http://www.calibayes.ncl.ac.uk/

Bayesian calibration of biological simulation models

D.A.Henderson, W.E.Wolski, R.J.Boys, T.B.L.Kirkwood and D.J.Wilkinson

Schools of Mathematics & Statistics and Clinical Medical Sciences
University of Newcastle upon Tyne, UK

UNIVERSITY OF NEWCASTLE

Summary

Modern Systems Biology is greatly concerned with using post-genomic experimental data sources in order to "fit" dynamic and predictive computer models of complex non-linear biological processes. If this is successful, scientists will be able to conduct "in silico" biological experiments that would be impossible to carry out "in vivo" and these could, in turn, transform our understanding of molecular biology. Unfortunately there are a great many obstacles in the way of this grand vision, and many are fundamentally statistical in nature. Here, an outline will be presented of the key issues involved in model calibration, and the advantages that Bayesian calibration offers over other commonly used approaches. The methodology is applied to a case study concerning the cell-cycle in frog eggs.

1 Introduction

- Large, complex, quantitative biological models are gaining increasing significance for research hypothesis formulation in biological and medical science (Kirkwood *et al.*, 2003; Proctor *et al.*, 2005).
- These models are typically quantified using a set of parameters, such as kinetic rate constants.
- Determining appropriate values for these parameters directly from experiments can be a challenging task.
- The process of determining parameter values based on a comparison of experimental data with output from a computer model is known as calibration (Kennedy and O'Hagan, 2001).
- Calibration of complex computer codes is a well-studied problem in a range of different scientific disciplines. It turns out that the calibration problem is particularly well-suited to a Bayesian non-parametric statistical analysis.

2 Example: cell cycle in frog eggs

We demonstrate a Bayesian calibration approach by analysing a model of the cell cycle in frog eggs (Zwolak, Tyson and Watson, 2005).

- The model consists of eight Michaelis-Menten type kinetic equations describing four reversible reactions.
- The total number of parameters in the model is 12, out of which four parameters are Michaelis K_m-constants which were preset. The collection of the eight parameters is denoted $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_8)$.
- Using the experimental data, made up of eight 2-4 timepoint short time series ($n = 25$ datapoints in all), we were interested in determining values for the eight parameters.
- The computer model (simulator) for this system is a set of ordinary differential equations (ODEs), for which a fast numerical solver was available.

2.1 Bayesian calibration

- The computer model is fast to run and so can be evaluated at many sets of parameter values.
- It can therefore be treated as a deterministic (non-linear) function, $f(\boldsymbol{x}; \boldsymbol{\theta})$, of the parameter values $\boldsymbol{\theta}$ and any covariate information $\boldsymbol{x}$, such as initial quantities, experimental conditions, etc.
- This deterministic function is embedded into a stochastic model, which relates the experimental data $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ to the computer model via the regression relationship:

$$y_i = f(x_i; \boldsymbol{\theta}) + \sigma \epsilon_i.$$

We take $\epsilon_i \sim t_\nu$ (independently) for $i = 1, 2, \ldots, n$, that is, the stochastic component of the model is independent, zero mean Student t–distributed noise. We present results for $\nu = 3$ degrees of freedom – a distribution with much heavier tails than a Gaussian distribution.

- With the above stochastic model, the likelihood $L(\boldsymbol{\theta}, \sigma; \boldsymbol{y}, \boldsymbol{x})$ is easy to write down.
- Prior uncertainty about parameter values is expressed through uniform $U(-10, 10)$ distributions for $\log(\theta_i)$, $i = 1, 2, \ldots, 8$, and for $\log(\sigma)$.
- Standard Markov chain Monte Carlo (MCMC) methods can be used to sample from the (analytically intractable) posterior distribution of the unknown parameters $\pi(\boldsymbol{\theta}, \sigma | \boldsymbol{y}, \boldsymbol{x})$.

2.2 Results

Results for two of the eight model parameters are shown in Figure 1.

