ABSTRACT

CAPALDI, ALEX Exploring the Inverse Problem with Infectious Disease Models. (Under the direction of Dr. Alun L. Lloyd).

In this dissertation, we explore multiple aspects of the inverse problem when applied to infectious disease models. First, we examine estimation of the parameters of Susceptible-Infective-Recovered (SIR) models in the context of least squares (LS). We review the use of asymptotic statistical theory and sensitivity analysis to obtain measures of uncertainty for estimates of the model parameters and basic reproductive number (R_0)—an epidemiologically significant parameter grouping. 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 with a view to better understanding when more frequent sampling (if possible) would prove to be most beneficial to the estimation process. We include a more general discussion of parameter identifiablility in both the epidemic and seasonal endemic SIR settings. We propose an algorithm to select parameter subset combinations that can be estimated using an LS inverse problem formulation with a given data set. The algorithm selects the parameter combinations that correspond to sensitivity matrices with full rank and it involves uncertainty quantification by using the inverse of the Fisher Information Matrix. We conclude with an application of the Akaike information criterion to select a model from a series of epidemic models fitted to an outbreak of influenza in a boy's boarding school in England. We find that an uncommonly used epidemic model, a Susceptible-Infective-Confined-Recovered (SICR), model is the best fitted model and produces an estimate of R_0 of 4.25.

Exploring the Inverse Problem with Infectious Disease Models

by
Alex Capaldi

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fullfillment of the
requirements for the Degree of
Doctor of Philosophy

Applied Mathematics

Raleigh, North Carolina

2010

APPROVED BY:

Dr. James Selgrade	Dr. Kevin Gross
Dr. Alun L. Lloyd Chair of Advisory Committee	Dr. Hien Tran

DEDICATION

To Elizabeth Pletos, my grandmother who gave me the inspiration and strength to complete this work and always has me in her prayers.

BIOGRAPHY

The author was born on August 27, 1982 in Detroit, Michigan to parents Thomas and Esther Capaldi. He grew up in Ferndale, Michigan, alongside his younger brother, Nicholas, and best friends David Mirfin Jr. and Ben Kabel. High school in Ferndale was a very musical experience for the author, as he participated in the various orchestras, bands, choirs and the musical theatre and he also composed his own works. College, however, was decidedly more scientific. After a brief flirtation with majoring in biotechnology and then pre-pharmacy, he eventually earned his B.S. in Applied Mathematics and his B.A. in Chemistry at Ferris State University in 2004.

The author moved to Raleigh, North Carolina in July 2004 where he learned to say "y'all" and earned his Master of Operations Research degree from North Carolina State University in 2006 and his Master of Science in Applied Mathematics in 2008. In the math department at NCSU, he met his beautiful, brilliant and beloved wife, now-Dr. Mindy Capaldi. They were married in the library of Georgetown College, her alma mater, in Georgetown, Kentucky on May 30, 2009.

Outside of mathematics, the author's interests lie in the theory and mechanics of game design (what makes them fun?), traveling the world and birding.

This thesis completes his requirements for a Doctor of Philosophy degree in Applied Mathematics from NCSU. The author begins work as an assistant professor of mathematics in the fall of 2010 at Valparaiso University in Valparaiso, Indiana.

ACKNOWLEDGMENTS

My most sincere thanks to my advisor, Dr. Alun Lloyd, who has been one of the most benevolent and *the* most helpful individual I have ever had the pleasure of working with or learning from. His mentorship has been invaluable and his patience with me has been saintly. I can only hope that I will be as great a mentor to others down the line as he has been to me.

Thanks go to Dr. Ariel Cintrón-Arias who has been a wonderful springboard of ideas. Our collaborations led to the paper [28], of which portions are included in Section 3.2. He also provided me with vital advice with my job search and thesis writing techniques.

My appreciation to the students in our summer 2006, 2007 and 2009 Research Experience for Undergraduates (REU) program at NCSU. Working with them has shown me the joys of mentoring and the experiences helped put me on my road towards professorship. Together, the 2007 REU students and I began initial investigations into issues with parameter estimation when parameters estimates are strongly correlated, which I later developed and turned into Chapter 2 of this work and the tech report [18].

Thank you to my committee, Dr.s Hien Tran, Kevin Gross and Jim Selgrade, whose advice has been useful and whose classes I found interesting, enjoyable and valuable.

Throughout my academic career, many of my professors have been instrumental in my development as a mathematician, scholar and human being. I would like to specifically thank Dr.s Kailash Misra, Melissa Bostrom, Mette Olufsen, Ralph Smith, H. T. Banks, and Ernie Stitzinger at NCSU for all that they have done for me. At Ferris State University, I give thanks to Dr. Kent Sun and Dr. Dan Adsmond who both helped me understand I should head to graduate school and prepared me well for it and to Fran Alegretto who first introduced me to advanced mathematics.

I greatly appreciate the support of my friends and colleagues of whom there are too many to list here.

Finally and most importantly, I give thanks to my beloved wife and parents, whom all have given me an incredible amount of motivation and the support needed to complete the arduous journey of graduate school. I love you three with all my heart.

TABLE OF CONTENTS

LIST OF TABLES . . . . . **viii**

LIST OF FIGURES . . . . . **x**

Chapter 1 Introduction **1**
1.1 Initial Remarks 1
1.2 Background 4
1.2.1 The Basic SIR Model 4
1.2.2 The SEIR Model 5
1.2.3 SIR Models for Endemic Infections 8

Chapter 2 Parameter Estimation **12**
2.1 Methodology: Asymptotic Statistical Theory 13
2.2 Generation of Synthetic Data, Model Fitting and Estimation 19
2.3 Results: Parameter Estimation 22
2.4 Results: Sampling Schemes and Uncertainty of Estimates 27
2.4.1 Sensitivity 29
2.4.2 Data Sampling 30
2.4.3 Data Sampling with Incidence Data 36
2.5 Discussion 38

Chapter 3 Parameter Identifiability and Subset Selection **45**
3.1 Parameter Identifiability 46
3.1.1 Application to Epidemic Scenario 47
3.1.2 Application to Endemic Scenario 49
3.2 Subset Selection Algorithm 51
3.2.1 Application to Endemic Scenario 58
3.3 Discussion 68

Chapter 4 Model Selection **71**
4.1 The Boarding School Data 73
4.1.1 Previous Work 76
4.2 Fitting Compartmental Models to the Data 77
4.2.1 The SEICR Model and Sub-models 77
4.2.2 Time-Dependent Parameters 79
4.3 Distributed Delays 84

4.3.1 Generally Distributed Delays 84
4.3.2 Gamma Distributed Delays 87
4.4 Discussion 91

Chapter 5 Conclusions and Future Directions 96
5.1 Concluding Remarks 96
5.2 Future Directions 97

Bibliography 98

LIST OF TABLES

Table 2.1 Parameter estimates of β, γ, R_0, and the correlation coefficient between estimates of β and γ, $\rho_{\hat{\beta},\hat{\gamma}}$, obtained using a Monte Carlo approach with $10,000$ realizations. The coefficients of variation (CV) obtained from the Monte Carlo were compared to those from the asymptotic stastical theory. The variance-covariance matrix Σ_0 was calculated exactly (*i.e.*, no curve-fitting was carried out) for the "Theory" value and was calculated directly from the realizations of the Monte Carlo for the "MC" value. Calculations were performed 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 $I_0 = 100$. 25

Table 3.1 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. $\xi = 1/2$, $N = 10,000$, $S_0 = 9900$, $I_0 = 100$, $\beta = 3$, $\gamma = 1$ and $\sigma_0^2 = 10^4$. 53

Table 3.2 Nominal parameter values for the SEIRS model. 66

Table 3.3 Feasible parameter vectors obtained while applying the subset selection algorithm for $p = 4, \ldots, 8$, using nominal values as listed earlier in the text. For each selected parameter vector $\theta \in \Theta_p$ the condition number of the sensitivity matrix $\kappa(\chi(\theta))$, and the selection score $\alpha(\theta)$ are displayed. 69

Table 3.4 Results of solving five inverse problems from a single synthetic data set generated as described in the text using nominal values listed earlier. For each parameter combination we display the estimate (Est.), the standard error (SE) and the coefficient of variation (standard error divided by the estimate, CV = SE/Est.). For notational convenience we use here the notation e to denote exponentiation to the base 10; i.e., $2.8e5$ denotes 2.8×10^5, etc. 71

Table 4.1 Here is a hypothetical example of model selection using AIC. Suppose we have four models fit to the same data set, M_i, with respective information criteria AIC_i. Using the Δ_i values for comparison we see that, model M_2, having the minimum AIC value is the best model; model M_3 is second best,

and is still strongly plausible; model M_3 is considerably less plausible and model M_4 has essentially no empirical support. 78

Table 4.2 Influenza epidemic data from a boys boarding school, which was garnered via the DataThief program from the figure in the 1978 paper [6]. The numbers given are those students who were "confined to bed." $N = 763$. 79

Table 4.3 A list of the results from four different models fit using OLS to the boarding school data. The listed values are the number of free structural parameters p for the model, the estimate of R_0, the cost functional value J and the model's AIC_c. The lowest AIC_c gives the model that fits the data the best while using the least number of free parameters, which is the SIR model. 85

Table 4.4 A list of the results from six different models with time-dependent parameters fit using OLS to the boarding school data. The listed values are the number of free structural parameters p for the model, the cost functional value J and the model's AIC_c. The lowest AIC_c gives the model that fits the data the best while using the least number of free parameters, which is the SIR model with piece-wise $\gamma(t)$. 90

Table 4.5 A list of the results from all of the models in this chapter that were fit using OLS to the boarding school data. The listed values are the number of free structural parameters p for the model, the estimate of the basic reproductive number R_0, the cost functional value J, the model's AIC_c and the AIC differences, Δ_i. The models are listed (top to bottom) most plausible to least plausible. 100

LIST OF FIGURES

Figure 1.1 Susceptibles (solid curve) and Infectives (dashed curve) against time for the SIR model. Parameter values used were $\beta = 3$, $\gamma = 1$, and $N = 10,000$ with initial conditions $S_0 = 9900$ and $I_0 = 100$. $R_0 = 3$. 6

Figure 1.2 Forward simulation of the SEIR model. The curves of S (solid), E (dot-dashed) and I (dashed) against time are shown. Parameter values used were $\beta = 3$, $\gamma = 1$, $\nu = 2$, and $N = 10,000$ with initial conditions $S_0 = 9900$, $E_0 = 0$ and $I_0 = 100$. $R_0 = 3$. 8

Figure 1.3 Prevalence versus time for the SIR model with demography. Notice that the system settles to its equilibrium after some time. Parameter values used were $\beta = 3$, $\gamma = 1$, $\mu = 1/10$, and $N = 10,000$ with initial conditions $S_0 = 9900$ and $I_0 = 100$. 10

Figure 1.4 The SIR model with demography and seasonal transmission when the system has reached its limit cycle, a period one attractor. Specifically, the transient behavior of the system has passed as we are looking at 900 years after the infection was first introduced. The first plot displays prevalence versus time and the second is a phase plot (I vs S). As time progresses, the trajectory in phase space is traversed in a counter-clockwise direction. Parameter values used were $a = 0.1$, $\beta_0 = 15/14$, $t_0 = 0$, $\gamma = 1/14$ days^{-1}, $\mu = 1/70$ years^{-1}, and $N = 100,000$. The value of γ and β_0 are such that the model approximates measles (with an $R_0 \approx 15$), a highly-infectious disease with seasonal transmission. 12

Figure 2.1 Synthetic prevalence data sets with R_0 equal to (a) 1.2, (b) 3 and (c) 10. Solid curves depict the prevalence $I(t)$ obtained from the SIR model, while the dots show the synthetic data generated by adding observational noise to $I(t)$ at discrete time points, as discussed in the text. Poisson noise was used ($\xi = 1/2$) where noise variance σ_0^2 equalled 1 and $n = 50$ data points. The initial conditions of the SIR model were $S_0 = 9900$, $I_0 = 100$, where $N = 10,000$ and γ was taken equal to one, so $\beta = R_0$. 22

Figure 2.2 Parameter estimates of β and γ obtained using a Monte Carlo approach with 1000 realizations. The true parameter value point is given by a

circle. Superimposed on the cloud of estimates is the 95% confidence ellipse. 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 $\beta = 3$, $\gamma = 1$, $N = 10,000$, $S_0 = 9900$, and $I_0 = 100$..... 27

Figure 2.3 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 $I_0 = 100$..... 28

Figure 2.4 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 $I_0 = 100$... 31

Figure 2.5 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 $\beta = R_0$..... 34

Figure 2.6 Standard errors of the estimates of β and γ as the number of observations, n, changes while maintaining a constant window of observation (fixed t_{end}). The points fall on a line of slope $-\frac{1}{2}$ on this log-log plot. 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 $I_0 = 100$..... 36

Figure 2.7 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. 38

Figure 2.8 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 $[0, t_{n_{\text{used}}}]$. 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. 39

Figure 2.9 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 (dashed curve). Standard errors were calculated using Equation (2.6) 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 $\sigma_0^2 = 10^2$, $S_0 = 9900$, $I_0 = 100$, $N = 10,000$, $\beta = 3$ and $\gamma = 1$. The additive noise structure, $\xi = 0$ was assumed. 41

Figure 2.10 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 (2.6), 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 standard error (without removing any points) also plotted (dashed curve), and the sensitivity of $I(t)$ with respect to β (dashed curve). Panel (b) shows the standard error for the estimate of γ (solid curve), together the baseline standard error also plotted (dashed curve), and with the sensitivity of $I(t)$ with respect to γ (dashed curve). All parameter values and other details are as in the previous figure. 42

Figure 2.11 Standard errors of estimates of β and γ as the number of observations, n, changes while maintaining a constant window of observation (fixed t_{end}). The points fall on a line of slope $-\frac{1}{2}$ on this log-log plot. Standard errors are calculated using Equation (2.6), 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 incidence 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 $I_0 = 100$..... 43

Figure 2.12 Impact of increasing the length of the observation window on standard errors of estimates of β (dashed curve) and γ (solid curve) when each is estimated separately from incidence 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 incidence curve, $z(t)$, is superimposed. Standard errors are plotted on a logarithmic scale. The exact formula for Σ_0 was used, with parameter values $\sigma_0^2 = 1$, $S_0 = 9900$, $I_0 = 100$, $N = 10,000$, $\beta = 3$ and $\gamma = 1$. The Poisson noise structure, $\xi = 1/2$, was employed..... 44

Figure 2.13 Standard errors for the estimation of β from incidence 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 (dashed curve). Standard errors were calculated using Equation (2.6) and each is plotted at the time t_i corresponding to the removed data point. For comparison, the sensitivity of $z(t)$ with respect to β is also shown (dotted curve). Synthetic data was generated using the parameter values $\sigma_0^2 = 10^2$, $S_0 = 9900$, $I_0 = 100$, $N = 10,000$, $\beta = 3$ and $\gamma = 1$. The additive noise structure, $\xi = 0$ was assumed..... 45

Figure 2.14 Standard errors for the simultaneous estimation of β and γ from incidence data using the single point removal method as discussed in the text (solid curves). Standard errors were calculated using Equation (2.6), 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 standard error (without removing any points) also plotted (dashed curve), and the sensitivity of $z(t)$ with respect to β (dashed curve). Panel (b) shows the standard error for the estimate of γ (solid curve), together the baseline standard error also plotted (dashed curve), and with the sensitivity of $z(t)$ with respect to γ (dashed curve). All parameter values and other details are as in the previous figure..... 46

Figure 3.1 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 $I_0 = 100$..... 52

Figure 3.2 Plots of the coefficients of variation for the model parameters β_0, γ, μ, a and R_0 versus the strength of seasonality a when the system has reached its limit cycle. The highest curve is the coefficient of variation of R_0, which is significantly decreased as the strength of seasonality increases. Parameter values used were $\beta_0 = 15/14$, $\gamma = 1/14$ days, $\mu = 1/70$ years, and $N = 100,000$. The value of γ and β_0 are such that the model approximates measles (with an $R_0 \approx 15$), a highly-infectious disease with seasonal transmission. The exact formula for Σ_0 was used with $\sigma_0^2 = 1$. The absolute noise structure, $\xi = 0$, was employed. 57

Figure 3.3 Parameter selection score $\alpha(\theta)$ versus the condition number $\kappa(\chi(\theta))$ of the $n \times p$ sensitivity matrix, for all parameter vectors $\theta \in \Theta_p$ with $p = 5$. Logarithmic scales are used on both axes. Nominal parameter values used are listed in Table 3.2. 68

Figure 3.4 Residual plots: $y_i - z(t_i; \theta_{OLS})$, versus time, t_i, for $i = 1, \ldots, n$. Graph (a) displays residuals obtained for $\theta = (S_0, N, L, D, M, \beta_0, a_1, b_1)$, while Graph (b) depicts residuals for $\theta = (L, D, \beta_0, a_1, b_1)$. 73

Figure 4.1 A plot of the data of an influenza outbreak in a boys boarding school [6] also presented in Table 4.2. The horizontal axis measures time in days while the vertical axis measures the number of students confined to bed on that day. 80

Figure 4.2 Plot of the number of students confined to bed from the OLS fit of the four models (solid curves) against time together with the boarding school data (dots). The number of students "confined to bed" corresponds to the I class for the SIR and SEIR models and the C class for the SICR and SEICR models. Estimates of R_0 can be found in Table 4.3. $N = 763$ and the initial conditions are $I_0 = 3$ for the SIR and SEIR models and $I_0 = 1$ for the SICR and SEICR models, $C_0 = 3$ and $S_0 = N - C_0 - I_0$. 86

Figure 4.3 The residuals against time and against the model values of prevalence from the OLS fit of the SIR model to the boarding school data. Notice in both plots that the values are centered around 0 and the variation from the mean appears to not follow a particular pattern dependent on t or $I(t)$ which does not provide any evidence that the assumptions about the noise are not upheld. $N = 763$ and the initial conditions are $S_0 = 760$, $I_0 = 3$. 87

Figure 4.4 Plots of the infectious class of the SIR and SICR models against time using the best fit parameters to the boarding school data. Notice that the

SICR model has prevalence right-skewed in comparison to the SIR model. For SIR, $\beta = 1.6924$, $\gamma = 0.4498$, $N = 763$, $S_0 = 760$ and $I_0 = 3$. For SICR, $\beta = 2.8519$, $\gamma = 0.9276$, $\omega = 0.4461$, $N = 763$, $S_0 = 759$, $I_0 = 1$ and $C_0 = 3$.. 88

Figure 4.5 Plot of the number of students confined to bed from the OLS fit of the SICR model with an $\omega(t)$ step-function (solid curve) against time together with the boarding school data (dots). The best fit parameters are $\beta = 2.7614$, $\gamma = 0.6490$, $\omega_1 = 0.4360$, $\omega_2 = 1.0875$, and $t_{\text{switch}} = 9.3362$. The model selection score is $AIC_c = 123.1824$ and the cost functional value is $J = 978.12$. Initial conditions are the same as previous SICR model fits. 91

Figure 4.6 Gamma distributed latency/infectious periods. The graphs represent the p.d.f.s of the gamma distributions with $a = 1$ (dotted), $a = 2$ (dot-dashed), $a = 5$ (dashed) and $a = 50$ (solid). The mean of the distribution in each case is two. When $a = 1$, we have the exponential distribution. When $a = 50$, the curve approaches the p.d.f. of the normal distribution. 94

Figure 4.7 The cost function surface J across values of the shape parameters a_E and a_I of the gamma distribution for the PDE SEIR model with gamma distributed duration of latency and infection. The model was fit to the boarding school data using OLS at the fixed a_E and a_I values, while fitting β, $1/\nu$ and $1/\gamma$. 98

Figure 4.8 Plot of predicted prevalence from the OLS fit of the SEIR PDE model (solid curve) against time together with the boarding school data (dots). The best fit parameters are $\beta = 27.3887$, $a_E = 2.5448$, $a_I = 1.9559$, $1/\nu = 2.8610$, and $1/\gamma = 2.0763$. The basic reproductive number was found to be $R_0 = 56.8670$. The model selection score is $AIC_c = 135.2282$ and the cost functional value is $J = 2312.4$. $N = 763$ and the initial conditions are $S(0) = 760$, $E(0;\tau) = 0$ and $I(0;\tau) = 3$. 99

Chapter 1

Introduction

1.1 Initial Remarks

The use of mathematical models to interpret disease outbreak data has provided many insights 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, [4] and [34]), 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 [20, 47, 57]).

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, [7, 23, 61, 77, 83, 84]), obtaining a set of parameters that provides the best match, in some sense, between model output and data (this process of finding optimal parameters is commonly referred to as solving the inverse problem). 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.

Mathematical models have been used to study questions in epidemiology as early as 1927 when Kermack and McKendrick introduced the SIR model [54]. However, a strong appreciation of the uncertainty of parameter estimates did not take a foothold in the field of mathematical epidemiology until much more recently. The studies of Anderson and others of the bovine spongiform encephalopathy (mad-cow disease) outbreak in the United Kingdom (UK) in 1996 was revolutionary in the field in terms of the statistical tools employed [2]. The field advanced as epidemiologists covered the emerging outbreaks of foot and mouth disease in the UK in 2001 [38, 53], SARS in 2003 [24, 25, 74], avian influenza in 2006-7 [55, 82], H1N1 (swine flu) in 2009 (for example, [42]). New analyses of older outbreaks for which there was quality data available also occurred; for instance, researchers have used statistical methods of uncertainty measurement when modeling the 1918 Spanish flu pandemic [22, 68] and the pre- and post-vaccination era measles data from the UK [21, 39, 41, 40].

The simultaneous estimation of several parameters raises questions of parameter identifiability (see, for example, [5, 12, 30, 37, 43, 48, 72, 85, 88, 89, 90]), 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. To address the question of parameter identifiability, we will propose an algorithm based on those introduced by [15, 17] to select parameter combinations (vectors) based on the sensitivity of model states to the parameters and on the uncertainty of estimates.

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 work we discuss the use of sensitivity analysis [36] and asymptotic statistical theory described in [80] and [8], to quantify the uncertainties associated with parameter estimates obtained by the use of least squares model fitting. 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.

Many model fits are often 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 often 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 has been shown that it can dwarf uncertainty due to the noise in the data. As an example, a number of authors have shown that estimates of the basic reproductive number based on the initial growth of an outbreak can be highly sensitive to assumptions made about the structure of the model [61, 71, 84]. We will present tools previously

developed that can be used for selecting an empirically best model from a set of similar models fit to a given data set. We will then use these tools on a set of outbreak data to obtain the best possible estimate of R_0.

The work is organized as follows: the different epidemiological models employed throughout our study are outlined in the following Background section. Chapter 2 presents a discussion on parameter estimation and uncertainty quantification in the case of a simple epidemic model. The selection of a subset of identifiable parameters when conducting the inverse problem is discussed in Chapter 3. Chapter 4 will give a case study of data from an influenza outbreak in a boys' boarding school to answer questions about how selecting the incorrect underlying model can severely affect the interpretation of both quantitative and qualitative results. We conclude with a discussion of the results and future work.

1.2 Background

1.2.1 The Basic SIR Model

We first 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, [4, 33, 46]) 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

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{1.1}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I, \tag{1.2}$$

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: $z(t_i) = \int_{t_{i-1}}^{t_i} \beta S(t)I(t)/N \, dt$. 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.

We present a forward simulation of the SIR model when $R_0 > 1$ in Figure 1.1 for illustrative purposes. Notice that once the epidemic has waned, there still remain some susceptible individuals. We will refer to this horizontal asymptote as $S(\infty)$.

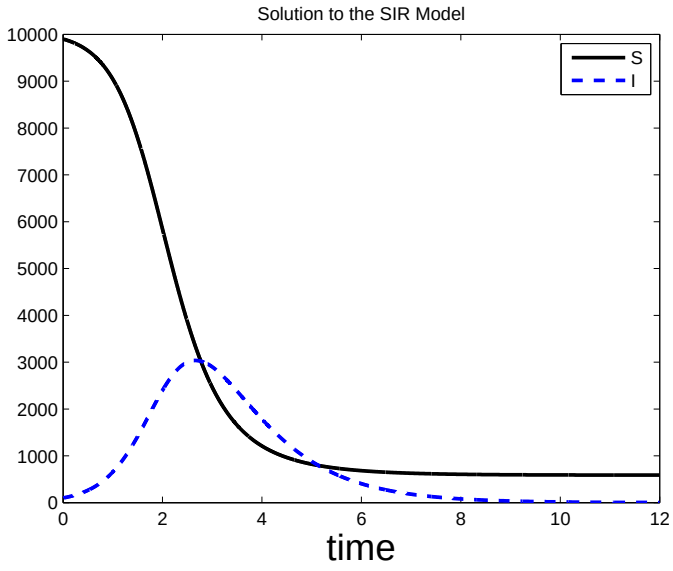

Figure 1.1: Susceptibles (solid curve) and Infectives (dashed curve) against time for the SIR model. Parameter values used were $\beta = 3$, $\gamma = 1$, and $N = 10,000$ with initial conditions $S_0 = 9900$ and $I_0 = 100$. $R_0 = 3$.

1.2.2 The SEIR Model

One assumption of the SIR model is that once infected, an individual immediately becomes infectious. In reality, there is often a latent period between getting infected and becoming infectious. This is typically due to the parasite needing to reproduce enough to spread between hosts (there is a minimum threshold of viral load necessary for probable transmission). To simulate this delay, we can add an exposed class to the SIR model to create the SEIR model.

If we take the latent period to be exponentially distributed with average duration $1/\nu$, then our model can be described as follows [63]

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{1.3}$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \nu E \tag{1.4}$$

$$\frac{dI}{dt} = \nu E - \gamma I. \tag{1.5}$$

Notice that the SIR model can be regained from this formulation simply by letting $\nu \to \infty$, which removes the delay.

The basic reproductive number for this SEIR model is again $R_0 = \beta/\gamma$, as neither the rate of transmission, nor the average duration of infection have changed from the SIR model. The fact that R_0 is the same for both models is result of choosing the simplest case for the SEIR model here. If there was be a way to leave the exposed class besides entering the infectious class (such as death, or some kind of treatment), we would have to multiple the R_0 value from the SIR model by the probability of continuing on to the infectious class to obtain the SEIR model's formula for R_0. With this simple SEIR model, though, that probability is 1.

We present a forward simulation of the SEIR model in Figure 1.2. Notice that despite similar parameter values as the plot of the SIR model, the prevalence curve for the SEIR model has a later peak than the SIR model because of the latent period. The peak prevalence value is also lower than in the SIR model, and the outbreak is more spread out over time. In contrast, $S(\infty)$ is the same here as it is for the SIR model. Even though the latent period changes the timing of the epidemic, it does not change the outbreak size. Again, this is because we do not have other methods of leaving the exposed class.

