An Introduction to Sequential Monte Carlo

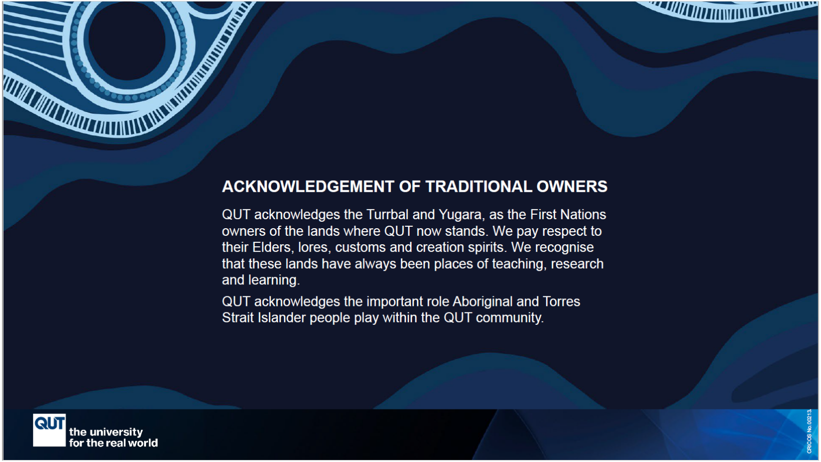
Lecture 1 - SMC Samplers

Matt Sutton

QUT Centre for Data Science (CDS)

Table of contents

Welcome


ACKNOWLEDGEMENT OF TRADITIONAL OWNERS
QUT acknowledges the Turrbal and Yugara, as the First Nations owners of the lands where QUT now stands. We pay respect to their Elders, lores, customs and creation spirits. We recognise that these lands have always been places of teaching, research and learning.
QUT acknowledges the important role Aboriginal and Torres Strait Islander people play within the QUT community.
QUT
the university
for the real world

Bayesian Inference

Aim of the Lecture

Aim

To provide an introduction into SMC samplers. Why to use them, how to use them, and how they work.

After this session, you should be able to code your own SMC sampler. The practical session will allow you to put these skills to practice!

What it is not: A complete introduction to Bayesian statistics, a deep dive into Monte Carlo methods, SMC theory (see the references for some good links).

Bayesian Statistics

Bayesian statistics combines information about θ from the prior with information obtained from the data y via Bayes rule.

This produces the *posterior distribution*.

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{Z},$$

where $Z = \int_\theta f(y|\theta)\pi(\theta)d\theta$ is the normalising constant or *marginal likelihood*.

Bayesian Statistics

Advantages: Coherent framework for incorporating domain expertise and historical data.

Full uncertainty quantification through the posterior.

Drawback: Computation associated with Bayesian statistics can be heavy.

We might care about estimating

- The posterior distribution itself (e.g. density estimation)
- Posterior expectations $\mathbb{E}_{\pi(\theta|y)}[h(\theta)] = \int_\theta h(\theta)\pi(\theta|y)d\theta$
  - What is a reasonable estimate of θ? $\rightarrow$ $\bar{\theta} = \mathbb{E}_{\pi(\theta|y)}[\theta]$
  - How sure are we about θ? $\rightarrow$ $\mathbb{E}_{\pi(\theta|y)}[(\theta - \bar{\theta})^2]$
- The normalising constant Z for model choice

We will talk about Monte Carlo methods to do this, which aim to get (weighted) samples from the posterior.

Sampling Methods

Key Ingredients:

- Markov chain Monte Carlo
- Importance sampling

(Rosenbluth) Metropolis-Hastings Algorithm:

Figure 1: Adrianna and Marshall Rosenbluth

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. & Teller, A. H., Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087–1092.

See also,

Gubernatis, J. E. (2005). Marshall Rosenbluth and the Metropolis algorithm. *Physics of Plasmas*, 12(5), 057303.

Markov chain Monte Carlo (MCMC)

Goal

Want to get samples $\xi \sim \pi(\theta|y)$

Generate a sequence (Markov chain)

$$\xi_1, \xi_2, \ldots, \xi_N,$$

that will converge in distribution to $\pi(\theta|y)$.

Given the current value ξ_i we make a proposal for ξ_{i+1} using a distribution $q(\cdot|\xi_i)$. A common choice is a random walk q

$$\xi^* \sim \mathcal{N}(\xi_i, \Sigma).$$

We choose to accept ξ^* or continue with ξ_i.

