Sequential Monte Carlo

Dr. Jarad Niemi

STAT 615 - Iowa State University

October 20, 2017

Outline

1. State-space models $p(y|\theta,\psi)p(\theta|\psi)$
   - Definition
   - Terminology
   - Notation
2. State inference $p(\theta|y,\psi)$
   - Exact inference
   - Importance sampling
   - Sequential importance sampling
   - Bootstrap filter - resampling
   - Auxiliary particle filter
3. State and parameter inference $p(\theta,\psi|y)$
   - Bootstrap filter
   - Kernel density
   - Sufficient statistics
4. Advanced SMC
   - SMC-MCMC
   - Fixed parameter
   - SMC for marginal likelihood calculations

Definition

Bayes' rule is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

In this rule, B represents what we know about the world and A represents what we don't.

Suppose $p(\theta_t, \psi|y_{1:t-1})$ is our current knowledge about the state of the world. We observe datum y_t then

$$p(\theta_t, \psi|y_{1:t}) = \frac{p(y_t|\theta_t, \psi)p(\theta_t, \psi|y_{1:t-1})}{p(y_t|y_{1:t-1})}.$$

where $y_{1:t} = (y_1, y_2, \ldots, y_t)$.

Definition

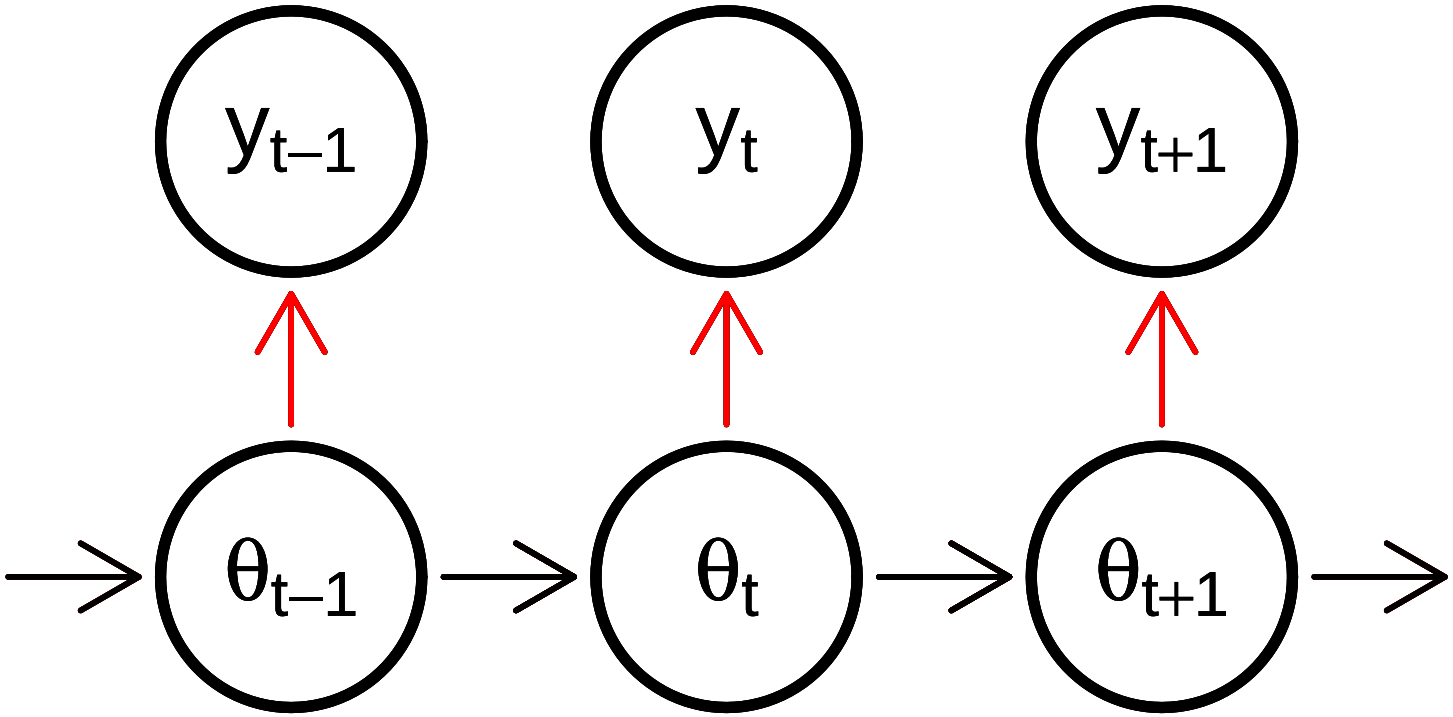
A state-space model can be described by these conditional distributions:

- an observation equation: $p_o(y_t|\theta_t, \psi)$,
- an evolution equation: $p_e(\theta_t|\theta_{t-1}, \psi)$, and
- a prior $p(\theta_0, \psi)$.

where

- y_t: an observation vector of length m
- θ_t: a latent state vector of length p
- ψ: a fixed parameter vector of length q

Graphical representation

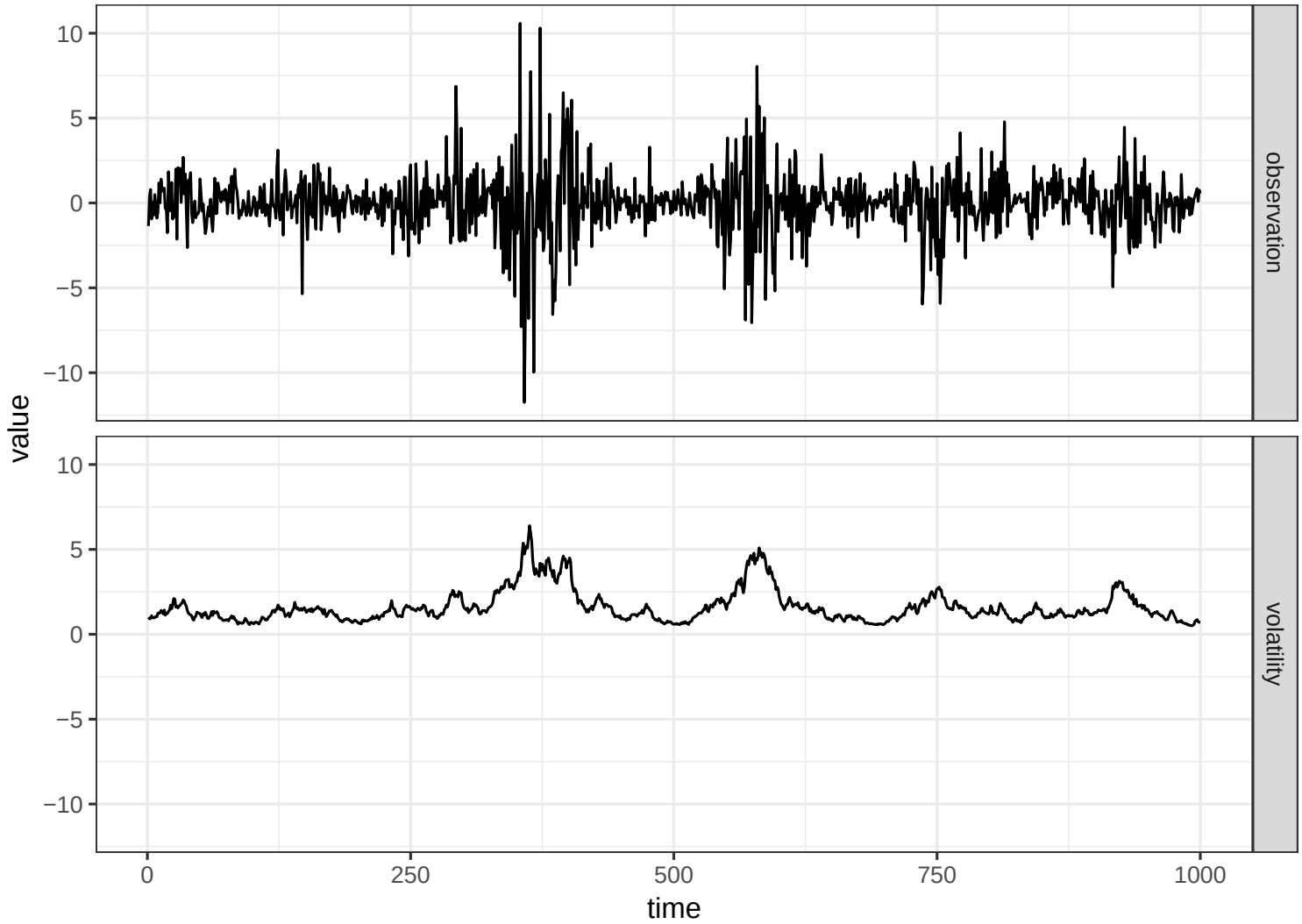


- $p(\theta_t|\theta_{t-1}, \psi)$
- $p(y_t|\theta_t, \psi)$

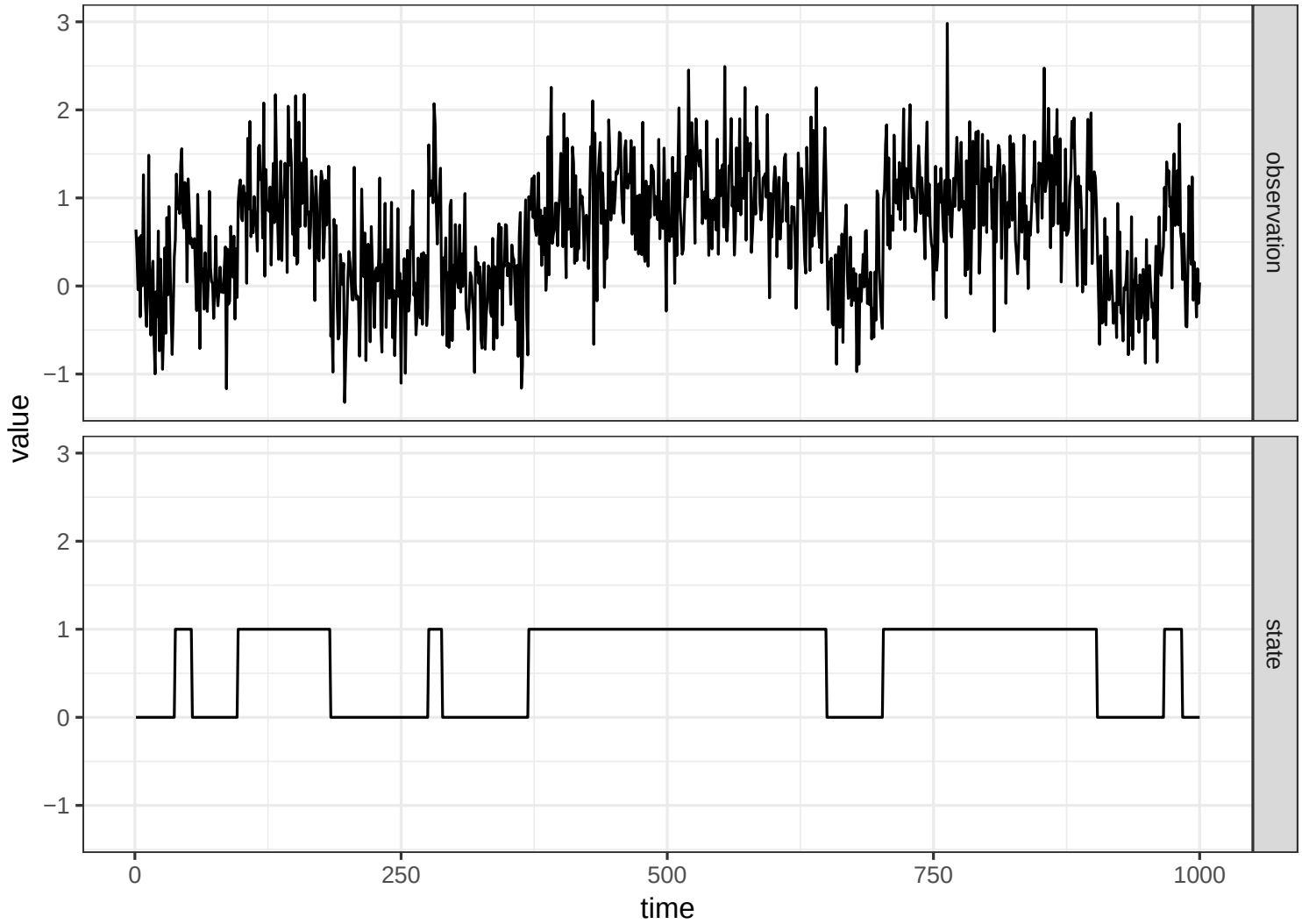
Interpretation

Model	State interpretation
Local level model	True level
Linear growth model	True level and slope
Seasonal factor model	Seasonal effect
Dynamic regression	Time-varying regression coefficients
Stochastic volatility	Underlying volatility in the market
Markov switching model	Influenza epidemic on/off

Stochastic volatility



Markov switching model



Inference

Definition

The state filtering distribution is the distribution for the state conditional on all observations up to and including time t, i.e.

$$p(\theta_t|y_{1:t}, \psi) = p(\theta_t|y_1, y_2, \ldots, y_t, \psi).$$

Definition

The state smoothing distribution is the distribution for the state conditional on all observed data, i.e.

$$p(\theta_t|y_{1:T}, \psi) = p(\theta_t|y_1, y_2, \ldots, y_T, \psi)$$

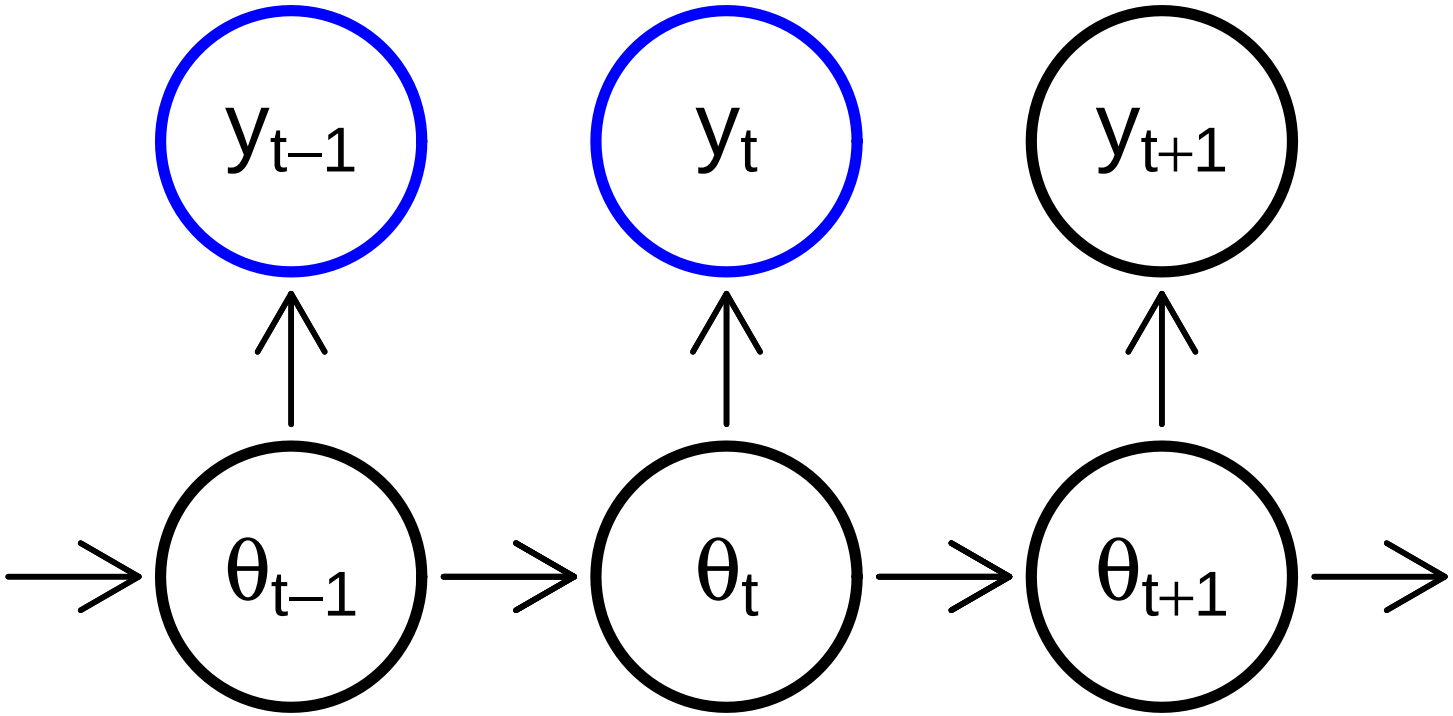
where $t < T$.