1.2.3 SIR Models for Endemic Infections

In an infectious disease model where the susceptible population is regularly replenished (*e.g.* due to births, immigration, loss of immunity, etc.), an endemic ensues. That is, the infection is able to persist indefinitely within the population. The simplest example of this is when the infectious population achieves an equilibrium. We shall examine the standard SIR model (Equation 1.1) with the inclusion of births and deaths.

As with the standard SIR model, the population is divided into three classes, susceptible, infectious and recovered, whose numbers are denoted by S, I, and R, respectively. We choose to have the per capita birth rate, μ, equal the per capita death

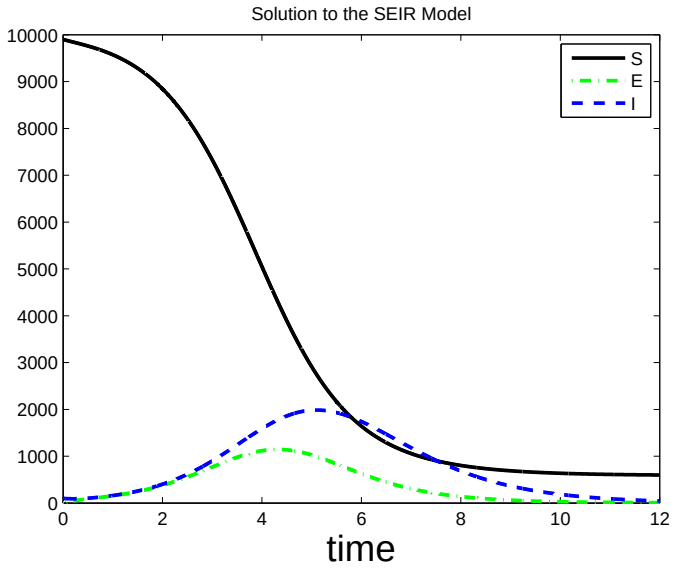

Figure 1.2: Forward simulation of the SEIR model. The curves of S (solid), E (dot-dashed) and I (dashed) against time are shown. Parameter values used were $\beta = 3$, $\gamma = 1$, $\nu = 2$, and $N = 10,000$ with initial conditions $S_0 = 9900$, $E_0 = 0$ and $I_0 = 100$. $R_0 = 3$.

rate to maintain a constant population. Such an assumption is a fair approximation of the demographics in a developed world country. Including births and deaths is typically only done on longer time scales (such as years)—it is typically acceptable to ignore for a single outbreak on a short time scale.

The basic reproductive number, R_0, can be expressed as the product of the transmission parameter and the average duration of infection. For this model, the average duration of infection is $1/(\gamma + \mu)$, (the duration of infection is shorter than in the standard SIR model since people can now leave the infectious class by two methods, recovery or death/emigration) thus,

$$R_0 = \frac{\beta}{\gamma + \mu}. \tag{1.6}$$

The model can be described by the following differential equations

$$\frac{dS}{dt} = \mu N - \frac{\beta SI}{N} - \mu S \tag{1.7}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I - \mu I, \tag{1.8}$$

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

To find the equilibrium of the model, we simply set $\frac{dS}{dt} = \frac{dI}{dt} = 0$ and solve for S and I. This gives us the equilibrium point

$$(S^*, I^*) = \left(\frac{N}{R_0}, \frac{\mu N(1 - 1/R_0)}{\gamma + \mu} \right). \tag{1.9}$$

Many infections that persist do not sit at an equilibrium, instead exhibiting some sort of periodic dynamics—recurrent epidemics. These limit cycles can be brought about from seasonal forcing, such as changes in transmission due to weather conditions or the school year cycle. Such models have been developed and studied by a number of authors (for example, [3, 4, 22, 29, 35, 39, 62]). We choose to study one of the simplest endemic models with seasonal forcing, the SIR model with demography and a periodic transmission parameter.

We modify the Equations 1.7-1.8 to introduce seasonal transmission by making β a dynamic function of time. One such way to do this, and the one we will employ is

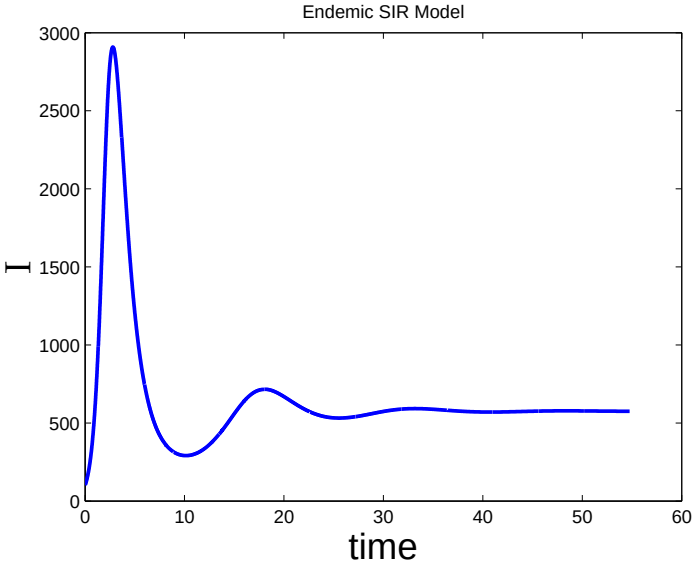

Figure 1.3: Prevalence versus time for the SIR model with demography. Notice that the system settles to its equilibrium after some time. Parameter values used were $\beta = 3$, $\gamma = 1$, $\mu = 1/10$, and $N = 10,000$ with initial conditions $S_0 = 9900$ and $I_0 = 100$.

to set

$$\beta(t) = \beta_0 \left(1 + a \sin 2\pi(t - t_0)\right), \tag{1.10}$$

where β_0 is the baseline transmission parameter, a is the strength of seasonality and t_0 is the transmission phase shift (which is used to align the baseline transmission value to the appropriate time point it occurs during the time course of a season). The period of this transmission function is one, however, depending on the model parameters, the *prevalence* curve could have a different period ("seasons" can be annual, biannual or any number of different kinds of cycles, some chaotic, depending on a [62]).

When there is weak seasonal forcing (a low value of a), the system exhibits small amplitude annual oscillations in phase space about the point (S^*, I^*) from Equation 1.9. This can be seen in Figure 1.4.

The basic reproductive number for the seasonal model is more complicated, but it is similar to that of our non-seasonal, demographic model,

$$R_0 \approx \frac{\beta(t)}{\gamma + \mu}. \tag{1.11}$$

There exists a maximum transmission potential in the above equation at any given point in time since β is dynamic, but a true formulation for R_0 involves some time averaging to account for this fluctuation. However, there is disagreement within the literature community whether that averaging should be arithmetic or geometric (for example, [64, 86]).

With epidemic models, 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. It is in this scenario where the basic reproductive number is useful to determine the strength of the outbreak. 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 quantity than R_0. For the SIR model, R_t is given by

$$R_t = R_0 \frac{S(t)}{N}. \tag{1.12}$$

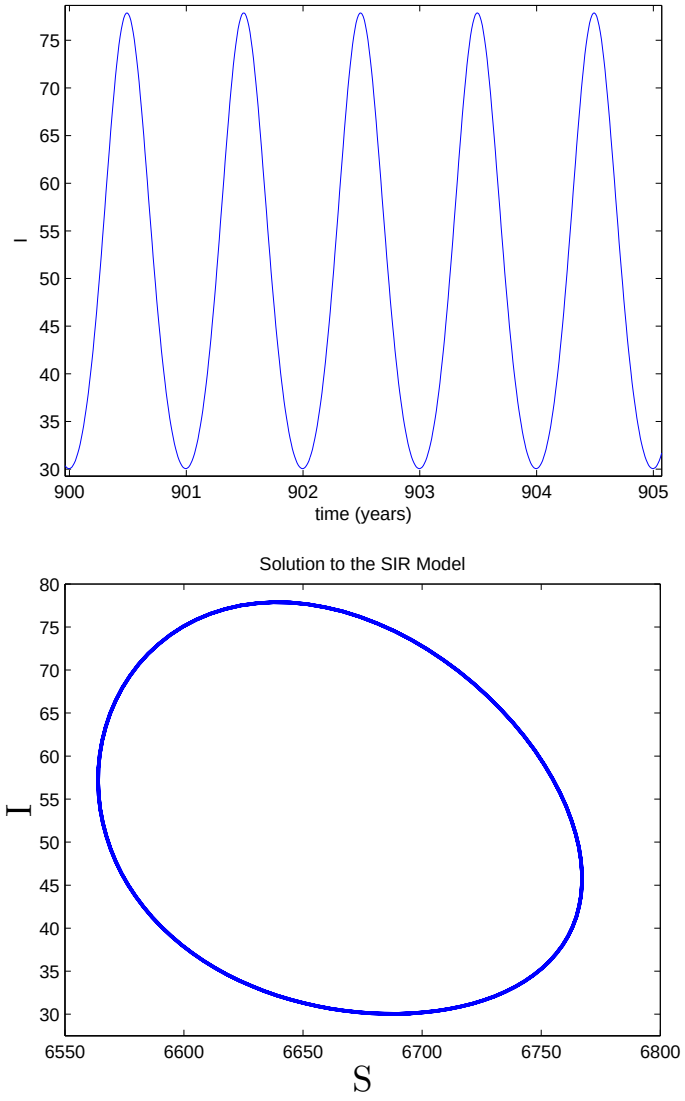

Figure 1.4: The SIR model with demography and seasonal transmission when the system has reached its limit cycle, a period one attractor. Specifically, the transient behavior of the system has passed as we are looking at 900 years after the infection was first introduced. The first plot displays prevalence versus time and the second is a phase plot (I vs S). As time progresses, the trajectory in phase space is traversed in a counter-clockwise direction. Parameter values used were $a = 0.1$, $\beta_0 = 15/14$, $t_0 = 0$, $\gamma = 1/14$ days^{-1}, $\mu = 1/70$ years^{-1}, and $N = 100,000$. The value of γ and β_0 are such that the model approximates measles (with an $R_0 \approx 15$), a highly-infectious disease with seasonal transmission.

In Chapter 3, we will investigate how much information about the model parameters we can obtain when we observe limit cycle behavior of an endemic infection.

Chapter 2

Parameter Estimation

This chapter is an expanded version of the paper "Parameter Estimation and Uncertainty Quantification for an Epidemic Model," *which has been accepted by the journal* Mathematical Biosciences and Engineering (MBE) *for publication pending minor revisions. The paper was co-authored by myself (the first author), Dr. Alun Lloyd (the senior author) and the four students from our 2007 REU, Sam Behrend, Ben Berman, Jason Smith and Justin Wright who conducted some of the initial investigations into the estimation issues described in this chapter.*

An understanding of the uncertainty in parameter estimates is crucial. Therefore, in this chapter, we present a methodology to calculate standard errors of parameter estimates and other measures of uncertainty. We then give some methods of how to sample data to increase the information one has about the model parameters, reducing uncertainty in the 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.

2.1 Methodology: Asymptotic Statistical Theory

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) [32]. Uncertainty quantification is performed using asymptotic statistical theory (see, for example, Seber and Wild [80]) then applied to the statistical model that describes the epidemiological data set. We provide a general summary of this theory here.

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

$$Y_i = M(t_i; \theta_0) + \mathcal{E}_i, \tag{2.1}$$

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 $\mathcal{E}_i$ depict the errors. We write $Y = (Y_1, \ldots, Y_n)^T$.

The appropriate estimation procedure depends on the properties of the errors $\mathcal{E}_i$.

We assume that the errors have the following form

$$\mathcal{E}_i = M(t_i;\theta_0)^{\xi}\epsilon_i, \tag{2.2}$$

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 $\mathcal{E}_i = \epsilon_i$, 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 referred to as absolute noise in the literature, which we will do. 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, which will we do. 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

$$J(\theta|Y) = \sum_{i=1}^{n} w_i(Y_i - M(t_i;\theta))^2, \tag{2.3}$$

in which the weights w_i are given by

$$w_i = \frac{1}{M(t_i;\theta)^{2\xi}}. \tag{2.4}$$

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

$$\hat{\theta}_{\mathrm{LS}} = \arg\min_{\theta} J(\theta|Y). \tag{2.5}$$

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 (2.3). 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 [32].

The asymptotic statistical theory, as detailed in [80], 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 [80]), this estimator has a p-dimensional multivariate normal distribution with mean θ_0 and variance-covariance matrix Σ_0 given by

$$\Sigma_0 = \lim_{n\to\infty} \Sigma_0^{(n)} = \lim_{n\to\infty} \sigma_0^2 \left(n\Omega_0^{(n)} \right)^{-1}, \tag{2.6}$$

where

$$\Omega_0^{(n)} = \frac{1}{n} \chi^{(n)}(\theta_0)^T W^{(n)}(\theta_0) \chi^{(n)}(\theta_0). \tag{2.7}$$

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

We note that the 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 (2.7), $W^{(n)}(\theta)$ 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

$$\chi^{(n)}(\theta)_{ij} = \frac{\partial M(t_i;\theta)}{\partial \theta_j}. \tag{2.8}$$

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 [76], for the general m-dimensional system

$$\dot{x} = F(x, t; \theta), \tag{2.9}$$

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

$$\frac{d}{dt}\frac{\partial x}{\partial \theta} = \frac{\partial F}{\partial x}\frac{\partial x}{\partial \theta} + \frac{\partial F}{\partial \theta}, \tag{2.10}$$

with initial conditions

$$\frac{\partial x(0)}{\partial \theta} = 0_{m\times p}. \tag{2.11}$$

Here, $\partial F/\partial x$ is the Jacobian matrix of the system. This initial value problem must be solved simultaneously with the original system (2.9).

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 (2.10) is absent and the appropriate matrix of initial conditions is $I_{m\times m}$.

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

$$\frac{d\phi_1}{dt} = -\frac{\beta I}{N}\phi_1 - \frac{\beta S}{N}\phi_3 - \frac{SI}{N} \tag{2.12}$$

$$\frac{d\phi_2}{dt} = -\frac{\beta I}{N}\phi_2 - \frac{\beta S}{N}\phi_4 \tag{2.13}$$

$$\frac{d\phi_3}{dt} = \frac{\beta I}{N}\phi_1 + \left(\frac{\beta S}{N} - \gamma\right)\phi_3 + \frac{SI}{N} \tag{2.14}$$

$$\frac{d\phi_4}{dt} = \frac{\beta I}{N}\phi_2 + \left(\frac{\beta S}{N} - \gamma\right)\phi_4 - I, \tag{2.15}$$

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

$$\frac{d\phi_5}{dt} = -\frac{\beta I}{N}\phi_5 - \frac{\beta S}{N}\phi_7 \tag{2.16}$$

$$\frac{d\phi_6}{dt} = -\frac{\beta I}{N}\phi_6 - \frac{\beta S}{N}\phi_8 \tag{2.17}$$

$$\frac{d\phi_7}{dt} = \frac{\beta I}{N}\phi_5 + \left(\frac{\beta S}{N} - \gamma\right)\phi_7 \tag{2.18}$$

$$\frac{d\phi_8}{dt} = \frac{\beta I}{N}\phi_6 + \left(\frac{\beta S}{N} - \gamma\right)\phi_8, \tag{2.19}$$

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

The observed data $y_1, y_2, \ldots, y_n$ represent a realization of the observation process, and our estimate of θ is a realization of the estimator $\hat{\theta}_{\text{LS}}$. Residuals for the model fit are defined as

$$r_i(\theta) = \frac{y_i - M(t_i;\theta)}{M(t_i;\theta)^{\xi}}, \tag{2.20}$$

and, when $\theta = \theta_0$, represent a realization of a set of independent draws from the ϵ_i distribution. Plots of residuals against time t, or against the model M, are often examined as a diagnostic test to determine if the noise structure and other assumptions of the statistical model are appropriate ([11, 29]).

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

$$\sigma^2 = \frac{1}{n-p}\sum_{i=1}^{n} w_i(M(t_i;\theta) - y_i)^2, \tag{2.21}$$

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

$$\Sigma = \sigma^2[\chi^T(\theta)W(\theta)\chi(\theta)]^{-1} \tag{2.22}$$

provides an approximation to the covariance matrix Σ_0.

Standard errors for the components of the estimator $\hat{\theta}_{\text{LS}}$ 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) given by

$$CV_{\hat{\theta}_i} = \frac{\sigma_{\hat{\theta}_i}}{\hat{\theta}_i}. \tag{2.23}$$

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

$$\rho_{\hat{\theta}_i,\hat{\theta}_j} = \frac{\mathrm{cov}(\hat{\theta}_i,\hat{\theta}_j)}{\sqrt{\mathrm{Var}(\hat{\theta}_i)\mathrm{Var}(\hat{\theta}_j)}}. \tag{2.24}$$

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 [56]), and obtain

$$\mathrm{E}\left(\frac{\hat{\beta}}{\hat{\gamma}}\right) \approx \frac{\beta_0}{\gamma_0}\left(1 - \frac{\mathrm{cov}(\hat{\beta},\hat{\gamma})}{\beta_0\gamma_0} + \frac{\mathrm{Var}(\hat{\gamma})}{\gamma_0^2}\right), \tag{2.25}$$

and

$$\mathrm{Var}\left(\frac{\hat{\beta}}{\hat{\gamma}}\right) \approx \left(\frac{\beta_0}{\gamma_0}\right)^2\left(\frac{\mathrm{Var}(\hat{\beta})}{{\beta_0}^2} + \frac{\mathrm{Var}(\hat{\gamma})}{{\gamma_0}^2} - \frac{2\mathrm{cov}(\hat{\beta},\hat{\gamma})}{\beta_0\gamma_0}\right). \tag{2.26}$$

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.* [26]. Equation (2.25), 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

$$\left(1 - \frac{\mathrm{cov}(\hat{\beta},\hat{\gamma})}{\beta_0\gamma_0} + \frac{\mathrm{Var}(\hat{\gamma})}{\gamma_0^2}\right) = \left(1 - \rho_{\hat{\beta},\hat{\gamma}}CV_{\hat{\beta}}CV_{\hat{\gamma}} + CV_{\hat{\gamma}}^2\right). \tag{2.27}$$

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.

2.2 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

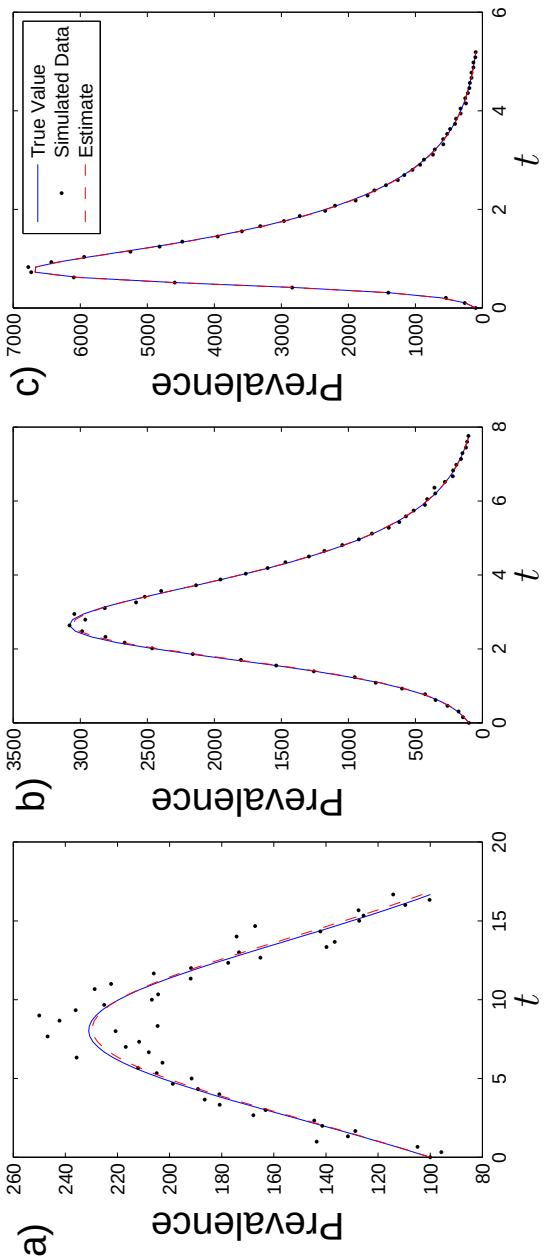

Figure 2.1: Synthetic prevalence data sets with R_0 equal to (a) 1.2, (b) 3 and (c) 10. Solid curves depict the prevalence $I(t)$ obtained from the SIR model, while the dots show the synthetic data generated by adding observational noise to $I(t)$ at discrete time points, as discussed in the text. Poisson noise was used ($\xi = 1/2$) where noise variance σ_0^2 equalled 1 and $n = 50$ data points. The initial conditions of the SIR model were $S_0 = 9900$, $I_0 = 100$, where $N = 10,000$ and γ was taken equal to one, so $\beta = R_0$.

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,

$$y(t_i, \theta_0) \equiv I(t_i, \theta_0) + I(t_i, \theta_0)^\xi e_i, \qquad i = 1, 2, \ldots, n, \tag{2.28}$$

satisfies the assumptions made in the previous section 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 [41]. 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 2.6) 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).

2.3 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 simple 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 chapter.

The least squares estimation procedure works well for synthetic data sets generated using the three different values of R_0 (see Figure 2.1). 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).

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.

During the Monte Carlo, it was verified that the bias factor for the estimation of R_0 (see Equation 2.27) was relatively insignificant, as its difference from 1 was on the order of 10^{-7} in the each case we studied ($R_0 = 1.2$, 3 and 10). It was summarily

Table 2.1: Parameter estimates of β, γ, R_0, and the correlation coefficient between estimates of β and γ, $\rho_{\hat{\beta},\hat{\gamma}}$, obtained using a Monte Carlo approach with $10,000$ realizations. The coefficients of variation (CV) obtained from the Monte Carlo were compared to those from the asymptotic stastical theory. The variance-covariance matrix Σ_0 was calculated exactly (*i.e.*, no curve-fitting was carried out) for the "Theory" value and was calculated directly from the realizations of the Monte Carlo for the "MC" value. Calculations were performed 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	**True Value**	**MC Estimate**	**MC CV**	**Theory CV**
β	1.2	1.200047	0.0121	0.0121
γ	1	0.999992	0.111	0.0110
R_0	1.2	1.200045	0.0023	0.0023
$\rho_{\hat{\beta},\hat{\gamma}}$	0.9837	0.9840	-	-
Parameter	**True Value**	**MC Estimate**	**MC CV**	**Theory CV**
β	3	2.999958	0.0020	0.0019
γ	1	0.999980	0.0034	0.0034
R_0	3	3.000051	0.0037	0.0037
$\rho_{\hat{\beta},\hat{\gamma}}$	0.1132	0.1302	-	-
Parameter	**True Value**	**MC Estimate**	**MC CV**	**Theory CV**
β	10	9.999670	0.0035	0.0035
γ	1	0.999993	0.0027	0.0027
R_0	10	9.999868	0.0050	0.0050
$\rho_{\hat{\beta},\hat{\gamma}}$	-0.3122	-0.3062	-	-

ignored and R_0 was estimated by the ratio of the point estimates β and γ.

In particular, it is seen in Figure 2.2 that roughly 95% of the estimates fall within the ellipse that is the level curve of the pdf of the bivariate normal distribution that contains 95% of the mass of the distribution, namely

$$h_{1-\alpha} = \frac{1}{1-\rho^2_{\hat{\beta},\hat{\gamma}}} \left(\frac{(\beta-\beta_0)^2}{\sigma^2_{\hat{\beta}}} - \frac{2\rho_{\hat{\beta},\hat{\gamma}}(\beta-\beta_0)(\gamma-\gamma_0)}{\sigma_{\hat{\beta}}\sigma_{\hat{\gamma}}} + \frac{(\gamma-\gamma_0)^2}{\sigma^2_{\hat{\gamma}}} \right), \tag{2.29}$$

where $h_{.95} = 5.991$. Both the shape and orientation of this ellipse depend on the correlation coefficient of $\hat{\beta}$ and $\hat{\gamma}$, *i.e.*, on the extent to which these two estimates are correlated [56].

Figure 2.3a demonstrates 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 2.3b).

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 2.4a). 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.

An explanation for the increased correlation of estimates of β and γ as $R_0 \to 1$ has to do with the increased symmetry of the epidemic curve as $R_0 \to 1$ (See Figure 2.1). The symmetry property of the epidemic curve is well known in the literature and can be seen in early works such as [54].

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 2.4b), allowing for easier identification of the two individual parameters.

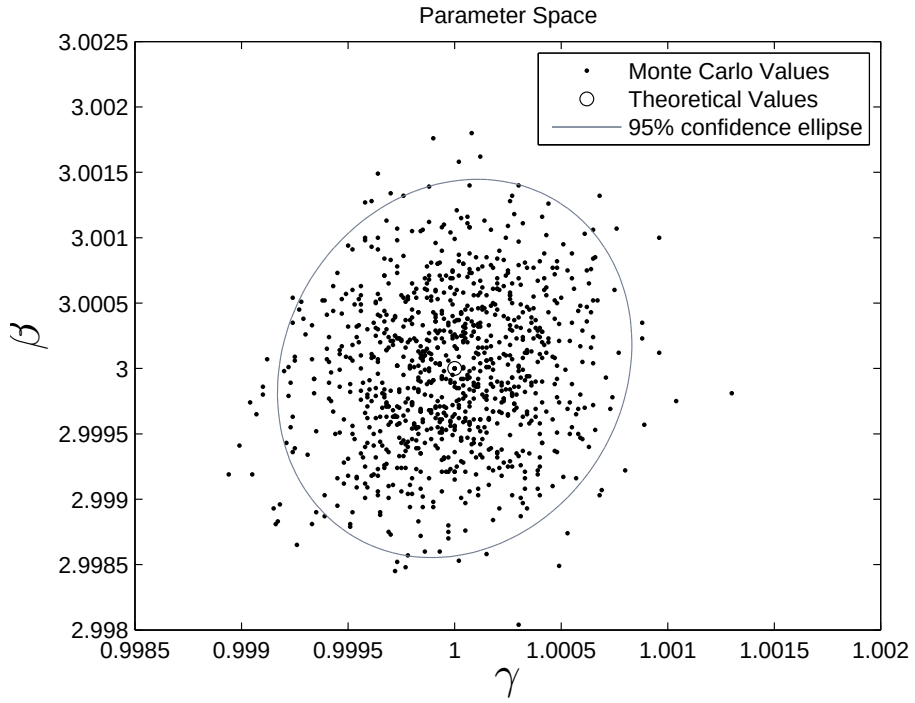

Figure 2.2: Parameter estimates of β and γ obtained using a Monte Carlo approach with 1000 realizations. The true parameter value point is given by a circle. Superimposed on the cloud of estimates is the 95% confidence ellipse. Calculations were done under a Poisson noise structure, $\xi = 1/2$, with $\sigma_0^2 = 1$, and $n = 50$ data points. Parameter values and initial conditions used were $\beta = 3$, $\gamma = 1$, $N = 10,000$, $S_0 = 9900$, and $I_0 = 100$.

