Sequential Monte Carlo Algorithms for Bayesian Sequential Design

Dr Chris Drovandi
Queensland University of Technology
c.drovandi@qut.edu.au

Collaborators: James McGree, Tony Pettitt, Gentry White
Acknowledgements: Australian Research Council Discovery Grant and organisers of MCMski

January 6, 2014
Adaptive decisions as new data are collected

More robust to parameter and model uncertainty

Natural to use Bayesian framework. Posterior becomes new prior

Next decision obtained by looking forward to all future decisions (backward induction)

Simplified by myopic design (one-at-a-time)

Next design point $d_{t+1} = \arg \max U(d|y_{1:t}, d_{1:t})$. $y_{1:t}$ collected data at design $d_{1:t}$. $U$ is utility function
Why SMC for Bayesian sequential design?

- Much more efficient than MCMC. Simple re-weighting step to incorporate new information.
- Parallel implementation possible (e.g. GPU)
- Increase in efficiency allows comparisons of utility functions
- Convenient estimation of important Bayesian utility functions (e.g. mutual information)
- Decisions in real time?
SMC For One static Model \( m \)

- Sample from sequence of targets
- Data annealing here

\[
\pi_t(\theta_m | y_{1:t}, d_{1:t}) = f(y_{1:t} | \theta_m, d_{1:t}) \pi(\theta_m) / Z_{m,t}, \text{ for } t = 1, \ldots, T.
\]

\( y_{1:t} \) (independent) data up to \( t \), \( d_{1:t} \) design points up to \( t \), \( \theta_m \) parameter for model \( m \). \( f \) is likelihood, \( \pi \) prior, \( \pi_t \) posterior

\[
f(y_{1:t} | m, d_{1:t}) = Z_{m,t} = \int_{\theta_m} f(y_{1:t} | \theta_m, d_{1:t}) \pi(\theta_m) d\theta_m.
\]

- SMC: Generate a weighted sample (particles) for each target in the sequence via steps
  - Reweight: particles as data comes in (efficient)
  - Resample: when ESS small
  - Mutation: diversify duplicated particles (can be efficient)
SMC For One STATIC Model \( m \) (Algorithm) Chopin (2002)

- Have current particles \( \{W^i_t, \theta^i_t\}_{i=1}^N \) based on data \( y_{1:t} \)
- Re-weight step to included \( y_{t+1} \)

\[
W^i_{t+1} \propto W^i_t f(y_{t+1} | \theta^i_t, d_{t+1}).
\]

- Check effective sample size: \( \text{ESS} = 1/\sum_{i=1}^N (W^i_{t+1})^2 \)
- If \( \text{ESS} > E \) (e.g. \( E = N/2 \)) go back to re-weight step for next observation
- If \( \text{ESS} < E \) do the following
  - Resample proportional to weights. Duplicates good particles
  - Mutation: Move all particles via MCMC kernel say \( R \) times (adaptive proposal)
It can be shown

\[ Z_{t+1}/Z_t = f(y_{t+1}|y_{1:t}, d_{t+1}) = \int_{\theta} f(y_{t+1}|\theta, d_{t+1}) \pi(\theta|y_{1:t}, d_{1:t}) d\theta. \]

Using SMC particles to approximate posterior at \( t \) gives estimator

\[ Z_{t+1}/Z_t \approx \sum_{i=1}^{N} W_i f(y_{t+1}|\theta_t^i, d_{t+1}). \]

Can then obtain approximation of \( Z_{t+1} \) through

\[ \frac{Z_{t+1}}{Z_0} = \frac{Z_{t+1}}{Z_t} \frac{Z_t}{Z_{t-1}} \ldots \frac{Z_1}{Z_0}. \]

Also gives estimate of posterior predictive probability of \( y_{t+1} \).
Advantage 1: Efficiently comparing utilities (Drovandi et al 2013)

- Discrete data (binary) example
- Need to compute utility for all possible $d$ (then for all $t = 1, \ldots, T$)

$$U(d|y_{1:t}, d_{1:t}) = \sum_{z \in \{0,1\}} f(z|y_{1:t}, d_{1:t}, d)U(d, z|y_{1:t}, d_{1:t}).$$

- The whole process requires the computation (sampling) of many posterior distributions
- SMC (IS) to rescue. Pretend $z$ is the ‘next’ observation collected at design $d$. Simple re-weight to incorporate this observation.
- Use weighted sample to estimate $U(d, z)$. SMC also provides estimate of posterior predictive.
- There may be different choices for $U(d, z)$, want to compare.
- Need to simulate the design many times.
The Design Problem

Estimating Maximum tolerated dose, minimum effective dose (clinical trials)

\[ E[Y_t] = g^{-1}(\eta_t) \text{ where} \]

\[ \eta_t = \theta_0 + \theta_1 \frac{d^\lambda_t - 1}{\lambda}, \]

where \( d^\lambda_t \) is the dose assigned to the \( t \)th subject.