Markov chain Monte Carlo (MCMC)

Random walk MCMC

1. Propose a candidate $\xi^* \sim \mathcal{N}(\xi_i, \Sigma)$
2. Calculate
$$\text{RMH} = \min\left\{1, \frac{\pi(\theta = \xi^*|y)}{\pi(\theta = \xi_i|y)}\right\}$$
3. With probability RMH set $\xi_{i+1} = \xi^*$ otherwise, set $\xi_{i+1} = \xi_i$

The values $\{\xi_1, \xi_2, ..., \xi_N\}$ form a Markov chain.

If $\xi_i \sim \pi(\theta|y)$ then $\xi_{i+1} \sim \pi(\theta|y)$.

This is a π-invariant Markov kernel, $\xi_{i+1} \sim K_\pi(\cdot|\xi_i)$

Markov chain Monte Carlo (MCMC)

MCMC with a general proposal distribution

1. Propose a candidate $\xi^* \sim q(\cdot|\xi_i)$
2. Calculate
$$\text{RMH} = \mathbf{min}\left\{1, \frac{\pi(\theta = \xi^*|y)q(\xi_i|\xi^*)}{\pi(\theta = \xi_i|y)q(\xi^*|\xi_i)}\right\}$$
3. With probability RMH set $\xi_{i+1} = \xi^*$ otherwise, set $\xi_{i+1} = \xi_i$

The values $\{\xi_1, \xi_2, ..., \xi_N\}$ form a Markov chain.

If $\xi_i \sim \pi(\theta|y)$ then $\xi_{i+1} \sim \pi(\theta|y)$.

This is a π-invariant Markov kernel, $\xi_{i+1} \sim K_\pi(\cdot|\xi_i)$

MCMC in action

https://chi-feng.github.io/mcmc-demo/app.html?algorithm=RandomWalkMH&target=banana

The issues (poor mixing or convergence)

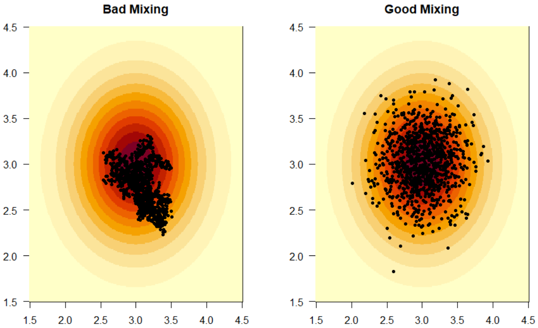


Figure 2: Two MCMC samplers run for $N = 1000$ iterations

Markov chain Monte Carlo

Suppose we have a Markov chain $\xi_1, \xi_2, ..., \xi_N$ with limiting distribution $\pi(\theta|y)$.

The Ergodic theorem states that,

$$\frac{1}{N}\sum_{i=1}^{N} h(\xi_i) \to \mathbb{E}_{\pi(\theta|y)}[h(\theta)], \qquad \text{as } N \to \infty$$

where h is some function of interest.

Key Ingredients:

- Markov chain Monte Carlo
- Importance sampling

Self normalised importance sampling

Choose a density q where $q(\theta) > 0$ whenever $\pi(\theta|y) \neq 0$. Then use

$$\tilde{w}_i = \frac{f(y|\xi_i)\pi(\xi_i)}{q(\xi_i)}, \;\; \xi_i \sim q(\theta)$$

for $i = 1, \ldots, N$ and

$$\hat{I}^{IS} = \sum_{i=1}^{N} h(\xi_i) W_i, \;\; W_i = \frac{\tilde{w}_i}{\sum_{i=1}^{N} \tilde{w}_i}.$$

The collection $\{W_i . \xi_i\}_{i=1}^{N}$ form a weighted sample from $\pi(\theta|y)$.