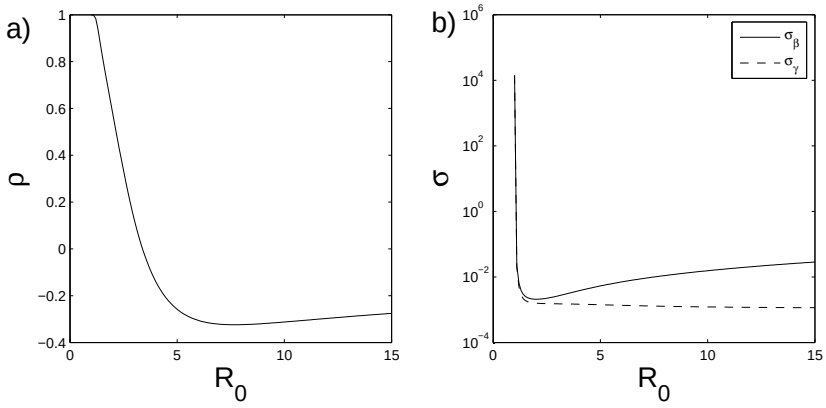

Figure 2.3: Dependence of the correlation coefficient and standard errors for estimates of β and γ on the value of R_0. Panel (a) displays the correlation coefficient, ρ, between estimates of β and γ for a range of R_0 values. Panel (b) shows, on a log scale, standard errors for estimates of β (solid 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 $I_0 = 100$.

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 2.5). 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.

A linear expansion of the cost functional $J(\beta, \gamma)$ about its minimum, (β_0, γ_0), will generically give elliptical contours, so it is not too unexpected that the elliptical contours are closely related to the $(1-\alpha)$-confidence ellipses in the local regime [50]. Specifically, if we expand J about its minimum, $(\beta_0,\ \gamma_0)$, we get:

$$\begin{aligned}
J(\beta,\gamma) &= \sum_{i=1}^{n} w_i (I(t_i;\beta,\gamma) - I(t_i;\beta_0,\gamma_0))^2 \\
&\approx \sum_{i=1}^{n} w_i \left(I(t_i;\beta_0,\gamma_0) + (\beta-\beta_0)\frac{\partial I(t_i;\beta_0,\gamma_0)}{\partial \beta} \right. \\
&\quad \left. +(\gamma-\gamma_0)\frac{\partial I(t_i;\beta_0,\gamma_0)}{\partial \gamma} - I(t_i;\beta_0,\gamma_0) \right)^2 \\
&= \sum_{i=1}^{n} w_i \left((\beta-\beta_0)\frac{\partial I(t_i;\beta_0,\gamma_0)}{\partial \beta} + (\gamma-\gamma_0)\frac{\partial I(t_i;\beta_0,\gamma_0)}{\partial \gamma} \right)^2 \\
&= (\beta-\beta_0)^2 \sum_{i=1}^{n} w_i \left(\frac{\partial I(t_i;\beta_0,\gamma_0)}{\partial \beta} \right)^2 + (\gamma-\gamma_0)^2 \sum_{i=1}^{n} w_i \left(\frac{\partial I(t_i;\beta_0,\gamma_0)}{\partial \gamma} \right)^2 + \\
&\qquad 2(\beta-\beta_0)(\gamma-\gamma_0) \sum_{i=1}^{n} w_i \left(\frac{\partial I(t_i;\beta_0,\gamma_0)}{\partial \beta} \right) \left(\frac{\partial I(t_i;\beta_0,\gamma_0)}{\partial \gamma} \right)
\end{aligned}$$

Using the finite n approximation to equations (2.6) and (2.7), we have that $\sigma_0^2 \Sigma_0^{-1} = \chi^T W \chi$. Specifying the elements of the inverse of the 2×2 covariance matrix in this

equation gives

$$\frac{\sigma_0^2}{\sigma_{\hat{\beta}}^2\sigma_{\hat{\gamma}}^2(1-\rho_{\hat{\beta},\hat{\gamma}}^2)} \begin{bmatrix} \sigma_{\hat{\gamma}}^2 & -\rho_{\hat{\beta},\hat{\gamma}}\sigma_{\hat{\beta}}\sigma_{\hat{\gamma}} \\ -\rho_{\hat{\beta},\hat{\gamma}}\sigma_{\hat{\beta}}\sigma_{\hat{\gamma}} & \sigma_{\hat{\beta}}^2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n w_i \left(\frac{\partial I}{\partial \beta}\right)^2 & \sum_{i=1}^n w_i \left(\frac{\partial I}{\partial \beta}\right)\left(\frac{\partial I}{\partial \gamma}\right) \\ \sum_{i=1}^n w_i \left(\frac{\partial I}{\partial \beta}\right)\left(\frac{\partial I}{\partial \gamma}\right) & \sum_{i=1}^n w_i \left(\frac{\partial I}{\partial \gamma}\right)^2 \end{bmatrix}.$$

The above matrix equation gives us three unique equations for the summations in our formulation of J. For example, $\sum_{i=1}^n w_i \left(\frac{\partial I}{\partial \beta}\right)^2 = \frac{\sigma_0^2}{\sigma_{\hat{\beta}}^2(1-\rho_{\hat{\beta},\hat{\gamma}}^2)}$. Thus, near its minimum, the cost function can be approximated by

$$J(\beta,\gamma) \approx \frac{\sigma_0^2}{1-\rho_{\hat{\beta},\hat{\gamma}}^2}\left(\frac{(\beta-\beta_0)^2}{\sigma_{\hat{\beta}}^2} - \frac{2\rho_{\hat{\beta},\hat{\gamma}}(\beta-\beta_0)(\gamma-\gamma_0)}{\sigma_{\hat{\beta}}\sigma_{\hat{\gamma}}} + \frac{(\gamma-\gamma_0)^2}{\sigma_{\hat{\gamma}}^2}\right). \tag{2.30}$$

We see that contours of the cost function are indeed of the same form as the confidence ellipses described above.

As seen in Table 2.1, 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) [27]—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 (2.26), and reflecting the earlier observation concerning the orientation of the contours of the cost functional along lines of the form $\beta = R_0\gamma$. To yield these lines with slope R_0 is why we choose to plot in the (γ,β) plane rather than (β,γ).

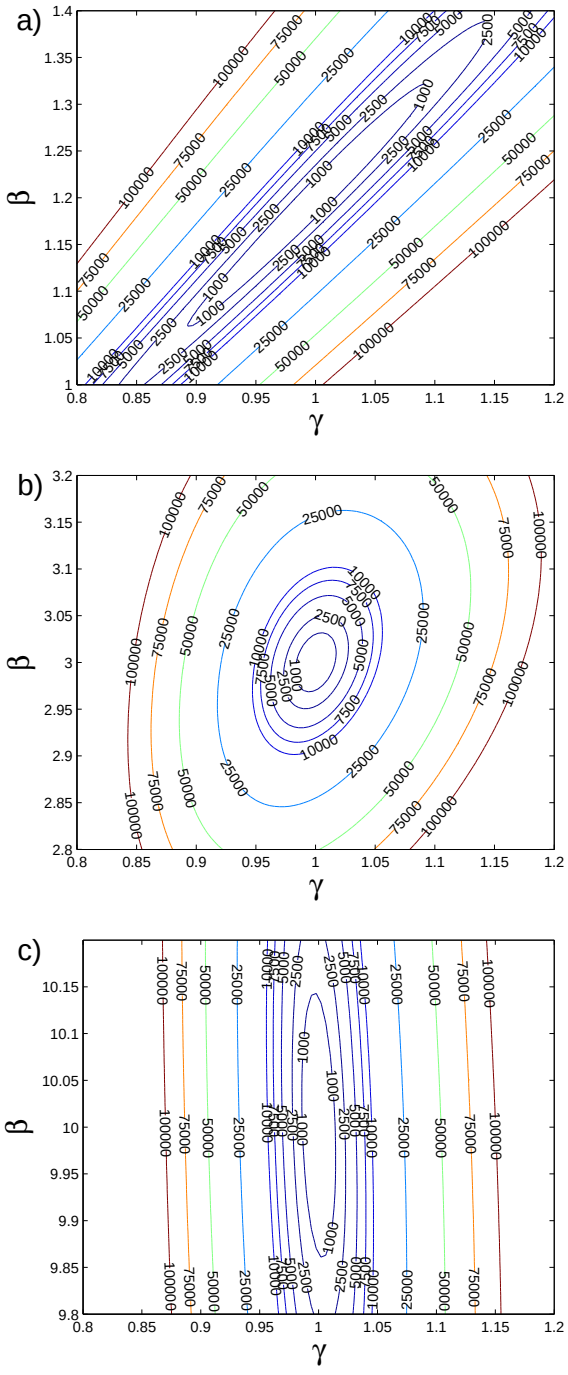

Figure 2.4: Contours of the cost functional J in the (γ, β)-plane (solid curves) for R_0 equal to (a) 1.2, (b) 3, and (c) 10. A Poisson noise structure was assumed ($\xi = 1/2$), with $\sigma_0^2 = 1$ and $n = 50$ data points. Parameter values and initial conditions used were $\beta = R_0$, $\gamma = 1$, $N = 10,000$, $S_0 = 9900$, and $I_0 = 100$.

2.4 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.

2.4.1 Sensitivity

The sensitivities of a system provide temporal information on how states of the system (data is usually one or more state variables) respond to changes in the parameters [76]. 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 [8]. 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 2.5a). 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 just 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 [81], 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 [8] and, given that we found that it provided little additional insight in the current setting, we shall not discuss it further here.

2.4.2 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 fre-

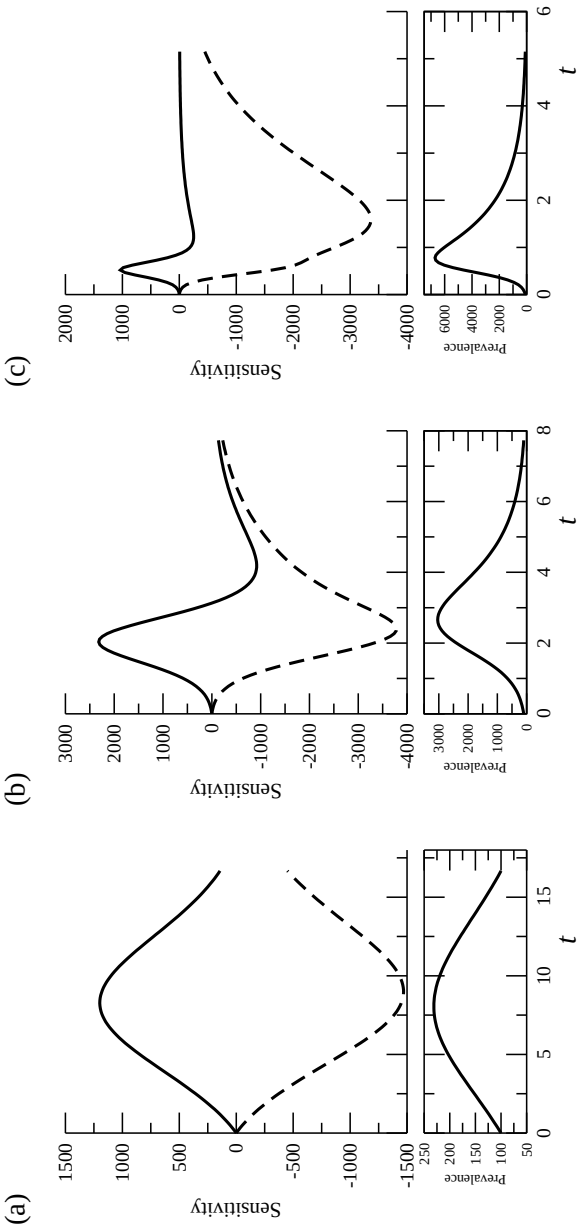

Figure 2.5: Sensitivities of $I(t)$ (*i.e.*, prevalence) with respect to the model parameters β (solid curves) and γ (dashed curves) are shown on the upper panels of the graphs for a) $R_0 = 1.2$, b) $R_0 = 3$ and c) $R_0 = 10$. The lower panel of each graph displays the corresponding prevalence-time 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 $\beta = R_0$.

quency at which data are sampled within a fixed observation window (*i.e.*, one 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 $t_n = t_{\text{end}}$. 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 2.6, in which prevalence data is used), in accordance with the asymptotic theory [80]. 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.

For real-time outbreak analysis, the amount of available data will increase over time as the epidemic unfolds [13]. Wearing *et al.* look at how estimates of R_0 varies as additional data points are obtained in [84], but there is no mention of uncertainty values in those estimates. However, 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. Previously, Cauchemez *et al.* have looked at confidence intervals of estimates of R_0 as an epidemic unfolds by using a Bayesian statistical framework in [20]. We will approach the problem using asymptotic statistical theory.

Using Equation (2.6) and the known values of the parameters, we calculated standard errors for parameter estimates based on the first n_{used} data points, where $p + 1 \leq n_{used} \leq n$. As seen in Figures 2.7a and 2.7b, when only one parameter is fitted, the standard error decreases rapidly at first, but its decrease slows significantly

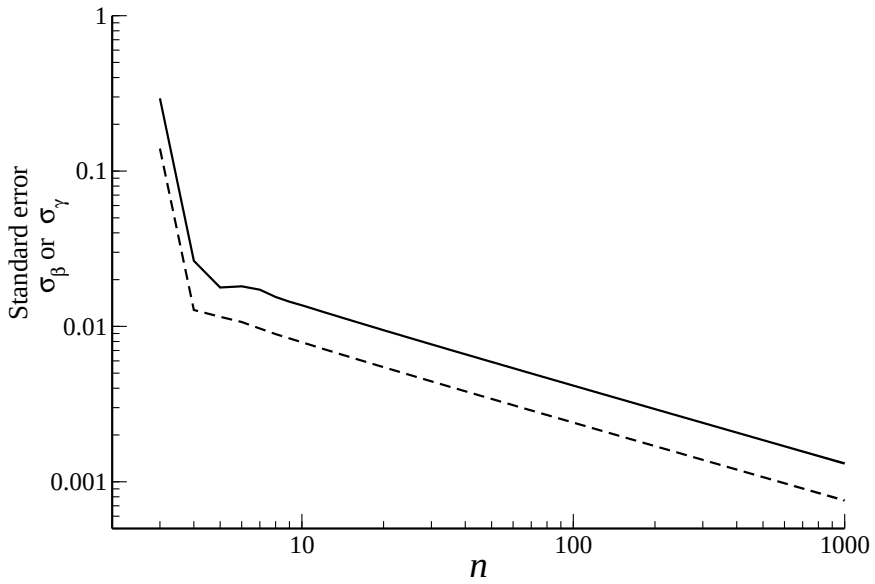

Figure 2.6: Standard errors of the estimates of β and γ as the number of observations, n, changes while maintaining a constant window of observation (fixed t_{end}). The points fall on a line of slope $-\frac{1}{2}$ on this log-log plot. The variance-covariance matrix Σ_0 was calculated exactly (*i.e.*, 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 $I_0 = 100$.

just before the peak of the epidemic. Once this point in time has been reached, subsequent data points provide significantly 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 2.8a. 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 information contained in the earliest portion of an outbreak is orders of magnitude higher than that contained in later portions.

Figure 2.8b 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 2.8a. 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 investigates 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. Fifty reduced data sets were created by removing, in turn, a single data point from the baseline data set. Standard errors were then computed using the true covariance matrix, Σ_0 for the reduced data set (Equation (2.6)). (For this experiment, use of the true covariance matrix allowed us to accurately observe these effects on standard errors that resulted from the removal of a single data point. Errors introduced by solving the inverse problem would make it impossible to ascertain trends.) 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 2.9 shows, when β is the only parameter fitted, the local maxima of the standard error curve occur at the same time as the local extrema of the sensitivity

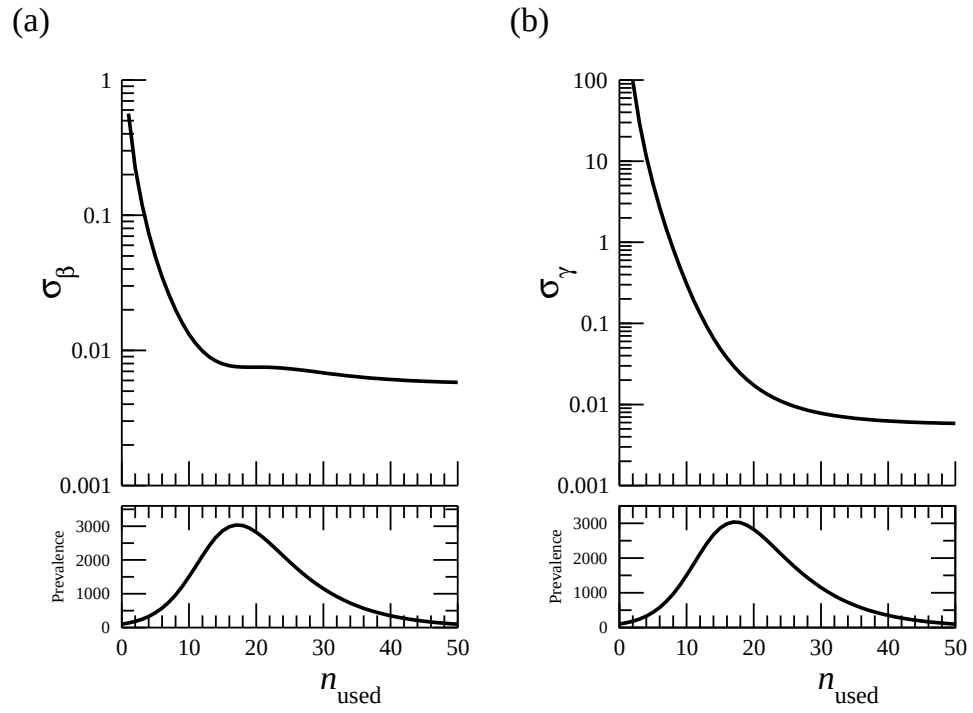

Figure 2.7: Impact of increasing the length of the observation window on standard errors of estimates of (a) β and (b) γ when each is estimated separately from prevalence data. The observation window is $[0, t_{n_{\text{used}}}]$, *i.e.*, estimation was carried out using n_{used} data points. Because 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.

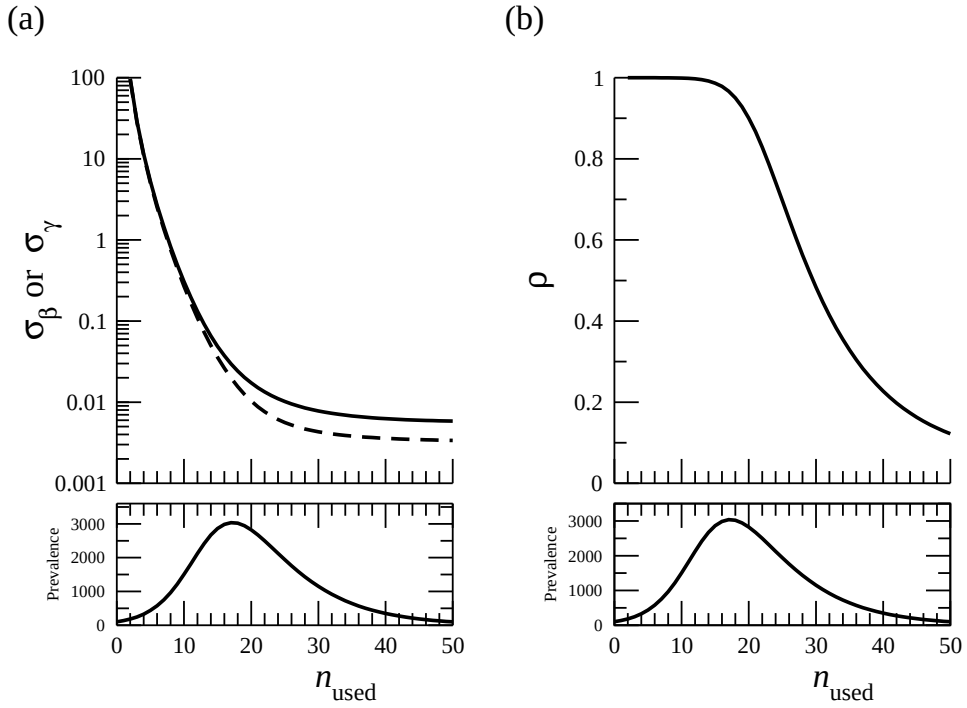

Figure 2.8: Illustrated in graph (a) is the impact of increasing the length of the observation window on standard errors of estimates of β (solid curve) and γ (dashed curve) when both are estimated simultaneously. Graph (b) displays the effect on the correlation coefficient between estimates of β and γ. The observation window consists of n_{used} data points in the time interval $[0, t_{n_{\text{used}}}]$. 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.

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. Figure 2.10 shows that the local maxima of the standard error curves no longer line up directly with the local extrema of the sensitivity curves. This is likely due 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 reduction method described here, based on the asymptotic theory, can identify when additional data should ideally be sampled.

2.4.3 Data Sampling with Incidence Data

Previously, we presented our three data sampling methods using prevalence data. In this section, we conduct the same experiments, but use incidence data.

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 $t_n = t_{\mathrm{end}}$. 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.

As before, adding additional data points in this way increases the accuracy of parameter estimates, with standard errors eventually decreasing as $n^{-1/2}$, in accordance with the asymptotic theory [80]. This is still the case for incidence data even when $\xi < 1$ where the signal to noise ratio decreases in n as can be seen in Figure 2.11.

The results of our second data sampling method, calculating standard errors for parameter estimates based on the first n_{used} data points, and third method, calculating standard errors for parameter estimates based on removing a single data point, when applied to incidence data can be seen in Figures 2.12-2.14.

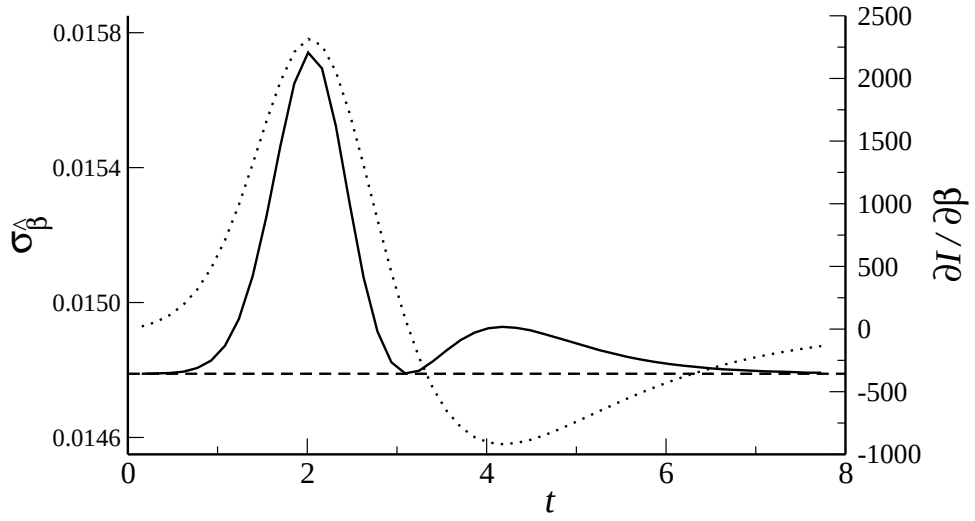

Figure 2.9: Standard errors for the estimation of β from prevalence data using the single point removal method as discussed in the text (solid curve) with the baseline standard error (without removing any points) also plotted (dashed curve). Standard errors were calculated using Equation (2.6) and each 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 $\sigma_0^2 = 10^2$, $S_0 = 9900$, $I_0 = 100$, $N = 10,000$, $\beta = 3$ and $\gamma = 1$. The additive noise structure, $\xi = 0$ was assumed.

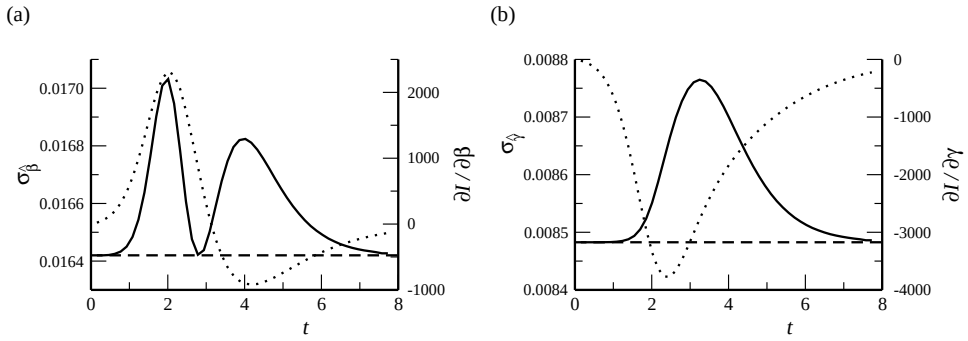

Figure 2.10: Standard errors for the simultaneous estimation of β and γ from prevalence data using the single point removal method as discussed in the text (solid curves). Standard errors were calculated using Equation (2.6), and each is plotted at the time t_i of the removed data point. Panel (a) shows the standard error for the estimate of β (solid curve), together with the baseline standard error (without removing any points) also plotted (dashed curve), and the sensitivity of $I(t)$ with respect to β (dashed curve). Panel (b) shows the standard error for the estimate of γ (solid curve), together the baseline standard error also plotted (dashed curve), and with the sensitivity of $I(t)$ with respect to γ (dashed curve). All parameter values and other details are as in the previous figure.

