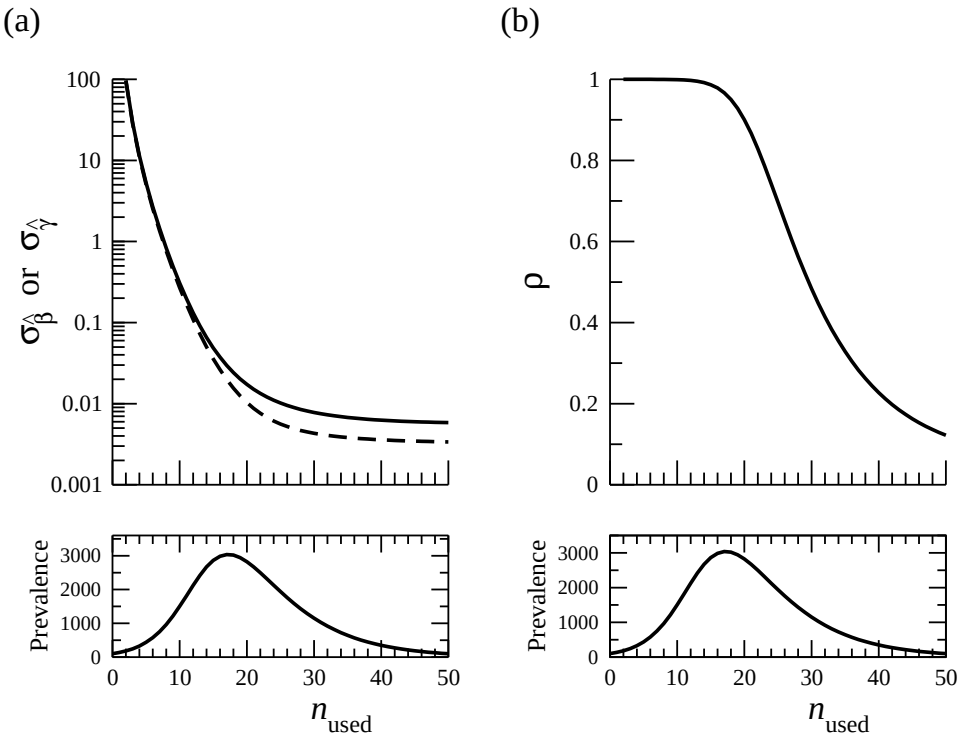

FIGURE 6. Illustrated in graph (a) is the impact of increasing the length of the observation window on standard errors of estimates of β (solid curve) and γ (dashed curve) when both are estimated simultaneously. Graph (b) displays the effect on the correlation coefficient between estimates of β and γ. The observation window consists of n_{used} data points in the time interval $[t_0, t_{n_{\text{used}}}]$. For reference, the prevalence curve, $I(t)$, is shown on the lower panels. All parameter values and other details are as in the previous figure.

the reduced data sets using the true covariance matrix Σ_0 (Equation (8)). (For this experiment, use of the true covariance matrix allowed us to accurately observe the small effects on standard errors that resulted from the removal of single data points. Errors introduced by solving the inverse problem would have obscured the patterns we observed.) The largest standard error values in this group of data sets correspond to the most informative data points since the removal of such points leads to the largest increase in uncertainty of the estimate.

As Figure 7 shows, when β is the only parameter fitted and ordinary least squares estimation is used, the local maxima of the standard error curve occur at the same times as the local extrema of the sensitivity curve, and the local minima occur when the sensitivity is close to zero. In this case, the sensitivity function correctly identifies subintervals in which data are most or least informative about β.

The picture is not quite as straightforward when β and γ are estimated simultaneously using ordinary least squares. Figure 8 shows that the local maxima of the standard error curves no longer line up directly with the local extrema of the sensitivity curves (this effect is more easily seen in Figure 8b). This is likely due