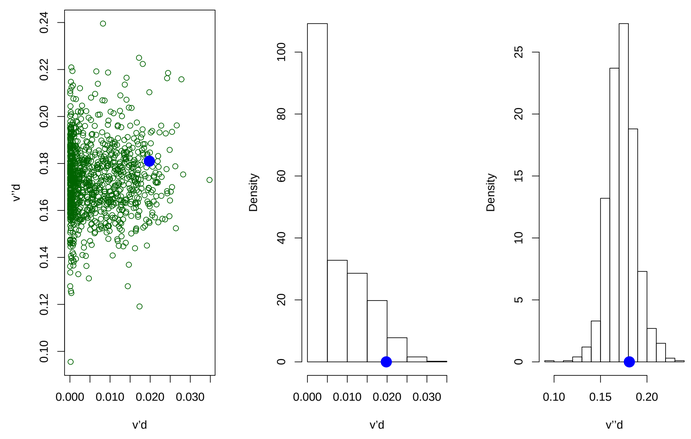

Figure 1: Scatterplot of sampled values from the joint posterior distribution of parameters $\theta_1 \equiv v'_d$ and $\theta_2 \equiv v''_d$, together with histograms representing their marginal posterior densities. The blue dots represent the values estimated in Zwolak *et al.* (2005).

The marginal posterior distributions highlight the uncertainty about the values of the parameters after observing the data in a way that is not easily obtained from a non-Bayesian analysis.

Our inference for the parameters v'_d and v''_d is largely consistent with that of Zwolak *et al.* (2005). This is true for most, but not all of the eight parameters.

However, our main focus is on predicting the output of the model, whilst accounting for the posterior uncertainty in the parameters. This allows us to perform *in silico* experiments; investigating the range of plausible outputs for a given set of experimental conditions.

Figure 2 illustrates how posterior uncertainty in the $\boldsymbol{\theta}$ values impacts on the output of the deterministic model $f(\boldsymbol{x}; \boldsymbol{\theta})$, for experimental conditions C and D from Zwolak *et al.* (2005).

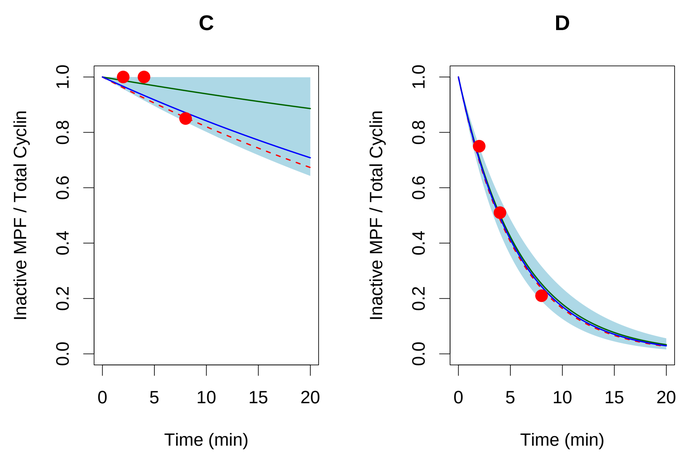

Figure 2: The red dots are the experimental data. The blue line is the simulator output corresponding to the set of parameters which gave the best match to the data (in terms of likelihood). The red dashed line is the output corresponding to the estimated parameters from Zwolak *et al.* (2005). The green line is the point-wise posterior mean of $f(\tilde{\boldsymbol{x}}; \boldsymbol{\theta})$, and the blue shading gives point-wise 95% posterior probability intervals for $f(\tilde{\boldsymbol{x}}; \boldsymbol{\theta})$.

As well as uncertainty, a point estimate of the parameter values can be obtained from the parameter values giving the best fit to the experimental data over the course of the MCMC simulation. The values we obtain give a better fit to the data (in terms of loglikelihood) than the estimated parameters from Zwolak *et al.* (2005).

Overall, this simple Bayesian approach seems to work well, even in this case-study involving sparse data, from different sources. The Bayesian approach allows the uncertainty in the parameter values and model output to be quantified in an appropriate way.

3 Inferential challenges

- The Bayesian calibration approach described in Section 2 relied on the fact that the computer model was fast to run, and so could be evaluated many times. For large and/or complex biological models, running the simulator at many input points can be prohibitive.
- Biological processes are intrinsically *stochastic*, and as such deterministic models can not always be used for making inferences. Perhaps the main obstacle in the way of calibration of stochastic simulators is the fact they are likely to be too slow to be used in a calibration process where many stochastic simulations are required.

