Parameters Estimation, Identification and Uncertainty Quantification for Epidemic Models

Conrad, J.*, Guan, L., and Geneus C.

Department of Mathematics

Tulane University

New Orleans, LA 70118

May 11, 2018

Abstract

We describe the general procedures for parameter estimation and identifiability along with uncertainty quantification for parameters in mathematical models. We start the process by fitting the model to the observable data. We analyze the fit by checking the residuals and correlations between parameters. We use bootstrap to compute the parameter uncertainties. We then discuss several contemporary approaches such as local parameter identification, profile likelihood analysis, and low discrepancy sampling method to carry out the parameter identification procedure and determine which parameters can be identified. Once again, we refit the model and estimate all of the identifiable parameters and compute the uncertainties. We also review the parameter sensitivity analysis and analyze the underling relationships among parameters. Finally, we choose simple Susceptible-Infectious-Susceptible (SIS) and Susceptible-Exposed-Infectious-Susceptible (SEIS) models with fixed total population as our examples. Both noise free data and noisy data are used for fitting the model parameters. Parameter identifiability analysis, sensitivity analysis and parameter uncertainty quantification are also included.

Keywords: Parameter Estimation, Uncertainty Quantification, Sensitivity Analysis, Model Diagnosis, Parameter Identification, Epidemic Model, Optimization

*Corresponding author. E-mail address: jconrad4@tulane.edu.

Contents

1 Introduction

Parameter identifiability and estimation play a critical role in accurately describing system behaviors through mathematical models. It allows us to extract both theoretical analysis and numerical simulation to better understand the underlying mechanisms behind the models.

We are aiming to find the model that best represents the path of our observations. To reach the goal, we start with fitting the model to data. Assuming that we have a good knowledge of where the data come from, we propose the model $\hat{y} = M(\beta)$ which we believe is a good representation of the data, y. We calculate the sum of the squares of the difference between our model output and the data. Let $f = \frac{1}{2}\sum(y - \hat{y})^2$ as a function of unknown parameters β. We look for the minimum of f by solving $\nabla f = 0$. The solution that we get for β is the parameter estimation and therefore, we find the model.

We follow up our model fit by two types of analysis: residual analysis and uncertainty analysis. There are two major criteria for residual analysis. Firstly, the residuals have to be relatively small, which measures whether our model path is close enough to the data points. Secondly, their uncertainties have to be within certain range which shows that our model assumptions are reasonable. Residual analysis could be performed by simply plotting the residuals against the data. Ideally, the plot should show no pattern or structure. Box plot and correlation plot can also be used to check the patterns between parameters. Uncertainty quantification can be realized by bootstrap technique. If the data are not time dependent or uncorrelated, we use resampling method to get distributions of parameters. If they are from a time series, then blocked bootstrap is the method to use.

Of course, we would like to be able to uniquely find the estimation for all of the unknown parameters. However, the process usually yields complicated situations where some of the parameters are correlated or combined that makes it impossible to identify every single one of them. Scientists have developed several techniques to determine the parameter identifiability in local, extended and global regions. There are certain ways such as checking the Hession matrix of f to see whether it is positive definite or not for locally identifiability analysis, computing the profile likelihood function (leave-one-out) for extended identifiability analysis, and finally using the low discrepancy sampling method for global identifiability analysis . All of these provide a good amount of information about parameter identification and therefore help us find the best model. The differential algebra approach (see e.g. [1]) is another way to structurally identify parameters, and the essential idea is to reformulate the model based on the input data and therefore check the new parameters as combinations of the old parameters. In the situation that we need to use the input-output differential equation, we go back to implement the estimations along with uncertainty quantification for the new

parameters again.

Local sensitivity is another way of measuring the impact of parameter of interest (POI) to the quantity of interest (QOI). Sensitivity index is calculated by computing the partial derivate of the QOI with respect to the POI and then evaluating at the baseline value of the POI. The index can be positive or negative or zero, meaning whether a small change of the POI around the baseline point has a positive, negative or no impact on the QOI. Sensitivity can also be analyzed in a wider region or the whole space, depending on the real situation of the problem.

In this report, we discuss all of the aforementioned techniques in detail and combine them together to work out some examples such as SIS and SEIS models in epidemiology. Although epidemic models, such as SIS and SEIS models, are useful tools that could provide guidance for preventing diseases from spreading, usually we have limited information about how the epidemic takes off and how it is transmitted and how people recover. In reality, what we usually have are limited to incidence data. In this report, we use both theoretical and numerical methods to determine the identifiable parameters and try to predict the future infectious population size from the available incidence data. In the meantime, we include a detailed discussion of the program outputs.

2 Methods

2.1 Fitting Model to Data

We begin our procedure with finding the best mathematical model $\hat{y} = X\beta$ as a function of the parameter β that represents the path to a series of data points. First we define the residual R as the distance between the mean of the distribution predicted by our model $\hat{y}$ and the actual data points observed by y. Notice that residuals can be positive or negative, so squaring the residuals is a natural way to measure the magnitude of it. Let $f = \frac{1}{2}R^T R$, where $R = \hat{y} - y$. Therefore, we end up with solving a system that minimizes $f = ||X\beta - y||_2^2$.

2.1.1 Linear Model

Let us consider the linear system $y = X\beta$, where our parameters of interest (POI) are $\beta \in R_n$ and our quantities of interest (QOI) are $y \in R_m$. $X \in R_{m\times n}$ is some given matrix.

Depending on the dimension of the problem, there are three major cases: full-rank, underdetermined and overdetermined systems. For a full rank system, one can solve it exactly, i.e. the solution is unique. For the other two cases, we propose the Moore-Penrose pseudo-inverse to help us find the optimal solution.

Let X be an $m \times n$ matrix over the field of real or complex numbers. Then the Moore-Penrose pseudo-inverse, X^+, associated with X is a unique $n \times m$ matrix over the same field, which satisfies

- $XX^+X = X$,
- $X^+XX^+ = X^+$,
- $(XX^+)^* = XX^+$,
- $(X^+X)^* = X^+X$,

where X^* denotes the conjugate transpose of X. According to [2],

$$\begin{aligned} X^+ &= \lim_{\delta \to 0} (XX^*X + \delta^2 I)^{-1} X^* \\ &= \lim_{\delta \to 0} X^*(XX^* + \delta^2 I)^{-1}. \end{aligned}$$

C1: Over-determined System

For an over-determined system, i.e., when $m > n$, with full rank, we know X^*X is invertible. Hence, according to the first equality above, we know $X^+ = (X^*X)^{-1}X^*$.

For any $n \times 1$ vector β, we have the decomposition

$$X\beta - y = X(\beta - X^+y) + (XX^+ - I)y. \tag{1}$$

We note that

$$\begin{aligned} & (X\beta - XX^+y)^*(XX^+y - y) \\ &= (\beta^*X^* - y^*(XX^+)^*) - (XX^+y - y) \\ &= \beta^*X^*XX^+y - \beta^*X^*y - y^*(XX^+)^*XX^+y + y^*(XX^+)^*y \\ &= \beta^*X^*(XX^+)^*y - \beta^*X^*y - y^*XX^+XX^+y + y^*XX^+y \\ &= \beta^*(XX^+X)^*y - \beta^*X^*y - y^*XX^+y + y^*XX^+y \\ &= \beta^*X^*y - \beta^*X^*y - y^*XX^+y + y^*XX^+y \\ &= 0, \end{aligned}$$

where we have used the first three identities satisfied by the Moore-Penrose pseudo-inverse. Hence, the two vectors on the right-hand side of (1) are orthogonal. Let $z = X^+y$. Then from Pythagorean theorem we know

$$\begin{aligned} \|X\beta - y\|_2^2 &= \|X(\beta - X^+y)\|_2^2 + \|(XX^+ - I)y\|_2^2 \\ &= \|X(\beta - z)\|_2^2 + \|Xz - y\|_2^2 \\ &\geq \|Xz - y\|_2^2. \end{aligned}$$

Hence, $\|Xz - y\|_2$ minimizes $\|X\beta - y\|_2$. We remark that this result holds for arbitrary matrices. In particular, it holds for overdetermined matrices.

C2: Under-detemined System

On the other hand, for an under-determined system, i.e., when $m < n$, with full rank, let β be a solution to the linear system. Define $R = XX^+X$. Then we know

$$Rz = X^+XXX^+b = X^+y = z \quad \text{and} \quad R^* = R,$$

which follow from the second and fourth identities satisfied by the Moore-Penrose pseudo-inverse. Since

$$\begin{aligned} z^*(\beta - z) &= (Rz)^*(\beta - z) \\ &= z^*(R^*\beta - R^*z) \\ &= z^*(R\beta - Rz) \\ &= z^*(X^+X\beta - X^+y) \\ &= z^*X^+(X\beta - y) \\ &= 0, \end{aligned}$$

it holds that

$$\begin{aligned} \|\beta\|_2^2 &= \|\beta - z\|_2^2 + 2Re[z^*(\beta - z)] + \|z\|_2^2 \\ &= \|\beta - z\|_2^2 + \|z\|_2^2 \\ &\geq \|z\|_2^2. \end{aligned}$$

Therefore, $\|z\|_2$ minimizes $\|\beta\|_2$ among solutions to the linear system $X\beta = b$ when $m < n$. In addition, it can be shown that for full rank systems, either of these methods can be used to solve for the unique solution of $\min_\beta ||X\beta - b||_2^2$.

2.2 Analysis of the Fit

2.2.1 Analysis of Residuals

To check whether our model has reached the goal to best represent our data while still respecting the model assumptions, we analyze the residuals of the model fit to the data. In other words, we analyze what is not explained by our model after explaining the variation of the data. In the ideal case, residuals should be relatively small and uncorrelated or unstructured. The analysis of residuals is a powerful tool for model diagnosis as it indicates whether all of the crucial features of the data are caught by model assumptions. The simplest way is plotting the residuals versus the data. If some pattern is observed, then we might

need to go back and change the model. If this is something more complicated to fix by simply changing the model, we go to the parameter identifiability analysis.

2.2.2 Analysis of Uncertainties – Bootstrap

Besides residual analysis, another approach to check the fit is using numerous sets of data as it provides more information and therefore produces multiple fits. As a result, instead of getting one single set of the parameter estimates, we get a distribution for each parameter. We then extract and analyze different kinds of information, such as mean, mode, 95% confidence interval and standard deviation *etc*, from the distribution. Unfortunately, we usually have limited amount of data due to collecting difficulty and extra expenses. In 1970, Bradley Efron developed a method called "bootstrap" that solves this problem [3]. It is a powerful resampling tool that can be used to learn more about the parameter behaviors and quantify the uncertainties without spending extra efforts on collecting data. One important constraint for applying bootstrap is that the observable data have to be uncorrelated. In other words, the data can not be a time series or the order of the data points must be serial independent. Once the condition is satisfied, a detailed procedure of bootstrap for parameter distribution analysis is listed as the following steps:

- **Step 1**. Start with observations (data) of size n, then use a generator to randomly generate a new sequence from the original data, which ranges from 1 to n with replacement.
- **Step 2**. Repeat **Step 1** for a large number of times (often 1,000 or 10,000 times) or until get desired number of data sets.
- **Step 3**. Use the same model to fit each data set. For each fit, label the parameter estimate as β_{ij} to denote the jth parameter estimate of the ith parameter in the model.
- **Step 4**. Plot the distribution for each parameter estimate β_i against the number of date set i.
- **Step 5**. Make the pairwise plot between parameters and see wether pattern exists.

The analysis of uncertainties will let us find out whether we are having a good knowledge of our data and making appropriate assumptions on the model. A good model should yield uncertainties of the parameters within a reasonable range. In the situation that out result is not satisfying, and yet we still believe that we are having an excellent model, we could go on with the step of parameter identification without going back to change the model.

2.3 Parameter Identifiability

Although we estimate our parameters through fitting the model to data, we might not get the correct results. Failures usually include but are not limited to: 1) The solution that minimizes the residual is not unique, which is often referred to as parameter non-identifiability. 2) The solution minimizing the residual is unique, but the uncertainties for some of the parameters do not look normal, e.g. one of the parameters is combined with another. The occurring of any of these problems makes it necessary for questioning the parameter identifiability. We consider the parameter identifiability analysis from two aspects: practical identifiability and structural identifiability. Practical identifiability is a numerical approach dealing with identifiability problems, while structural identifiability usually refers to theoretical approaches, such as differential algebra. There are different types of parameter identifiability, depending on the specific problems under consideration, namely, local identifiability, extended identifiability, and global identifiability. Next, we discuss them in detail.

