Profile likelihood based analyses of infectious disease models

Christian Tönsing 1, Jens Timmer 1,2,3 and Clemens Kreutz 1,2

Abstract
Ordinary differential equation models (ODEs) are frequently applied to describe the temporal evolution of epidemics. However, ODEs are also utilized in other scientific fields. We summarize and transfer state-of-the-art approaches from other fields like Systems Biology to infectious disease models. For this purpose, we use a simple SIR model with data from an influenza outbreak at an English boarding school in 1978 and a more complex model of a vector-borne disease with data from the Zika virus (ZIKV) outbreak in Colombia in 2015/16. Besides parameter estimation using a deterministic multistart optimization approach, a multitude of analyses based on the profile likelihood are presented comprising identifiability analysis and model reduction. The analyses were performed using the freely available modeling framework Data2Dynamics (data2dynamics.org) which has been awarded as best performing within the DREAM6 parameter estimation challenge and in the DREAM7 network reconstruction challenge.

Keywords
Modeling, dynamical systems, ordinary differential equations, parameter estimation, uncertainty analysis, profile likelihood, identifiability analysis, model reduction, vector-borne disease models, Zika virus (ZIKV) disease

1 Institute of Physics, University of Freiburg, Freiburg im Breisgau, Germany
2 Center for Biosystems Analysis (ZBSA), University of Freiburg, Freiburg im Breisgau, Germany
3 BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg im Breisgau, Germany

Corresponding author:
Christian Tönsing, Institute of Physics University of Freiburg, Hermann-Herder-Str. 3, 79104 Freiburg, Germany.
Email: christian.toensing@fdm.uni-freiburg.de
1 Introduction

Mathematical models can be used to understand mechanisms of spreading of infectious diseases. If the populations are large enough, deterministic models like ordinary differential equation models (ODEs) can be utilized to describe the dynamics of the different stages and for deriving strategies for controlling an epidemic or for predicting future outbreaks\(^1\).

Most models describe the number or fraction of individuals in several population subsets, usually termed as compartments. In the well-known SIR model

\[
S \rightarrow I \rightarrow R, \tag{1}
\]

as an example, individuals are passing from susceptibles \(S\) to the infected subset \(I\) before becoming recovered \(R\). The velocity of transitions between compartments is determined by rate constants, usually considered as unknown parameters. Schemes like Equation (1) can be directly translated into sets of ODEs in generally yielding first order derivatives which are nonlinear with respect to the dynamic states.

A variety of approaches has been established to estimate parameters in such models from experimental observation\(^2\), using bayesian inference networks\(^3–6\), Markov chain Monte Carlo methods\(^7\) in combination with Gaussian Processes\(^8\) and with focus on robust bayesian inference\(^9\), as well as agent based methods\(^10\), spatio temporal autoregression models\(^11\), but also stochastic maximum likelihood analyses\(^12\) as well as multiple shooting for stochastic systems\(^13\). The major challenge is that the solution of ODE models is nonlinear with respect to parameters like initial conditions and transition rate constants. Therefore, many classical methods for parameter fitting and assessing uncertainties are only applicable to a strongly limited manner\(^14;15\).

The profile likelihood is a generalization of classical approaches like standard errors which are based on the Fisher-Information matrix and it has been shown that the profile likelihood provides reasonable confidence intervals for parameter estimation of ODE models. Moreover, the profile likelihood has been used to assess the identifiability of model parameters\(^16\), for the determination of informative observations\(^17\) and for the calculation of uncertainties for predictions\(^18\). Moreover, the profile likelihood pinpoints model components which can be removed without loosing statistical agreement with observations. It thereby provides a strategy for reducing ODE models down to a completely identifiable level of detail\(^19\).

In this article, these well-established modeling methods are illustrated in the setting of infectious disease modeling. Moreover, insights concerning numerically efficient optimization are provided. The analyses are performed using the MATLAB-based Data2Dynamics modeling environment\(^20;21\). The presented methodology and its implementation has been awarded twice within the Dialogue on Reverse-Engineering Assessment and Methods (DREAM) challenges for parameter estimation\(^22\) and network
reconstruction\textsuperscript{23}, there mentioned as ’Team Crux’. The software package is used for all further analyses. Setup scripts, model definitions and data files for all presented examples are available within the Data2Dynamics software on www.data2dynamics.org.

2 Methodology

We introduce the methodology using a canonic SIR model with the well-known data set of an influenza outbreak at an English boarding School in 1978. In a second step, we apply the same methods and the profile likelihood analyses to a more sophisticated model of the vector-borne disease model with data from the ZIKV outbreak in Colombia in 2015/16. Hereby the goal is illustration of profile likelihood analyses in a standard example and application study, rather than focusing on credible biological discoveries.

In both examples we only use raw data time series although additional data, i.e. prior information for model parameters would be available. The intention is to introduce a methodology which is also applicable in applications where less or no prior knowledge is available. Nevertheless, the presented approach can also account for multiple priors but it does not necessarily require it.

2.1 ODE models

In infectious disease models, a set of coupled ordinary differential equations (ODEs)

\[
\dot{x} = f(x, u, k)
\]

is often used to describe the dynamics of the compartments represented by the vector \(x\). The term compartment is equivalent to the terms species, compound or node which are more commonly used in other fields. In general the function \(f\) can be nonlinear and depends on the compartment states \(x(t)\) at time \(t\), on a possibly time depend input function \(u(t)\) and on the model parameters \(k\), such as transmission rate constants. The initial states of the system at time \(t = 0\) are described by \(x^{\text{init}} = x(t = 0)\). Thus, the time dependence of the system is entirely described by a set of kinetic parameters \(\theta = (k, x^{\text{init}})\) which contains both, rate constants \(k\) and initial conditions \(x^{\text{init}}\).

