Todos

University of Auckland

Thesis Proposal

Parametrisation and Identification Methods for Epidemic Systems

David Wu

supervised by

Dr. Oliver Maclaren

and

Dr. Vinod Suresh

Department of Engineering Science

November 2019

Abstract

Although the field of mathematical epidemiology is well-trodden, there still exists questions on how to derive useful inferences from mathematical models. Conclusions drawn from mathematical analysis are useful for describing overarching behaviours of the system, but these results are often not particularly useful for prediction. For those purposes, estimation of the parameters and state of the model is required. Typically, maximum likelihood methods that involve evaluating the least squares error between the solution of the forward model with the data. However, this process is subject to instability and identifiability problems that can be hard or expensive to detect.

The aim of this project is to take a known method in functional data analysis for fitting ordinary differential equations to data, and apply it in the field of epidemic modelling. Utilising existing automatic differentiation tools, initial analyses of toy models and real systems have already been performed. We also consider the method's usefulness for solving identifiability problems and develop ways of performing uncertainty quantification. In future, we are interested in the extension of this method to multi-dimensional models, in particular to answer questions about epidemic propagation. Additionally, we show that this method is a special case of a more general hierarchical Bayesian model, allowing comparisons between different statistical frameworks, and ways of combining their properties.

Contents

List of Figures

1 Background

1.1 Problem Overview

In much of mathematical biology, real-world biological systems are often idealised in the modelling process. For understanding the overarching dynamics and structure of such biological systems, this idealisation is both inevitable and invaluable. However, this can lead to problems when trying apply the results of theoretical analysis to real-world scenarios.

Modelling is a powerful tool for answering problems about the behaviour of systems. For epidemiological applications, these problems fall roughly into two categories: prevention and control. Though analytical results, such as the ones elaborated by **rass2003**, **capasso1993**, or **diekmann2000**, are useful in providing general guidelines for asymptotic behaviours, the transient dynamics may be more useful to practitioners. Therefore a more specific prediction must be able to be made.

Such predictions can be made by estimating the parameters of the mathematical model, and using these estimates to generate an estimate of the state of the system in the past and into the future. From this, control strategies can be implemented and simulated, creating meaningful predictions, and allowing for informed decisions to be made.

This thesis hopes to explore the use of optimisation methods to perform the parametrisation of mathematical models, as a complementary method to typical Bayesian methods. In particular, we are interested in fitting some (epidemiological) model of the form:

$$\mathcal{D}u = f(u;\theta) \tag{1.1}$$

to some data y, where u is the state variable, $\mathcal{D}$ is some differential operator, f some arbitrary function, and θ a finite-dimensional vector of parameters. We note that that this model can be solved to produce a solution for u, if given θ, initial, and boundary conditions. We term this the *forward problem*.

An example of a model of this form is the ordinary differential equation:

$$\frac{du}{dt} = f(u;\theta) \tag{1.2}$$

which is the typical form of compartmental models in mathematical epidemiology. These compartmental models of epidemiology have a relatively long history, dating back to the seminal **kermack1927** papers. The general idea is to partition the population into several compartments that describe their state of infection. The simplest type, the SIR model as presented in (**??**), is still in wide use as a baseline model that some practitioners use to fit their data to.

$$\begin{aligned} \dot{S} &= -\beta SI/N \\ \dot{I} &= \beta SI/N - \alpha I \\ (\dot{R} &= \alpha I) \\ N &= S + I + R \end{aligned} \tag{1.3}$$

Then, to make predictions using this model, we would wish to determine the values of the parameters α and β, as well as initial conditions of the state variables S, I and R (we could assume that $I(0) = 1$, $S(0) = N - 1$, $R(0) = 0$ to reduce this to estimating just one more value). This would then allow us to solve the problem forward in time for prediction, or perturb the system to model control. We term this estimation of the parameters the *inverse problem*.

1.2 State of the Art

The estimation of the parameters (also known as model calibration) for such models is usually solved as a consequence of estimating the true state of the system, assuming some form of error. A common approach is to assume a normally distributed, additive error in the observation of the system**chowell2017**, **dsilva2017**, **smirnova2016**, **smirnova2017**. Taking a log-likelihood of this error model gives what can be termed *trajectory matching* **ramsay2017**, where the sum of squares error between a model-generated trajectory and the data is minimised. This is also known as the nonlinear least-squares approach, and can be written as

$$\theta_{opt} = \inf_{\theta} \|y - u(\theta, u_0)\|^2 \tag{1.4}$$

where y is the observations, u is the estimated state from integrating the forward model, θ the parameters, and u_0 the initial state of the system (often also estimated). Note that in solving the inverse problem, that the forward problem must be solved multiple times.

