Improving epidemic size prediction through stable reconstruction of disease parameters by reduced iteratively regularized Gauss–Newton algorithm

Article *in* Journal of Inverse and Ill-Posed Problems · January 2017

DOI: 10.1515/jiip-2016-0053

CITATION
1

READS
89

5 authors, including:

Gerardo Chowell
Georgia State University
392 PUBLICATIONS **11,491** CITATIONS

SEE PROFILE

Research Article

Alexandra Smirnova*, Gerardo Chowell-Puente, Linda deCamp, Seyed Moghadas and Michael Jameson Sheppard

Improving epidemic size prediction through stable reconstruction of disease parameters by reduced iteratively regularized Gauss–Newton algorithm

DOI: 10.1515/jiip-2016-0053
Received August 22, 2016; revised November 21, 2016; accepted January 21, 2017

Abstract: Classical compartmental epidemic models of infectious diseases track the dynamic transition of individuals between different epidemiological states or risk groups. Reliable quantification of various transmission pathways in these models is paramount for optimal resource allocation and successful design of public health intervention programs. However, with limited epidemiological data available in the case of an emerging disease, simple phenomenological models based on a smaller number of parameters can play an important role in our quest to make forward projections of possible outbreak scenarios. In this paper, we employ the generalized Richards model for stable numerical estimation of the epidemic size (defined as the total number of infections throughout the epidemic) and its turning point using case incidence data of the early epidemic growth phase. The minimization is carried out by what we call the Reduced Iteratively Regularized Gauss–Newton (RIRGN) algorithm, a problem-oriented numerical scheme that takes full advantage of the specific structure of the non-linear operator at hand. The convergence analysis of the RIRGN method is suggested and numerical simulations are conducted with real case incidence data for the 2014–15 Ebola epidemic in West Africa. We show that the proposed RIRGN provides a stable algorithm for early estimation of turning points using simple phenomenological models with limited data.

Keywords: Iteratively regularized Gauss–Newton algorithm, epidemiology, parameter estimation

MSC 2010: 47A52, 65F22

1 Introduction

Emerging and re-emerging infectious diseases continue to cause significant morbidity and mortality around the world [13]. Despite many successes in the application of mathematical models to understanding disease transmission and its control, the key challenge remains: a stable estimation of important disease parameters that will enable accurate forecasting [12]. Estimating such parameters early on can help determine whether the invading pathogen is capable of generating sustained local or global outbreaks. Therefore, it is critical to develop both suitable models and methods to reliably quantify disease-specific parameters, particularly in the face of limited epidemiological data and substantial uncertainty [21, 22].

***Corresponding author: Alexandra Smirnova:** Department of Mathematics and Statistics, Georgia State University, Atlanta, USA, e-mail: asmirnova@gsu.edu
Gerardo Chowell-Puente: School of Public Health, Georgia State University, Atlanta, USA, e-mail: gchowell@gsu.edu
Linda deCamp, Michael Jameson Sheppard: Department of Mathematics and Statistics, Georgia State University, Atlanta, USA, e-mail: ldecamp1@student.gsu.edu, msheppard6@student.gsu.edu
Seyed Moghadas: Agent Based Modelling Laboratory, York University, Toronto, Canada, e-mail: moghadas@yorku.ca

For an emerging disease, the principal goal is to construct a reliable computational algorithm in order to quantify the most significant parameters describing the nature of impending epidemic. This may require, for example, fitting model predictions to a short-term data set comprised of aggregated time series of case incidence. Another crucial question is to understand how soon after the emergence of a new disease, key parameters such as the epidemic size and the epidemic turning point can be projected. Here, we propose a regularized numerical method for early estimation of such parameters, and investigate its characteristics for convergence and computational stability. We apply this method to the generalized Richards model [20]

$$\frac{dC}{dt} = rC^p\left[1 - \left(\frac{C}{K}\right)^a\right], \tag{1.1}$$

where r is the intrinsic growth rate, a measures the extent of deviation from the S-shaped dynamics of the classical logistic growth model [19], and K represents the epidemic final size, defined as the total number of infections throughout the epidemic. When $p = 1$, (1.1) is known as the Richards model [20] and has the analytical solution for the cumulative incidence, given by

$$C(t) = \frac{K}{[1 + ae^{-ar(t-\tau)}]^{1/a}},$$

with τ being the inflection point of C. If $p \neq 1$, (1.1) has no closed form solution and must be solved numerically, although a solution may be expressed in the form of an infinite series [19]. At the early stages of the epidemic, this model allows the capture of different growth profiles ranging from constant incidence ($p = 0$), polynomial (or sub-exponential [7]) growth ($0 < p < 1$), to exponential growth ($p = 1$). The maximum incidence in the generalized Richards model is

$$C'(\tau) = \frac{raK^p}{p}\left(\frac{p}{a+p}\right)^{1+\frac{p}{a}}.$$

Estimation of the time (τ) at which this maximum occurs for an emerging outbreak provides important information on the time-window available to implement the necessary intervention policies to reduce the number of infections. Past the peak-time of the epidemic, public health measures may have little effect on reducing the epidemic final size. We will apply our method to estimating the inflection point and the epidemic final size using the generalized Richards model.

The paper is organized as follows. In Section 2, the original least squares problem with respect to parameters r, p, a, and K, is discussed and the lack of stability in the reconstruction of K is highlighted. In Section 3, the problem is reformulated in a more stable manner with the unknown parameters having closer levels of magnitude. Advantages and limitations of the new formulation in case of both least-squares curve fitting trust-region algorithm in Matlab and our own implementation of iteratively regularized Gauss–Newton solver are presented. Further analysis of the optimization algorithm is proposed in Section 4. It is followed by the introduction of the Reduced Iteratively Regularized Gauss–Newton (RIRGN) method and numerical simulations demonstrating its efficiency in Section 5. The convergence analysis of the RIRGN is carried out in Section 6. Finally, in Section 7 we outline conclusions and directions for future work.

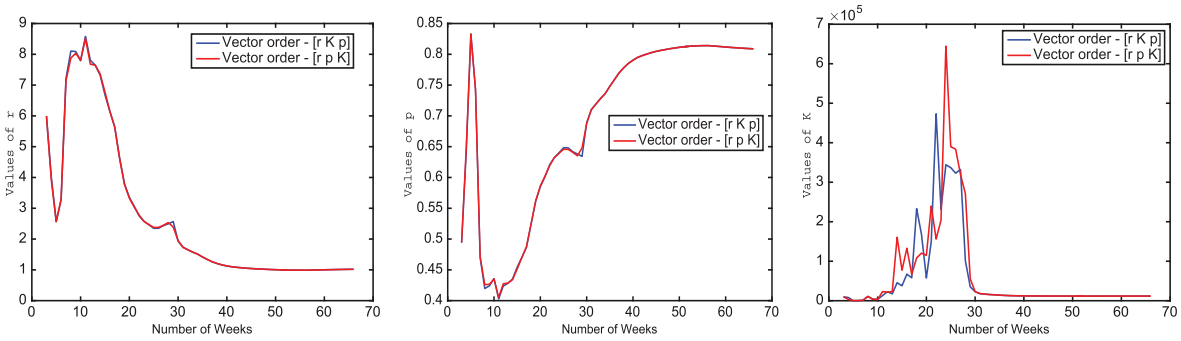
2 The least squares problem

In this section we use the most natural formulation of the inverse problem aimed at the recovery of parameters r, p, a, and K in equation (1.1). Given the cumulative data for a particular outbreak, $D = [D_1, D_2, \ldots, D_m]$, we obtain a numerical solution, $C = C(r, p, a, K)$, to the initial value problem

$$\frac{dC}{dt} = rC^p\left[1 - \left(\frac{C}{K}\right)^a\right], \quad C(t_1) = D_1,$$

at the same points $\{t_1, t_2, \ldots, t_m\}$ where the data are given. Optimizing the values of the unknown parameters to fit the corresponding data, we now have the following non-linear least squares problem

$$\min_{r,p,a,K} \frac{1}{2}\|C(r, p, a, K) - D\|^2. \tag{2.1}$$



(a) Parameter r. **(b)** Parameter p. **(c)** Parameter K.

Figure 1. Sierra Leone EVD outbreak: Impact of coding differences.

The simulation results hint to a substantial noise propagation in the reconstructed values of K prior to the inflection point, which undermines their reliability. We illustrate this phenomenon using cumulative data for the most recent outbreak of Ebola Virus Disease (EVD) in West Africa, predominately affecting Guinea, Liberia, and Sierra Leone. This EVD outbreak, which began in early 2014, has received wide attention due to its scale, scope, location and alarming potential. The largest previous Ebola outbreak was in Uganda in 2000, with a total of 425 cases. The West African outbreak surpassed the size of that outbreak by the first week of June, 2014. The World Health Organization (WHO) declared the latest Ebola outbreak a public health emergency on August 8th, 2014 [26]. By the 21st of that month the case count exceeded the total of all other previous outbreaks combined – 2,387 cases. As of the most recent WHO situation report (March 30th, 2016) there have been 28,646 Ebola cases with 11,323 fatalities [27], and these numbers are widely believed to be underreported.

