Calibrating Complex Stochastic Models using Emulation and History Matching

IDDConf 2023

Andrew Iskauskas

Durham University
In collaboration with LSHTM

May 23, 2023

Complex Models of Real-World Phenomena

Complex computer models (or *simulators*) are used in a variety of fields, including

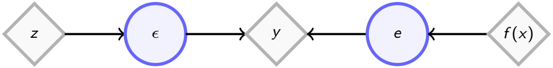
- Oil Industry (oil reservoir and geology models) [4]
- Climate Science (climate models of global warming) [11]
- Systems Biology (genetic and metabolic network models) [10]
- Cosmology (galaxy formation simulations) [9]
- Nuclear Physics (quantum many-body models of nuclei) [5]
- **Epidemiology** HIV, TB, Covid, . . . [1, 7]

Simulators are often computationally expensive: a full exploration of the parameter space using only the simulator is infeasible.

Uncertainty Structure for Models

Consider a simulator $f(x)$ that represents a physical process y, from which we may obtain observed quantities z. Two main sources of uncertainty are

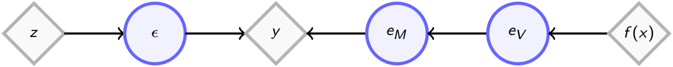
- **Observational error**. Our observations z of y are made imperfectly: $z = y + \epsilon$;
- **Model discrepancy**. Our simulator $f(x)$ cannot faithfully represent the process y: $y = f(x) + e$.



Uncertainty Structure for Models

Consider a **stochastic** simulator $f(x)$ that represents a physical process y, from which we may obtain observed quantities z. Two main sources of uncertainty are

- **Observational error**. Our observations z of y are made imperfectly: $z = y + \epsilon$;
- **Model discrepancy**. Our simulator $f(x)$ cannot faithfully represent the process y: $y = f(x) + e$. *Moreover, repeated evaluations of $f(x)$ at the same point x give different values.*



The Emulator

An *emulator* is a statistical approximation of a complex computer simulator [3].

Let $f(x)$ be an output from the simulator at a given parameter set $x \in \mathbb{R}^d$, corresponding to some real physical process y. Then we define a emulator for output $f(x)$ as

$$g(x) = \sum_i \beta_i h_i(x_A) + u(x_A) + w(x)$$

The $h_i(x_A)$ are a collection of basis functions in the *active variables* x_A, β_i the coefficients, $u(x_A)$ a weakly stationary process in the active variables, and $w(x)$ a 'nugget term'.

Pragmatic choice: consider Bayes Linear emulators, so only need prior beliefs for expectations, variances, and covariances.

Stochastic Emulation

The quantity $\text{Var}[g(x)]$ encodes the uncertainty of the emulator prediction. For stochastic models, we apply a *hierarchical* approach to accurately account for model variability.

- Train emulator $g_V(x) = \sum_i \beta_{Vi} h_{Vi}(x_{VA}) + u_V(x_{VA}) + w_V(x)$ to the *stochasticity* of the model output;
- Use $\text{E}[g_V(x)]$ as an informed prior for $\text{Var}[g(x)]$, and create output emulators $g(x)$.

Can extend this framework further – emulating covariances between outputs, not just variance.

Emulation

Emulators are **fast to evaluate**, requiring only matrix multiplication. For complex models which can take anywhere from minutes to months to evaluate a limited ensemble of runs, an emulator can quickly investigate model behaviour across the entire parameter space.

Emulators **have uncertainty statements built-in**. Each prediction comes with a corresponding uncertainty, $\text{Var}[g(x)]$, which depends on the data provided to it and the proximity of the unseen points thereof.

History Matching

Given observed data corresponding to a simulator output, what combinations of input parameters could give rise to output consistent with this observation?

History matching works on the principle of complementarity: a point x is considered unsuitable if **even accounting for the uncertainties in the system**, the prediction $\mathrm{E}_D[g(x)]$ cannot be 'close' to the observed value z. Closeness is defined via an *implausibility* measure

$$I^2(x) = (\mathrm{E}[g(x)] - z)^T \left(\mathrm{Var}[g(x)] + \mathrm{Var}[e] + \mathrm{Var}[\epsilon]\right)^{-1} (\mathrm{E}[g(x)] - z)$$

Emulation and HM: Summary

Emulators can **efficiently** and **robustly** predict simulator output at unseen points, given a small collection of known runs.