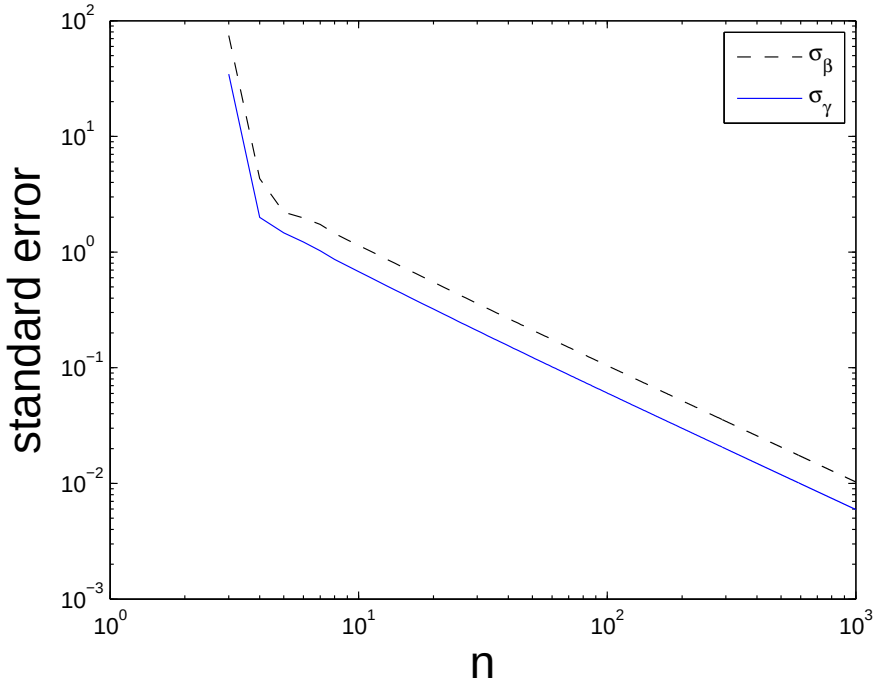

Figure 2.11: Standard errors of estimates of β and γ as the number of observations, n, changes while maintaining a constant window of observation (fixed t_{end}). The points fall on a line of slope $-\frac{1}{2}$ on this log-log plot. Standard errors are calculated using Equation (2.6), using the true values of the parameters. The variance-covariance matrix Σ_0 was calculated exactly (*i.e.*, no curve-fitting was carried out) with the disease incidence 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 $I_0 = 100$.

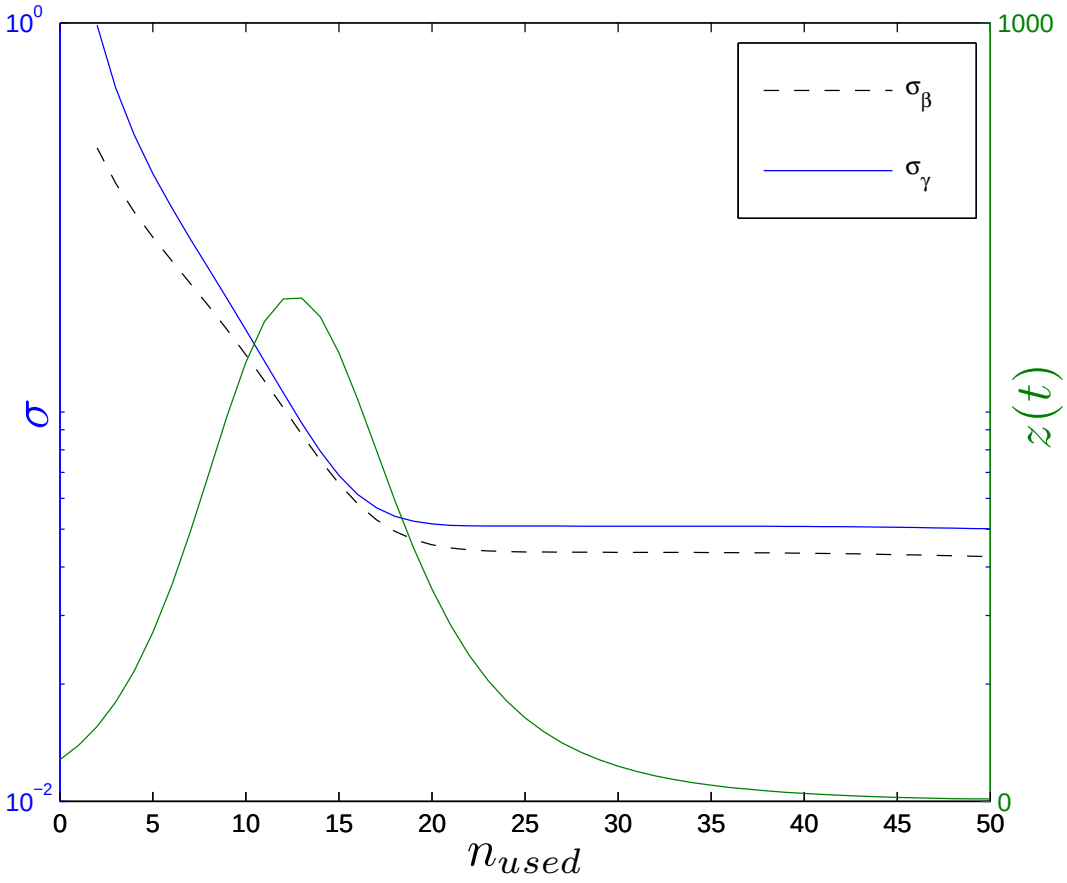

Figure 2.12: Impact of increasing the length of the observation window on standard errors of estimates of β (dashed curve) and γ (solid curve) when each is estimated separately from incidence data. The observation window is $[0, t_{n_{\text{used}}}]$, *i.e.*, estimation was carried out using n_{used} 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 incidence curve, $z(t)$, is superimposed. Standard errors are plotted on a logarithmic scale. The exact formula for Σ_0 was used, with parameter values $\sigma_0^2 = 1$, $S_0 = 9900$, $I_0 = 100$, $N = 10,000$, $\beta = 3$ and $\gamma = 1$. The Poisson noise structure, $\xi = 1/2$, was employed.

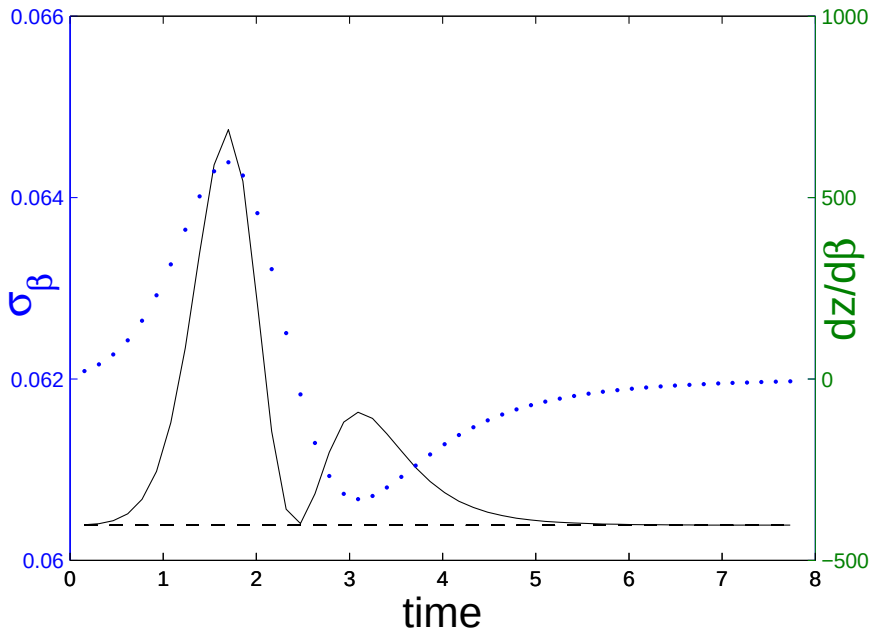

Figure 2.13: Standard errors for the estimation of β from incidence data using the single point removal method as discussed in the text (solid curve) with the baseline standard error (without removing any points) also plotted (dashed curve). Standard errors were calculated using Equation (2.6) and each is plotted at the time t_i corresponding to the removed data point. For comparison, the sensitivity of $z(t)$ with respect to β is also shown (dotted curve). Synthetic data was generated using the parameter values $\sigma_0^2 = 10^2$, $S_0 = 9900$, $I_0 = 100$, $N = 10,000$, $\beta = 3$ and $\gamma = 1$. The additive noise structure, $\xi = 0$ was assumed.

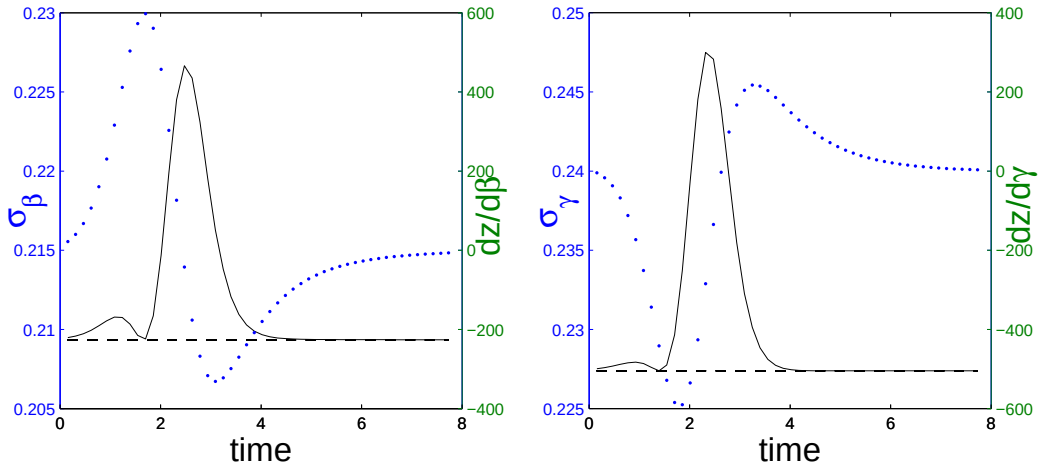

Figure 2.14: Standard errors for the simultaneous estimation of β and γ from incidence data using the single point removal method as discussed in the text (solid curves). Standard errors were calculated using Equation (2.6), and each is plotted at the time t_i of the removed data point. Panel (a) shows the standard error for the estimate of β (solid curve), together with the baseline standard error (without removing any points) also plotted (dashed curve), and the sensitivity of $z(t)$ with respect to β (dashed curve). Panel (b) shows the standard error for the estimate of γ (solid curve), together the baseline standard error also plotted (dashed curve), and with the sensitivity of $z(t)$ with respect to γ (dashed curve). All parameter values and other details are as in the previous figure.

2.5 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 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 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.

It should be noted that all experiments presented here were conducted with knowledge that the underlying model was correct. However, in scenarios with real data this assumption is not valid and results in a further layer of uncertainty. The effects on estimation when using the wrong model have been explored by authors such as Wearing and Lloyd [61, 84] but many questions still remain. We shall explore this issue further in a later chapter.

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 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 (we shall investigate problems arising from parameter identifiability more in the following chapter). 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.

Chapter 3

Parameter Identifiability and Subset Selection

The first section of this chapter is an expanded version of one section of the paper "Parameter Estimation and Uncertainty Quantification for an Epidemic Model," *which has been accepted by the journal* Mathematical Biosciences and Engineering (MBE) *for publication pending minor revisions. The second section of this chapter is portions of the paper* "A Sensitivity Matrix Based Methodology for Inverse Problem Formulation" *published in the* Journal of Inverse and Ill-Posed Problems *written together with Dr. Ariel Cintrón-Arias with senior authors Dr. H. T. Banks and Dr. Alun. L. Lloyd [28]. My contributions to the paper include development of the subset selection algorithm and the justification for the use of the condition number of the sensitivity matrix as a metric of parameter identifiability.*

Simultaneously estimating several parameters can bring into question the identifiability of those parameters. 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. Thus, there may not be uniqueness to the solution to the inverse problem.

When parameter identifiability is questionable, it may be prudent to perform subset selection. That is, fixing insensitive parameters at some nominal values and then estimating the remainder.

Section 3.1 of this chapter highlights specific situations with simple infectious disease models where parameter identifiability can be problematic and offers a quantitative means of measuring identifiability. In Section 3.2, we present an algorithm for performing subset selection.

3.1 Parameter Identifiability

In the previous chapter, it was shown that in the setting of R_0 approaching one, there can be considerable difficulty in independently estimating a pair of parameters. 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. In the past, this concept has been explored by a number of authors (for example, [15], [17], [28], [50]).

It has been shown by Evans *et al.* in [37] that the SIR model with demography is identifiable for all model parameters and initial conditions (which are treated as parameters). 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 separately. In this chapter, we use quantitative methods to assess ease of parameter identifiability in the context of subset selection.

It was stated in the previous chapter 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 [17], 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 $\Sigma^{(n)}$ (this is elaborated on in Section 3.2). 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 [66].

3.1.1 Application to Epidemic Scenario

Initially, we investigate the case where only β and γ are fitted. In this situation, we are able to find an expression for $\kappa(\Sigma)$ by taking the ratio of the eigenvalues of the 2×2 Σ matrix as follows

$$\kappa(\Sigma) = \frac{\sigma_{\hat{\beta}}^2 + \sigma_{\hat{\gamma}}^2 + \sqrt{\sigma_{\hat{\beta}}^4 + \sigma_{\hat{\gamma}}^4 - 2\sigma_{\hat{\beta}}^2\sigma_{\hat{\gamma}}^2 + 4\rho_{\hat{\beta},\hat{\gamma}}^2\sigma_{\hat{\beta}}^2\sigma_{\hat{\gamma}}^2}}{\sigma_{\hat{\beta}}^2 + \sigma_{\hat{\gamma}}^2 - \sqrt{\sigma_{\hat{\beta}}^4 + \sigma_{\hat{\gamma}}^4 - 2\sigma_{\hat{\beta}}^2\sigma_{\hat{\gamma}}^2 + 4\rho_{\hat{\beta},\hat{\gamma}}^2\sigma_{\hat{\beta}}^2\sigma_{\hat{\gamma}}^2}}. \tag{3.1}$$

For given standard errors, it is easy to show that as the correlation between estimates of β and γ approaches one, the condition number goes to infinity. However, it is not practical to construct a situation where the correlation changes while keeping standard errors of the parameters constant, thus we examine the condition number across multiple values of R_0 in Figure 3.1. As Figure 3.1 illustrates, it is more difficult to rely on estimates of β and γ when R_0 approaches one. This corroborates what we have previously seen for the correlation coefficient (see Figure 2.3a).

Numerical experiments show that when more parameters are fitted to the data, identifiability becomes a more serious issue. In such cases, while we can no longer give a simple expression for $\kappa(\Sigma_0)$ since it is a function of the parameters 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 3.1. As we increase the number of parameters fitted, the condition number can increase by multiple orders of magnitude.

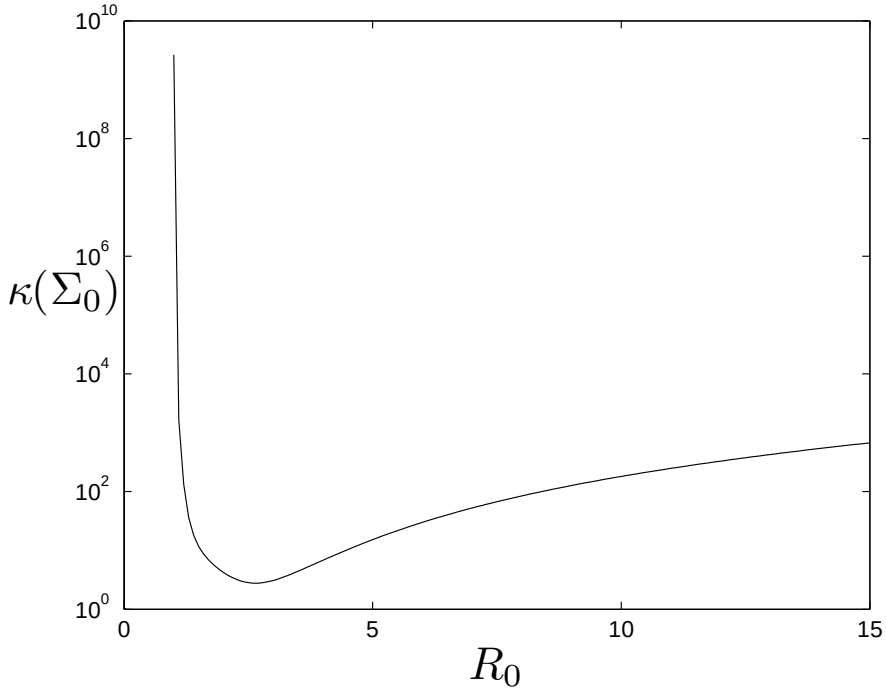

Figure 3.1: Dependence of the condition number of the 2×2 variance-covariance matrix (fitting β and γ) on the value of R_0. The condition number is displayed on a log scale. The variance-covariance matrix Σ_0 was calculated exactly (*i.e.*, no curve-fitting was carried out) under a Poisson noise structure, $\xi = 1/2$, 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 3.1: 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. $\xi = 1/2$, $N = 10,000$, $S_0 = 9900$, $I_0 = 100$, $\beta = 3$, $\gamma = 1$ and $\sigma_0^2 = 10^4$.

Parameters Fitted	σ_β	σ_γ	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}

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 between estimates of β and γ. 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. Knowing exactly which parameters should be removed from θ (and set to nominal values) to have an identifiable set is called subset selection, and is the focus of Section 3.2.

3.1.2 Application to Endemic Scenario

Up to this point, we have considered the informativeness of data from a single outbreak—*i.e.* an epidemic. As discussed in Subsection 1.2.3, many infections exhibit endemic behavior, which could take the form of a stable equilibrium, periodic oscillations, or more complex dynamics. We now ask the question of how much information about model parameters can be extracted from data describing such endemic settings. Many infections such as the seasonal flu, or measles outbreaks prior to national vaccination campaigns, are persistent, so it is likely that data from these outbreaks contain periodic oscillations. If this is the case, it would be very valuable to understand how much information these data contain about the transmission parameter, average duration of infection and other model parameters.

We will use the demographic SIR model (Equation 1.7) with seasonal forcing

(Equation 1.10), since it is a one of the simplest models that contain the dynamics we would like to study. We will assume the transmission phase shift t_0 is zero for simplicity, which merely corresponds to measuring time from some particular point in the year. Note that $a = 0$ corresponds to an endemic equilibrium rather than a limit cycle.

If we have the demographic SIR model at equilibrium, our only datum is value I^* from Equation 1.9. Although this value is a combination of all model parameters ($R_0 = \beta/\gamma$, γ, μ, N) and thus would contain information about each of them, we still have a dramatically underdetermined system. However, it is plausible to have a priori information about N, $1/\gamma$, and $1/\mu$. In such a case, obtaining R_0 from this single point is achieveable. If we have data about the approach to the equilibrium, such as from a system perturbation, we gain some more information about the model parameters. It is this fact that makes one wonder how much information can be gleaned from a system that does not sit at equilibrium, but follows a limit cycle.

To perform uncertainty quantification on estimates of model parameters as explained in Section 2.1, the sensitivity equations must be obtained whether analytically or numerically. Even though this system is non-autonomous, we are able to express the sensitivities of this SIR model analytically.

Writing the sensitivities of the state variables with respect to the model parameters as $\phi_1 = \partial S/\partial\beta_0$, $\phi_2 = \partial S/\partial\gamma$, $\phi_3 = \partial S/\partial\mu$, $\phi_4 = \partial S/\partial a$ $\phi_5 = \partial I/\partial\beta_0$, $\phi_6 = \partial I/\partial\gamma$,

$\phi_7 = \partial I/\partial\mu$, and $\phi_8 = \partial I/\partial a$, the following sensitivity equations are obtained

$$\frac{d\phi_1}{dt} = -\left(\mu + \frac{\beta(t)I}{N}\right)\phi_1 - \left(\frac{\beta(t)S}{N}\right)\phi_5 - \frac{SI}{N}(1 + a\sin 2\pi t) \tag{3.2}$$

$$\frac{d\phi_2}{dt} = -\left(\mu + \frac{\beta(t)I}{N}\right)\phi_2 - \left(\frac{\beta(t)S}{N}\right)\phi_6 \tag{3.3}$$

$$\frac{d\phi_3}{dt} = -\left(\mu + \frac{\beta(t)I}{N}\right)\phi_3 - \left(\frac{\beta(t)S}{N}\right)\phi_7 + N - S \tag{3.4}$$

$$\frac{d\phi_4}{dt} = -\left(\mu + \frac{\beta(t)I}{N}\right)\phi_4 - \left(\frac{\beta(t)S}{N}\right)\phi_8 - \frac{\beta_0 SI\sin 2\pi t}{N} \tag{3.5}$$

$$\frac{d\phi_5}{dt} = \frac{\beta(t)I}{N}\phi_1 + \left(\frac{\beta(t)S}{N} - (\mu+\gamma)\right)\phi_5 + \frac{SI}{N}(1 + a\sin 2\pi t) \tag{3.6}$$

$$\frac{d\phi_6}{dt} = \frac{\beta(t)I}{N}\phi_2 + \left(\frac{\beta(t)S}{N} - (\mu+\gamma)\right)\phi_6 - I \tag{3.7}$$

$$\frac{d\phi_7}{dt} = \frac{\beta(t)I}{N}\phi_3 + \left(\frac{\beta(t)S}{N} - (\mu+\gamma)\right)\phi_7 - I \tag{3.8}$$

$$\frac{d\phi_8}{dt} = \frac{\beta(t)I}{N}\phi_4 + \left(\frac{\beta(t)S}{N} - (\mu+\gamma)\right)\phi_8 + \frac{\beta_0 SI\sin 2\pi t}{N}, \tag{3.9}$$

with the initial conditions $\phi_i(0) = 0$ for $i = 1, ..., 8$.

We ran multiple forward simulations of the seasonal model at various values of the strength of seasonality to see how this parameter's value affects the estimation process. To ensure that we had arrived at the limit cycle, we initialized the system at the endemic equilibrium point (Equation 1.9) and then allowed the system to run for 900 years to run off its transient behavior before we began using the model output. We found that as the strength of seasonality parameter, a, is increased from 0 to a qualitatively large value, 0.5, the magnitude of the correlations between pairs of estimates the parameters β_0, γ, μ and a all remain very close to 1 (results not shown). That is, they are not easily uniquely identifiable. Although the correlation between estimates of all the parameters remains high, the coefficients of variation of all the model parameters are reduced as the strength of seasonality increases. Specifically, when examining the CV for the parameter combination β_0/γ (the baseline transmissibility), we see that it becomes substantially easier to estimate as the strength of seasonality increases (see Figure 3.2). So as with the simple SIR model from the

previous chapter when R_0 approached 1, we had that the constituent parts of R_0 were difficult to estimate, yet we still had a relatively easier time estimating R_0. A similar scenario is true in this instance, where we have a seasonal model at its limit cycle. Thus, for seasonal endemics at limit cycles, it can be difficult, if not impossible, to uniquely identify the basic model parameters but obtaining a useful estimate of β_0/γ is still possible.

3.2 Subset Selection Algorithm

In the last section, we highlighted situations where some parameters provide redundant information about the data. Redundant information can imply that some parameters may be impossible to determine uniquely—this is the issue of parameter identifiability. When one attempts to solve the inverse problem for a parameter set that is non-identifiable, one can obtain parameter estimates that are dramatically different, but still produce the same, or nearly the same, states. This can be problematic when conducting the inverse problem to determine parameters that have real interpretations, such as wanting to determine $1/\gamma$, the average duration of infection. We have given some ways to recognize when parameters are non-identifiable (correlation between two parameter estimates is near one in magnitude or the condition number of the covariance matrix is large). It would be beneficial, however, to formalize a quantitative approach to find, a priori, a subset of parameters that are identifiable, so that the inverse problem can be solved more easily. This section will present a subset selection algorithm, which was developed by the author together with Ariel Cintrón-Arias in [28]. The remainder of this chapter is in large part based on that work. To remain consistent with the original paper some introductory material is reiterated and some notation may differ slightly (we will highlight this when it occurs).

In particular, in this section we investigate the problem of finding multiple solutions for unknown parameters from observations with a statistical error structure (a more practical setting than one assuming noise free observations). We address parameter identifiability by exploiting properties of *both* the sensitivity matrix and

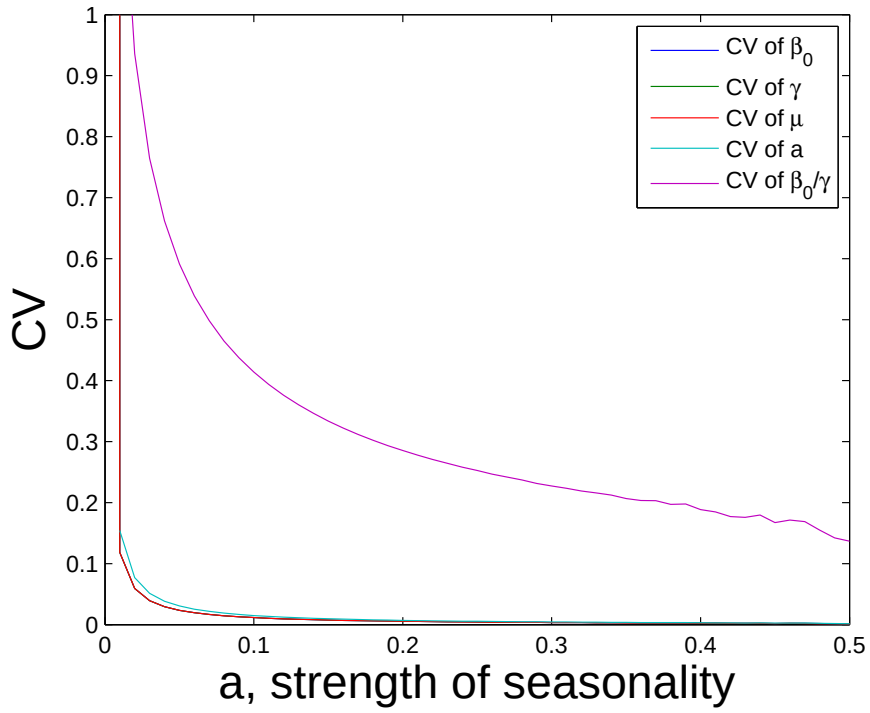

Figure 3.2: Plots of the coefficients of variation for the model parameters β_0, γ, μ, a and R_0 versus the strength of seasonality a when the system has reached its limit cycle. The highest curve is the coefficient of variation of R_0, which is significantly decreased as the strength of seasonality increases. Parameter values used were $\beta_0 = 15/14$, $\gamma = 1/14$ days, $\mu = 1/70$ years, and $N = 100,000$. The value of γ and β_0 are such that the model approximates measles (with an $R_0 \approx 15$), a highly-infectious disease with seasonal transmission. The exact formula for Σ_0 was used with $\sigma_0^2 = 1$. The absolute noise structure, $\xi = 0$, was employed.

uncertainty quantifications in the form of standard errors. We propose an algorithm inspired by [15, 17], to select parameter combinations (vectors) in two stages. In the first stage, all possible parameter combinations (i.e., subsets of all parameters) are considered and only those with a full rank sensitivity matrix are selected. In the second stage, a score involving uncertainty quantification (standard errors) is calculated for each parameter vector selected in the first stage. Then parameter subset combinations are examined in view of their score and the condition number of corresponding sensitivity matrices. We believe that some form of this type of *practical identifiability analysis* [5] could be carried out a priori, *i.e.*, before any attempt to solve inverse problems (from experimental observations) is made. We illustrate the ideas and methodology with a seasonal epidemic model.