Human-to-human EVD transmission results from direct contact through broken skin or mucous membranes with the blood and other bodily fluids of infected people. The incubation period, or the time interval from infection to onset of symptoms, is from 2 to 21 days. The patients become contagious once they begin to show symptoms [25]. They are not contagious during the incubation period. Individuals remain infectious as long as their blood and secretions contain the virus [1, 6]. Additionally, humans get infected from improperly handled corpses of infected individuals. The EVD data are notoriously noisy due to substantial underreporting and differing reporting periods. As such, this data provide a unique opportunity to investigate the stability of least squares problem (2.1).

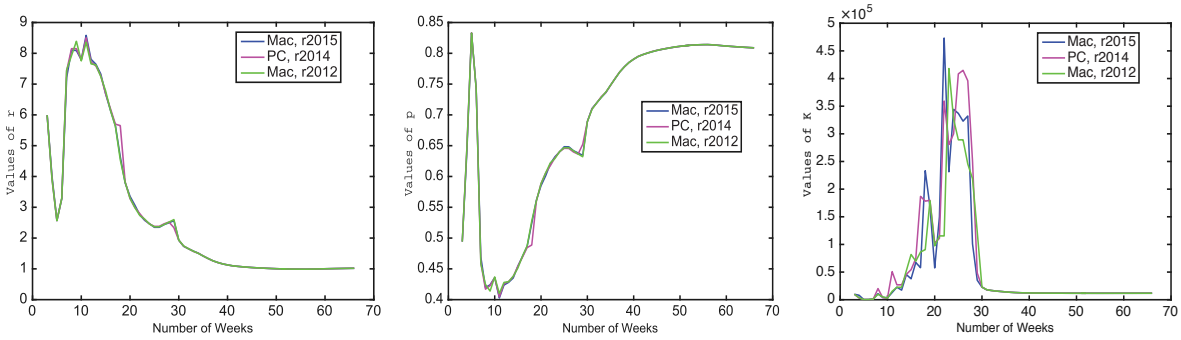
In Figure 1, the impact of programming differences on the approximate values of system parameters is illustrated. Values along the horizontal axis show the number of weeks, for which cumulative data is available to recover the unknown parameters. The corresponding values of the function represent the computed parameters r, p, and K. For each partial data set, system parameters are assumed to be constant.

To enforce stability, we reduce the size of solution space by keeping parameter a fixed and equal to 1, since our statistical analysis of various nested models shows that a is the least important as far as the fit is concerned. This yields what is known as p-Logistic model. Restricting a helps with the recovery of p and r, but not K, which is our main target. Matlab codes developed independently produce small differences in p and r and more significant differences in K. These codes use the same initial values: $r = 1$, $p = 1$, $K = 10000$, but the vectors are coded in different orders: $[r, K, p]$ and $[r, p, K]$. In both cases, the built-in least-squares curve fitting (lsqcurvefit) Matlab sub-function implements the trust-region optimization procedure.

Computer architecture and the version of Matlab utilized also have an effect on the values of K with this being a reflection of instability of the least squares problem. Figure 2 gives parameter outputs under three scenarios:

(1) 1.7 GHz Intel Core i5 MacAir under OS 10.11.2 running Matlab r2015;
(2) late model HP under Windows 9 and Matlab r2014;
(3) 2.26GHz Intel Core 2 Duo under OS 10.7.1 and Matlab r2012a.

Again we observe minor differences in p and r with more significant variation in K.



(a) Parameter r. **(b)** Parameter p. **(c)** Parameter K.

Figure 2. Sierra Leone EVD outbreak: Impact of computer architectures.

3 A creative formulation

The above experiments indicate that approximation of all parameters except K is rather stable. Therefore our next step is to eliminate K from the least squares problem (LSP) [5], and to replace it with another closely related (and equally important) parameter, τ, the disease turning point, which is much closer to other parameters in its level of magnitude. To this end, instead of using the initial condition at t_1 to identify the desired solution curve of (1.1), we take the value of C at τ, $C(\tau) = K(\frac{p}{a+p})^{1/a}$, that one can easily verify by computing the second derivative of C and equating it to zero. To reformulate the LSP, we divide both sides of the equation and of the boundary condition by K:

$$\frac{1}{K}\frac{dC}{dt} = \frac{r}{K^{1-p}}\left(\frac{C}{K}\right)^p\left[1-\left(\frac{C}{K}\right)^a\right], \quad \frac{C}{K}(\tau) = \left(\frac{p}{a+p}\right)^{\frac{1}{a}},$$

and define

$$b := \frac{r}{K^{1-p}}, \quad H(t) := \frac{C(t)}{K}.$$

Hence, we arrive at the boundary value problem (BVP)

$$\frac{dH}{dt} = bH^p(1-H^a), \quad H(\tau) = \left(\frac{p}{a+p}\right)^{\frac{1}{a}}. \tag{3.1}$$

The BVP is solved at every step of the optimization algorithm on some interval $[t_1, t_m]$ that may or may not contain τ. Numerically, we solve it as an IVP by built-in ode23s in the following sense with two cases to be considered:

(1) If $\tau < t_m$, then we solve the ODE forward on the interval $[\tau, t_m]$ using the initial value condition at τ. Subsequently, we solve the ODE backwards on $[\tau, t_1]$ with a negative step size, again utilizing the initial condition. This yields numerical solutions at the grid points $\{t_1, t_2, \ldots, t_m\}$.

(2) If $\tau \geq t_m$, then we solve the ODE backwards on the interval $[\tau, t_1]$ as before. The extraneous entries from the solution vector (after t_m) are deleted so that we have numerical solutions at $\{t_1, t_2, \ldots, t_m\}$ only.

To ensure that early cases do not dominate over the later ones (that are usually less noise contaminated), we replace cumulative data for an epidemic with the incidence data $I = [I_1, I_2, \ldots, I_m]$. By solving (3.1) as stated above, one obtains numerical values of the derivative, $\frac{dH}{dt}$, at $\{t_1, t_2, \ldots, t_m\}$. Since these values approximate the corresponding normalized incidence data, we have the following non-linear least squares problem:

$$\min_{b,p,a,K,\tau} \frac{1}{2}\left\|K\frac{dH}{dt}(b,p,a,\tau) - I\right\|^2.$$

Seemingly, we face a more challenging problem as we now have five parameters rather than four. However, K can be eliminated. Indeed, let

$$f := \frac{1}{2}\left\|K\frac{dH}{dt}(b,p,a,\tau) - I\right\|^2 = \frac{1}{2}K^2\left\|\frac{dH}{dt}\right\|^2 - K\left(\frac{dH}{dt}, I\right) + \frac{1}{2}\|I\|^2.$$

By the first order necessary condition,

$$\frac{\partial f}{\partial K} = K\left\|\frac{dH}{dt}\right\|^2 - \left(\frac{dH}{dt}, I\right) = 0,$$

which implies

$$K = \frac{1}{\left\|\frac{dH}{dt}\right\|^2}\left(\frac{dH}{dt}, I\right). \tag{3.2}$$

Thus, we have the LSP with respect to parameters b, p, a, τ only, since $K = K(b, p, a, \tau)$. From now on, we denote $\mathbf{q} := [b, p, a, \tau]^T$ as the parameter vector. Formally, the revised least squares problem is

$$\min_{\mathbf{q}} \frac{1}{2}\left\|K(\mathbf{q})\frac{dH}{dt}(\mathbf{q}) - I\right\|^2. \tag{3.3}$$

Once the LSP is solved, we can compute K by (3.2) and obtain $r = bK^{1-p}$.

Figures 3–5 illustrate the behavior of turning point parameter τ as a function of the number of weeks of incidence data, computed using the revised least squares problem and Matlab built-in lsqcurvefit solver. Black vertical bars represent the confidence intervals (CIs) evaluated with Matlab built-in nlparci sub-function, which employs a method based on asymptotic normal distribution for the parameter estimate to obtain the CIs. The outline of this algorithm is as follows [24]. Let $\bar{\mathbf{q}}$ be an approximate minimizer of (3.3). Calculate the residual variance as

$$S_2 = \left\|K(\bar{\mathbf{q}})\frac{dH}{dt}(\bar{\mathbf{q}}) - I\right\|^2 \frac{1}{df},$$

where df is the residual degree of freedom:

$$df = \text{length}(I) - \text{length}(\mathbf{q}).$$

The residual variance, S_2, along with a suitable approximation for the Jacobian matrix, J, yields the estimate for the coefficient variance $v = S_2(J^*J)^{-1}$. At the final step, the coefficient variance, v, is used to find the upper and lower confidence bounds,

$$\bar{\mathbf{q}} \pm \text{tinv}(0.975, df)\sqrt{\text{diag}(v)}, \tag{3.4}$$

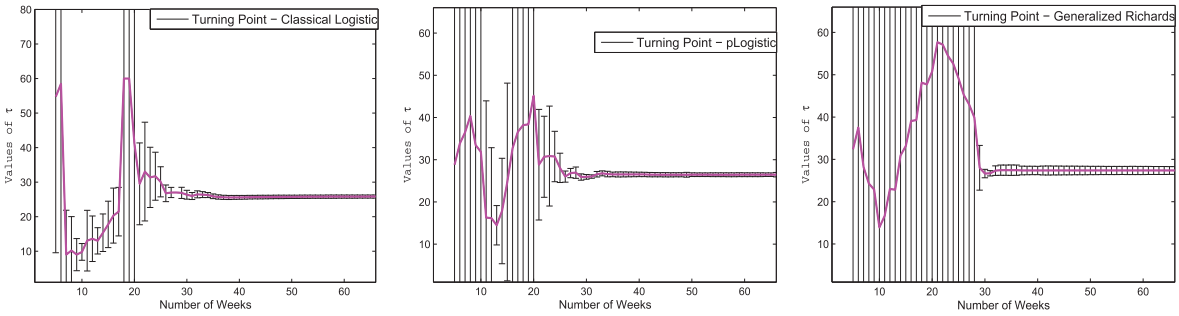
respectively. In (3.4), $\text{tinv}(\rho, n)$ is the inverse of *t-cdf* (cumulative distribution function) with its first parameter, ρ, being the desired probability, and the second parameter, n, representing the degree of freedom. In Section 5 below, when $\bar{\mathbf{q}}$ is approximated by the Reduced Iteratively Regularized Gauss–Newton scheme, the Jacobian matrix in nlparci is replaced with its reduced version.

