Advanced Statistical Computing

Fall 2016

Steve Qin

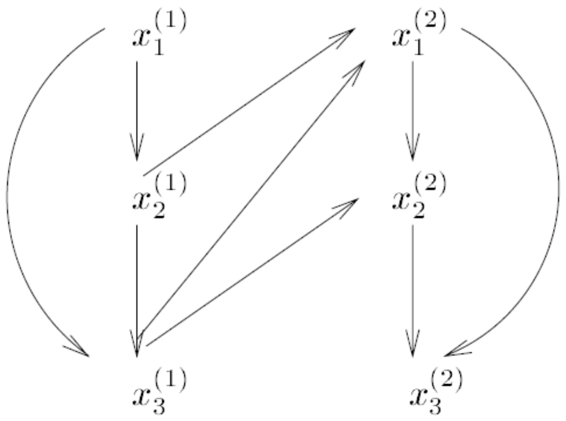
Outline

- Collapsing, predictive updating
- Sequential Monte Carlo

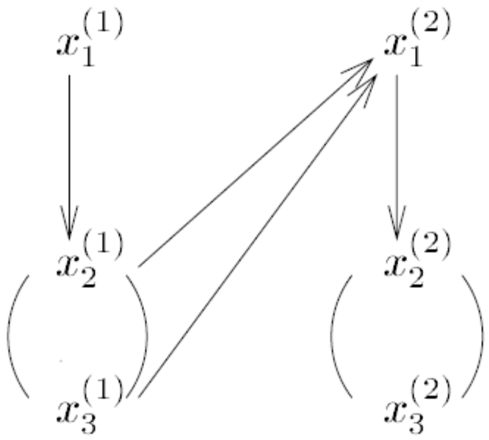
Collapsing and grouping

- Want to sample from $\mathbf{x} = (x_1, x_2, \ldots, x_d)$
- Regular Gibbs sampler:
  - Sample $x_1^{(t+1)}$ from $\pi(x_1^{(t+1)} \mid x_2^{(t)}, x_3^{(t)}, \ldots, x_d^{(t)})$,
  - Sample $x_2^{(t+1)}$ from $\pi(x_2^{(t+1)} \mid x_1^{(t)}, x_3^{(t)}, \ldots, x_d^{(t)})$,
  - …
  - Sample $x_d^{(t+1)}$ from $\pi(x_d^{(t+1)} \mid x_2^{(t)}, x_3^{(t)}, \ldots, x_{d-1}^{(t)})$,
- Alternatively:
  - Grouping: $\mathbf{x}_{d-1}' = (x_{d-1}, x_d)$.
  - Collapsing, i.e., integrate out x_d: $\mathbf{x}^{-} = (x_1, x_2, \ldots, x_{d-1})$

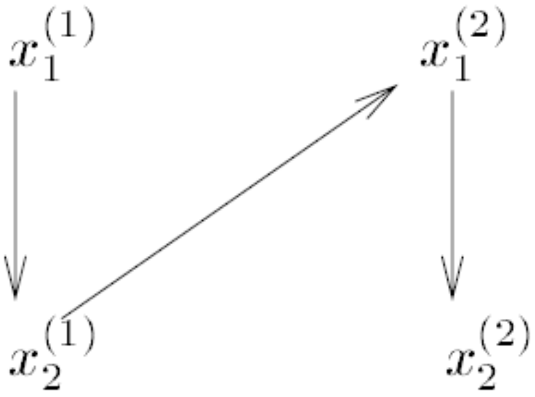
The three-schemes



standard



grouping



collapsing

Some theory

- Hilbert space $L_2(\pi)$ of functions $h()$.
- Define $\langle h, g \rangle = E_\pi \{h(x)g(x)\}$, thus $\|h\| = \text{var}_\pi(h)$.
- Define forward operator ***F*** as

$$Fh(x) = \int K(x,y)h(y)dy = E_\pi \left\{h\left(x^{(t+1)}\right) | \, x^{(t)} = x\right\}.$$

$$\|F\| = \sup_h \|Fh(x)\| \text{ for all fucntions with } E(h^2) = 1.$$

- The convergence of Markov chains is tied to the norms of the corresponding forward operators.

Three-scheme theorem

- Standard $\boldsymbol{F_s}$: $x_1 \to x_2 \to \cdots \to x_d$;
- Grouping $\boldsymbol{F_g}$: $x_1 \to x_2 \to \cdots \to \{x_{d-1}, x_d\}$;
- Collapsing $\boldsymbol{F_c}$: $x_1 \to x_2 \to \cdots \to x_{d-1}$.