The Fisher information matrix,

$$F = \chi^T W \chi, \tag{3.10}$$

whose inverse appears in Equation 2.22, has historically been the target of studies of identifiability (for example, [9, 81]). If $\chi^T(\theta_0)\chi(\theta_0)$ (henceforth in this chapter we will use the OLS approach which, sets $W = I$) is nearly singular, then $\hat{\theta}$ may be very sensitive to observation errors. Moreover, near-singularity (or ill-conditioning [44]) of F may also affect the approximation of the covariance matrix Σ, and consequently the calculation of standard errors for estimated parameters as we have seen in Section 3.1. Yet, while there are many experiments that can be done using the Fisher information matrix, our approach will rely on properties of the sensitivity matrix χ and the standard errors calculated using Equation 2.22.

We discuss the singular value decomposition approach and justify the use of the condition number of χ, rather than $\chi^T\chi$ as a diagnostic tool below.

To motivate the role singular value decomposition plays in uncertainty assessment, we consider another linearization that relates the estimator $\hat{\theta}$ to the singular values of the rectangular sensitivity matrix χ.

Suppose the model output $M(\theta)$ is well approximated by its linear Taylor expan-

sion around θ_0, i.e.,

$$M(\theta) \approx M(\theta_0) + \chi(\theta_0)(\theta - \theta_0). \tag{3.11}$$

This first order Taylor expansion can be used to reduce $Y - M(\theta)$ to an affine transformation of θ, by using Equations 3.11 and 2.1:

$$Y - M(\theta) = -\chi(\theta_0)(\theta - \theta_0) + \mathcal{E}, \tag{3.12}$$

where $\chi(\theta_0) \in \mathbb{R}^{n \times p}$, $\theta - \theta_0 \in \mathbb{R}^p$, $\mathcal{E}$ is an $\mathbb{R}^n$-valued random variable, and $n > p$.

The singular value decomposition (SVD) of the sensitivity matrix $\chi(\theta_0)$ is denoted as

$$\chi(\theta_0) = U \begin{bmatrix} \Lambda \\ \mathbf{0} \end{bmatrix} V^T, \tag{3.13}$$

where U is an $n \times n$ orthogonal matrix, Λ is a $p \times p$ diagonal matrix and V denotes an orthogonal $p \times p$ matrix. Specifically, since U is orthogonal, $U^T U = UU^T = I_n$, and U_1 contains the first p columns of U and U_2 contains the last $n - p$ columns, $U = [U_1\ U_2]$; we define $\Lambda = \text{diag}(s_1, \ldots, s_p)$, with $s_1 \geq s_2 \geq \cdots \geq s_p \geq 0$; $\mathbf{0}$ denotes an $(n - p) \times p$ matrix of zeros; and since V is orthogonal, $V^T V = VV^T = I_p$ (more details about SVD can be found in [44, 66] and references therein).

The Euclidean norm is invariant under orthogonal transformations. In other words, for any vector $w \in \mathbb{R}^n$ we have that $|w|^2 = w^T w = w^T I w = w^T UU^T w = |U^T w|^2$. According to [44, 70] this invariance of the Euclidean norm implies

$$\begin{aligned} |-\chi(\theta_0)(\theta - \theta_0) + \mathcal{E}|^2 &= \left|U^T \left(-\chi(\theta_0)(\theta - \theta_0) + \mathcal{E}\right)\right|^2 && (3.14) \\ &= \left| - \begin{bmatrix} \Lambda \\ \mathbf{0} \end{bmatrix} V^T(\theta - \theta_0) + \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \mathcal{E} \right|^2 && (3.15) \\ &= \left|-\Lambda V^T(\theta - \theta_0) + U_1^T \mathcal{E}\right|^2 + \left|U_2^T \mathcal{E}\right|^2. && (3.16) \end{aligned}$$

The estimator $\hat{\theta}_{OLS}$ minimizes $|Y - M(\theta)|^2$ and according to equations (3.12) and (3.16) can be calculated by solving $\left|-\Lambda V^T(\theta - \theta_0) + U_1^T \mathcal{E}\right|^2 = 0$, for θ and thus obtaining

$$\hat{\theta}_{OLS} = \theta_0 + V\Lambda^{-1} U_1^T \mathcal{E} = \theta_0 + \sum_{i=1}^{p} \frac{1}{s_i} v_i u_i^T \mathcal{E}, \tag{3.17}$$

where $v_i \in \mathbb{R}^p$ and $u_i \in \mathbb{R}^n$ denote the ith columns of V and U, respectively (the matrix V has column partitioning $V = [v_1, \ldots, v_p] \in \mathbb{R}^{p \times p}$, while $U = [u_1, \ldots, u_n] \in \mathbb{R}^{n \times n}$).

At this point we need a couple of definitions. The range of a matrix $C \in \mathbb{R}^{n \times p}$ with column partitioning $C = [c_1, \ldots, c_p]$ is defined as the subspace spanned by its columns, i.e.,

$$\mathcal{R}(C) = \left\{ \sum_{j=1}^{p} q_j c_j \in \mathbb{R}^n : q_j \in \mathbb{R} \right\}. \tag{3.18}$$

The rank of a matrix $C \in \mathbb{R}^{n \times p}$ is equal to the dimension of $\mathcal{R}(C)$:

$$\text{rank}(C) = \dim(\mathcal{R}(C)). \tag{3.19}$$

If $\text{rank}(C) < p$ (because we are assuming there are more observations than parameters, i.e., $n > p$) the matrix $C \in \mathbb{R}^{n \times p}$ is said to be rank deficient. On the other hand, if $\text{rank}(C) = p$ we say the matrix $C \in \mathbb{R}^{n \times p}$ has full (column) rank [44].

For a full rank sensitivity matrix $\chi(\theta_0) \in \mathbb{R}^{n \times p}$ (assuming $\text{rank}(\chi(\theta_0)) = p$ and $s_1 \geq s_2 \geq \cdots \geq s_p > 0$) its *condition number* κ is defined as the ratio of the largest to smallest singular value [44]:

$$\kappa(\chi(\theta_0)) = \frac{s_1}{s_p}. \tag{3.20}$$

We note that if the matrix $\chi(\theta_0)$ has full rank and a large condition number (a feature known as ill-conditioning [44]), then the Fisher information matrix $F = \chi(\theta_0)^T \chi(\theta_0)$ inherits a large condition number. Equation 3.13 implies the SVD of $\chi(\theta_0)^T \chi(\theta_0)$ is

$$\chi(\theta_0)^T \chi(\theta_0) = V \Lambda^2 V^T, \tag{3.21}$$

and therefore

$$\kappa(\chi(\theta_0)^T \chi(\theta_0)) = \frac{s_1^2}{s_p^2} = \left[\frac{s_1}{s_p} \right]^2 = \kappa(\chi(\theta_0))^2. \tag{3.22}$$

As discussed in [44], the columns of $\chi(\theta_0)$ are nearly dependent if and only if $\kappa(\chi(\theta_0))$ is large. In other words, if $\kappa(\chi(\theta_0))$ is not large (the matrix $\chi(\theta_0)$ is well-conditioned) then the columns of the sensitivity matrix are not nearly dependent,

suggesting one could use the condition number of $\chi(\theta_0)$ as a criterion to select parameter combinations.

Now we propose an algorithm for parameter selection which is based on the rank and condition number of the sensitivity matrix rather than the Fisher information matrix.

The identifiability analyses developed by Brun, *et al.*, [15], and Burth, *et al.*, [17], motivate the subset selection algorithm introduced in this section. Both of these approaches use submatrices of the Fisher information matrix in their selection procedures. Burth, *et al.*, implemented a reduced-order estimation by determining which parameter axes lie closest to the ill-conditioned directions of the Fisher information matrix, and then by fixing the associated parameter values at priori estimates throughout an iterative estimation process. The subset selection keeps the well-conditioned parameters (those that can be estimated with little uncertainty from given measurements) active in the optimization, subject to having the corresponding Fisher information submatrix with a small condition number. Brun, *et al.*, determine identifiability of parameter combinations using the eigenvalues of submatrices that result from excluding columns from of the Fisher information matrix. They quantify the near dependence of columns in the sensitivity submatrix using the smallest eigenvalue of the Fisher information submatrix.

We propose an algorithm that searches all possible parameter combinations and selects some of them, based on two main criteria: the full rank of the sensitivity matrix, and uncertainty quantification as embodied in asymptotic standard errors.

Our approach is numerical and we illustrate its use with the SEIRS model introduced in the next section. To carry out the algorithm we require prior knowledge of nominal variance and nominal parameter values.

Henceforth, we use the terms "parameter combination" and "parameter vector" interchangeably. Parameter vectors $\theta \in \mathbb{R}^p$ will be considered for different fixed values of p.

The set

$$\mathcal{S}_p = \{\theta = (\lambda_1, \lambda_2, \ldots, \lambda_p) \in \mathbb{R}^p |\ \lambda_k \in \mathcal{I},\ \lambda_k \neq \lambda_m \forall\ k, m = 1, \ldots, p\} \tag{3.23}$$

collects the parameter vectors explored by a combinatorial search.

We define the set

$$\Theta_p = \{\theta |\ \theta \in \mathcal{S}_p \subset \mathbb{R}^p,\ \text{rank}(\chi(\theta)) = p\}, \tag{3.24}$$

where $\chi(\theta)$ denotes the $n \times p$ sensitivity matrix, and its rank is defined by Equation 3.19. By construction, the elements of Θ_p are parameter vectors that give sensitivity matrices with independent columns.

An important step in the selection procedure involves the calculation of standard errors (uncertainty quantification) using the asymptotic theory described in Section 2.1. For every $\theta \in \Theta_p$, we define a vector of *coefficients of variation* $\nu(\theta) \in \mathbb{R}^p$ such that for each $i = 1, \ldots, p$,

$$\nu_i(\theta) = \frac{\sqrt{(\Sigma(\theta))_{ii}}}{\theta_i},$$

and

$$\Sigma(\theta) = \sigma_0^2 \left[\chi(\theta)^T \chi(\theta)\right]^{-1} \in \mathbb{R}^{p \times p}.$$

Next, define

$$\alpha(\theta) = |\nu(\theta)|.$$

We call $\alpha(\theta)$ the *parameter selection score*, and remark that $\alpha(\theta)$ near zero indicates lower uncertainty possibilities in the estimation while large values of $\alpha(\theta)$ suggest that one could expect to find wide uncertainty in at least some of the estimates.

In the optimization literature the term "feasible" usually denotes a vector satisfying inequality or equality constraints. Here we use this term in the context of identifiability: a feasible parameter vector denotes a combination that can be estimated from data with reasonable to little uncertainty. More precisely, we say a given $\theta \in \Theta_p$ is a *feasible parameter vector* if both $\alpha(\theta)$ and $\kappa(\chi(\theta))$ are relatively small.

We summarize the steps of the algorithm as follows:

- **Combinatorial Search** Let I be the set of all possible parameters λ_i. For a fixed p calculate the set

$$S_p = \{\theta = (\lambda_1, ..., \lambda_p) \in \mathbb{R}^p | \lambda_k \in I, \lambda_k \neq \lambda_m \forall\, k, m = 1, ..., p\}\,. \tag{3.25}$$

  The set S_p collects all the parameter vectors obtained from a combinatorial search.

- **Full Rank Test** Calculate the set of viable parameters Θ_p as

$$\Theta_p = \{\theta | \theta \in S_p \subset \mathbb{R}^p, rank(\chi(\theta)) = p\}\,. \tag{3.26}$$

- **Standard Error Test** For every $\theta \in \Theta_p$ calculate a vector of coefficients of variation $\nu(\theta) \in \mathbb{R}^{p \times p}$ by

$$\nu_i(\theta) = \frac{\sqrt{\Sigma(\theta)_{i,i}}}{\theta_i}, \tag{3.27}$$

  for $i = 1, ..., p$, and $\Sigma(\theta) = \sigma_0^2 \left[\chi^T(\theta)\chi\right]^{-1} \in \mathbb{R}^p$. Calculate the parameter selection score as $\alpha(\theta) = |\nu(\theta)|$.

We will now apply this algorithm in the next section.

3.2.1 Application to Endemic Scenario

We introduce a specific model, the standard *Susceptible-Exposed-Infective-Recovered-Susceptible* (SEIRS) model, to illustrate the algorithm from the previous section. In particular we consider a seasonal model for disease spread and progression in a population. Seasonal patterns of disease incidence are observed in epidemics of influenza [35], meningococcal meningitis [73], measles [3], and rubella [87], to mention a few. Many temporal factors play a role in the formation of cyclical patterns, for instance [45]: (i) survival of the pathogen outside the host, (ii) host behavior and (iii) host immune function.

Cyclical incidence patterns are often modeled with a transmission parameter being a function of time. We denote the time-dependent transmission parameter by $\beta(t)$;

it is traditionally defined by [35, 52]

$$\beta(t) = \beta_0 \left[1 + \beta_1 \cos(2\pi(t - t_0))\right], \tag{3.28}$$

where β_0 is called the baseline level of transmission, β_1 is known as the amplitude of seasonal variation or simply the strength of seasonality, and t_0 denotes the transmission parameter phase shift. (This is similar to the transmission function described by Equation 1.10, only that we previously employed the sine curve and $\beta_1 = a$. We shall remain consistent here with the notation of the original publication [28].) We may, for convenience, derive an equivalent formulation. Because

$$\beta_1 \cos(2\pi(t - t_0)) = a_1 \cos(2\pi t) + b_1 \sin(2\pi t),$$

where $a_1 = \beta_1 \cos(2\pi t_0)$ and $b_1 = \beta_1 \sin(2\pi t_0)$, we may re-write Equation 3.28 as

$$\beta(t) = \beta_0 \left(1 + a_1 \cos(2\pi t) + b_1 \sin(2\pi t)\right). \tag{3.29}$$

The time-dependent transmission parameter $\beta(t)$, as defined in Equation 3.29, is used in the seasonal epidemic model introduced here.

Four main epidemiological events are described: latent infection, active infection, recovery, and loss of immunity. It is assumed that individuals becoming infected undergo latency, a period of time during which they are incapable of effectively transmitting the infectious agent, before progressing into active infection. People recover from active infection and develop temporary immunity (they will eventually become susceptible once again). Four epidemiological classes are considered, and at time t the number of: susceptible is denoted by $S(t)$; latent or exposed is denoted by $E(t)$; infectious is denoted by $I(t)$; and recovered or temporarily immune is denoted by

$R(t)$. The nonlinear differential equations [59, 79]

$$\begin{aligned}
\frac{dS}{dt} &= \frac{1}{P}N + \frac{1}{L}R(t) - \beta(t)S(t)\frac{I(t)}{N} - \frac{1}{P}S(t) && (3.30)\\
\frac{dE}{dt} &= \beta(t)S(t)\frac{I(t)}{N} - \frac{1}{M}E(t) - \frac{1}{P}E(t) && (3.31)\\
\frac{dI}{dt} &= \frac{1}{M}E(t) - \frac{1}{D}I(t) - \frac{1}{P}I(t) && (3.32)\\
\frac{dR}{dt} &= \frac{1}{D}I(t) - \frac{1}{L}R(t) - \frac{1}{P}R(t) && (3.33)\\
N &= S(t) + E(t) + I(t) + R(t) && (3.34)\\
S(t_0) &= S_0 && (3.35)\\
E(t_0) &= E_0 && (3.36)\\
I(t_0) &= I_0 && (3.37)\\
R(t_0) &= N - S_0 - E_0 - I_0, && (3.38)
\end{aligned}$$

define the epidemic dynamics known as an SEIRS model. This formulation takes into account demographic processes (the birth rate is N/P and the average life span is P) while assuming the total population size N remains constant.

The mean latency period is denoted by M, while the average length of active infection is denoted by D. It is also assumed immunity lasts an average of L units of time.

(Notice that $1/P = \mu$, $1/M = \nu$ and $1/D = \gamma$ from our earlier formulation of the SEIR model and SIR model with demography in Chapter 1.)

As in the previous section, we consider a scenario where the initial conditions of the SEIRS model (S_0, E_0, and I_0) may be unknown, and may need to be estimated, along with all the other model parameters. We apply inverse problem methodologies to determine estimates of the vector parameter

$$\theta = (S_0, E_0, I_0, N, L, D, M, P, \beta_0, a_1, b_1)^T \in \mathbb{R}^p = \mathbb{R}^{11}, \tag{3.39}$$

according to an ordinary least squares criterion.

The subset selection algorithm is illustrated first by solving inverse problems from synthetic observations. We construct a synthetic data in a similar way to the process

Table 3.2: Nominal parameter values for the SEIRS model.

Parameter	**Nominal Value**	**Units**
S_0	2.78×10^5	people
E_0	1.08×10^{-1}	people
I_0	1.89×10^{-1}	people
N	1×10^6	people
L	5	years
D	9.59×10^{-3}	years
M	5.48×10^{-3}	years
P	75	years
β_0	375	years^{-1}
a_1	2×10^{-2}	1
b_1	-2×10^{-2}	1

in Section 2.2. We suppose a nominal parameter vector and a nominal error variance are equal to θ_0 (true parameter vector) and σ_0^2 (true variance), respectively. Random noise is then added to the model output (incidence, in this case) as follows:

$$Y_i = z(t_i; \theta_0) + \epsilon_i, \tag{3.40}$$

where ϵ_i is a normal random variable, i.e., $\epsilon_i \sim \mathcal{N}(0, \sigma_0^2)$. A realization y_i of the observation process Y_i, is calculated by drawing independent samples e_i from the normal distribution so that

$$y_i = z(t_i; \theta_0) + e_i \quad \text{for } i = 1, \ldots, n.$$

We assume the observation error variance is $\sigma_0^2 = 500$, and assume the nominal parameter values for the SEIRS model that are presented in Table 3.2.

For the purposes of this example, we slightly modify the algorithm presented in the previous section as follows. When $p = 11$ the parameter combination

$$\theta = (S_0, E_0, I_0, N, L, D, M, P, \beta_0, a_1, b_1) \in \mathbb{R}^{11}, \tag{3.41}$$

with the nominal parameter values given in the next section, produces a rank deficient sensitivity matrix $\chi(\theta)$ for the SEIRS model. For $p = 3$ the only parameter

combination considered here is that of the transmission parameters, i.e.,

$$\theta = (\beta_0, a_1, b_1) \in \mathbb{R}^3. \tag{3.42}$$

Other parameter vectors for fixed values of $p = 4, \ldots, 10$ are considered in the following way. For each fixed $j = 1, \ldots, 7$, and therefore fixed $p = 3 + j$, we explore parameter vectors of the form

$$\theta = (\lambda_1, \lambda_2, \ldots, \lambda_j, \beta_0, a_1, b_1) \in \mathbb{R}^p, \tag{3.43}$$

where for $k = 1, \ldots, j$,

$$\lambda_k \in \{S_0, E_0, I_0, N, L, D, M, P\} = \mathcal{I},$$

such that no entries of θ in Equation 3.43 are repeated.

To illustrate the algorithm we consider several values of p, while using the MATLAB (The Mathworks, Inc.) routine `rank` (this routine computes the number of singular values that are greater than "machine tolerance").

Results for $p = 5$ (using the nominal parameter values) are displayed in Figure 3.3 (on logarithmic scales), where $\alpha(\theta)$ is depicted as a function of $\kappa(\chi(\theta))$ for all $\theta \in \Theta_5$. The pairs in the lower-left corner of Figure 3.3 correspond to feasible parameter vectors, because $\alpha(\theta)$ and $\kappa(\chi(\theta))$ are here relatively small.

The subset selection algorithm was applied for $p = 4, \ldots, 10$, while using the nominal variance and parameter values. We find that there is not a single parameter combination with $p = 10$ that has a full rank sensitivity matrix. For $p = 9$, only three parameter vectors pass the full rank test, and none of which can be considered feasible. We summarize the feasible parameter vectors in Table 3.2.1 for $p = 4, \ldots, 8$, where each feasible $\theta \in \Theta_p$ is displayed along with $\kappa(\chi(\theta))$ and $\alpha(\theta)$. The cutoffs used to select the parameter combinations in Table 3.2.1 were somewhat arbitrary but relative to the smallest values computed for the two criteria (condition number and selection score) in each example.

The OLS inverse problems were solved by implementing a subspace trust region method (based on an interior-reflective Newton method [70]). We used the MATLAB

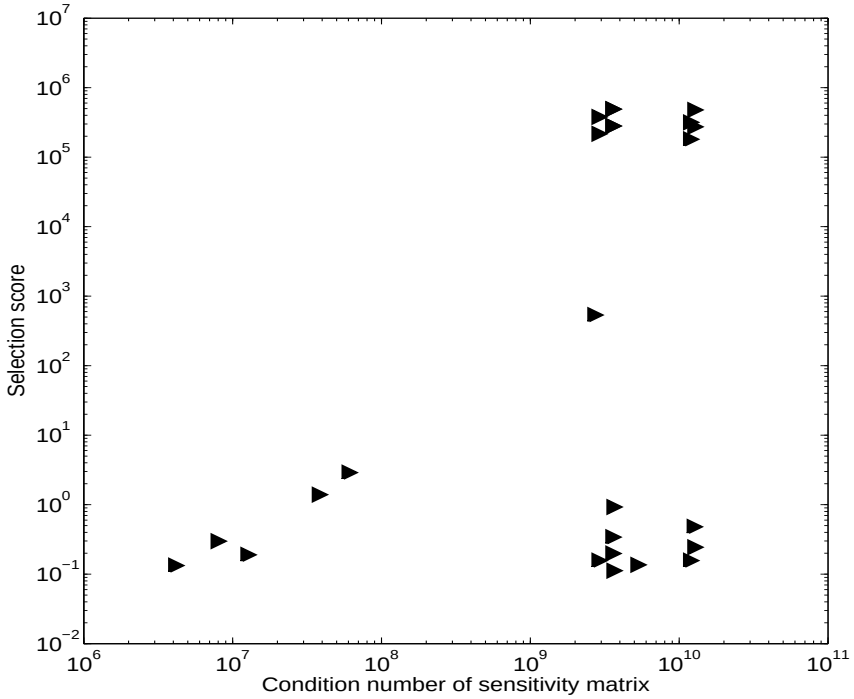

Figure 3.3: Parameter selection score $\alpha(\theta)$ versus the condition number $\kappa(\chi(\theta))$ of the $n \times p$ sensitivity matrix, for all parameter vectors $\theta \in \Theta_p$ with $p = 5$. Logarithmic scales are used on both axes. Nominal parameter values used are listed in Table 3.2.

Table 3.3: Feasible parameter vectors obtained while applying the subset selection algorithm for $p = 4, \ldots, 8$, using nominal values as listed earlier in the text. For each selected parameter vector $\theta \in \Theta_p$ the condition number of the sensitivity matrix $\kappa(\chi(\theta))$, and the selection score $\alpha(\theta)$ are displayed.

Parameter vector θ	Condition number $\kappa(\chi(\theta))$	Selection score $\alpha(\theta)$
(L, β_0, a_1, b_1)	2.047×10^{5}	5.019×10^{-2}
(M, β_0, a_1, b_1)	1.420×10^{5}	6.386×10^{-2}
(P, β_0, a_1, b_1)	3.176×10^{5}	7.044×10^{-2}
$(L, D, \beta_0, a_1, b_1)$	4.034×10^{6}	1.332×10^{-1}
$(D, M, \beta_0, a_1, b_1)$	1.233×10^{7}	1.897×10^{-1}
$(D, P, \beta_0, a_1, b_1)$	7.781×10^{6}	2.987×10^{-1}
$(N, L, D, \beta_0, a_1, b_1)$	1.829×10^{10}	1.670×10^{-1}
$(S_0, N, D, \beta_0, a_1, b_1)$	1.454×10^{10}	2.026×10^{-1}
$(S_0, L, D, \beta_0, a_1, b_1)$	1.828×10^{10}	2.375×10^{-1}
$(S_0, D, M, \beta_0, a_1, b_1)$	2.152×10^{10}	3.301×10^{-1}
$(S_0, D, P, \beta_0, a_1, b_1)$	1.828×10^{10}	4.832×10^{-1}
$(N, D, M, \beta_0, a_1, b_1)$	2.166×10^{10}	5.739×10^{-1}
$(N, D, P, \beta_0, a_1, b_1)$	1.829×10^{10}	9.658×10^{-1}
$(N, L, D, M, \beta_0, a_1, b_1)$	2.166×10^{10}	5.960×10^{0}
$(S_0, L, D, M, \beta_0, a_1, b_1)$	2.167×10^{10}	5.970×10^{0}
$(N, D, M, P, \beta_0, a_1, b_1)$	2.166×10^{10}	1.153×10^{1}
$(S_0, D, M, P, \beta_0, a_1, b_1)$	2.167×10^{10}	1.159×10^{1}
$(S_0, N, L, D, M, \beta_0, a_1, b_1)$	6.333×10^{12}	5.044×10^{1}
$(S_0, N, D, M, P, \beta_0, a_1, b_1)$	6.561×10^{12}	2.950×10^{2}

(The Mathworks, Inc.) routine `lsqnonlin`. For the purposes of this demonstration we initialized every optimization routine at the nominal parameter vector θ_0. This prevents the optimization routine from stopping in a local minimum that is not the global minimum, should multiple local minima exist.

The nominal error variance and nominal parameter values are those given above. The parameter vectors estimated from synthetic data are those appearing on top of each subtable in Table 3.2.1, for each value of p, where parameter combinations are sorted in ascending order of their selection score (from top to bottom). In other words, all the parameter vectors estimated from synthetic observations have reasonable condition numbers and relatively small selection scores. Five inverse problems (for $p = 8, 7, 6, 5, 4$) were solved from the same realization of the observation process, to estimate the parameter vectors

$$\theta = (S_0, N, L, D, M, \beta_0, a_1, b_1),$$
$$\theta = (N, L, D, M, \beta_0, a_1, b_1),$$
$$\theta = (N, L, D, \beta_0, a_1, b_1),$$
$$\theta = (L, D, \beta_0, a_1, b_1),$$
$$\theta = (L, \beta_0, a_1, b_1).$$

Results of these numerical experiments are summarized in Table 3.2.1.