Our experiments aimed at the recovery of τ from partial data sets reaffirm that models with fewer parameters (such as the classical Logistic model) have shorter intervals of large uncertainty in the reconstructed values of τ. However, the accuracy of τ, prior to the actual turning point, approximated by p-Logistic ($a = 1$) and the generalized Richards model tends to be higher. In particular, for Sierra Leone in Figure 3 the p-Logistic model clearly gives the best result, while for Guinea in Figure 4 and for Liberia in Figure 5, the generalized Richards model outperforms.

Very similar results have been obtained with optimization executed by the classical iteratively regularized Gauss–Newton (IRGN) algorithm [2–4, 9–11, 15], which provides us with more control over regularization compared to the Matlab lsqcurvefit built-in procedure. Thus while replacing parameter K with τ does improve the efficiency of the numerical scheme, the instability of K is essentially carried into the new parameter τ and, therefore, further analysis of the numerical method is required.

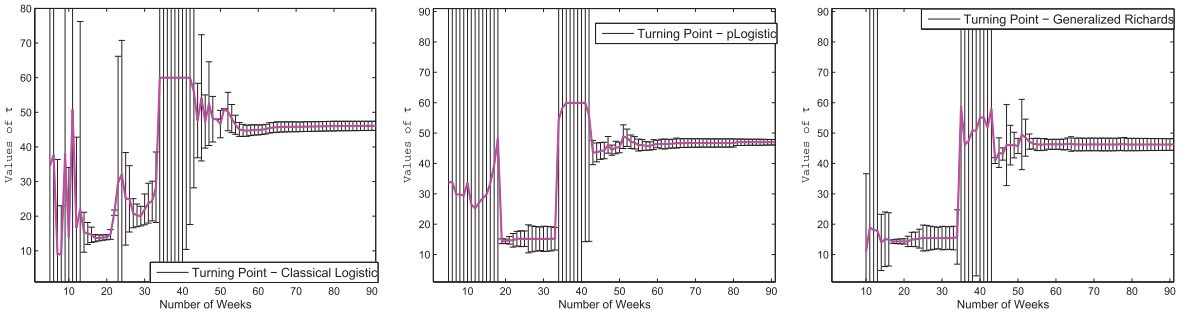
4 Motivation for truncating the Jacobian

Due to severe noise propagation in the parameter τ prior to the actual turning point, which is evident from the large confidence intervals, sporadic behavior, and ill-conditioned Jacobians at each step, our next goal is to consider the computational properties of the gradient and Hessian approximation and to design a more problem-oriented regularized procedure to estimate τ at the early stages of an epidemic.



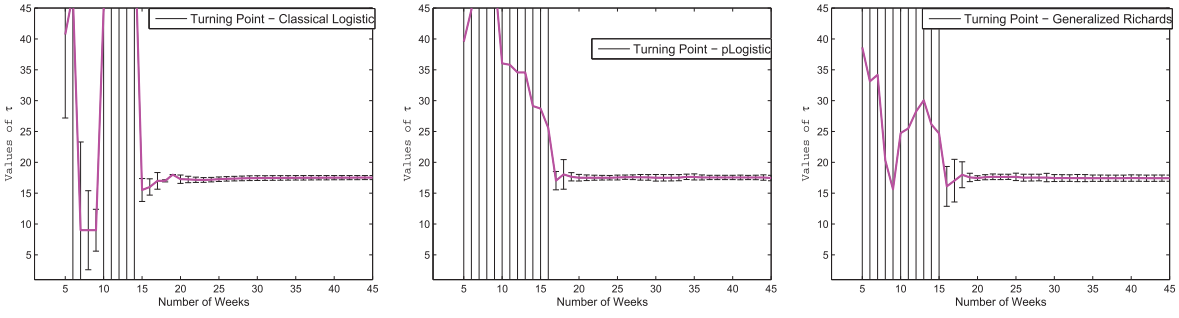
(a) Classical logistic. **(b)** p-logistic. **(c)** Generalized Richards.

Figure 3. Turning point numerical results for Sierra Leone – MATLAB lsqcurvefit.



(a) Classical logistic. **(b)** p-logistic. **(c)** Generalized Richards.

Figure 4. Turning point numerical results for Guinea – MATLAB lsqcurvefit.



(a) Classical logistic. **(b)** p-logistic. **(c)** Generalized Richards.

Figure 5. Turning point numerical results for Liberia – MATLAB lsqcurvefit.

Recall that to approximate τ and other unknown parameters, we consider the constrained least squares problem

$$\min_{\mathbf{q},H} \frac{1}{2}\left\| K(\mathbf{q})\frac{dH}{dt} - I \right\|^2, \quad \text{subject to } F(\mathbf{q}, H) = \mathbf{0},$$

where the operator F is defined by the ODE and by the boundary value condition at τ. When BVP (3.1) is solved numerically, we define

$$\Phi(\mathbf{q}) := K(\mathbf{q})\frac{dH}{dt}(\mathbf{q}), \quad \Phi : \mathbb{R}^4 \to \mathbb{R}^m, \tag{4.1}$$

and penalize the cost functional to obtain the unconstrained regularized least squares problem (variational Tikhonov's regularization) [17, 18, 23]

$$\min_{\mathbf{q}} f_\alpha(\mathbf{q}) := \min_{\mathbf{q}} \frac{1}{2}\|\Phi(\mathbf{q}) - I\|^2 + \frac{\alpha}{2}\|L(\mathbf{q} - \tilde{\mathbf{q}})\|^2. \tag{4.2}$$

Here L is a linear operator, $L : \mathbb{R}^4 \to \mathbb{R}^n$, $n \geq 4$. By solving (4.2) with the Gauss–Newton algorithm and

updating α iteratively, we get the classical IRGN procedure [3]

$$\begin{aligned}[\Phi'^{*}(\mathbf{q}_k)\Phi'(\mathbf{q}_k) + \alpha_k L^{*}L]\mathbf{p}_k &= -[\Phi'^{*}(\mathbf{q}_k)(\Phi(\mathbf{q}_k) - I) + \alpha_k L^{*}L(\mathbf{q}_k - \tilde{\mathbf{q}})],\\ \mathbf{q}_{k+1} &= \mathbf{q}_k + \lambda_k \mathbf{p}_k, \quad \lambda_k > 0.\end{aligned} \tag{4.3}$$

In (4.3), α_k is a regularizing sequence that converges to zero as k approaches infinity, $\tilde{\mathbf{q}}$ is a reference value of $\mathbf{q}$, and λ_k is a line search parameter. In order to compute the Jacobian $\Phi'(\mathbf{q}_k)$, we evaluate partials of Φ with respect to q_i

$$\frac{\partial \Phi_j}{\partial q_i} = \frac{\partial K}{\partial q_i}\frac{dH(t_j)}{dt} + K\frac{d}{dt}\frac{\partial H(t_j)}{\partial q_i}, \tag{4.4}$$

where

$$\frac{\partial K}{\partial q_i}(\mathbf{q}) = \frac{(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), I) - 2K(\mathbf{q})(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), \frac{dH}{dt}(\mathbf{q}))}{(\frac{dH}{dt}(\mathbf{q}), \frac{dH}{dt}(\mathbf{q}))}, \tag{4.5}$$

and $(\cdot, \cdot)$ stands for the scalar product in $\mathbb{R}^m$. To find partials of H, we differentiate the ODE in (3.1) with respect to each parameter to form a system of five differential equations to be solved numerically. In this system, the first differential equation is the original ODE with its corresponding boundary condition at τ. The remaining four differential equations are for $\frac{\partial H}{\partial q_i}$. They take the form

$$\begin{aligned}
\frac{d}{dt}\Big(\frac{\partial H}{\partial b}\Big) &= bH^p\Big[\frac{1 - H^a}{b} + \frac{\partial H}{\partial b}\Big(\frac{p}{H} - (a+p)H^{a-1}\Big)\Big],\\
\frac{d}{dt}\Big(\frac{\partial H}{\partial p}\Big) &= bH^p\Big[\ln(H) + \frac{p}{H}\frac{\partial H}{\partial p} - H^a\Big(\ln(H) + \frac{a+p}{H}\frac{\partial H}{\partial p}\Big)\Big],\\
\frac{d}{dt}\Big(\frac{\partial H}{\partial a}\Big) &= bH^p\Big[\frac{p}{H}\frac{\partial H}{\partial a} - H^a\Big(\ln(H) + \frac{a+p}{H}\frac{\partial H}{\partial a}\Big)\Big],\\
\frac{d}{dt}\Big(\frac{\partial H}{\partial \tau}\Big) &= bH^p\frac{\partial H}{\partial \tau}\Big[\frac{p}{H} - (a+p)H^{a-1}\Big],
\end{aligned}$$

and the boundary conditions at τ are obtained by using the boundary condition in (3.1):

$$\begin{aligned}
\frac{\partial H(\tau)}{\partial b} &= 0,\\
\frac{\partial H(\tau)}{\partial p} &= \frac{p^{\frac{1}{a}-1}}{(a+p)^{\frac{1}{a}+1}},\\
\frac{\partial H(\tau)}{\partial a} &= -\Big(\frac{p}{a+p}\Big)^{\frac{1}{a}}\frac{(a+p)\ln(\frac{p}{a+p}) + a}{a^2(a+p)},\\
\frac{\partial H(\tau)}{\partial \tau} &= -b\Big(\frac{p}{a+p}\Big)^{\frac{p}{a}}\frac{a}{a+p}.
\end{aligned}$$