Unfortunately, this method displays some problems typical of nonlinear, nonconvex objective functions. Namely, the parameter estimate recovered is sensitive to the initial iterate, or *unstable*. Some methods have been proposed for the selection of this initial iterate **dattner2015**; typically though, this problem is side-stepped by simply performing the optimisation many times with a large number of initial iterates **gabor2015**, **goeyvaerts2015**, **dsilva2017**.

This problem is exacerbated in biological systems, where the model is generally inadequate for capturing the dynamics of the observations **brynjarsdottir2014**, or we have low confidence in our model. There are explicit ways of formulating a model which can account for this sort of error **morrison2018**, but the forward problem is generally computationally expensive to solve. Even without this additional complexity, the forward problem can be expensive to solve, for example, if there are many states, i.e u is high-dimensional, such as in PDE models. Indirect methods can be used to solve this problem, by performing least-squares on the vector field, instead of the trajectory, and avoiding solving the forward problem **ramsay2017**. However, these methods require full observability of all state variables, or make very strong assumptions about the trajectories of the unobserved states. This makes them difficult to use in practice, since typically data is only measured for a subset of the states of the system. Methods have been developed to solve this partial observation problem, such as the all-at-once method in the geophysics literature, or the generalised profiling (also parameter cascading) method in the statistical literature, the latter of which we explore in detail in Section **??**. These attempt to estimate the underlying state and the parameters at the same time, which

One other problem with the above optimisation methods is that they require strong assumptions about the form of error in the model in order to quantify the uncertainty, such as having only independent, additive, Gaussian observation noise. To address this, the standard tool is Bayesian analysis. This consists of computing the conditional probability of the parameters given the data, from a likelihood model (probability of the data conditioned on the parameters) and a prior on the parameters. Typically, this is not an analytically tractable problem, due to the probability

distributions used, or the form of the likelihood model. Thus, numerical methods are used to generate a set of samples that converge onto the posterior distribution, allowing it to be characterised. Sampling methods such as Monte Carlo Markov Chain are now use extensively for this purpose. However, these methods are computationally expensive as they require a high number of forward models solves in order to generate the sample distribution.

A more thorough review of the literature and associated theory is provided in the attached literature review **litrev2019**, along with some description of epidemiological models.

2 Scope

The objectives of this research project are to:

1. Develop a reusable package for the parameterisation and identifiability analysis of epidemic models, based on generalised profiling
2. Analyse the extension of the methods to multidimensional problems
3. Analyse a range of different epidemiological models including
    - 2019 measles outbreak in New Zealand
    - Relations with and imbued immunity from the 2017 mumps outbreak
4. Explore the Bayesian equivalents for this method, and determine efficiency gains from the optimisation simplifications

The following tasks will be undertaken as a part of the proposed research:

1. Construct a classification of relevant epidemiological models and their applicability to specific diseases
2. Develop a package for fitting one-dimensional-domained dynamic models
3. Extend the package to multi-dimensional domains
4. Collate data for relevant disease outbreaks and controls
5. Perform parameter estimation studies and hypothetical modelling of relevant disease outbreaks and controls
6. Implement the "Bayesianisation" of the method
7. Compare the uncertainty of estimates under the two statistical frameworks
8. Explore hybrid/multi-phase methods that blend the two frameworks

3 Methodology

3.1 Generalised Profiling

The generalised profiling method was introduced by **ramsay2007** as a way of parametrising differential equation models. It is a relaxed version of the trajectory matching problem, through the use of a spline approximation to the true state:

$$\mathcal{L}(\theta, c, y) = \|y - c\Phi\|^2 + \rho\|\mathcal{D}(c\Phi) - f(c\Phi, \theta)\|^2 \tag{3.1}$$

In the literature**ramsay2007**, **hooker2011**, **campbell2013**, **xun2013**, the above objective is used in an inner loop to estimate c conditioned on a proposed θ and given y:

$$\hat{c}(\theta) = \inf_c \mathcal{L}(c|\theta, y) \tag{3.2}$$

This is then used to perform least-squares optimisation in the "outer" optimisation:

$$\theta_{opt} = \inf_\theta H(\theta) \coloneqq \inf_\theta \|y - \hat{c}(\theta)\Phi\|^2 \tag{3.3}$$

This allows for the standard properties of ordinary least squares to be used on the outer objective to recover confidence intervals.

One of the underlying problems of the generalised profiling method is that the solution is dependent on the regularisation parameter ρ. If the model f is linear, then those properties can be exploited in order to derive an explict form of the generalised cross validation criterion, which can then be used to select ρ **ramsay2017**. However, most biological models do not possess this property. This means that other methods must be used.