2.2 Observations

In addition to prior knowledge, empirical data is used in order to estimate the unknown parameters \(\theta\). Usually not all compartments are observed individually and linear combinations like the sum of several compartments may occur as observations. In order to compare the model dynamics simulated
for a candidate set of parameter values, empirical data

\[ y_{\text{obs}}(t_i) = g(x(t_i)) + \epsilon(t_i) \]  (3)

is mapped to the dynamics \( x(t_i) \) at time \( t_i \) via observation functions \( g \) which are not necessarily bijective and invertible. If a single compartment \( x_j \) is directly observed, \( g \) is the identity function, i.e. \( y_{\text{obs}} = x_j(t_i) + \epsilon(t_i) \). In other cases, \( g \) may include linear combinations of compartments or contain scaling parameters in the case of relative data or offset parameters if background correction is required. In our illustration examples, independent additive Gaussian errors \( \epsilon(t_i) \sim \mathcal{N}(0, \sigma^2) \) with unknown standard deviation \( \sigma \) is assumed. However, the methodology is discussed in terms of the likelihood which comprises also other error distributions. Moreover, \( \sigma \) can be described by an error model, e.g. \( \sigma^2 = \sigma^2_{\text{rel}} x^2 + \sigma^2_{\text{abs}} \) as it is assumed for the Zika virus example.

### 2.3 Maximum likelihood estimation

Maximum likelihood estimation is a general approach for estimating model parameters. The so-called maximum likelihood estimate

\[
\hat{\theta} = \arg \max_\theta L(y|\theta) = \arg \min \theta -2 \log (L(y|\theta))
\]  (4)

is asymptotically unbiased and has several beneficial statistical properties like consistency and efficiency. In practice, \(-2LL(\theta) = -2 \log (L(y|\theta))\) is usually minimized instead of maximizing the likelihood because of numerical reasons and since this is equivalent to least-squares estimation in the case of Gaussian observation errors.

For model calibration the model’s dynamics, i.e. the observables \( g \) are compared to the available data \( y_{\text{obs}} \). If prior knowledge about the model parameters is available it can be incorporated by penalizing the log-likelihood \( LL \). For Gaussian priors \( \bar{\theta}_i \) with standard errors \( \bar{\sigma}_i \) the penalized log-likelihood is given by

\[
-2 LL_{\text{pen}}(\theta) = -2 \log (L(y|\theta)) + \sum_i \left( \theta_i - \bar{\theta}_i \right)^2 \bar{\sigma}_i^2 .
\]  (5)

For ODE models, optimization of Equations (4) or (5) has to be performed numerically. In the following, the unpenalized log-likelihood Equation (4) is used for all methodology and analyses.
2.4 Integration and optimization

The solution of an ODE system is usually nonlinear with respect to the parameters. In consequence, also the objective function is strongly nonlinear and, thus, parameter estimation in ODE models is numerically challenging. In addition, evaluating the objective function requires numerical integration of the ODEs which is only feasible with limited accuracy. Therefore, it is not possible to calculate derivatives based on finite differences. Moreover, it has been shown that it is usually more efficient to optimize the logarithms of parameters. Another aspect is that even ODE integrators for stiff systems struggle with numerical instabilities for processes with extremely large differences in time scales. Therefore, the parameter space is typically restricted, e.g. to eight orders of magnitude. Although analytical solutions for small ODE systems exist, numerical integration is chosen throughout the manuscript to not restrict applicability of the methodology on small special cases. Due to the nonlinearity of the ODEs and because of the restricted amount of empirical data, likelihood landscapes typically exhibit multiple local optima.

Although it is theoretically possible for stochastic optimization algorithms to find the global optimum (potentially requiring an infinite number of steps), it has been shown that they are not applicable without tuning in the setting of signaling pathway models. Instead, the applicability of deterministic optimization in combination with a multistart approach using a sufficient number of different initial guesses has been proven in several applications. This strategy is implemented in the Data2Dynamics modeling environment and has been awarded in scientific benchmark challenges in the setting of parameter estimation as well as for network reconstruction and is applied in this article. Here we use a Gauss-Newton gradient-based trust region optimizer as implemented as function \texttt{lsqnonlin} in the optimization toolbox of MATLAB in combination with the CVODES initial value problem solver which is implemented in C. Gradient information of the likelihood with respect to the parameters is calculated by solving the so-called sensitivity equations

\[
\frac{d}{dt} \frac{dx}{d\theta} = \frac{\partial f}{\partial x} \frac{dx}{d\theta} + \frac{\partial f}{\partial \theta}.
\]

Because the partial derivatives $\partial f/\partial x$ and $\partial f/\partial \theta$ can be calculated analytically, the parameter sensitivities $\partial x/\partial \theta$ occur as solutions of another ODE system of dimension $\text{dim}(\theta) \times \text{dim}(x)$ which can be attached to the original ODE system (Equation 2) and then numerically integrated simultaneously.

2.5 Profile likelihood

Traditionally, confidence intervals for point estimates of parameters are calculated from the Fisher information which is derived from second derivatives of the log-likelihood around the maximum. This approach yields exact results for linear models but does not account for higher order derivatives as they
occur in the nonlinear setting. In contrast, the profile likelihood approach comprehensively evaluates the shape of the log-likelihood enabling the calculation of reliable confidence intervals of the estimated parameters also for nonlinear settings\textsuperscript{16,18,34–36}.

To calculate the profile likelihood, each parameter \( \theta_i \) is iteratively changed around its maximum likelihood estimate while reoptimizing all other parameters \( \theta_i \neq j \). Thus, the profile likelihood is defined as

\[
PL_i(p) = \min_{\{\theta | \theta_i = p\}} -2LL(\theta).
\] (7)

The uncertainty of a parameter can be assessed by the profile likelihood-based confidence interval

\[
CI_{PL}(\theta_i) = \{p | PL_i(p) \leq -2LL(\hat{\theta}) + \Delta_\alpha(\chi^2_1)\}
\] (8)

which represents all acceptable parameter values \( \theta_i \) with \( PL_i \) below the threshold defined by the \( \alpha \) quantile \( \Delta_\alpha(\chi^2_1) \) of the \( \chi^2_1 \) statistics\textsuperscript{37}.

Optimization of Equation (7) for each \( i \) and \( p \) yields parameter vectors which are suited as representative candidates for sampling the dynamics within the high-dimensional confidence region given by the \( \Delta_\chi^2_1 \) threshold. It has been shown, that using these representative parameters enables evaluation of prediction uncertainties\textsuperscript{17} and the identification of informative experimental designs\textsuperscript{22}.

### 2.6 Identifiability analysis

Identifiability of model parameters related to the model structure and the experimental data has been extensively discussed in the literature\textsuperscript{38–44}. Here, we summarize and apply a numerical approach which only requires numerical implementation of parameter optimization but has no limitations in terms of model size, model structure or mathematical prerequisites.