Definition

The state forecasting distribution is the distribution for future states conditional on all observed data, i.e.

$$p(\theta_{T+k}|y_{1:T}, \psi) = p(\theta_{T+k}|y_1, y_2, \ldots, y_T, \psi)$$

where $k > 0$.



- Filtering
- Smoothing
- Forecasting

Filtering

Goal: $p(\theta_t|y_{1:t})$ (filtered distribution)

Recursive procedure:

- Assume $p(\theta_{t-1}|y_{1:t-1})$
- Prior for θ_t

$$\begin{aligned} p(\theta_t|y_{1:t-1}) &= \int p(\theta_t, \theta_{t-1}|y_{1:t-1})d\theta_{t-1} \\ &= \int p(\theta_t|\theta_{t-1}, y_{1:t-1})p(\theta_{t-1}|y_{1:t-1})d\theta_{t-1} \\ &= \int p(\theta_t|\theta_{t-1})p(\theta_{t-1}|y_{1:t-1})d\theta_{t-1} \end{aligned}$$

- One-step ahead predictive distribution for y_t

$$\begin{aligned} p(y_t|y_{1:t-1}) &= \int p(y_t, \theta_t|y_{1:t-1})d\theta_t \\ &= \int p(y_t|\theta_t, y_{1:t-1})p(\theta_t|y_{1:t-1})d\theta_t \\ &= \int p(y_t|\theta_t)p(\theta_t|y_{1:t-1})d\theta_t \end{aligned}$$

- Filtered distribution for θ_t

$$p(\theta_t|y_{1:t}) = \frac{p(y_t|\theta_t, y_{1:t-1})p(\theta_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} = \frac{p(y_t|\theta_t)p(\theta_t|y_{1:t-1})}{p(y_t|y_{1:t-1})}$$

Start from $p(\theta_0)$.

Smoothing

Goal: $p(\theta_t|y_{1:T})$ for $t < T$

- Backward transition probability $p(\theta_t|\theta_{t+1}, y_{1:T})$

$$\begin{aligned} p(\theta_t|\theta_{t+1}, y_{1:T}) &= p(\theta_t|\theta_{t+1}, y_{1:t}) \\ &= \frac{p(\theta_{t+1}|\theta_t, y_{1:t})p(\theta_t|y_{1:t})}{p(\theta_{t+1}|y_{1:t})} \\ &= \frac{p(\theta_{t+1}|\theta_t)p(\theta_t|y_{1:t})}{p(\theta_{t+1}|y_{1:t})} \end{aligned}$$

- Recursive smoothing distributions $p(\theta_t|y_{1:T})$ assuming we know $p(\theta_{t+1}|y_{1:T})$

$$\begin{aligned} p(\theta_t|y_{1:T}) &= \int p(\theta_t, \theta_{t+1}|y_{1:T})d\theta_{t+1} \\ &= \int p(\theta_{t+1}|y_{1:T})p(\theta_t|\theta_{t+1}, y_{1:T})d\theta_{t+1} \\ &= \int p(\theta_{t+1}|y_{1:T})\frac{p(\theta_{t+1}|\theta_t)p(\theta_t|y_{1:t})}{p(\theta_{t+1}|y_{1:t})}d\theta_{t+1} \\ &= p(\theta_t|y_{1:t})\int \frac{p(\theta_{t+1}|\theta_t)}{p(\theta_{t+1}|y_{1:t})}p(\theta_{t+1}|y_{1:T})d\theta_{t+1} \end{aligned}$$

Start from $p(\theta_T|y_{1:T})$

Forecasting

Goal: $p(y_{T+k}, \theta_{T+k}|y_{1:T})$

$$p(y_{T+k}, \theta_{T+k}|y_{1:T}) = p(y_{T+k}|\theta_{T+k})p(\theta_{T+k}|y_{1:T})$$

Recursively, given $p(\theta_{T+(k-1)}|y_{1:T})$

$$\begin{aligned} p(\theta_{T+k}|y_{1:T}) &= \int p(\theta_{T+k}, \theta_{T+(k-1)}|y_{1:T})\, d\theta_{T+(k-1)} \\ &= \int p(\theta_{T+k}|\theta_{T+(k-1)}, y_{1:T})p(\theta_{T+(k-1)}|y_{1:T})d\theta_{T+(k-1)} \\ &= \int p(\theta_{T+k}|\theta_{T+(k-1)})p(\theta_{T+(k-1)}|y_{1:T})d\theta_{T+(k-1)} \end{aligned}$$

Start with $k = 1$.

Outline

1. State-space models $p(y|\theta,\psi)p(\theta|\psi)$
   - Definition
   - Terminology
   - Notation
2. State inference $p(\theta|y,\psi)$
   - Exact inference
   - Importance sampling
   - Sequential importance sampling
   - Bootstrap filter - resampling
   - Auxiliary particle filter
3. State and parameter inference $p(\theta,\psi|y)$
   - Bootstrap filter
   - Kernel density
   - Sufficient statistics
4. Advanced SMC
   - SMC-MCMC
   - Fixed parameter
   - SMC for marginal likelihood calculations

Exact inference

Our goal for most of today is to find filtering methods.

- We assume $p(\theta_{t-1}|y_{1:t-1})$ is known
- and try to obtain $p(\theta_t|y_{1:t})$ using
- $p(\theta_t|\theta_{t-1})$ and $p(y_t|\theta_t)$.

Then, starting with $p(\theta_0|y_0) = p(\theta_0)$ we can find $p(\theta_t|y_{1:t})$ for all t.

There are two important state-space models when the filtering updating is availably analytically:

- Hidden Markov models
- Dynamic linear models

Hidden Markov models

Definition

A hidden Markov model (HMM) is a state-space model with an arbitrary observation equation and an evolution equation that can be represented by a transition probability matrix, i.e.

$$p(\theta_t = j|\theta_{t-1} = i) = p_{ij}.$$

Filtering in HMMs

Suppose we have a HMM with p states. Let $q_i = p(\theta_{t-1} = i|y_{1:t-1})$, then

$$\begin{aligned} p(\theta_t = j|y_{1:t-1}) &= \sum_{i=1}^{p} q_i p_{ij} \\ p(\theta_t = j|y_{1:t}) &\propto p(y_t|\theta_t = j)p(\theta_t = j|y_{1:t-1}). \end{aligned}$$

If $p_i \propto a_i$ for $i \in \{1, 2, \ldots, p\}$, then $p_i = \frac{a_i}{\ldots}$

Definition

A dynamic linear model (DLM) is a state-space model where both the observation and evolution equations are linear in the states and have additive Gaussian errors and the prior is Gaussian, i.e.

$$\begin{aligned} y_t &= F_t\theta_t + v_t & v_t &\sim N(0, V_t) \\ \theta_t &= G_t\theta_{t-1} + w_t & w_t &\sim N(0, W_t) \\ \theta_0 &\sim N(m_0, C_0) \end{aligned}$$

where v_t, w_t, and θ_0 are independent of each other and mutually independent through time.

- Kalman filter
- Kalman smoother

Suppose $\theta_{t-1}|y_{1:t-1} \sim N(m_{t-1}, C_{t-1})$, then

- The one-step-ahead prior distribution for θ_t is $\theta_t|y_{1:t-1} \sim N(a_t, R_t)$ where

$$\begin{aligned} a_t &= E(\theta_t|y_{1:t-1}) = G_t m_{t-1}, \\ R_t &= Var(\theta_t|y_{1:t-1}) = G_t C_{t-1} G_t' + W_t. \end{aligned}$$

- The one-step-ahead predictive distribution for y_t is $y_t|y_{1:t-1} \sim N(f_t, Q_t)$ where

$$\begin{aligned} f_t &= E(y_t|y_{1:t-1}) = F_t a_t, \\ Q_t &= Var(y_t|y_{1:t-1}) = F_t R_t F_t' + V_t. \end{aligned}$$

- The filtering distribution of θ_t is $\theta_t|y_{1:t} \sim N(m_t, C_t)$ where

$$\begin{aligned} m_t &= E(\theta_t|y_{1:t}) = a_t + R_t F_t' Q_t^{-1} e_t, \\ C_t &= Var(\theta_t|y_{1:t}) = R_t - R_t F_t' Q_t^{-1} F_t R_t, \end{aligned}$$

where $e_t = y_t - f_t$ is the forecast error.

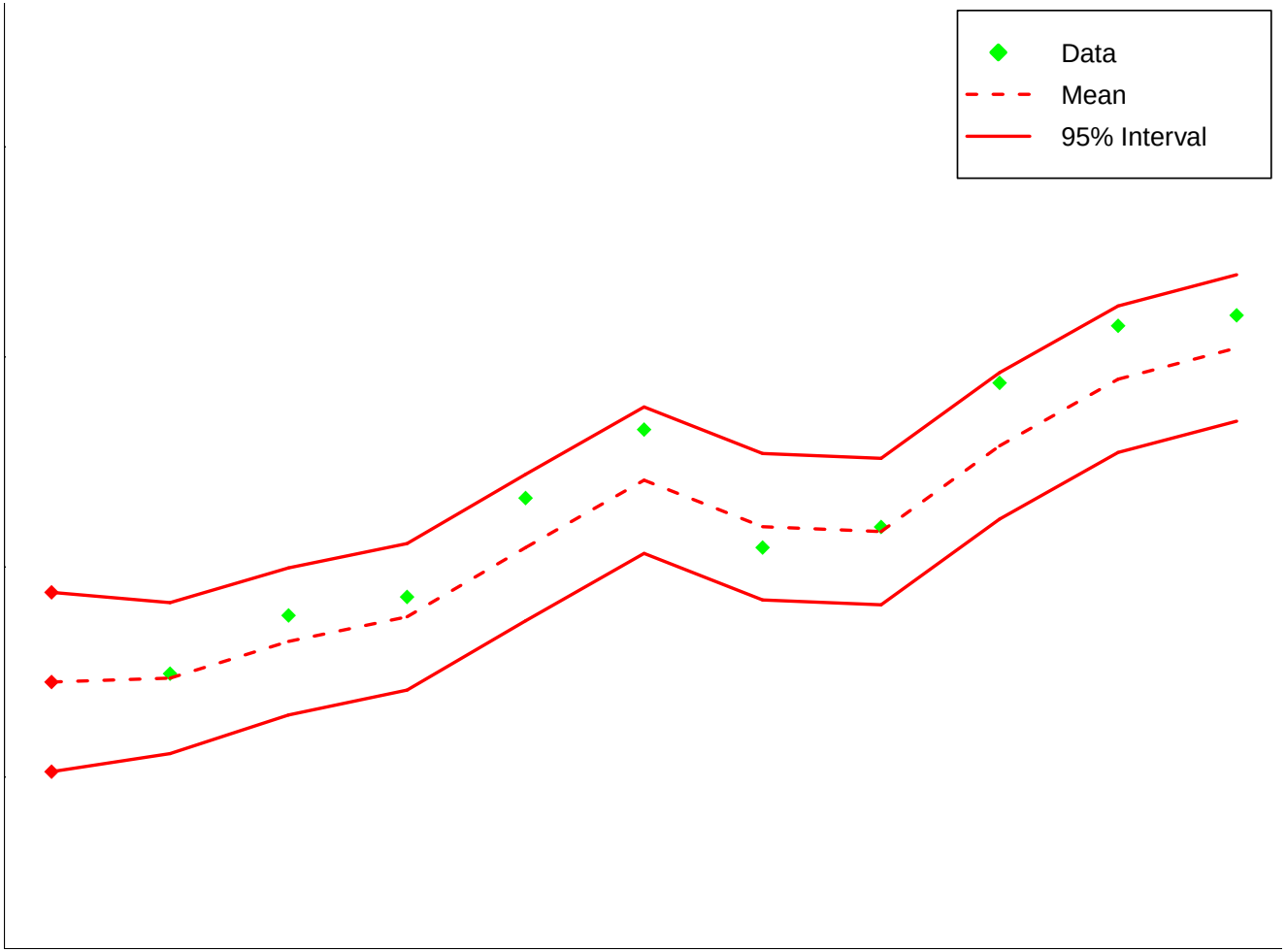
Test model:

$$
\begin{aligned}
y_t &= \theta_t + v_t \\
\theta_t &= \alpha + \beta\theta_{t-1} + w_t \\
v_t &\overset{ind}{\sim} N(0, V) \\
w_t &\overset{ind}{\sim} N(0, W) \\
\theta_0 &\sim N(m_0, C_0)
\end{aligned}
$$

Assume

$$\begin{aligned}
\theta_{t-1}|y_{1:t-1} &\sim N(m_{t-1}, C_{t-1}) \\
\\
\theta_t|y_{1:t-1} &\sim N(a_t, R_t) \\
a_t &= \alpha + \beta m_{t-1} \\
R_t &= \beta^2 C_{t-1} + W \\
\\
y_t|y_{1:t-1} &\sim N(f_t, Q_t) \\
f_t &= a_t \\
Q_t &= R_t + V \\
\\
\theta_t|y_{1:t} &\sim N(m_t, C_t) \\
C_t &= \left(\frac{1}{R_t} + \frac{1}{V}\right)^{-1} \\
m_t &= C_t\left(\frac{a_t}{R_t} + \frac{y_t}{V}\right)
\end{aligned}$$

Kalman filter updating



Dynamic linear models are a rich class of models:

- Trend
- Seasonal
- Dynamic regression
- ARIMA
- Seeming unrelated time series equations
- Seemingly unrelated regression models
- Hierarchical DLMs
- Multivariate ARMA models

- Petris, Petrone, Campagnoli. (2009) Dynamic Linear Models with R.
- West and Harrison. (1997) Bayesian Forecasting and Dynamic Models.