Upon obtaining the partials, we can evaluate (4.4) at each point in time t_j to form the Jacobian Φ', which enables us to construct both the gradient and the Hessian approximation. As we transition from full Newton's to the Gauss–Newton method, the approximate Hessian matrix, G, is calculated as follows:

$$\begin{aligned}
G_{ij}(\mathbf{q}) &= \Big(\frac{\partial \Phi}{\partial q_i}(\mathbf{q}), \frac{\partial \Phi}{\partial q_j}(\mathbf{q})\Big)\\
&= \Big(\frac{\partial K}{\partial q_i}(\mathbf{q})\frac{dH}{dt}(\mathbf{q}) + K(\mathbf{q})\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), \frac{\partial K}{\partial q_j}(\mathbf{q})\frac{dH}{dt}(\mathbf{q}) + K(\mathbf{q})\frac{d}{dt}\frac{\partial H}{\partial q_j}(\mathbf{q})\Big).
\end{aligned}$$

Substituting expression (4.5) for $\frac{\partial K}{\partial q_i}(\mathbf{q})$ in the above identity, we obtain

$$\begin{aligned}
G_{ij}(\mathbf{q}) = \frac{1}{(\frac{dH}{dt}(\mathbf{q}), \frac{dH}{dt}(\mathbf{q}))}\Big\{&\Big(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), I\Big)\Big(\frac{d}{dt}\frac{\partial H}{\partial q_j}(\mathbf{q}), I - K(\mathbf{q})\frac{dH}{dt}\Big)\\
&- \Big(\frac{d}{dt}\frac{\partial H}{\partial q_j}(\mathbf{q}), I\Big)\Big(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), K(\mathbf{q})\frac{dH}{dt}\Big)\Big\} + K^2\Big(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), \frac{d}{dt}\frac{\partial H}{\partial q_j}(\mathbf{q})\Big).
\end{aligned} \tag{4.6}$$

By the Cauchy–Schwarz inequality, the second term inside the braces in (4.6) can be estimated from below as follows:

$$-\Big(\frac{d}{dt}\frac{\partial H}{\partial q_j}(\mathbf{q}), I\Big)\Big(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), K(\mathbf{q})\frac{dH}{dt}\Big)$$
$$\geq -K\Big(\frac{d}{dt}\frac{\partial H}{\partial q_j}(\mathbf{q}), \frac{d}{dt}\frac{\partial H}{\partial q_j}(\mathbf{q})\Big)^{1/2}(I,I)^{1/2}\Big(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), \frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q})\Big)^{1/2}\Big(\frac{dH}{dt}(\mathbf{q}), \frac{dH}{dt}(\mathbf{q})\Big)^{1/2}.$$

For the diagonal entries of the Hessian approximation, G, this yields

$$G_{ii}(\mathbf{q}) \geq \frac{1}{(\frac{dH}{dt}(\mathbf{q}), \frac{dH}{dt}(\mathbf{q}))}\Big(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), I\Big)\Big(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), I - K(\mathbf{q})\frac{dH}{dt}\Big)$$
$$-\Big(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), \frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q})\Big)\Big[\Big\{\frac{(I - K(\mathbf{q})\frac{dH}{dt}, I) + K(\frac{dH}{dt}, I)}{(\frac{dH}{dt}(\mathbf{q}), \frac{dH}{dt}(\mathbf{q}))}\Big\}^{1/2}K - K^2\Big].$$

Taking into consideration (3.2) and (4.1), one concludes

$$G_{ii}(\mathbf{q}) \geq \frac{1}{(\frac{dH}{dt}(\mathbf{q}), \frac{dH}{dt}(\mathbf{q}))}\Big(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), I\Big)\Big(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), I - \Phi(\mathbf{q})\Big)$$
$$-\Big(\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), \frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q})\Big)\Big[\Big\{\frac{(I - \Phi(\mathbf{q}), I)}{(\frac{dH}{dt}(\mathbf{q}), \frac{dH}{dt}(\mathbf{q}))} + K^2\Big\}^{1/2}K - K^2\Big].$$

This lower bound plays an important role in our understanding of the ill-posedness of the inverse problem at hand. Indeed, it is clear that, as we iterate, the residual, $I - \Phi(\mathbf{q})$, is decreasing (provided that the algorithm converges). Therefore the diagonal elements of the Hessian approximation can become close to zero making the process computationally unstable with G being singular to working precision or highly ill-conditioned. We encountered this problem in the course of our numerical simulations with some limited data sets.

5 Numerical study of the Reduced Iteratively Regularized Gauss–Newton (RIRGN) algorithm

Evidently, we are facing a dilemma of either using the above Hessian approximation coupled with a large penalty term, which reduces the accuracy of the method, or modifying the Hessian to make the algorithm more stable, yet accurate. First we evaluate the gradient of f, $\nabla f(\mathbf{q}) = \Phi'(\mathbf{q})^*(\Phi(\mathbf{q}) - I)$, with the ith component being

$$\frac{\partial f}{\partial q_i}(\mathbf{q}) = \Big(\frac{\partial \Phi}{\partial q_i}(\mathbf{q}), \Phi(\mathbf{q}) - I\Big) = \Big(\frac{\partial K}{\partial q_i}(\mathbf{q})\frac{dH}{dt}(\mathbf{q}) + K(\mathbf{q})\frac{d}{dt}\frac{\partial H}{\partial q_i}(\mathbf{q}), K(\mathbf{q})\frac{dH}{dt}(\mathbf{q}) - I\Big).$$

The Jacobian $\Phi'(\mathbf{q})$ can be expressed as the sum of two matrices $\Phi_1'(\mathbf{q}) + \Phi_2'(\mathbf{q})$ in the following manner

$$\Phi'(\mathbf{q}) = \Big[\frac{\partial K}{\partial q_1}\frac{dH}{dt} \quad \cdots \quad \frac{\partial K}{\partial q_4}\frac{dH}{dt}\Big] + K\Big[\frac{d}{dt}\frac{\partial H}{\partial q_1} \quad \cdots \quad \frac{d}{dt}\frac{\partial H}{\partial q_4}\Big] =: \Phi_1'(\mathbf{q}) + \Phi_2'(\mathbf{q}).$$

Note that for each i, one has

$$\Big(\frac{\partial K}{\partial q_i}(\mathbf{q})\frac{dH}{dt}(\mathbf{q}), K(\mathbf{q})\frac{dH}{dt}(\mathbf{q}) - I\Big) = \frac{\partial K}{\partial q_i}K\Big\|\frac{dH}{dt}\Big\|^2 - \frac{\partial K}{\partial q_i}\Big(\frac{dH}{dt}, I\Big)$$
$$= \frac{\partial K}{\partial q_i}\Big[\frac{1}{\|\frac{dH}{dt}\|^2}\Big(\frac{dH}{dt}, I\Big)\Big\|\frac{dH}{dt}\Big\|^2 - \Big(\frac{dH}{dt}, I\Big)\Big] = 0.$$

Hence, the residual, $\Phi(\mathbf{q}) - I$, is in the kernel of matrix $\Phi_1'(\mathbf{q})$, which yields a simplified form of the gradient, $\nabla f(\mathbf{q}) = \Phi_2'^*(\mathbf{q})(\Phi(\mathbf{q}) - I)$ with a reduced number of operations and, therefore, a reduced noise propagation due to unnecessary rounding. Moreover, the above observation implies that the Hessian approximation comes down to

$$G(\mathbf{q}) = \Phi_2'^*(\mathbf{q})\Phi'(\mathbf{q}) = \Phi_2'^*(\mathbf{q})(\Phi_1'(\mathbf{q}) + \Phi_2'(\mathbf{q})). \tag{5.1}$$

