Parameter Identification and Forecast with a Biased Model

Amadi Miracle, Haario Heikki

DOI: 10.1007/978-3-031-11818-0_30

Parameter identification and forecast with a biased model

Miracle Amadi and Heikki Haario

Abstract A well known practical issue is to ascertain how well the parameters of a model can be identified so as to allow a legitimate inference. In most cases, models are biased and may not contain all the necessary features needed to fit the data well. Employing the simplest Ross model as an example, we illustrated that parameter identifiability can be a problem of three factors: model specification, noisy data and partially observed model. Kalman filtering technique was employed in order to produce an optimal estimate of the evolving state of the system based on the model and other information such as rainfall, while simultaneously estimating the model parameters using the Kalman filter likelihood. Markov Chain Monte Carlo (MCMC) was employed as a general tool to diagnose parameter identifiability. To show the performance of the methods, an illustrative example was given with malaria data from Kalangala district, Uganda. In the end, the parameters were more or less well identified although the posterior is larger than when a synthetic data was used.

1 Introduction

Determining how well the parameters of a model can be distinctly identified with the help of available data, has been a common practical issue. When model parameters are not identifiable, there is little reason to believe that estimated values are close to the actual values. Even for noiseless data, the data can be fit arbitrarily well by different combinations of parameter values for some model/data combinations, and the uncertainties in the model parameter estimations are boundless. This follows

Miracle Amadi
LUT School of Engineering Science, LUT University, Yliopistonkatu 34, Lappeenranta, Finland, e-mail: miracle.amadi@lut.fi

Heikki Haario
LUT School of Engineering Science, LUT University, Yliopistonkatu 34, Lappeenranta, Finland, e-mail: heikki.haario@lut.fi

from the fact that although some parameter estimates can be obtained from a given model, these estimates could easily be local estimates or an arbitrary set of estimates that can over-fit the observation data if proper identification of the actual values of the parameters is not done.

The Ross model [5] has since played a key role in the development of mosquito-borne pathogen transmission studies and has had major influence on the development of strategies for malaria control. Using two differential equations for the human and mosquito, the model presents the time evolution of the fraction of individuals in infected classes (i_h, i_m):

$$\begin{aligned} di_h &= mabi_m(1-i_h) - i_h r \\ di_m &= aci_h(1-i_m) - \mu i_m, \end{aligned} \tag{1}$$

where i_h and i_m represents the fractions of infected humans and mosquitoes, correspondingly, m denotes the mosquito-to-human ratio, b and c denote the transmission probabilities during mosquito contact with the human, μ is the mosquito mortality rate, a is the contact rate and r represents the recovery rate for humans. Based on benchmarks described in [6], the simpler models, such as the Ross model, appear to do a better job of matching data and heuristics than the more complex models. Here, we demonstrate how well the parameters of the simple Ross model can be identified based on available data and parameter selections. The data on reported monthly malaria cases for Kalangala district for six years (2006-2011) from Uganda, and the corresponding mean monthly rainfall data were employed in this study from World Weather Online.

2 MCMC parameter identification

Parameter identifiability is usually diagnosed using MCMC approach. This method is based on Bayesian inference and can be used to determine the reliability of parameter estimates as well as to quantify parameter confidence. Thus, by generating distributions of parameter values consistent with the available data, this method gives reliable estimates of model parameters (and associated uncertainties) and may be used to check whether those estimates are unique. Adaptive MCMC is used in this study since we may not be able to determine a well-working proposal distribution at the outset [2]. The Adaptive MCMC is an improved version of Metropolis-algorithm that updates the proposal covariance during the MCMC run, by using information of the previously sampled points. To evaluate the fit with the data, we utilized the cost function which returns the sum of squared differences between observations and model outputs while accounting for measurement error variance. The structure of the posterior distribution shows if the observables uniquely bound the model parameters. A helpful practice for seeing how well the chain is mixing, is to make a plot of the autocorrelation functions of the parameter chain, from which one can see the degree to which samples that are k steps away correlate with each other [2]. We would expect successive points to correlate more with each other than points further apart because in MCMC, next points are dependent on the previous points. Again, the model parameters are considered to be identifiable if the parameter values that are in the best agreement with the data are bound to a small region of the parameter space.