2.3.1 Local Identifiability

We define a parameter to be locally identifiable if there exists a unique solution for this parameter to minimize $f = R^T R$ in a relatively small region that contains the baseline/best guess. This property is determined by calculating the eigenvalues of the Hessian matrix of f, where

$$\mathbf{H} = J(\nabla f) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial r_1^2} & \dfrac{\partial^2 f}{\partial r_1 \partial r_2} & \cdots & \dfrac{\partial^2 f}{\partial r_1 \partial r_n} \\ \dfrac{\partial^2 f}{\partial r_2 \partial r_1} & \dfrac{\partial^2 f}{\partial r_2^2} & \cdots & \dfrac{\partial^2 f}{\partial r_2 \partial r_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial r_n \partial r_1} & \dfrac{\partial^2 f}{\partial r_n \partial r_2} & \cdots & \dfrac{\partial^2 f}{\partial r_n^2} \end{bmatrix}, \qquad R = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ y_3 - \hat{y}_3 \\ \vdots \\ y_{n-1} - \hat{y}_{n-1} \\ y_n - \hat{y}_n \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_{n-1} \\ r_n \end{bmatrix}.$$

If $\mathbf{H}$ is a positive definite matrix, in other words, all of the eigenvalues of $\mathbf{H}$ are positive near the baseline/best guess, then we say that the parameter is locally identifiable. On the other hand, if $\mathbf{H}$ has at least one zero or close to zero eigenvalue, then we conclude that the parameters are non-identifiable.

2.3.2 Extended Identifiability – Profile Likelihood

Local identifiability analysis is limited from describing the whole space, and so we continue with the extended identifiability analysis in a wider region. Profile likelihood (or leave-one-out profiling) is a type of analysis that has been widely used for extended identifiability of model parameters (see e.g. [6]). Roughly speaking, the idea of the approach is that one holds a parameter of interest, p_i, as constant and minimize f over the other parameters p_j where $i \neq j$. The parameter p_i is then iterated over a range of values, i.e., $p_i \in [l, u]$, where l and u are the lower and upper limits of p_i, respectively. Such a range can be determined from the literature or from the results after applying bootstrap. After the iteration one then plots f versus the range of p_i and looks for a minimum. The existence of the minimum indicates that the parameter is identifiable in the extended region.

2.3.3 Global Identifiability

Practical global identifiability extends the idea of profile likelihood further by selecting a feasibility region for all of the parameters $p_i \in [l_i, u_i]$, where $i = 1, 2, ...$ We sample points from this region using low discrepancy sampling methods, such as the Latin hypercube [4], Halton [5], or Sobol methods [7], and then evaluate the system and calculate the mean squared error (MSE). The resulting contour plots can reveal both structural and practical identifiability properties. Figure 1 shows example plots of what we would expect to see when parameters are structurally or practically non-identifiable, versus identifiable.

2.3.4 Structural Identifiability – Differential Algebra

One of the analytic approaches to observe global identifiability is to transform a system into its input-output equation using differential elimination (cf. [1]). For a given state space model, we want to eliminate the state variables and their derivatives such that the resulting equation describes the system entirely in terms of inputs, outputs, and parameters. By using this approach, one can determine whether the system under consideration is generically globally identifiable or not.

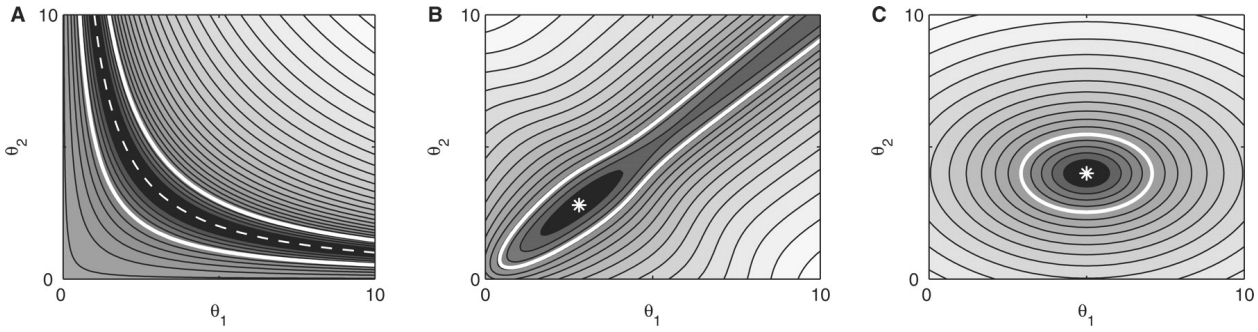

Figure 1: Contour plots for a two-dimensional parameter space, shown in non-logarithmic scale for illustrative reasons. Left panel: a structural non-identifiability. Center panel: a practical non-identifiability. Right panel: both parameters are identifiable (**?**).

2.4 Parameters Sensitivity Analysis

In this section we outline various levels of sensitivity analysis and related methods for capturing the uniqueness of parameters. Sensitivity analysis quantifies the uncertainty in the output that is attributable to the uncertainty in the input.

When we consider a system with multiple parameters, we want to quantify the impact of changes in these parameters on the relevant outputs of the model. Usually, the parameters of interest (POI) in a mathematical model are categorized as

- the parameters that can be controlled;
- the parameters that can only be approximated;
- the parameters that cannot be defined to be specific values because of randomness.

We would like to understand how the uncertainties of parameters affect the predictions of the QOI. Fortunately, often the relative ranking of the response of the QOI to the POI is a more robust measurement, even though the exact model predictions can be in error. Below we explore various methods for quantifying and understanding such an uncertainty in local, extended local, and global parameter spaces.

2.4.1 Local Sensitivity

In its simplest form, local sensitivity analysis defines the derivative of the QOI in a model as a function of the model parameters for a particular reference (baseline) solution. The sensitivity indices (derivatives) can quantify how small changes in the input POI cause variability in the output QOI without a scaling for the units or magnitude of each. These

values can be found in the Jacobian, also known as the unscaled sensitivity index. Relative sensitivity indices determine the relative importance of the model parameters on the model predictions.

In local sensitivity analysis, we perturb a reference (baseline) solution to quantify how the QOI change in response to small changes in the POI. The sign of the index indicates the direction of the response, and its magnitude tells us the relative importance of each parameter in model predictions.

Local sensitivity can be analyzed by the following method: Consider that we have been given a set of parameters p that have some amount of noise δp. We estimate the local sensitivity of p by perturbing $\hat{p}$ by $\delta\hat{p}$. Our output quantity of interest q, therefore can be estimated by the following:

$$\begin{aligned}\hat{q} = q(\hat{p} + \delta\hat{p}) &\approx q(\hat{p}) + \delta\hat{p}\frac{dq}{dp}\Big|_{\hat{p}} + O(\delta^2) \\ &\approx q(\hat{p}) + \delta\hat{p}\frac{dq}{dp}\Big|_{\hat{p}}\end{aligned}$$

where $\frac{dq}{dp} \approx \frac{q(p+\delta p)-q(p)}{\delta p}$.

Specifically, the relative sensitivity index of q at $\hat{p}$ is defined as:

$$S_p^q: \quad \lim_{\delta p \to 0} \frac{\frac{\delta q}{q}}{\frac{\delta p}{p}} = \left(\frac{\hat{p}}{\hat{q}}\right)\left(\frac{dq}{dp}\Big|_{\hat{p}}\right), \quad \hat{q} \neq 0.$$

Because such kind of analysis is based on the linearization of the solution around the baseline parameters, the sensitivity indices are only valid for POI very close to these baseline values, i.e., the local sensitivity analysis indices are only valid in a small neighborhood of the reference POI and related QOI. The local normalized relative sensitivity index, S_p, is the percent change in the output given the percent change in an input parameter. That is, if the parameter, p, changes by $x\%$, then the QOI will change by $S_p x\%$. Note that the sign of the sensitivity index indicates whether the QOI increases (> 0), decreases (< 0), or insensitive ($= 0$) with respect to the POI.

2.4.2 Extended Sensitivity – One-at-a-Time

As with local identifiability, local sensitivity analysis is limited from describing the whole space. We continue the sensitivity analysis in a wider region using One-at-a-Time Analysis. For this approach, one holds a parameter of interest p_i, as a constant and evaluates the value of a quantity of interest q. As before with the Profile Likelihood analysis, the parameter p_i is then iterated over a range of values, i.e., $p_i \in [l, u]$, where l and u are the lower and upper limits of p_i, respectively. The one plots q versus the range of p_i and analyzes the trend.

2.4.3 Global Sensitivity

As with Global Identifiability Analysis, we select a feasibility region for all of the parameters $p_i \in [l_i, u_i]$, where $i = 1, 2, ...$ We sample points from this region using low discrepancy sampling methods, such as the Latin hypercube [4], Halton [5], or Sobol methods [7], assuming the parameters live in a uniform distribution. We then map these sample points onto the distribution we believe the parameters p_i actually have, and then evaluate the system for these values of p_i and calculate the quantity of interest q. The resulting correlation plots can reveal the dependencies of the quantities of interest. Additionally, we can now evaluate the distribution of the quantity of interest given the distributions of our parameters of interest.

2.5 Cross-validation

In general, cross-validation aims to refit the model of interest $M(p)$ to various training sets, in order to obtain additional information about the fitted model. It is a technique for measuring the predictive ability of a model and helps us pick out the best one. The idea of cross-validation is to test a model on a set of data that has not been used in selection. This is referred to as "test set" and the data set used for fitting is called the "training set". For example, one can measure the predictive accuracy of a model by the mean squared error (MSE) on the test set, which generally is larger than the MSE on the training set because the test data were not used for selection. We have to tradeoff the bias-variance with the computational cost, that is, as we lower the number of folds in cross-validation to save the time and effort on validation, we increase the variance of our estimators as well. In order to quantify the uncertainties of the estimators and have real control of measuring the accuracy, we need to use the idea of bootstrap resampling again.

In the next section, we illustrate the ideas of the aforementioned methodologies by applying them to specific mathematical models in epidemiology.

3 Example – Simple Epidemic Models

3.1 Models

With all of the major techniques discussed above, we choose to study some cases in mathematical epidemiology. Infectious diseases have been a big threat for human and animal lives. They result in millions of deaths each year, especially in developing countries. Mathematicians have been using compartmental models to study the spread of infectious diseases for a long history, starting from the pioneering work of Daniel Bernoulli in 1760 on smallpox. In

general, compartmental models partition the population into several non-overlapping classes. A system of first-order ordinary differential equations is used to describe the dynamics of each class over time. However, due to the lack of sufficient real-world data, we often face the challenge of finding good estimates for the model parameters. Thus, parameter estimation, identification and uncertainty quantification become crucial in this journey.

Consider a Susceptible-Infectious-Susceptible (SIS) compartmental model, where infected individuals do not confer immunity and therefore can have repeated or reoccurring infections. After they recovered from the infectious state, they return to the susceptible state. Some of the diseases that fall into this category include influenza and sexually transmitted diseases such as gonorrhea and chlamydia. Mathematically, the system of ordinary differential equations (ODE) modeling the process with fixed total population can be analytically solved to understand the disease dynamics with accurate parameter values. In its simplest form, the SIS model reads as follows:

$$\begin{aligned}\dot{S} &= -\beta\frac{I}{N}S + \gamma I\\ \dot{I} &= \beta\frac{I}{N}S - \gamma I.\end{aligned}$$

Here β represents the total transmission rate through contacts with infectious individuals; γ is called the recovery rate which is defined as the probability of getting rid of the disease per time but then at which infected individuals become susceptible again; $N(t)$ is the total population size at time t and is given by $N(t) = S(t) + I(t)$. By adding the two equations in the SIS model, we find that $\dot{N}(t) = 0$, which indicates that the population is conserved, i.e., $N(t) = N(0)$. Hence, with the initial conditions at time $t = 0$, we have $S(t) = N(0) - I(t)$. Thus, we can reduce the system of equations to the single equation:

$$\begin{aligned}\dot{I} &= \beta\frac{I}{N}(N - I) - \gamma I\\ &= (\beta - \gamma)\, I\left(1 - \frac{I}{\frac{\beta-\gamma}{\beta}N}\right),\end{aligned}$$

which is indeed the logistic equation and can be solved analytically.