The approach based on the profile likelihood allows for a data based identifiability analysis of the estimated parameters\textsuperscript{16,17}, where three classifications are available. A parameter is considered as identifiable if its likelihood profile exceeds the statistical threshold given by \( -2LL(\hat{\theta}) + \Delta_\alpha(\chi^2_1) \), yielding finite confidence intervals. If the amount or quality of the available data is not sufficient to indicate a confidence bound in both directions, the parameter is termed as practically non-identifiable. Here, a unique point estimate may be obtained, but the profile does not exceed the statistical threshold in at least one direction and thus does not have a confined confidence interval.

For structural non-identifiability, there are several definitions in the literature. In the context of continuous noise-free observations, a model is commonly defined as structurally non-identifiable if for \( \theta \) and \( \theta' \) equality of the model equations \( x(t, \theta) = x(t, \theta') \) and \( g(t, \theta) = g(t, \theta') \) for all times...
t does not imply equality of the two parameter vectors $\theta$ and $\theta'$\textsuperscript{41}. Instead, we use the conceptually different definition from\textsuperscript{16} which is more appropriate for a finite number of time points with observation errors. Here, parameters are already termed as structurally non-identifiable if the estimated parameters for a given data set are not unique. This means that there are several maximum likelihood estimates for given model outputs $g(x_j)$ and observation times $t_i$. Except for special cases, this means that the respective non-identifiable parameters cannot be constrained at all and the profile is constant for all values of the parameter. Such a behavior indicates redundant parametrization of the model structure which cannot be resolved by available observations. Consequently, the individual influence of a structurally non-identifiable parameter on the experimentally observed model outputs can be compensated by re-optimization of the other parameters. These redundancies may be obvious in some cases, e.g. if two parameters only appear as a product. Nontrivial structural non-identifiability can be identified by Lie group theory\textsuperscript{45}. Identifiability analysis may also be applicable when using parameter priors via the penalized log-likelihood (Equation 5).

2.7 Model reduction

As every model is a simplification of complex natural phenomena, where some features may be over-represented while others are neglected, it can be reasonable to adjust the complexity of the model to the available information in experimental data. Too complex models are more likely to overfit measurement errors and then providing inaccurate predictions. The model reduction strategy discussed here, utilizes the profile likelihood based identifiability analysis for indicating simplification which are not rejected statistically by available data.

For structural non-identifiabilities, re-parameterisation based on the underlying symmetry transformation, e.g. by fixing the parameter to a reasonable, user-defined value yields the desired outcome. For practical non-identifiabilities, the profile likelihood can be utilized as a data-based method for reducing nonlinear models\textsuperscript{19} since the profile likelihood calculation drives the model parameters towards their extreme values, i.e. minimal and maximal values, while keeping the model’s output in agreement with the data. Furthermore, the definition of the profile likelihood $PL_i$ (Equation 7) coincides with the test statistic of the likelihood ratio test\textsuperscript{46} which is a well-established statistical test for assessing whether reducing the model is rejected by the data. A parameter profile which flattens out to a constant level below the threshold for the limit of small values of the parameter close to zero, points to a simplification where the parameter value is set to zero, i.e. the corresponding reaction is removed from the model. Iterative application of such a procedure enables elimination of all practically non-identifiable parameters, yielding a desirable fully identifiable model\textsuperscript{19}. As in every step-wise model selection procedure, the order of the reduction steps might have an impact on the result. Solving such
ambiguities depends on the questions which have to be answered by the model and are beyond the scope of this paper.

3 Influenza Model

The applicability of the introduced methodology is demonstrated in the following. Therefore, it is applied to the frequently analyzed standard data example of an influenza outbreak at an English boarding school using a basic SIR model\(^{47-51}\).

3.1 Model equations

The so-called SIR model

\[ S \xrightarrow{\beta I} I \xrightarrow{\gamma} R \]  

was first introduced by Kermack and McKendrick\(^ {52}\) and consists of three compartments: susceptible individuals \(S\) which are not yet infected, infected individuals \(I\) showing symptoms and are able to transmit the disease, as well as recovered individuals \(R\) which, after the infection, are assumed to be immune to reinfection. The change of the population numbers in the compartments is described by the model equations

\[
\begin{align*}
\dot{S} &= -\frac{\beta SI}{N} \\
\dot{I} &= \frac{\beta SI}{N} - \gamma I \\
\dot{R} &= \gamma I \\
N &= S + I + R,
\end{align*}
\]

where \(N\) is the total number of individuals. Susceptibles \(S\) are infected by infected \(I\) with transmission rate \(\beta\) and velocity of transmission from the infected \(I\) to recovered \(R\) is described by rate constant \(\gamma\). Both rate parameters have the physical unit \(1/time\). The initial value of recovered individuals \(R^{init}\) is set to zero, whereas the initial values \(S^{init} = N_S\), which represents the total number of humans which can be infected and \(I^{init} = N_I\), which is the number of initially infected humans, i.e. the source of infection, are considered as parameters to be estimated and have the unit \(\text{number of humans}\).

Although, analytical solutions are available for the simple SIR model, we use numerical methods to integrate the ODEs to not restrict applicability to analytically solvable special cases.

Prepared using sagej.cls
3.2 Data

The data set from an influenza outbreak at an English boarding school in 1978 contains a total of 763 boys. As the school administration kept track about daily numbers of boys staying at bed, a complete, direct, and time-resolved record of the model compartment of infected individuals $I$ is available. One observation per day is available. As the number of infected boys is directly observed, the observation function is the identity and the model is linked to the data via

$$y_{\text{obs}}(t_i) = I + \epsilon(t_i),$$

where $\epsilon \sim \mathcal{N}(0, \sigma_{\text{abs}}^2)$ with a priori unknown standard deviation $\sigma_{\text{abs}}$ which is treated as an additional parameter to be estimated.

The total number of susceptible individuals $N^\dagger_S = 762$ was published with the data and a single infected individual $N^\dagger_I = 1$ has been assumed as starting point of the epidemic. However, this information will not be included into the presented analysis, but in contrast is predicted from the model only by the time-resolved empirical data of infected individuals $I$.

3.3 Parameter estimation

Figure 1 shows the data as well as the best fit and the respective trajectories for the compartments of the susceptibles and the recovered.

