Unlocking data sets by calibrating populations of models to data density: A study in atrial electrophysiology

Dr Christopher Drovandi

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

Collaborators: Brodie Lawson, Kevin Burrage, Pamela Burrage, Nicole Cusimano, Blanca Rodriguez
Motivating Example: Atrial Electrophysiology

- Interested in variability in heart function: structure, electrophysiology, response to stress, response to medication etc.
- Sophisticated (deterministic) heart models now exist, denote as $M(\theta)$ with parameter $\theta$.
- Variability can be captured with parameter distribution $g(\theta)$ (referred to as ‘population of models’ in applied literature).
- Wish to calibrate $g(\theta)$ to real data.
Data

Single cardiac cell, that generates an action potential

Figure: Action potential and biomarkers.

Biomarker data available for 469 cells from 363 patients. Two groups of patients - normal and arrhythmia (unhealthy)
The Mathematical Problem

- Have a deterministic model $M(\theta)$ where $\theta \in \mathbb{R}^p$.
- Have data $y = (y_1, y_2, \ldots, y_n)$ where $y_i \in \mathbb{R}^d$. Denote distribution of $y$ as $f(y)$ (obtained from density estimation).
- Denote $g(\theta)$ as a distribution on $\theta$.
- Denote $h(y|g(\theta))$ as the distribution on $y$ when pushing $g(\theta)$ through $M(\theta)$.
- Wish to find $g^*(\theta)$ such that

$$g^*(\theta) = \arg \min_{g(\theta)} \int_y \ln \left( \frac{h(y|g(\theta))}{p(y)} \right) h(y|g(\theta)) \, dy.$$  

Too computationally intensive too solve when $M$ is expensive to simulate.
Previous approaches ignore data distribution and match only to range of biomarkers.

**X** – data  ● – selected by range calibration  ● – what we’d like to select
Outline of Approach

To reduce computation we proposed a pragmatic approach:

- Use Monte Carlo methods to generate samples from 
  \[ p(\theta) \propto f(y(\theta))\pi(\theta) \] 
  where \( \pi(\theta) \) defines initial range.
- The above step is wrong as it doesn’t account for the transformation \( y \rightarrow \theta \).
- Apply a refinement step so that we better match \( f(y) \).
For the Monte Carlo method we use sequential Monte Carlo (SMC):

- SMC moves a population of $N$ particles through a sequence of distributions (starting with one easy to sample from and finishing at the target).
- Here we define the sequence of targets as: $p_t(\theta) \propto \pi(\theta) f(y)^{\gamma_t}$. Here $0 = \gamma_0 < \gamma_1 < \cdots < \gamma_T = 1$
- Output of SMC is a set of weighted samples (particles) $\{\theta_t^i, W_t^i\}_{i=1}^N$ from each $p_t(\theta)$.
- Can measure the ‘quality’ of sample through effective sample size, $\text{ESS}_t = 1/\sum_i (W_t^i)^2$. 
1. Sample $N$ particles from ‘prior’ $\pi(\theta)$ and set $\gamma_0 = 0$, $t = 1$.
2. Determine $\gamma_t$ so that $\text{ESS}_t \approx N/2$.
3. Re-weight: $W^i_t \propto f(y(\theta^i_t))^{\gamma_t - \gamma_{t-1}}$
4. Resample particle set according to their weights.
5. Apply (easy to adapt) MCMC kernel with invariant distribution $p_t(\theta)$ to each particle.
6. Set $t = t + 1$. Repeat from Step 2 until $\gamma_t = 1$. 
Advantages of SMC

- Embarrassingly Parallel
- Easy to adapt
- Effective in sampling complex parameter distributions
As mentioned before, SMC does not have the correct target. 
SMC produces initial population \( \{ \theta_i \}_{i=1}^N \).
Use simulated annealing to drop/add parameter values from population to better match data distribution.
To decide whether to drop/add parameter values we use the Jensen-Shannon distance (\( JSD \))

\[
\frac{1}{2} \int_y p(y) \ln \left( \frac{p(y)}{\frac{1}{2} p(y) + \frac{1}{2} q(y)} \right) dy + \frac{1}{2} \int_y q(y) \ln \left( \frac{y}{\frac{1}{2} p(y) + \frac{1}{2} q(y)} \right) dy \right]^{1/2}.
\]

Compute all marginal \( JSD_i \)’s and bivariate \( JSD_{i,j} \)’s and put in matrix \( P \).
Criterion for dropping/adding is based on \( \rho = \| P \|_2 \).
Results: After SMC

Figure: Results after SMC. Data = black, ranges = red, SMC = blue. Source: http://advances.sciencemag.org/content/4/1/e1701676
Results: After Refinement

Figure: Results after SMC. Data = black, SMC = red, refine = blue. Source: http://advances.sciencemag.org/content/4/1/e1701676
Results: Predict Effect of Drug

Figure: Before drug = red, After drug = gold. Source: http://advances.sciencemag.org/content/4/1/e1701676
Developed method a bit improvised but computationally feasible and effective for illustrating the benefit of matching to data distribution.
References

- Lawson et al (2018). Science Advances. 4: e1701676.
- Drovandi et al (2016). Journal of the Royal Society Interface. 13, 20160214.
- Britton et al (2017). Front. Physiol. 8, 278.
- Sanchez et al (2012). Am. J. Physiol. Heart Circ. 302, H1146–H1159.

Email: c.drovandi@qut.edu.au
Web: chrisdrovandi@weebly.com
Twitter: @chris_drovandi