\( Y_t \sim Binary(E[Y_t]) \). Uninformative prior

\[ \pi(\theta_0, \theta_1, \lambda) \propto N(\theta_0; 0, 100)N(\theta_1; 0, 100)U(\lambda; 0, 1)1\left(\frac{\lambda \eta^* - \theta_0}{\theta_1} + 1 > 0\right), \]

Objective is precise estimation of

\[ d^* = \left(\frac{\lambda^T \eta^* - \theta_0^T}{\theta_1^T} + 1\right)^{1/\lambda^T}. \]

Let \( \theta = (\theta_0, \theta_1, \lambda) \).
The Utility Functions

Possible choices for $U(d, z|y_{1:t}, d_{1:t})$

1. Posterior precision of $d^*$ (Natural choice, but posterior of $d^*$ can be unstable when little information is available)
2. Kullback-Leibler Divergence between prior and posterior for $\theta$
3. Determinant of Posterior Covariance matrix of $\theta$
Some Results

Figure: Distributions of the estimated target stimulus over 10 to 100 subjects for the true parameter configuration of $\theta^T = (0,3,1)$ producing $d^* = 0.538$. Solid horizontal line is the true $d^*$. Shown are the 2.5%, 50% and 97.5% quantiles over the 500 runs for each utility function.
Advantage 2: Estimating Difficult Utilities

- E.g. Mutual Information for Model Discrimination (Drovandi et al. 2014)
- Have set of $K$ proposed models $M = 1, \ldots, K$. Select design to maximise ability to discriminate between models
- Consider mutual information between model indicator $M$ and predicted observation $Z$ for $y_{t+1}$ Box and Hill (1967).

\[
I(M; Z | y_{1:t}, d) = H(M | y_{1:t}) - H(M | Z; y_{1:t}, d).
\]

Therefore $U(d | y_{1:t}) = -H(M | Z; y_{1:t}, d)$ which is equal to

\[
U(d | y_{1:t}) = \sum_{m=1}^{K} \pi(m | y_{1:t}) \int f(z | m, y_{1:t}, d) \log \pi(m | y_{1:t}, z, d) d\mu(z).
\]
SMC for multiple models

- Effectively run an SMC algorithm for each model $m = 1, \ldots, K$
- Have set of $N$ particles for each model $\{W_{m,t}^i, \theta_{m,t}^i\}_{i=1}^N$.
- ESS for each model $m$
- resampling and within-model updates when required
- Design part: use data up to $t$, $y_{1:t}$, and particles of all models to compute the next design $d_{t+1}$
Estimating the Utility

\[ U(d | y_{1:t}) = \sum_{m=1}^{K} \pi(m | y_{1:t}) \int f(z | m, y_{1:t}, d) \log \pi(m | y_{1:t}, z, d) d\mu(z), \]

- Borth (1975) notes difficult computation
- SMC to the rescue.
- Potential observation \( z \) at potential design point \( d \).
- Pretend this the observation for \( y_{t+1} \).
Estimating the Utility (cont...)

- Estimate predictive probability using weights

\[ w_{m,t}^i(d, z) = W_{m,t}^i f(z | \theta_{m,t}^i, d), \]

\[ \hat{f}(z | m, y_{1:t}, d_{1:t}, d) = \sum_{i=1}^{N} w_{m,t}^i(d, z). \]

- \( Z_{m,t} \) denotes current evidence for model \( m \), which integrates out posterior of \( \theta \) at \( t \).

- Estimate evidence including \((d, z) Z_{m,t}(d, z)\) using

\[ \log \hat{Z}_{m,t}(d, z) = \log \hat{Z}_{m,t} + \log \hat{f}(z | m, y_{1:t}, d_{1:t}, d). \]

- Convert to \( \hat{\pi}(m | y_{1:t}, d, z) \)
Therefore estimate of utility for discrete $z$ is

$$
\hat{U}(d|y_{1:t}) = \sum_{m=1}^{K} \hat{\pi}(m|y_{1:t}) \sum_{z \in S} \hat{f}(z|m, y_{1:t}, d) \log \hat{\pi}(m|y_{1:t}, z, d).
$$

In continuous case approximate integral via MC integration. Draw $z_{m,t}^{i} \sim f(z|m, \theta_{m,t}^{i}, d)$ for $i = 1, \ldots, N$. Weighted sample $\{W_{m,t}^{i}, \theta_{m,t}^{i}, z_{m,t}^{i}\}$ from joint $p(z, \theta|d, m, y_{1:t}, d_{1:t})$. Therefore estimate of utility for continuous $z$ is

$$
\hat{U}(d|y_{1:t}, d_{1:t}) = \sum_{m=1}^{K} \hat{\pi}(m|y_{1:t}, d_{1:t}) \sum_{i=1}^{N} W_{m,t}^{i} \log \hat{\pi}(m|y_{1:t}, d_{1:t}, z_{m,t}^{i}, d).
$$