The model was fitted by optimization of the likelihood from 100 different initial guesses leading to a clearly identifiable global optimum. The uniformly distributed initial guesses for optimization are drawn randomly from the parameter search space which spans nine orders of magnitudes from $10^{-5}$ to $10^4$ in every dimension. Figure 2 shows the likelihood of the initial guess as well as after optimization. The optimization runs are ordered by the final likelihood value. 67 of the 100 fits converge to the same optimum which is indicated by the same final value of the objective function and, as shown in the lower panel, by the same parameter values (up to numerical tolerances). All other fits start from regions in the parameters space from which the deterministic optimizer is not able to locate the global optimum but, instead, converges to local optima with substantial inferior likelihood values. The lower panel of Figure 2 indicates, that for the suboptimal local optima, some parameters values like $\gamma$ and $\sigma_{\text{abs}}$ exhibit patterns, while others like $N_S, N_I$ and $\beta$ seem to be rather randomly distributed. The parameter values of the best fit are provided in Table 1.
Figure 1. Model dynamics of the three compartments for the best fit parameters provided in Table 1 with data for infected individuals. Gray bands represent estimated observation error.

Table 1. Best fit parameter values as well as confidence intervals for the SIR model for the English boarding school influenza data.

<table>
<thead>
<tr>
<th>parameter</th>
<th>physical unit</th>
<th>parameter search region</th>
<th>estimated parameter value</th>
<th>profile likelihood based confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_I$</td>
<td>humans</td>
<td>$10^{-5} - 10^4$</td>
<td>2.89</td>
<td>[0.54 – 8.88]</td>
</tr>
<tr>
<td>$N_S$</td>
<td>humans</td>
<td>$10^{-5} - 10^4$</td>
<td>854</td>
<td>[565 – 2324]</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$1/t$</td>
<td>$10^{-5} - 10^4$</td>
<td>1.74</td>
<td>[1.57 – 2.06]</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$1/t$</td>
<td>$10^{-5} - 10^4$</td>
<td>0.51</td>
<td>[0.34 – 1.10]</td>
</tr>
<tr>
<td>$\sigma_{abs}$</td>
<td>humans</td>
<td>$10^{-5} - 10^4$</td>
<td>19.7</td>
<td>[14.2 – 30.1]</td>
</tr>
</tbody>
</table>

3.4 Profile likelihood and identifiability analysis

Once the global optimum is found, likelihood profiles according to Equation (7) are calculated to derive confidence intervals for the estimated parameters. The resulting likelihood profiles shown in Figure 3 exceed the threshold for 95% confidence intervals indicating identifiability of all parameters in this example.

The literature values of susceptible individuals $N_S^\dagger = 762$ and the infected individual $N_I^\dagger = 1$ meet the estimated values for parameters $N_S$ and $N_I$ as they lie within their 95% profile likelihood based confidence intervals, c.f. Table 1.
Figure 2. Results from multistart optimization with 100 initial guesses. The initial likelihood values are plotted as crosses, the end-points after the optimization are depicted as triangles. For illustration, likelihood values are shifted to the baseline by subtraction of the likelihood value for the global minimum. The lower panel shows the estimated parameter values of the respective optimization runs. The 67 best performing optimization runs converge not only to the same likelihood value, but also to the same global optimum in the parameter space, as indicated by the horizontal pattern of the parameter values. In contrast, the other optimization runs converging to the suboptimal minima share the same final likelihood value, but converge to different regions in the parameter space. Some of these regions are characterized by similar end-points for some parameters, while other parameters spread over the whole parameter search space.

Each point on the profile likelihood corresponds to an individual parameter vector obtained by fixing one parameter to a specific value and reoptimizing the other parameters. Within the confidence intervals, the objective function of these parameter vectors does not exceed the threshold given by the selected significance level. Therefore, all the respective trajectories shown in Figure 4 are statistically in accordance with the examined data. For observed quantities, the trajectories differ only slightly from observations. In contrast, predictions for unobserved quantities can have large uncertainties indicated by a large spread of the trajectories. Figure 4B shows prediction trajectories for the number of susceptibles $S$ and recovered $R$ individuals as well as for the two transition fluxes (Figure 4C). The trajectories for the
susceptibles and recovered show that the measurements for the infected individuals alone do not provide accurate information about the total number of individuals at risk $N_S$.

### 3.5 Conclusions

The SIR model can describe the data of infected school boys, using the best fit parameters from a deterministic multistart optimization sequence. Likelihood profiles indicate a fully identifiable model and yield finite confidence intervals for all estimated parameters. Furthermore, trajectories for unobserved compartments can be used to predict their dynamics.

As the used data set is a standard basic example in the literature, the estimated parameter values can be compared to earlier analyses. Our estimates parameters shown in Table 1 are close to the reference values from $^{48-51}$ which are in agreement with the profile likelihood based confidence intervals.

In the case of the simple SIR model, the basic reproduction number $R_0$, i.e. the number of secondary cases from a primary infected individual $^{53}$ can be easily computed from the estimates as

$$R_0 = \frac{\beta}{\gamma} = 3.4,$$  \hspace{1cm} (12)

which is close to values estimated from literature for the same data and model. The estimated values of $R_0$ for the English boarding school data set is slightly larger than assumed for influenza in general and the estimated infection period is slightly shorter than expected $^{54;55}$. However, this can be explained by the ideal circumstances for an epidemic because of higher infection rates for school-aged individuals $^{56}$ and

---

*Figure 3. Likelihood profiles for all estimated parameters (solid lines). Best fit parameter values are indicated by gray asterisk. Parameter values at which the profile hits the $\Delta_{\alpha}(\chi^2_1)$ threshold (dashed line) are identifiable with lower and upper confidence bounds.*
the well-mixing of the individuals in a boarding school in contrast to other observed influenza outbreaks, e.g. in larger cities or regions.

Although, priori information for the estimated parameters would be available and could be integrated, it was intentionally not included into the analysis. For instance, the number of susceptibles $N_S$ was published with the data and there are also other data sources which would, e.g. provide detailed information on infection period durations for the specific influenza type. Nevertheless, published values from earlier analyses for single parameters as well as for the basic reproduction number $R_0$ are in accordance with the profile likelihood based confidence intervals$^{48–51}$.

