Recovering the Time-Dependent Transmission Rate in an SIR Model from Data and an Overfitting Danger

Mark Pollicott\textsuperscript{1}, Hao Wang\textsuperscript{2}, and Howie Weiss\textsuperscript{3}

\textsuperscript{1}Mathematics Institute, University of Warwick, Coventry, CV4 7AL, United Kingdom
\textsuperscript{2}Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, T6G 2G1, Canada
\textsuperscript{3}School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia, 30332, United States

(Dated: May 26, 2009)

The transmissibility of many infectious diseases varies significantly in time, but is impossible to measure directly. We provide an explicit formula to obtain the time-dependent transmission rate in a SIR model from data. This construction also illustrates some dangers of overfitting a transmission model.

I. INTRODUCTION

The SIR epidemiological model was proposed by Kermack and McKendrik \cite{1} and extensively developed by Anderson and May \cite{2}. One of the key parameters is the transmission rate, which is the product of the daily number of contacts a susceptible individual has with infected individuals and the probability of transmission during each contact. In Section 3.4.9 of Anderson and May \cite{2}, the authors state that “... the direct measurement of the transmission rate is essentially impossible for most infections. But if we wish to predict the changes wrought by public health programmes, we need to know the transmission rate ... .” The transmission rate of many infectious diseases varies significantly in time \cite{3}. For instance, one is more likely to contract the flu during the winter than during the summer. For this reason, a number of transmission models use a time varying transmission rate, that is almost always assumed to be a simple trigonometric \cite{5} or Haar \cite{14} function.

We consider the SIR transmission model and allow the transmission rate to be a time varying
function, i.e., there is a positive function \( \beta(t) \) such that

\[
\frac{dS}{dt} = -\beta(t)SI, \quad (1)
\]

\[
\frac{dI}{dt} = \beta(t)SI - \nu I, \quad (2)
\]

\[
\frac{dR}{dt} = \nu I. \quad (3)
\]

Here \( S(t), I(t), \) and \( R(t) \) are the proportions of susceptible, infected, and removed individuals.

We ask the question, given a continuous epidemiological data set, can one choose a transmission rate function \( \beta(t) \) and removal rate \( \nu \) such a model always perfectly fits the data? And if this is possible, what does it imply about the infectious disease? We prove that this is always possible, and we provide an explicit formula for \( \beta(t) \). The simple construction also clearly illustrates some dangers of overfitting a transmission model.

To make the question much more realistic, suppose in addition that the removal rate \( \nu \) is prescribed. Can one always choose a transmission rate function \( \beta(t) \) such a model always perfectly fits the continuous data with the given \( \nu \)? We prove that subject to a mild restriction, this is also always possible.

Of course epidemiological data is discrete, not continuous. We show that one can robustly estimate \( \beta(t) \) by smoothly interpolating the data with a spline or rational function and then applying the formula for continuous data.

Finally, we discuss analogous formula for standard variations of the standard SIR model, and apply the algorithm to a measles dataset.

II. MATHEMATICAL RESULTS

Mathematically, we pose the following inverse question: given an arbitrary smooth positive function \( f(t) \), does there exist a positive function \( \beta(t) \) and recovery rate \( \nu > 0 \) such that the solution \( I(t) \) of (2) coincides with \( f(t) \) on a large time interval? We prove that this is always possible.

**Theorem II.1** Given a smooth positive function \( f(t), b > 0, \) and \( T > 0, \) there exists \( \nu_0 > 0 \) such that if \( \nu > \nu_0 \) there is a solution \( \beta(t) \) with \( \beta(0) = b \) and such that \( I(t) = f(t) \) for \( 0 \leq t \leq T \).

Theorem II.1 implies that there are infinitely many solutions \( \beta(t) \) for each sufficiently large \( \nu \). A more challenging question is whether this is possible if \( \nu \) is prescribed. The answer is provided by the next theorem.
**Theorem II.2** Given a smooth positive function \( f(t) \), \( \nu > 0 \), and \( T > 0 \), there exists \( \beta_0 > 0 \) such that if \( \beta(0) < \beta_0 \) there is a solution \( \beta(t) \) with this initial condition such that \( I(t) = f(t) \) for \( 0 \leq t \leq T \) if and only if \( f'(t)/f(t) > -\nu \) for \( 0 \leq t \leq T \).