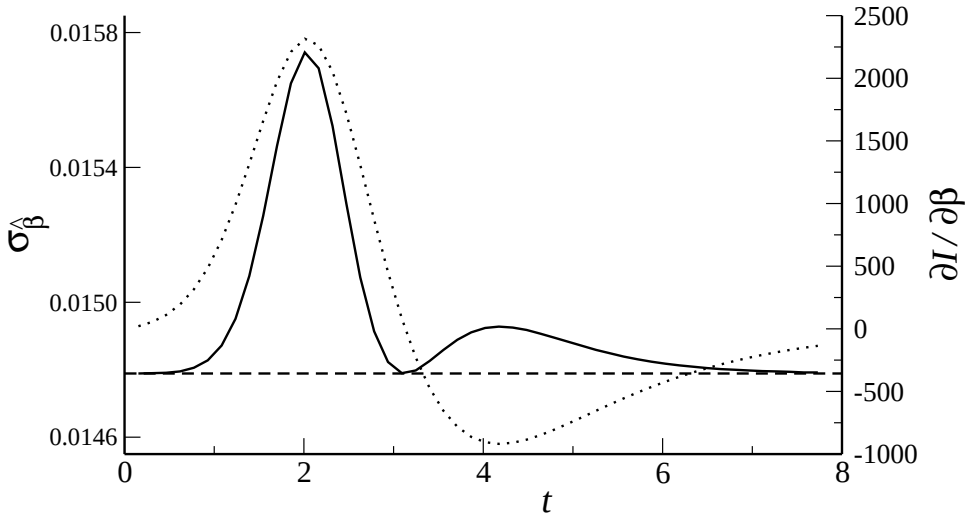

FIGURE 7. Standard errors for the estimation of β from prevalence data using the single point removal method as discussed in the text (solid curve) with the baseline standard error (without removing any points) also plotted (horizontal dashed line). Standard errors were calculated using Equation (8) and each is plotted at the time t_i corresponding to the removed data point. For comparison, the sensitivity of $I(t)$ with respect to β is also shown (dotted curve). Synthetic data was generated using the parameter values $\sigma_0^2 = 10^4$, $S_0 = 9900$, $I_0 = 100$, $N = 10,000$, $\beta = 3$ and $\gamma = 1$. The additive noise structure, $\xi = 0$ was assumed.

to the correlation between the estimates of β and γ: the off-diagonal terms of $\chi^{\mathrm{T}}(\theta)W(\theta)\chi(\theta)$ involve products of sensitivities with respect to the two different parameters. As a consequence, it is no longer sufficient to examine individual sensitivity curves, but, as we have seen, the selective reductive method described here, based on the asymptotic theory, can identify when additional data should ideally be sampled.

Similarly, having a weight matrix other than the identity (*i.e.*, when GLS, rather than OLS, is to be used) leads to the sensitivity curves misidentifying the subintervals in which data are most or least informative for parameter estimation (results not shown; see [9]). This occurs whether single or multiple parameters are estimated, and happens because the sensitivity curves do not, by themselves, account for the relative importance placed on different data points. Again, the selective reduction method accounts for this effect and correctly identifies time intervals when additional data would be most informative.

7. **Results: Parameter identifiability.** Until now, we have only considered the introduction of an infection into a virgin population, assuming a known initial number of infectives in an otherwise susceptible population. For an endemic infection, such as seasonal flu, only a fraction of the population would be susceptible at the start of an outbreak. In such instances, the general reproductive number, R_t, the average number of secondary infections at any point in time, is a more relevant