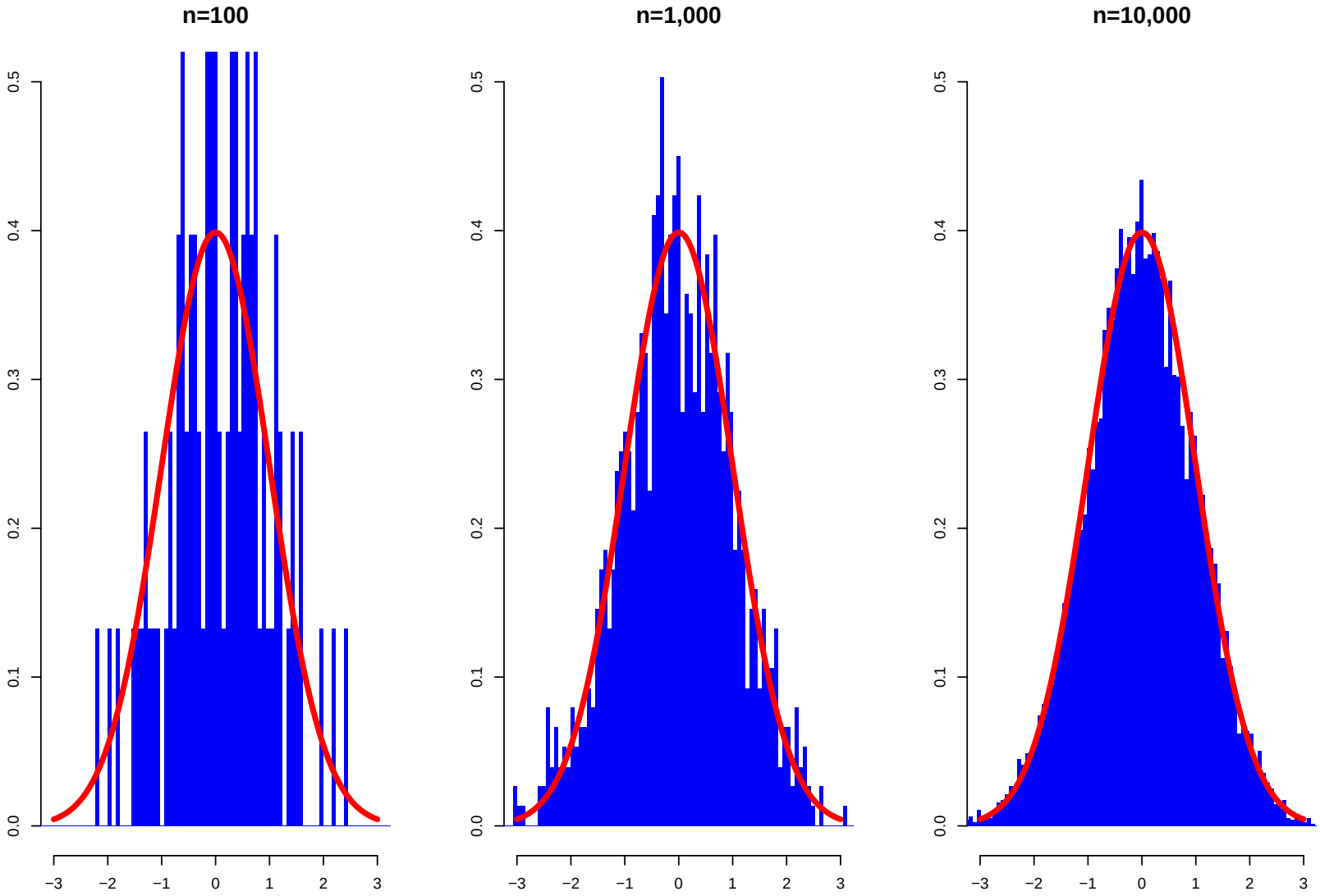
Approximate inference

HMMs and DLMs are the main classes of models with closed form updating of the filtering distributions.

Generally, no closed form expression exists and we must use an approximation.

- Numerical approximations
  - Extended Kalman filter
  - Bound optimal filter
  - Gaussian sum filter
  - Quadrature filter
- Monte Carlo approximations
  - Markov chain Monte Carlo (MCMC)
  - Sequential Monte Carlo (SMC)
    - Bootstrap filter
    - Auxiliary particle filter

Suppose we want to approximate some density $f(\theta)$, e.g. $p(\theta_t|y_{1:t})$. Draw samples from $f(\theta)$.

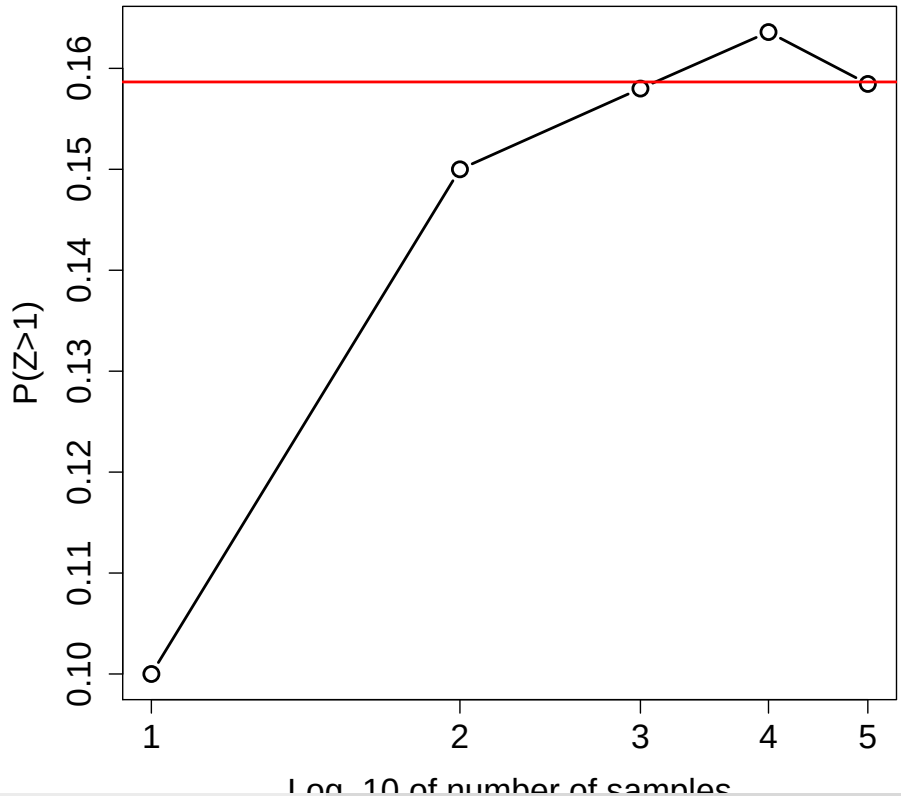


Suppose $Z \sim N(0,1)$ and we are trying to estimate $P(Z > 1) \approx 0.1586553$. If $Z_i \overset{ind}{\sim} N(0,1)$, then

$$P(Z > 1) \approx \frac{1}{J} \sum_{i=1}^{J} \mathrm{I}(Z_i > 1)$$

is a standard MC approach.

# Samples	$P(Z > 1)$
10	0.10
100	0.15
1000	0.158
10000	0.1636
100000	0.15846



Sequential MCMC

Suppose we want to sample from

$$p(\theta_{t+1}|y_{1:t+1}) = \int p(\theta_{0:t+1}|y_{1:t+1})d\theta_{0:t}.$$

An MCMC approach says to iterate through draws of full conditionals, e.g.

$$\theta_s \sim p(\theta_s|y_{1:t}, \theta_{-s}) = p(\theta_s|y_s, \theta_{s-1}, \theta_{s+1})$$

where θ_{-s} indicates $\theta_{0:t}$ with the s component removed and $s = 0, 1, 2, \ldots, t$.

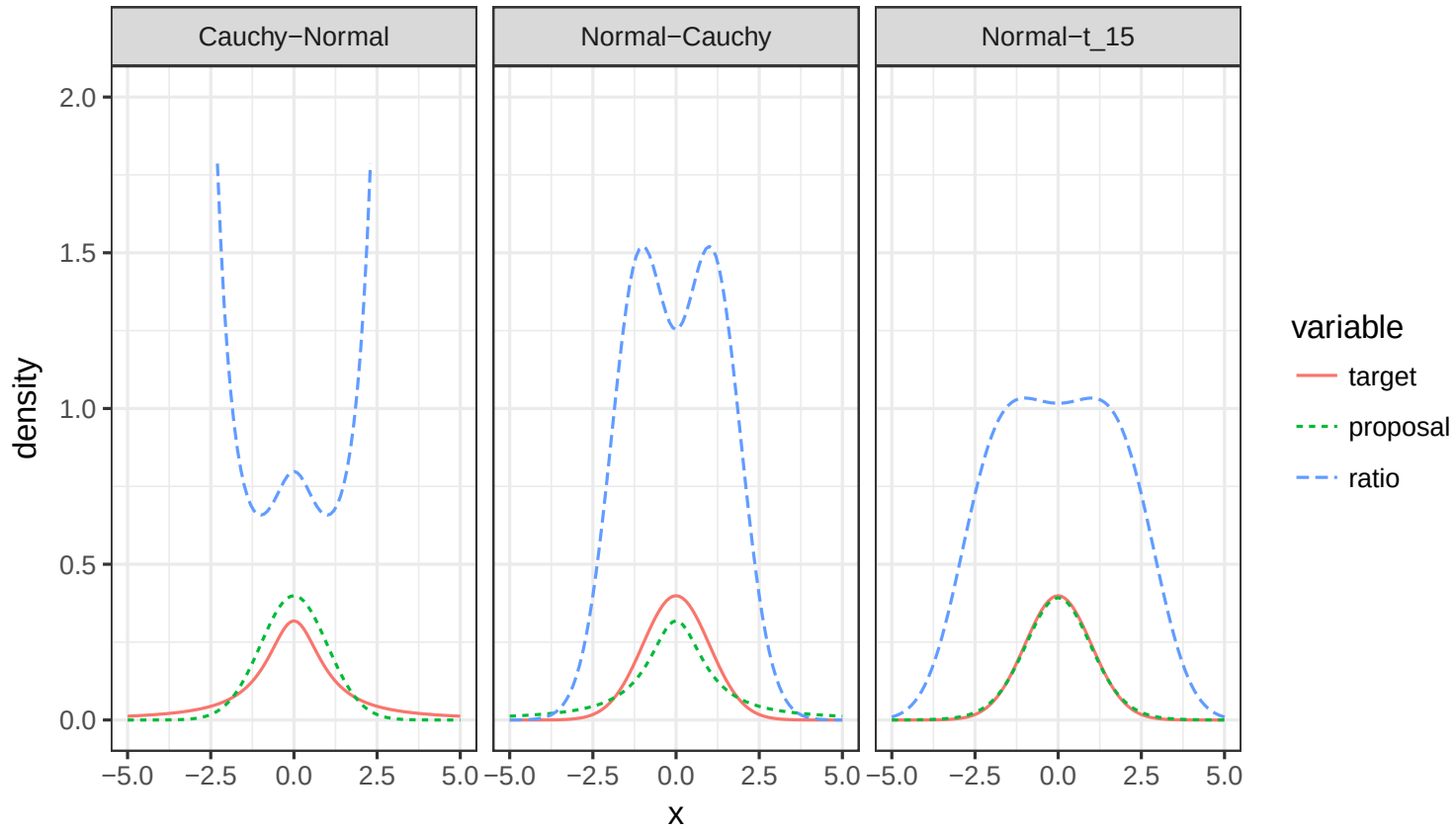
Now, you just obtained y_{t+1}. You need to redo the analysis, e.g.

$$\theta_s \sim p(\theta_s|y_s, \theta_{s-1}, \theta_{s+1})$$

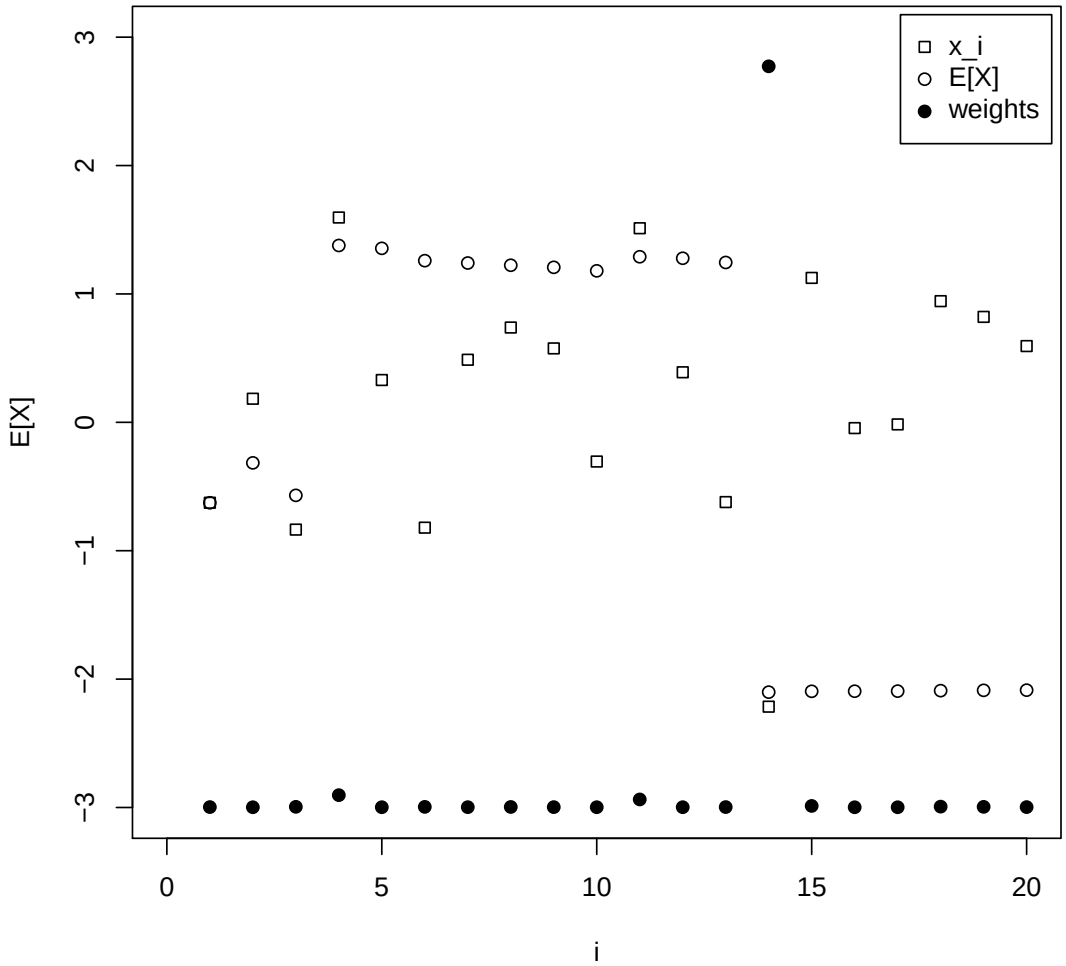
for $s = 0, 1, 2, \ldots, t, t+1$.