The SIS model can be made more complex by adding an Exposed (E) compartment:

$$\dot{S} = -\beta\frac{I}{N}S + \gamma I \tag{1}$$

$$\dot{E} = \beta\frac{I}{N}S - \omega E \tag{2}$$

$$\dot{I} = \omega E - \gamma I. \tag{3}$$

Sometimes called a staged infection model, this can be thought of as a Susceptible-Exposed-Infected-Susceptible (SEIS) model where individuals have an incubation period after getting infected, yet infectious. Here ω represents the progression rate. Again the total population is conserved in this system such that for some constant N, $S(t) + E(t) + I(t) = N$. This allows us to reduce the system to the following:

$$\dot{E} = \beta \frac{I}{N}(N - E - I) - \omega E$$
$$\dot{I} = \omega E - \gamma I.$$

We note that with the reduction of model dimension, the complexity of analysis is reduced, but the number of parameters that need to be estimated remains unchanged.

Besides the transmission rate, progression rate and recovery rate, we also consider the reporting rate k as an extra parameter. This is because in reality incidence data are based on the infectious cases that are being reported. That is, the incidence data may only reflect a fraction of the total infected population, and hence the reporting rate is a positive parameter bounded above by 1. To make our results as accurate as possible, we are going to estimate the reporting rate as well.

3.2 Parameter Estimation

To create a study example, we generate "fake" data by using the solutions of our ODE models. Parameter inputs are based on baseline values which we believe are our best guess. We then start fitting the model to data by finding the optimal solution that minimizes the residuals between data and the model. This also serves the purpose of checking whether our algorithm is working correctly.

To compare the results, we consider the following additive noise values for assessing practical identifiability properties: 0% and 3 % on both SIS and SEIS models. For parameter estimation and identifiability, since our data are considered as time sensitive, we cannot use bootstrapping in the whole time space. Nevertheless, we are able to divide our data into small blocks, and we assume that within each block (small time interval) the order of the data does not matter.

3.3 Parameter Identifiability

To analyze our fit, we present the local and profile likelihood methods for practical identifiability analysis, and differential algebra approach for structural identifiability analysis to exam our parameter estimations. Most importantly, due to the fact that the incidence data

for studying the spread of infectious diseases are based on reported cases, we present detailed derivations of the differential algebra approach for structural identifiability of the SIS and SEIS models including reporting rate. The idea is to use substitution and differentiation to eliminate all variables in the original models except for observed output.

3.3.1 Local and Extended Identifiability Analysis

Given parameter estimates from model fit, we carry out local and extended identifiability analysis to verify whether the parameters are identifiable or not, and therefore determine if we have found an optimal unique parameter set for the given model and data set. We first check the eigenvalues and eigenvectors of the Hessian matrix to investigate the local identifiability and then apply profile likelihood to exam the identifiability in an extended region.

3.3.2 Differential Algebra – Structural Identifiability Analysis

We can check the structural identifiability by converting the model into a differential equation with respect to the data. For the SIS model, suppose we can only observe y, a fraction k of the infected I compartment, such that $y = kI$. Then by plugging $y = kI$ and $\dot{y} = k\dot{I}$ into the logistic equation satisfied by $I(t)$, i.e.,

$$\dot{I} = (\beta - \gamma)\, I \left(1 - \frac{I}{\frac{\beta-\gamma}{\beta} N}\right),$$

we can show that

$$\frac{\dot{y}}{k} = (\beta - \gamma)\, \frac{y}{k} \left(1 - \frac{\frac{y}{k}}{\frac{\beta-\gamma}{\beta} N}\right),$$

which can be simplified as

$$\dot{y} = (\beta - \gamma) y - \frac{\beta}{Nk} y^2.$$

This is a second order Bernoulli equation that one can solve analytically. According to the general principle of differential algebra (see e.g. [1]), we conclude that the identifiable parameter combinations are $\frac{\beta}{Nk}$ and $\beta - \gamma$. We note that there are three single parameters: β, γ and k, assuming the total population size N is known. Hence, the SIS model with the output information $y = kI$ is not globally structurally identifiable. Nevertheless, with two relationships, the model parameters can be uniquely identified if one of them is known.

By applying the same approach to the SEIS model, we obtain

$$\frac{\dot{y}}{k} = \omega E - \gamma \frac{y}{k},$$

which yields

$$E = \frac{\dot{y} + \gamma y}{\omega k}.$$

By substituting the above expression for E into the equation for I, we have

$$\frac{\ddot{y} + \gamma \dot{y}}{\omega k} = \beta \frac{y}{Nk} \left(N - \frac{\dot{y} + \gamma y}{\omega k} - \frac{y}{k} \right) - \frac{\dot{y}}{k} - \frac{\gamma y}{k}.$$

After simplification, we obtain

$$\ddot{y} = -\frac{\beta}{Nk} y \dot{y} - (\omega + \gamma)\dot{y} - \frac{\beta}{Nk}(\omega + \gamma) y^2 + \omega(\beta - \gamma) y,$$

which implies that the combinations of parameters:

$$\frac{\beta}{Nk}, \qquad \omega + \gamma, \qquad \omega(\beta - \gamma)$$

are identifiable. Assuming the initial total population (conserved) is known, we end up with three relationships bridging four parameters: β, γ, ω, k. Therefore, the SEIS model with the output information $y = kI$ is not globally structurally identifiable. Again, as in the situation for the SIS model, the parameters in the SEIS model with $y = kI$ can be uniquely identified if one of them is known.

3.4 Parameter Sensitivity

To analyze the effect of parameter uncertainty on quantities of interest, we present sensitivity indices, One-at-a-Time plots, and contour plots of POIs versus QOIs.

The quantities of interest are as follows:

- $R_0 = \frac{\beta}{\gamma}$
- Number of Infected at Time Step 2
- Total Infected

The basic reproduction number is defined as the expected number of secondary cases produced by a single (typical) infection in a completely susceptible population [8]. This is commonly used by public health professionals to see whether an epidemic is increasing ($R_0 > 1$) or decreasing ($R_0 < 1$).

3.4.1 Local and Extended Sensitivity Analysis

Given parameter estimates from model fit, we carry out local and extended sensitivity analysis to understand how uncertainties in parameters effect predictions of the QOIs. We first check the unscaled and relative sensitivity indices for local sensitivity, and create One-at-a-Time plots for extended sensitivity.

3.4.2 Global Sensitivity Analysis

Using assumptions of the underlying parameter distributions, we take a sample of parameter values and evaluate the resulting values for the QOIs to generate correlation plots. We evaluate these for global trends.

4 Results

To apply the methods that we have discussed in Section 2, now we would like to perform our algorithms step by step by following our descriptions in Section 3 and see how the methods work in practice.

4.1 Generating Data

We start with generating data from our ODE models. Parameter baseline values are given in Table 1. These values will also be used to check our estimation results.

Parameter	**Description**	**Baseline Value**	**Dimension**
β	transmission rate	0.5	time^{-1}
ω	incubation rate	0.3	time^{-1}
γ	recovery rate	0.1	time^{-1}
k	probability of reporting	0.7	None

Table 1: Parameter baseline values. Note that in epidemiology the probability of reporting is often called the reporting rate.

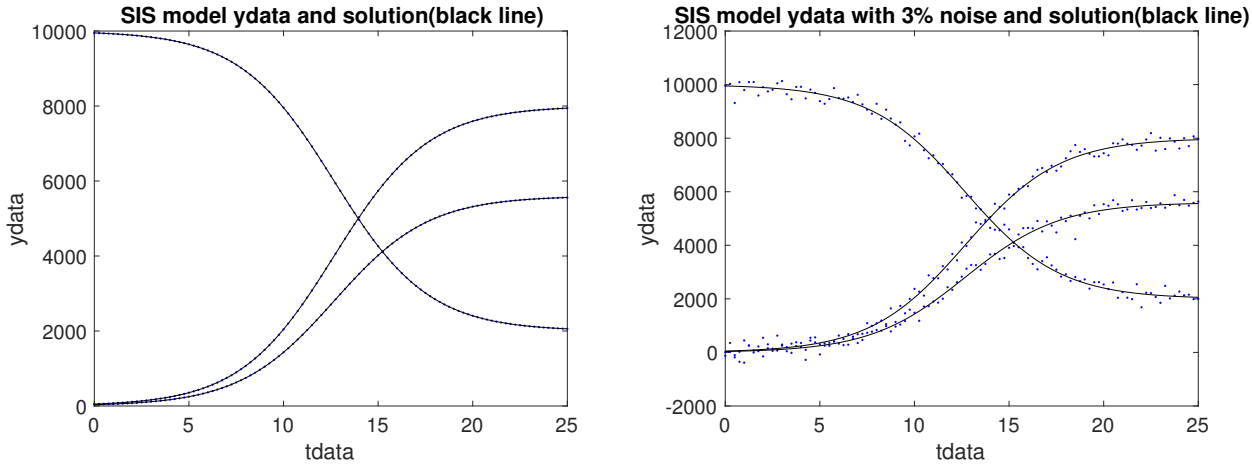

Figure 2: SIS model solution and data with no noise and 3% noise. Initial conditions: $S_0 = 9950, I_0 = 50, y_0 = 35, t = 25$ days.

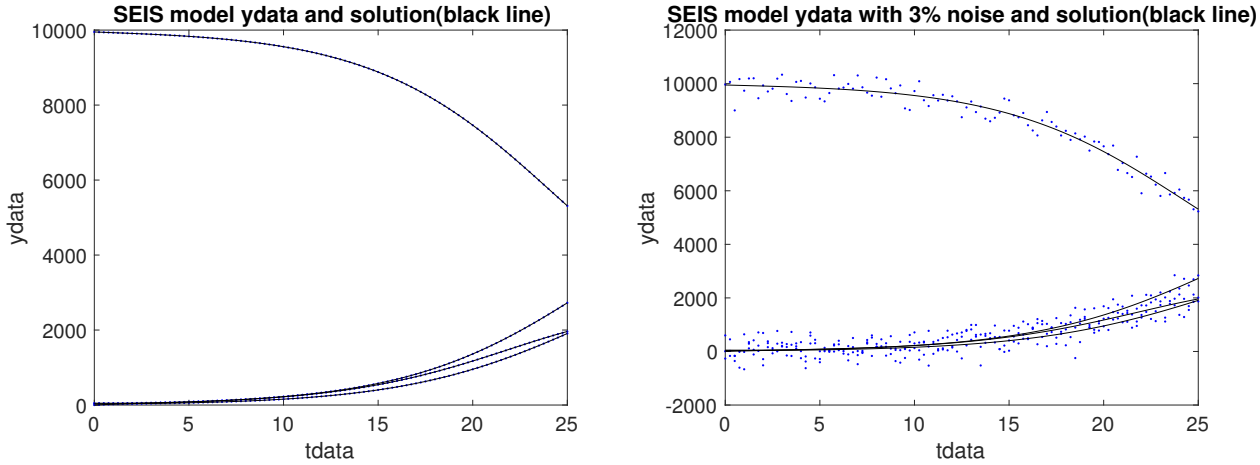

Figure 3: SEIS model solution and data with no noise and 3% noise. Initial conditions: $S_0 = 9950, E_0 = 0, I_0 = 50, y_0 = 35, t = 25$ days.

4.2 Parameter Estimation and Quantification

The following Boxplots visualize our models fitted to both clean and noisy data.