The corresponding emulator uncertainty is a **natural extension** to existing sources of uncertainty in our model, and can easily account for stochasticity in the models.

History matching allows us to leverage the uncertainty structure to find **all** acceptable matches to data arising from our model.

The hmer Package

The hmer package [6] was developed to make the tools of emulation and history matching accessible for modellers. It allows

- Careful prior specifications to be determined and emulators to be trained
- Diagnostics to be performed to assess suitability
- Appropriate choices of implausibility measure and design for further waves to be made.
- Try it: `cran.r-project.org/web/packages/hmer/`

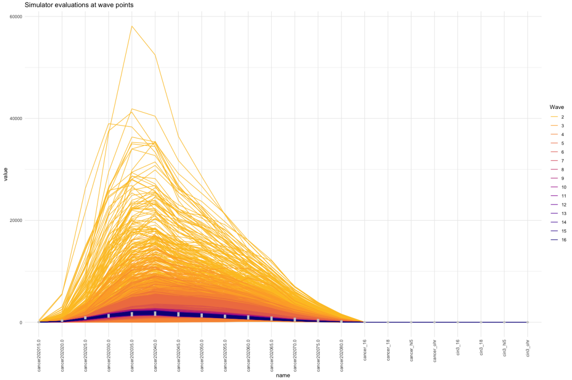
The HPVsim model

- Agent-based model of HPV transmission and natural history developed by IDM [8]
- Can simulate multiple different HPV genotypes, sexual behaviours, demographics, . . .
- Input space anywhere from 6 to 60+ parameters, any relevant outputs available via analyzers
- Run time for a single parameter set between 2 minutes and 1 hour.
- Planned use in evaluating screening and vaccination strategies worldwide.

HPVSim Emulation

Considered four “classes” of genotype (29 dimensional parameter space); matching to data collected from Nigeria comprising 22 outputs [2].

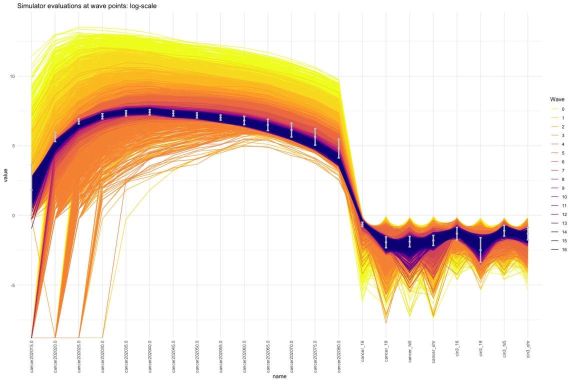
16 waves of emulation; each wave used 16 repetitions at each of 290 parameter sets.



HPVSim Emulation

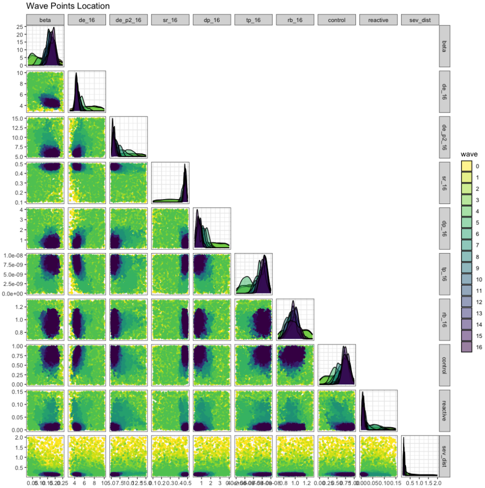
Considered four "classes" of genotype (29 dimensional parameter space); matching to data collected from Nigeria comprising 22 outputs [2].

16 waves of emulation; each wave used 16 repetitions at each of 290 parameter sets.



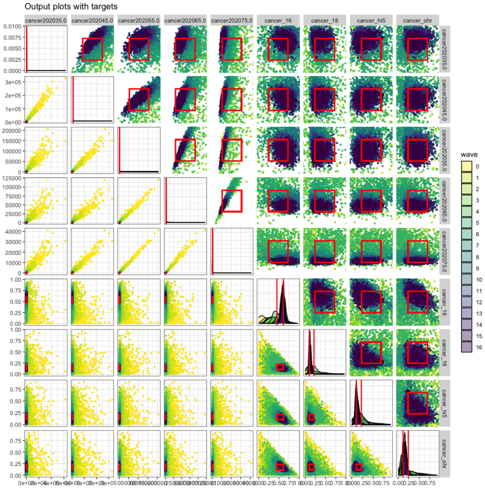
HPVSim Input Space

History matching gives the full collection of acceptable input parameter sets, allowing for all uncertainties and discrepancies.