Importance sampling

Suppose we want to approximate some density $f(\theta)$, e.g. $p(\theta_t|y_{1:t})$, but we cannot simulate from $f(\theta)$. Draw $\theta_i \overset{ind}{\sim} g(\theta)$ and give each draw a weight $w_i = \frac{f(\theta_i)}{g(\theta_i)}$.



Suppose we are trying to estimate $E[\theta]$ when $\theta \sim t_2$. We draw samples from $\theta_i \sim N(0, 0.5^2)$ and give a weight $w_i = \frac{t_2(\theta_i)}{N(\theta_i;0,0.5^2)}$ to each sample.



Suppose $Z \sim N(0,1)$ and we are trying to estimate $P(Z > 4.5) \approx 3.398 \times 10^{-6}$. If $Z_i \sim N(0,1)$, then

$$P(Z > 4.5) \approx \frac{1}{J} \sum_{i=1}^{J} \mathrm{I}(Z_i > 4.5)$$

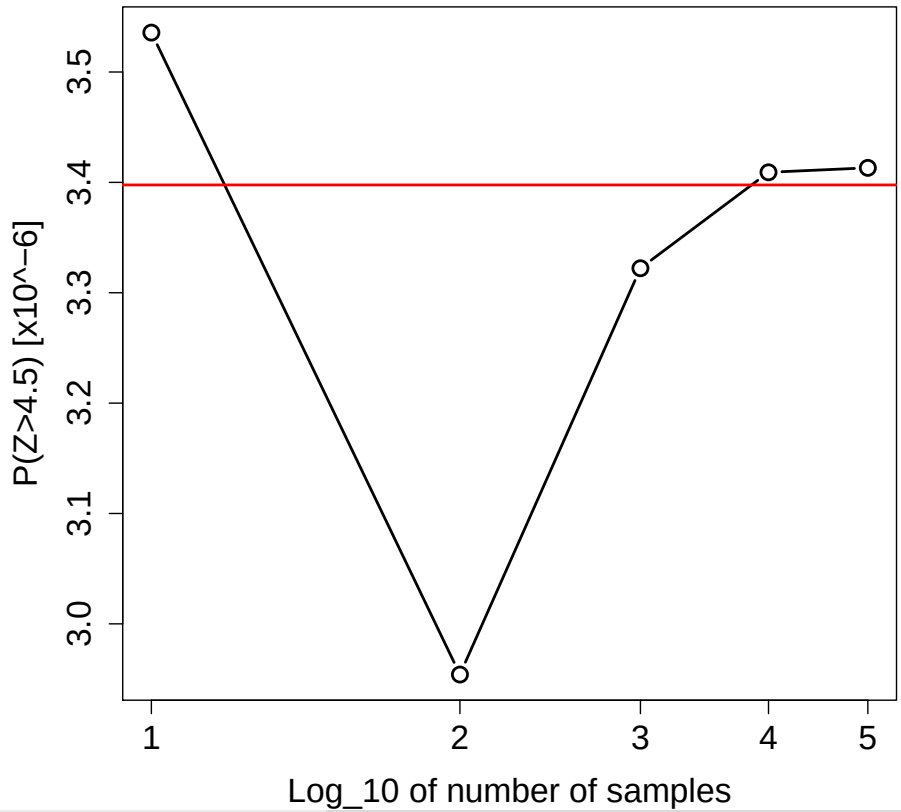
is a standard MC approach. If J=100,000 usually the indicator function is all zeros.

Suppose $Z \sim N(0,1)$ and we are trying to estimate $P(Z > 4.5) \approx 3.398 \times 10^{-6}$. If $\theta_i \sim Exp(1) + 4.5$, then

$$P(Z > 4.5) \approx \frac{1}{J} \sum_{i=1}^{J} \frac{N(\theta_i; 0, 1)}{Exp(\theta_i - 4.5; 1)} \mathrm{I}(\theta_i > 4.5)$$

is an importance sampling approach.

# Samples	$P(Z > 4.5)[\times 10^{-6}]$
10	3.53
100	2.95
1000	3.32
10000	3.41
100000	3.41



Importance sampling summary

Importance sampling summary:

- Importance sampling can be vastly superior to Monte Carlo sampling.
- When we are trying to estimate an entire density, we want the
  - tails of our proposal density to be heavier than our target density and
  - the proposal density to be as close to the target density as possible.

Suppose we have a general state-space model

$$p(y_t|\theta_t)$$
$$p(\theta_t|\theta_{t-1})$$

and a current filtered distribution $p(\theta_{t-1}|y_{1:t-1})$. Our goal is to approximate $p(\theta_t|y_{1:t})$. Let

$$\begin{aligned}
f(\theta_t) &= p(\theta_t|y_{1:t}) = \frac{p(y_t|\theta_t)p(\theta_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} \propto p(y_t|\theta_t)p(\theta_t|y_{1:t-1}) \\
g(\theta_t) &= p(\theta_t|y_{1:t-1}) = \int p(\theta_t|\theta_{t-1})p(\theta_{t-1}|y_{1:t-1})d\theta_{t-1} \\
\frac{f(\theta_t)}{g(\theta_t)} &\propto \frac{p(y_t|\theta_t)p(\theta_t|y_{1:t-1})}{p(\theta_t|y_{1:t-1})} = p(y_t|\theta_t) \\
\theta_{t-1}^{(i)} &\sim p(\theta_{t-1}|y_{1:t-1}) \\
\theta_t^{(i)} &\sim p(\theta_t|\theta_{t-1}^{(i)}) \\
w_t^{(i)} &\propto p(y_t|\theta_t^{(i)}) \\
p(\theta_t|y_{1:t}) &\approx \sum_{i=1}^{J} w_t^{(i)} \delta_{\theta_t^{(i)}}
\end{aligned}$$

The pair $\left(w_t^{(i)}, \theta_t^{(i)}\right)$ is called a particle.

Sequential importance sampling

Sequential importance sampling procedure:

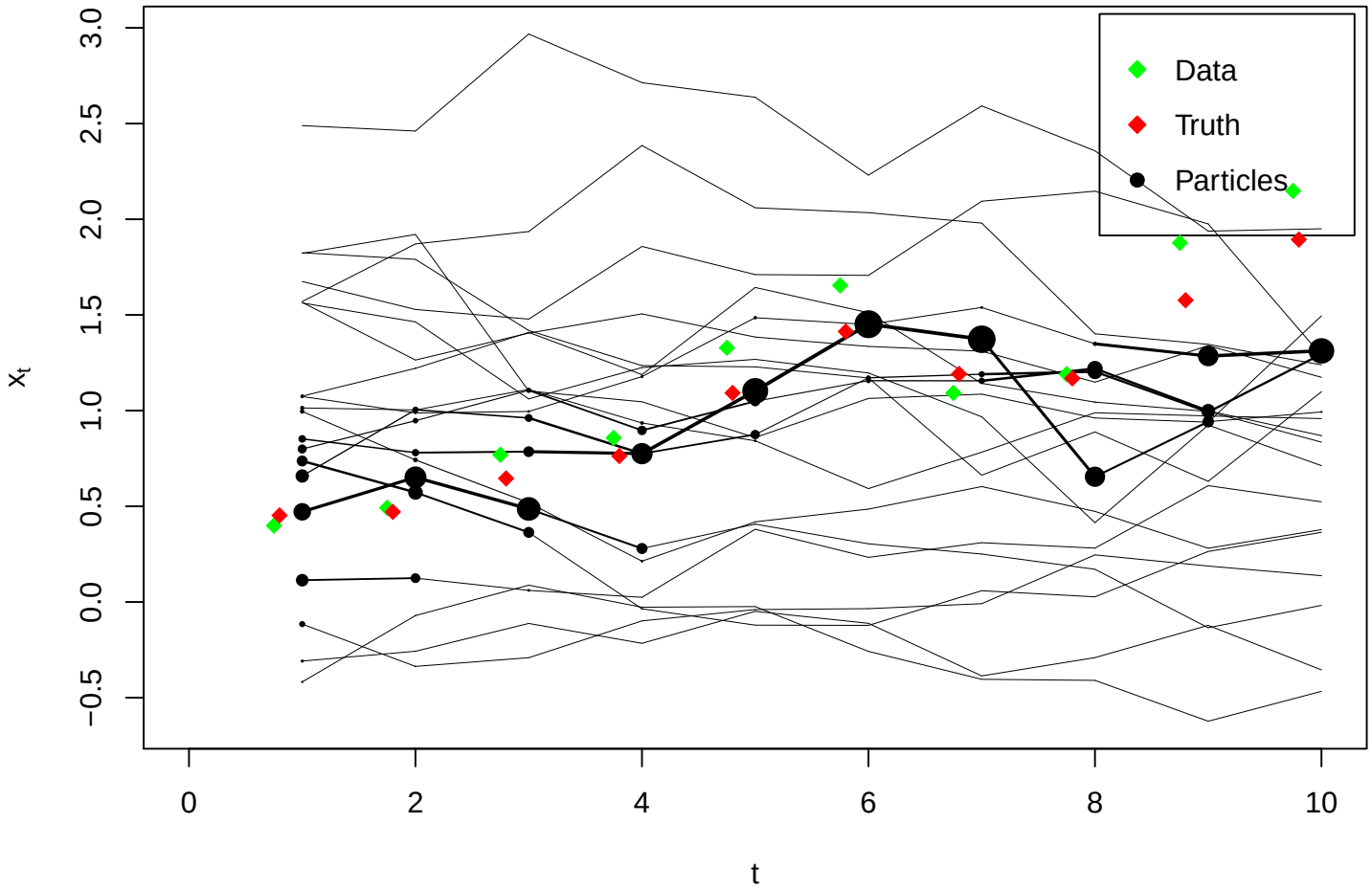
1. Suppose we have a particle approximation to our density at time $t-1$, i.e.

$$p(\theta_{t-1}|y_{1:t-1}) \approx \sum_{i=1}^{J} w_{t-1}^{(i)} \delta_{\theta_{t-1}^{(i)}}.$$

2. For $i \in \{1, 2, \ldots, J\}$
   a. Sample $\theta_t^{(i)} \sim p(\theta_t|\theta_{t-1}^{(i)})$
   b. Set $w_t^{(i)} \propto w_{t-1}^{(i)} p(y_t|\theta_t^{(i)})$
3. We now have a particle approximation to our density at time t, i.e.

$$p(\theta_t|y_{1:t}) \approx \sum_{i=1}^{J} w_t^{(i)} \delta_{\theta_t^{(i)}}.$$

Sequential importance sampling (SIS)



SIS summary

Sequential importance sampling (SIS) summary:

- Positives
  - As the number of particles J increases, the accuracy increases.
- Negatives
  - Inference is dominated by a few particles with high weight
  - Many particles are kept that are irrelevant

Why don't we eliminate particles with low weight in favor of particles with large weight?

Let's approximate $p(\theta_{t-1}|y_{1:t-1})$ by sampling with replacement proportional to the weights $w_{t-1}^{(i)}$:

$$p(\theta_{t-1}|y_{1:t-1}) \approx \left\{ \begin{array}{c|cccc} i & 1 & 2 & \cdots & J \\ \hline w_{t-1}^{(i)} & 0.02 & 0.05 & \cdots & 0.03 \\ \theta_{t-1}^{(i)} & 1.91 & 0.63 & \cdots & -0.12 \end{array} \right.$$

$$\Downarrow$$

$$p(\theta_{t-1}|y_{1:t-1}) \approx \left\{ \begin{array}{c|cccc} i & 1 & 2 & \cdots & J \\ \hline w_{t-1}^{(i)} & 1/J & 1/J & \cdots & 1/J \\ \theta_{t-1}^{(i)} & 0.63 & 0.63 & \cdots & -0.12 \end{array} \right.$$

(Gordon, Salmond, and Smith 1993)

Sequential importance sampling with resampling (SIR) procedure:

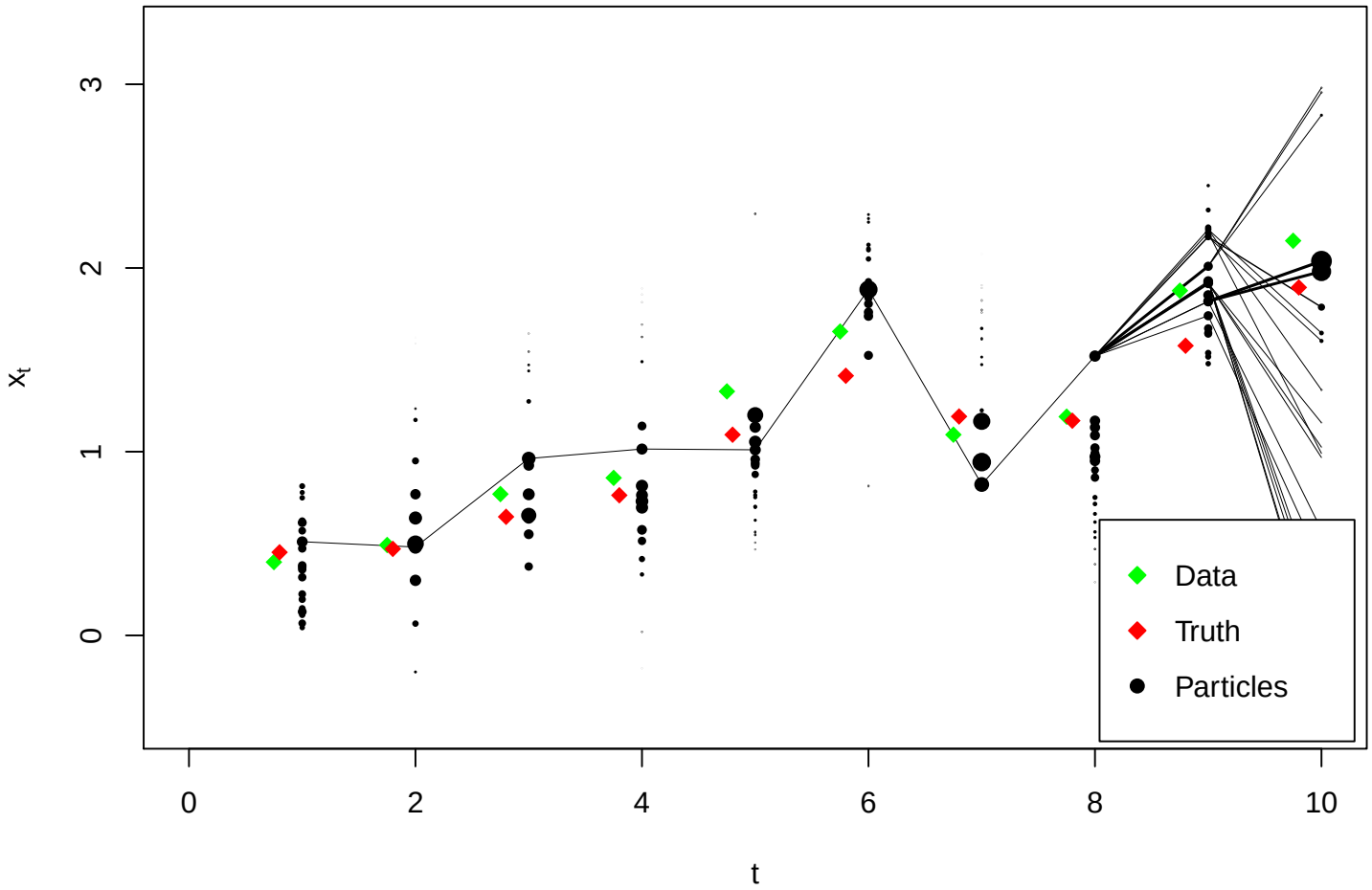
1. Suppose we have a particle approximation to our density at time $t-1$, i.e.

$$p(\theta_{t-1}|y_{1:t-1}) \approx \sum_{i=1}^{J} w_{t-1}^{(i)} \delta_{\theta_{t-1}^{(i)}}.$$

2. For $i \in \{1, 2, \ldots, J\}$
   a. Sample $j \in \{1, 2, \ldots, J\}$ with probability $w_{t-1}^{(j)}$
   b. Sample $\theta_t^{(i)} \sim p(\theta_t|\theta_{t-1}^{(j)})$
   c. Set $w_t^{(i)} \propto 1 p(y_t|\theta_t^{(i)})$
3. We now have a particle approximation to our density at time t, i.e.

$$p(\theta_t|y_{1:t}) \approx \sum_{i=1}^{J} w_t^{(i)} \delta_{\theta_t^{(i)}}.$$

Sequential importance sampling with resampling (SIR)



(Douc, Cappé, and Moulines 2005)

Constraints on resampling:

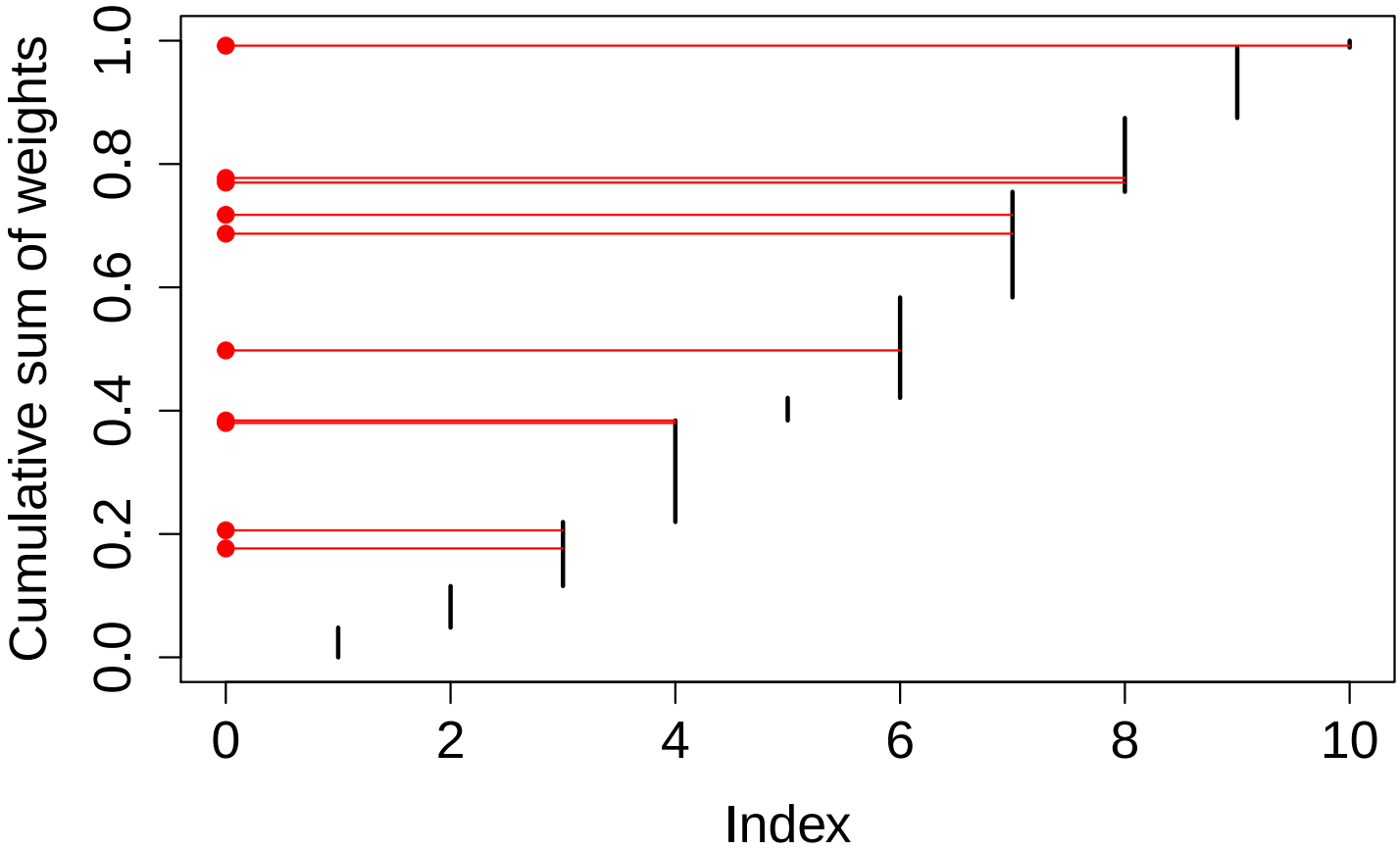
- Number of resulting particles (J) is fixed
- Resulting weights are uniform ($1/J$)
- Number of repeats is unbiased ($E[N_j] = Jw^{(j)}$)

Schemes that meet these requirements:

- Multinomial sampling
- Residual sampling
- Stratified sampling
- Systematic sampling

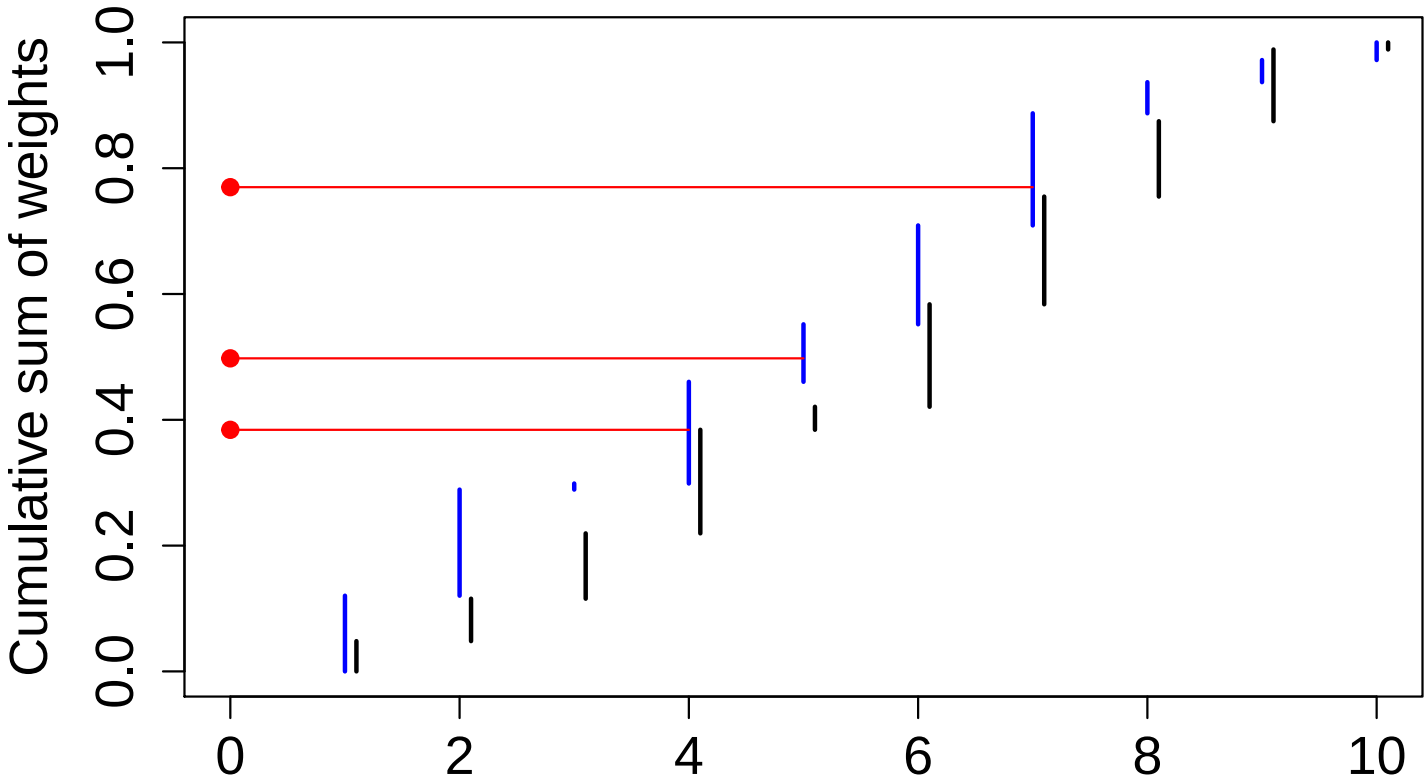
Multinomial sampling

1. Draw $U_1, \ldots, U_J \overset{ind}{\sim} Unif(0,1)$
2. Invert cumulative sum of weights



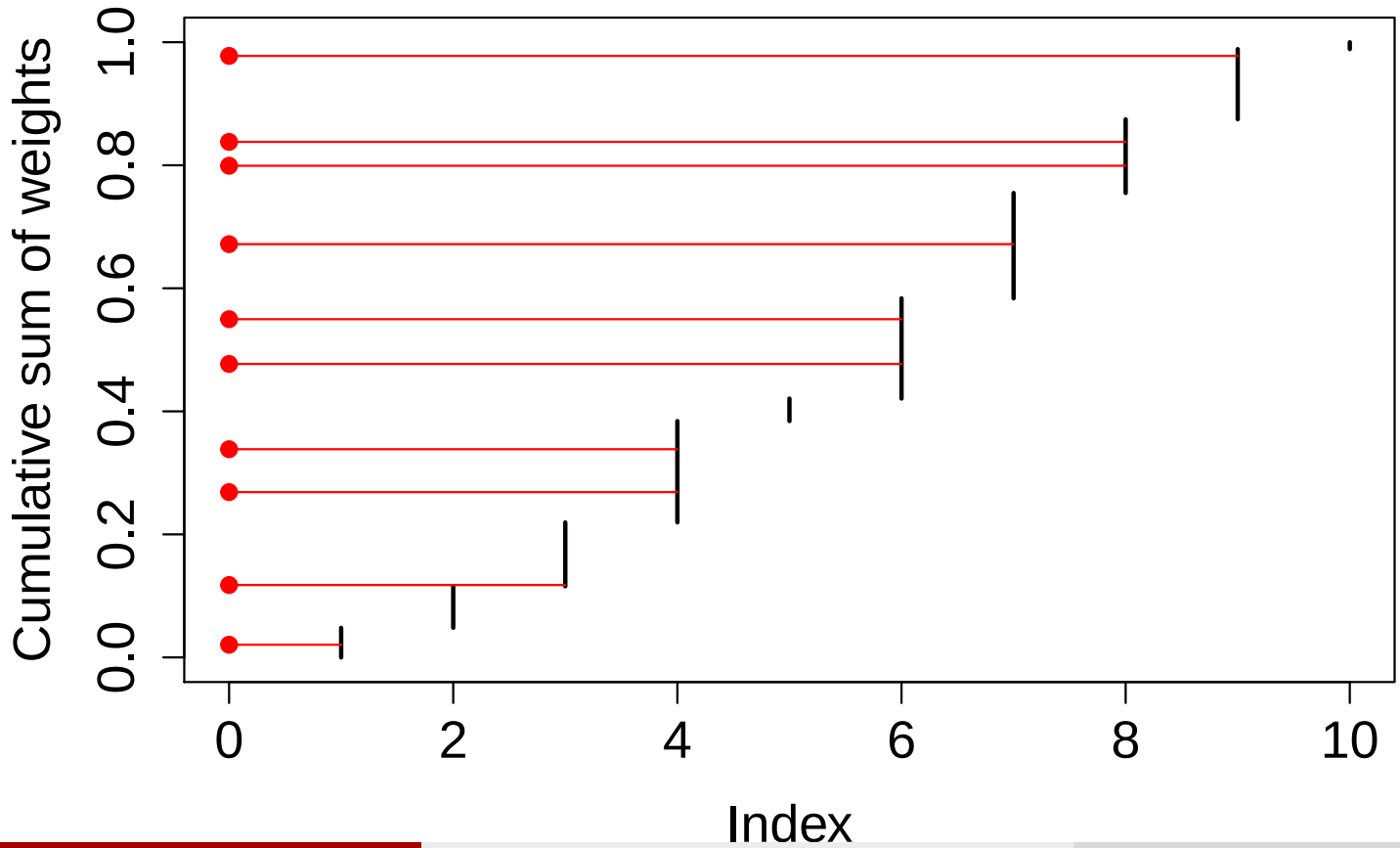
Residual sampling

1. Keep $n_j = \lfloor J w^{(j)} \rfloor$ repeats of particle j
2. Update remaining probability $w^{(j)'} \propto J w^{(j)} - n_j$
3. Multinomial sampling on remaining $J - \sum_{j=1}^{J} n_j$ particles with probabilities $w^{(j)'}$