We analyze the results using the coefficient of variation. For instance in Table 3.2.1, when $\theta = (S_0, N, L, D, M, \beta_0, a_1, b_1)$ it is seen for D that the standard error is nearly one third of the estimate, suggesting lower uncertainty. For the other parameters S_0, N, L, M, β_0, a_1, and b_1 the standard error can be nearly four times (and up to eleven times) the estimate (for b_1 its SE is $|4 \times \text{Est}|$, because $b_1 < 0$). These values indicate substantial uncertainty. Figure 3.4(a) displays the residual plot (see [7] for a discussion of the effective use of residual plots) for this parameter combination: $y_j - z(t_j; \theta_{OLS})$ versus time t_j, where $j = 1, \ldots, n$. The temporal pattern in the residuals together with large standard errors suggest that estimation of this parameter combination from observations (with a statistical error structure) would be meaningless.

Table 3.4: Results of solving five inverse problems from a single synthetic data set generated as described in the text using nominal values listed earlier. For each parameter combination we display the estimate (Est.), the standard error (SE) and the coefficient of variation (standard error divided by the estimate, CV = SE/Est.). For notational convenience we use here the notation e to denote exponentiation to the base 10; i.e., $2.8e5$ denotes 2.8×10^5, etc.

Parameter vector $\theta = (S_0, N, L, D, M, \beta_0, a_1, b_1)$								
	S_0	N	L	D	M	β_0	a_1	b_1
Est.	2.8e5	1.0e6	5.0e0	9.6e-3	5.5e-3	3.7e2	2.0e-2	-2.0e-2
SE	1.5e6	5.0e6	4.5e1	3.1e-3	6.2e-2	3.4e3	7.7e-2	8.4e-2
CV	5.5e0	5.0e0	9.1e0	3.2e-1	1.1e1	9.0e0	3.8e0	-4.2e0
Parameter vector $\theta = (N, L, D, M, \beta_0, a_1, b_1)$								
Est.		1.0e6	5.0e0	9.6e-3	5.5e-3	3.7e2	2.0e-2	-2.0e-2
SE		2.7e4	2.7e0	2.5e-3	2.2e-2	5.9e2	3.1e-2	2.5e-2
CV		2.7e-2	5.4e-1	2.6e-1	4.1e0	1.6e0	1.6e0	-1.3e0
Parameter vector $\theta = (N, L, D, \beta_0, a_1, b_1)$								
Est.		1.0e6	5.0e0	9.6e-3		3.8e2	2.0e-2	-2.0e-2
SE		2.7e4	1.7e-1	5.8e-4		1.5e1	1.3e-3	1.2e-3
CV		2.7e-2	3.4e-2	6.1e-2		3.9e-2	6.3e-2	-6.1e-2
Parameter vector $\theta = (L, D, \beta_0, a_1, b_1)$								
Est.			5.0e0	9.6e-3		3.8e2	2.0e-2	-2.0e-2
SE			7.4e-2	5.8e-4		9.8e0	1.2e-3	1.2e-3
CV			1.5e-2	6.1e-2		2.6e-2	6.2e-2	-6.0e-2
Parameter vector $\theta = (L, \beta_0, a_1, b_1)$								
Est.			5.0e0			3.8e2	2.0e-2	-2.0e-2
SE			1.4e-2			2.6e0	2.0e-4	7.9e-4
CV			2.7e-3			6.8e-3	9.9e-3	-4.0e-2

The residual plots for all the other parameter combinations in Table 3.2.1 do not have temporal structure. For the sake of illustration we display in Figure 3.4(b) the residuals versus time for $\theta = (L, D, \beta_0, a_1, b_1)$.

Improvements in uncertainty quantification are observed with the removal of some key parameters. It is not just reducing the number p of parameters, but rather which parameters are to be estimated that matters. The near dependence in the columns of the sensitivity matrix χ reflects correlations between parameter estimates which make a parameter combination unsuitable for estimation. This can be seen rather obviously by the magnitude and temporal dependence of the residuals in Figure 3.4(a). However, consider the removal of S_0 from the estimation, and compare $\theta = (S_0, N, L, D, M, \beta_0, a_1, b_1)$ with $\theta = (N, L, D, M, \beta_0, a_1, b_1)$ in Table 3.2.1. The standard error for N is seen to drop from 500% to approximately 3% of the estimate. Another substantial improvement when dropping S_0 is obtained for L, for which its standard error reduces from being nine times the estimate to one half of its value. Lower uncertainty improvements are also obtained for the parameters M, β_0, a_1, and b_1.

The next numerical experiment considered here is the removal of S_0 and M. We compare the results for $\theta = (S_0, N, L, D, M, \beta_0, a_1, b_1)$ with those for $\theta = (N, L, D, \beta_0, a_1, b_1)$, in Table 3.2.1. There are uncertainty improvements for all parameters. The least (but still substantial) improvement is for D, where its standard error drops from being nearly 30% to being just 6% of the estimate. For the parameters N, L, β_0, a_1, and b_1 an improvement of two orders of magnitude is seen. Improvements in uncertainty are more pronounced after removing S_0, N, and M: for this we compare $\theta = (S_0, N, L, D, M, \beta_0, a_1, b_1)$ and $\theta = (L, D, \beta_0, a_1, b_1)$ in Table 3.2.1.

Undoubtedly, the best case scenario of uncertainty quantification we obtained is that of estimating $\theta = (L, \beta_0, a_1, b_1)$ from the same synthetic data set. In Table 3.2.1, it is seen that the standard errors reduce to less than 1% of the estimates for L, β_0, and a_1, and to 4% from nearly 400% of the estimate for b_1.

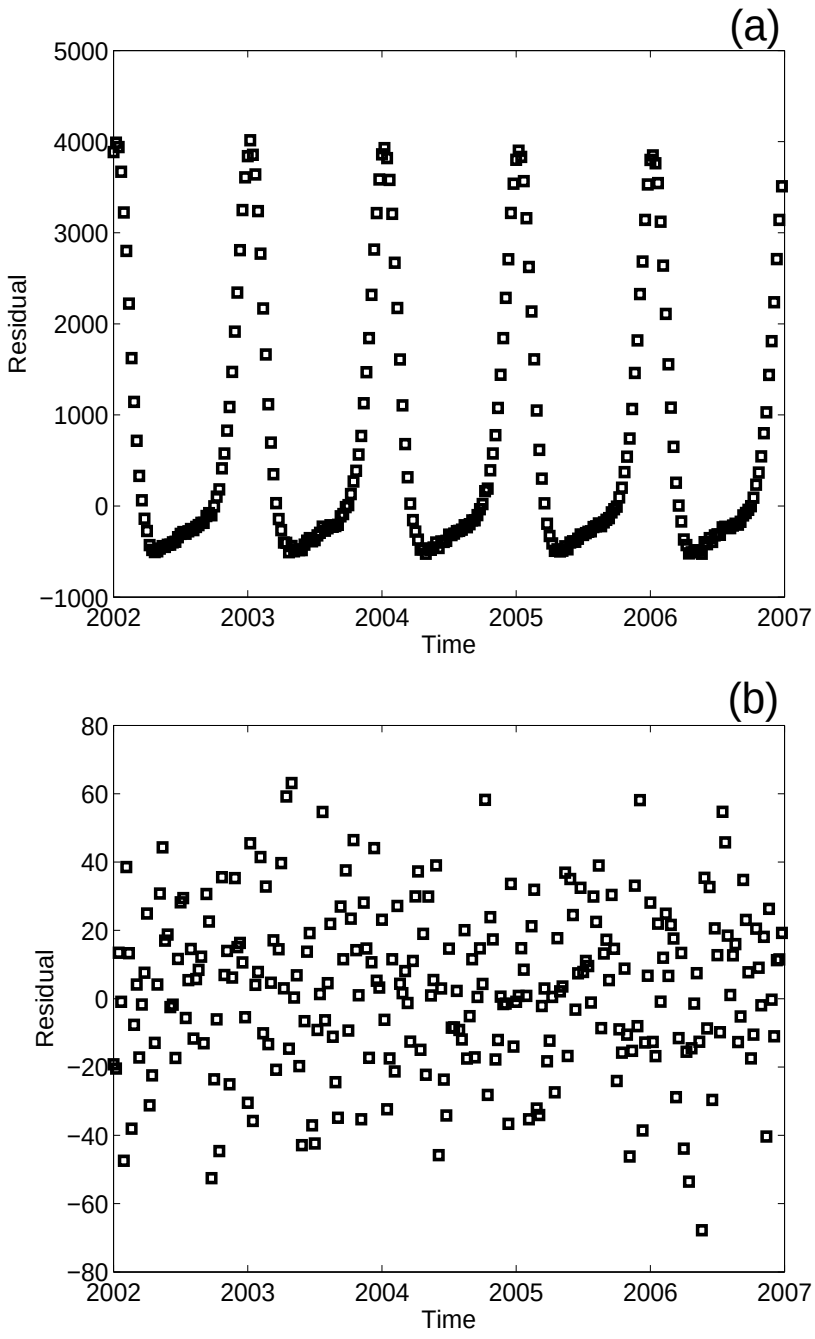

Figure 3.4: Residual plots: $y_i - z(t_i; \theta_{OLS})$, versus time, t_i, for $i = 1, \ldots, n$. Graph (a) displays residuals obtained for $\theta = (S_0, N, L, D, M, \beta_0, a_1, b_1)$, while Graph (b) depicts residuals for $\theta = (L, D, \beta_0, a_1, b_1)$.

3.3 Discussion

We have discussed a computational methodology for inverse problem formulation in the context of parameter identifiability. Using an OLS scheme based on a constant variance statistical model for the observation process and a seasonal SEIRS epidemics model for illustration, we have proposed a prior-analysis algorithm that we believe might profitably precede efforts on parameter estimation from data. The algorithm can be used if reasonable ranges for the sought after parameters are either known a priori, or can be assumed by the user much in the same way one must assume reasonable ranges in inverse problem formulations and initiation of algorithms for the resulting estimation procedures.

The subset selection [67] algorithm we gave is based on two main criteria for a fixed number of parameters: (i) full rank of the sensitivity matrix; and (ii) calculation of standard errors. We proposed to first select according to the sensitivity matrix rank, because those parameter combinations for which χ has full rank will have a non-singular Fisher information matrix $\chi^T\chi$, and its inverse is used in the calculation of the standard errors (see Equation 2.22).

The near dependence of the sensitivity matrix columns can be a fingerprint of parameter correlations—a pertinent feature for subset selection [67]. In chapter 2, we determined identifiability of parameters in a simple SIR model, and showed how correlation between parameter estimates can impede the estimation of parameters and parameter combinations, such as the basic reproductive number. Moreover, Brun, *et al.*, [15] explain that if the columns of χ are nearly dependent, then changes in the model output due to small changes in a single parameter can be compensated by appropriate changes in other parameters.

We have presented illustrations of the how the removal of nearly dependent columns of the sensitivity matrix can provide substantial improvements in uncertainty quantification. This feature involves more than just reducing the number p of parameters, it relates to excluding certain key parameters. For instance, if we assume a linear Taylor expansion of the model output, the estimator $\hat{\theta}_{OLS} \in \mathbb{R}^p$ is given by Equation

3.17, where the sensitivity matrix $\chi(\theta_0)$ has singular values $s_1 \geq \cdots \geq s_{p-1} \geq s_p > 0$. If $s_p \approx 0$ and $s_{p-1} > 1$, then submatrices with singular values $s_2 \geq \cdots \geq s_p > 0$, and $s_1 \geq \cdots \geq s_{p-1}$, have different conditioning when quantifying the sensitivity of *reduced order* estimations that only involve $p-1$ parameters. The condition number of the former submatrix is s_2/s_p, which is large if $s_p \approx 0$, while for the latter submatrix the condition number satisfies $1 \leq s_1/s_{p-1} < s_1$, because $s_{p-1} > 1$.

In our numerical experiments, we calculate sensitivity matrices $\chi(\theta)$ evaluated at different realizations of the estimator $\theta = \hat{\theta}_{OLS}$. When $\theta = (S_0, N, L, D, M, \beta_0, a_1, b_1)$ the singular values of the sensitivity matrix range from 4.7×10^6 to 4.6×10^{-6} while for $\theta = (L, \beta_0, a_1, b_1)$ the singular values of $\chi(\hat{\theta}_{OLS})$ range from 1.9×10^6 to 9.3×10^0. The smallest singular value changes from 4.6×10^{-6} to 9.3×10^0 while the largest remain on the order of 10^6. This improvement in conditioning is reflected in the standard error for L, β_0, and a_1, which reduces to less than 1% of the estimate, from nearly 900% and 380% (see Table 3.2.1).

Although in this section we only discuss OLS, the selection algorithm can be easily applied when using a generalized least squares scheme as seen in Section 2.1. We also carried out numerical experiments (for brevity not discussed here) involving use of synthetic nonconstant variance data sets in GLS formulations, and obtained results absolutely consistent with those of the OLS formulation presented here.

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 inference and Markov Chain Monte Carlo (*e.g.* [60] and [21]) are two other commonly used methods to solve the inverse problem. Yet, since identifiability is a feature of the mathematical model and not the statistical model nor the fitting process, switching estimation techniques does not remove the problem of parameter identifiability, so it remains an important concern when solving the inverse problem in any respect.

Chapter 4

Model Selection

In the previous chapters we discussed the pitfalls associated with estimating model parameters. All previous discussions, however, hinged on the assumption that the underlying model was correct. This was the primary reason for the use of synthetic data for the experiments done in Chapters 2 and 3. Yet, with real data, the assumption that the model chosen to fit to the data is, in fact, the true underlying process has the potential to be wrong. Depending on the question being asked, such structural uncertainty has the potential to outweigh the uncertainty due to noise [61, 71, 84].

Rather than choosing a single model to describe a process in question, one can come up with a set of potential models and then use an iterative modeling process where one uses information gained from the inverse problem to tell how good (in some sense) each model is. There are many measures used throughout the theory of statistics to do so. The more useful measures are ones that reward goodness of fit while also penalizing the number of parameters used in a model. The latter is necessary since as the number of parameters increases, the easier it is to fit a model to data. (As John von Neumann is attributed to saying, "with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.") Two of the more common such measures are the Bayes information criterion (BIC) [65] and the Akaike information criterion (AIC) [1].

The Bayes information criterion is calculated by

$$BIC = k \ln n - 2 \ln L, \tag{4.1}$$

where k is the number of estimated parameters, n is the number of data points and L is the maximum likelihood for the model. Similarly, the Akaike information criterion is calculated by

$$AIC = 2k - 2 \ln L. \tag{4.2}$$

The number of estimated parameters, k, for the type of problem we consider is typically $p+1$, as one estimates σ_0^2 from Equation 2.21 in addition to the p free parameters in the model (often called the structural parameters).

Recall that when using the ordinary least squares theory, our statistical model of the observations Y_i is given by the mathematical model $M(t;\theta)$ plus some noise ϵ_i. Under our previous noise assumptions, the ϵ_i are assumed to be independent, identically distributed random variables with zero mean and (finite) constant variance σ_0^2. Here we shall add the assumption that the ϵ_i are normally distributed, which then makes the MLE equivalent to the OLS estimate of the structural parameters. While the estimates of the structural parameters are the same under both optimization schemes with the aforementioned assumptions, one should note that the least squares estimate for σ_0^2 differs from the maximum likelihood estimate (which equals J/n), though this difference is trivial for large sample sizes. To stay consistent with information theory, which is based on likelihood theory, we shall use the MLE for σ^2 in such a context and shall denote it σ_{MLE}^2. We continue to use the unbiased least squares estimate σ^2 in the same contexts used in previous chapters.

To calculate these selection scores, we need the maximum of the likelihood function of θ, which is given according to [16]

$$L = \left(\frac{1}{\sqrt{2\pi}\sigma_{MLE}} \right)^n \exp\left(\frac{-n}{2} \right) \tag{4.3}$$

$$\ln L = \frac{-n}{2} \ln \sigma_{MLE}^2 - \frac{n}{2} \ln 2\pi - \frac{n}{2}. \tag{4.4}$$

There also exists a version of the AIC for small samples. It is a second-order (from the Taylor expansion) information criterion, is named the AIC_c [49], and is

Table 4.1: Here is a hypothetical example of model selection using AIC. Suppose we have four models fit to the same data set, M_i, with respective information criteria AIC_i. Using the Δ_i values for comparison we see that, model M_2, having the minimum AIC value is the best model; model M_3 is second best, and is still strongly plausible; model M_3 is considerably less plausible and model M_4 has essentially no empirical support.

Model	AIC	Δ_i
M_1	180	6
M_2	174	0
M_3	175	1
M_4	185	11

calculated by

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1} = 2k - 2\ln L + \frac{2k(k+1)}{n-k-1}. \tag{4.5}$$

Burnham and Anderson [16] suggest using AIC_c when the ratio $n/k < 40$. We will use AIC_c when performing our model selection, as we have a small sample size. This data set is discussed in the following section.

When multiple models fit to the same data are compared, the model with the lowest (ordinally, not in magnitude) value of selection score (*e.g.* BIC, AIC or AIC_c) is the one that best explains the data while using the least number of free parameters. The actual value of the selection score is unimportant. It is the relative differences that describe the plausibility of each model. We will calculate the AIC differences using

$$\Delta_i = AIC_i - AIC_{\min}. \tag{4.6}$$

Burnham and Anderson [16] suggest the following rules of thumb for determining model plausibility (which is "particularly useful for nested models"): For $0 \leq \Delta_i \leq 2$ there is a substantial level of empirical support, $4 \leq \Delta_i \leq 7$ has considerably less support while $\Delta_i > 10$ has essentially no empirical support. We present a brief example of model selection in Table 4.1.

Table 4.2: Influenza epidemic data from a boys boarding school, which was garnered via the DataThief program from the figure in the 1978 paper [6]. The numbers given are those students who were "confined to bed." $N = 763$.

Day	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Confined	3	8	28	75	222	291	256	236	191	125	70	28	12	5

4.1 The Boarding School Data

Throughout this chapter, we will examine a case study of data obtained from an outbreak of influenza in an English boys boarding school [6] and see how well a number of different epidemic models fit these data. The data given (see Table 4.2 and Figure 4.1) is the number of students who were confined to bed over the course of the epidemic, which lasted a fortnight. The total population was 763 boys. In addition to the 763 boys, the author mentions that there were 130 adults with limited contact with the boys and only one adult became ill, so the adults were omitted from the data. The boys resided in eleven different dormitory houses.

Beyond being confined to bed, the treatment the boys received was minimal. Ten of the 512 ill boys received antibiotics. 630 of the 763 boys had received an influenza vaccine about four months earlier, but it was for strains different than that which caused this outbreak.

The boys reportedly spent between three and seven days away from class, but this does not necessarily tell us the time spent in the infectious or confined to bed class, as the author also reports that symptoms quickly ended once the boys were confined to bed and that boys were typically allowed out of bed 36 hours after symptoms subsided. If this information did apply directly to information about average duration of infection or other model parameters, it would be possible to incorporate such information into the fitting process by using a Bayesian method.

This is a famous data set and is well-worn, having been used in the past by a number of authors (including but not limited to [68, 69, 75, 84]). The main reason why this data has seen so much use is because of the conditions under which it was

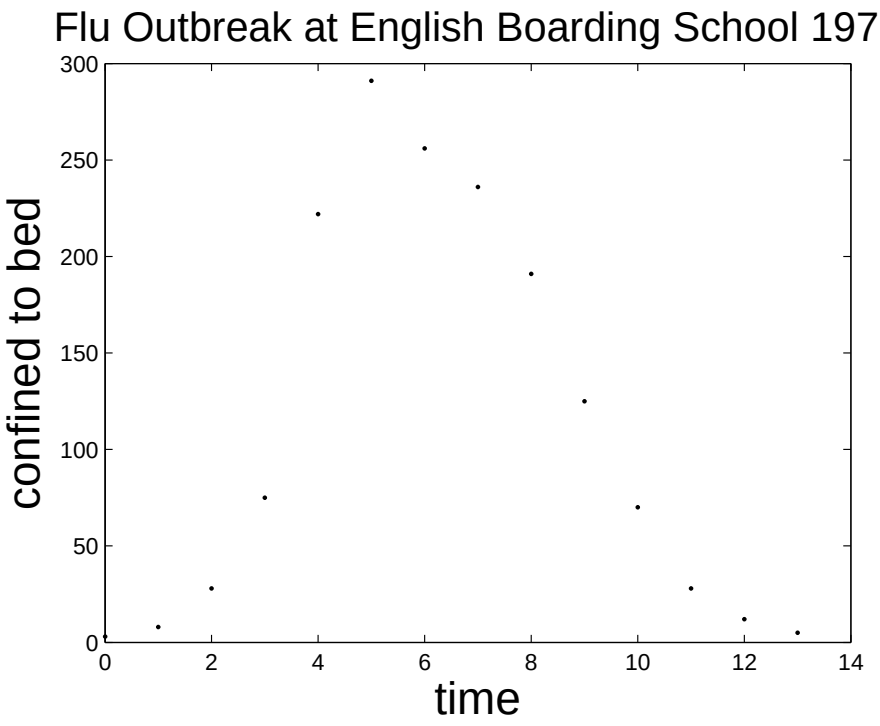

Figure 4.1: A plot of the data of an influenza outbreak in a boys boarding school [6] also presented in Table 4.2. The horizontal axis measures time in days while the vertical axis measures the number of students confined to bed on that day.

collected. This outbreak was in a closed population, a scenario that is typically very unlikely, and helpful when most compartmental models assume a closed population. Also the boys were closely monitored, which substantially reduces, if not removes, uncertainty in the data with regards to under reporting of cases of infection. This data is probably much better than is typically obtained from outbreaks across cities, states or countries. While better than average, this data is not without difficulties in interpretation.

The fact that the data is students who were "confined to bed" is somewhat problematic as its interpretation of which state variable of a model is rather vague. This is clearly not incidence data, as the sum of the values exceeds N. This could be prevalence data, however the exact nature of their confinement is unclear. If this was a true quarantine, the confined students could not infect the susceptibles remaining in the population, which would imply that the data does not represent infectious individuals (at least not infectious individuals in a well mixed population). Additionally, the symptomatic individuals in the population are not necessarily the infectious individuals, which complicates the issues of using this data. We shall try several models and see which model fits best with a hope that it will shed some light on what interpretation is likely.

The author also reports numbers of convalescent students, which perhaps represents the number of individuals in a recovered class. As with infectious individuals not being biologically identical to symptomatic individuals, convalescent individuals are not necessarily the same as recovered, so there remains a level of uncertainty with this class and fitting to this class as well as the confined to bed individuals increases the difficulty of an already difficult model selection problem. Finally, epidemic data is usually fitted to incidence or prevalence data rather than convalescent individuals, so we will not investigate this portion of the data further.

Another concern about this data set is that it is only 14 data points, which is a relatively small amount. With such few data points, the complexity of the model to be fit to the data is significantly hindered (as we want the ratio of p/n to be as close to 0 as possible to avoid the effects of bias). With this in mind, we will restrict

ourselves to models containing $p \leq 5$ structural parameters.

4.1.1 Previous Work

Previous analyses of this data were performed by a number of authors who attempt to estimate R_0. Murray [69] finds the equivalent of $R_0 = 3.77$ for the SIR model. Wearing *et al.* [84] also fits an SIR and a variant on the SEIR model. They posit that their SEIR model fits significantly better than the SIR model, and have obtained an estimate of R_0 of 35.9 and 3.74 respectively. Ross [75] uses density dependent Markov population processes (rather than OLS) to fit variants of the SIR and SEIR models and receive estimates of R_0 of 4.38 and 16.9 respectively. It is also possible to estimate R_0 from the total outbreak size [54] (*i.e.* number of people that ever get infected), $y_{\text{total}} = 1 - \frac{S(\infty)}{N}$, by using the formula

$$\exp\left(-R_0 y_{\text{total}}\right) = 1 - y_{\text{total}}. \tag{4.7}$$

Mills *et al.* point out in the supplement to [68] that an R_0 estimate of much higher than 3 is unlikely when one considers the fact that only 67% of the population was eventually infected, which results in $R_0 = 1.66$ from Equation 4.7. Thus, Wearing's estimate of 35.9 seems extremely high; the authors attribute this to the possibilities for stronger mixing in the small population size in the boarding school compared to the level of mixing in cities or countries, from which most R_0 estimates have been obtained. So, there remains controversy over not just the interpretation of data (though all four authors assume the data represents prevalence), but also what the value of R_0 is for the epidemic. We attempt to resolve these questions by using the AIC_c model selection technique on a series of models.

4.2 Fitting Compartmental Models to the Data

4.2.1 The SEICR Model and Sub-models

We would like to gain an understanding of what underlying process governs the dynamics of the boarding school influenza outbreak. To do so, we begin by comparing a set of ordinary differential equation models. The most general model we employ in this subsection is a Susceptible-Exposed-Infective-Confined-Recovered, an SEICR, model. The exposed class contains all individuals who have contracted the infection, but are not yet infectious, while the confined class is everyone that has been confined to bed, which we assume is a true quarantine meaning they are no longer infectious. The S, I and R classes remain as before from SIR model (see Chapter 1). The infection leads to permanent immunity and we have a closed population size, N, and ignore the effects of births and deaths (this is acceptable because of the short time scale of the outbreak).

We will fit the SEICR model and its sub-models, the SEIR, SICR and SIR model to the data. Whenever the C class is present (when we are dealing with the SICR or SEICR models), we assume that the data refers to that class. In the absense of the C class (when we are dealing with the SEIR or SIR models), we assume the data refers to the I class.

The parameters are as follows: β is the transmission parameter, $1/\nu$ is the average duration of latency, $1/\gamma$ is the average duration of infectiousness, and $1/\omega$ is the average duration of confinement. This yields an R_0 of β/γ for the all four models.

The SEICR model can be described by the following differential equations

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{4.8}$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \nu E \tag{4.9}$$

$$\frac{dI}{dt} = \nu E - \gamma I \tag{4.10}$$

$$\frac{dC}{dt} = \gamma I - \omega C, \tag{4.11}$$

together with the initial conditions $S(0) = S_0$, $E(0) = E_0$, $I(0) = I_0$ and $C(0) = C_0$.