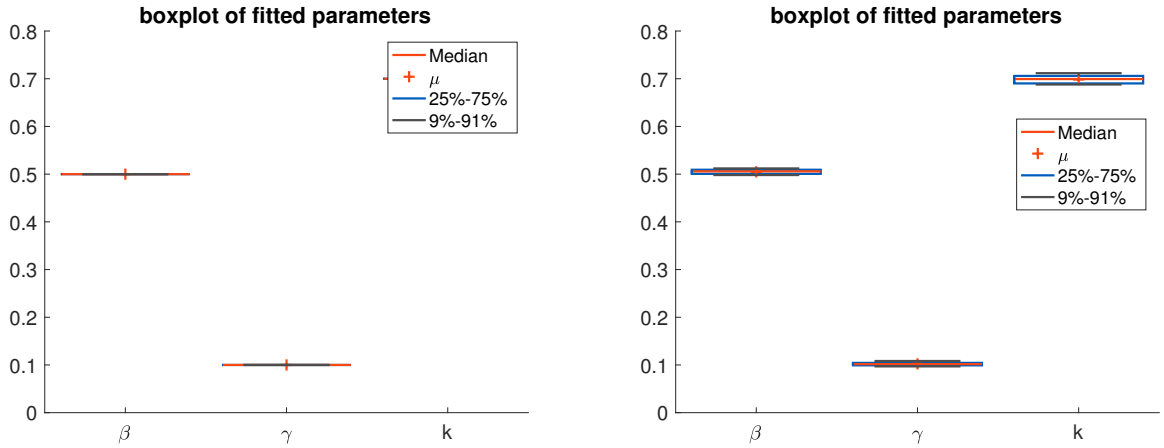

Figure 4: SIS model boxplot of the parameter fits from bootstrap. Left: data with no noise. Right: 3% noise.

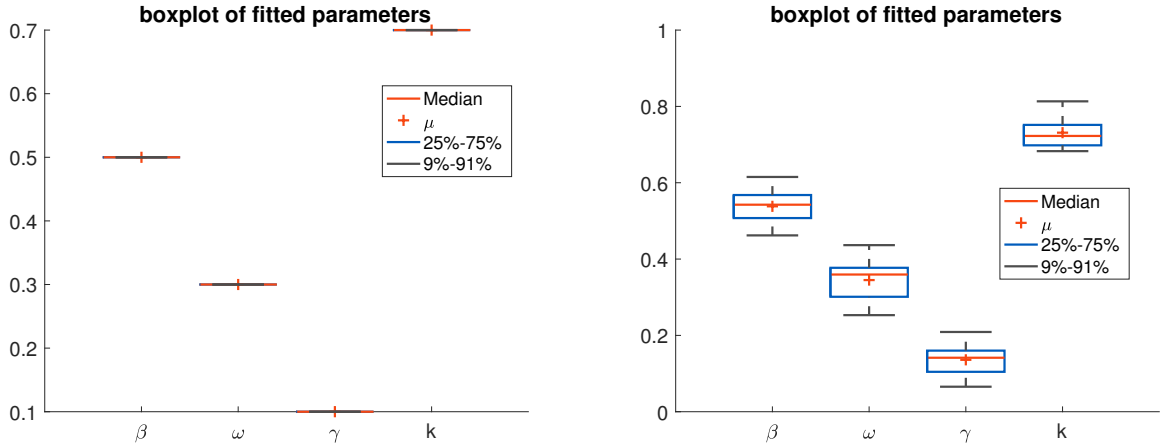

Figure 5: SEIS model boxplot of the parameter fits from bootstrap. Left: data with no noise. Right: 3% noise.

The following Table 2 shows our parameter estimation results along with the 95% confidence interval to quantify the uncertainties from bootstrap process.

Model	**Noise**	$\beta \pm 95\%CI$	$\omega \pm 95\%CI$	$\gamma \pm 95\%CI$	$k \pm 95\%CI$
SIS	None	0.5000±4.5669e-16	None	0.1000±4.4087e-16	0.7000±1.1329e-16
SIS	3%	0.50054±2.3163e-04	None	0.10052±2.2197e-04	0.69998±1.5630e-05
SEIS	None	0.5000±6.6388e-16	0.3000±6.0855e-16	0.1000±5.8281e-16	7e-01±7.4046e-17
SEIS	3%	0.53819±2.4149e-02	0.34497±2.8540e-02	0.13600±2.2468e-02	0.73119±1.9627e-02

Table 2: Uncertainty quantification for parameter estimates.

From the above table we see that our estimations match the baseline values pretty well for noise-free data. This is a sign showing that the program is functioning. With the 3% noise, we see that all of the parameter estimates are also comparably precise.

Correlation plots shown below examine the possible correlations between parameters. Some linear relations between β, γ and ω are shown especially in the SEIS model. All parameters appear to be approximately normally distributed.

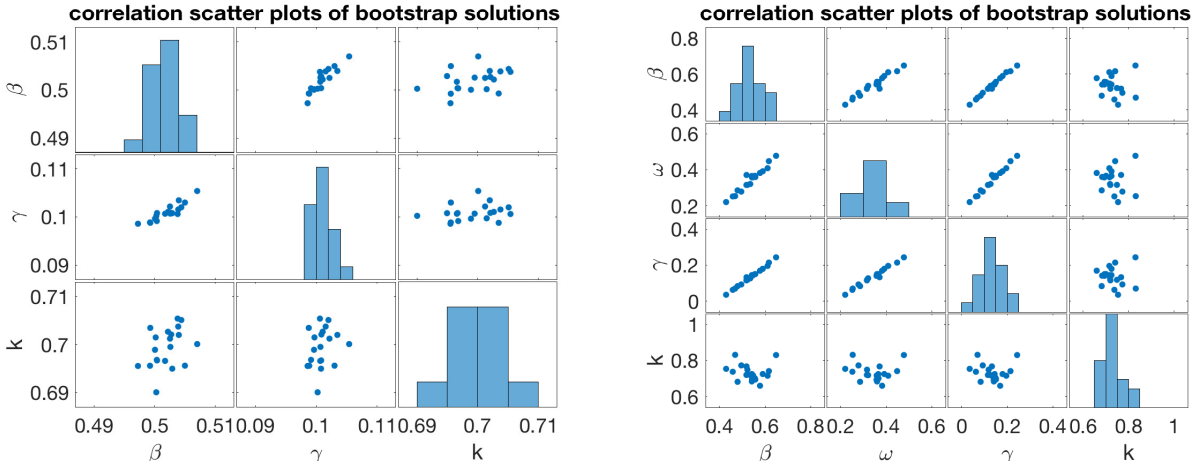

Figure 6: Correlation between parameters. Left: SIS model. Right: SEIS model.

4.3 Identifiability Analysis

We then check the local identifiability for the parameters by computing the eigenvalues of the Hessian matrix of the mean squared errors. Remember that we can confirm the local identifiability if the Hessian is positive definite. Eigenvalues associated with the SIS and SEIS models are listed in the following table.

Model	Noise	Eigenvalues
SIS	0	1.6953e+11, 3.8789e+09, 1.9142e+09
SIS	0.03	1.6515e+11, 3.8700e+09, 1.9327e+09
SEIS	0	8.3393e+10, 4.8049e+08, 7.4241e+07, 1.2190e+07
SEIS	0.03	7.8719e+10, 4.3191e+08, 7.3767e+07, 1.2574e+07

Table 3: Eigenvalues of Hessian matrix of mean squared errors of fitted models.

From the table above we see that all of the eigenvalues associated with the models are positive and therefore, the parameters are locally identifiable. We then extend the identifiability analysis by using profile likelihood. The following leave-one-out (LOO) plots show where the minima of the MSE are, which demonstrate the extended identifiability of the parameters.

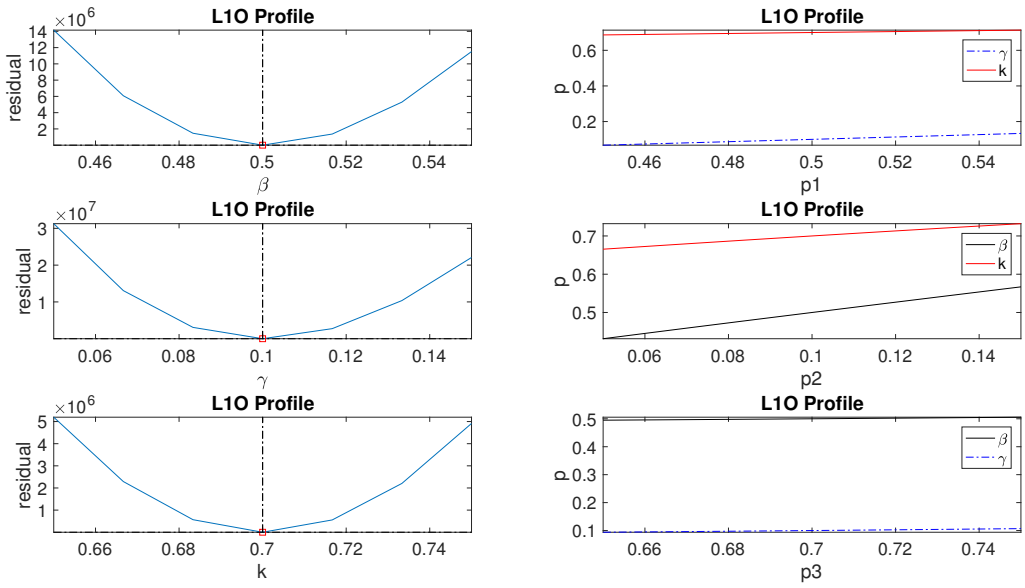

Figure 7: SIS leave-one-out profile analysis with no noise.

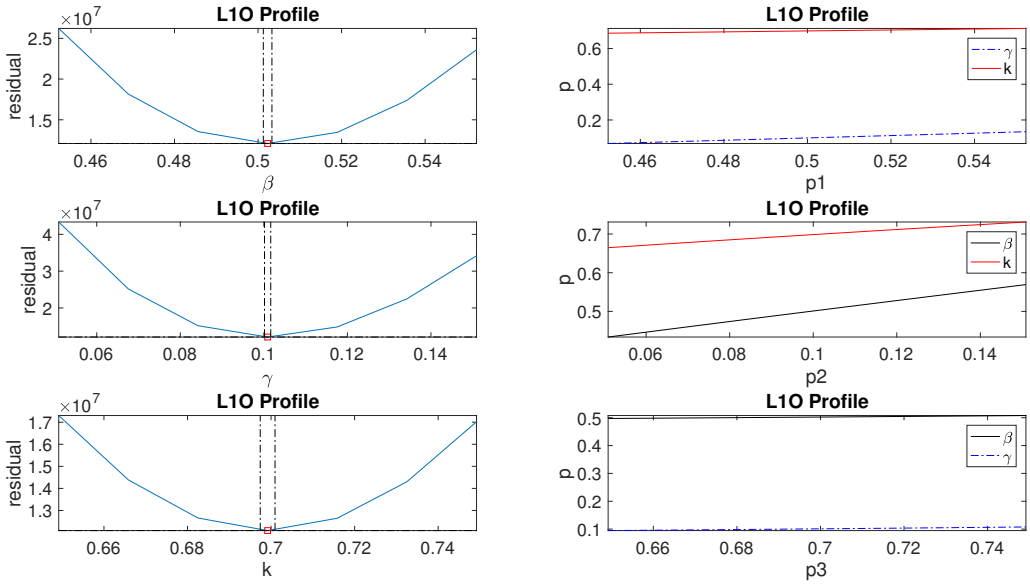

Figure 8: SIS leave-one-out profile analysis with 3% additive noise.

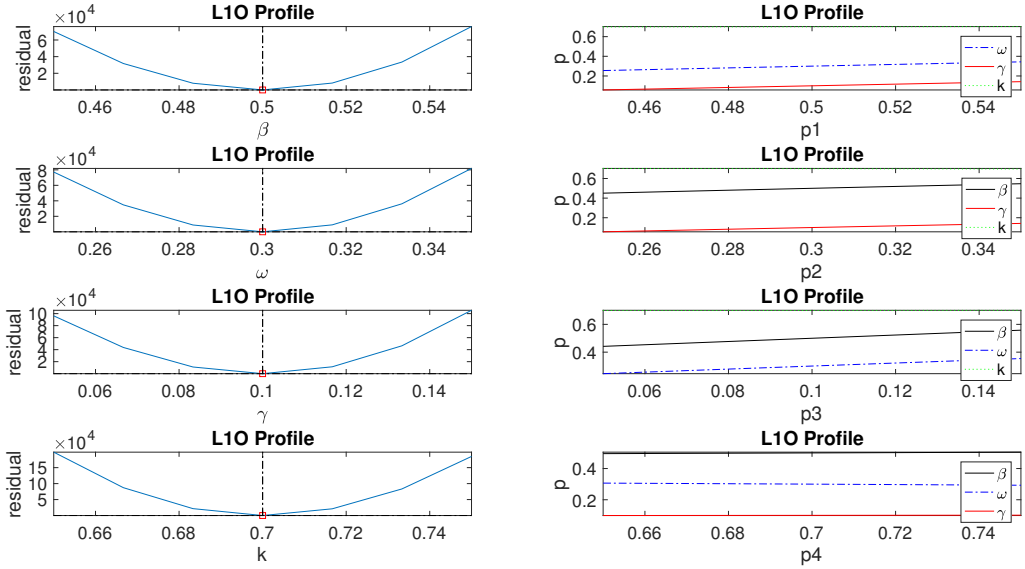

Figure 9: SEIS leave-one-out profile analysis with no noise.