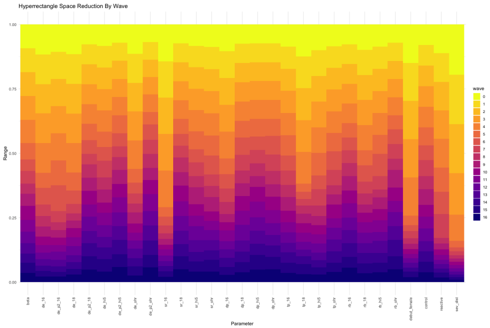
HPVSim Output Space

We may also consider the dependencies between outputs.



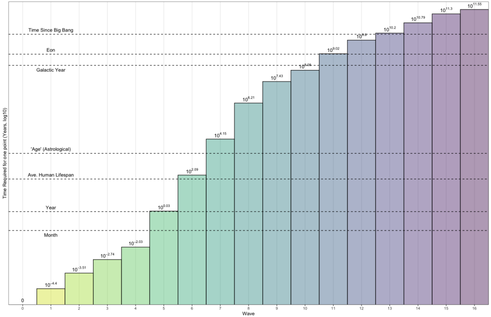
HPVSim Output Space

Final parameter space is $\sim 7 \times 10^{-18}$ of the original volume; final wave generated 55 points matching all targets according to the simulator.



HPVSim Output Space

Were we to want to characterise the parameter space equivalently by naive methods: **2×10^{13} years** of simulator run-time.



Summary

- Complex stochastic simulators can be slow to run and difficult to meaningfully analyse.
- Emulators can make robust and fast predictions across a large-dimensional parameter space.
- HM uses the induced uncertainty structure to determine the *full* set of parameter combinations that could give rise to the observed data.
- Flexible, low specificational burden, extensible via hierarchical emulation, multistate emulation, known boundary emulation, . . .
- The `hmer` package is designed to make these tools more readily available for modellers.

References I

[1] Ioannis Andrianakis et al. "Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda". In: *PLoS computational biology* 11.1 (2015), e1003968.

[2] HPV Information Centre. *Human Papillomavirus and Related Diseases Report: Nigeria*. 2021.

[3] Peter S Craig et al. "Constructing Partial Prior Specifications for Models of Complex Physical Systems". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1 (1998), pp. 37–53.

[4] Jonathan A Cumming and Michael Goldstein. "Bayes Linear Uncertainty Analysis for Oil Reservoirs Based on Multiscale Computer Experiments". In: *The Oxford handbook of applied Bayesian analysis* (2010), pp. 241–270.

[5] Baishan Hu et al. "Ab Initio Predictions Link the Neutron Skin and 208Pb to Nuclear Forces". In: *Nature Physics* 18.10 (2022), pp. 1196–1200.

[6] Andrew Iskauskas et al. "Emulation and History Matching using the hmer Package". In: *arXiv preprint arXiv:2209.05265* (2022).

[7] Danny Scarponi et al. "Demonstrating Multi-country Calibration of a Tuberculosis Model Using New History Matching and Emulation Package - hmer". In: *Epidemics* 43 (2023), p. 100678.

References II

[8] Robyn M Stuart et al. "HPVsim: An Agent-Based Model of HPV Transmission and Cervical Disease". In: *medRxiv* (2023).

[9] Ian Vernon, Michael Goldstein, and Richard Bower. "Galaxy Formation: A Bayesian Uncertainty Analysis". In: *Bayesian analysis* 5.4 (2010), pp. 619–669.

[10] Ian Vernon et al. "Bayesian Uncertainty Analysis for Complex Systems Biology Models: Emulation, Global Parameter Searches and Evaluation of Gene Functions". In: *BMC systems biology* 12.1 (2018), pp. 1–29.

[11] Daniel Williamson et al. "History Matching for Exploring and Reducing Climate Model Parameter Space using Observations and a Large Perturbed Physics Ensemble". In: *Climate dynamics* 41.7 (2013), pp. 1703–1729.