Self normalised importance sampling

Can obtain an estimate of the unknown normalising constant:

$$Z = \int_\theta f(y|\theta)\pi(\theta)d\theta \approx \frac{1}{N}\sum_{i=1}^{N} \tilde{w}_i$$

Importance sampling for sampling from target distributions

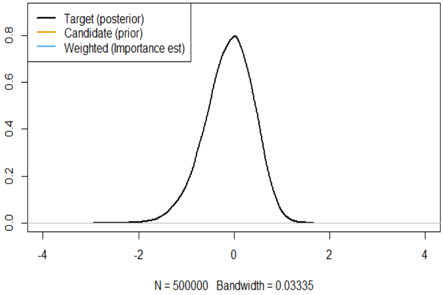


Figure 3: Example importance sampling

Importance sampling for sampling from target distributions

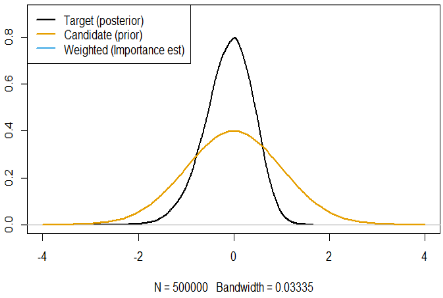


Figure 4: Example importance sampling

Importance sampling for sampling from target distributions

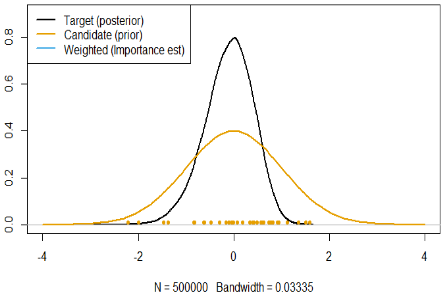


Figure 5: Example importance sampling

Importance sampling for sampling from target distributions

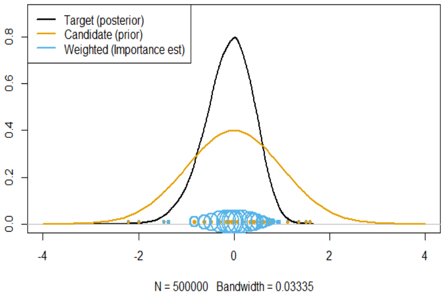


Figure 6: Example importance sampling

Importance sampling for sampling from target distributions

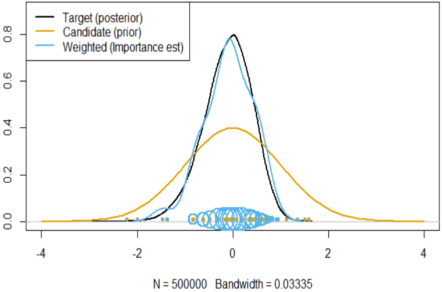


Figure 7: Example importance sampling

Effective Sample Size

These are not ‘perfect’ samples

Efficiency of the samples can be assessed via Effective Sample Size (ESS).

$$ESS = N \frac{\mathrm{var}_{\pi}[\hat{I}]}{\mathrm{var}_{q}[\hat{I}^{IS}]}$$

The exact definition is intractable in practice but the following Monte Carlo estimator can be used

$$\widehat{ESS} = 1 / \sum_{i=1}^{N} (W_i)^2$$

We can see from this formulation that $1 \leq \mathrm{ESS} \leq N$.

Challenge for importance sampling

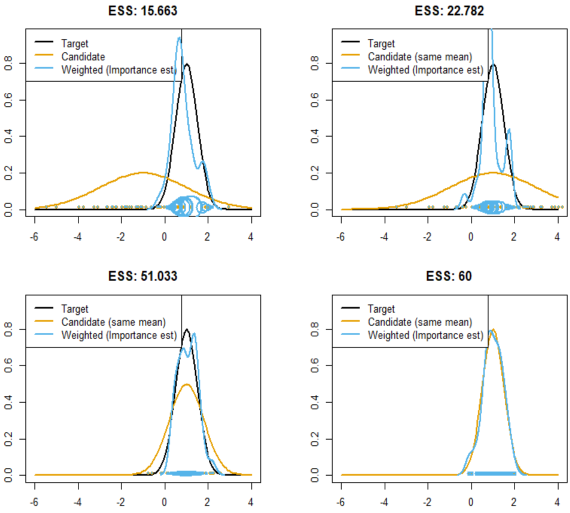


Figure 8: Efficiency for N=60 samples

Sequential Monte Carlo Samplers

Introduction to SMC samplers

We'll now introduce SMC samplers [3][2]

Benefits can include:

- Better sampling from complex posteriors*
- Can be parallelised
- Estimates the normalising constant of the posterior as a byproduct
- Can handle streaming data**

* for likelihood annealing SMC

** For data annealing SMC

Sequential Monte Carlo for static models

An extension of importance sampling

Requires a sequence of slowly evolving distributions, $\pi_1(\theta), \pi_2(\theta), \ldots, \pi_T(\theta)$, where π_1 is easy to sample from and π_T is the ultimate target.