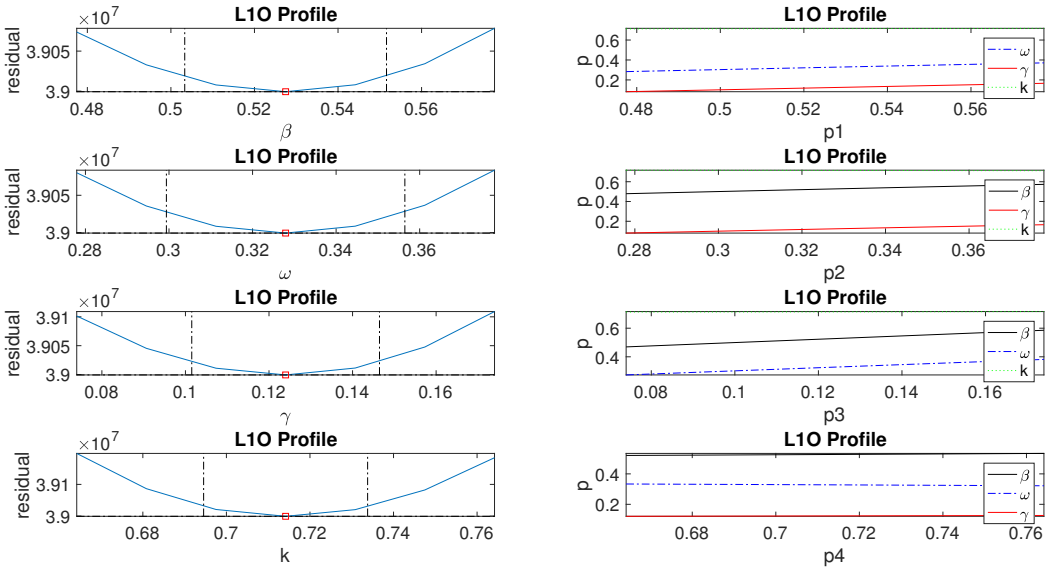

Figure 10: SEIS leave-one-out profile analysis with 3% additive noise.

We also generate two dimensional contour plots to check structural and practical identifiability properties. For all contour plots generated from our algorithms, we find visible structural identifiability issues for some of the parameters. The plots are consistent with our analytical proof presented in Section 3.3.2.

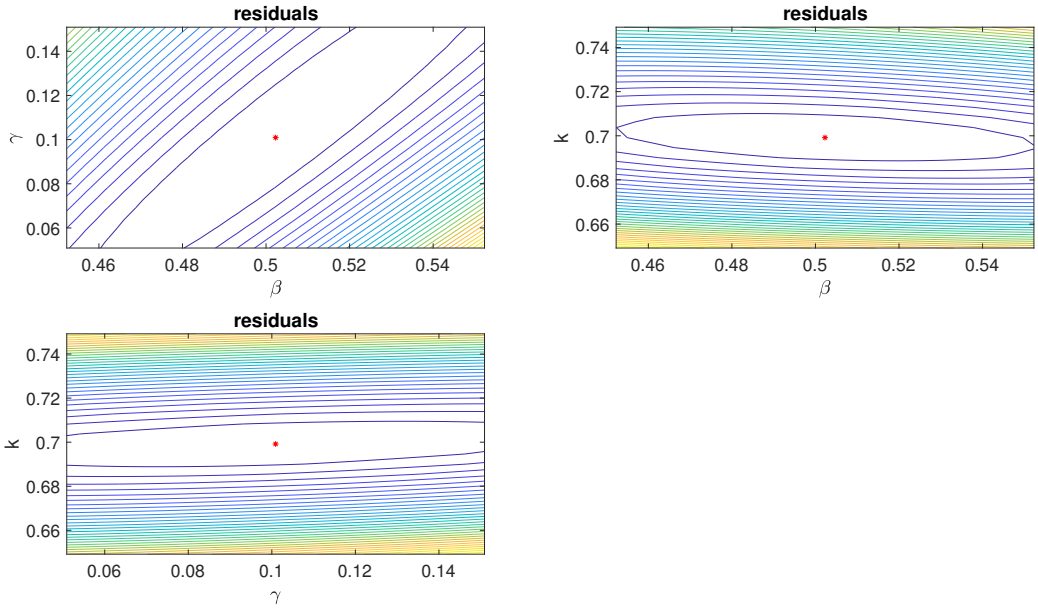

Figure 11: SIS two dimensional contour plots with noise.

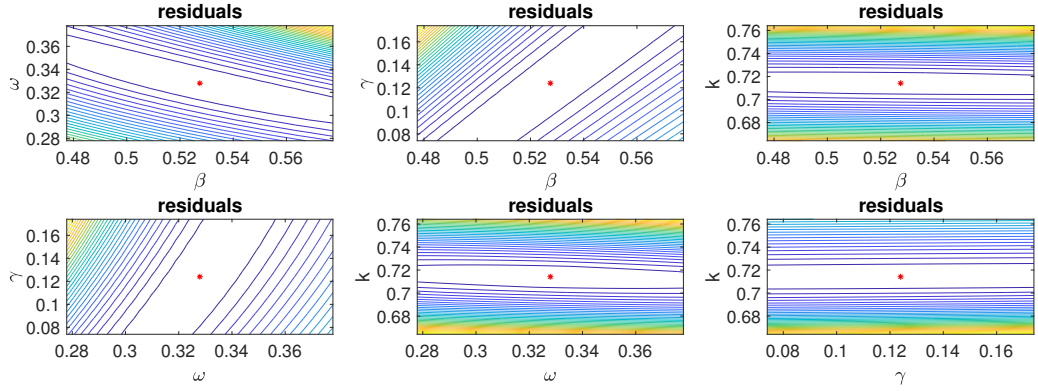

Figure 12: SEIS two dimensional contour plots with noise.

4.4 Sensitivity Analysis

We generate our sensitivity indices (unscaled and relative) to understand the impact of changes in parameters to quantity of interest values.

	β	γ
Total Infected	7.4884e+00	4.2558e+01
Infected($t = 2$)	5.0648e-04	-5.0656e-04
R_0	1.0000e+01	-5.0000e+01

Table 4: SIS Sensitivity Indices (unscaled).

	β	γ
Total Infected	6.0332e-01	6.8576e-01
Infected($t = 2$)	1.1165e+00	-2.2333e-01
R_0	1.0000e+00	-1.0000e+00

Table 5: SIS Relative Sensitivity Indices (scaled).

Without using the relative scaled sensitivity indices, we may falsely assume that γ has a larger effect on R_0 than β when in fact they have the same effect on this QOI. Continuing to focus on the relative sensitivity indices, we can see similarly that the effect of γ is more important for the early development of an epidemic, as evidenced by the QOI number of infected at time step $t = 2$. Both parameters appear to have an equal effect on the final total number of infected.

	β	ω	γ
Total Infected	5.2921e-01	-7.4346e-01	8.7134e-01
Infected($t = 2$)	1.5828e-05	-1.1388e-04	6.4687e-05
R_0	1.0000e+01	0.0000e+00	-5.0000e+01

Table 6: SEIS Sensitivity Indices (unscaled).

	β	ω	γ
Total Infected	9.1621e+00	-7.7229e+00	3.0171e+00
Infected($t = 2$)	1.2845e-01	-5.5451e-01	1.0499e-01
R_0	1.0000e+00	0.0000e+00	-1.0000e+00

Table 7: SEIS Relative Sensitivity Indices (scaled).

As expected, ω has no effect on R_0, since it is not included in the calculation of this QOI. For total infected, β has the largest effect, followed by ω, then γ. For the number of infected in the early development of the epidemic, ω has the strongest effect on the QOI number of infected at time step $t = 2$.

Then we generate One-at-a-Time Plots, shown in Figures 23, 14, 24, 13, **??**, and 15. For the parameters, we assumed they all followed a triangle distribution. The distributions were constructed according to the following:

	Mode	Min	Max
β	0.50	0.40	0.60
ω	0.30	0.10	0.40
γ	0.10	0.00	0.20

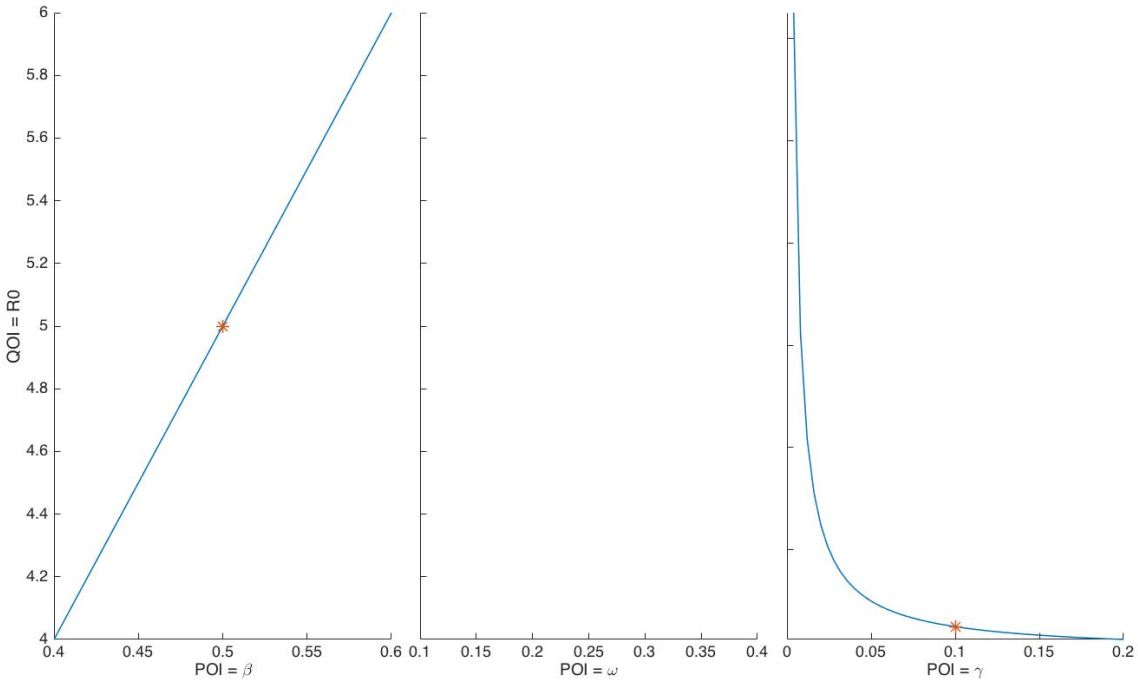

Figure 13: SEIS One-at-a-Time analysis with no noise for R_0.

We can see that β has a linear relationship with R_0, while γ has an inverse relationship. For the SEIS mode, we see the same figures as ω has no effect on R_0.