However, $\Phi_2'^*(\Phi_1' + \Phi_2')$ inherits poor computational properties from $\Phi'^*\Phi'$, as one can easily verify by deriving a similar lower bound for operator (5.1) using the Cauchy–Schwarz inequality. Besides, (5.1) is no longer symmetric non-negative definite. This consideration coupled with the evidence from our numerical experiments suggest further reduction of the Hessian approximation by eliminating Φ_1' from (5.1) and setting

$$G(\mathbf{q}) \approx \Phi_2'^*(\mathbf{q})\Phi_2'(\mathbf{q}). \tag{5.2}$$

This operator is symmetric and non-negative definite. Approximation (5.2) results in the following iterative scheme:

$$\begin{gathered}[\Phi_2'^*(\mathbf{q}_k)\Phi_2'(\mathbf{q}_k) + \alpha_k L^*L]\mathbf{p}_k = -[\Phi_2'^*(\mathbf{q}_k)(\Phi(\mathbf{q}_k) - I) + \alpha_k L^*L(\mathbf{q}_k - \tilde{\mathbf{q}})], \\ \mathbf{q}_{k+1} = \mathbf{q}_k + \lambda_k \mathbf{p}_k, \quad \lambda_k > 0.\end{gathered} \tag{5.3}$$

To optimize the step size in (5.3), we use a version of the Armijo–Goldstein line search strategy [14], i.e., a backtracking with $\lambda_k = \frac{1}{2}, \frac{1}{4}, \ldots$ until

$$\|\Phi(\mathbf{q}_k + \lambda_k \mathbf{p}_k) - I\|^2 < \|\Phi(\mathbf{q}_k) - I\|^2 + \lambda_k \beta(\Phi_2'^*(\mathbf{q}_k)(\Phi(\mathbf{q}_k) - I), \mathbf{p}_k),$$

which is commonly implemented for Gauss–Newton-type algorithms. In (5.3), we assume that L^*L is invertible and

$$(L^*L\mathbf{h}, \mathbf{h}) \geq c\|\mathbf{h}\|^2, \quad c > 0, \tag{5.4}$$

for any $\mathbf{h} \in \mathbb{R}^4$. The upgrade from the identity operator to a general linear operator L allows, when necessary, the placement of more regularization on some unknown parameters and less on others. Condition (5.4) includes a bound which depends on the finite dimensional operator L in terms of calculable parameter $c = \min \varsigma_i^2$, which in turn enters in (6.5), used for the convergence analysis in the next section, through its inverse. Here ς_i are the singular values of L ordered from largest to smallest. For the version of L suggested below, $c = 1$.

We call (5.3) *the Reduced IRGN*. For our specific problem, this algorithm is more stable compared to classical IRGN, and most solution curves obtained by (5.3) are superior to those produced by IRGN or Matlab built-in lsqcurvefit in terms of accuracy and stability as one can see by comparing reconstructions of τ in Figures 3–5 and Figures 6–8.

In particular, Figures 6–8 illustrate the advantages of the Reduced IRGN in leading to much smaller confidence intervals (CIs) in the decisive time period before the turning point. The new method does lead slightly larger CIs after the turning point, but this period is much less important in practice.

Apart from the turning points, τ, the initial results for Sierra Leone, Guinea, and Liberia presented in Figures 6–8 illustrate saturation levels, K, and comparison of the forecasting curves with parameters recovered at (a) the earliest possible moment (black curves), (b) at the actual turning point (green curves), and (c) from full disease data (red curves). For the parameter K, RIRGN yields considerably more accurate upper bounds prior to the turning points as compared to our initial reconstructions shown in Figures 1 and 2. All experiments presented in Figures 6–8 have been conducted for the generalized Richards model.

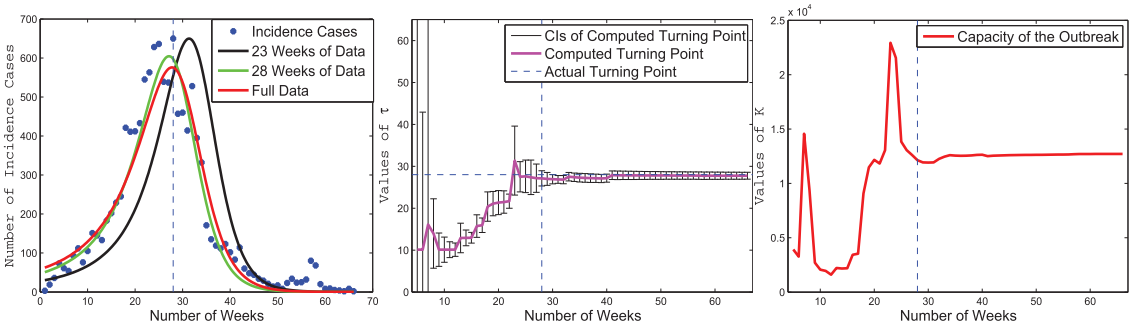
As we implement algorithm (5.3) in practice, at every step of the iterative process $K^2(\mathbf{q}_k)$ is canceled on both sides of (5.3), which yields the following system:

$$\left[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L\right]\mathbf{p}_k = -\left[A'^*(\mathbf{q}_k)\left(A(\mathbf{q}_k) - \frac{I_\delta}{K(\mathbf{q}_k)}\right) + \tilde{\alpha}_k L^*L(\mathbf{q}_k - \tilde{\mathbf{q}})\right], \tag{5.5}$$

where

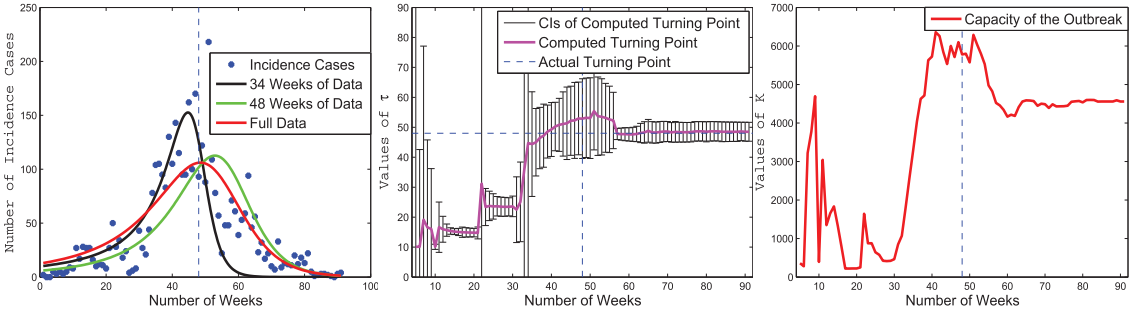
$$A(\mathbf{q}) := \frac{dH}{dt}(\mathbf{q}), \quad \|I - I_\delta\| \leq \delta, \quad \text{and} \quad \tilde{\alpha}_k := \frac{\alpha_k}{K^2(\mathbf{q}_k)}.$$

Recall that $K(\mathbf{q}_k)$ in (3.2) is evaluated from noisy data. Hence division by $K^2(\mathbf{q}_k)$ enables us to move all noise from the matrix of system (5.5) to its right-hand side, which makes (5.5) computationally more stable. It also normalizes the residual and allows the use of the same regularization sequence, $\{\tilde{\alpha}_k\}$, for multiple data sets.



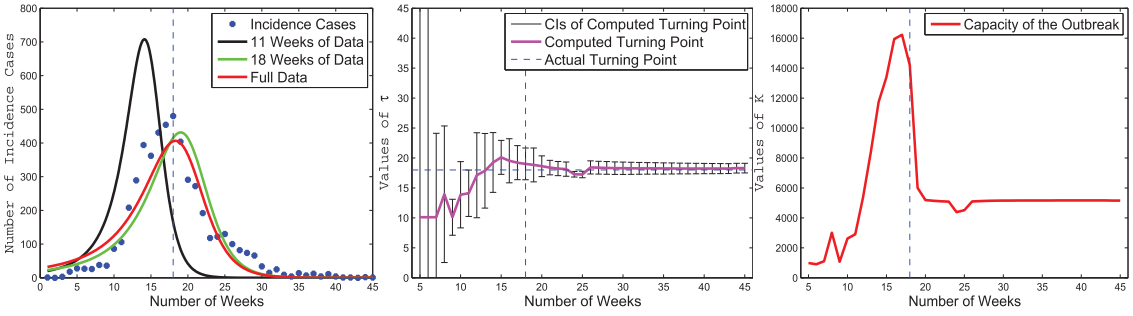
(a) Forecasting curves. **(b)** Turning point. **(c)** Saturation level.

Figure 6. Numerical results for Sierra Leone – Reduced IRGN.



(a) Forecasting curves. **(b)** Turning point. **(c)** Saturation level.

Figure 7. Numerical results for Guinea – Reduced IRGN.



(a) Forecasting curves. **(b)** Turning point. **(c)** Saturation level.

Figure 8. Numerical results for Liberia – Reduced IRGN.

The only adjustment that needs to be made is for $\tilde{\alpha}_0$, since data sets with higher noise level require more regularization. In all experiments shown in Figures 6–8, $\tilde{\alpha}_k = \tilde{\alpha}_0 \exp(-\frac{k}{2})$ with $\tilde{\alpha}_0 = 5 \cdot 10^{-4}$ for Sierra Leone and Liberia, and $\tilde{\alpha}_0 = 10^{-3}$ for Guinea. The choice $\tilde{\alpha}_k = \tilde{\alpha}_0 \exp(-\frac{k}{2})$ provides the most aggressive convergence rate for the regularized algorithm. At the same time, it preserves stability at every step of the iterative process until it is terminated by stopping rule (6.2) below. The stopping rule guarantees that, while our numerical solution does fit the data, we do not over-fit and ensure approximation of the exact solution to the noise-free problem rather than solution to the problem with noise-contaminated data. This phenomena is called semi-convergence, and stopping at the right moment is crucial for an unstable model. Stopping rule (6.2) is explained in Section 6.