Involves traversing a set of N weighted samples ('particles') through the sequence of distributions.

Sequence of distributions (data annealing)

Need to set-up a sequence of distributions to get us to the posterior:

Data Annealing (bring in data one-at-a-time or in batches) [1]

Start from the prior $\pi_1(\theta) = \pi(\theta)$,

$$\pi_t(\theta) \propto f(y_{1:(t-1)}|\theta)\pi(\theta), \quad \text{for } t = 2, \ldots, T,$$

for data $y = (y_1, y_2, ..., y_{T-1})$.

Great for streaming data!

Sequence of distributions (likelihood annealing)

Likelihood Annealing (power up the likelihood)

$$\pi_t(\theta|y) \propto f(y|\theta)^{\gamma_t}\pi(\theta) \text{ for } t = 1, \ldots, T,$$

Start from the prior $\pi_1(\theta) = \pi(\theta)$ (i.e. $\gamma_1 = 0$) evolve to the posterior $\pi_T(\theta) = \pi(\theta|y)$ (i.e. $\gamma_T = 1$).

Great for complex posteriors!

Sequential Monte Carlo for static models

Involves traversing a set of N weighted samples ('particles') through the sequence of distributions by iteratively applying the following steps:

- re-weighting (importance sampling)
- re-sampling
- diversifying (MCMC update)

Reweighting

At the start of iteration t, we have a population of N particles, $\{\xi_{t-1}^i, W_{t-1}^i\}_{i=1}^N$, approximating π_{t-1}.

Need to 're-weight' these samples so that they reflect target t

$$\tilde{w}_t^i = W_{t-1}^i \frac{\pi_t(\xi_{t-1}^i)}{\pi_{t-1}(\xi_{t-1}^i)} = W_{t-1}^i f(y|\theta = \xi_{t-1}^i)^{\gamma_t - \gamma_{t-1}}, \text{ for } i = 1, \ldots, N.$$

where $\tilde{w}_t^i$ are the unnormalised weights. Set

$$W_t^i = \frac{\tilde{w}_t^i}{\sum_{j=1}^N \tilde{w}_t^j}$$

and $\xi_t^i = \xi_{t-1}^i$ for $i = 1, \ldots, N$ then $\{\xi_t^i, W_t^i\}_{i=1}^N$ represents a weighted sample from π_t.

Resampling

Then we apply a ’resampling’ step to boost ESS back to N. Draw N samples from $\{\xi_t^i\}_{i=1}^N$ with probability proportional to the weights $\{W_t^i\}_{i=1}^N$.

Resampling duplicates promising particles (high weight) and eliminates particles with very low weight.

Resampling

Then we apply a ’resampling’ step to boost ESS back to N. Draw N samples from $\{\xi_t^i\}_{i=1}^N$ with probability proportional to the weights $\{W_t^i\}_{i=1}^N$.
Resampling duplicates promising particles (high weight) and eliminates particles with very low weight.

Sample ancestors $A_t^i \sim \text{Cat}(W_t^1, \dots, W_t^N)$ for $i \in \{1, \dots, N\}$

The new particles set is $\{\xi_t^i = \xi_t^{A_t^i}, W_t^i = \frac{1}{N}\}_{i=1}^N$ which targets π_t.

Sumeetpal Singh will have more to say on this!

Move (Diversification)

Following re-sampling step, the estimated ESS is N, but there will be some duplicated particles.

Want to diversify particles, but can't just move them any old way! Need to preserve the current target.

Solution: Move particles with an MCMC kernel with invariant distribution π_t. May want to repeat move step R times on each particle (not guaranteed move with MCMC kernel)

Likelihood annealing SMC Samplers

Input: $\gamma = (\gamma_1 = 0, \ldots, \gamma_T = 1)$, random walk covariance Σ, number MCMC steps R.

1. Sample initial $\xi_1^i \sim \pi_1(\theta)$ and set $W_1^i = \frac{1}{N}$ for $i \in \{1, ..., N\}$
2. For each time $t = 2, ..., T$
   (i) Calculate weights (Importance sampling)

$$\tilde{w}_t^i = W_{t-1}^i f(y|\theta = \xi_{t-1}^i)^{\gamma_t - \gamma_{t-1}} \qquad W_t^i = \frac{\tilde{w}_t^i}{\sum_{j=1}^N \tilde{w}_t^j},$$

   for $i = 1, \ldots, N$.