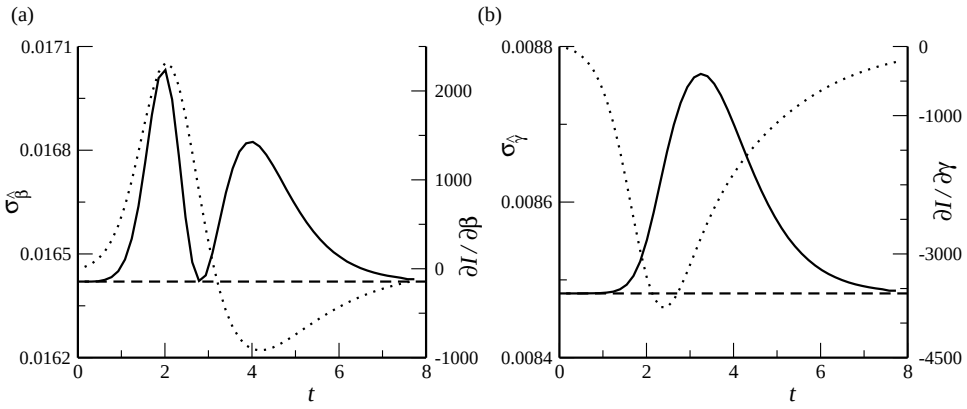

FIGURE 8. Standard errors for the simultaneous estimation of β and γ from prevalence data using the single point removal method as discussed in the text (solid curves). Standard errors were calculated using Equation (8), and each is plotted at the time t_i of the removed data point. Panel (a) shows the standard error for the estimate of β (solid curve), together with the baseline (*i.e.*, without removing any points) standard error (horizontal dashed line) and the sensitivity of $I(t)$ with respect to β (dotted curve). Panel (b) shows the standard error for the estimate of γ (solid curve), together with the baseline standard error (horizontal dashed line) and the sensitivity of $I(t)$ with respect to γ (dotted curve). All parameter values and other details are as in the previous figure.

quantity than R_0. For the SIR model, R_t is given by

$$R_t = R_0 \frac{S(t)}{N}. \tag{21}$$

In the virgin population considered above, we saw that as R_0 approached one there was considerable difficulty in independently estimating a pair of parameters. In the endemic setting, this phenomenon occurs as R_t approaches one, so the parameter identifiability issue can arise even if R_0 is significantly greater than one.

In the endemic setting, we would be unlikely to know the initial numbers of infectives and susceptibles, so we would also need to estimate the values of S_0 and I_0. Given the difficulty in estimating a pair of parameters that has already been illustrated, it seems reasonable to expect that parameter identfiability would become a more delicate issue if larger sets of parameters were estimated. In this section we shall explore the identifiability of parameters when combinations of β, γ, S_0 and I_0 are estimated. This method is generally referred to as subset selection and has been explored by in the context of identifiability by a number of authors (for example, [7, 8, 15, 28]).

It has been shown by Evans *et al.* in [22] that the SIR model with demography is identifiable for all model parameters and initial conditions. They use a strict definition of non-identifiability, where in such a model, a change in one parameter can be compensated by changes in other parameters. However, the authors also concede that while the model may be identifiable, that property alone does not give insight into the ease of estimation of certain subsets of parameters. For example,

by their definition, two parameters whose estimates have a correlation coefficient of 0.99 would be identifiable, yet they may not be easily estimated. In this section, we use quantitative methods to assess ease of parameter identifiability in the context of subset selection.

It was stated above that the asymptotic statistical theory requires the limiting matrix Ω_0 to be invertible. With a finite-sized sample, we instead require this of $\Omega_0^{(n)}$. Non-identifiability leads to these matrices being singular, or close to singular [8], and so one method for determining whether model parameters are identifiable involves calculating the condition number of $\Omega_0^{(n)}$, or, equivalently the condition number of the matrix $\Sigma^{(n)}$ [15]. The condition number, $\kappa(X)$, of a nonsingular matrix X is defined to be the product of the norm of X and the norm of X^{-1}. If we take the norm to be the usual induced matrix 2-norm, we have that the condition number of X is the ratio of the largest singular value (from a singular value decomposition) of X to the smallest singular value of X [34].

Initially, we investigate the case where only β and γ are fitted. In this situation, we are able to find an expression for $\kappa(\Sigma)$

$$\kappa(\Sigma) = \frac{\sigma_{\hat{\beta}}^2 + \sigma_{\hat{\gamma}}^2 + \sqrt{\sigma_{\hat{\beta}}^4 + \sigma_{\hat{\gamma}}^4 - 2\sigma_{\hat{\beta}}^2\sigma_{\hat{\gamma}}^2 + 4\rho_{\hat{\beta},\hat{\gamma}}^2\sigma_{\hat{\beta}}^2\sigma_{\hat{\gamma}}^2}}{\sigma_{\hat{\beta}}^2 + \sigma_{\hat{\gamma}}^2 - \sqrt{\sigma_{\hat{\beta}}^4 + \sigma_{\hat{\gamma}}^4 - 2\sigma_{\hat{\beta}}^2\sigma_{\hat{\gamma}}^2 + 4\rho_{\hat{\beta},\hat{\gamma}}^2\sigma_{\hat{\beta}}^2\sigma_{\hat{\gamma}}^2}}. \tag{22}$$

If the standard errors were fixed, Equation 22 shows that as the correlation between estimates of β and γ approaches one, the condition number goes to infinity. However, in reality standard errors do depend on the values of β and γ; Figure 9 provides a more complete picture of how the condition number changes over a range of R_0 values. As the figure shows, it is more difficult to rely on estimates of β and γ when R_0 approaches one, corroborating what we have previously seen for the correlation coefficient (see Figure 1a).