4 Zika virus disease model

In the following, the presented methodology is applied to empirical data from the Zika virus disease outbreak in Colombia in 2015/16 using a complex vector-borne disease model.
The Zika virus disease is caused by the the Zika flavivirus (ZIKV) and is transmitted mainly by bites of *Aedes aegypti* mosquitoes\textsuperscript{57}. Typical acute symptoms include fever, maculopapular skin rashes, conjunctivitis, retro-orbital pain and headache. Humans overcome ZIKV infection usually after days to weeks and do not have long-term effects\textsuperscript{58}. However, a causal link to microcephaly\textsuperscript{59} and Guillain-Barré syndrome\textsuperscript{60} has been reported\textsuperscript{61}. The latest Zika virus disease outbreak in the Americas in 2015 thus became of global interest and was declared as an emergency of international concern by the World Health Organisation (WHO)\textsuperscript{62}.

In order to describe and possibly intervene the spread of the disease, many different models were discussed in the literature. Mainly models adapted from other mosquito-borne diseases such as malaria\textsuperscript{63;64}, dengue fever or chikungunya virus (CHIKV) have been adapted in order to analyze the dynamics of ZIKV outbreaks\textsuperscript{65–67}. Amongst others, the motif of human-to-human infection is discussed and integrated into the models in order to analyze its influence on the epidemic. Presumably, because of the global public interest and as publicly accessible empirical data of ZIKV infected humans is available from the national health institutes of most affected countries, dynamical ZIKV infectious models became a rapidly growing research interest.

### 4.1 Model

To describe the ZIKV transmission via humans and mosquitos, we use a model, which includes several of the discussed features of mosquito-borne infection resulting in a complex ZIKV model as shown in Figure 5A. The core of vector-borne disease models is a so-called *SEIR* submodel for humans combined with an *SEI* submodel for mosquitos: It is commonly assumed that a human individual that is infected with the Zika virus belongs in a so-called exposed ($E$) compartment for the incubation time of a couple of days\textsuperscript{58;68} before the virus can be transmitted to other individuals such as mosquitos or humans. Thus, the classical *SIR* model is extended by an intermediate compartment $E$, leading to the so-called *SEIR* model.

Analogously, an exposed compartment is assumed within the propagation of the disease in mosquitos. However, since the lifespan of mosquitos is short compared to typical timescales of recovering from a Zika virus infection, there is no recovered compartment for mosquitos, resulting in the mentioned *SEI* model. A mosquito may progress from susceptible $S_v$ to exposed $E_v$ when infected by an infected human with transmission rate $\beta_{hv}$ and converts to infected $I_v$ with rate $\nu_v$ afterwards.

Additionally to the classical *SEIR*-SEI model, the following extensions are introduced into the model. Birth of susceptible mosquitos is assumed to be relative to the total number $N_v = S_v + E_v + I_v$ with rate $\mu_v$. Since global changes of the total mosquito population should not be covered by the model, death with the same rate $\mu_v$ is assumed to all three mosquito compartments\textsuperscript{65;70}. When biting susceptible humans $S_h$, infected mosquitos $I_v$ transmit the virus with rate $\beta_{vh}$.
Figure 5. A: Model structure of the full ZIKV infection model as described by Equations (13). Mosquitos progress from susceptible $S_v$ to exposed $E_v$ when infected by a contagious human and convert to infected $I_v$ with rate $\nu_v$. Birth and death of mosquitos is described by rate $\mu_v$. When bitten by an infected mosquito, susceptible humans $S_h$ progress with rate $\beta_{vh}$ either to asymptomatically infected $I_{h,a}$ or symptomatically infected $I_{h,s}$, before they convert to convalescent $I_{h,c}$ with rate $\gamma_h$ and finally to recovered $R_h$. The proportion between symptomatically and asymptomatically infected is controlled by $\kappa_{as}$. All infected humans can transmit the virus to susceptible humans with rate $\beta_{hh}$, whereas only symptomatically and asymptomatically infected humans infect susceptible mosquitos with transmission rate $\beta_{hv}$. Only the cumulative incidence of symptomatically infected humans is observed (c.f. Equation 14). B: Data of newly infected humans and calculated cumulative sum of infected humans.

For the ZIKV there is a considerable amount of asymptomatically infected humans of approximately 80% reported\textsuperscript{58}, which do not show any symptoms but are able to spread the disease\textsuperscript{71}. This is included in the model by splitting up the compartment of infected humans $I_h$ to asymptomatically infected $I_{h,a}$ and symptomatically infected $I_{h,s}$ humans. The proportion of exposed humans becoming asymptomatically infected is represented by the parameter $\kappa_{as}$. Since this parameter is of great interest in practice and only a rough estimate is available a priori, $\kappa_{as}$ is estimated simultaneously with all other model parameters.

As suggested in\textsuperscript{68}, compartments $I_{h,S}$ and $I_{h,A}$ both transmit to an additional compartment of so-called convalescent humans $I_{h,c}$ with rate $\gamma_{h1}$ before they proceed with rate $\gamma_{h2}$ to the compartment of recovered humans $R_h$\textsuperscript{68}. Sexually transmitted direct human-to-human infection has been reported\textsuperscript{72,73}.
and is incorporated into the model by human-to-human transmission rate $\beta_{hh}^{68,74}$. It has been shown, that the virus persists longer in semen than in serum$^{75,76}$. Thus, besides asymptomatically infected $I_{h,a}$ and symptomatically infected $I_{h,s}$ humans, in the model also convalescent humans $I_{h,c}$ are able to transmit the virus to susceptible humans, whereas mosquitos cannot be infected by biting convalescent humans$^{68}$.

