Kinetic models of biological systems: robust identifiability analysis

Ana Arias-Méndez, Julio R. Banga, Eva Balsa-Canto

(Bio)process Engineering Group, IIM-CSIC, Spain
ebalsa@iim.csic.es

One of the most challenging tasks in systems biology is the reconciliation of the mathematical models with experimental data. Biological models usually contain unknown and unmeasurable parameters whose values must be estimated from data. Parameter estimation can be carried out by minimizing a cost function of the fit (like the log-likelihood) in combination with an adequate global optimiser (in order to avoid convergence to local solutions).

However, in the presence of scarce and noisy data a good fit does not guarantee unique parameter estimations. In other words, the parameters may vary significantly despite providing fits of equivalent quality. Practical identifiability analysis enables the possibility to asses the reliability of the estimates, both in terms of confidence regions and correlation. This information is critical to define the confidence on model predictions.

In linear models confidence intervals are proportional to the square root of the eigenvalues of the covariance matrix. Since, systems biology models are non-linear, and there is no general method to obtain the covariance matrix, it is usually approximated by the inverse of the Fisher information matrix. Unfortunately this approximation is only valid under specific conditions that are not fulfilled in most practical cases. In addition, high correlation of the parameters may induce rank deficient Fisher information matrices, thus precluding the computation of confidence intervals. More reliable results can be achieved with Monte Carlo based approaches. The underlying idea is to simulate a sufficiently large number of experiments (several hundreds), and then perform parameter estimation for all data realisations. The resulting population of parameter solutions is then analysed to obtain information about confidence and correlation.

Both the Fisher information matrix and Monte Carlo based approaches require information about the type and amount of experimental noise. In practise, Gaussian noise distribution is assumed with a given standard deviation. Nevertheless, whether the Gaussian assumption or the standard deviation are correct is, in most cases, unknown.

In this work, we investigate the possibility of using the Leave One Out (LOO) and Leave Several Out (LSO) methods, typically used for cross-validation, to study identifiability. In addition, we propose the use of a residuals based bootstrap approach (RBB) to enable the use of bootstrap methods to general (non stationary) time series data. The LOO method solves successive parameter estimation problems eliminating a single data point in each estimation, this translates in the necessity of solving n_d (number of data) parameter estimation problems. The LSO method was designed to solve a sufficiently large number of

parameter estimation problems (500 in this work) neglecting a random selection of $n_{lso} \geq \sqrt{n_d}$ data points in each estimation. The RBB approach generates a sufficiently large number of pseudo-experimental data realisations (500 in this work) by perturbing the residuals of the model at the best fit, the parameter estimation problem is then solved for all realisations. The populations of parameter solutions achieved in all cases are analysed to obtain information about confidence and correlation.

To study the properties of the proposed approaches we have selected a collection of examples representative of biological systems, covering different sizes and types of non-linear terms. Results achieved were compared to those reported in the literature or those achieved by either the Fisher information matrix or the Monte Carlo based approaches under the assumption that exact information about the experimental noise is available.

Results

Our results indicate that the LOO approach is not suitable to estimate confidence intervals, although it is quite reliable for the estimation of correlation between parameters. In the examples considered, this method tended to overestimate the confidence intervals when the number of data is limited and to underestimate them when the number of data is high.

The LSO and RBB alternatives provide pretty similar results, in agreement with those from the literature or those obtained with the Fisher information matrix and Monte Carlo based methods with exact information about the experimental noise. In addition both methods are capable of detecting lack of identifiability.

It should be noted that both methods are equally successful in those cases in which the ratio between the number of data and the number of parameters (n_d/n_θ) is large, i.e. $(n_d - \sqrt{n_d})/n_\theta >> 2$. The LSO method will tend to overestimate confidence intervals in those cases in which the data is limited, since lack of identifiability in each realisation may be induced by the reduction of $\sqrt(n_d)$ data.

As the main conclusion, the proposed RBB method presents clear advantages since it does not require prior knowledge of the experimental error or the parameters distribution and will provide confidence information even in highly correlated scenarios.

References

1. Banga, J.R., Balsa-Canto E.: Parameter estimation and optimal experimental design. Essays in Biochemistry 45:195210.(2008)
2. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. New York: Chapman & Hall, (1994)
3. Vanlier J., Tiemann C.A. , Hilbers P.A.J., van Riel N.A.W.: Parameter uncertainty in biochemical models described by ordinary differential equations. Mathematical Biosciences 2:305-314 (2013)