2.1 Factors influencing parameter identifiability

2.1.1 Partially observed model

Even for very basic models, partial observation of state variables frequently results in structural non-identifiability of model parameters. [4]. The Ross model employed for this study has a compartment for infected mosquitoes population which is hardly measurable. Thus, such data is not available for this study. One approach for tailoring the model complexity to the information content of the data is to reduce the model complexity in accordance with the available data, resulting in a reduction in the ODE system's dimension [4]. A method for addressing this problem was proposed in [7], based on the practical necessity that parameters be written as functions of the known quantities of the ODE system. In this work, considering that the presence of mosquito dynamics gives an additional degree of freedom, a reduced model which has only the infectious human compartment is proposed. Therefore, the equilibrium solution of infected mosquitoes given as

$$i_m^* = \frac{i_h}{i_h + \kappa}, \text{ where } \kappa = \frac{\mu}{ac}, \tag{2}$$

is plugged in and parameterised as

$$\frac{di_h}{dt} = mab\frac{i_h}{i_h + \kappa}(1 - i_h) - i_h r. \tag{3}$$

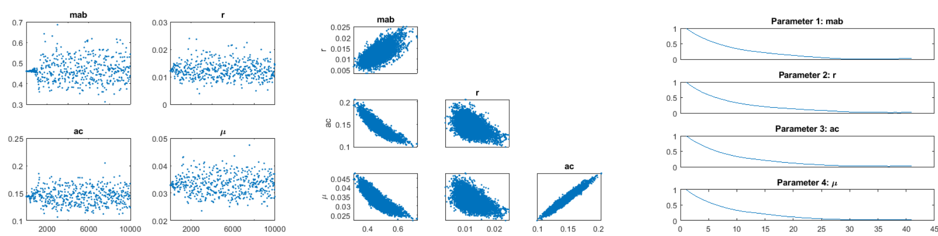
The parameter κ denotes the ration of mosquito mortality and infection rate. It can be small or large depending on the size of the infected mosquito population. In our preliminary analysis, the dynamics of the original and the reduced model are the same.

2.1.2 Interdependence and lack of influence of parameters

Non-identifiability can be caused by lack of influence of a parameter on the observables, as well as interdependence among the parameters. It is obvious that if a parameter has no effect on the observables, it is not possible to determine its value. On the other hand if a change in one parameter can be matched by a corresponding change in another, parameter identification can be difficult, since they may not be individually identifiable [1].

Despite the fact that there is no way to absolutely establish a model's structure, unsuccessful models can be ruled out if they fail to fit the available data for any set of parameters. When the available data lack the power to constrain a model's parameters significantly, it is possible that several other models of equivalent complexity are likely to match the data well. Thus, diagnosing identifiability is a first stage in the model selection process, in which possible models are ruled out if they are unable to be bound by available data. Given that models and parameters are evaluated simultaneously, the MCMC method for detecting parameter non-identifiability may also be employed for model selection.

We tested this using synthetic data generated by adding a gaussian noise to the output of the Ross model which was initially computed by employing values for the first set of parameters and initial conditions given in [6]. We found that all the six parameters are not well identified since the uncertainties in most of the model parameters are unbounded. However, from the nature of the mixing of the chains, it appears that some of the parameters are better expressed as products, with those products, taken as new parameters (see[1]). Following this approach, other parameterizations were evaluated, and we finally came up with one having four parameters as shown in Fig. 1. With this parameterization, the parameter chains have a very good mixing, with their associated levels of uncertainty, uniquely identified as can be seen in Fig. 1. Thus, we use this model parameterization in fitting the real data.



(a) (b) (c)

Fig. 1: **MCMC results for the chosen parameterization: (a) trace plot (b) pairwise correlation plot (c) autocorrelation plot.**

2.1.3 Biased Model

Bayesian identification procedure takes a long time to converge when the noise level is high [2]. Besides model parameter estimation, where the goal is to estimate static parameters, it is of interest to estimate the dynamically changing state of the system since the initial values of the system are not known. However, in some cases, the model state is not known precisely and it has to be estimated along with the parameters. State estimation in dynamical models can be done using filtering methods, where the distribution of the model state evolves along with the dynamical model and updated sequentially as new observations become available [3]. Another rationale behind filtering is that the model can have bias and may not contain all the necessary features required to fit the data well. For instance the Ross model alone does not have a provision to include the changing weather information. Thus, the filtering in this study incorporated the rainfall observations to the numerical model. The ODE was solved with the mosquito density m periodically following the rainfall with a linear model, using the time lag calculated by cross correlations and regression. The estimation of the time-lag was done by a separate analysis.