Since virus transmission rates are assumed to be orders of magnitudes faster than human birth or death rates, they are completely neglected in the model. The initial value of convalescent humans is $I_{h,c}^{\text{init}} = 0$, as well as for recovered humans $R_{h}^{\text{init}} = 0$. All other initial values are estimated comprehensively with the model parameters. The model equations for the described model read

\begin{align}
\dot{S}_v &= \mu_v N_v - \frac{\beta_{hv}S_v(I_{h,a} + I_{h,s})}{N_h} - \mu_v S_v \\
\dot{E}_v &= \frac{\beta_{hv}S_v(I_{h,a} + I_{h,s})}{N_h} - \nu_v E_v - \mu_v E_v \\
\dot{I}_v &= \nu_v E_v - \mu_v I_v \\
\dot{S}_h &= -\frac{\beta_{vh}S_h I_v}{N_v} - \frac{\beta_{hh}S_h(I_{h,a} + I_{h,s} + I_{h,c})}{N_h} \\
\dot{E}_h &= \frac{\beta_{vh}S_h I_v}{N_v} + \frac{\beta_{hh}S_h(I_{h,a} + I_{h,s} + I_{h,c})}{N_h} - \nu_h E_h \\
\dot{I}_{h,a} &= \kappa_{as}\nu_h E_h - \gamma_{h1}I_{h,a} \\
\dot{I}_{h,s} &= (1 - \kappa_{as})\nu_h E_h - \gamma_{h1}I_{h,s} \\
\dot{I}_{h,c} &= \gamma_{h1}(I_{h,a} + I_{h,s}) - \gamma_{h2}I_{h,c} \\
\dot{R}_h &= \gamma_{h2}I_{h,c} \\
N_v &= S_v + E_v + I_v \\
N_h &= S_h + E_h + I_{h,a} + I_{h,s} + I_{h,c} + R_h.
\end{align}

The unit of the rates $\beta_{hh}, \beta_{vh}, \beta_{hv}, \nu_v, \nu_h, \mu_v, \gamma_{h1}$ and $\gamma_{h2}$ is the inverse of the time $t^{-1}$, for the initial values $S_h^{\text{init}}, E_h^{\text{init}}, I_{h,a}^{\text{init}}, I_{h,s}^{\text{init}}$ and $I_{h,c}^{\text{init}}$ the unit is number of humans, for initial values $S_v^{\text{init}}, E_v^{\text{init}}$ and $I_v^{\text{init}}$ it is number of mosquitos and $\kappa_{as}$ is dimensionless.

### 4.2 Data

Empirical data of newly infected humans is published by Colombia’s National Health Institute on a weekly basis. These numbers contain also delayed reported cases from earlier weeks. As numbers of delayed reported cases are in the same order of magnitude as the current new infections, they should not be neglected. Hence, being the only continuously updated data source, graphical data from the weekly
reports had to be used. For reliably extracting the approximate number of reported cases, we used the WebDigitizer tool. The extracted data is depicted in Figure 5B.

In contrast to the first example, the current number of infected humans is not observed directly. Instead, the weekly number of newly infected individuals is recorded. Thus, the cumulative sum of newly infected is calculated and compared to the model, although fitting of cumulative incidence data can cause issues concerning underestimation of confidence intervals sizes. However, identifiability analyses and model reduction methods should not be affected. The cumulative sum of symptomatically infected humans is assessed in the model via the integral over the influx to $I_{h,s}$ so that the observation function reads

$$y_{obs}(t_i) = x_{obs} + \epsilon(t_i) = \int (1 - \kappa_{as}) \nu_h E_h dt + \epsilon(t_i).$$

An error model is chosen for which $\epsilon \sim \mathcal{N}(0, \sigma^2)$, with

$$\sigma^2 = \sigma^2_{rel} x_{obs}^2 + \sigma^2_{abs}$$

with relative $\sigma^2_{rel}$ and absolute $\sigma^2_{abs}$ error parameters, which are treated as additional parameters to be comprehensively estimated with other model parameters. In the presented approach, only information contained in the data is utilized to estimate the parameters and to assess their uncertainties, so that no prior knowledge from the literature, e.g. for transmission rates or average duration times of infection states will be integrated into the analysis.

### 4.3 Parameter Estimation

In order to obtain parameter estimates using the introduced model and data, the multistart fitting procedure with 1000 initial guesses was performed. In addition to several suboptimal local minima, the global optimum according to minimal objective function could be identified in 163 runs (Figure 6A). Using the best fit parameters, the model is able to describe the data adequately, as shown in Figure 6B.

Inspection of the likelihood profiles indicates that the estimated parameters have large uncertainties, as many profiles are flat and reveal non-identifiabilities (Figure 7). As a consequence of structural non-identifiability of parameters $\gamma_{h2}$ and $\kappa_{as}$ and the practical non-identifiability of $\beta_{hh}, \beta_{vh}, \gamma_{h1}, E_v^{init}, I_v^{init}, S_h^{init}, S_v^{init}, \mu_v$ and $\sigma_{abs}$, the estimated values within the global minimum are not unique as depicted in the lower panel of Figure 6A. Even for the global likelihood optimum, there are very few horizontal structures in the parameter plot indicating many flat regions in the parameters space. Flat likelihood profiles shown in Figure 7 indicate large ranges of parameter values which are in accordance with the data. Therefore, the mere estimated values and the corresponding trajectories are of
Figure 6. A: Multistart optimization with 1000 fits for the full ZIKV infection model (Equations 13) with optimization end-point parameters. For illustration, likelihood values are shifted to the baseline by subtraction of the likelihood value for the global minimum. Search regions for the estimated parameters are: initial values $\in [10^{-5}, 10^{10}]$, kinetic parameters $\in [10^{-10}, 10^{5}]$, $\sigma_{abs} \in [10^{-5}, 10^{3}]$, $\sigma_{rel} \in [10^{-5}, 10^{-0.3}]$, $\kappa_{as} \in [0, 1]$. Computation time for 1000 fits was 56.9 min, i.e. approximately 3.4 s for an average single fit on a 3.4 GHz quad-core CPU. B: Best fit of observation function with data and model dynamics of compartments for the estimated parameters.

limited impact without assessing uncertainties. Since uncertainties of parameters translate to uncertainties of the compartment trajectories, the quality of the drawn conclusions is restricted.

4.4 Model reduction

In several circumstances, it is reasonable to reduce the model to a level of complexity which is identifiable for given data. In the following, model reduction based on the likelihood profiles is exemplified. Since a
Figure 7. Likelihood profiles for the full Zika model (Equations 13). Flat profiles of parameters $\gamma_h$ (panel e) and $\kappa_{as}$ (panel o) reveal structural non-identifiabilities, whereas the profiles of $\beta_{hh}$, $\beta_{vh}$, $\gamma_{h1}$, $E_{v}^{\text{init}}$, $I_{v}^{\text{init}}$, $S_{h}^{\text{init}}$, $S_{v}^{\text{init}}$, $\mu_v$ and $\sigma_{abs}$ exhibit practical non-identifiabilities. Computation time was approximately 40.4 min on a 3.4 GHz quad-core CPU.