When a (deterministic) simulator is slow to evaluate, one possible solution is to emulate it using a Gaussian process (GP) model.

3.1 Emulation via Gaussian processes

- Gaussian processes provide a flexible model for the output from a deterministic computer simulation model that can be fitted on relatively few data points.
- The GP can then be used as a (fast) emulator of the computer model.
- To illustrate this, we fit a Gaussian process to simulated data from the frog model using the software `PErK` (Santner *et al.*, 2003).
- The scatterplot in Figure 3 indicates that based on only 650 evaluations of the computer model, the GP can be used to emulate the computer model reasonably well.
- In the case of the frog model, evaluating the GP model at a set of parameter values $\boldsymbol{\theta}$ is approximately 4 times faster than running the fast ODE solver. The speed improvement will become larger for more complex models.
- Once estimated, the GP can be used as a surrogate for the computer model in methods to estimate model parameters.

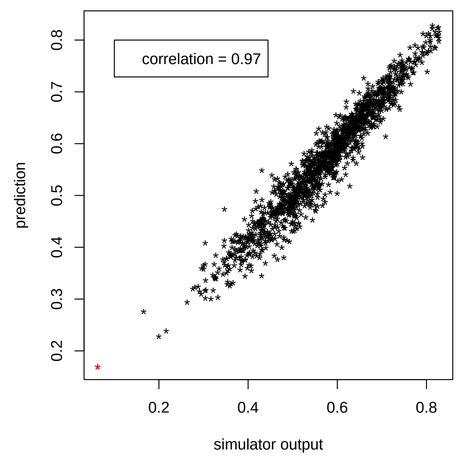

Figure 3: Scatterplot of predicted against true agreement with experimental data for a set of 1500 input parameters sampled uniformly from the parameter space.

4 Future work

- The ultimate goal of the CaliBayes project is to develop *fully* Bayesian methods which allow us to quantify all sources of uncertainty in the calibration process.
- The Bayesian calibration of *stochastic* models is of particular interest due to the inherent variability in biological processes which cannot be represented by a deterministic model.
- Progress has been made using exact and approximate Bayesian models (e.g. Boys *et al.* (2004), Golightly and Wilkinson (2005)) but these methods are not likely to be applicable when data are sparse, from different sources and of different types, and when the models are large and contain a mixture of low and high copy-number species.
- When a (black box) stochastic simulator is too slow to make many thousands of simulations, some form of emulation using a fast surrogate will almost certainly be required.

References

Boys, R. J., Wilkinson, D. J. and Kirkwood, T. B. L. (2004) Bayesian inference for a discretely observed stochastic kinetic model. Statistics preprint STA04.5, University of Newcastle.

Golightly, A. and Wilkinson, D. J. (2005) Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, **61**, 781–788.

Kennedy, M. C. and O'Hagan, A. (2001) Bayesian calibration of computer models (with discussion). *J. R. Statist. Soc.* B, **63**, 425–464.

Kirkwood, T. L. B., Boys, R. J., Gillespie, C. S., Proctor, C. J., Shanley, D. P. and Wilkinson, D. J. (2003) Towards an e-biology of ageing: integrating theory and data. *Nat. Rev. Mol. Cell Biol.*, **4**(3), 243–9.

Proctor, C. J., Soti, C., Boys, R. J., Gillespie, C. S., Shanley, D. P., Wilkinson, D. J. and Kirkwood, T. L. B. (2005) Modelling the actions of chaperones and their role in ageing. *Mech. Ageing Dev.*, **126**(1), 119–31.

Santner, T. J., Williams, B. J. and Notz, W. I. (2003) *The Design and Analysis of Computer Experiments*. New York: Springer.

Zwolak, J. W., Tyson, J. H. and Watson, L. T. (2005) Parameter estimation for a mathematical model of the cell cycle in frog eggs. *J. Comput. Biol.*, **12**, 48–63.

BBSRC Bioinformatics and e-Science Programme II, BBS/B/16550, January 2005 – December 2007