Another important aspect is the choice of L in (5.5). In the vector $\mathbf{q} := [b, p, a, \tau]^T$, the value of τ is between one and two orders of magnitude larger than b, p, and a. This suggest that the regularization applied to b, p, and a should be appropriately weighted in order to balance the sensitivity of the cost functional to all

four parameters. Thus we take

$$L = \begin{bmatrix} \omega & 0 & 0 & 0 \\ 0 & \omega & 0 & 0 \\ 0 & 0 & \omega & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \omega > 1. \tag{5.6}$$

An arbitrary choice, say, $\omega = 10$ gives stable computational results, but $\omega = 1$ yields a very poor accuracy of the approximate solutions, since either τ tends to be over-regularized or there is lack of stability in b, p, and a. For this reason, the use of a general linear operator L, rather than the identity operator, is crucial for the success of the proposed algorithm.

The choice of the test function, $\tilde{\mathbf{q}}$, that is meant to bring *a priori* information in the penalty term, is very difficult. In the beginning of an emerging outbreak, it is hard to have an accurate *a priori* estimate as to when the peak is going to occur. For a fair comparison to lsqcurvefit, in all our experiments we take $\tilde{\mathbf{q}} = [1, 1, 1.5, 60]$, i.e., we assume the incidence curve will turn after 60 weeks, which puts Liberia at huge disadvantage, since the actual turning point there appears to be 18. However, the use of a general matrix L in the penalty term allows us to reduce the weight on τ and, nevertheless, maintain stability. As the result, the poor *a priori* value of $\tilde{\mathbf{q}}$ does not hinder the recovery of τ in case of Liberia. In fact, the reconstruction of the Liberia turning point is the most accurate, partly due to a smaller noise level (compared to, say, Guinea) and partly because of a shorter time frame. The most difficult case is Guinea due to a high noise level in the reported incidence data. But even for Guinea, the τ curve does not bounce all the way towards 60 (as opposed to lsqcurvefit where 60 is enforced as upper bound on τ) and, starting with week 34, we get a reliable estimate of the actual turning point, 48.

For all numerical experiments presented in this paper, the confidence intervals have been computed with Matlab built-in nlparci sub-function that estimates uncertainty in the recovered parameters using residual and Hessian approximation for a Newton-type iterative method at hand. The iterations are terminated by the generalized discrepancy principle as outlined in the convergence analysis below. Encouraged by the numerical simulations presented in this section, we move to the theoretical study of the RIRGN procedure.

6 Convergence analysis of the RIRGN method

In order to show that the RIRGN algorithm is well defined and convergent, we use the general scheme developed for the analysis of the original IRGN [2, 3, 9, 16]. Assume that $\{\tilde{\alpha}_k\}$ in (5.5) is a regularization sequence satisfying the conditions

$$\tilde{\alpha}_k \geq \tilde{\alpha}_{k+1} > 0, \quad \sup_{k \in \mathcal{N} \cup \{0\}} \frac{\tilde{\alpha}_k}{\tilde{\alpha}_{k+1}} = d < \infty, \quad \lim_{k \to \infty} \tilde{\alpha}_k = 0,$$

and $\{\lambda_k\}$ is a step size sequence such that

$$0 < \lambda \leq \lambda_k \leq 1.$$

Let $\hat{\mathbf{q}}$ be a global minimum of the functional $f(\mathbf{q})$ with the noise-free data and let

$$\|I - I_\delta\| \leq \delta, \quad \delta \geq 0.$$

It has been established in [3, 9, 16] that, if the estimate

$$\|\mathbf{q}_{k+1} - \hat{\mathbf{q}}\| \leq (1 - \gamma\lambda_k)\|\mathbf{q}_k - \hat{\mathbf{q}}\| + \frac{\lambda_k \beta}{\sqrt{\tilde{\alpha}_k}}\|\mathbf{q}_k - \hat{\mathbf{q}}\|^2 + \lambda_k \sqrt{\tilde{\alpha}_k}\sigma + \frac{\lambda_k \kappa \delta}{\sqrt{\tilde{\alpha}_k}}, \quad k = 0, 1, \ldots, \tag{6.1}$$

holds for $\{\mathbf{q}_k\}$ in a Hilbert space $\mathbb{H}$ with some non-negative constants β, γ, σ, and κ (with σ being sufficiently small and $\gamma\lambda < 1$), then there exists $l > 0$, $l = l(\beta, \gamma, \sigma, \kappa, d)$, such that

$$\frac{\|\mathbf{q}_k - \hat{\mathbf{q}}\|}{\sqrt{\tilde{\alpha}_k}} \leq l \quad \text{for } k = 0, 1, \ldots, \mathcal{K}(\delta),$$

provided $\|\mathbf{q}_0 - \hat{\mathbf{q}}\|$ is sufficiently small, and $\mathcal{K} = \mathcal{K}(\delta)$ is evaluated by the discrepancy-type stopping rule

$$\left\|A(\mathbf{q}_{\mathcal{K}(\delta)}) - \frac{I_\delta}{K(\mathbf{q}_{\mathcal{K}(\delta)})}\right\|^2 \le \rho\kappa\delta < \left\|A(\mathbf{q}_k) - \frac{I_\delta}{K(\mathbf{q}_k)}\right\|^2, \quad 0 \le k \le \mathcal{K}(\delta),\ \rho > 1, \tag{6.2}$$

and the sequence $\mathcal{K} = \mathcal{K}(\delta)$ is admissible, that is,

$$\lim_{\delta\to 0} \|\mathbf{q}_{\mathcal{K}(\delta)} - \bar{\mathbf{q}}\| = 0,$$

for

$$\bar{\mathbf{q}} = \operatorname*{argmin}_{\mathbf{q}\in\mathbb{H}} \left\|A(\mathbf{q}) - \frac{I}{K(\mathbf{q})}\right\|.$$

Remark 6.1. A stronger version of the above stopping rule has been proposed for IRGN in [8] under the assumption that L is the identity operator.

In what follows, we will verify that for $\{\mathbf{q}_k\}$, defined in (5.5), inequality (6.1) holds. Assume as before that $A(\mathbf{q}) := \frac{dH}{dt}(\mathbf{q})$, and let $A : \mathbb{R}^4 \to \mathbb{R}^m$, where m is the number of data points. Clearly, the matrix $A'(\mathbf{q})$ is Lipschitz continuous in a neighborhood $\mathcal{O}(\hat{\mathbf{q}})$, which does not contain negative values of b, p, a, and τ. Negative values of b, p, a, and τ are not relevant for our particular application. Assume that for any $\mathbf{u}, \mathbf{v} \in \mathcal{O}(\hat{\mathbf{q}})$,

$$\|A'(\mathbf{u})\| \le M_1, \quad \|A'(\mathbf{u}) - A'(\mathbf{v})\| \le M_2\|\mathbf{u} - \mathbf{v}\|, \quad \text{and} \quad \frac{\|A(\mathbf{u})\|}{|(A(\mathbf{u}), A(\mathbf{v}))|} \le N. \tag{6.3}$$

The last inequality in (6.3) underscores that, while $A(\mathbf{u}) > 0$, $\mathbf{u} \in \mathcal{O}(\hat{\mathbf{q}})$, may get close to zero as C approaches K, we do not consider the case when $t \to \infty$ or gets too large. The time of an outbreak is finite, and $1 - \frac{C(t)}{K} \ge 1 - \frac{C(t_m)}{K} > 0$. For early data, t_m is even smaller than in the case when the entire outbreak is investigated. Condition (6.3) yields

$$A(\hat{\mathbf{q}}) = A(\mathbf{q}_k) + A'(\mathbf{q}_k)(\hat{\mathbf{q}} - \mathbf{q}_k) + \mathcal{B}(\hat{\mathbf{q}}, \mathbf{q}_k),$$

where

$$\|\mathcal{B}(\hat{\mathbf{q}}, \mathbf{q}_k)\| \le \frac{M_2}{2}\|\hat{\mathbf{q}} - \mathbf{q}_k\|^2.$$

By (5.5) and (6.3), one concludes that

$$\begin{aligned}\mathbf{q}_{k+1} - \hat{\mathbf{q}} = \mathbf{q}_k - \hat{\mathbf{q}} &- \lambda_k[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L]^{-1}\{A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L\}(\mathbf{q}_k - \hat{\mathbf{q}})\\ &- \lambda_k[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L]^{-1}A'^*(\mathbf{q}_k)\left\{A(\hat{\mathbf{q}}) - \frac{I_\delta}{K(\mathbf{q}_k)} - \mathcal{B}(\hat{\mathbf{q}}, \mathbf{q}_k)\right\}\\ &- \lambda_k\tilde{\alpha}_k[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L]^{-1}L^*L(\hat{\mathbf{q}} - \tilde{\mathbf{q}}).\end{aligned} \tag{6.4}$$