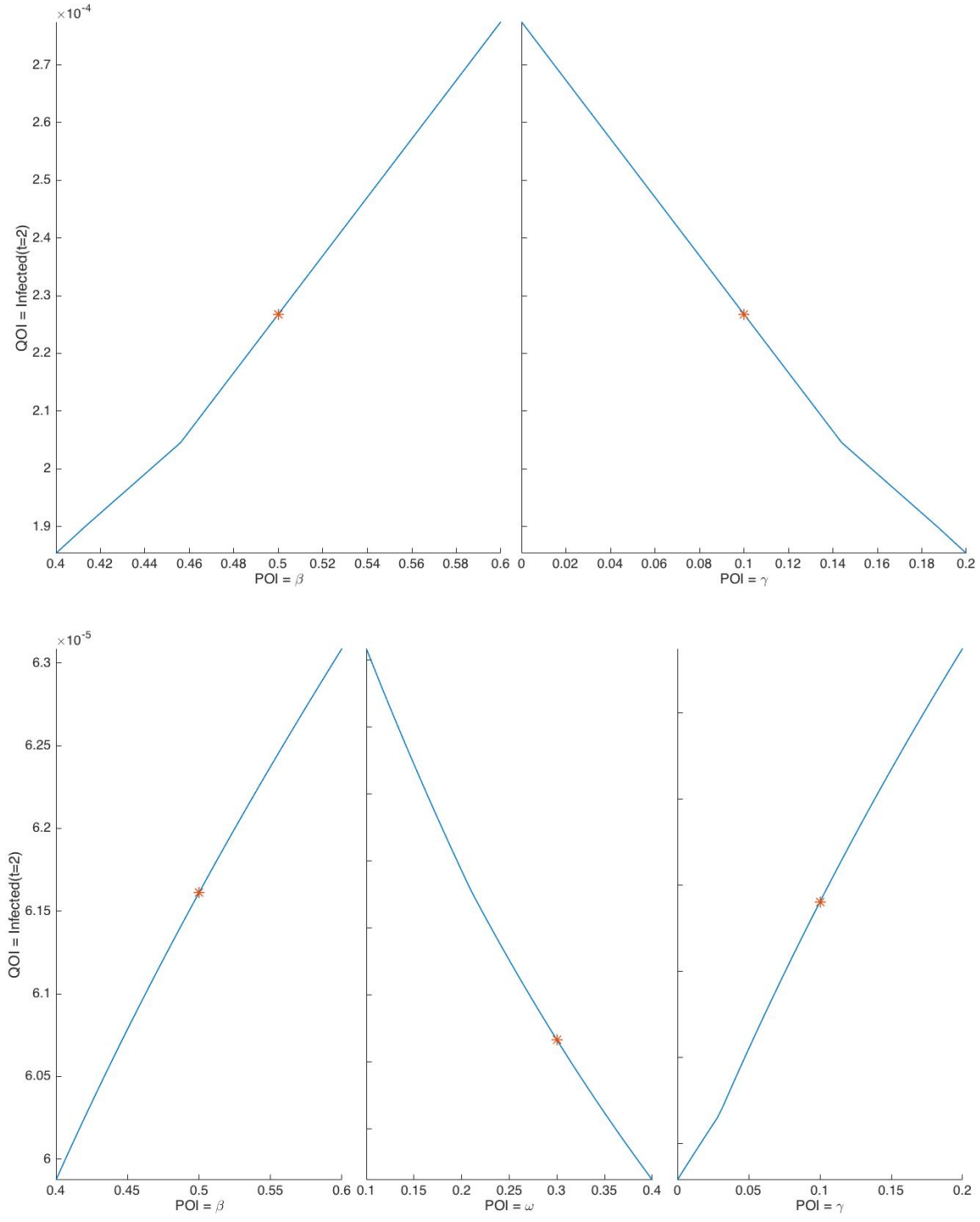

Figure 14: One-at-a-Time analysis with no noise for Infected (t=2). **Top:**SIS **Bottom:**SEIS

For the both models, β, ω, and γ appear to have a nearly linear relationship with the number of infected at time step $t = 2$. For both models, β has a positive relationship with

initial number of infected. This corresponds to an increasing number of susceptible people entering the infection stage.

For the SIS model, γ has a negative relationship with the number of infected at time step $t = 2$. This indicates that as individuals recover faster and leave the infection stage, the initial number of infected drops. Conversely, for the SEIS model, γ has a positive relationship with the initial number of infected individuals and ω has a negative relationship with the initial number of infected individuals. As the incubation period becomes longer, less individuals enter the infected compartment and therefore the number of total infected individuals decreases.

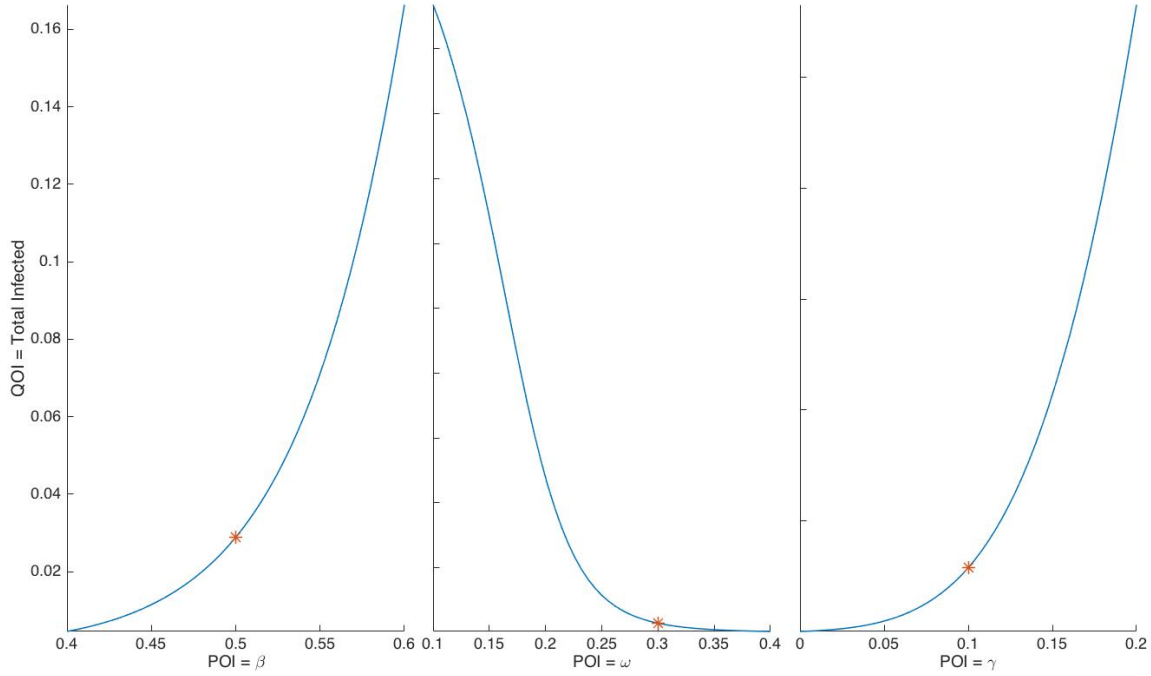

Figure 15: SEIS One-at-a-Time analysis with no noise for Total Infected.

For the SEIS and SIS models, β and γ appear to have a strong positive relationship with the total number of infected. For the SEIS model, ω has a negative relationship with the total number of infected, indicating that long incubation periods decrease the overall number of infected individuals in a population.

We can then look at the correlation between POI vs POI, POI vs QOI, and QOI vs QOI to further assess trends, shown in Figures 16, 25, 17, 26, and 18.

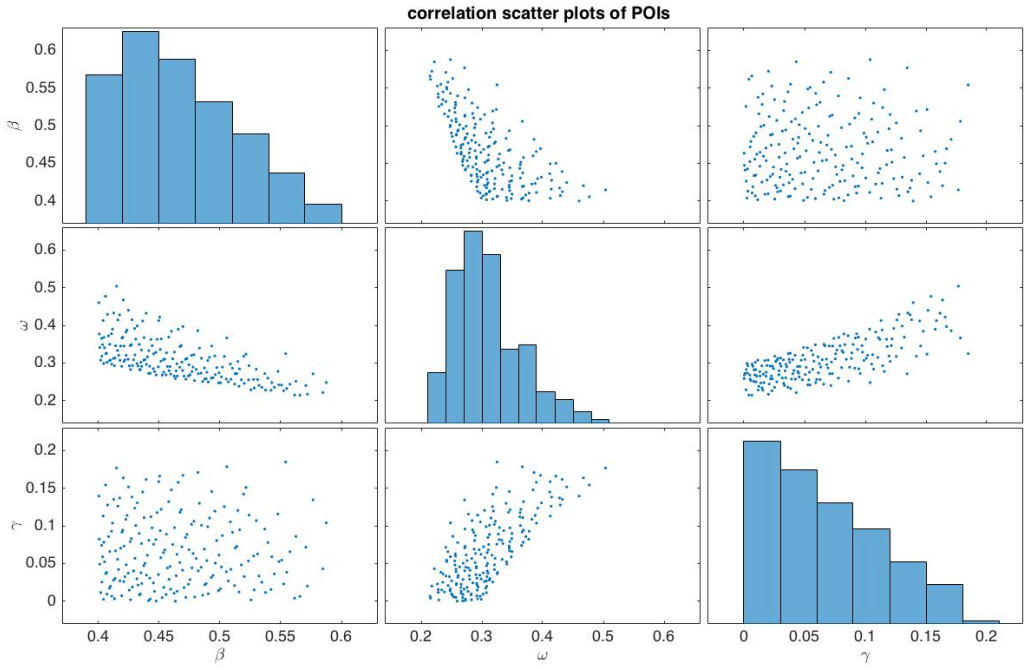

Figure 16: Correlations between parameters of interest and parameters of interest for the SEIS model.

While there doesn't appear to be a trend between the parameters and themselves for the SIS model which only involves the interactions of β and γ, the SEIS model shows there may be a relationship between ω and γ, and ω and β.

All the parameters have triangle distributions, as predetermined by the authors. A better method would be to use the distributions for the parameters found from the bootstrap analysis.

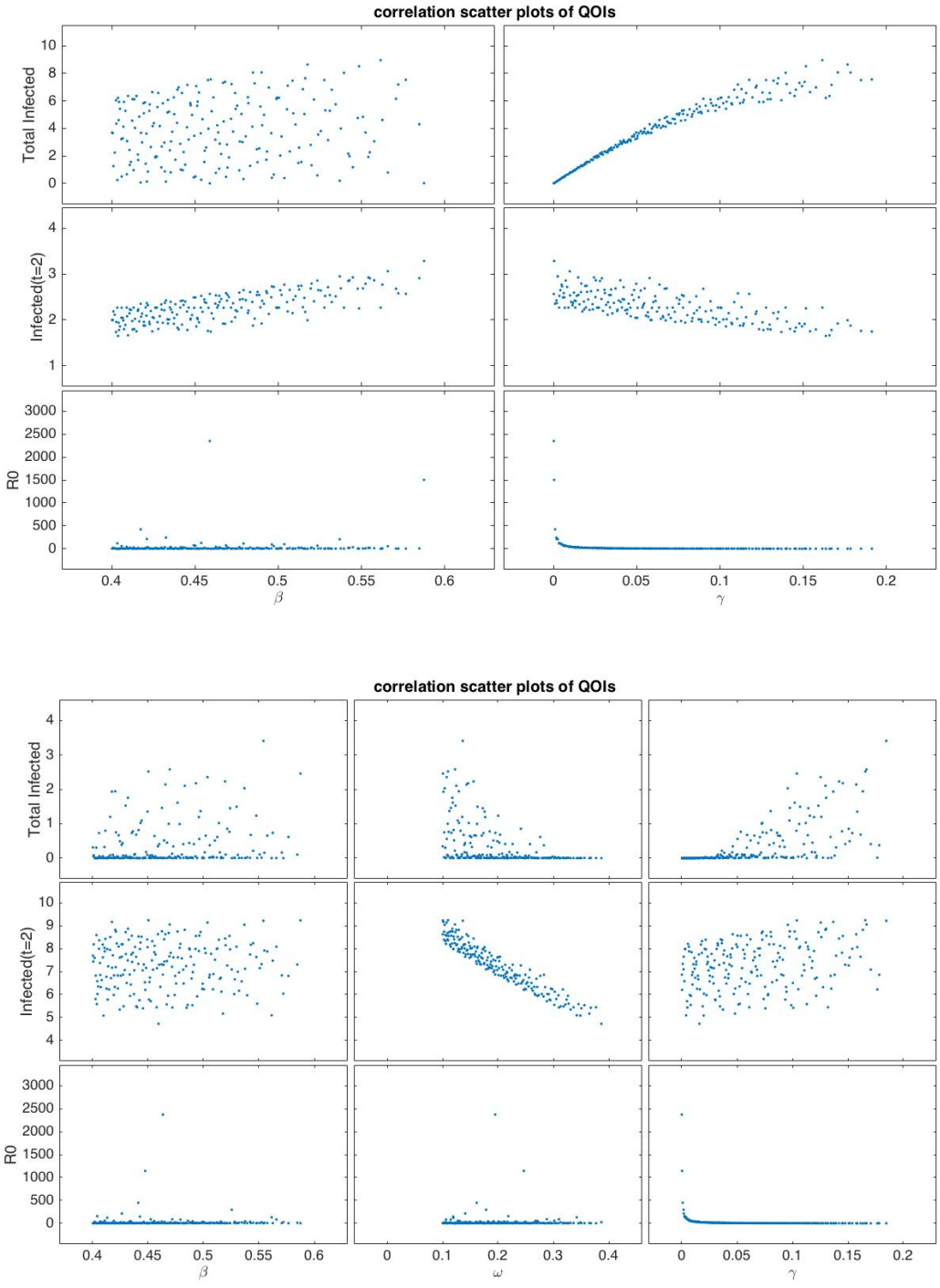

Figure 17: Correlations between quantities of interest and parameters of interest. **Top:** SIS **Bottom:**SEIS

For the SIS model, there is a strong positive trend between γ and the total number of

infected. The inverse relationship observed between γ and R_0 is observed again in this plot for SIS and SEIS models. For the SEIS model, ω appears to have a negative trend with number of infected at time step $t = 2$, as observed with the One-at-a-Time analysis.