The growth condition in Theorem II.2 imposes no restrictions on how \( f(t) \) increases, but requires that \( f(t) \) cannot decrease too fast, in the sense that its logarithmic derivative is always bounded below by \(-\nu\).

The proofs of both theorems are nearly identical and elementary. The details are expanded in Appendix A.

### III. ALGORITHM

Conditions:

- \( f'(t)/f(t) > -\nu \), where \( f(t) \) is the smooth curve generated from a patient dataset and \( \nu \) is the recovery rate.
- \( \beta(0) < 1 / \int_0^T e^{P(s)} f(s) ds \).

Procedure:

**Step 1.** Smooth the patient dataset to obtain a smooth function \( f(t) \) which has at least continuous second derivative.

**Step 2.** Apply the formula \( p(t) = \frac{f''(t) f(t) - f'(t)^2}{f(t)(f'(t) + \nu f(t))} \) to obtain \( p(t) \).

**Step 3.** Apply the formula \( P(t) = \int_0^t p(\tau) d\tau \) to obtain \( P(t) \).

**Step 4.** Obtain \( \beta(t) \) on the given interval \([0 \ T]\) by applying the formula

\[
\beta(t) = 1 / \left[ e^{-P(t)} / \beta(0) - e^{-P(t)} \int_0^t e^{P(s)} f(s) ds \right].
\]

To obtain \( \beta(t) \) on the given interval \([0 \ T]\), one should know one value of \( \beta \) at some time such as \( \beta(0) \), or at least the reasonable range of \( \beta \).

For extensions of the basic SIR model, the conditions and the procedure need to be changed accordingly (see Section V).
IV. SIMULATIONS: RECOVERING THE TRANSMISSION RATE FROM DATA

We first illustrate explicit solutions of the inverse problem with two simulated continuous data sets. To avoid the problems of reasonable units and ranges, we use the “simulation year” as the unit for $t$, choose the functions as the fractions of infected population, and explain $\beta(t)$ as the actual transmission rate divided by the potentially maximum transmission rate.

The first simulates an infectious disease with periodic outbreaks, as observed in measles (before mass vaccination) and cholera [6, 10]. We choose the periodic function $f(t) = 0.25[1.4 + \cos(1.5t)]$, and Figure 1(a) contains plots of both $f(t)$ (solid green) and an associated scaled transmission rate function $\beta(t)$ (dotted blue).

The second data set simulates an infectious disease with periodic outbreaks that decay in time, as observed in influenza [13]. We choose the function $g(t) = 0.5[1.1 + \sin(t)] \exp(-0.1t)$ and Figure 2(a) contains plots of both $g(t)$ (solid green) and an associated scaled transmission rate function $\beta(t)$ (dotted blue).

The observant reader will observe that the peaks for $\beta(t)$ in Figure 1 are slightly increasing over time. This is, in part, a manifestation of the choice of $\beta(0)$ and $\nu$. Different values of $\beta(0)$ and $\nu$ may lead to the peaks increasing, decreasing, or non-monotonic over time.

Formula (8) for $\beta(t)$ requires that the data be smooth, since it involves the first and second derivatives of $f(t)$. In real life, this is never the case: epidemiological data is discrete. We now explain how the algorithm can be used to recover the transmission rate with discrete data. The idea is simple: first smoothly interpolate the data, and then apply the algorithm to the smooth interpolation function.
FIG. 2:

The small black squares in Figure 1 illustrate the function \( f(t) \) sampled at equi-spaced intervals. To this discrete time series we apply three well known interpolation algorithms (trigonometric approximation, sixth order spline approximation, tenth order rational approximation). Figure 2(b) contains plots obtained by substituting the three smooth interpolating functions into formula (8). Comparing the three plots with the actual \( \beta(t) \) in Figure 1(b), one observes that all three functions provide good approximations to the actual \( \beta(t) \). The approximations are even better in Figure 2 for \( \beta(t) \) recovered from the function \( g(t) \) sampled at equi-spaced intervals. Our simulations show that the recovery algorithm is robust with respect to white noise up to 10% of the data mean, as well as the number of sample points.

V. EXTENSIONS OF THE BASIC MODEL

Analogous results hold for all standard variations of the standard SIR model, and combinations. The proofs are very similar to the proofs of Theorems (II.1) and (II.2).