Theorem The norms of the three forward operators are ordered as

$$\|F_c\| \leq \|F_g\| \leq \|F_s\|$$

Examples

- Murray’s data
- Bivariate Gaussian with mean 0 and unknown covariance matrix *Σ*

1	1	−1	−1	2	2	−2	−2	*	*	*	*
1	−1	1	−1	*	*	*	*	2	2	−2	−2

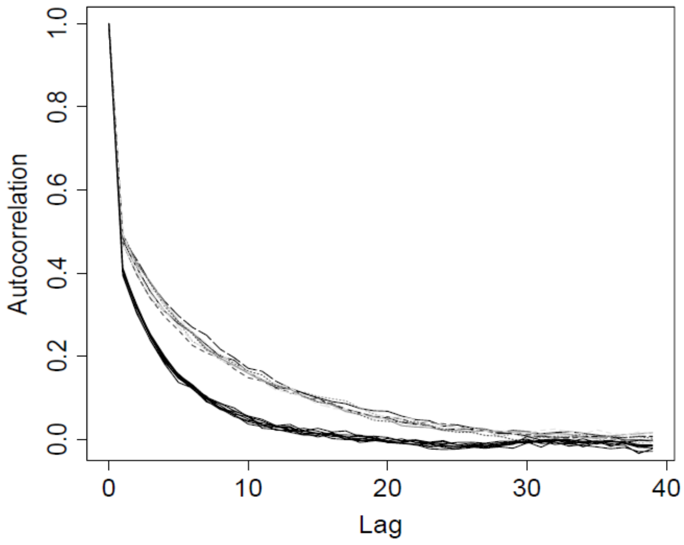
standard

$$\Sigma \mid y_{obs}, y_{mis},$$

$$y_{mis} \mid y_{obs}, \Sigma.$$

collapsing

$$y_{mis,i} \mid y_{obs}, y_{mis,[-i]}.$$



Remarks

- Avoid introducing unnecessary parameters into a Gibbs sampler,
- Do as much analytical work as possible,
- However, introducing some clever auxiliary variables can greatly improve computation efficiency.

Sequential Monte Carlo

- We wish to evaluate an integral

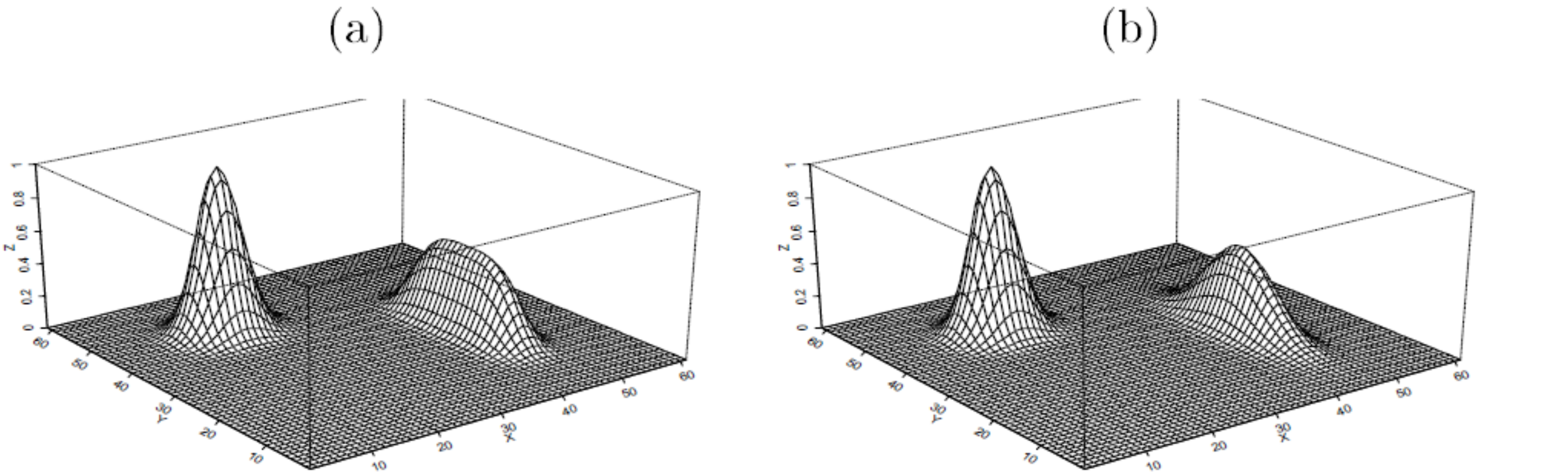
$$\theta = \int_{\aleph} h(x)\pi(x)dx = E_{\pi}[h(X)].$$

assume $h(x) \geq 0$.

- Riemann sum (on grid points) as approximation.
- Alternatively, use Monte Carlo. Select random samples uniformly on its support.

An example

$$f(x,y) = 0.5e^{-90(x-0.5)^2-45(y+0.1)^4} + e^{-45(x+0.4)^2-60(y-0.5)^2}$$



- Both grid-point method and vanilla Monte Carlo methods wasted resources on “boring” desert area.