Numerical experiments indicate that when more parameters are fitted to the data, identifiability becomes a more serious issue. In such a case, while we can no longer give a simple expression for $\kappa(\Sigma_0)$ since it is a function of the parameters, the initial conditions and even the data, it provides insight into parameter identifiability. We examine $\kappa(\Sigma_0)$ across different subsets of fitted parameters as seen in Table 2. As we increase the number of parameters fitted, the condition number can increase by multiple orders of magnitude. This is evident whenever we fit both β and S_0. Notice that for the larger κ values, the magnitude of ρ is very near to one, indicating strong correlation. Thus, we can surmise that as we increase the number of fitted parameters, our ability to identify individual parameters decreases, especially if the parameters added to θ have correlated estimates.

In this example, if we assume the initial conditions are known, our ability to estimate β and γ is good. Yet, once we have to estimate one or both initial coniditions, our ability to estimate either β or γ worsens considerably. Given that in most situations initial conditions are not known exactly, parameter identifiability has the potential to be of widespread concern.

8. **Discussion.**

Parameter values estimated from real-world data will always be accompanied by some uncertainty. Estimates of this uncertainty allow us to judge how reliable the parameter estimates are and how much faith should be put in any predictions made on their basis. As such, uncertainty estimates should always

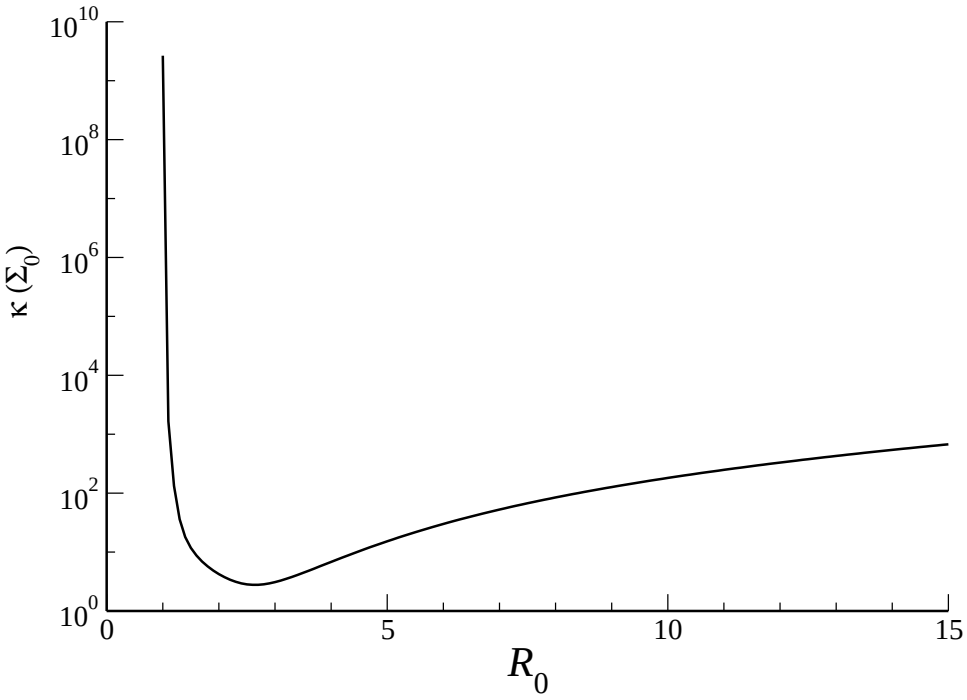

FIGURE 9. Dependence of the condition number of the 2×2 variance-covariance matrix (fitting β and γ) on the value of R_0. The condition number is displayed on a log scale. The variance-covariance matrix Σ_0 was calculated exactly (*i.e.*, no curve-fitting was carried out) under a Poisson noise structure, $\xi = 1/2$, with $\sigma_0^2 = 1$, and $n = 250$ data points. Parameter values and initial conditions used were $\beta = R_0$, $\gamma = 1$, $N = 10,000$, $S_0 = 9900$, and $I_0 = 100$.

TABLE 2. Standard errors of β and γ, the correlation coefficient between estimates of β and γ and the condition number of the $\chi^T W \chi$ matrix when $R_0 = 3$ when fitting different sets of parameters. $\xi = 1/2$, $N = 10,000$, $S_0 = 9900$, $I_0 = 100$, $\beta = R_0$, $\gamma = 1$ and $\sigma_0^2 = 10^4$.