As proven in [3, 9, 16], under assumption (5.4)

$$\begin{aligned}\left\|[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L]^{-1}\right\| &\le \frac{1}{\tilde{\alpha}_k c},\\ \left\|[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L]^{-1}A'^*(\mathbf{q}_k)\right\| &\le \frac{1}{2\sqrt{\tilde{\alpha}_k c}},\\ \left\|[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L]^{-1}A'^*(\mathbf{q}_k)A'(\mathbf{q}_k)\right\| &\le 1.\end{aligned} \tag{6.5}$$

Consider the term $A(\hat{\mathbf{q}}) - \frac{I_\delta}{K(\mathbf{q}_k)}$. Note that by (3.2)

$$K(\hat{\mathbf{q}}) = \frac{(A(\hat{\mathbf{q}}), I)}{(A(\hat{\mathbf{q}}), A(\hat{\mathbf{q}}))} = \frac{(A(\hat{\mathbf{q}}), I - K(\hat{\mathbf{q}})A(\hat{\mathbf{q}}))}{(A(\hat{\mathbf{q}}), A(\hat{\mathbf{q}}))} + K(\hat{\mathbf{q}})\frac{(A(\hat{\mathbf{q}}), A(\hat{\mathbf{q}}))}{(A(\hat{\mathbf{q}}), A(\hat{\mathbf{q}}))}. \tag{6.6}$$

Hence $A(\hat{\mathbf{q}}) = \frac{I}{K(\hat{\mathbf{q}})}$. One has

$$A(\hat{\mathbf{q}}) - \frac{I_\delta}{K(\mathbf{q}_k)} = A(\hat{\mathbf{q}}) - \frac{I}{K(\mathbf{q}_k)} + \frac{I - I_\delta}{K(\mathbf{q}_k)}. \tag{6.7}$$

To complete the estimate, one writes $K(\mathbf{q}_k)$ as

$$K(\mathbf{q}_k) = \frac{(A(\mathbf{q}_k), I)}{(A(\mathbf{q}_k), A(\mathbf{q}_k))} = K(\hat{\mathbf{q}})\frac{(A(\mathbf{q}_k), A(\hat{\mathbf{q}}))}{(A(\mathbf{q}_k), A(\mathbf{q}_k))}. \tag{6.8}$$

Based on (6.8), one derives

$$A(\hat{\mathbf{q}}) - \frac{I}{K(\mathbf{q}_k}) = A(\hat{\mathbf{q}}) - \frac{I(A(\mathbf{q}_k), A(\mathbf{q}_k))}{K(\hat{\mathbf{q}})(A(\mathbf{q}_k), A(\hat{\mathbf{q}}))} = A(\hat{\mathbf{q}}) - A(\hat{\mathbf{q}})\frac{(A(\mathbf{q}_k), A(\mathbf{q}_k))}{(A(\mathbf{q}_k), A(\hat{\mathbf{q}}))}. \tag{6.9}$$

If one replaces $\hat{b}$ with zero in $\hat{\mathbf{q}} := [\hat{b}, \hat{p}, \hat{a}, \hat{\tau}]^T$ and introduces the vector $\mathbf{q}^{(b)} := [0, \hat{p}, \hat{a}, \hat{\tau}]^T$, then by the equation in (3.1)

$$0 = A(\mathbf{q}^{(b)}) = A(\hat{\mathbf{q}}) + A'(\xi^{(b)})(\mathbf{q}^{(b)} - \hat{\mathbf{q}}).$$

Suppose that $A'(\xi^{(b)})$ takes the form $A'(\xi^{(b)}) = A'(\hat{\mathbf{q}})R(\xi^{(b)}, \hat{\mathbf{q}})$, with $R(\xi^{(b)}, \hat{\mathbf{q}})$ being a 4×4 matrix and $\|R(\xi^{(b)}, \hat{\mathbf{q}})(\mathbf{q}^{(b)} - \hat{\mathbf{q}})\| := \mu$, a small positive constant. This assumption is justified, since $\|\mathbf{q}^{(b)} - \hat{\mathbf{q}}\| = |\hat{b}| = \hat{b}$, and parameter b is normalized by K^{1-p}, $1 - p > 0$. Therefore for all data sets we consider, $0 < \hat{b} < 1$ (usually it is between 0.2 and 0.3) and in all experiments we take $b_0 = 0.5$. Hence from (6.9) one concludes

$$\begin{aligned} A(\hat{\mathbf{q}}) - \frac{I_\delta}{K(\mathbf{q}_k)} &= A'(\xi^{(b)})(\hat{\mathbf{q}} - \mathbf{q}^{(b)})\frac{(A(\mathbf{q}_k), A(\hat{\mathbf{q}}) - A(\mathbf{q}_k))}{(A(\mathbf{q}_k), A(\hat{\mathbf{q}}))} \\ &= A'(\hat{\mathbf{q}})R(\xi^{(b)}, \hat{\mathbf{q}})(\hat{\mathbf{q}} - \mathbf{q}^{(b)})\frac{(A(\mathbf{q}_k), A'(\xi_k)(\hat{\mathbf{q}} - \mathbf{q}_k))}{(A(\mathbf{q}_k), A(\hat{\mathbf{q}}))} + \frac{I - I_\delta}{K(\mathbf{q}_k)}. \end{aligned} \tag{6.10}$$

Representation (6.10) implies

$$\begin{aligned} &\left\|[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L]^{-1}A'^*(\mathbf{q}_k)\left\{A(\hat{\mathbf{q}}) - \frac{I_\delta}{K(\mathbf{q}_k)}\right\}\right\| \\ &\quad \le \{\|[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L]^{-1}A'^*(\mathbf{q}_k)\|\,\|A'(\hat{\mathbf{q}}) - A'(\mathbf{q}_k)\| + \|[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L]^{-1}A'^*(\mathbf{q}_k)A'(\mathbf{q}_k)\|\} \\ &\qquad \times \|R(\xi^{(b)}, \hat{\mathbf{q}})(\hat{\mathbf{q}} - \mathbf{q}^{(b)})\|\frac{\|A(\mathbf{q}_k)\|\,\|A'(\xi_k)\|\,\|\hat{\mathbf{q}} - \mathbf{q}_k\|}{|(A(\mathbf{q}_k), A(\hat{\mathbf{q}}))|} \\ &\qquad + \|[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L]^{-1}A'^*(\mathbf{q}_k)\|\frac{\|I - I_\delta\|}{|K(\mathbf{q}_k)|}. \end{aligned} \tag{6.11}$$

Given the nature of $K(\mathbf{q}_k)$, we assume that $0 < \tilde{K} \le K(\mathbf{q}_k)$ for any $k = 0, 1, 2, \ldots$. Inequality (6.11) combined with (6.3) and (6.5) yields

$$\begin{aligned} &\left\|[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L]^{-1}A'^*(\mathbf{q}_k)\left\{A(\hat{\mathbf{q}}) - \frac{I_\delta}{K(\mathbf{q}_k)} - \mathcal{B}(\hat{\mathbf{q}}, \mathbf{q}_k)\right\}\right\| \\ &\quad \le \left\{\frac{M_2\|\hat{\mathbf{q}} - \mathbf{q}_k\|}{2\sqrt{\tilde{\alpha}_k c}} + 1\right\}\mu N M_1\|\hat{\mathbf{q}} - \mathbf{q}_k\| + \frac{\delta}{2\tilde{K}\sqrt{\tilde{\alpha}_k c}} + \frac{M_2\|\hat{\mathbf{q}} - \mathbf{q}_k\|^2}{4\sqrt{\tilde{\alpha}_k c}} \\ &\quad = [2\mu N M_1 + 1]\frac{M_2\|\hat{\mathbf{q}} - \mathbf{q}_k\|^2}{4\sqrt{\tilde{\alpha}_k c}} + \mu N M_1\|\hat{\mathbf{q}} - \mathbf{q}_k\| + \frac{\delta}{2\tilde{K}\sqrt{\tilde{\alpha}_k c}}. \end{aligned} \tag{6.12}$$

To complete the estimate for $\|\mathbf{q}_{k+1} - \hat{\mathbf{q}}\|$, we assume that L and $\tilde{\mathbf{q}}$ are chosen according to the modified source condition [16]

$$L^*L(\hat{\mathbf{q}} - \tilde{\mathbf{q}}) \in A'^*(\hat{\mathbf{q}})S, \quad S := \{w \in \mathbb{R}^m : \|w\| \le \varepsilon\}, \tag{6.13}$$

where ε is a small non-negative constant. If L is the identity operator, this is equivalent to the Hölder-type condition with exponent being $\frac{1}{2}$, cf. [3, 9].

Remark 6.2. To see if assumption (6.13) is reasonable in our case, note that identity (5.5) implies

$$L^*L(\mathbf{q}_{k+1} - \tilde{\mathbf{q}}) = A'^*(\mathbf{q}_k)\frac{[A(\mathbf{q}_k) - \frac{I_\delta}{K(\mathbf{q}_k)} - A'(\mathbf{q}_k)(\mathbf{q}_{k+1} - \mathbf{q}_k)]}{\tilde{\alpha}_k} =: A'^*(\mathbf{q}_k)w_k. \tag{6.14}$$

Hence, in terms of structure, it is only natural to require that L and $\tilde{\mathbf{q}}$ satisfy (6.13). With our particular choice of L according to (5.6), condition (6.13) does not restrict the unknown parameters to any subspace. On the contrary, it enforces appropriate scaling, which results in a more effective regularization. The appearance of $\tilde{\alpha}_k$ in the denominator of (6.14) highlights the importance of driving $\tilde{\alpha}_k$ to zero at a rate that is not too fast to ensure that w_k remains bounded (accuracy and stability are well balanced).