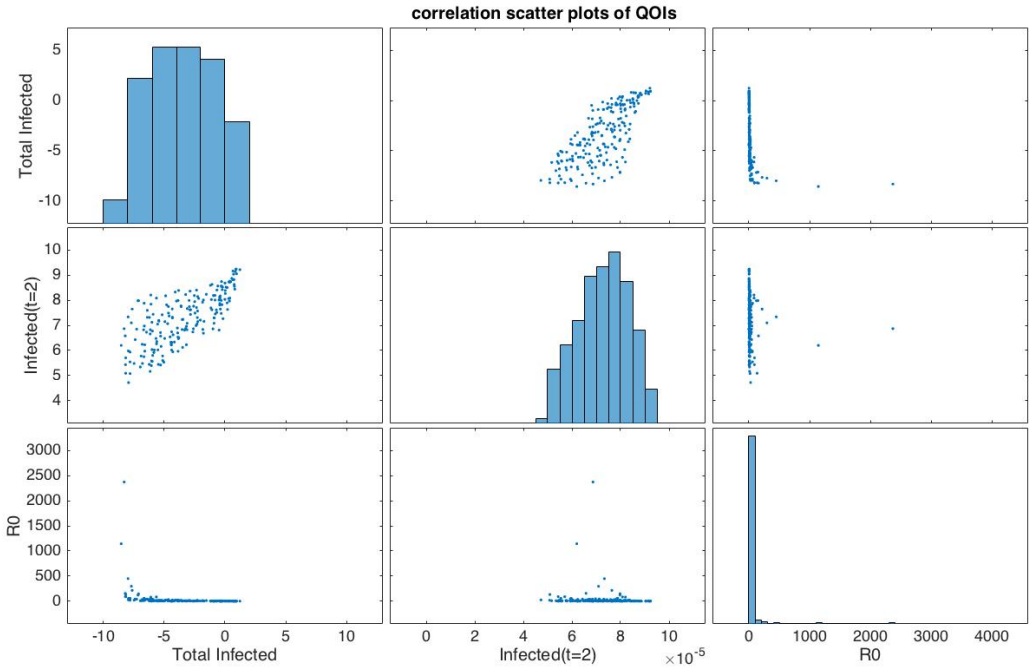

Figure 18: Correlations between quantities of interest and quantities of interest for the SEIS model.

For the SIS and SEIS models, we observe that there appears to be an inverse relationship between the total number of infected and R_0 is preserved. R_0 and Total number of infected appear to have uniform distributions for both the SIS and SEIS models, and there appears to be a trend between the number of infected at time step $t = 2$ and the total number of infected individuals.

The number of infected at time step $t = 2$ appears to be normally distributed, while the total number of infected is approximately uniform. The distribution for R_0 is unclear, and has a strong skew to the right. This indicates that under certain conditions it can have a very high value, though the majority of the values appear to be less than 500.

4.5 Final Fit and Prediction

Finally, we conclude our study by showing the extrapolated plot.

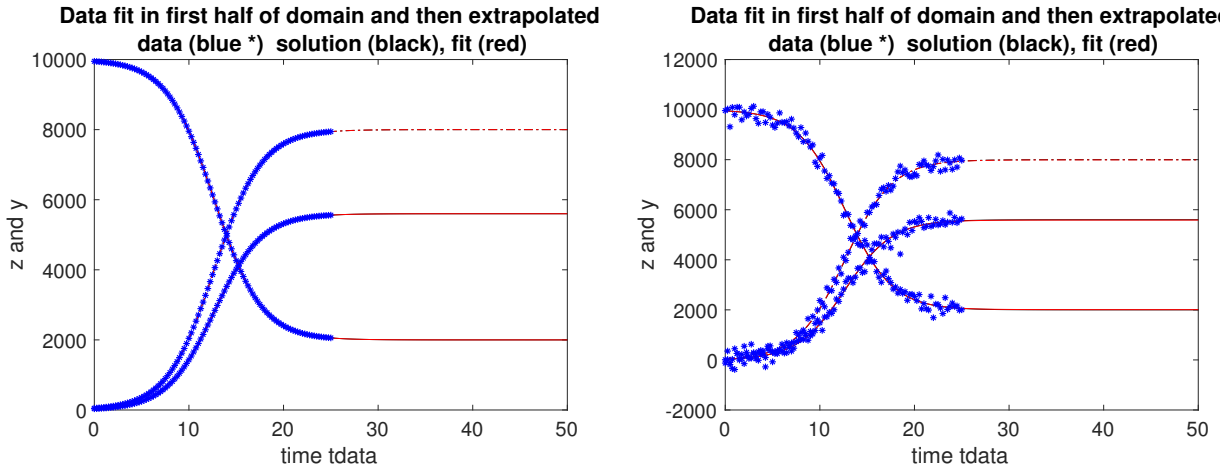

Figure 19: SIS model extrapolation with no noise and 3% noise.

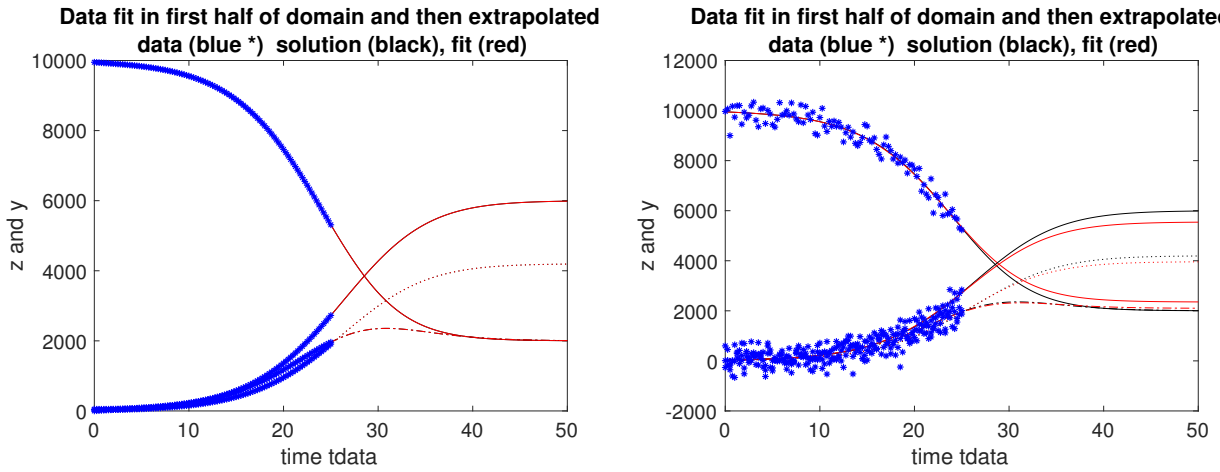

Figure 20: SEIS model extrapolation with no noise and 3% noise.

5 Discussion

In this report, we went over the main procedures for parameter estimation and identifiability. We created test data by solving the SIS and SEIS models with the baseline parameter values which we believe are our best guess. The advantage of using fake data is that we are able to make sure that our algorithm is working correctly before applying it to real-world noisy data. We then added noise to the clean data for testing. After data become available, we took the following actions:

- First, we fit the models to data by minimizing the sum of the squared residuals. A detailed analysis of the fit was then followed. We performed several procedures to

analyze the model fit, including comparing the estimates with baseline values and plotting the residuals against the data to detect any observable structure. We also discussed bootstrap technique which consists of resampling data to gain a distribution of the parameter estimation. Since general bootstrap only works on non-correlated data sets, we explained the blocked bootstrap for resampling data that are time or order sensitive. Uncertainty quantification for the estimates was extracted from the distribution of the parameter estimates by computing the standard error on a 95% level.

- Secondly, we explored the methods for determining the identifiability of model parameters. For practical identifiability, we used several numerical approaches including Hessian matrix for local identifiability, profile likelihood for extended identifiability, and low discrepancy sampling for global identifiability. In the case when data are not directly related to the quantities of interest, we converted the models into single input-output differential equations and then investigated the structural identifiability of the parameters by applying the approach of differential algebra. It is worth mentioning that all of the methods that we have discussed can be used at the same time to determine the parameter identifiability from different aspects.

- Thirdly, we discussed the sensitivity analysis for checking the impact of the parameters of interest to the quantities of interest. Again we used the SIS and SEIS epidemic models to illustrate the method.

References

[1] S. Audoly, G. Bellu, L. D'Angiò, M.P. Saccomani and C. Cobelli, Global Identifiability of Nonlinear Models of Biological Systems, *IEEE Trans. Biomed. Engi.*, Vol. **48**, No. 1: 55–66, 2001.

[2] A. Ben-Israel and T.N.E. Greville, *Generalized Inverses: Theory and Applications*, 2nd ed., New York, NY: Springer, 2003.

[3] B. Efron, Bootstrap Methods: Another Look at the Jackknife, *Annals Statistics*, Vol. **7**, No. 1: 1–26, 1979.

[4] M.D. McKay, R.J. Beckman and W.J. Conover, A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, *Technometrics*, Vol. **21**, No. 2: 239–245, 1979.

[5] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, p. 29, 1992.

[6] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller and J. Timmer, Structural and Practical Identifiability Analysis of Partially Observed Dynamical Models by Exploiting the Profile Likelihood, *Bioinformatics*, Vol. **25**, No. 15: 1923–1929, 2009.

[7] I. Sobol, Sensitivity Analysis for Nonlinear Mathematical Models, *Math. Modeling Comp. Exp.*, Vol. **1**: 407–414, 1993.

[8] Jones, J. H., (2007) Notes on R0. Department of Anthropological Sciences, Stanford University.

6 Appendix

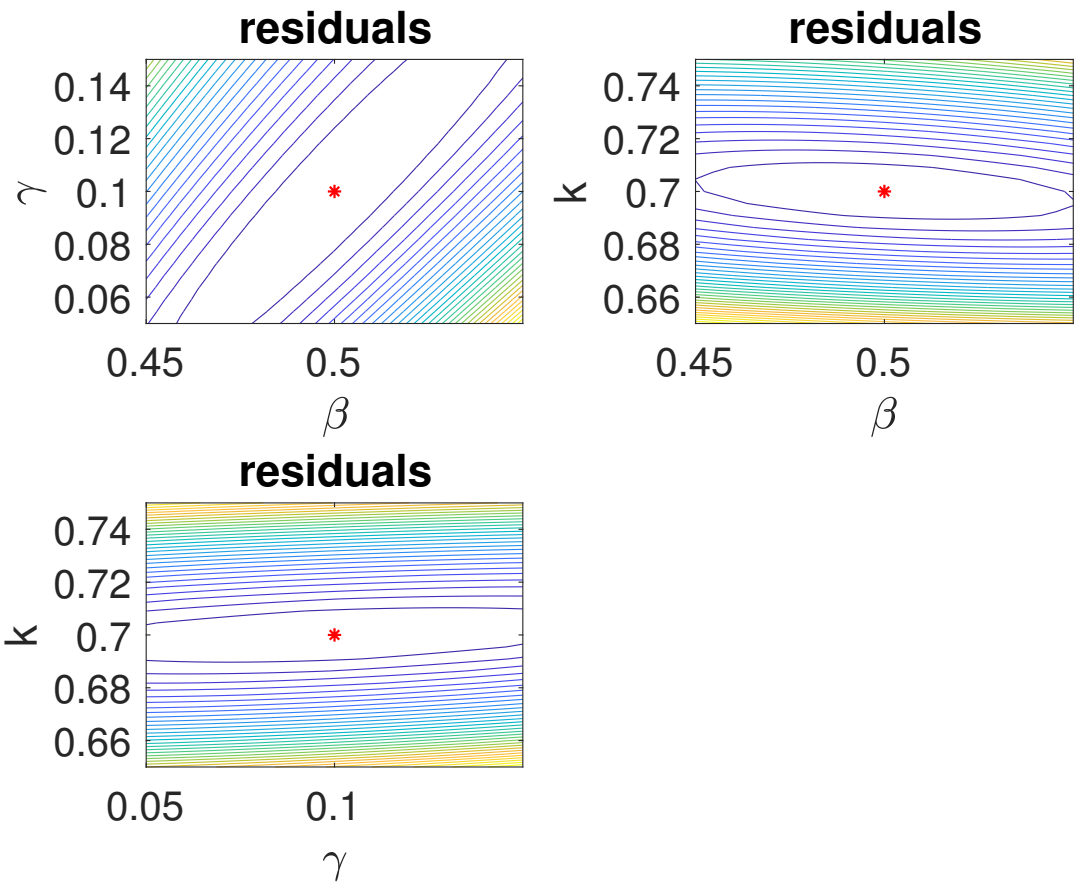

Figure 21: SIS two dimensional contour plots.