Parameters Fitted	$\sigma_{\hat{\beta}}$	$\sigma_{\hat{\gamma}}$	ρ	κ
β, γ	0.3419	0.1142	-0.2067	5.2211×10^0
β, γ, $S(0)$	17.094	1.9936	-0.9984	5.5760×10^9
β, γ, $I(0)$	1.7536	0.1176	0.3534	1.2000×10^6
β, γ, $S(0)$, $I(0)$	44.655	3.3060	-0.9548	1.4383×10^{10}

accompany estimates of parameter values. The asymptotic statistical theory employed here provides a reasonably straightforward way to obtain such information when least-squares fitting is used as the estimation process.

The use of a number of synthetic data sets, generated under a number of different scenarios concerning the transmissibility of infection, has allowed us to get a broader

understanding of the parameter estimation process than would have been possible if we had limited attention to a single data set. As we have demonstrated, the uncertainties that accompany parameter estimation, and even our ability to separately identify parameters—even with this simplest of SIR models—can be extremely varied based on the underlying parameter values and the parameter set being fitted. A primary reason for difficulties in estimation and identifiability stems from correlations between parameter estimates. Even if individual parameter estimates have large uncertainties it can still be possible to estimate epidemiologically important information, *e.g.*, the basic reproductive number R_0, with much less uncertainty.

Increasing the number of observations made at critical times during the epidemic can provide a substantial gain in the precision of the estimation process. While the sensitivity equations of the model provide a general idea of times at which additional data will be most informative, they do not tell the whole story. The asymptotic statistical theory, together with the data point removal technique, can be used to guide data collection. This approach can be employed once a parameter set is known: this might be one based on a preliminary set of estimates, expert opinon, or even a best-guess. Some aspects of our discussion do, however, require more detailed information on the magnitude and nature of the noise in the data.

We have focused on identifiability in the least squares context, but one cannot escape a lack of parameter identifiability simply by using a different method of parameter estimation. Bayesian methods, including Markov Chain Monte Carlo, (see, for example, [11] and [31]), provide an alternative suite of approaches that are commonly used to solve the inverse problem. Yet, since identifiability is primarily a feature of the mathematical model and less dependent on the fitting process, switching estimation techniques often does not remove the problem of parameter identifiability, so it remains an important concern when solving the inverse problem in any respect.

It should be noted that all experiments presented here were conducted with knowledge of the underlying model, that is, the correct model was fit to the data. However, in scenarios with real data this assumption is not valid and results in a further layer of uncertainty. This type of structural uncertainty has received far less attention but in some circumstances it can dwarf uncertainty due to noise in the data. As an example, a number of authors have shown that estimates of the basic reproductive number obtained by fitting models to data on the initial growth of an outbreak can be highly sensitive to model assumptions [32, 33, 35, 42].

We chose to focus our attention on perhaps the simplest possible setting for the estimation process, one for which the SIR model was appropriate. Unfortunately, few real-world disease transmission processes are quite this simple; in most instances, a more complex epidemiological model, accompanied by a larger set of parameters and initial conditions, would be more realistic. It is not hard to imagine that many of the issues discussed here would be much more delicate in such situations: parameter identifiability, in particular, could be a major concern. The approach employed here would reveal whether such problems would accompany estimation using a given model, and indeed can be used to guide the selection of models and/or parameter sets that can be used or estimated reliably. Again, this emphasizes the need for the estimation process to be accompanied by some account of the uncertainties, but not only in terms of uncertainties of individual estimates but also of correlation between estimates.

Acknowledgments. We would like to thank the referees for their valuable comments and suggestions. This work was funded by Research Experiences for Undergraduates grants from the National Science Foundation (DMS-0552571) and the National Security Agency (H98230-06-1-0098), and by the Center for Quantitative Sciences in Biomedicine, North Carolina State University. Funding support also came from the Research and Policy for Infectious Disease Dynamics (RAPIDD) program of the Science and Technology Directory, Department of Homeland Security, and Fogarty International Center, National Institutes of Health. Preliminary results of Sections 5 (Parameter Estimation) and 6.1 (Sensitivity) originated from a summer REU project. These sections, and the remainder of the work, were then developed by the first author.