   (ii) Resample $A_t^i \sim \text{Cat}(W_t^1, ..., W_t^N)$ and set $W_t^i = 1/N$ for $i \in \{1, ..., N\}$

   (iii) Diversify using R MCMC iterations with proposal covariance Σ

$$\xi_t^i \sim K_{\pi_t}(\cdot \mid \theta = \xi_{t-1}^{A_t^i}), \text{ for } i = 1, \ldots, N.$$

Likelihood annealing SMC Samplers

Input: $\gamma = (\gamma_1 = 0, \ldots, \gamma_T = 1)$, random walk covariance Σ, number MCMC steps R.

1. Sample initial $\xi_1^i \sim \pi_1(\theta)$ and set $W_1^i = \frac{1}{N}$ for $i \in \{1, ..., N\}$
2. For each time $t = 2, ..., T$
   (i) Calculate weights (Importance sampling)

$$\tilde{w}_t^i = W_{t-1}^i f(y|\theta = \xi_{t-1}^i)^{\gamma_t - \gamma_{t-1}} \qquad W_t^i = \frac{\tilde{w}_t^i}{\sum_{j=1}^N \tilde{w}_t^j},$$

   for $i = 1, \ldots, N$.
   (ii) Resample $A_t^i \sim \mathrm{Cat}(W_t^1, ..., W_t^N)$ and set $W_t^i = 1/N$ for $i \in \{1, ..., N\}$
   (iii) Diversify using R MCMC iterations with proposal covariance Σ

$$\xi_t^i \sim K_{\pi_t}(\cdot \mid \theta = \xi_{t-1}^{A_t^i}), \text{ for } i = 1, \ldots, N.$$

Francesca Crucinio will talk more about parallelism in SMC

Summary

Goal: Generate samples from a distribution $\pi(\theta)$

SMC Samplers give:

- Particles $\xi_T^{1:N}$ targeting $\pi(\theta|y)$ to use for posterior inference
- An estimate of the normalising constant:

$$\prod_{t=2}^{T}\sum_{i=1}^{N}\tilde{w}_t^i \approx \int_\theta f(y|\theta)\pi(\theta)d\theta$$

by iterating a sequence of distributions by **re-weighting** (importance sampling), **re-sampling** to duplicate promising particles and **moving** (MCMC) to diversify the particles.

Adaptation

Adaptation

Some choices:

C1 How many tempering distributions π_t should we use? What should I pick for γ_t for $t = 2, \ldots$?

C2 What should we use for the random walk Σ proposal (MCMC kernel)?

Adaptation in SMC can automate these considerations.

Even with some adaptation, the samples give consistent estimates.

Saifuddin Syed will have more to talk about on this!

Choosing the tempering sequence (C1)

Recall the weights for target t are given by

$$W_t^i \propto f(y|\theta = \xi_{t-1}^i)^{\gamma_t - \gamma_{t-1}}$$

We know that ESS = $1/\sum_{i=1}^{N}(W_t^i)^2$. Can view ESS as a function of γ_t (ESS(γ_t)). Thus we want to 'find' a γ_t so that ESS is close to the ESS threshold, e.g. $N/2$.

Amounts to solving $h(\gamma_t) = \text{ESS}(\gamma_t) - N/2 = 0$. One strategy is to use bisection method with initial bounds (γ_{t-1},1).

Choosing the MCMC kernel (C2)

After the re-sample step at iteration t, we have a particle set $\{\xi_t^i, 1/N\}_{i=1}^N$ distributed according to π_t (but with some duplicated particles).

Can use these particles to form an efficient MCMC proposal distribution for the MCMC step. For example, compute the sample covariance matrix of the particles and use it in a multivariate normal random walk proposal

The adaptive likelihood annealing SMC sampler

Input: Number MCMC steps R.