In a first instance, the profile of the initial value parameter $S_{v}^{\text{init}}$ (Figure 7k) of susceptible mosquitos is investigated. It is only bounded to lower values as the likelihood profile is flat towards extreme large values. Also both remaining initial values of the mosquito population $E_{v}^{\text{init}}$ and $I_{v}^{\text{init}}$ (Figure 7g,i) exhibit practical non-identifiability but with profiles being flat to minus infinity. Obviously, the data of the infected humans does not provide information about the upper bound of the mosquito population size nor...
Table 2. Model reduction steps from the full Zika model (Equations 13) to the reduced Zika model (Equations 17).

<table>
<thead>
<tr>
<th>step</th>
<th>reduction</th>
<th>reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>introduction of $\kappa_{as}$ profiles $E_v^{init}$ and $I_v^{init}$ flat to $-\infty$, $S_v^{init}$ flat to $+\infty$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$\sigma_{abs} = 0$</td>
<td>profile flat to $-\infty$</td>
</tr>
<tr>
<td>3</td>
<td>$\mu_v = 0$</td>
<td>profile flat to $-\infty$</td>
</tr>
<tr>
<td>4</td>
<td>$\gamma_h2 = 0$</td>
<td>structural non-identifiable</td>
</tr>
<tr>
<td>5</td>
<td>$\gamma_h1 = 0$</td>
<td>profile flat to $-\infty$</td>
</tr>
<tr>
<td>6</td>
<td>$\kappa_{as} = 0.8$</td>
<td>structural non-identifiable</td>
</tr>
<tr>
<td>7</td>
<td>$\beta_{hh} = 0$</td>
<td>profile flat to $-\infty$</td>
</tr>
</tbody>
</table>

A lower bound for the exposed and infected mosquitoes. The only possibility to remove this redundancy is to use prior knowledge about the ratio of mosquitoes to humans in the initial states at time $t = 0$. The parameter $\kappa_{hv}$ is therefore introduced by

\[
S_v^{init} = \kappa_{hv}S_h^{init} \\
E_v^{init} = \kappa_{hv}E_h^{init} \\
I_v^{init} = \kappa_{hv}I_h^{init}
\]

with prior value $\bar{\kappa}_{hv} = 5$ and prior uncertainty $\bar{\sigma}_{\kappa_{hv}} = 5$ as suggested in the literature\textsuperscript{78,79}. This first step introduces a link between the total population size of humans and mosquitoes in order to facilitate the further investigation and is the only additional prior information included into the analysis.

As described in Section 2.7, non-identifiable parameters with likelihood profiles being flat to minus infinity allow for reduction in a way that parameters are set to zero. Consequently, a purely relative error model (c.f. Equation 15) is indicated by the likelihood profile (Figure 7p) of $\sigma_{abs}$ (reduction step 2). Likewise, $\mu_v$ (Figure 7l) can be set to zero and so the mosquito’s birth and death dynamics are removed from the model, as they cannot be constrained to lower values by the analyzed data (reduction step 3).

The practical non-identifiable parameter $\gamma_{h1}$ and the structural non-identifiable parameter $\gamma_{h2}$ (Figure 7d,e) describe the progression of infected humans $I_{h,a}$ and $I_{h,s}$ via the convalescent state $I_{h,c}$ into the recovered compartment $R_h$. As the human-to-human infection rate $\beta_{hh}$ (Figure 7a) is also in accordance with zero, no interaction of the dead end $\rightarrow I_{h,c} \rightarrow R_h$ with the rest of the model is left. Thus, the distribution of individuals and their progression within this submodule cannot be resolved by the data. In order to resolve this over-complexity of the model, the structural non-identifiable parameter $\gamma_{h2}$ can be set to an arbitrary value (reduction step 4) and $\gamma_{h1}$ can be fixed to zero due to the flat likelihood profile towards minus infinity (reduction step 5). Setting $\gamma_{h1} = \gamma_{h2} = 0$ is equivalent with the introduction
A structural non-identifiability is observed for $\kappa_{as}$, as its profile is flat over the entire parameter range and moreover shows a functional relationship with parameters $S_{h}^{init}$, $E_{h}^{init}$ and $I_{h}^{init}$, as illustrated in Figure 8A. This behavior originates from the fact that $\kappa_{as}$ represents the proportion of exposed humans $E_{h}$ moving to the compartment of asymptomatically infected humans $I_{h,a}$ and $1 - \kappa_{as}$ is the proportion moving to symptomatically infected humans $I_{h,a}$. The proportion parameter $\kappa_{as}$ can be tuned to arbitrary values because the same total amount of symptomatically infected is obtained when compensating with the coupled initial value of exposed humans $E_{h}^{init}$. As long as the branch of asymptomatically infected humans cannot be compared to any observation, the parameters cannot be decoupled and $\kappa_{as}$ remains structurally non-identifiable. Therefore, model reduction can be performed by either remove the module of asymptomatically infected humans $I_{h,a}$ from the model or to fix this parameter to a literature value. Here, we choose the latter and fix $\kappa_{as} = 0.8$ as reported in the literature $^{58}$ to be still able to interpret $\nu_{h}$ in terms of a general transmission rate from exposed to infected humans (reduction step 6).
The likelihood profile of $\beta_{hh}$ in Figure 8B indicates that the frequently discussed infection via human-human transmission rate $\beta_{hh}$ is not required to explain the available data and is statistically in accordance with zero (final reduction step 7). Thus, based on this analysis, there is no significant relevance of the human-to-human infection of the Zika virus, as already suggested in the literature.