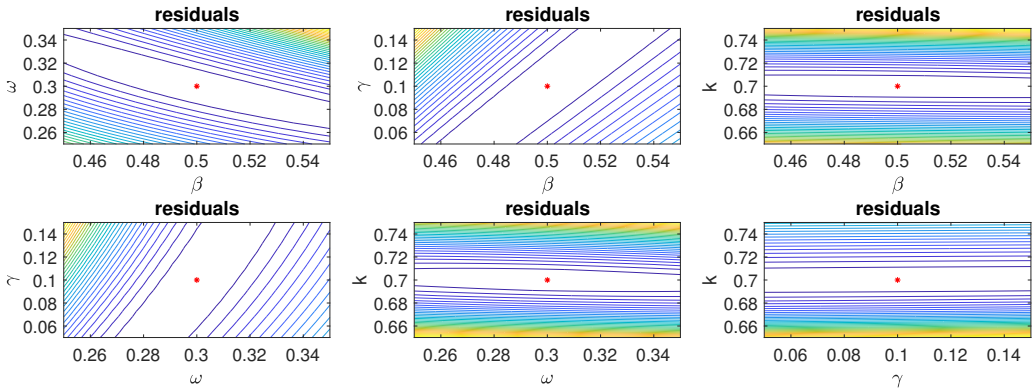

Figure 22: SEIS two dimensional contour plots.

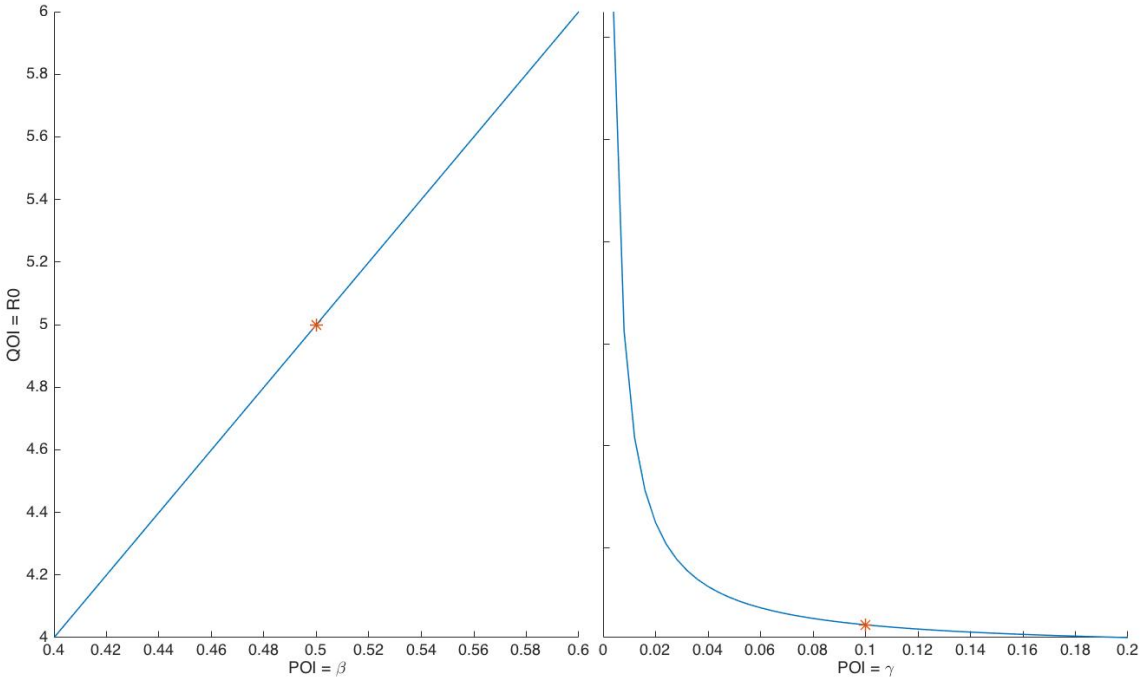

Figure 23: SIS One-at-a-Time analysis with no noise for R_0.

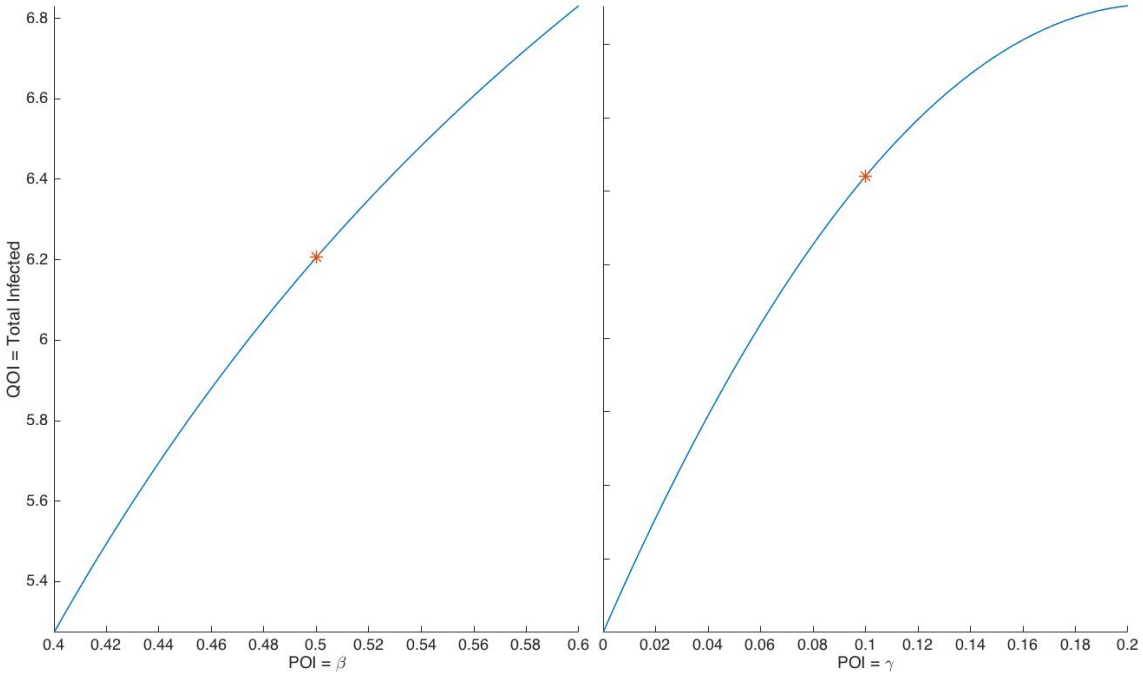

Figure 24: SIS One-at-a-Time analysis with no noise for Total Infected.

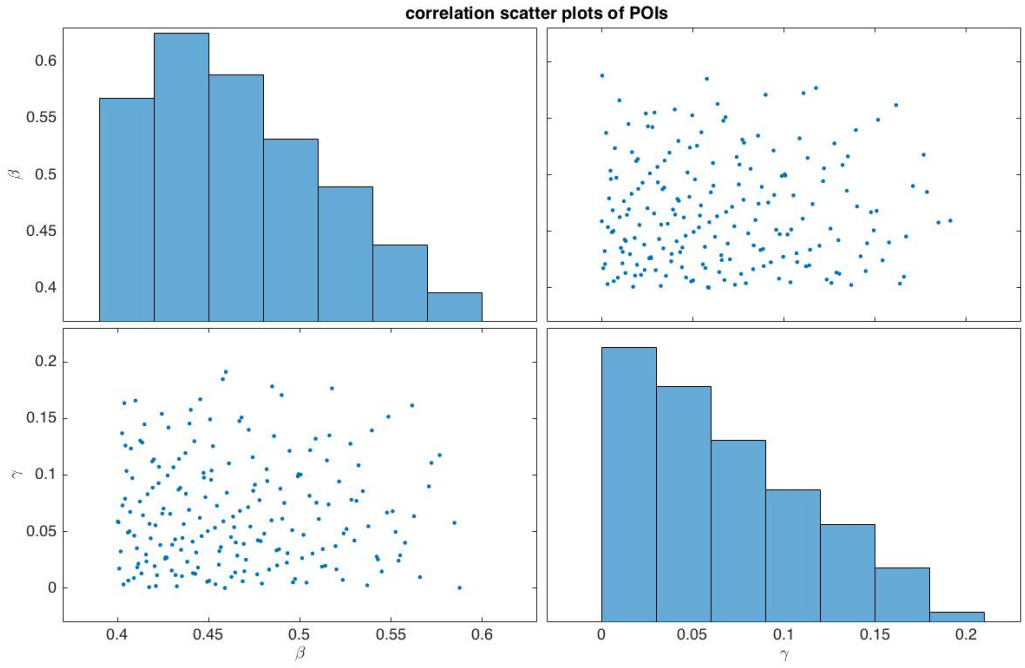

Figure 25: Correlation between parameters of interest and parameters of interest for the SIS model.

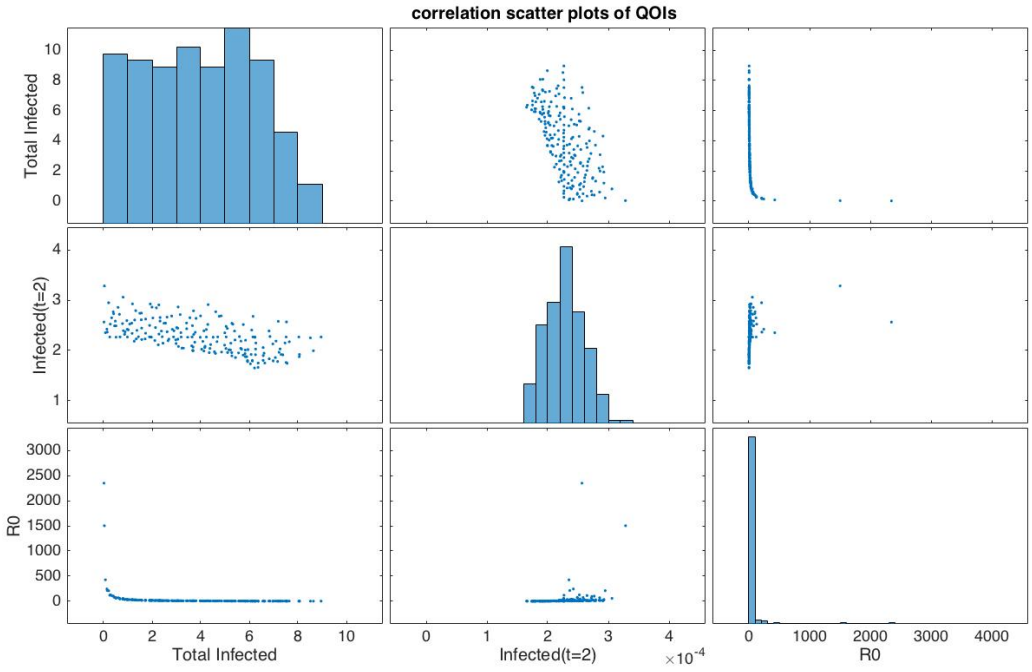

Figure 26: Correlation between quantities of interest and quantities of interest for the SIS model.