Vignette MOSAIC$_{\text{growth}}$

D. Wu, A. Ratier, G. Multari, C. Lopes, S. Charles

September 23, 2020

Contents

This document includes two parts. The first part describes the model used in the MOSAIC$_{\text{growth}}$ application to analyse the toxic effect of a contaminant on growth-type data. The second part describes how to account for the uncertainty on the x% effective concentration or rate (EC$_x$ or ER$_x$) in order to adequately censor it for future (species sensitivity distribution) SSD analyses. The EC$_x$/ER$_x$ is the concentration resulting in x% of inhibition compared to the organism growth in the control.

1 Dose-response analysis of growth inhibition toxicity tests

1.1 Growth inhibition toxicity tests

In a growth inhibition toxicity test, organisms are exposed to a series of concentrations / rates of a contaminant over a period of time and the growth data (such as length, weight of organisms, …) is collected at given time points during exposure. In the end, a growth data set is collected. For a chosen time point, observations can be described as $\{X_i, Y_i\}$, where X_i are the tested concentrations / rates, and Y_i the growth measurements.

1.2 Dose-response modelling

The dose-response model is defined as follows. Assuming that Y_i is normally distributed with mean μ and standard deviation σ, and that the mean μ is defined as a function f of the contaminant concentration / rate, we obtain:

$$Y_i \sim \mathcal{N}(f(X_i), \sigma^2)$$

There may be various possibilities for f. In the MOSAIC$_{\text{growth}}$ application, we assume a three-parameter log-logistic function for f:

$$f(x) = \frac{d}{1 + \left(\frac{x}{e}\right)^b}$$

where b, d and e are positive parameters. Parameter b is the shape parameter (the "slope" of the dose-response curve, corresponding to the effect intensity of the contaminant), d corresponds to growth in control data (i.e., in absence of contaminant) and e corresponds to the EC_{50}/ER_{50}.

1.3 Bayesian inference

The Bayesian approach considers that data are fixed and that the parameters are unknown random variables following a probabilistic distribution. These results in the following practical implications: *(i)* the Bayesian process optimises the probability of parameter vector θ given the data set Y used for calibration (the so-called posterior distribution) not only the likelihood (see below); *(ii)* there is a need to provide reasonable prior information, then updating this information by accounting for the data. Below is a short introduction to Bayesian principles[1].

In short, the Bayesian approach requires the following steps:
- Choose the prior distributions on parameters based on previous results, literature or expert knowledge (without looking at the data to fit): $\text{P}(\theta)$;
- Define the probabilistic model from the data, that is the random variable whose data would be one realisation assuming known values of parameters, namely the likelihood: $\text{P}(\text{Y} \mid \theta)$;
- Calculate the joint posterior distribution of the parameters given the data via the Bayes formula: $\text{P}(\theta \mid \text{Y})$;
- Provide statistical summaries of parameter estimates (namely, appropriate quantiles);
- Get any function of the parameter estimates as posterior probability distribution, like for example EC_x/ER_x calculations or predictions of new observations.

Basic principles
The keystone of the Bayesian approach is the Bayes formula:

$$\text{P}(\theta \mid \text{Y}) = \frac{\text{P}(\theta)\text{P}(\text{Y} \mid \theta)}{\text{P}(\text{Y})}$$

where Y are the observed data; $\mathrm{P}(\theta \mid \mathrm{Y})$ is the joint posterior distribution of parameter vector θ; $\mathrm{P}(\mathrm{Y} \mid \theta)$ is the likelihood of the data given the parameters; $\mathrm{P}(\theta)$ is the joint prior distribution of parameter vector θ. Given that $\mathrm{P}(\mathrm{Y})$ is known and fixed, it is often not considered as it does not depend on θ and will not influence the posterior distribution. Hence:

$$\mathrm{P}(\theta \mid \mathrm{Y}) \sim \mathrm{P}(\theta)\mathrm{P}(\mathrm{Y} \mid \theta)$$

with $\mathrm{P}(\theta)\mathrm{P}(Y|\theta)$ the unormalised posterior density and:

$$\mathrm{P}(\mathrm{Y}) = \int \mathrm{P}(\theta)\mathrm{P}(\mathrm{Y} \mid \theta)\, d\theta$$