By (6.13), there is $w \in S$ such that

$$L^*L(\hat{\mathbf{q}} - \tilde{\mathbf{q}}) = (A'(\hat{\mathbf{q}}) - A'(\mathbf{q}_k))^* w + A'^*(\mathbf{q}_k)w.$$

This yields the following inequality:

$$\tilde{\alpha}_k \|[A'^*(\mathbf{q}_k)A'(\mathbf{q}_k) + \tilde{\alpha}_k L^*L]^{-1} L^*L(\hat{\mathbf{q}} - \tilde{\mathbf{q}})\| \le \frac{M_2\varepsilon}{c}\|\mathbf{q}_k - \hat{\mathbf{q}}\| + \frac{\varepsilon}{2}\sqrt{\frac{\tilde{\alpha}_k}{c}}. \tag{6.15}$$

Taking into account (6.4)–(6.15), one arrives at the estimate

$$\|\mathbf{q}_{k+1} - \hat{\mathbf{q}}\| \le (1 - \lambda_k)\|\mathbf{q}_k - \hat{\mathbf{q}}\| + \lambda_k[2\mu NM_1 + 1]\frac{M_2\|\hat{\mathbf{q}} - \mathbf{q}_k\|^2}{4\sqrt{\tilde{\alpha}_k c}} + \lambda_k \mu NM_1\|\hat{\mathbf{q}} - \mathbf{q}_k\|$$
$$+ \frac{\lambda_k \delta}{2\tilde{K}\sqrt{\tilde{\alpha}_k c}} + \frac{\lambda_k M_2 \varepsilon}{c}\|\mathbf{q}_k - \hat{\mathbf{q}}\| + \frac{\lambda_k \varepsilon}{2}\sqrt{\frac{\tilde{\alpha}_k}{c}}.$$

Introducing the notations

$$\gamma := 1 - \mu NM_1 - \frac{M_2\varepsilon}{c}, \quad \kappa := \frac{1}{2\tilde{K}\sqrt{c}}, \quad \beta := \frac{M_2}{4\sqrt{c}}[2\mu NM_1 + 1], \quad \sigma := \frac{\varepsilon}{2\sqrt{c}},$$

we obtain (6.1), which shows convergence of the RIRGN algorithm.

7 Concluding remarks

An inherently challenging problem in infectious disease modeling is parameter estimation, especially in the presence of limited data. At the onset of an epidemic, quantification of key parameters can help understand the epidemiology of invading pathogen, make predictions of the likely morbidity and mortality impact, as well as disease transmissibility and incidence over time, which in turn could guide a timely implementation of the most effective intervention strategies. For example, as evident from phenomenological models studied here, there is strong correlation between the final size of an epidemic and its turning point, a critical parameter for disease forecasting during the early epidemic growth phase. These models describe the epidemic dynamics in two phases of fast and slow infection spread with a transition (turning) point, at which the maximum rate of disease incidence occurs. In the slow phase of infection spread (after the turning point), the epidemic peaks and subsequently declines, and therefore the cumulative number of cases eventually saturates at the epidemic final size. However, the challenge in parameter estimation generally arises in the fast phase of epidemic spread before the turning point where the amount of data is inadequate given the number of unknown parameters.

In this paper, our goal was to explore the nature of instability of classical regularized Gauss–Newton-type algorithms for the estimation of important disease parameters at the fast phase of epidemic spread. To enhance computational properties of the Hessian approximation, we introduced a modified problem-oriented optimization procedure, which yields a substantial progress in the recovery of two crucial epidemiological parameters, namely, the epidemic size of an emerging disease and the expected turning point of the outbreak. The convergence analysis of the new method is proposed under sufficient conditions that are fully justified for the generalized Richards model used to recover these and other unknown parameters.

In our future work, a systematic comparison across various phenomenological models will be conducted in order to assess which model is the most effective for stable parameter estimation from early outbreak data.

In optimization algorithm, the regularization will be enforced through a special non-linear penalty term quantifying a sub-exponential growth rate of an emerging outbreak. Further studies will make use of incidence data for Avian Influenza, Middle East Respiratory Syndrome (MERS), and Zika virus among others.

References

[1] C. L. Althaus, Estimating the reproduction number of Zaire Ebolavirus (EBOV) during the 2014 outbreak in West Africa, (2014), DOI 10.1371/currents.outbreaks.91afb5e0f279e7f29e7056095255b288.

[2] A. Bakushinsky, Iterative methods for nonlinear operator equations without regularity. New approach, *Dokl. Russian Acad. Sci.* **330** (1993), 282–284.

[3] A. Bakushinsky and M. Kokurin, *Iterative Methods for Ill-Posed Operator Equations with Smooth Operators*, Springer, Dordrecht, 2005.

[4] F. Bauer, T. Hohage and A. Munk, Iteratively regularized Gauss–Newton Method for nonlinear inverse problems with random noise, *SIAM J. Numer. Anal.* **47** (2009), 1827–1846.

[5] F. Cavallini, Fitting a logistic curve to data, *College Math. J.* **24** (1993), no. 3, 247–253.

[6] G. Chowell, N. W. Hengartner, C. Castillo-Chavez, P. W. Fenimore and J. M. Hyman, The basic reproductive number of Ebola and the effects of public health measures: The cases of Congo and Uganda, *J. Theoret. Biol.* **229** (2004), 119–126.

[7] G. Chowell, C. Viboud, J. M. Hyman and L. Simonsen, The Western Africa Ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates, *PLOS Currents* **2015** (2015), DOI 10.1371/currents.outbreaks.8b55f4bad99ac5c5db3663e916803261.

[8] Q. Jin, Further convergence results on the general iteratively regularized Gauss–Newton methods under the discrepancy principle, *Math. Comp.* **82** (2013), no. 283, 1647–1665.

[9] B. Kaltenbacher, A. Neubauer and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, Radon Ser. Comput. Appl. Math. 6, Walter de Gruyter, Berlin, 2008.

[10] M. Kokurin and A. Bakushinsky, Iteratively regularized Gauss–Newton methods under random noise, in: *Inverse Problems and Applications* (Stockholm 2013), Springer Proc. Math. Stat. 48, Springer, Cham (2015), 1–14.

[11] S. Langer and T. Hohage, Convergence analysis of an inexact iteratively regularized Gauss–Newton method under general source conditions, *J. Inverse Ill-Posed Probl.* **15** (2007), 19–35.

[12] J. O. Lloyd-Smith, S. Funk, A. R. McLean, S. Riley and J. L. Wood, Nine challenges in modelling the emergence of novel pathogens, *Epidemics* **10** (2015), 35–39.

[13] S. Morse, J. Mazet, M. Woolhouse, C. Parrish, D. Carroll, W. Karesh and P. Daszak, Prediction and prevention of the next pandemic zoonosis, *The Lancet* **380** (2012), no. 9857, 1956–1965.

[14] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2000.

[15] A. Smirnova, On convergence rates for iteratively regularized procedures with a linear penalty term, *Inverse Problems* **28** (2012), no. 8, Article ID 085005.

[16] A. Smirnova, R. Renaut and T. Khan, Convergence and application of a modified iteratively regularized Gauss–Newton algorithm, *Inverse Problems* **23** (2007), no. 4, 1547–1563.

[17] A. Tikhonov, A. Goncharsky, V. Stepanov and A. Yagola, *Numerical Methods for the Solution of Ill-Posed Problems*, Kluwer Academic, Dordrecht, 1995.

[18] A. Tikhonov, A. Leonov and A. Yagola, *Nonlinear Ill-Posed Problems. Vol. 1 and 2*, Chapman & Hall, London, 1998.

[19] A. Tsoularis and J. Wallace, Analysis of logistic growth models, *Math. Biosci.* **179** (2001), 21–55.

[20] M. Turner, E. Bradley, K. Kirk and K. Pruitt, A theory of growth, *Math. Biosci.* **29** (1976), 367–373.

[21] C. Viboud, L. Simonsen and G. Chowell, A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks, *Epidemics* **15** (2016), 27–37.

[22] J. Weitz and J. Dushoff, Modeling post-death transmission of Ebola: Challenges for inference and opportunities for control, *Sci. Rep.* **5** (2015), DOI 10.1038/srep08751.

[23] A. Yagola, Ill-posed problems and methods for their numerical solution, in: *Optimization and Regularization for Computational Inverse Problems and Applications* (Beijing 2008), Springer, Berlin (2010), 17–34.

[24] MathWorks Matlab R2016b, https://www.mathworks.com/products/matlab/?requestedDomain=www.mathworks.com.

[25] World Health Organization, Frequently asked questions on Ebola virus disease, 2015, http://www.who.int/csr/disease/ebola/faq-ebola/en/.

[26] World Health Organization, Statement on the 1st meeting of the IHR Emergency Committee on the 2014 Ebola outbreak in West Africa, 2015, http://www.who.int/mediacentre/news/statements/2014/ebola-20140808/en/.

[27] World Health Organization, Ebola situation report – 30 March 2016, http://apps.who.int/ebola/current-situation/ebola-situation-report-30-march-2016.