For completeness, we also present the SICR model,

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{4.12}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I \tag{4.13}$$

$$\frac{dC}{dt} = \gamma I - \omega C, \tag{4.14}$$

with its initial conditions $S(0) = S_0$, $I(0) = I_0$ and $C(0) = C_0$; the SIR and SEIR models appear in Sections 1.2.1 and 1.2.2 respectively.

It should be noted that in the SIR model, the transition from the infectious class to the recovered class is not necessarily the process of recovery, but could also simply be considered "removal from infectiousness". This could imply that the SIR model, where the R class is fit to data could be a possibility for some situations where the data is the cumulative number of people put into quarantine. However, such a situation then does not allow for removal from this removed class, whether by recovery from the disease or ceasing quarantine. This is why we will use the SICR model here, to allow for loss from the confined class.

To fit the models to the data, we use the MATLAB Nelder Mead minimization tool `fminsearch` on the typical OLS cost function (see Equation 2.3 where $w_i = 1$). We report the results of the model fits in Table 4.3 and Figure 4.2. As our methodology in Section 2.1 suggests, we also display the residual plots for the fit of the SIR model in Figure 4.3 to verify that the zero mean and constant variance assumptions about the noise are valid.

From the results, it can be seen that the SIR model is the best from this set of models as it has the lowest AIC_c value. However, both the SEIR and SICR models have higher AIC_c, but the difference from the minimum is less than one. The model selection process does not make much of a distinction between these two model. In such a case, it is important to use the biological knowledge of the system to make decisions on model selection. The fact that a latent period is known to biologically exist makes us seriously consider choosing the SEIR model. Interestingly though, the SICR model has an even lower value of R_0 than the "optimistically" low value from

Table 4.3: A list of the results from four different models fit using OLS to the boarding school data. The listed values are the number of free structural parameters p for the model, the estimate of R_0, the cost functional value J and the model's AIC_c. The lowest AIC_c gives the model that fits the data the best while using the least number of free parameters, which is the SIR model.

Model	p	R_0	J	AIC_c
SIR	2	3.7627	3960.8	127.1624
SEIR	3	11.6263	3157.9	128.0351
SICR	3	3.0745	3147.5	127.9891
SEICR	4	10.5652	2816.8	131.4903

SIR, bringing it more in line with the estimate of R_0 from Equation 4.7. When we increase the model dimension one more to the SEICR model, we notice that the AIC_c value increases by more than 4, making it a less likely choice of a model.

Just as the the addition of the exposed class imposes a latency period causing the epidemic to be slower to start, adding the confined class after the infectious class causes the infectious class to peak earlier, but the individuals are removed from infectiousness much more quickly.

4.2.2 Time-Dependent Parameters

By including, and fitting to, a confined class, the epidemic is clearly right-skewed in terms of infectious individuals (see Figure 4.4). Perhaps this phenomenon could be better explained by a model that has different transmission parameters for the early portion and late portion of the epidemic. Such a model also has the ability to account for changes in behavior in the population, such as a decreased amount of interaction between individuals or increased preventative hygiene activities once the population is aware of the presence of the disease.

Alternatively, we could create a model that has different recovery parameters for the early and late portions of the outbreak. This model could account for, in a model without a confined class, treatment being applied. In a model with a confined class,

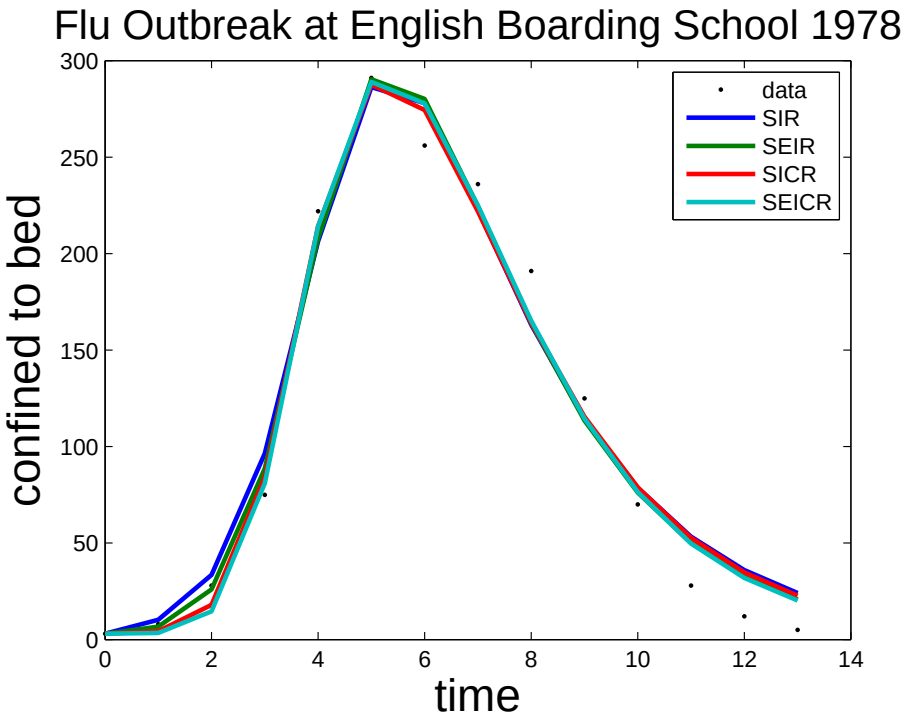

Figure 4.2: Plot of the number of students confined to bed from the OLS fit of the four models (solid curves) against time together with the boarding school data (dots). The number of students "confined to bed" corresponds to the I class for the SIR and SEIR models and the C class for the SICR and SEICR models. Estimates of R_0 can be found in Table 4.3. $N = 763$ and the initial conditions are $I_0 = 3$ for the SIR and SEIR models and $I_0 = 1$ for the SICR and SEICR models, $C_0 = 3$ and $S_0 = N - C_0 - I_0$.

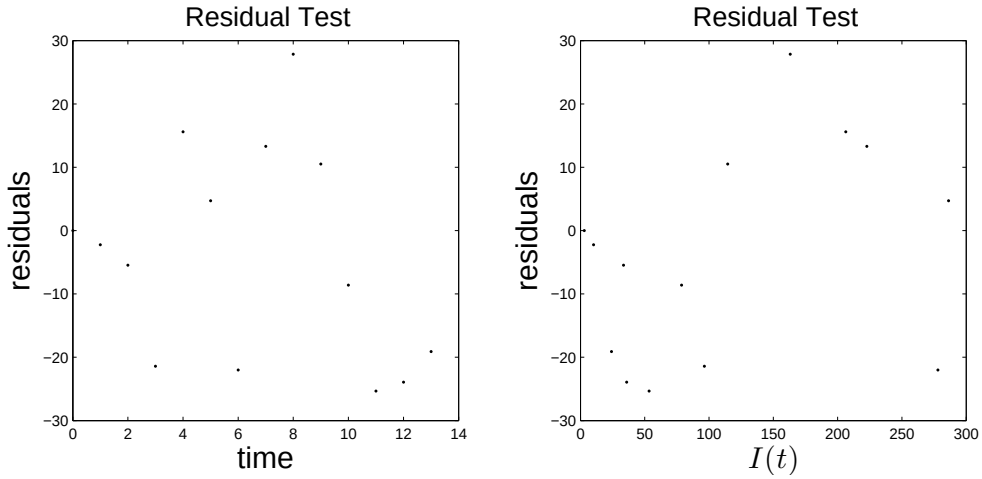

Figure 4.3: The residuals against time and against the model values of prevalence from the OLS fit of the SIR model to the boarding school data. Notice in both plots that the values are centered around 0 and the variation from the mean appears to not follow a particular pattern dependent on t or $I(t)$ which does not provide any evidence that the assumptions about the noise are not upheld. $N = 763$ and the initial conditions are $S_0 = 760$, $I_0 = 3$.

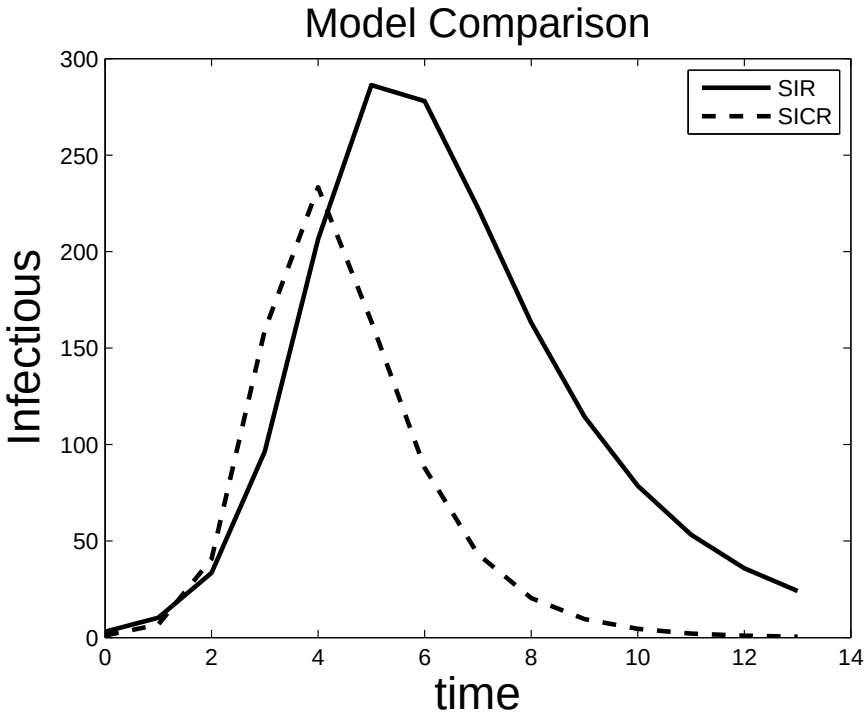

Figure 4.4: Plots of the infectious class of the SIR and SICR models against time using the best fit parameters to the boarding school data. Notice that the SICR model has prevalence right-skewed in comparison to the SIR model. For SIR, $\beta = 1.6924$, $\gamma = 0.4498$, $N = 763$, $S_0 = 760$ and $I_0 = 3$. For SICR, $\beta = 2.8519$, $\gamma = 0.9276$, $\omega = 0.4461$, $N = 763$, $S_0 = 759$, $I_0 = 1$ and $C_0 = 3$

the rate leaving the infectious class is not recovery, but is the rate of confining a sick student so they do not continue being infectious; this rate would likely change as attitudes and/or quarantine policies about the disease change. Finally, we can also construct models with varying latency parameters and durations of confinement.

To explore this first scenario, we fit SIR, SEIR and SICR models with a step-function for the transmission parameter to the boarding school data. The transmission parameter will be written as

$$\beta(t) = \begin{cases} \beta_1 & : t < t_{\text{switch}} \\ \beta_2 & : t \geq t_{\text{switch}} \end{cases} .$$

Thus, this adds two additional structural parameters. This SIR model has four free parameters: β_1, β_2, γ and t_{switch}; the SEIR and SICR have five (adding ν and ω respectively). The $SEIR$ and $SICR$ models with step-function transmission are not very plausible models as they have AIC_c values of 136.0361 and 139.9037, respectively. However, with an AIC_c score of 128.1192, the step-function transmission SIR model is a very plausible choice of model, though not necessarily the best. Though, when compared to the SEICR model which is the only other model with the same number of structural parameters fitted thusfar, the piece-wise transmission model is empirically much superior. Interestingly, the best fit parameter values are not ones that we would expect. We find that β_1 is less than β_2, which means people are more infectious later in the outbreak while we might have expected more people would be aware of the disease and would likely be taking precautionary measures. Also notice that the best fit value of $t_{\text{switch}} = 6.3423$ is just after the peak of the epidemic, meaning the two β values correspond to the increasing and decreasing portions of the epidemic. Perhaps the model is simply trying to find the two rates for the (approximately) exponential growth and exponential decay portions of the curve.

We shall now attempt to fit SIR, SEIR and SICR models with a constant transmission parameter β, but with a step-function for the per capita recovery rate to the boarding school data. The recovery parameter is defined by

$$\gamma(t) = \begin{cases} \gamma_1 & : t < t_{\text{switch}} \\ \gamma_2 & : t \geq t_{\text{switch}} \end{cases}.$$

The results of fitting these models can be seen in Table 4.4. The SIR model with a step-function $\gamma(t)$ has the lowest AIC_c value of any presented model thusfar, so it is the best (presented) model to describe the data. The best fit two γ values were $\gamma_1 = 0.4351$ and $\gamma_2 = 0.9637$, which translates to an average duration of infectiousness of about 2.3 days for the first period and about 1 day for the second period. Since t_{switch} was found to be 9.3651, which is rather late in the outbreak, it is unlikely that this threshold time could be interpreted as people noticing that there is an outbreak. Perhaps the reduction in average duration of being confined to bed could be attributed to the nurses knowing that the epidemic has waned, so they are likely to let students

Table 4.4: A list of the results from six different models with time-dependent parameters fit using OLS to the boarding school data. The listed values are the number of free structural parameters p for the model, the cost functional value J and the model's AIC_c. The lowest AIC_c gives the model that fits the data the best while using the least number of free parameters, which is the SIR model with piece-wise $\gamma(t)$.

Model	p	J	AIC_c
SIR (piece-wise β)	4	2214.0	128.1192
SEIR (piece-wise β)	5	2449.8	136.0361
SICR (piece-wise β)	5	3229.3	139.9037
SIR (piece-wise γ)	4	1974.5	126.5165
SEIR (piece-wise γ)	5	1647.2	130.4788
SICR (piece-wise γ)	5	1444.6	128.6420

free earlier. It should be noted, though, that the duration of confinement was said to be 3 to 7 days in most cases, which are far longer than the model predicts.

Finally, we fit an SEIR model with a similarly defined step-function for ν and an SICR model with a step-function for ω. Again, we find a new best model, the new SICR model with an AIC_c of 123.1824 and an $R_0 = 4.2547$. The new SEIR model has an AIC_c of 124.9049 ($R_0 = 7.8621$), which is still lower than the previous models presented. The fit of this SICR model can be seen in Figure 4.5.

4.3 Distributed Delays

4.3.1 Generally Distributed Delays

It is unlikely that an individual's duration of infection is exponentially distributed [78]. Since R_0 can be taken as the product of the strength of transmission times the average duration of infection, to accurate estimate the basic reproductive number, we require knowledge of the distribution of recovery time. Thus, we shall discard the assumption of a constant rate of recovery. A more general distribution of recovery

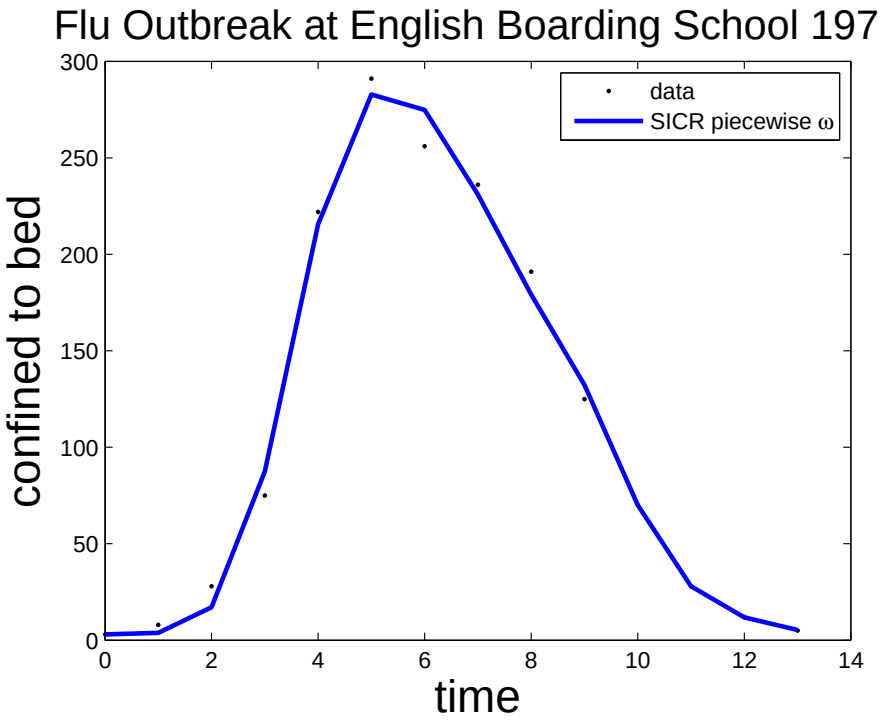

Figure 4.5: Plot of the number of students confined to bed from the OLS fit of the SICR model with an $\omega(t)$ step-function (solid curve) against time together with the boarding school data (dots). The best fit parameters are $\beta = 2.7614$, $\gamma = 0.6490$, $\omega_1 = 0.4360$, $\omega_2 = 1.0875$, and $t_{\text{switch}} = 9.3362$. The model selection score is $AIC_c = 123.1824$ and the cost functional value is $J = 978.12$. Initial conditions are the same as previous SICR model fits.

times is desired. There are many ways to approach distributed delays of the latency and recovery processes in an SEIR model.

If we take $f_E(\tau_E)$ and $f_I(\tau_I)$ as the probability density functions (p.d.f.s) of the distribution of latency time and recovery time, respectively, we can convolve these distributions with the rate of entering the class τ time units ago. This results in the following system of integro-differential equations:

$$\frac{dS}{dt} = -\frac{\beta S(t)I(t)}{N} \tag{4.15}$$

$$\frac{dE}{dt} = \frac{\beta S(t)I(t)}{N} - \int_0^\infty \frac{\beta S(t-\tau_E)I(t-\tau_E)}{N} f_E(\tau_E)\, d\tau_E \tag{4.16}$$

$$\frac{dI}{dt} = \int_0^\infty \frac{\beta S(t-\tau_E)I(t-\tau_E)}{N} f_E(\tau_E)\, d\tau_E \tag{4.17}$$

$$- \int_0^\infty \int_0^\infty \frac{\beta S(t-\tau_E-\tau_I)I(t-\tau_E-\tau_I)}{N} f_E(\tau_E)\, d\tau_E f_I(\tau_I)\, d\tau_I, \tag{4.18}$$

the initial condition $E(0) = E_0$ and we are required to know the entire previous history of $S(t)$ and $I(t)$ for $t \leq 0$. Similar models have been used by in the past in disease settings (for example, [14, 33, 58]) and solved by a variety of different methods, including transforming the integro-differential equation system into abstract evolution equations [10].

For an individual that has been in the E class for τ time units, the hazard function gives the rate of change of the probability that we remain in the E class. It is a central concept in survival analysis that the hazard function can be written as [19]

$$\lambda_E(\tau) = \frac{f_E(\tau)}{1 - F_E(\tau)}, \tag{4.19}$$

where $f_E(\tau)$ is the p.d.f. of the distribution of time spent in the E class and $F_E(\tau)$ is the cumulative density function (c.d.f.) of said distribution.

Using hazard functions, we can create an alternative formulation of the distrubted delay SEIR model. Rather than having just $E(t)$ and $I(t)$ we can account for the time since entering the class, and model the distribution of numbers of E (and I) that have been in the class for that amount of time. We call that time τ and now we have the classes E and I dependent on two independent variables, t and τ. To obtain

$E(t)$ and $I(t)$, we integrate across all delay times like so

$$E(t) = \int_0^\infty E(t;\tau)\, d\tau \tag{4.20}$$

$$I(t) = \int_0^\infty I(t;\tau)\, d\tau. \tag{4.21}$$

We then use the following system of partial differential equations (PDEs) to describe the distributed delay SEIR model:

$$\frac{dS}{dt} = -\frac{\beta S(t) I(t)}{N} \tag{4.22}$$

$$\frac{\partial E}{\partial t} + \frac{\partial E}{\partial \tau} = -\lambda_E(\tau) E(t;\tau) \tag{4.23}$$

$$\frac{\partial I}{\partial t} + \frac{\partial I}{\partial \tau} = -\lambda_I(\tau) I(t;\tau), \tag{4.24}$$

together with the boundary conditions $E(t;0) = \frac{\beta S(t) I(t)}{N}$ and $I(t;0) = \int_0^\infty \lambda_E(\tau) E(t;\tau)\, d\tau$ and initial conditions $S(0) = S_0$, $E(0;\tau) = \phi_E(\tau)$ and $I(0;\tau) = \phi_I(\tau)$. The functions λ_E and λ_I are the hazard functions for the distributions of times spent in the respective state.

The benefit of the above model (in either formulation) is that there is far less restriction in the distribution associated with the latency and recovery times. However, both formulations have difficulties in terms of their implementation or simulation. The integro-differential equation form is rather complicated to solve even from a numerical perspective when delays are nonlinear and the PDE is computationally intensive, potentially requiring a significant amount of time when one has to perform on the order of thousands or more iterations to solve the inverse problem. To proceed with the inverse problem, we shall try to add a level of specification in our distributions to circumvent numerical difficulties—we can assume gamma distributed delays.

4.3.2 Gamma Distributed Delays

The gamma distribution has the p.d.f.

$$g(t;a,b) = \frac{1}{\Gamma(a) b^a} t^{a-1} e^{-t/b}, \tag{4.25}$$

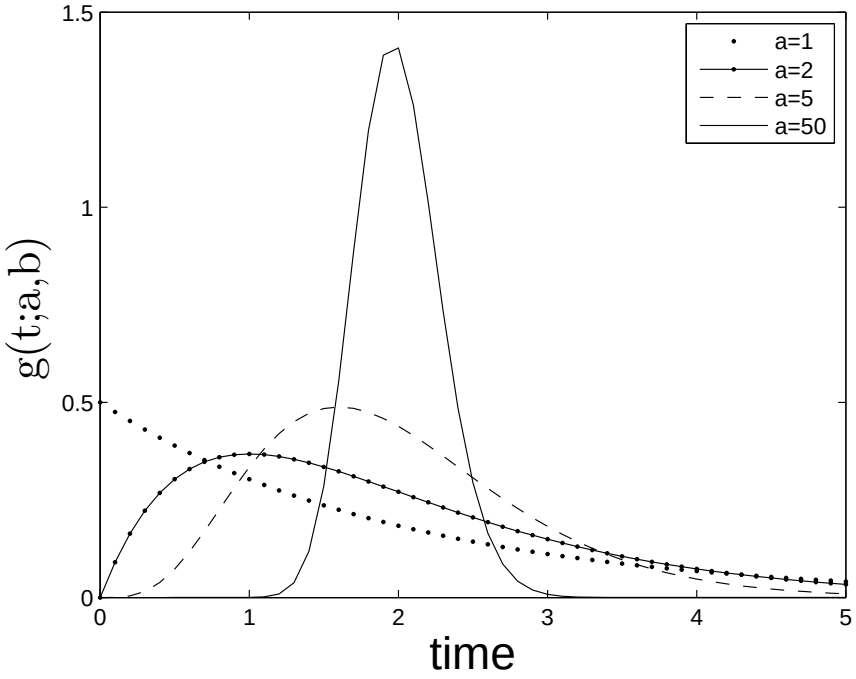

Figure 4.6: Gamma distributed latency/infectious periods. The graphs represent the p.d.f.s of the gamma distributions with $a = 1$ (dotted), $a = 2$ (dot-dashed), $a = 5$ (dashed) and $a = 50$ (solid). The mean of the distribution in each case is two. When $a = 1$, we have the exponential distribution. When $a = 50$, the curve approaches the p.d.f. of the normal distribution.

where $t \geq 0$, $\Gamma(a)$ is the Gamma function, and a is called the shape parameter and b is called the scale parameter and are both positive. The reciprocal of b is sometimes called the rate parameter. The expected value of a gamma distributed random variable is ab and the variance is ab^2 [19]. Example plots of the p.d.f of the gamma distribution are illustrated in Figure 4.6.

Conveniently, a gamma distributed random variable with integral shape parameter is the sum of independent, identically distributed exponential random variables [19]. It is possible to create a gamma distributed duration of infection by subdividing the infective class into stages in series, each with an exponentially distributed sojourn

(with identical means). Specifically, when an individual becomes infectious, they enter the I_1 compartment and depart it with rate $n\gamma I_1$ where n is the number of stages in the aggregate I class. The individual would proceed through each of the n stages until they recover by leaving the last compartment I_n. This is referred to as the method of stages [31, 51, 63, 84]. It should be noted that this subdivision of the I class into n compartments is simply a mathematical construct to produce the desired distribution and does not necessarily have a biological interpretation.

This same process can be used to obtain a gamma distributed latent period. Doing so yields the $\mathrm{SE}_m\mathrm{I}_n\mathrm{R}$ [63, 84] model

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{4.26}$$
$$\frac{dE_1}{dt} = \frac{\beta SI}{N} - m\nu E_1 \tag{4.27}$$
$$\frac{dE_2}{dt} = m\nu E_1 - m\nu E_2 \tag{4.28}$$
$$\vdots \tag{4.29}$$
$$\frac{dE_m}{dt} = m\nu E_{m-1} - m\nu E_m \tag{4.30}$$
$$\frac{dI_1}{dt} = m\nu E_m - n\gamma I_1 \tag{4.31}$$
$$\frac{dI_2}{dt} = n\gamma I_1 - n\gamma I_2 \tag{4.32}$$
$$\vdots \tag{4.33}$$
$$\frac{dI_n}{dt} = n\gamma I_{n-1} - n\gamma I_n. \tag{4.34}$$

Note that the standard SIR and SEIR models are nested within the $\mathrm{SE}_m\mathrm{I}_n\mathrm{R}$ model; let $\nu \to \infty$ and set $n = 1$ to obtain the SIR and set $m = n = 1$ to obtain the SEIR.

While this formulation is convenient from a forward problem perspective, it has a glaring issue when conducting the inverse problem. The gamma shape parameters m and n must be whole numbers. Wearing *et al.* fit the $\mathrm{SE}_m\mathrm{I}_n\mathrm{R}$ model when $m \in \{0, 1, 2, 3\}$ and $n \in \{1, 2, 3, 4\}$. When attempting to find a globally optimal parameter

set, however, it is useful to have $m, n \in \mathbb{R}^+$. To conduct such an experiment, we require a slightly more general formulation of the model with gamma distributed delays.

We choose to return to the PDE formulation (Equations 4.22-4.24) and accept the computational time required to gain the desired versatility. We maintain our assumption of gamma distributions, so λ_E and λ_I represent the hazard functions of the gamma distribution with shape parameters a_E and a_I and scale parameters $(a_E\nu)^{-1}$ and $(a_I\gamma)^{-1}$ respectively. Thus, the average duration of latency is $1/\nu$ and $1/\gamma$ is the average duration of infectiousness.

