New Insights into History Matching via Sequential Monte Carlo

Dr Christopher Drovandi

School of Mathematical Sciences
ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)
Queensland University of Technology

Collaborators: David Nott (National University of Singapore) and Dan Pagendam (CSIRO Data 61)
Acknowledgments: Leah South (QUT), ACEMS and ARC funding

Parameter Estimation

Denote θ as the parameter of a deterministic model $M(\theta)$ or stochastic model $M(\theta, u)$ where u are the random numbers needed in the simulation.

Have some observed data y believed to be generated by M.

Wish to obtain a collection of plausible θ based on y.

However, the model M is too expensive to use say Markov chain Monte Carlo or sequential Monte Carlo.

History Matching

Aim: Determine ‘plausible’ region of the parameter space relatively quickly.

History matching (eg Craig et al 1997) determines a non-implausible region of the parameter space relatively quickly by using an *emulator* of the model outputs or ‘distance’ between model outputs and y.

The emulator not only provides predictions at untrained points but also quantifies the uncertainty in the predictions.

History Matching Notation/Details

Denote $\pi(\theta)$ as a 'prior' distribution.

For simplicity we will assume there is only $n = 1$ model output, eg:

- The model really does have only 1 output
- The output is a distance between outputs and data.
- The output is a (approximate) likelihood function.

Parameter θ is deemed as non-implausible if $\mathcal{I}(\theta) < c$ for cut-off c where $\mathcal{I}(\theta)$ includes emulation uncertainty, eg Andrianakis et al 2015

$$\mathcal{I}(\theta) = \frac{|y_\theta - y_{\mathrm{obs}}|}{\sqrt{s_{m,\theta}^2 + s_{e,\theta}^2 + s_d^2}}. \qquad (1)$$

Uncertainty: $s_{m,\theta}^2$ simulator, $s_{e,\theta}$ emulator, s_d^2 model.

History Matching Procedure

Steps involved in the history matching algorithm:

1. Generate N_w training samples $\{\theta_j\}_{j=1}^{N_w} \sim \pi(\theta)$ using a space filling design and simulate the model at each θ_j to generate the collection of outputs $\{y_{\theta_j}\}_{j=1}^{N_w}$.
2. Fit an emulator E_w to the training data $\{\theta_j, y_{\theta_j}\}_{j=1}^{N_w}$.
3. Use the emulator E_w to define an implausibility function $\mathcal{I}_w(\theta)$. If $\mathcal{I}_w(\theta) > c_w$ for some chosen c_w then θ is deemed as implausible by emulator E_w.
4. Use all emulators $\{E_r\}_{r=1}^{w}$ to define the non-implausible region $\Theta_w = \{\theta \in \Theta | \cap_{r=1}^{w} \mathcal{I}_w(\theta) < c_w\}$.
5. Increase wave counter $w = w + 1$.
   Generate N_w training samples $\{\theta_j\}_{j=1}^{N_w}$ from Θ_w and simulate the model at each θ_j to generate $\{y_{\theta_j}\}_{j=1}^{N_w}$.
6. If the stopping rule is satisfied then finish otherwise return to Line 2.

Issues with History Matching

History matching does have at least two issues:

1. The cut-off values c_w may not be easy to select in practice and there is no existing automated method for doing so.
2. Sampling uniformly from Θ_w as w increases becomes increasingly difficult.

Solution: We use sequential Monte Carlo (SMC, eg Chopin 2002) to help address the above 2 issues.

Sequential Monte Carlo

SMC samples from a sequence of distributions (connecting easy and target distributions) by iteratively applying re-weighting, re-sampling and move steps.

Advantages of SMC approach over MCMC:

- Naturally adaptive
- Easily parallelisable
- More capable of dealing with multimodal or complex posterior distributions

SMC History Matching

For history matching, we define the sequence of distributions as

$$p_w(\theta) \propto \pi(\theta) \prod_{k=1}^{w} \mathbb{I}(\mathcal{I}_k(\theta) \leq c_k).$$

Assume that we have a collection of 'particles', $\{W_w^i, \theta_w^i\}_{i=1}^M$ from $p_w(\theta)$. To push the particles to target $w+1$ we apply re-weighting step:

$$W_{w+1}^i \propto W_w^i \mathbb{I}(\mathcal{I}_{w+1}(\theta_w^i) \leq c_{w+1}).$$

Here W_w are all equal, so the weights W_{w+1} will either be constant or equal to zero. Thus we can select c_{w+1} so that a certain proportion, α, have non-zero weight. Same as ensuring that the effective sample size (ESS) $= 1/\sum_{i=1}^M (W_{w+1}^i)^2$, is αM.

SMC History Matching

After re-weighting the ESS drops to roughly αM.

Resampling M times from the surviving particles allows the ESS to return to M. However, the drawback is that some particles will be duplicated.