Could also be used for large counts.
Consider chemical reaction $A \rightarrow B$. Four competing models for the fraction of $A$ remaining after time $t$ minutes at temperature $T$

Model 1: $\mu_1 = \exp \left( -\theta_1 t \exp \left( -\frac{\theta_2}{T} \right) \right)$

Model 2: $\mu_2 = \left[ 1 + \theta_1 t \exp \left( -\frac{\theta_2}{T} \right) \right]^{-1}$

Model 3: $\mu_3 = \left[ 1 + 2\theta_1 t \exp \left( -\frac{\theta_2}{T} \right) \right]^{-1/2}$

Model 4: $\mu_4 = \left[ 1 + 3\theta_1 t \exp \left( -\frac{\theta_2}{T} \right) \right]^{-1/3}$

$y|m \sim N(\mu_m, \sigma^2)$ Two design variables $d = (t, T)$: 
$t \in \{0, 25, 50, 75, 100, 125, 150\}$ and $T \in \{450, 475, 525, 575, 600\}$ yielding 35 possible choices.
Chemical Example Continued

- 15 independent observations
- Four cases. Each model allowed to be true with parameter $(\theta_1, \theta_2, \sigma) = (400, 5000, 0.1)$
- Prior distribution

\[
\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N \left( \begin{bmatrix} 400 \\ 5000 \end{bmatrix}, \begin{bmatrix} 70 & 0 \\ 0 & 500 \end{bmatrix} \right) 1(\theta_1 > 0)1(\theta_2 > 0).
\]

\[\sigma \sim IG(10, 1)\]

- Comparing 3 different utility functions: Random design, mutual information and total separation (Masoumi et al 2013)
Chemical Reaction Results (model 1 true)

**Figure:** First order reaction as true.
Chemical Reaction Results (model 2 true)

Figure: Second order reaction as true.
Chemical Reaction Results (model 3 true)

Figure: Third order reaction as true.
Chemical Reaction Results (model 4 true)

Figure: Fourth order reaction as true.
Advantage 3: Embarrassingly Parallel

- Extend to design in presence of random effects model
- Pharmacokinetics Example. For subject $t$ the model for samples collected at design $d_t = (d_{1t}, d_{2t})$ is:
  \[
  y_t \sim \text{MVN}(g(\beta_t, d_t), \delta_0 I),
  \]
  \[
  \beta_t \sim \text{MVN}(\mu, \Omega),
  \]
- Here $\beta_t$ is random effect for $t$th subject
  \[
  g(\beta_t, d_t) = \frac{100}{\exp(\beta_{2t})} \exp\left(-\frac{\exp(\beta_{1t})}{\exp(\beta_{2t})} d_t \right),
  \]
- Priors:
  \[
  \mu \sim \text{MVN}(0, \Sigma), \text{ for } \Sigma \text{ known.}
  \]
  \[
  \Omega \sim \text{InvWish}(\Psi, \nu), \text{ for } \Psi \text{ and } \nu \text{ known}
  \]
  \[
  \delta_0 \sim \text{U}(a, b), \text{ } 0 < a \leq b < \infty,
  \]
- Design objective is to learn about parameters: $\theta = (\mu, \Omega, \delta_0)$.  

Chris Drovandi  
MCMski 2014, Chamonix, France
Exact-Approximate SMC

- The (marginal) likelihood

\[ f(y_t | \theta, d_t) = \int f(y_t | \beta_t, \delta_0, d_t) \pi(\beta_t | \mu, \Omega) d\beta_t \]

- Can be estimated unbiasedly. For each particle \( \theta^{(i)} \)

\[ f(y_t | \theta^{(i)}, d_t) = \frac{1}{M} \sum_{j=1}^{M} f(y_t | \beta^{(j)}_t, \delta^{(i)}_0, d_t) \] (1)

where \( \beta^{(j)}_t \sim MVN(\mu^{(i)}, \Omega^{(i)}) \), \( j = 1, \ldots, M \).

- SMC with unbiased estimate of likelihood \( \rightarrow \) an exact-approximate algorithm! (Duan and Fulop 2013)

- Caveat: SMC can degenerate quicker compared to using ‘exact’ likelihood (need large enough \( M \))
Serial implementation of SMC design too slow here

Many ways to parallelise algorithm (over particles, $M$ and/or design choices $d$)

Here we chose calculating the likelihood over particles in parallel on GPU (required in all aspects of algorithm)

Order of magnitude speed improvement over C implementation of likelihood

Work still in progress...
Key References


- Drovandi, C. C. et al. (2014). A Sequential Monte Carlo Algorithm to Incorporate Model Uncertainty in Bayesian Sequential Design. JCGS To Appear.