Appendix: Sensitivity equations for the SIR model. Here we present the sensitivity equations that are relevant for SIR model-based estimation. If prevalence data is being used, then the relevant sensitivities are $\partial I(t_i)/\partial\theta$. Analysis of incidence data would instead make use of $\partial S(t_{i-1})/\partial\theta - \partial S(t_i)/\partial\theta$. (Recall that, for the SIR model considered here, the number of cases that occur over a time interval is equal to the decrease in the number of susceptibles over that time).

Writing the sensitivities of the state variables with respect to the model parameters as $\phi_1 = \partial S/\partial\beta$, $\phi_2 = \partial S/\partial\gamma$, $\phi_3 = \partial I/\partial\beta$, and $\phi_4 = \partial I/\partial\gamma$, the following sensitivity equations are obtained

$$\frac{d\phi_1}{dt} = -\frac{\beta I}{N}\phi_1 - \frac{\beta S}{N}\phi_3 - \frac{SI}{N} \tag{23}$$

$$\frac{d\phi_2}{dt} = -\frac{\beta I}{N}\phi_2 - \frac{\beta S}{N}\phi_4 \tag{24}$$

$$\frac{d\phi_3}{dt} = \frac{\beta I}{N}\phi_1 + \left(\frac{\beta S}{N} - \gamma\right)\phi_3 + \frac{SI}{N} \tag{25}$$

$$\frac{d\phi_4}{dt} = \frac{\beta I}{N}\phi_2 + \left(\frac{\beta S}{N} - \gamma\right)\phi_4 - I, \tag{26}$$

with the initial conditions $\phi_1(0) = \phi_2(0) = \phi_3(0) = \phi_4(0) = 0$.

For the sensitivities of the state variables with respect to initial conditions, writing $\phi_5 = \partial S/\partial S_0$, $\phi_6 = \partial S/\partial I_0$, $\phi_7 = \partial I/\partial S_0$, and $\phi_8 = \partial I/\partial I_0$, we have that

$$\frac{d\phi_5}{dt} = -\frac{\beta I}{N}\phi_5 - \frac{\beta S}{N}\phi_7 \tag{27}$$

$$\frac{d\phi_6}{dt} = -\frac{\beta I}{N}\phi_6 - \frac{\beta S}{N}\phi_8 \tag{28}$$

$$\frac{d\phi_7}{dt} = \frac{\beta I}{N}\phi_5 + \left(\frac{\beta S}{N} - \gamma\right)\phi_7 \tag{29}$$

$$\frac{d\phi_8}{dt} = \frac{\beta I}{N}\phi_6 + \left(\frac{\beta S}{N} - \gamma\right)\phi_8, \tag{30}$$

together with the initial conditions $\phi_5(0) = \phi_8(0) = 1$, and $\phi_6(0) = \phi_7(0) = 0$.

REFERENCES

[1] R. M. Anderson and R. M. May, "Infectious Diseases of Humans," Oxford University Press, Oxford, 1991.

[2] D. T. Anh, M. P. Bonnet, G. Vachaud, C. V. Minh, N. Prieur, L. V. Duc and L. L. Anh, *Biochemical modeling of the Nhue River (Hanoi, Vietnam): Practical identifiability analysis and parameters estimation*, Ecol. Model., **193** (2006), 182–204.