A. SIR model with vital rates

\[
\begin{align*}
\dot{S} &= \delta - \beta(t)SI - \delta S, \\
\dot{I} &= \beta(t)SI - \nu I - \delta I, \\
\dot{R} &= \nu I - \delta R.
\end{align*}
\] (4) (5) (6)

The necessary and sufficient condition for solving for \( \beta(t) \) given \( \nu \) and \( \delta \) is \( f'(t)/f(t) > -(\nu + \delta) \).
B. SIR model with waning immunity

\[ \dot{S} = mR - \beta(t)SI, \]
\[ \dot{I} = \beta(t)SI - \nu I, \]
\[ \dot{R} = \nu I - mR, \]

where \(1/m\) is the memory period of immunity. The necessary and sufficient condition for solving for \(\beta(t)\) given \(\nu\) is \(f'(t)/f(t) > -\nu\).

C. SIR model with time-dependent indirect transmission rate (Richard et al. [12])

\[ \dot{S} = -\omega(t)S, \]
\[ \dot{I} = \omega(t)S - \nu I, \]
\[ \dot{R} = \nu I, \]

where \(\omega(t)\) is the time-dependent indirect transmission rate. The necessary and sufficient condition for solving for \(\beta(t)\) given \(\nu\) is \(f'(t)/f(t) > -\nu\).

D. SEIR model

\[ \frac{dS}{dt} = -\beta(t)SI \]
\[ \frac{dE}{dt} = \beta(t)SI - \alpha E \]
\[ \frac{dI}{dt} = \alpha E - \nu I \]
\[ \frac{dR}{dt} = \nu I, \]

where \(1/\alpha\) is the mean latent period for the disease and has a similar proof. The necessary and sufficient condition for solving for \(\beta(t)\) given \(\nu\) and \(\alpha\) is \(f'(t)/f(t) > -\nu\).

E. SEIR model with vital rates

\[ \frac{dS}{dt} = \delta - \beta(t)SI - \delta S, \]
\[ \frac{dE}{dt} = \beta(t)SI - \alpha E - \delta E, \]
\[ \frac{dI}{dt} = \alpha E - \nu I - \delta I, \]
\[ \frac{dR}{dt} = \nu I - \delta R. \]
The necessary and sufficient condition for solving for $\beta(t)$ is
\[ f'(t) + (\nu + \delta)f(t) > 0 \quad \text{and} \quad f''(t) + (\nu + 2\delta + a)f'(t) + (\delta + a)(\nu + \delta)f(t) > 0. \] (21)

In this case, $\beta(t)$ satisfies the Bernoulli equation
\[ \beta' + p(t)\beta + q(t)\beta^2 = 0, \] (22)

where
\[
p = \frac{-af'''f - a(\nu + 2\delta + a)f''f - a(\delta + a)(\nu + \delta)f'f + af''f' + a(\nu + 2\delta + a)f'^2}{af[f'' + (\nu + 2\delta + a)f' + (\delta + a)(\nu + \delta)f]} + \frac{a(\delta + a)(\nu + \delta)f'f - \delta af''f - \delta a(\nu + 2\delta + a)f'f - \delta a(\delta + a)(\nu + \delta)f^2}{af[f'' + (\nu + 2\delta + a)f' + (\delta + a)(\nu + \delta)f]},
\]

and
\[
q = \frac{\delta a^2 f^2 - a f'' f^2 - a(\nu + 2\delta + a)f'f^2 - a(\delta + a)(\nu + \delta)f^3}{af[f'' + (\nu + 2\delta + a)f' + (\delta + a)(\nu + \delta)f]}. \]

VI. RECOVERING THE TRANSMISSION RATE FROM MEASLES DATA

We now apply the recovery algorithm for the SEIR model with vital rates to measles data from England and Wales (1948-1958) [11]. The original data is weekly data with a substantial amount of noise (see Figure 3(a)). We convert it to monthly data with the fact that part of infected people among all weeks of each month are overlapped. We use the smoothed monthly data to obtain $\beta(t)$ from our recovery algorithm. Eventually, we convert all units back to yearly based.

In this recovering process of the transmission rate, we estimate parameters for measles from Anderson and May [2]: $\nu = 52/\text{year}= 52/12/\text{month}$, $a = 52/\text{year}= 52/12/\text{month}$, $\delta = 1/70/\text{year}= 1/70/12/\text{month}$, where $1/\nu$ is the recovery period, $1/a$ is the latent period, $\delta$ is the birth/death rate.