Interestingly, the objective function bears striking resemblance to the regularised inverse problem:

$$\mathcal{L}(\theta|y) = \|y - A(\theta)\|^2 + \lambda R(\theta) \tag{3.4}$$

This leads to the interpretation of the generalised profiling method as a data interpolation problem regularised by a model fitting penalty. Using this framework we can repurpose methods of determining regularisation parameters for the choice of the tuning parameter in the generalised profiling method. A priori methods, such as the Morozov discrepancy principle, can be used if the magnitude of the least-squares error is known, but this is rarely the case in practice. A posteriori methods, such as the L-curve criterion**hansen2000**, cross validation**picard1984** or the Lepskii balancing principle**mathe2006** analyse the behaviour of the objective as λ is varied. All these methods have some ad hoc element - the L-curve criterion assumes some trend of regularisation trade-off to determine a regulariastion parameter; cross validation is typically expensive, so its approximations are performed with ad hoc termination; the balancing principle chooses arbitrary threshold functions to determine the regularisation parameter. We choose to use the L-curve criterion for its simplicity, and for its ability to extend to the determination of multiple regularisation parameters **belge1998**, **belge2002**. Furthermore, this method also corresponds in spirit to the technique used in **campbell2013**. We also make the choice to propagate forward the solutions of c and θ as ρ is varied, as we expect the objective minima to vary smoothly as ρ is changed.

One of the oddities of the generalised profiling method is the discrepancy between the outer and inner objective functions. The reasoning for the form of the outer objective, as given in **ramsay2007** is:

> Because $\hat{c}_i(\theta, \sigma; \lambda)$ is already regularized, criterion H does not require further regularization and is a straightforward measure of fit such as error sum of squares, log-likelihood or some other measure that is appropriate given the distribution of the errors e_{ij}.

with the context that there exist observation errors $e_{ij} \sim \mathcal{N}(0, \sigma^2)$. We take a slightly different philosophical standpoint. Because the model penalty models the process error that is not captured in trajectory matching, it is an integral part of the objective function, as opposed to being a simple regularisation term. Thus the outer objective is computed as the inner objective, with c being profiled out at its minima. This also allows for computational savings, since there is little need to perform a nested optimisation, and c and θ can be optimised simultaneously.

3.2 Identifiability

Regardless of the approach for parameter estimation, the problems of identifiability and estimability will present themselves. These refer to the questions:

1. Given an infinite amount of perfect data, can the parameters be uniquely recovered?
2. Given a finite sample of imperfect data, can the parameters be recovered uniquely?

These questions relate to the problems of structural identifiability and practical identifiability, respectively.

Structural identifiability questions can generally be answered a priori, and an array of differential algebra techniques have been developed and utilised. Two prominent techniques are Differential Algebra Identifiability of Systems (DAISY) and Exact Arithmetic Rank (EAR). These both utilise computational differential algebra techniques to determine global and local structural identifiability respectively. Consider the process (**??**) and observation (**??**) models:

$$\dot{x}(t) = f(x(t), u(t), \theta), \tag{3.5}$$

$$y(t) = g(x(t), u(t), \theta). \tag{3.6}$$

DAISY uses Ritt's algorithm to eliminate the state x to generate algebraic functions of the parameters θ, which can then be analysed for identifiability globally as well as locally**bellu2007**. EAR methods instead construct a Jacobian matrix of the output y with respect to the initial state and parameters (about arbitrary states and parameters), and evaluate its rank to determine local identifiability**karlsson2012**.

Practical identifiability problems generally are data-dependent. **raue2009** introduces the profile likelihood technique to determine practical identifiability. Recall the trajectory matching objective as introduced in Equation (**??**) and denote it $H(\theta)$. Then the profile likelihood p with respect to some parameter θ_i is

$$p(\theta_i) = \min_{\theta_j, \theta_j \neq \theta_i} \{H(\theta)\}. \tag{3.7}$$

This represents the subset of the parameter space that minimises the objective function, for every value of θ_i. As θ_i is varied, if the model is non-identifiable, then the profile will be flat, as the underlying combination of non-identifiable parameters can remain unchanged. The idea

of finite confidence intervals can also be applied to the profile to give a quantitative threshold of "flatness". If the log-likelihood (i.e. objective function) remains below the confidence interval threshold for a given quantile, then the parameter is said to be practically identifiable. Another powerful application of the profile likelihood method is to confirm whether specific combinations of parameters are identifiable, such as $\mathcal{R}_0 = \frac{\beta}{\alpha}$ in the epidemiological literature.