The prior distribution $\mathrm{P}(\theta)$ expresses the available parameter information without knowing the observed data, while the posterior distribution $\mathrm{P}(\theta \mid \mathrm{Y})$ combines this prior information (which may be more or less informative depending on what is known about the value of the parameters beforehand) with evidence from the data (expressed through the likelihood) into a posterior density probability distribution for the parameters. The overall expectation is to get a narrower posterior distribution compared to the prior one: the difference between the two distributions reflects the information provided by the data. When the non-informative prior is vague (translated into a flat uniform distribution), and the data sufficiently informative, the results are similar to those obtained by the frequentist approach.

Joint posterior distribution

The joint posterior distribution has the dimension of the number of parameters times the number of iterations within the MCMC chains, and it can be plotted in planes of parameter pairs to visualise correlations between parameters. In an example case with two binormally distributed parameters, the joint posterior distribution can be plotted in the 2D-parameter space as illustrated by ellipses on **Fig. 1**; in this example, parameters θ_1 and θ_2 appear slightly correlated. From the joint posterior distribution, the marginal posterior distributions for each parameter (as illustrated by grey normal distributions on bottom and left sides of **Fig. 1**) can be extracted. Then, from the marginal posterior distributions, some statistical summaries on parameter estimates can be extracted, usually the median (illustrated by vertical and horizontal plain grey lines on **Fig. 1**) as well as 2.5% and 97.5% quantiles to serve as 95% credible intervals (illustrated by vertical and horizontal dotted grey lines on **Fig. 1**). Another advantage of having the joint posterior distribution is that any posterior distribution of any function of the parameters can be obtained.

Parameter uncertainties

One implication of adopting a Bayesian approach is that the uncertainty on a parameter is expressed as a probability distribution from which a credible interval (also called a Bayesian confidence interval) can be extracted. For example, the 95% credible interval delimits a range of values where the parameters should lie with a 95% probability, whereas the calculation of a confidence interval used in a frequentist approach (usually a 95% confidence interval) is based on hypothetical repeated sampling from the model: if samples of the same population size are repeatedly obtained and a 95% confidence interval for each of the samples is calculated, it is expected that 95% of the confidence intervals will contain the true value of the parameter. Another key point is that the credible interval is conditional to the data used to estimate the parameters.

Numerical computation

Many numerical methods have been developed to approximately compute the posterior distribution in challenging cases, mainly based on simulations by Monte Carlo Markov Chain (MCMC) sampling methods used to generate random numbers from complex joint distributions. MCMC algorithms are a general method based on drawing values of parameter vector θ from approximate distributions and then correcting those draws to better approximate the target posterior distribution $\mathrm{P}(\theta \mid Y)$. The sampling is done sequentially, with the distribution of the sampled draws depending only on the last value drawn; hence, the draws form a Markov Chain. The key to the method success, however, is not the Markov property but rather the fact that

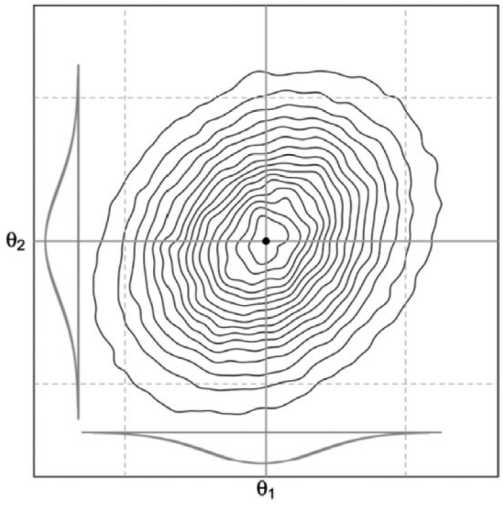

Figure 1. *Theoretical binormal joint posterior distribution of parameter vector (θ_1, θ_2). Ellipses correspond to isoclines of the joint posterior distribution; grey distributions are marginal posterior distributions of both parameters; solid horizontal and vertical lines correspond to the medians of these marginal distributions; dashed horizontal and vertical lines correspond to the 2.5% and 97.5% quantiles of the marginal distributions.*