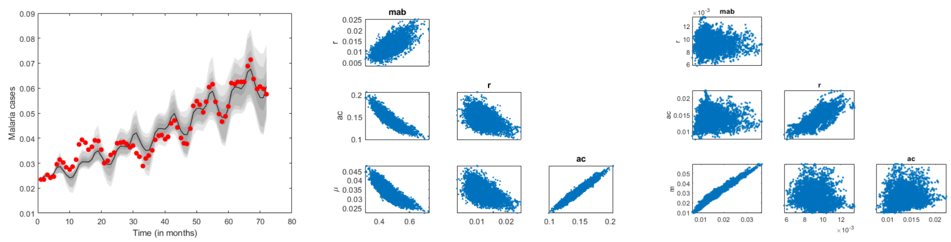
We considered the likelihood approach of implementing parameter estimation within a data assimilation system. The likelihood of a parameter value is computed by running a state estimation procedure over a specified data set while keeping the parameter value unchanged. The likelihood is computed using the filter residuals [3]. This is similar to traditional parameter estimation, but a state estimation technique is used to "integrate out" the uncertainty in the model state. Thus, two stages are involved in order to obtain the parameter estimates:

- a filtering method for computing the posterior density for a parameter value
- a parameter estimation algorithm for obtaining the estimates.

For the first task, we use the extended Kalman filter technique, since the model is non-linear. For the second task, we use the MCMC algorithm. For further reading on this approach, see [3].

3 Results

The result of the Kalman filtering done with the new ODE parameterization is given in this section. It can be seen from Fig. 2(a) that the parameters are properly identified. Overall, parameter identifiability improved at each step of rectifying the issues posed by the influencing factors. However, it can be seen from the plot of the two dimensional marginal distributions in Fig. 2(b-c) that the case with real data has a larger posterior density as compared to the case with synthetic data.



(a) (b) (c)

Fig. 2: **(a)The predictive posterior distribution of the state variable calculated from MCMC (light gray), a single prediction by MAP estimate (black bold line) and the data (red circles); The pairwise distribution plots for the case with (b) synthetic data (c) Real data**

4 Conclusion

In general, we acknowledge that identifiability could also be a property of likelihood and suggest that the nature of the proposed model in relation to the available data be studied before embarking on full MCMC implementation. Apart from allowing for the on-line estimation of model states with relevant sources of information, the Kalman filtering conducted reduces uncertainties and bias, and thereby improve forecasting. The present work could be regarded as a proof of concepts that can be employed to improve parameter identifiability and forecasting.

References

1. Brouwer, A.F., Meza, R., Eisenberg, M.C.: Parameter estimation for multistage clonal expansion models from cancer incidence data: A practical identifiability analysis. PLoS. Comput. Biol. **13**, e1005431 (2017)
2. Haario, H., Laine, M., Mira, A., Saksman, E.: DRAM: efficient adaptive MCMC. Stat. Comput. **16**, 339–354 (2006)
3. Hakkarainen, J., Ilin, A., Solonen, A., Laine, M., Haario, H., Tamminen, J., Oja, E., Jarvinen, H.: On closure parameter estimation in chaotic systems. Nonlinear. Process. Geophys. **19**, 127-143 (2012)
4. Miao, H., Xia, X., Perelson, A.S., Wu, H.: On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM. Rev. Soc. Ind. Appl. Math. **53**, 3–9 (2011)
5. Ross, R.: The prevention of malaria. London, UK (1911)
6. Wallace, D.I., Southworth, B.S., Shi, X., Chipman, J.W., Githeko, A.K.: A comparison of five malaria transmission models: benchmark tests and implications for disease control. Malar. J. **13**, 268 (2014)
7. Xia, X., Moog, C.H.: Identifiability of nonlinear systems with application to HIV/AIDS models. IEEE. Trans. Automat. Contr. **48**, 330–336 (2003)