[3] H. T. Banks, M. Davidian, J. R. Samuels Jr. and K. L. Sutton, *An inverse problem statistical methodology summary*, in "Mathematical and Statistical Estimation Approaches in Epidemiology" (eds. G. Chowell, J. M. Hyman, L. M. A. Bettencourt and C. Castillo-Chávez), Springer, New York, (2009), 249–302.
[4] H. T. Banks, S. Dediu and S. L. Ernstberger, *Sensitivity functions and their uses in inverse problems*, Tech. Report CRSC-TR07-12, Center for Research in Scientific Computation, North Carolina State Unversity, July 2007.
[5] H. T. Banks, S. L. Ernstberger and S. L. Grove, *Standard errors and confidence intervals in inverse problems: Sensitivity and associated pitfalls*, J. Inverse Ill-Posed Probl., **15** (2007), 1–18.
[6] R. Bellman and K. J. Åström, *On structural identifiability*, Math. Biosci., **7** (1970), 329–339.
[7] R. Brun, M. Kühni, H. Siegrist, W. Gujer and P. Reichert, *Practical identifiability of ASM2d parameters–systematic selection and tuning of parameter subsets*, Water Res., **36** (2002), 4113–4127.
[8] M. Burth, G. C. Verghese and M. Vélez-Reyes, *Subset selection for improved parameter estimation in on-line identification of a synchronous generator*, IEEE Trans. Power Syst., **14** (1999), 218–225.
[9] A. Capaldi, S. Behrend, B. Berman, J. Smith, J. Wright and A. L. Lloyd, *Parameter estimation and uncertainty quantification for an epidemic model*, Tech. Report CRSC-TR09-18, Center for Research in Scientific Computation, North Carolina State Unversity, August 2009.
[10] S. Cauchemez, P.-Y. Böelle, G. Thomas and A.-J. Valleron, *Estimating in real time the efficacy of measures to control emerging communicable diseases*, Am. J. Epidemiol., **164** (2006), 591–597.
[11] S. Cauchemez and N. M. Ferguson, *Likelihood-based estimation of continuous-time epidemic models from time-series data: Application to measles transmission in London*, J. R. Soc. Interface, **5** (2008), 885–897.
[12] G. Chowell, C. E. Ammon, N. W. Hengartner and J. M. Hyman, *Estimating the reproduction number from the initial phase of the Spanish flu pandemic waves in Geneva, Switzerland*, Math. Biosci. Eng., **4** (2007), 457–470.
[13] G. Chowell, N. W. Hengartner, C. Castillo-Chávez, P. W. Fenimore and J. M. Hyman, *The basic reproductive number of ebola and the effects of public health measures: The cases of Congo and Uganda*, J. Theor. Biol., **229** (2004), 119–126.
[14] G. Chowell, M. A. Miller and C. Viboud, *Seasonal influenza in the United States, France and Australia: Transmission and prospects for control*, Epidemiol. Infect., **136** (2007), 852–864.
[15] A. Cintrón-Arias, H. T. Banks, A. Capaldi and A. L. Lloyd, *A sensitivity matrix based methodology for inverse problem formulation*, J. Inv. Ill-Posed Problems, **17** (2009), 545–564.
[16] A. Cintrón-Arias, C. Castillo-Chávez, L. M. A. Bettencourt, A. L. Lloyd and H. T. Banks, *The estimation of the effective reproductive number from disease outbreak data*, Math. Biosci. Eng., **6** (2009), 261–282.
[17] C. Cobelli and J. J. DiStefano, III, *Parameter and structural identifiability concepts and ambiguities: A critical review and analysis*, Am. J. Physiol. (Regulatory Integrative Comp. Physiol. 8), **239** (1980), R7–R24.
[18] M. Davidian and D. M. Giltinan, "Nonlinear Models for Repeated Measurement Data," Chapman & Hall, 1996.
[19] O. Diekmann and J. A. P. Heesterbeek, "Mathematical Epidemiology of Infectious Diseases. Model Building, Analysis and Interpretation," Wiley Series in Mathematical and Computational Biology, John Wiley & Sons, Ltd., Chichester, 2000.
[20] K. Dietz, *The estimation of the basic reproduction number for infectious diseases*, Stat. Meth. Med. Res., **2** (1993), 23–41.
[21] M. Eslami, "Theory of Sensitivity in Dynamic Systems. An Introduction," Springer-Verlag, Berlin, 1994.
[22] N. D. Evans, L. J. White, M. J. Chapman, K. R. Godfrey and M. J. Chappell, *The structural identifiability of the susceptible infected recovered model with seasonal forcing*, Math. Biosci., **194** (2005), 175–197.
[23] B. Finkenstädt and B. Grenfell, *Empirical determinants of measles metapopulation dynamics in England and Wales*, Proc. R. Soc. Lond. B, **265** (1998), 211–220.
[24] K. Glover and J. C. Willems, *Parametrizations of linear dynamical systems: Canonical forms and identifiability*, IEEE Trans. Auto. Contr., **AC-19** (1974), 640–646.
[25] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Rev., **42** (2000), 599–653.