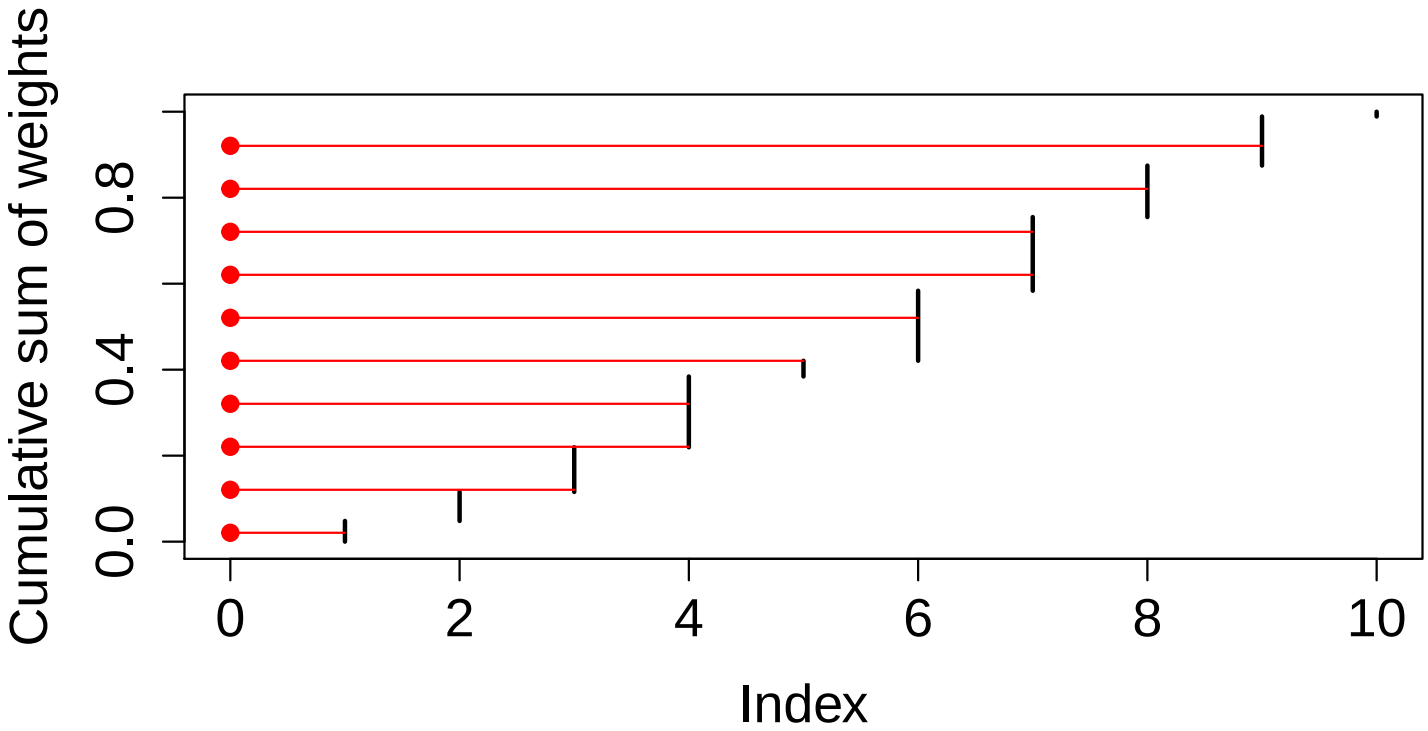
Stratified sampling

1. Draw $U_j \overset{ind}{\sim} Unif\left(\frac{j-1}{J}, \frac{j}{J}\right)$ for $j = 1, 2, \ldots, J$
2. Invert cumulative sum of weights



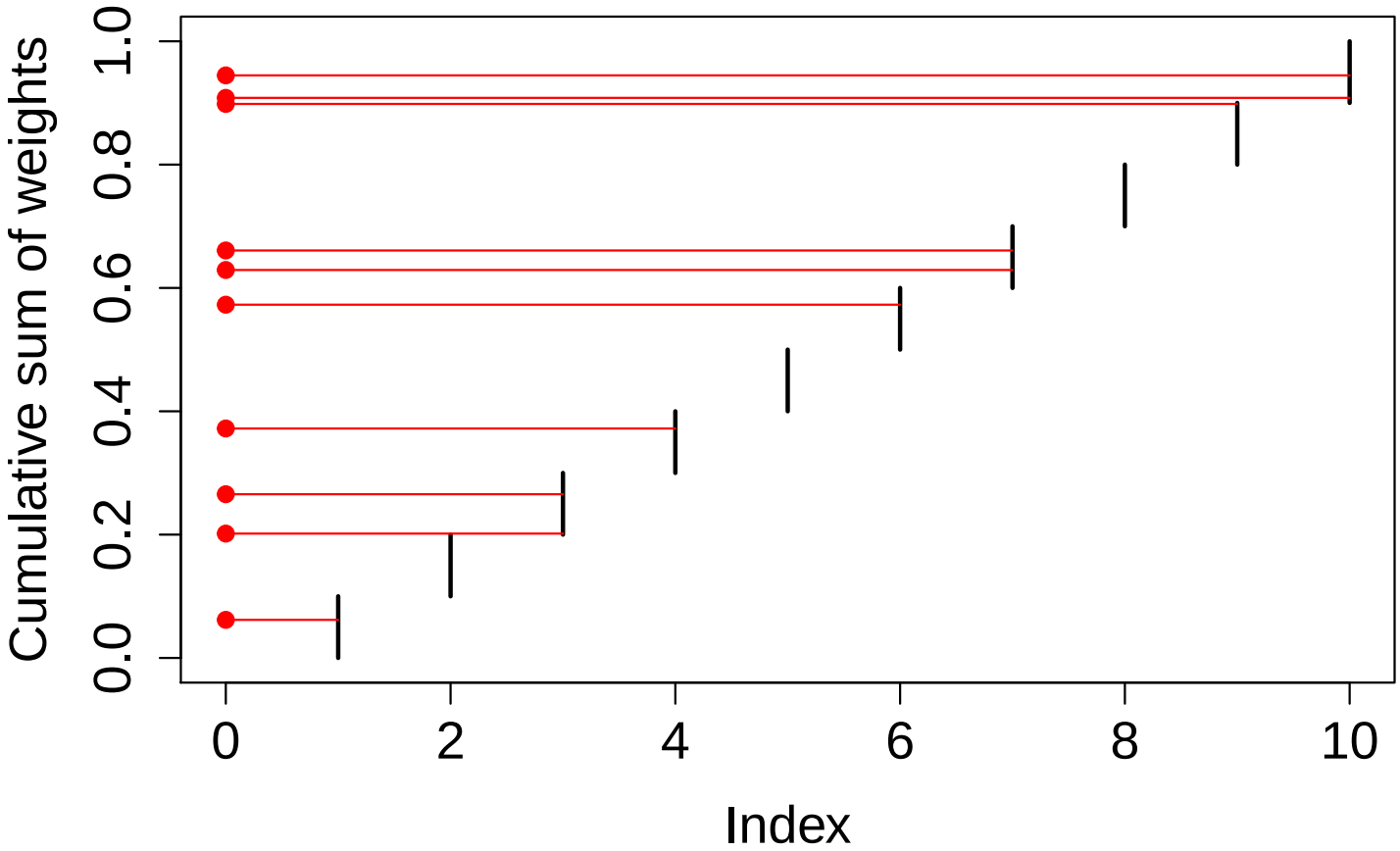
Systematic sampling

1. Draw $U_1 \sim Unif(0, 1/J)$
2. Set $U_j = U_{j-1} + \frac{1}{J}$ for $j = 1, 2, \ldots, J$
3. Invert cumulative sum of weights



A counter example shows that this can be worse than stratified and residual sampling.

Resampling adds variability



Methods for determining when to resample

- Effective sample size

$$ESS = \left(\sum_{i=1}^{J} (w^{(j)})^2 \right)^{-1}$$

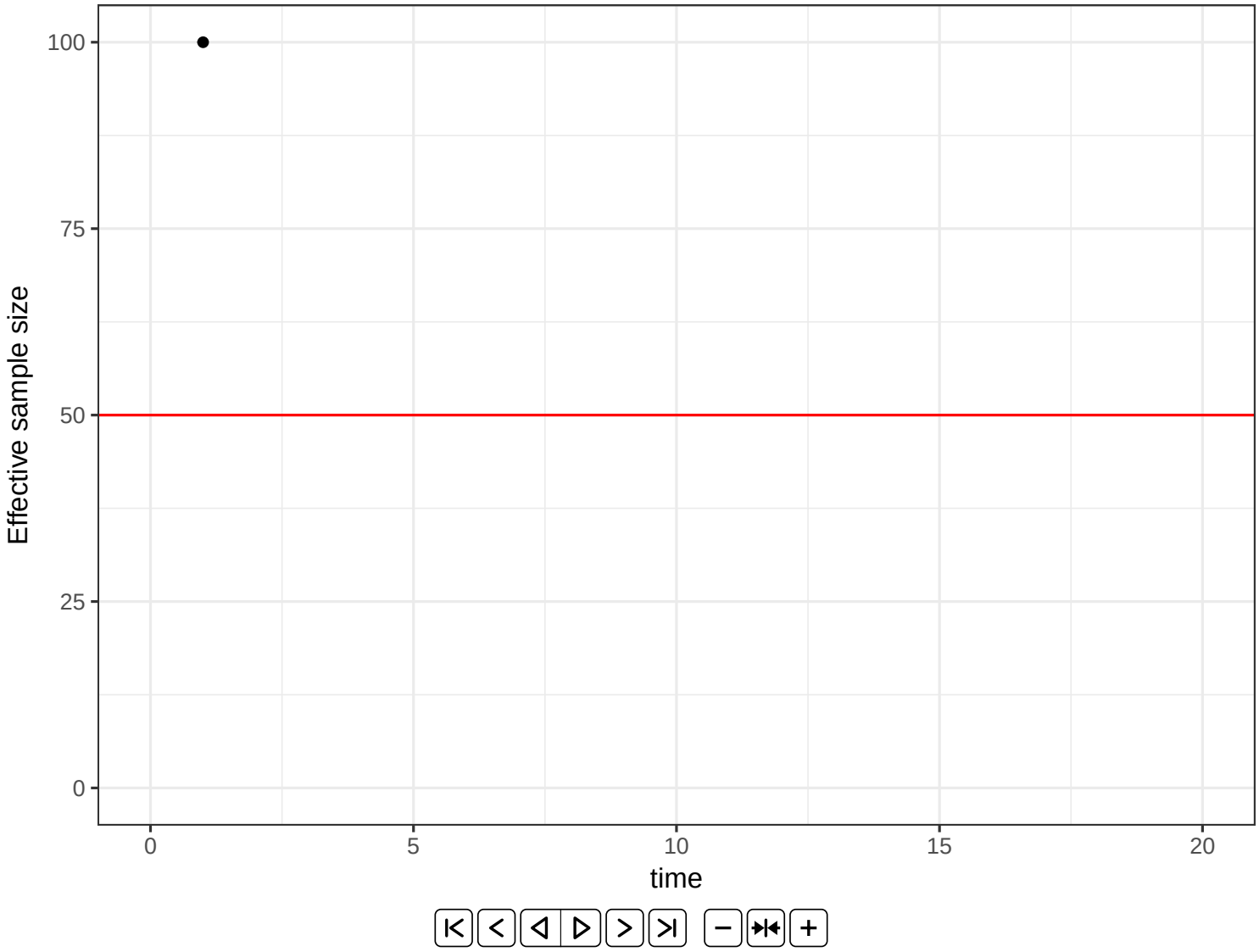
- Coefficient of variation

$$CoV = \left(\frac{1}{J} \sum_{i=1}^{J} (Jw^{(j)} - 1)^2 \right)^{1/2}$$

- Entropy

$$Ent = -\sum_{j=1}^{J} w^{(j)} \log_2(w^{(j)})$$

Dynamic resampling



Better than resampling would be to avoid the need altogether

- Resample-move (Gilks and Berzuini 2001)
- Auxiliary particle filter (Pitt and Shepherd 1999)

Resample-move procedure:

1. Suppose we have a particle approximation to our density at time $t-1$, i.e.
$$p(\theta_{t-1}|y_{1:t-1}) \approx \sum_{i=1}^{J} \frac{1}{J}\delta_{\theta_{t-1}^{(i)}}.$$
2. For $i \in \{1, 2, \ldots, J\}$
   a. Sample $j \in \{1, 2, \ldots, J\}$ with probability proportional to $p(y_t|\theta_{t-1}^{(j)})$
   b. Sample $\theta_t^{(i)} \sim p(\theta_t|\theta_{t-1}^{(j)})$
3. We now have a particle approximation to our density at time t, i.e.
$$p(\theta_t|y_{1:t}) \approx \sum_{i=1}^{J} \frac{1}{J}\delta_{\theta_t^{(i)}}.$$

Evaluating $p(y_t|\theta_{t-1})$ requires solving the integral

$$p(y_t|\theta_{t-1}) = \int p(y_t|\theta_t)p(\theta_t|\theta_{t-1})d\theta_t.$$

Auxiliary particle filter (APF) procedure:

1. Suppose we have a particle approximation to our density at time $t-1$, i.e.

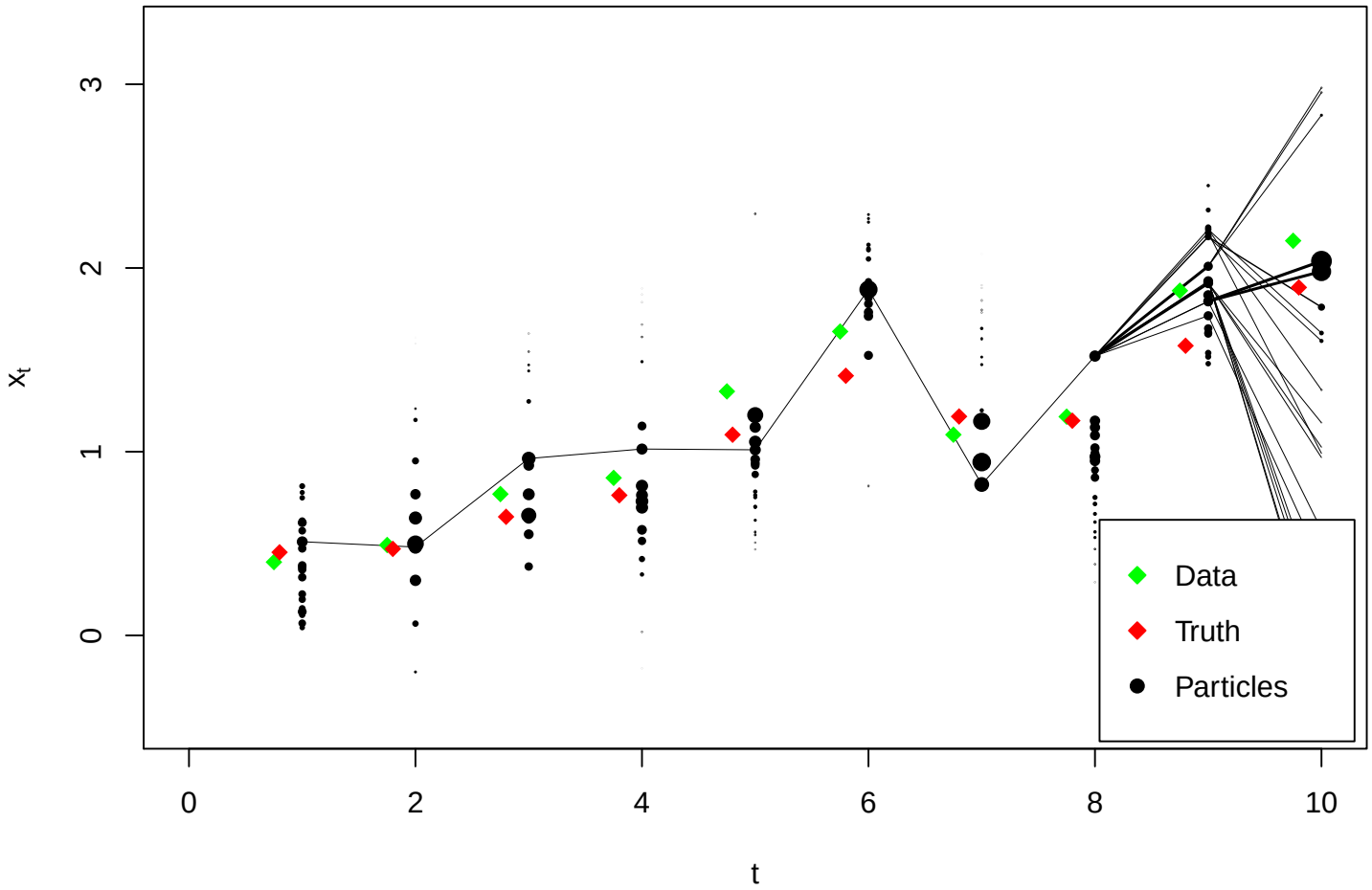
$$p(\theta_{t-1}|y_{1:t-1}) \approx \sum_{i=1}^{J} w_{t-1}^{(i)} \delta_{\theta_{t-1}^{(i)}}.$$

2. For $i \in \{1, 2, \ldots, J\}$
   a. Sample $j \in \{1, 2, \ldots, J\}$ with prob. proportional to $w_{t-1}^{(j)} p(y_t|\mu_t^{(j)})$
   b. Sample $\theta_t^{(i)} \sim p(\theta_t|\theta_{t-1}^{(j)})$
   c. Set $w_t^{(i)} = \frac{p(y_t|\theta_t^{(i)})}{p(y_t|\mu_t^{(j)})}$
3. We now have a particle approximation to our density at time t, i.e.

$$p(\theta_t|y_{1:t}) \approx \sum_{i=1}^{J} w_t^{(i)} \delta_{\theta_t^{(i)}}.$$

where $\mu_t^{(j)}$ is a point estimate of $\theta_t^{(j)}$, usually $\mu_t^{(j)} = E[\theta_t|\theta_{t-1}^{(j)}]$.

Auxiliary Particle Filter (APF)



Summary of $p(\theta_t|y_t, \psi)$:

- Avoid particle degeneracy if possible
  - Resample-move
  - Auxiliary particle filter
- When resampling
  - Use either stratified or residual
  - and only resample when necessary, e.g. ESS

Outline

1. State-space models $p(y|\theta,\psi)p(\theta|\psi)$
   - Definition
   - Terminology
   - Notation
2. State inference $p(\theta|y,\psi)$
   - Exact inference
   - Importance sampling
   - Sequential importance sampling
   - Bootstrap filter - resampling
   - Auxiliary particle filter
3. State and parameter inference $p(\theta,\psi|y)$
   - Bootstrap filter
   - Kernel density
   - Sufficient statistics
4. Advanced SMC
   - SMC-MCMC
   - Fixed parameter
   - SMC for marginal likelihood calculations

Unknown fixed parameters

What if the fixed parameters are unknown?

- incorporate into state with degenerate evolutions, e.g. $\psi_t = \psi_{t-1}$
- incorporate into state with evolutions, e.g. $\psi_t = \psi_{t-1} + \epsilon_t$
- use kernel density approximation to regenerate parameter values
- use sufficient statistics to regenerate parameter values
- use Markov chain Monte Carlo to regenerate parameter values

Unknown fixed parameters

Our goal has changed slightly

- We assume $p(\theta_{t-1}, \psi | y_{1:t-1})$ is known
- and try to obtain $p(\theta_t, \psi | y_{1:t})$ using
- $p(\theta_t | \theta_{t-1}, \psi)$ and $p(y_t | \theta_t, \psi)$.

Then, starting with $p(\theta_0, \psi | y_0) = p(\theta_0, \psi)$ we can find $p(\theta_t, \psi | y_{1:t})$ for all t.

Resampling with fixed parameters

Modify SIR to include fixed parameters:

1. Suppose we have a particle approximation to our density at time $t-1$, i.e.

$$p(\theta_{t-1}, \psi | y_{1:t-1}) \approx \sum_{i=1}^{J} w_{t-1}^{(i)} \delta_{(\theta_{t-1},\psi)^{(i)}}.$$

2. For $i \in \{1, 2, \ldots, J\}$
   a. Sample $j \in \{1, 2, \ldots, J\}$ with probability $w_{t-1}^{(j)}$.
   b. Set $\psi^{(i)} = \psi^{(j)}$.
   c. Sample $\theta_t^{(i)} \sim p(\theta_t | \theta_{t-1}^{(j)}, \psi^{(i)})$.
   d. Set $w_t^{(i)} \propto p(y_t | \theta_t^{(i)}, \psi^{(i)})$.
3. We now have a particle approximation to our density at time t, i.e.

$$p(\theta_t, \psi | y_{1:t}) \approx \sum_{i=1}^{J} w_t^{(i)} \delta_{(\theta_t,\psi)^{(i)}}.$$

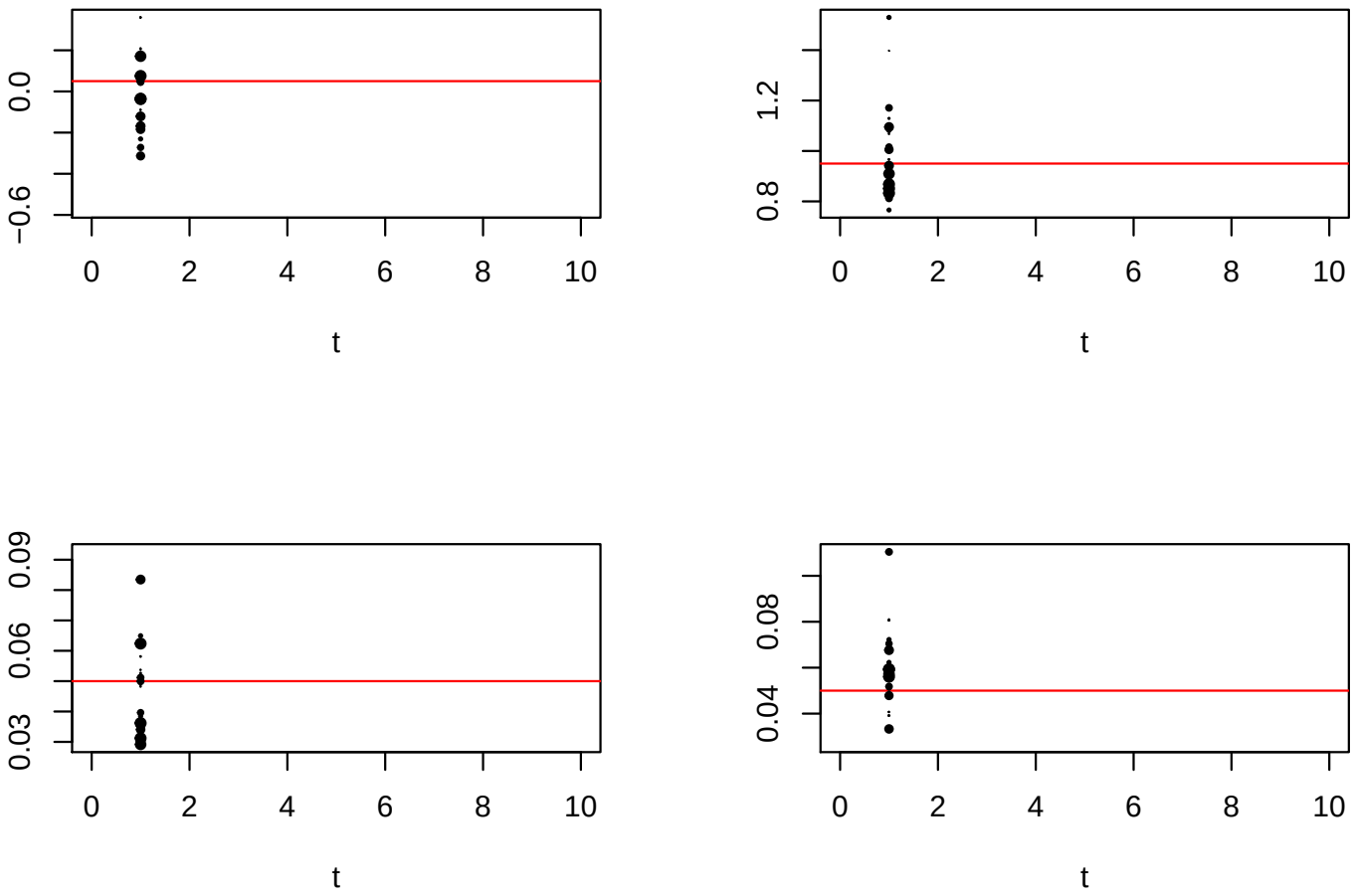
Reampling with fixed parameters

Test model:

$$
\begin{aligned}
y_t &= \theta_t + v_t \\
\theta_t &= \alpha + \beta\theta_{t-1} + w_t \\
v_t &\overset{ind}{\sim} N(0, V) \\
w_t &\overset{ind}{\sim} N(0, W) \\
\theta_0 &\sim N(m_0, C_0)
\end{aligned}
$$

$$
\psi = (\alpha, \beta, V, W)
$$

SIR with fixed parameters



Clearly we need to regenerate parameter values. One idea:

- Approximate our particle approximation by a kernel density approximation
- Draw new parameter values from this kernel density approximation

$$\begin{aligned} p(\psi|y_{1:t-1}) &\approx \sum_{i=1}^{J} w_{t-1}^{(i)} \delta_{\psi^{(i)}} \\ &\approx \sum_{i=1}^{J} w_{t-1}^{(i)} N(a\psi^{(i)} + (1-a)\overline{\psi}, h^2 V) \end{aligned}$$

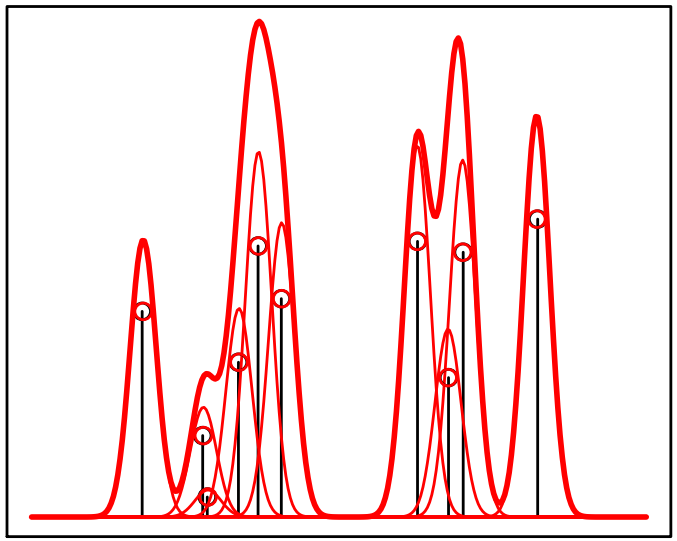
where

$$h^2 = 1 - a^2 = 1 - \left(\frac{3\tau - 1}{2\tau}\right)^2$$

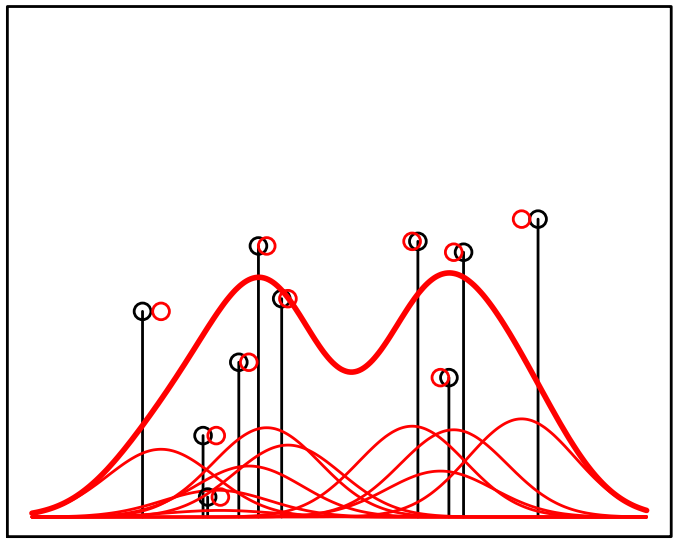
and $\overline{\psi}$ and V are the Monte Carlo estimate of the mean and covariance.

Kernel density approximation

0.99



0.85



Sufficient statistics

Another idea:

- Rather than storing draws of parameters, store sufficient statistics for the parameters.