4.5 Fully identifiable model after reduction

Multistart optimization of the fully reduced model

$$
\begin{align*}
\dot{S}_v &= -\frac{\beta_{hv}S_v(I_{h,a} + I_{h,scR})}{N_h} \\
\dot{E}_v &= \frac{\beta_{hv}S_v(I_{h,a} + I_{h,scR})}{N_h} - \nu_v E_v \\
\dot{I}_v &= \nu_v E_v \\
\dot{S}_h &= -\frac{\beta_{vh}S_h I_v}{N_v} \\
\dot{E}_h &= \frac{\beta_{vh}S_h I_v}{N_v} - \nu_h E_h \\
\dot{I}_{h,a} &= 0.8 \cdot \nu_h E_h \\
\dot{I}_{h,scR} &= 0.2 \cdot \nu_h E_h \\
N_v &= S_v + E_v + I_v \\
N_h &= S_h + E_h + I_{h,a} + I_{h,scR} \\
I_{h,a}^{init} &= 0.8 \cdot I_h^{init} \\
I_{h,scR}^{init} &= 0.2 \cdot I_h^{init} \\
S_v^{init} &= \kappa_{hv} S_h^{init} \\
E_v^{init} &= \kappa_{hv} E_h^{init} \\
I_v^{init} &= \kappa_{hv} I_h^{init}
\end{align*}
$$

with the same initial guesses as for the full model in the previous section exhibits a clear global optimum with unique parameter values (Figure 9A). Horizontal structures of the fitted parameters in the lower panel of Figure 9A support the uniqueness of the global optimum in the parameter space. Consequently, Figure 10 reveals identifiability of all estimated model parameters as well as finite profile likelihood-based confidence intervals, c.f. Table 3. Figure 9B shows the agreement of data and model.
Figure 9. A: Multistart optimization with 1000 runs for the fully reduced model. The global minimum is found in almost 50% of the runs and the parameter values within this optimum are unique. For illustration, likelihood values are shifted to the baseline by subtraction of the likelihood value for the global minimum. Computation time for 1000 fits was 44.1 min, i.e. approximately 2.6 s for an average single fit on a 3.4 GHz quad-core CPU. B: Agreement of model and data for the best fit parameters as well as trajectories of all compartments.

5 Summary

In this article, parameter estimation and profile likelihood based analyses have been applied to two examples of infectious disease models. Keeping the usage of prior knowledge for parameter values to a minimum, solely information contained in the data of infected individuals was used in order to estimate parameters and assess their uncertainties. The first example was a basic SIR model where all parameters were identifiable.
Figure 10. Likelihood profiles of the fully reduced and identifiable model (Equations 17). All parameters are identifiable, although the information about $\kappa_{as}$ exclusively originates from the assumed prior. Computation time was approximately 3.4 min on a 3.4 GHz quad-core CPU.

Table 3. Estimated parameter values for reduced Zika model (Equations 17).

<table>
<thead>
<tr>
<th>parameter</th>
<th>physical unit</th>
<th>search region</th>
<th>estimated parameter value</th>
<th>profile likelihood based confidence interval</th>
<th>assumed prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{hv}$</td>
<td>$1/t$</td>
<td>$10^{-10}$ - $10^5$</td>
<td>1.57</td>
<td>[1.17, 1.96]</td>
<td>-</td>
</tr>
<tr>
<td>$\beta_{vh}$</td>
<td>$1/t$</td>
<td>$10^{-10}$ - $10^5$</td>
<td>0.087</td>
<td>[0.016, 0.092]</td>
<td>-</td>
</tr>
<tr>
<td>$E_h^{init}$</td>
<td>humans</td>
<td>$10^{-5}$ - $10^{10}$</td>
<td>$1.30 \times 10^3$</td>
<td>[0.84 $\times 10^3$, 1.85 $\times 10^3$]</td>
<td>-</td>
</tr>
<tr>
<td>$I_h^{init}$</td>
<td>humans</td>
<td>$10^{-5}$ - $10^{10}$</td>
<td>189</td>
<td>[170, 211]</td>
<td>-</td>
</tr>
<tr>
<td>$S_h^{init}$</td>
<td>humans</td>
<td>$10^{-5}$ - $10^{10}$</td>
<td>$5.3 \times 10^5$</td>
<td>[5.1 $\times 10^5$, 5.6 $\times 10^5$]</td>
<td>-</td>
</tr>
<tr>
<td>$\nu_h$</td>
<td>$1/t$</td>
<td>$10^{-10}$ - $10^5$</td>
<td>0.04</td>
<td>[0.02, 0.07]</td>
<td>-</td>
</tr>
<tr>
<td>$\nu_v$</td>
<td>$1/t$</td>
<td>$10^{-10}$ - $10^5$</td>
<td>0.0038</td>
<td>[0.0013, 0.0144]</td>
<td>-</td>
</tr>
<tr>
<td>$\kappa_{hv}$</td>
<td>mosquitoes to human</td>
<td>0 - 20</td>
<td>5</td>
<td>[0, 15]</td>
<td>$\mathcal{N}(5, 5^2)$</td>
</tr>
<tr>
<td>$\sigma_{rel}$</td>
<td>-</td>
<td>$10^{-5}$ - $10^{-0.3}$</td>
<td>0.06</td>
<td>[0.05, 0.07]</td>
<td>-</td>
</tr>
</tbody>
</table>

As a second example, a comprehensive model of Zika virus infection via mosquitoes and with data of infected humans from Colombia was analyzed. Profile likelihood based analysis shows non-identifiability of several parameters. It can be concluded that the data does not provide information about
the duration of the acute or convalescent phase of the ZIKV infection disease in humans, the proportion of asymptptomatically infected humans or the population size of the mosquitos. Moreover, the data alone does not provide evidence for human-human infections. After elimination of non-identifiabilities, a minimal model with identifiable parameters could be derived.

It should be noted, that here we focused on a data based model reduction scheme, which points to a result where all aspects of the model can be restricted to finite confidence intervals by the analyzed data. The validity and benefit of such a strategy or the number of reasonable steps certainly strongly depend on the application setting. Therefore, model reduction always has to be augmented with plausibility checks to prevent wrong conclusions. However, the method nicely illustrates the amount and quality of information contained in the analyzed data. Furthermore, this procedure demonstrates the commonly strong dependence of infectious disease models on prior knowledge and the risk of drawing conclusions predominately based on prior information rather than on recorded data.

In both investigated examples, there is a good agreement between data and model and it was demonstrated how uncertainties about model parameters translate to compartment trajectories. Since the profile likelihood constitutes an intuitive generalization of classical approaches for quantification of uncertainties to nonlinear settings as they occur in ODE models, we enforce usage of the reviewed methodology in the infectious disease modeling field.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors acknowledge funding of the projects EA:sys (031L0080), LungSysII (0316042A) and LiSyM (031L0042) by the German Federal Ministry of Education and Research (BMBF).

References


Prepared using sagej.cls