Other methods, such as Monte Carlo simulation and sensitivity analysis techniques are also used to analyse practical identifiability**tuncer2018**. The first of these is similar in intent to the profile likelihood method. It generates parameter estimates over a random sample of trajectories with known true parameters and magnitude of noise, and uses this to identify correlations between parameters - and thus nonidentifiability. As a side-effect of this computation, the relative error of the parameter estimates can also be used to detect unidentifiable parameters, which will have a large amount of error compared to the noise in the trajectory. This method suffers from the need to be performed on known synthetic data, which makes assumptions, such as model adequacy, that cannot be guaranteed for real datasets. Sensitivity analysis techniques use the inverse Fisher Information Matrix as a surrogate for the sample covariance to estimate correlation. The Fisher Information Matrix can be computed as the Hessian of the system, and can be extracted from automatic differentiation techniques.

4 Applications

The regularised generalised profiling technique has been implemented for ordinary differential equations in Python using the CasADi framework. The code is available at https://github.com/dwu402/self-harm.

The implementation has then been applied to synthetic data generated from toy ODE models and standard epidemiological models. The fitting is done with partially observed, noisy data. This method was also applied to a mechanistically-motivated model of the immune system, fitting against malaria data in mice specimens; and to a dimensional SEIR model, fitting against the reported measles case data in Auckland 2019.

4.1 A Toy Model

To demonstrate the properties of the methods, we examine the system:

$$\begin{aligned} \dot{x}_1 &= cx_2(K - x_1), \\ \dot{x}_2 &= -cx_1. \end{aligned} \tag{4.1}$$

We consider the observation model

$$y = x_1 + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{4.2}$$

Trajectories of the system and a single realisation of the observation model are provided in Figure **??**.

toy$_t raj$ toy$_o bsv$

(a) Trajectories of x_1, x_2 **(b)** Realisation of y

Figure 4.1: Toy model trajectories and observation model realisation for $c = 1, K = 3, \sigma = 0.15$, where x_1 is sample uniformly at 15 points in $[0, 20]$

Fitting the generated data using the generalised profiling objective function with IPOPT for the tuning parameters $\rho \in [1 \times 10^{-6}, 1 \times 10^{6}]$ and $\alpha = (\log_{10}(p) + 6) \times 10^{-5}$, we can produce an L curve by plotting the two major error components the data fit error and model fit error against each other on log-log scales as in Figure **??**.

toy$_l$ toy$_{lb} ack$

(a) L Curve **(b)** Back-Propagated L Curve

Figure 4.2: L curve generated by fitting the toy model for $\rho \in [1 \times 10^{-6}, 1 \times 10^{6}]$ on the left, and the back-propagated curve overlay on the right.

We observe an unexpected discontinuity in the L curve. This is likely to do with a nonconvexity leading to multiple near-optima, which switch on some bifurcation of the tuning parameters. We

can propagate the solution backwards (in the direction of decreasing ρ) to examine if this is the case (Figure **??**).

We also profile the likelihoods of the objectives, to examine the identifiability of the parameters.

$$\mathrm{toy}_p 1_p rofile$$

Figure 4.3: Profile of the parameter K with $\rho \approx 0.1335$, and $\alpha = 5 \times 10^{-5}$ for regularisation.

We see that the parameter K is one-sidedly identifiable - all we know is that the value of K must be large. The regularisation therefore specifies the parameter value by choosing the tolerable amount of error as we increase the parameter from the left. We note that the true parameter value is still smaller than the recovered parameter — but now the confidence interval is finite, and still will contain the true value.

4.2 Synthetic Epidemics

To examine the properties of the generalised profiling and profile likelihood methods in the field of interest, we generate synthetic data with the simple SIR model introduced in Equation (**??**). We impose the noise model:

$$\begin{aligned} \hat{I}(t) &\sim \mathcal{N}(I(t), \Gamma), \\ \Gamma_{ij} &= a \exp\left(-\frac{1}{2\delta^2}(t_i - t_j)^2\right), \\ \hat{R}(t) &= \alpha \frac{1}{\Delta t} \sum_{\tau=0}^{t} I(\tau), \\ t &= \{t_0, t_0 + \Delta t, t_0 + 2\Delta t, \dots\} \end{aligned} \tag{4.3}$$

That is, the noisy infected values are modelled as a Gaussian process centred around the true infected values, and the noisy removed values are a cumulative sum of the fraction of infected that are removed. A realisation is provided in Figure **??**

$$\mathrm{synth}_{it} raj \qquad \mathrm{synth}_{rt} raj$$

(a) Realisation of $\hat{I}$ **(b)** Realisation of $\hat{R}$

Figure 4.4: Realisation of Synthetic SIR model with Gaussian process I, $\delta = 2$, $a = 10^4$, $\alpha = 0.75$, $\Delta t = 0.5$

Trajectories are recovered with $\rho = 0.77$, which is chosen to minimise the validation error computed as

$$\sum_i \left((S_{recovered}(t_i) - S_{true}(t_i))^2 + (I_{recovered}(t_i) - I_{true}(t_i))^2 \right). \tag{4.4}$$