[26] T. D. Hollingsworth, N. M. Ferguson and R. M. Anderson, *Will travel restrictions control the international spread of pandemic influenza?*, Nature Med., **12** (2006), 497–499.
[27] A. Holmberg, *On the practical identifiability of microbial growth models incorporating Michaelis-Menten type nonlinearities*, Math. Biosci., **62** (1982), 23–43.
[28] J. A. Jacquez and P. Greif, *Numerical parameter identifiability and estimability: Integrating identifiability, estimability and optimal sampling design*, Math. Biosci., **77** (1985), 201–227.
[29] S. Kotz, N. Balakrishnan, C. Read and B. Vidakovic, eds., "Encyclopedia of Statistics," 2^{nd} edition, Wiley-Interscience, Hoboken, New Jersey, 2006.
[30] M. Kretzschmar, S. van den Hof, J. Wallinga and J. van Wijngaarden, *Ring vaccination and smallpox control*, Emerg. Inf. Dis., **10** (2004), 832–841.
[31] P. E. Lekone and B. F. Finkenstädt, *Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study*, Biometrics, **62** (2006), 1170–1177.
[32] A. L. Lloyd, *The dependence of viral parameter estimates on the assumed viral life cycle: Limitations of studies of viral load data*, Proc. R. Soc. Lond. B, **268** (2001), 847–854.
[33] ———, *Sensitivity of model-based epidemiological parameter estimation to model assumptions*, in "Mathematical and Statistical Estimation Approaches in Epidemiology" (eds. G. Chowell, J. M. Hyman, L. M. A. Bettencourt and C. Castillo-Chavez), Springer, New York, (2009), 123–141.
[34] C. D. Meyer, "Matrix Analysis and Applied Linear Algebra," SIAM, Hoboken, New Jersey, 2000.
[35] M. A. Nowak, A. L. Lloyd, G. M. Vasquez, T. A. Wiltrout, L. M. Wahl, N. Bischofberger, J. Williams, A. Kinter, A. S. Fauci, V. M. Hirsch and J. D. Lifson, *Viral dynamics of primary viremia and antiretroviral therapy in simian immunodeficiency virus infection*, J. Virol., **71** (1997), 7518–7525.
[36] J. G. Reid, *Structural identifiability in linear time-invariant systems*, IEEE Trans. Auto. Contr., **AC-22** (1977), 242–246.
[37] A. Saltelli, K. Chan and E. M. Scott, eds., "Sensitivity Analysis," Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd., Chichester, 2000.
[38] M. A. Sanchez and S. M. Blower, *Uncertainty and sensitivity analysis of the basic reproductive rate*, Am. J. Epidemiol., **145** (1997), 1127–1137.
[39] G. A. F. Seber and C. J. Wild, "Nonlinear Regression," John Wiley & Sons, Hoboken, New Jersey, 2003.
[40] K. Thomaseth and C. Cobelli, *Generalized sensitivity functions in physiological system identification*, Ann. Biomed. Eng., **27** (1999), 607–616.
[41] J. Wallinga and M. Lipsitch, *How generation intervals shape the relationship between growth rates and reproductive numbers*, Proc. R. Soc. Lond. B, **274** (2007), 599–604.
[42] H. J. Wearing, P. Rohani and M. Keeling, *Appropriate models for the management of infectious diseases*, PLoS Med., **2** (2005), e174.
[43] L. J. White, N. D. Evans, T. J. G. M. Lam, Y. H. Schukken, G. F. Medley, K. R. Godfrey and M. J. Chappell, *The structural identifiability and parameter estimation of a multispecies model for the transmission of mastitis in dairy cows*, Math. Biosci., **174** (2001), 77–90.
[44] H. Wu, H. Zhu, H. Miao and A. S. Perelson, *Parameter identifiability and estimation of HIV/AIDS dynamic models*, Bull. Math. Biol., **70** (2008), 785–799.
[45] X. Xia and C. H. Moog, *Identifiability of nonlinear systems with application to HIV/AIDS models*, IEEE Trans. Auto. Contr., **48** (2003), 330–336.
[46] H. Yue, M. Brown, F. He, J. Jia and D. B. Kell, *Sensitivity analysis and robust experimental design of a signal transduction pathway system*, Int. J. Chem. Kinet., **40** (2008), 730–741.

Received December 27, 2009; Accepted April 10, 2012.

E-mail address: alex.capaldi@valpo.edu
E-mail address: sbehrend@email.unc.edu
E-mail address: bpberman@math.arizona.edu
E-mail address: jqrs42@msn.com
E-mail address: jwright3@ncsu.edu
E-mail address: alun_lloyd@ncsu.edu