We shall approximate a solution to the PDE using the Foward Euler (FE) method with step-size identical in both the t and τ direction ($\Delta t = \Delta\tau$). The FE scheme for this SEIR PDE model is

$$
\begin{aligned}
t_i &= t_0 + i\Delta t && (4.35)\\
\tau_j &= \tau_0 + j\Delta t && (4.36)\\
S_0 &= S(0) && (4.37)\\
S_{i+1} &= S_i - \Delta t\frac{\beta S_i}{N}\sum_{j=0}^{\infty} I_{i,j} && (4.38)\\
E_{0,0} &= E(0) && (4.39)\\
E_{0,j>0} &= 0 && (4.40)\\
E_{i+1,0} &= \Delta t\frac{\beta S_i}{N}\sum_{j=0}^{\infty} I_{i,j} && (4.41)\\
E_{i+1,j+1} &= E_{i,j} - \Delta t\lambda_E(\tau_j)E_{i,j} && (4.42)\\
I_{0,0} &= I(0) && (4.43)\\
I_{0,j>0} &= 0 && (4.44)\\
I_{i+1,0} &= \Delta t\sum_{j=0}^{\infty}\lambda_E(\tau_j)E_{i,j} && (4.45)\\
I_{i+1,j+1} &= I_{i,j} - \Delta t\lambda_I(\tau_j)I_{i,j}, && (4.46)
\end{aligned}
$$

where S_i, $E_{i,j}$ and $I_{i,j}$ are the approximations of $S(t_i)$, $E(t_i, \tau_j)$ and $I(t_i, \tau_j)$ respec-

tively. We also consider that the population is entirely susceptible except for E_0 newly exposed individuals and I_0 newly infectious individuals who arrived at time 0. The convergence rate of FE is dependent on the step-size, Δt, and is $O(\Delta t^2)$. To test the accuracy of this FE scheme, we used synthetic data generated from the $\mathrm{SE}_m\mathrm{I}_n\mathrm{R}$ model with various sets of parameters. We chose $\Delta t = 0.005$, which gave a relative error of about 0.5% and only requires a reasonable amount of computational time (on the order of 2 seconds for one model solution, of which thousands could be necessary for the inverse problem).

We first examine the cost function J of the SEIR model when fit to the boarding school data while fixing the scale parameters at specific values. The parameters β, $1/\nu$ and $1/\gamma$ were fitted using OLS at each point (a_E, a_I). The results can be seen in Figure 4.7. Notice that the surface is relatively smooth and convex, allowing the minimization routine an easy time (at least in the a_E and a_I directions) an easy time to find the minimum.

We then fit the SEIR PDE model to the boarding school data allowing a_E and a_I to also be fitted by the OLS routine. The results of the fitting procedure can be seen in Figure 4.8. For the same number of fitted parameters as Wearing *et al.* used in their $\mathrm{SE}_2\mathrm{I}_2\mathrm{R}$ model, this model was able to obtain a better fit to the boarding school data. Our estimated value of $R_0 = 56.8670$, is significantly higher than Wearing *et al.*'s estimate of 35.9. The estimated value of the average duration of latency, 2.8610 days, seems reasonable, as there was about a three day period between when the sick boy returned from Hong Kong and the initial three cases were reported in the boarding school. The estimate for the average duration of infection seems low, however, at 2.0763 days when the article states that in most cases boys were bed-ridden for three to seven days.

In the end, our model selection score, $AIC_c = 135.2282$, is more than 12 points above the minimum AIC_c value obtained, which was for the step-function $\omega(t)$ SICR model which also has 5 structural parameters, implying that this is not a model well-suited to describing the data.

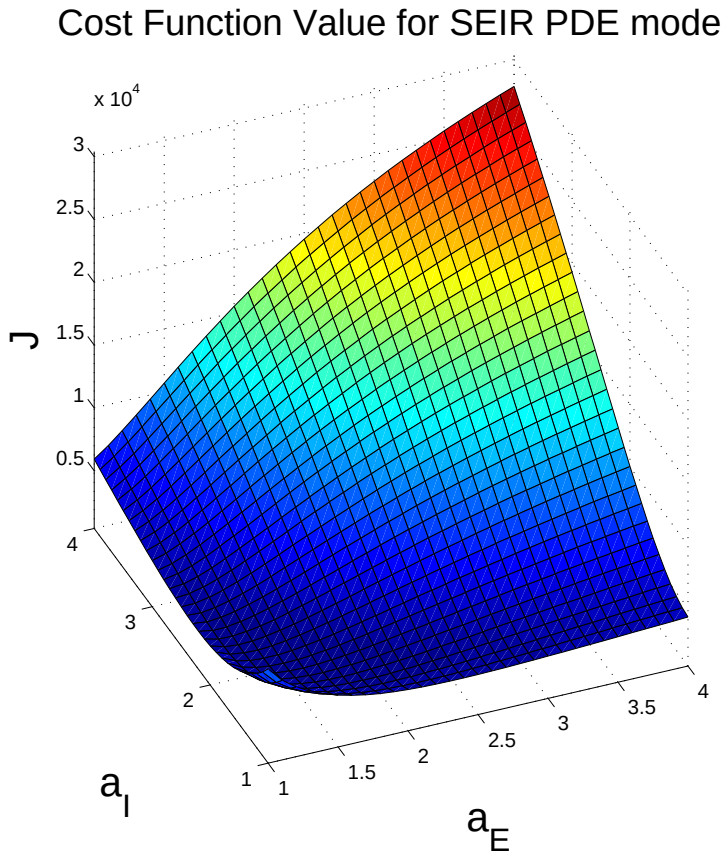

Figure 4.7: The cost function surface J across values of the shape parameters a_E and a_I of the gamma distribution for the PDE SEIR model with gamma distributed duration of latency and infection. The model was fit to the boarding school data using OLS at the fixed a_E and a_I values, while fitting β, $1/\nu$ and $1/\gamma$.

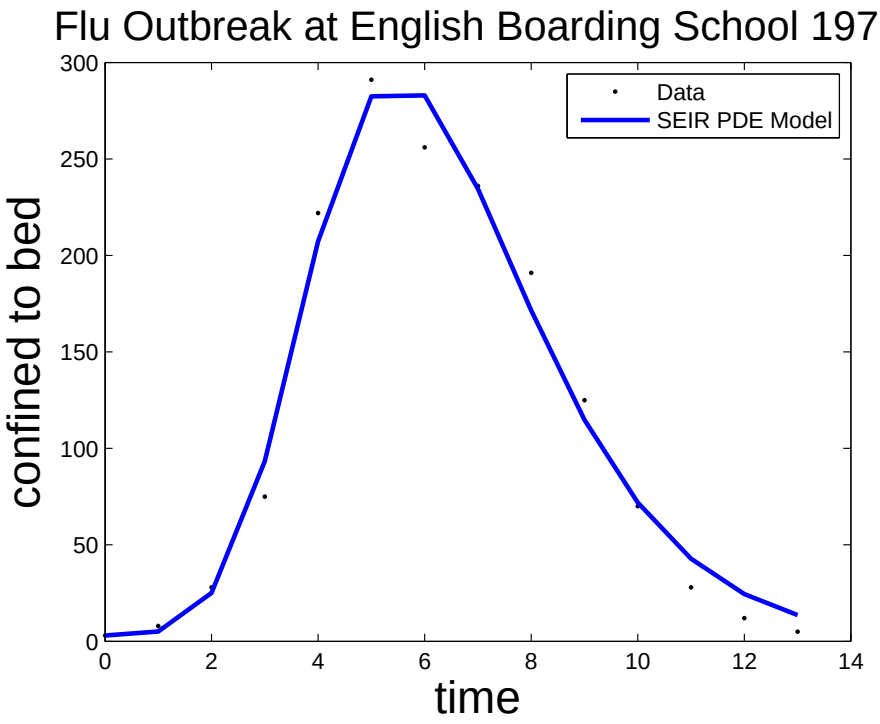

Figure 4.8: Plot of predicted prevalence from the OLS fit of the SEIR PDE model (solid curve) against time together with the boarding school data (dots). The best fit parameters are $\beta = 27.3887$, $a_E = 2.5448$, $a_I = 1.9559$, $1/\nu = 2.8610$, and $1/\gamma = 2.0763$. The basic reproductive number was found to be $R_0 = 56.8670$. The model selection score is $AIC_c = 135.2282$ and the cost functional value is $J = 2312.4$. $N = 763$ and the initial conditions are $S(0) = 760$, $E(0;\tau) = 0$ and $I(0;\tau) = 3$.

Table 4.5: A list of the results from all of the models in this chapter that were fit using OLS to the boarding school data. The listed values are the number of free structural parameters p for the model, the estimate of the basic reproductive number R_0, the cost functional value J, the model's AIC_c and the AIC differences, Δ_i. The models are listed (top to bottom) most plausible to least plausible.

Model	p	R_0	J	AIC_c	Δ_i
SICR (piece-wise ω)	5	4.2547	978.1	123.1824	0
SEIR (piece-wise ν)	5	7.8621	1106.2	124.9049	1.7225
SIR (piece-wise γ)	4	-	1974.5	126.5165	3.3341
SIR	2	3.7627	3960.8	127.1624	3.9800
SICR	3	3.0745	3147.5	127.9891	4.8067
SEIR	3	11.6263	3157.9	128.0351	4.8527
SIR (piece-wise β)	4	-	2214.0	128.1192	4.9368
SICR (piece-wise γ)	5	-	1444.6	128.6420	5.4596
SEIR (piece-wise γ)	5	-	1647.2	130.4788	7.2964
SEICR	4	10.5652	2816.8	131.4903	8.3079
SEIR (gamma delays)	5	56.8670	2312.4	135.2282	12.0458
SEIR (piece-wise β)	5	-	2449.8	136.0361	12.8537
SICR (piece-wise β)	5	-	3229.3	139.9037	16.7213

4.4 Discussion

Model selection methods give us a means of reducing structural uncertainty when fitting a model to data. The boarding school data is a classic example of a data set having a vague underlying process, so it was an obvious choice for a case study of model selection. We employed the use of the small sample version of the Akaike information criterion AIC_c to a fit series of 13 compartmental models with both constant and time dependent step-function parameters as well as a gamma distributed latency and recovery SEIR model. The results of the OLS fitting for all 13 models have been summarized in Table 4.5, where they are ranked according to AIC_c.

It can be seen that the SICR model with piece-wise ω was the best fitted model and carried an estimate of $R_0 = 4.2547$. However, many of the models are in the plausible range given their Δ_i values. We can only safely reject the gamma distributed delay

SEIR model (which estimated an extremely high value of R_0) and the piece-wise β SEIR and SICR models. Of the remaining models, there is still a relatively wide-range of estimates of R_0 from 3.0745 to 11.6263.

It is interesting that the gamma distributed SEIR model was rejected by the model selection process. Biologically, there is known to exist a latent period before an individual becomes infectious, thus a strong justification for the E class. Also, it is well-known that recovery (and likely the latency) process is not expoentially distributed and that gamma is a likely distribution. Burnham and Anderson [16] suggest that if there is a priori system information known (such as the existence of latency), then that should certainly be included in the model selection process. However, this makes the process more subjective. Perhaps the models with superior descriptions of the system produced poorer fits simply because of the small size of the data set.

Chapter 5

Conclusions and Future Directions

5.1 Concluding Remarks

This dissertation has primarily focused on presenting the utility of uncertainty quantifying tools that accompany the inverse problem. These methods were given in the context of infectious disease models. However, we have attempted to present these methods in such a way that it would be easy to extend them to other systems.

In the first chapter, we give the motivation for studying uncertainty quantifying techniques, parameter identifiability and subset and model selection, especially in the context of epidemic models. We also present a brief background of some compartmental epidemic models used throughout the work, the basic SIR model, SEIR model and two endemic SIR models.

Chapter 2 contains a methodology we apply throughout the dissertation, asymptotic statistical theory. A method of using synthetic data to study difficulties in the parameter estimation process for a given model is presented. We use sensitivity functions and data sampling techniques to discern the informativeness of individual data points and use this to guide how future data could be optimally sampled.

In Chapter 3, we highlight the difficulties of identifying parameters whose estimates are correlated. We then present a subset selection algorithm that is used to prune less informative parameters from the set of fitted parameters.

In Chapter 4, we discuss the issue of structural uncertainty and how one can reduce it by using proper model selection techniques. We used a case study of data from an influenza outbreak in a boarding school to which we fit a series of epidemic models to find the best fit model. We found that the most general models were rejected, including a gamma distributed SEIR model while the SICR model with piece-wise ω was the best fitted model and estimated of R_0 for the outbreak at 4.2547.

5.2 Future Directions

As the methods presented in this dissertation are primarily data analysis driven, it is important that we know that the data we have is representative of the system we are trying to study and describe. That is why we chose to use synthetic data in Chapters 2 and 3 and performed model selection on data from a well-defined population in Chapter 4. However, in the infectious disease scenario, it is very likely that not all infected individuals report to doctors or health centers, like the Centers for Disease Control. Thus, data is oftentimes under reported, and thus not necessarily representative of the actual number of cases. Not taking such a phenomenon into account could possibly dramatically alter the results of parameter estimation. We leave this issue as an open question that we desire to pursue in the future.

Another question that remains is the impact of having so few data points in the case study in Chapter 4. Could the fact that some models with more structural parameters received very low AIC_c scores simply due to the fact that adding a single additional model parameter dramatically reduces that score when there is so little data? Perhaps if the outbreak had more data points available, but was otherwise the same, different models would have been selected.

Bibliography

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. on Auto. Cont. **6** (1974), 716–723.

[2] R. M. Anderson, C. A. Donnelly, N. M. Ferguson, M. E. J. Woolhouse, C. J. Watt, H. J. Udy, S. MaWhinney, S. P. Dunstan, T. R. E. Southwood, J. W. Wilesmith, J. B. M. Ryan, L. J. Hoinville, J. E. Hillerton, A. R. Austin, and G. A. H. Wells, *Transmission dynamics and epidemiology of bse in british cattle*, Nature **382** (1996), 779–788.

[3] R. M. Anderson, B. T. Grenfell, and R. M. May, *Oscillatory fluctuations in the incidence of infectious disease and the impact of vaccination: time series analysis*, J. Hygiene **93** (1984), 587–608.

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

[5] 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.

[6] Anonymous, *Influenza in a boarding school*, Brit. Med. J. **1** (1978), no. 6112, 587.

[7] H. T. Banks, M. Davidian, J. R. Samuels Jr., and K. L. Sutton, *An inverse problem statistical methodology summary*, Mathematical and statistical estimation

approaches in epidemiology (G. Chowell, J. M. Hyman, L. M. A. Bettencourt, and C. Castillo-Chávez, eds.), Springer, New York, 2009, pp. 249–302.

[8] 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.

[9] H. T. Banks, S. L. Ernstberger, and S. L. Grove, *Standard errors and confidence intervals in inverse problems: Sensitivity and associated pitfalls*, J. Inv. Ill-posed Problems **15** (2006), 1–18.

[10] H. T. Banks and F. Kappel, *Spline approximations for functional differential equations*, J. Diff. Eqns. **34** (1979), 496–522.

[11] H. T. Banks and H. T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, Boca Raton, FL, 2009.

[12] R. Bellman and K. J. Åström, *On structural identifiability*, Math. Biosci. **7** (1970), 329–339.

[13] L. M. A. Bettencourt, R. M. Ribeiro, G. Chowell, T. Lant, and C. Castillo-Chavez, *Towards real time epidemiology: data assimilation, modeling and anomaly detection of health surveillance data streams.*

[14] D. M. Bortz, *Modeling, analysis and estimation of an in vitro hiv infection using functional differential equations*, Ph.d. thesis, N. C. State University, 2002.

[15] 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.

[16] K. P. Burnham and D. Anderson, *Model Selection and Multi-Model Inference*, Springer, 2002.

[17] 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.

[18] 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.

[19] G. Casella and R. L. Berger, *Statistical Inference*, 2 ed., Duxbury, Pacific Grove, CA, 2002.

[20] 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.

[21] 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.

[22] G. Chowell, C. E. Ammon, N. W. Hengartner, and J. M. Hyman, *Transmission dynamics of the great influenza pandemic of 1918 in geneva, switzerland: Assessing the effects of hypothetical interventions*, J. Theor. Biol. **241** (2006), 193–204.

[23] ———, *Estimating the reproduction number from the initial phase of the Spanish flu pandemic waves in Geneva, Switzerland*, Math. Biosci. Eng. **4** (2007), 457–470.

[24] G. Chowell, C. Castillo-Chavez, P. W. Fenimore, C. Kribs-Zaleta, L. Arriela, and J. M. Hyman, *Implications of an uncertainty and sensitivity analysis for SARS' basic reproductive number for general public health measures.*

[25] G. Chowell, P. W. Fenimore, M. A. Castillo-Garsow, and C. Castillo-Chavez, *SARS outbreaks in Ontario, Hong Kong and Singapore: the role of diagnosis and isolation as a control mechanism*, J. Theor. Biol. **224** (2003), 1–8.

[26] 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.

[27] 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** (2008), 852–864.

[28] 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.

[29] 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.

[30] 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.

[31] D. R. Cox and H. D. Miller, *The theory of stochastic processes*, Methuen, London, 1965.

[32] M. Davidian and D. M. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman & Hall, 1996.

[33] O. Diekmann and J. A. P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*, Wiley, 2000.

[34] K. Dietz, *The estimation of the basic reproduction number for infectious diseases*, Statistical Methods in Medical Research **2** (1993), 23–41.

[35] J. Dushoff, J. B. Plotkin, S. A. Levin, and D. J. D. Earn, *Dynamical resonance can account for seasonality of influenza epidemics*, Proc. Natl. Acad. Sci. USA **101** (2004), 16915–16916.

[36] M. Eslami, *Theory of Sensitivity in Dynamic Systems, An Introduction*, Springer-Verlag, New York, NY, 1994.

[37] 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.

[38] N. M. Ferguson, C. A. Donnelly, and R. M. Anderson, *Transmission intensity and impact of control policies on the foot and mouth epidemic in great britain*, Nature **413** (2001), 542–548.

[39] P. E. M. Fine and J. A. Clarkson, *Measles in England and Wales - II: The impact of the measles vaccination programme on the distribution of immunity in the population*, Int. J. Epidemiol. **11** (1982), 15–25.

[40] B. Finkenstadt, O. N. Bjornstad, and B. T. Grenfell, *A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks*, Biostat. **3** (2002), 493–510.

[41] B. Finkenstädt and B. Grenfell, *Empirical determinants of measles metapopulation dynamics in England and Wales*, Proc. R. Soc. Lond. B **265** (1998), 211–220.

[42] C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, T. Jombart, W. R. Hinsley, N. C. Grassly, F. Balloux, A. C. Ghani, N. M. Ferguson, A. Rambaut, O. G. Pybus, H. Lopez-Gatell, C. M. Apluche-Aranda, I. B. Chapela, E. P. Zavala, D. Ma. E. Guevara, F. Checchi, E. Garcia, S. Hugonnet, C. Roth, and The WHO Pandemic Assessment Collaboration, *Pandemic potential of a strain of influenza a (h1n1): Early findings*, Science (2009).

[43] K. Glover and J. C. Willems, *Parametrizations of linear dynamical systems: Canonical forms and identifiability*, IEEE Trans. Auto. Contr. **48** (1974), no. AC-19, 640–645.

[44] G. H. Golub and C. F. Van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, MD, 1996.

[45] N. C. Grassly and C. Fraser, *Seasonal infectious disease epidemiology*, Proc. R. Soc. Lond. B **273** (2006), 2541–2550.

[46] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Review **42** (2000), 599–653.

[47] 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.

[48] A. Holmberg, *On the practical identifiability of microbial growth models incorporating Michaelis-Menten type nonlinearities*, Math. Biosci. **62** (1982), 23–43.

[49] C. H. Hurvich and C. Tsai, *Regression and time series model selection in small samples*, Biometrika **76** (1989), 297–307.

[50] J. A. Jacquez and P. Greif, *Numerical parameter identifiability and estimability: Integrating identifiability, estimability and optimal sampling design*, Math. Biosci. **77** (1985), 201–227.

[51] A. Jensen, *An elucidation of erlang's statistical works through the theory of stochastic processes*, The Life and Works of A. K. Erlang (E. Brockmeyer, H. L. Halstrøm, and A. Jensen, eds.), The Copenhagen Telephone Company, Copenhagen, 1948, pp. 23–100.

[52] M Kalivianakis, S. L. J. Mous, and J. Grasman, *Reconstruction of the seasonally varying contact rate for measles*, Math. Biosci. **124** (1994), 225–234.

[53] M. J. Keeling, M. E. J. Woolhouse, D. J. Shaw, L. Matthews, M. Chase-Topping, D. T. Haydon, S. J. Cornell, J. Kappey, J. Wilesmith, and B. T. Grenfell, *Dynamics of the 2001 uk foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape*, Science **294** (2001), no. 5543, 813–817.

[54] W. O. Kermack and A. G. McKendrick, *A contribution to the mathematical theory of epidemics*, Proc. R. Soc. A **115** (1927), 700–721.

[55] A. M. Kilpatrick, A. A. Chmura, D. W. Gibbons, R. C. Fleisher, P. P. Marra, and P. Daszak, *Predicting the global spread of h5n1 avian influenza*, Proc. Natl. Acad. Sci. USA **103** (2006), 19368–19373.

[56] S. Kotz, N. Balakrishnan, C. Read, and B. Vidakovic (eds.), *Encyclopedia of Statistics*, 2 ed., Wiley-Interscience, Hoboken, New Jersey, 2006.

[57] M. Kretzschmar, S. van den Hof, J. Wallinga, and J. van Wijngaarden, *Ring vaccination and smallpox control*, Emerg. Inf. Dis. **10** (2004), 832–841.

[58] Y. Kuang, *Delay Differential Equations With Applications in Population Dynamics*, Acadmeic Press Inc., San Diego, CA, 1993.

[59] Y. A. Kuznetsov and C. Piccardi, *Bifurcation analysis of periodic SEIR and SIR epidemic models*, J. Math. Biol. **32** (1994), 109–121.

[60] P. E. Lekone and B. F. Finkenstadt, *Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study*, Biometrics **62** (2006), 1170–1177.

[61] 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. **268** (2001), 847–854.

[62] ———, *Destabilization of epidemic models with the inclusion of realistic distributions of infections periods*, Proc. R. Soc. Lond. B **268** (2001), 985–993.

[63] ———, *Sensitivity of model-based epidemiological parameter estimation to model assumptions*, Mathematical and statistical estimation approaches in epidemiology (G. Chowell, J. M. Hyman, L. M. A. Bettencourt, and C. Castillo-Chavez, eds.), Springer, New York, 2009, pp. 123–141.

[64] C. C. Lord, M. E. J. Woolhouse, J. A. P. Heesterbeek, and P. S. Mellor, *Vector-borne diseases and the basic reproduction number: a case study of african horse sickness*, Med. Vet. Entomol. **10** (1996), 19–28.

[65] A. D. R. McQuarrie and C. Tsai, *Regression and Time Series Model Selection*, World Scientific, 1998.

[66] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Hoboken, New Jersey, 2000.

[67] A. J. Miller, *Subset Selection in Regression*, Chapman & Hall, New York, 1990.

[68] C. E. Mills, J. M. Robins, and M. Lipsitch, *Transmissibility of 1918 pandemic influenza*, Nature **432** (2004), 904–906.

[69] J. D. Murray, *Mathematical Biology I: An Introduction*, 3 ed., Springer, New York, NY, 2002.

[70] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, New York, NY, 1999.

[71] 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), no. 10, 7518–25.

[72] J. G. Reid, *Structural identifiability in linear time-invariant systems*, IEEE Trans. Auto. Contr. **22** (1977), 242–246.

[73] F. Riedo, B. Plikaytis, and C. Broome, *Epidemiology and prevention of meningococcal disease*, Pediatr. Infect. Dis. J. **14** (1995), 643–657.

[74] S. Riley, C. Fraser, C. A. Donnelly, A. C. Ghani, L. J. Abu-Raddad, A. J. Hedley, G. M. Leung, L.-M. Ho, T.-H. Lam, T.-Q. Thach, P. Chau, K.-P. Chan, S.-V. Lo, P.-Y. Leung, T. Tsang, W. Ho, K.-H. Lee, E. M. C. Lau, N. M. Ferguson, and R. M. Anderson, *Transmission dynamics of the etiological agent of sars in hong kong: Impact of public health interventions*, Science **300** (2003), 1961–1966.

[75] J. V. Ross, D. E. Pagendam, and P. K. Pollett, *On parameter estimation in population models ii: Multi-dimensional processes and transient dynamics*, Theor. Popul. Biol. **75** (2009), 123–132.

[76] A. Saltelli, K. Chan, and E. M. Scott (eds.), *Sensitivity Analysis*, John Wiley & Sons, New York, NY, 2000.

[77] M. A. Sanchez and S. M. Blower, *Uncertainty and sensitivity analysis of the basic reproductive rate*, Am. J. Epidemiol. **145** (1997), no. 12, 1127–37.

[78] P. E. Sartwell, *The incubation period and the dynamics of infectious disease*, Am. J. Epidemiol. **83** (1966), 204–216.

[79] I. B. Schwartz and H. L. Smith, *Infinite subharmonic bifurcation in an SEIR epidemic model*, J. Math. Biol. **18** (1983), 233–253.

[80] G. A. F. Seber and C. J. Wild, *Nonlinear Regression*, John Wiley & Sons, Hoboken, New Jersey, 2003.

[81] K. Thomaseth and C. Cobelli, *Generalized sensitivity functions in physiological system identification*, Ann. Biomed. Eng. **27** (1999), 607–616.

[82] M. van Boven, M. Koopmans, M. Du Ry van Beest Holle, A. Meijer, D. Klinkenberg, C. A. Donnelly, and H. Heesterbeek, *Detecting emerging transmissibility of avian influenza virus in human households*, PLoS Comput. Biol. **3** (2007), e145.

[83] 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.

[84] H. J. Wearing, P. Rohani, and M. Keeling, *Appropriate models for the management of infectious diseases*, PLoS Medicine **2** (2005), no. 7, e174.

[85] 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.

[86] B. G. Williams and C. Dye, *Infectious disease persistence when transmission varies seasonally*, Math. Biosci. **145** (1997), 77–88.

[87] J. Witte, A. Karchmer, M. Case, K. L. Hermann, E. Abrutyn, and I. Kassanof et al., *Epidemiology of rubella*, Am. J. Dis. Child. **118** (1969), 107–111.

[88] 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.

[89] X. Xia and C. H. Moog, *Identifiability of nonlinear systems with application to HIV/AIDS models*, IEEE Trans. Auto. Contr. **48** (2003), 330–336.

[90] 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.