Plots of the recovered trajectory for the observed (R) and unobserved (S, I) states are provided in Figure **??**

$\text{synth}_r ecover_r$

(a) Recovery of R

$\text{synth}_r ecover_s i$

(b) Recovery of S, I

Figure 4.5: Recovered trajectories of the model with $\rho = 0.77$, separated by state observability

We also generate profiles for three parameters: β, $S(0) = N + 1$ and α. We expect these to be nonidentifiable, due to the appearance of $\frac{\beta}{N}$ in the model, and the form of the observation model (for α). Indeed, examining the profile (log-)likelihoods of *beta* and $S(0)$ in Figure **??** we can see that they possess a one-sided identifiability. We also note that finite confidence intervals can be constructed in the regularised case, but that is a result of the regularisation imposed, as opposed any qualitative changes in the identifiability. The regularisation effect on confidence intervals

$\text{synth}_{bp} rofile$

(a) Profile of β

$\text{synth}_s 0_p rofile$

(b) Profile of $S(0)$

Figure 4.6: Profiles $p(\theta_i)$ and χ^2 confidence intervals, computed as $1.96 + \min(p(\theta_i))$, are shown for non-regularised and regularised (Tikhonov) forms of the objective function as β and $S(0)$ are varied.

is characterised in more detail for the α profile, where the regularisation parameter is tuned as $\{0, 0.01, 0.5, 10\}$, respectively unregularised, weakly regularised, regularised, and strongly regularised in Figure **??**.

$\text{synth}_{ap} rofile$

Figure 4.7: Profile $p(\alpha)$ and χ^2 confidence intervals $(1.96 + \min(p(\alpha)))$ shown for various levels of regularisation ($\{0, 0.01, 0.5, 10\}$) as the parameter α is varied

We note that as the regularisation increases, the size of the confidence interval shrinks. However, this confidence interval seems to converge onto a value that is different to the true parameter value. This suggests, again, that the regularisation parameter must be tuned carefully to avoid underfitting the data. We also note that sufficiently weak regularisation is not conducive generating stable parameters estimates, as it does not significantly shrink the confidence interval of the estimate as compared to the unregularised estimate - the parameter is still effectively not estimable.

4.3 Measles

The 2019 outbreak of measles is an sobering reminder of the effects of New Zealand's less than optimal vaccination policies and attitudes. Concentrated in the Auckland and Upper North Island regions, particularly in the Counties Manukau DHB area, the number of reported cases reached over 1500 as of the end of September 2019 **esr2019**.

Data on the number of reported cases collated by the ESR, and weekly reports were made available to the public. This was used as the data for fitting to an SEIR model, consistent with the latency (and initial non-infectivity) patterns of measles infections:

$$
\begin{aligned}
\dot{S} &= -\beta SI/N, \\
\dot{E} &= \beta SI/N - \gamma E, \\
\dot{I} &= \gamma E - \alpha I, \\
\dot{R} &= \alpha R, \\
N &= S + E + I + R, \\
\theta &= \{\alpha, \beta, \gamma\}.
\end{aligned}
\tag{4.5}
$$

We impose the observation model:

$$
y = R + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{4.6}
$$

where y is the cumulative weekly incidence count. The interpretation is that the reported cases correspond to the number of newly reported symptomatic cases. Assuming a 100% reporting rate, this will correspond to the number of recovered cases. Due to the fact that the system is autonomous and thus invariant under time shifts, the fact that the reported cases should lag behind the infected cases is can be neglected.

We make a note here that this is not standard. Typically, the data is interpreted as the number of infected present at any given time point ($y \sim \mathcal{N}(\dot{I}, \sigma^2)$) **hooker2011 chatzilena2019**. We find this incongruous with our understanding of the reporting system.

Regularisation was applied to two roughly known parameters $\alpha \approx 7/8, \gamma \approx 7/10$, representing an latent period of approximately 8 days ($7/\gamma$), and an infectious period of 8 days ($7/\alpha$). We choose to impose this prior knowledge in a less strict way to allow for deviations in the qualities of the disease. This also allows for a supplementary ad hoc criterion for determining correct regularisation parameters.