the approximate distributions are improved at each step of the simulation, in the sense it finally converges to the target posterior distribution after an enough number of iterations. Indeed, with such algorithms, the simulation process must run long enough so that the distribution of the current draws is close enough to the desired target posterior distribution.
MCMC algorithms use random walk algorithms. Among them, the Metropolis algorithm (and its generalisation, the Metropolis–Hasting algorithm) is an adaptation of a random walk with an acceptance/rejection rule to converge to the specified target distribution[2,3]. The Gibbs sampler is a special case of the Metropolis–Hastings algorithm applicable when the joint distribution is not known explicitly, or where it is difficult to directly sample from, while the conditional distribution of each parameter is known and it is easy (or at least, easier) to sample from[4].
Several tools are available to automatically perform these computations. In MOSAIC$_{\text{growth}}$, JAGS[5] (version 4.3.0. (2017-08-10)) and R software[6] (version 4.0.2 (2020-06-22)) are used. The models are fitted to growth data using Bayesian inference via Monte Carlo Markov Chain (MCMC) sampling based on a Gibbs-type algorithm. For each model, we start by running a short sampling with three chains (5,000 iterations after a burn-in phase of 10,000 iterations) using the Raftery and Lewis[7] method to set the necessary thinning and number of iterations to reach an accurate estimation of the joint posterior distribution.

1.4 Choice of prior distributions

In MOSAIC$_{\text{growth}}$, prior choice is hidden to the user. However, here we give some information to help the user to understand the model behind. Before conducting an experimental study, prior distributions are defined for each parameter according to information available from the literature and/or previous experiments. Depending on the sources where the information come from, informative, semi-informative or non-informative prior distribution can be used. If a parameter was already estimated in previous studies or if previous data are available, a prior distribution can easily be characterized with an appropriate probability distribution. However, if no information is available but an order of magnitude is (positive only, for example), it is possible to use a weakly informative prior. If any information is available on the order of magnitude of a parameter, its prior can be defined on a decimal logarithm scale in order to consider with equal probability both low or high expected values.

As MOSAIC$_{\text{growth}}$ application has to be the most generic as possible, priors were assumed to be

as follows:

- Quasi-non-informative for parameter b:

$$log_{10}b \sim \mathcal{U}(-2, 2)$$

- Uniform on parameter d with the following bounds:

$$d \sim \mathcal{U}(0, 2 \times maxY)$$

where $maxY$ equals the highest observed Y for the species under consideration so that this observation is excluded from the data for the analysis.

- For parameter e, we assume that the range of tested concentrations / rates in experiment is chosen to contain the ER_{50} with a high probability. Hence we use a normal prior distribution for parameter e as follows:

$$log_{10}e \sim \mathcal{N}(\frac{log_{10}(maxX) + log_{10}(minX)}{2}, \frac{log_{10}(maxX) - log_{10}(minX)}{4})$$

where $minX$ and $maxX$ are the smallest and the highest tested concentrations / rates, respectively.

- The prior distribution on parameter σ is chosen as follows:

$$\sigma \sim \mathcal{U}(0, maxY)$$

where $maxY$ is the same as defined above.

2 Options of censoring on EC_x/ER_x to account for the uncertainty

Under the Bayesian framework, we get a posterior probability distribution of the EC_x/ER_x (see example in **Fig. 2. A**) quantifying the uncertainty of the EC_x/ER_x that can be summarized as a median and a 95% credible interval (CI95), representing the range of values within which the EC_x/ER_x has 95% of chance to lie.

This uncertainty on the EC_x/ER_x can then be used to adequately censor these values if required for future SSD analyses. A key question to consider is the CI95 of the EC_x/ER_x estimate always precise enough to be used as it is or regarding the range of tested concentrations / rates, should we have to right-censor it? In practice, we assume that the true EC_x/ER_x value has 95% of chance to be greater than the lowest bound of the CI95 (LCI95), depending on the relative position of max_rate compared to the LCI95. Hence, we propose a criterion based on the following ratio of probabilities:

$$ratio = \frac{P(LCI95 \leq ER_x \leq max\_rate)}{P(LCI95 \leq ER_x \leq UCI95)}$$

This is the ratio of the probability that the EC_x/ER_x lies within LCI95 and max_rate over the probability that the EC_x/ER_x lies within its CI95 (this latter equals 95%); based on **Fig. 2. B**, this criterion is calculated as the ratio of the orange surface divided by the (orange + grey) surface.