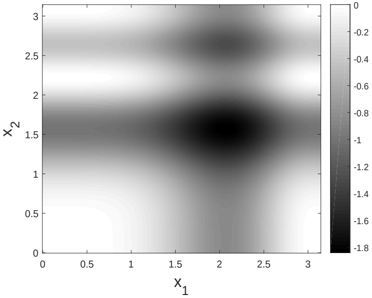
Diversify particles with an MCMC kernel R times to each of the resampled particles. Can adaptively choose R. Can use population of particles to inform efficient MCMC proposal. This move step only uses the emulator (no expensive model simulations).

Simple Example

Consider the function

$$y = -\sin(x_1)\sin(x_1^2/\pi)^2 - \sin(x_2)\sin(2x_2^2/\pi)^2,$$

where $\theta = (x_1, x_2) \in (0, \pi) \times (0, \pi)$.
We wish to find the regions where y is small.



History Matching Parameters

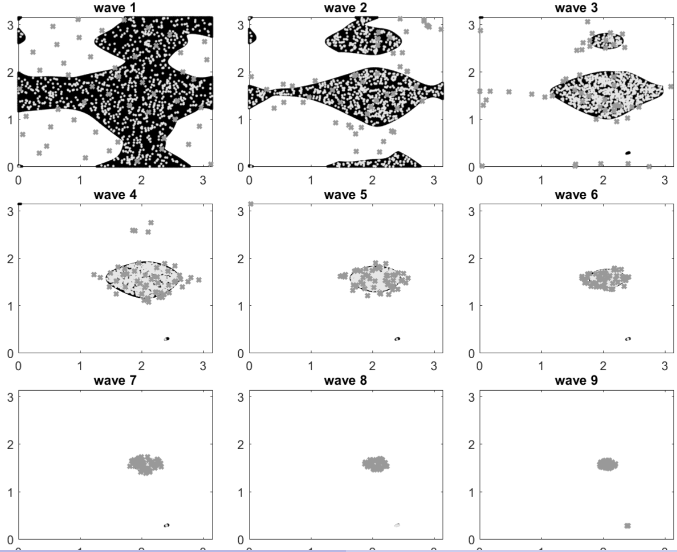
We use $N = 50$ training samples at each wave.

Implausability measure $\mathcal{I}(\theta) = y_p(\theta) - r \times s_p(\theta)$ where $y_p(\theta)$ and $s_p(\theta)$ is the prediction and the standard deviation from the currently fitted GP, respectively. Set $r = 3$.

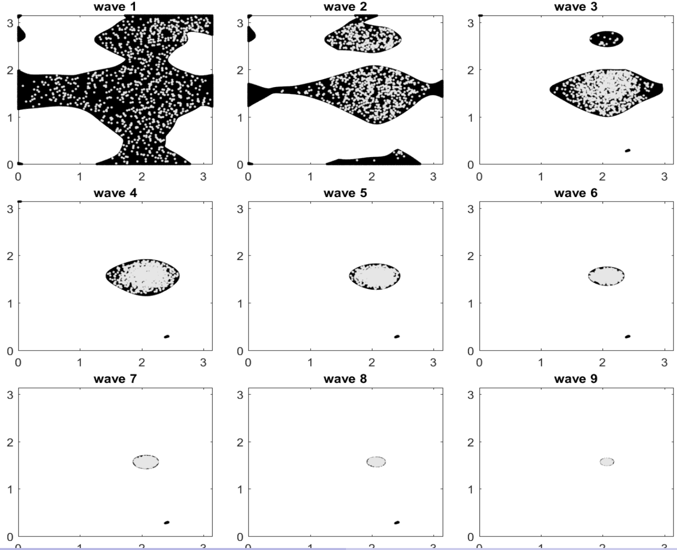
We solve the history matching problem 'perfectly' by taking 2^{20} draws from parameter space and use rejection sampling. The cut-offs c_w are chosen so that half the remaining particles survive at each wave.

We investigate how well SMC can uniformly sample from non-implausible space implied by 'perfect' history matching.

Results from SMC History Matching



Results from Typical History Matching Approach



Conclusions

Our algorithm reveals the significant complexity of the non-implausible probability distributions arising from history matching.

The SMC algorithm provides a semi-automated and principled method for performing history matching.

More work needs to be done on designing efficient MCMC kernels when derivative information about the target distribution is not available.

References

Andrianakis et al (2015). Bayesian history matching of complex infectious disease models using emulation: a tutorial and a case study on hiv in uganda. PLOS Computational Biology, 11(1):e1003968.

Chopin (2002). A sequential particle filter method for static models. Biometrika, 89(3):539-552.

Craig et al (1997). Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments. In Case Studies in Bayesian Statistics, pages 37-93. Springer.

Drovandi et al (2017). New insights into history matching via sequential Monte Carlo. https://eprints.qut.edu.au/112175/.

Email: c.drovandi@qut.edu.au

Twitter: @chris_drovandi