Performing the fitting, we see that the deviation from the regularisation happens when $\rho \approx 1e-3$ in Figure **??**, which corresponds closely with being in the region of a turn in the L-curve in Figure **??**. A trajectory for this value of ρ is produced in Figure **??**.

$$\text{measles}_l curv_w dot$$

Figure 4.8: L-curve of the relative misfits of data and model for the measles dataset when $\alpha = 0.1$ and ρ is varied. The red vertical line corresponds to $\rho = 0.005$.

$$\text{parameter}_v alues_m easles_w red$$

Figure 4.9: Plot of the parameters as ρ is varied. $\rho = 0.005$ is shown in red, the black horizontal lines are the regularisation values of α, γ (lower, known) and β (upper, unknown).

$$\text{measles}_i dx30_t raj$$

Figure 4.10: Recovered trajectory for the measles case when $\rho = 0.005$, $\alpha = 0.1$.

4.3.1 Identifiability and Other Notes

When solving for this system, we have elected to also estimate the total at-risk population N. Sue to this decision, we see that the profiles of the parameters display signs of practical identifiability problems (Figure **??**). This is consistent with the findings in **tuncer2018**, where the SEIR model is found to be practically unidentifiable, despite its supposed structural identifiability (when $S(0)$ and N are known).

$\text{measles}_p rofile_a$ $\text{measles}_p rofile_b$ $\text{measles}_p rofile_g$

Figure 4.11: Profiles of parameters $\{\alpha, \beta, \gamma\}$ with $\rho = 0.005$ for the measles case.

In fact, if the at-risk population is not known, it can be shown that the model is actually also structurally non-identifiable. Using DAISY (Ritt's algorithm), the coefficients of normalised the input-output equation for the SEIR model are given in **tuncer2018** as:

$$\frac{\beta}{N}, \qquad \alpha + \gamma, \qquad \frac{\alpha\beta + \beta\gamma}{N}, \qquad \frac{\alpha\beta\gamma}{N}$$

which can be reduced to show that β is free, with $N = \frac{N_q}{\beta_q}\beta$, alongside the previously known pair of solutions $(\alpha, \gamma) \in \{(\alpha_q, \gamma_q), (\gamma_q, \alpha_q)\}$. To resolve this non-identifiability, either the at-risk population must be known, or a value of $\mathcal{R}_0$ must be specified; alongside knowledge of α or γ. This means that our prior knowledge of the presentation of measles is not sufficient, even if it is strictly imposed. Unfortunately, the value of $\mathcal{R}_0$ is also not specified narrowly enough by previous outbreaks and studies **guerra2017** that we get a usable estimate for it.

However, there is a glimmer of hope. If we consider the final size relation from classical epidemiology, we find a surprising result. The final size relation gives the (theoretical) number of individuals that become infected over the course of the infection:

$$\log\left(\frac{S_\infty}{S_0}\right) = \mathcal{R}_0 \frac{S_\infty - N}{N} \tag{4.7}$$

$$R_\infty = N - S_\infty \tag{4.8}$$

where R_∞ is the final size, S_0 is the initial susceptible population, and $\mathcal{R}_0 \coloneqq \frac{\beta}{\alpha}$ is the basic reproduction number. We can see that the final epidemic size is relatively stable even as ρ is changed in Figure **??**. This is despite the fact that the estimated size of the initial population is relatively volatile as ρ is varied. This may show that the final size of the epidemic is an identifiable parameter combination of the system.

$\text{final}_s ize$

Figure 4.12: Final size of the measles epidemic as ρ is varied, as calculated from the final size relation.

5 Further Work

5.1 Control Strategy Efficacy

One of the curiosities of the 2019 Measles outbreak was the prior implementation of a vaccination campaign in the Waitemata and Auckland District Health Board (WDHB, ADHB) regions for a previous outbreak of mumps in 2017. A question that the two DHBs have is what effect the campaign had on the measles in the two regions **mcqueen2019**, considering they did not experience the same intensity of outbreak as in the Counties Manukau DHB region. The ultimate goal is to use this information to inform cost-benefit analyses of similar vaccination strategies, augmenting analyses performed at a holistic level by **hayman2017b**

Considering the difficulty in estimating the at-risk population, there is a high likelihood that there are still a significant number of susceptible individuals, particularly in the adolescent and young adult age group **reynolds2015**. The concern is that outbreaks are not of sufficient size to affect the entire at-risk population. This makes the prediction of epidemic recurrence difficult, as outbreaks will be of potentially unknown size. This could also mark the importance of contact networks, as there may be sufficiently "disjoint" sub-networks such that the outbreak will exhaust one of in one outbreak, but leave the rest susceptible to further outbreaks.

Other complicating mechanisms that could affect the modelling of this type of include the (external) sources of infection **hayman2017a**. Particularly in Pasifika communities, the amount of non-endemic contact induced by familial visitation, for example, can introduce new sources of infection, or even create a pool of infected. An example of the latter is the recent Samoan measles epidemic that has resulted from contact with infected populations in New Zealand, which can later retransmit to the same population.