Figure 3(b) plots the yearly based $\beta(t)$ from the 10-year data set. The range of $\beta(t)$ is biologically reasonable compared to previous studies [4, 8], which include measles models with periodic forcing $\beta(t)$. There are four dominant frequencies, and they are all comparable. Their frequencies are $0.1520/\text{year}$, $0.2400/\text{year}$, $0.3440/\text{year}$, $0.4560/\text{year}$, with the corresponding periods 6.5790 years, 4.1667 years, 2.9070 years, 2.1930 years. The last one is close to the frequency of the monthly
data, since the monthly data has the dominant frequency 0.468/year whose corresponding period is 2.14 years. Actually, the second dominant frequency of the data set is 0.99/year, approximately corresponding to 1 year. This second dominant peak is non-negligible, although it is not the most dominant one.

![Graph of infected population with time]

**FIG. 3:** (a) Weekly measles data from England and Wales in 1948-1958; (b) $\beta(t)$ obtained from converted monthly data.

The lowest point of $\beta(t)$ in each year occurs in August, which is the last month of summer vacation. This is consistent to what scientists anticipated due to the school years. Most high peaks of $\beta(t)$ occur sometime from September to February.

![Graph of beta(t) with different beta(0)]

**FIG. 4:** $\beta(t)$ computed from our algorithm with different $\beta(0)$. All means are within the range of Earn et al. [4].

We plot the dependence of $\beta(t)$ on $\beta(0)$ in Figure 4. When $\beta(0)$ is in the reasonable range, all three curves of $\beta(t)$ have almost the same shape. Hence, the results we observed for the high and low points of $\beta(t)$ are robust to the chosen of $\beta(0)$. Furthermore, all $\beta(t)$’s are in the reasonable range of the realistic transmission rate.
APPENDIX A: PROOF

It is easy to see that $f'(t)/f(t) > -\nu$ is a necessary condition, since Equation (2) implies that $f'(t) + \nu f(t) = \beta(t)S(t)f(t)$, which must be positive for $0 \leq t \leq T$. We now show this condition is also sufficient. We rewrite Equation (2) as

$$S = \frac{f' + \nu f}{\beta f}, \quad (A1)$$

then compute $(d/dt)S$, and then equate with Equation (1) to obtain

$$\frac{d}{dt} \left( \frac{f' + \nu f}{\beta f} \right) = -\beta \left( \frac{f' + \nu f}{\beta f} \right) f. \quad (A2)$$

Calculating the derivative and simplifying the resulting expression yields the following Bernoulli differential equation for $\beta$

$$\beta' - p\beta - f\beta^2 = 0, \quad \text{where} \quad p = \frac{f''f - f'^2}{f(f' + \nu f)}. \quad (A3)$$

The change of coordinates $x = 1/\beta$ transforms this nonlinear ODE into the linear ODE

$$x' - p(t)x - f(t) = 0. \quad (A4)$$

The method of integrating factors provides the explicit solution

$$\frac{1}{\beta(t)} = x(t) = x(0) e^{-P(t)} - e^{-P(t)} \int_{0}^{t} e^{P(s)} f(s) ds, \quad \text{where} \quad P(t) = \int_{0}^{t} p(\tau) d\tau. \quad (A5)$$

The only problem that could arise with this procedure is for the denominator of $p(t)$ to be zero. If there is no restriction on $\nu$, this can be prevented by choosing $\nu$ sufficiently large such that

$$\nu \cdot \min\{f(t) : 0 \leq t \leq T\} > -\max\{f'(t) : 0 \leq t \leq T\}. \quad (A6)$$

The initial condition $\beta(0)$ is arbitrary.

If $\nu$ is prescribed, a singular denominator is prevented by requiring that the denominator be always positive, i.e., $f' + \nu f > 0$. Having done this, to ensure that $\beta(t)$ is positive, one must choose $\beta(0)$ sufficiently small such that

$$\int_{0}^{T} e^{P(s)} f(s) ds < 1/\beta(0). \quad (A7)$$

There are clearly infinitely many choices of $\beta(0)$ and thus infinitely many transmission functions $\beta(t)$. □


[11] The weekly OPCS (Office of Population Censuses and Surveys) reports, the Registrar General’s Quarterly or Annual Reports, & various English census reports.