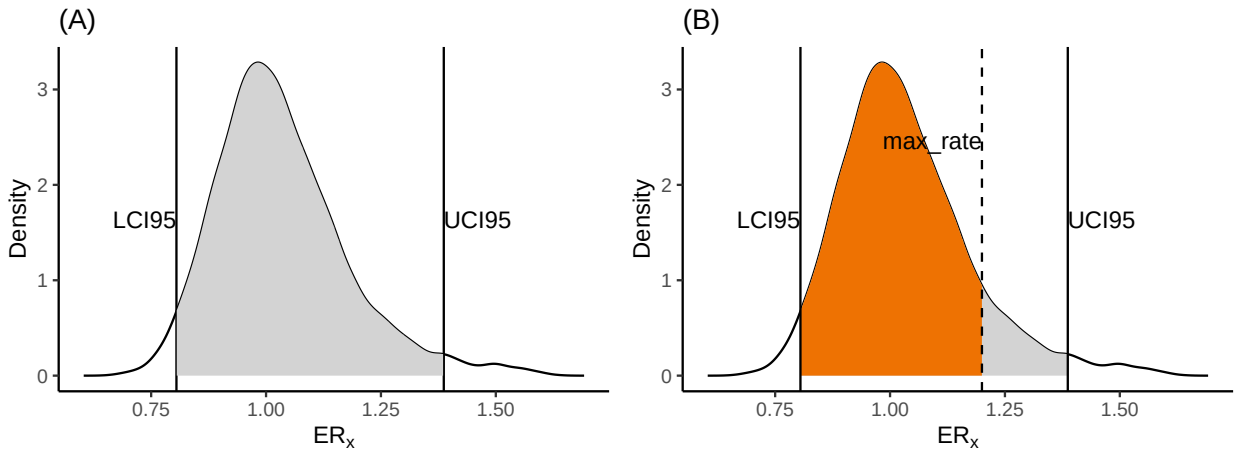

Figure 2. *Options of censoring on ER_x to account for the uncertainty: (A) posterior distribution of ER_x, (B) ratio of probabilities.*

If the calculated ratio is greater than a given threshold T chosen to be equal to 0.5, we keep an interval-censored EC_x/ER_x corresponding to the whole CI95; otherwise, we consider a right-censored EC_x/ER_x with a lower bound being the minimum between the lower bound of the CI95 and the maximum tested rate (max_rate):

$$\text{censored } EC_x/ER_x = \begin{cases} [LCI95, UCI95] & \text{ratio} > T \\ [min(LCI95, max\_rate), +\infty) & \text{ratio} \leq T \end{cases}$$

References

(1) Ockleford, C., Adriaanse, P., Berny, P., Brock, T., Duquesne, S., Grilli, S., Hernandez-Jerez, A. F., Bennekou, S. H., Klein, M., Kuhl, T., Laskowski, R., Machera, K., Pelkonen, O., Pieper, S., Smith, R. H., Stemmer, M., Sundh, I., Tiktak, A., Topping, C. J., Wolterink, G., Cedergreen, N., Charles, S., Focks, A., Reed, M., Arena, M., Ippolito, A., Byers, H., and Teodorovic, I. (2018) Scientific Opinion on the state of the art of Toxicokinetic/Toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms.

(2) Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953) Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21, 1087–1092.

(3) Hastings, W. K. (1970) Monte carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

(4) Geman, S., and Geman, D. (1984) Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741.

(5) Plummer, M. (2003) JAGS: A program for analysis of Bayesian models using Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing; Vienna, Austria.

(6) R Core Team. (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

(7) Raftery, A. E., and Lewis, S. M. (1992) [Practical Markov chain Monte Carlo]: Comment: one long run with diagnostics: implementation strategies for Markov chain Monte Carlo. Statistical Science 7, 493–497.