Previous modelling work has been done for the optimal timing of the MMR vaccine **roberts2000**, **tobias1998**, using a deterministic SIR model, and generating predictions using the R_0 and R_v (secondary cases with vaccination) constants, and a trajectory matching method. Due to the data availability and computational limits at the time, analysis based on age or ethnicity structure was not performed, which would be the aim of this project.

5.2 Extension to Multiple Dimensions

One clear extension we can make to the model is to introduce more complex differential operators than $\frac{d}{dt}$. This would mean an extension into spatial dimensions.

It is rather simple to construct a reaction-diffusion equation for fitting:

$$\frac{\partial u}{\partial t} = D\nabla^2 u + f(u,\theta) \iff \left(\frac{\partial}{\partial t} - D\nabla^2\right) u = f(u,\theta) \tag{5.1}$$

where D is an element of θ.

This can then be written in the standard generalised profiling form with

$$\mathcal{D}(\cdot) = \left(\frac{\partial}{\partial t} - D\nabla^2\right)(\cdot) \tag{5.2}$$

and utilising the standard tensor product definition of multi-dimensional B-splines **hollig2013**:

$$B(\mathbf{x}) = \prod_i B_i(x_i),\ \mathbf{x} = \{t, x_1, x_2, \dots\} \tag{5.3}$$

where B_i represents a univariate B-spline in the i-th dimension of $\mathbf{x}$, and B is the multidimensional spline.

In fact, this method (as well as a Bayesian near-equivalent) has already been investigated by **xun2013** for a 1D, single-state, linear problem. However, the models used in epidemiology typically have nonlinear reaction terms, as well as being naturally 2D and has multiple states.

One avenue of exploration is the use of non-spline bases, for example, employing classical finite element techniques. Indeed, this has been done for a homogenous anisotropic diffusion model in **bernardi2018** to model rainfall. The generalisation to reaction-diffusion equations for epidemiological modelling would mean that the linear properties of the problem disappear, and also inherit the identifiability problems seen in Section **??**.

Another technique for spatial modelling in the epidemiological literature is the use of metapopulation models, where individuals are also categorised into spatial compartments in addition to infection compartments **arino2003**, **hyman2003**. This allows for the spatio-temporal system to be written as a purely temporal differential equation. This for of model is a likely candidate for the model to be used in the control efficacy study above (Section **??**).

5.3 Bayesian Framework Interpretation

Of course, there is the ability to interpret the generalised profiling objective in the Bayesian framework. Typically, the likelihood function used for Bayesian parameter estimation is derived from the typical trajectory matching objective, i.e.

$$\begin{aligned} \mathcal{L}(y|\theta) &\sim P(\nu_t), \\ \nu_t &= \left\| y - \int_{t-\delta t}^{t} f(x;\theta) \right\|^2, \\ x(t=0) &= x_0 \end{aligned} \tag{5.4}$$

where P is some distribution that can parametrised by ν, for example, in the work of **chatzilena2019**, that P is taken as the Poisson distribution.

If instead we make some more typical assumptions about the error structures in the process and observation models, we can arrive at a Bayesian equivalent of the generalised profiling method.

Consider the observation model to be:

$$y = g(x) + e_o,\ e_o \sim \mathcal{N}(0, \sigma_o^2 I) \tag{5.5}$$

i.e. there is Gaussian, iid additive noise. Then we can say that

$$\pi(y|x) = \pi_{e_o}(y - g(x)) = \mathcal{N}(y - g(x), \sigma_o^2 I) \tag{5.6}$$

by integrating the observation model over e_o.

Next, consider a similar form for the process model:

$$\mathcal{D}x = f(x;\theta) + e_p,\ e_p \sim \mathcal{N}(0, \sigma_p^2 I) \tag{5.7}$$

which gives rise to

$$\pi(x|\theta) = \pi_{e_p}(\mathcal{D}x - f(x;\theta)) = \mathcal{N}(\mathcal{D}x - f(x;\theta), \sigma_p^2 I) \tag{5.8}$$

Recall that the multivariate normal distribution has the density function

$$pdf(x;\mu,\Gamma) = \frac{1}{\sqrt{(2\pi)^k|\Gamma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Gamma^{-1}(x-\mu)\right) \tag{5.9}$$

for $x, \mu \in \mathbb{R}^k$.

Then, taking the log of the posterior gives

$$\begin{aligned}
\log \pi(\theta|y) &= \log\left(\pi(y|x)\pi(x|\theta)\pi(\theta)\right) && (5.10)\\
&= \log \pi(y|x) + \log \pi(x|\theta) + \log \pi(\theta) \\
&= -\frac{1}{2}\left[\sigma_o^{-2}\|y-x\|^2 + \sigma_p^{-2}\|\mathcal{D}x - f(x;\theta)\|^2 + \sigma_\theta^{-2}\|\theta-\theta_0\|^2\right] + const && (5.11)
\end{aligned}$$

where we have introduced the prior $\pi(\theta) \sim \mathcal{N}(\theta_0, \sigma_\theta^2 I)$.