1. Draw $\xi_1^i \sim \pi_1(\theta)$ for $i = 1, \ldots, N$.
2. While $\gamma_t \neq 1$
   (i) If ESS($\gamma = 1$) $> N/2$ set $\gamma_t = 1$ otherwise find γ_t such that ESS(γ) $\approx N/2$.
   (ii) Calculate weights

$$\tilde{w}_t^i = W_{t-1}^i f(y|\theta)^{\gamma_t - \gamma_{t-1}} \qquad W_t^i = \frac{\tilde{w}_t^i}{\sum_{j=1}^N \tilde{w}_t^j}, \text{ for } i = 1, \ldots, N.$$

   (iii) Resample $A_t^i \sim$Cat$(W_t^1, ..., W_t^N)$ and set $W_t^i = 1/N$ for $i \in \{1, ..., N\}$
   (iv) Compute $\hat{\Sigma}$ as the sample covariance of the particles $\{\xi_{t-1}^{A_t^i}\}_{i=1}^N$.
   (v) Diversify using R MCMC iterations with proposal covariance $\hat{\Sigma}$

$$\xi_t^i \sim K_t(\theta_t \mid \theta_{t-1} = \xi_{t-1}^{A_t^i}), \text{ for } i = 1, \ldots, N.$$

Example

Coding Section

To start off head to: **https://github.com/bonStats/HelloSMC.jl/**

Hit the launch binder button.

Try out the following:

1. lvsmc_algo/
    - The notebook smc_sampler_lanneal.ipynb contains details introducing the SMC sampler and a primer for writing your own adaptive version.
    - The notebook lv_smc_example.ipynb contains some example code running the adaptive version of the sampler.
2. lvsmc/
    - smc-sampler-lotka-volterra.ipynb contains details on how to set up an SMC sampler using the SequentialMonteCarlo.jl package

Ex 1) Inference for population dynamics

Aim

To learn a model for the population dynamics of a predator-prey relationship.

$X_{1,t}$ the population of the prey, $X_{2,t}$ the population of the predator at time t.

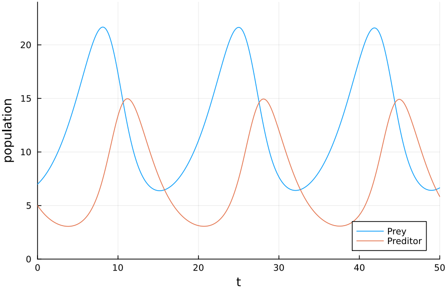


Figure 9: Predator-Prey Dynamics (Lotka Volterra)

Ex 1) Inference for population dynamics

$X_{1,t}$ the population of the prey, $X_{2,t}$ the population of the predator at time t evolve according to:

$$\frac{dx_1}{dt} = (\alpha - \beta x_2)x_1 \qquad \frac{dx_2}{dt} = (\beta x_1 - \gamma)x_2$$

Observe a noisy measurement at times $t = 0, ..., 50$.

$$Y_{1,t} \sim \mathcal{N}(X_{1,t}, \sigma^2), \qquad Y_{2,t} \sim \mathcal{N}(X_{2,t}, \sigma^2)$$

We want to estimate $\theta = (\alpha, \beta, \gamma, \sigma)$ for the model.

Assign some priors $\alpha, \gamma \sim U(0, 1)$ $\beta \sim U(0, 0.1)$ and $\sigma \sim U(0.1, 7)$.

Ex 1) Inference for population dynamics

Observe a noisy measurement at times $t = 0, \ldots, 50$.

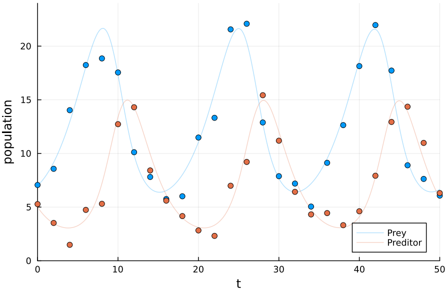


Figure 10: Noisy observations

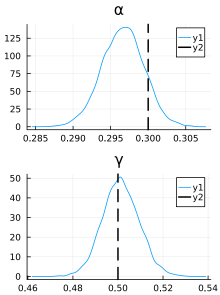


Figure 11: Marginal posterior distribution $\pi(\theta|y)$

References i

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

C. Dai, J. Heng, P. E. Jacob, and N. Whiteley.
An invitation to sequential monte carlo samplers.
Journal of the American Statistical Association, 117(539):1587–1600, July 2022.

P. Del Moral, A. Doucet, and A. Jasra.
Sequential monte carlo samplers.
Journal of the Royal Statistical Society. Series B, Statistical methodology, 68(3):411–436, May 2006.