- Suppose the model admits a sufficient statistic representation, i.e.

$$p(\theta_t, s_t, \psi | y_{1:t}) = p(\psi | s_t) p(\theta_t, s_t | y_{1:t}).$$

- Then, each particle stores a distribution for the parameters
- and the sufficient statistics can be updated deterministically via

$$s_t = \mathcal{S}(s_{t-1}, \theta_t, \theta_{t-1}, y_t).$$

For example, in our model

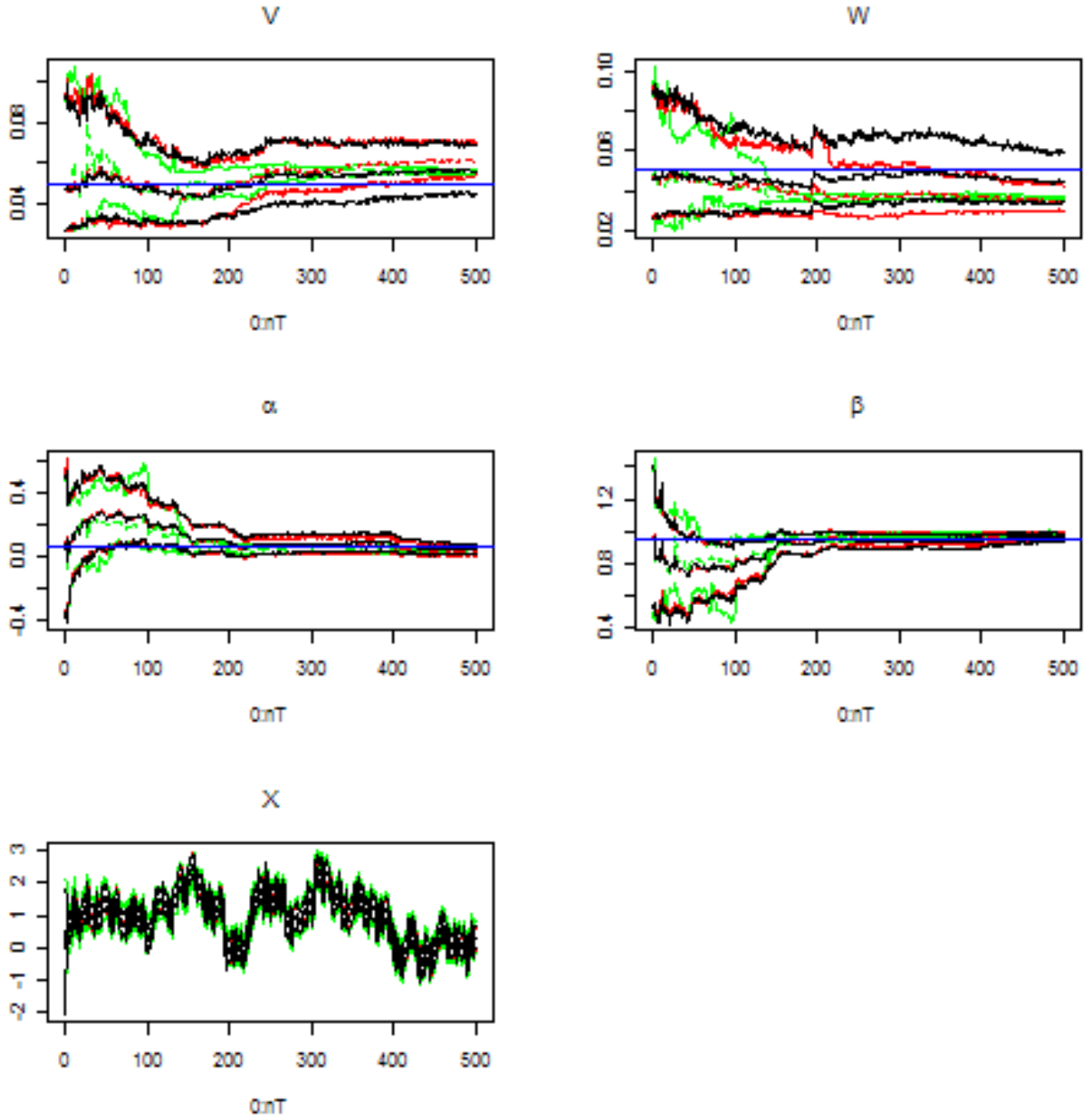
$$\begin{aligned} y_t &= \theta_t + v_t & v_t &\sim N(0, V) \\ \theta_t &= \alpha + \beta\theta_{t-1} + w_t & w_t &\sim N(0, W) \end{aligned}$$

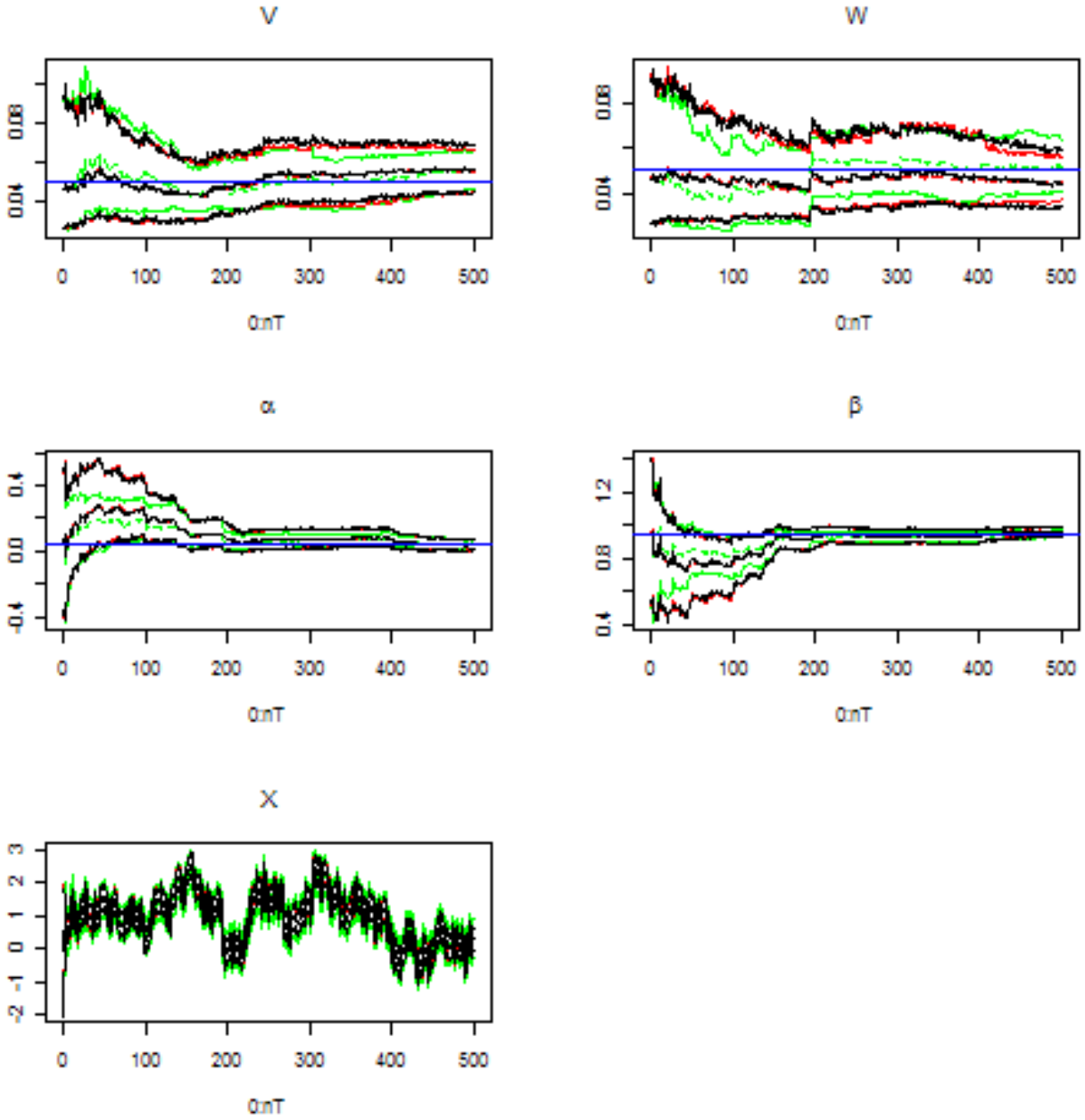
if

$$\begin{aligned} V|y_{1:t-1} &\sim IG(a_{V,t-1}, b_{V,t-1}) \\ \alpha, \beta, W|y_{1:t-1} &\sim N - IG(m_{t-1}, S_{t-1}, a_{W,t-1}, b_{W,t-1}) \\ s_{t-1} &= (a_{V,t-1}, b_{V,t-1}, m_{t-1}, S_{t-1}, a_{W,t-1}, b_{W,t-1}) \end{aligned}$$

then

$$\begin{aligned} V|y_{1:t} &\sim IG(a_{V,t}, b_{V,t}) \\ \alpha, \beta, W|y_{1:t}, \theta_t, \theta_{t-1} &\sim N - IG(m_t, S_t, a_{W,t}, b_{W,t}) \\ s_t &= (a_{V,t}, b_{V,t}, m_t, S_t, a_{W,t}, b_{W,t}) \end{aligned}$$


V
0.04
0.06
0
100
200
300
400
500
0:nT
W
0.02
0.06
0.10
0
100
200
300
400
500
0:nT
α
-0.4
0.0
0.4
0
100
200
300
400
500
0:nT
β
0.4
0.8
1.2
0
100
200
300
400
500
0:nT
X
-2
-1
0
1
2
3
0
100
200
300
400
500
0:nT


V
0.08
0.04
0
100
200
300
400
500
0:nT
W
0.08
0.04
0
100
200
300
400
500
0:nT
α
0.4
0.0
-0.4
0
100
200
300
400
500
0:nT
β
12
08
0.4
0
100
200
300
400
500
0:nT
X
3
2
1
0
-1
-2
0
100
200
300
400
500
0:nT

MCMC within SMC

Final idea for fixed parameter regeneration:

- Stop the SMC algorithm
- Perform one (or more) iterations of MCMC on each particle
- Pros
  - Does not affect SMC theory
  - Will move fixed parameters around
- Cons
  - Requires entire data and state history
  - If not, introduces bias

Summary

Summary of $p(\theta_t, \psi | y_t)$:

- Regenerating fixed parameters is necessary
  - When possible use sufficient statistics
  - Otherwise use kernel density or MCMC step

Outline

1. State-space models $p(y|\theta,\psi)p(\theta|\psi)$
   - Definition
   - Terminology
   - Notation
2. State inference $p(\theta|y,\psi)$
   - Exact inference
   - Importance sampling
   - Sequential importance sampling
   - Bootstrap filter - resampling
   - Auxiliary particle filter
3. State and parameter inference $p(\theta,\psi|y)$
   - Bootstrap filter
   - Kernel density
   - Sufficient statistics
4. Advanced SMC
   - SMC for marginal likelihood calculations
   - Theoretical results for fixed parameters
   - SMC to generate Metropolis proposals

Marginal likelihood calculations

In Bayesian data analysis, model comparison and hypothesis testing involves the marginal likelihood:

$$p(y) = \int p(y|\psi)p(\psi)d\psi.$$

For example, Bayes' factors depend on the marginal likelihood of for both models:

$$BF(0:1) = \frac{p(y|M_0)}{p(y|M_1)} = \frac{\int p(y|\psi_0)p(\psi_0|M_0)d\psi_0}{\int p(y|\psi_1)p(\psi_1|M_1)d\psi_1}.$$

The marginal likelihood can be decomposed as

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1})$$

This can be approximated using SMC methods by

$$p(y_t|y_{1:t-1}) \approx \frac{1}{J}\sum_{i=1}^{J} \tilde{w}_t^{(i)}$$

where $\tilde{w}_t^{(i)}$ are the normalized weights

Theoretical results

For $p(\theta_{1:t}|y_{1:t},\psi)$, we have

$$E\left[\left|\int h(\theta_{T:T-L})[\hat{p}(\theta_{T:T-L}|y_{1:T})-p(\theta_{T:T-L}|y_{1:T})]d\theta_{T:T-L}\right|^p\right]^{1/p}\leq\frac{c(L)}{\sqrt{J}}.$$

As long as we are only interested in a fixed time into the past, we can use a constant number of particles and maintain accuracy.

While for $p(\theta_{1:t},\psi|y_{1:t})$, we have

$$E\left[\left|\int h(\psi)[\hat{p}(\psi|y_{1:T})-p(\psi|y_{1:T})]d\psi\right|^p\right]^{1/p}\leq\frac{c(T)}{\sqrt{J}}.$$

So, as time increases, we need more and more particles to maintain accuracy.

MCMC within SMC

How about using SMC to generate proposal draws for the states within a Metropolis sampling scheme?

Particle Markov chain Monte Carlo methods

Christophe Andrieu,

University of Bristol, UK

Arnaud Doucet

Institute of Statistical Mathematics, Tokyo, Japan, and University of British Columbia, Vancouver, Canada

and Roman Holenstein

University of British Columbia, Vancouver, Canada

Summary

- If your goal is $p(\theta_t|y_{1:t}, \psi)$, then
  - Analytically integrate anything you can
  - Use a point estimate to reduce particle degeneracy
  - Use stratified or residual resampling
- If your goal is $p(\theta_t, \psi|y_{1:t})$, then
  - Analytically integrate anything you can
  - Use kernel density approximation to regenerate particles
  - Use lots of particles
  - Stop every once in a while and run MCMC
- Lots of areas for open research
  - Marginal likelihood calculation
  - Theoretical results
  - SMC for Metropolis proposals

References

- Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993), Novel approach to nonlinear/non-Gaussian Bayesian state estimation, IEEE Proceedings Part F: Communications, Radar and Signal Processing, 140, 107113.
- Pitt, M. K. and Shephard, N. (1999), Filtering via simulation: auxiliary particle filters, Journal of the American Statistical Association, 94, 590599.
- Liu, J. and West, M. (2001), Combined parameter and state estimation in simulation-based filtering, in Sequential Monte Carlo Methods in Practice, eds. A. Doucet, J. F. G. De Freitas, and N. J. Gordon, pp. 197217, Springer-Verlag, New York.
- Doucet, A., De Freitas, N., and Gordon, N. (2001), Sequential Monte Carlo Methods in Practice, Springer-Verlag, New York.
- Gilks, W. R. and Berzuini, C. (2001) "Following a Moving Target-Monte Carlo Inference for Dynamic Bayesian Models," Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 63, No. 1, 127-146
- Fearnhead, P. (2002), Markov chain Monte Carlo, sufficient statistics, and particle filters, Journal of Computational and Graphical Statistics, 11, 848862.
- Storvik, G. (2002), Particle filters in state space models with the presence of unknown static parameters, IEEE Transactions on Signal Processing, 50, 281289.
- R. Douc, O. Capp, and E. Moulines, (2005) "Comparison of Resampling Schemes for Particle Filtering," In 4th International Symposium on Image and Signal Processing and Analysis (ISPA), Zagreb

Computation

- R::KFAS
- R::dlm
- R::pomp
- R::SMC (not updated sinced 2011-12-11)
- R::smcUtils
- LiBbi `libbi.org/` (RBi `https://github.com/pierrejacob/RBi`)