An estimator of the posterior distribution is the *maximum a posteriori* (MAP) estimate, which maximises the posterior likelihood:

$$\begin{aligned}
\theta_{MAP} &= \arg\max_\theta \{\pi(\theta|y)\} && (5.12)\\
&= \arg\max_\theta \{\log \pi(\theta|y)\} \\
&= \arg\min_\theta \{-2\log \pi(\theta|y)\} \\
&= \arg\min_\theta \left\{\sigma_o^{-2}\|y-x\|^2 + \sigma_p^{-2}\|\mathcal{D}x - f(x;\theta)\|^2 + \sigma_\theta^{-2}\|\theta-\theta_0\|^2\right\} && (5.13)
\end{aligned}$$

recovering the generalised profiling objective function. Thus, we see that the generalised profiling method can be thought of as an estimator of a specific case of a much more general error model.

This raises questions about the properties of the estimator if different error structures are used. For example, if we use the Poisson observation model (like in **chatzilena2019**), we can derive a different likelihood, and thus log-likelihood function. Does this allow for different objective function structures to be equally valid estimators? And how does the frequentist approximations of identifiability and confidence breakdown, as compared to the distributions generated by Bayesian methods? Can optimisation methods be used to make Bayesian methods more efficient?

This final question is of particular interest, when considering the inherent computational tradeoff of using a Bayesian method. With optimisation methods, we traverse the posterior distribution in a particular direction to converge onto the MAP estimate. However, to quantify the uncertainty in a Bayesian sampling framework, we are interested in the density of samples in a large area around the MAP estimate. This requires a larger number of samples, and due to needing them to be representative of the posterior, this also means a large number of samples generated that are rejected by the algorithm of choice. Some of these problems can be mitigated by using optimisation methods to get an idea of the behaviour of the posterior around its maxima — its location and local behaviour are represented by the MAP estimate and the profile likelihood respectively. An alternate use of optimisation could be to generate informative priors in order to better regularise the problem. Of course, there is also the obvious benefit of using optimisation methods to detect non-identifiability before attempting the expensive sampling procedures.

6 Resources

6.1 Facilities

It is expected that the bulk of the work for this thesis will be computational in nature. For this, the sufficient resources would be a computer or network of computers able to run code in a reasonable time horizon. The bulk of implementation is expected to be in Python, with dabblings in other languages as required by other tools and plugins. There is potential for the use of a compiled language to accelerate computational speed. Regardless, standard computational resources should be sufficient.

It is expected that for some portions of the project, sensitive healthcare data will need to handled and analysed. For this, it is expected ethics approval to be required. Up to the current point in time, this has not been a requirement, since data has been sourced from publicly accessible databases.

In terms of practical experimentation, it would be impractical, if not wholly unethical, to run realisations of the models. Research relationships with labs and projects more focused on (micro-)biological aspects of infection will be sought. We are currently in contact with epidemiologists in the School of Public Health and representatives from the Auckland and Waitemata DHBs with regard to data and modelling guidance.

6.2 Budget

It is expected that the bulk of non-conference-related expenditure for this project would come from data acquisition, and any outsourced computational power. We do not have firm beliefs on expected costs for these, as we are not pursuing these avenues as of yet.

The budget for this project should be covered by the PRESS allocation.

7 Deliverables

By End of November 2019

Completed by the time of provisional candidate assessment.

1. Implementation of the generalised profiling optimisation method for ODEs
2. Parameter estimation for a simple epidemic model (measles) with synthetic and real data
3. Literature review on epidemiological modelling and parameter estimation.

7.1 In the Next Year

By May 2020

1. Validation study of the method on a synthetic model
2. Implementation and release of a general package for generalised profiling and identifiability analysis of ODEs

By November 2020

1. Analysis of effect of mumps control programme on measles outbreak
2. Basic implementation of generalised profiling methods for PDEs
3. Preliminary analysis of the spatial spread of measles

Expected Outputs and Events

1. Conference: ANZIAM 2020 (February)
2. Conference: SMB 2020 (September)
3. Output: Submission of article on generalised profiling/identifiability for measles

7.2 Within 2 Years

By End of May 2021

1. Validation experiments for the PDE methods
2. Implementation of generalised profiling in a Bayesian sampling method
3. Implementation of generalised profiling as a pre-analysis for Bayesian methods

By End of November 2021

1. Packaged generalised profiling method for PDEs
2. Comparison of uncertainty estimates between Bayesian and frequentist implementations of generalised profiling