The basic idea

- Marshall (1956) suggested that one should focus on the region(s) of "importance" so as to save computational resources—*importance sampling*.
- Essential in high-dimensional models.

The algorithm

- To evaluate $\mu = E_\pi[h(X)] = \int_{\aleph} h(x)\pi(x)dx.$
  - Draw $x^{(1)},\ldots,x^{(m)}$ from a trial distribution $g()$.
  - Calculate the *importance weight*

    $$w^{(j)} = \pi(x^{(j)}) / g(x^{(j)}), \quad \text{for } j = 1,\ldots,m.$$

  - Approximate μ by $\hat{\mu} = \dfrac{w^{(1)}h(x^{(1)}) + \cdots + w^{(m)}h(x^{(m)})}{w^{(1)} + \cdots + w^{(m)}}.$

- Remark: $\hat{\mu}$ is better than the unbiased estimator $\tilde{\mu} = \dfrac{1}{m}\{w^{(1)}h(x^{(1)}) + \cdots + w^{(m)}h(x^{(m)})\}.$ why?

An example (cont.)

- Use proposal function

$$g(x, y) \propto 0.5e^{-90(x-0.5)^2-10(y+0.1)^2} + e^{-45(x+0.4)^2-60(y-0.5)^2},$$

with $(x,y) \in [-1,1]$ x $[-1,1]$, a truncated mixture of bivariate Gaussian

$$0.46\mathcal{N}\left[\begin{pmatrix} 0.5 \\ -0.1 \end{pmatrix}, \begin{pmatrix} \frac{1}{180} & 0 \\ 0 & \frac{1}{20} \end{pmatrix}\right]+0.54\mathcal{N}\left[\begin{pmatrix} -0.4 \\ 0.5 \end{pmatrix}, \begin{pmatrix} \frac{1}{90} & 0 \\ 0 & \frac{1}{120} \end{pmatrix}\right]$$

Vanilla Monte Carlo	Importance Sampling
$\hat{\mu} = 0.1307$	$\hat{\mu} = 0.1259$
$std(\hat{\mu}) = 0.009$	$std(\hat{\mu}) = 0.0005$

Rao-Blackwellization

- Basic principle in Monte Carlo:

  carry out analytical computation as much as possible.

Rao-Blackwellization

- Estimating $E[h(x)]$.

  draw independent samples: $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, …

$$\hat{I} = \frac{1}{m}\left\{h(\mathbf{x}^{(1)}) + \cdots + h(\mathbf{x}^{(m)})\right\}.$$

- If $\mathbf{X}$=($\mathbf{x}_1$, $\mathbf{x}_2$) and $E[h(\mathbf{x})|x_2]$ can be carried out analytically.

$$\tilde{I} = \frac{1}{m}\left\{E[h(\mathbf{x}) \mid x_2^{(1)}] + \cdots + E[h(\mathbf{x}) \mid x_2^{(m)}]\right\}.$$

Rao-Blackwellization

- Both $\hat{I}$ and $\tilde{I}$ are unbiased.

$$E_\pi h(\mathbf{x}) = E_\pi[E\{h(\mathbf{x}) \mid x_2\}].$$

- But $\tilde{I}$ should be preferred becasue

$$\mathrm{var}(\hat{I}) = \frac{\mathrm{var}\{h(\mathbf{x})\}}{m} \geq \frac{\mathrm{var}\{E[h(\mathbf{x}) \mid x_2]\}}{m} = \mathrm{var}(\tilde{I}).$$

due to

$$\mathrm{var}\{h(\mathbf{x})\} = \mathrm{var}\{E[h(\mathbf{x}) \mid x_2]\} + E\{\mathrm{var}[h(\mathbf{x}) \mid x_2]\},$$

Rao-Blackwellization

- Conditioning an inferior estimator on the vale of sufficient statistics leads to the optimal estimator.

Sequential importance sampling

- For high dimensional problem, how to design trial distribution is challenging.
- Suppose the target density of $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ can be decomposed as

$$\pi(\mathbf{x}) = \pi(x_1)\pi(x_2 \mid x_1)\cdots\pi(x_d \mid x_1, .., x_{d-1})$$

then constructed trial density as

$$g(\mathbf{x}) = g_1(x_1)g_2(x_2 \mid x_1)\cdots g_d(x_d \mid x_1, .., x_{d-1})$$

Sequential importance sampling

$$w(\mathbf{x}) = \frac{\pi(x_1)\pi(x_2 \mid x_1)\cdots\pi(x_d \mid x_1,..,x_{d-1})}{g_1(x_1)g_2(x_2 \mid x_1)\cdots g_d(x_d \mid x_1,..,x_{d-1})}$$

Suggest a recursive way of computing and monitoring importance weight. Denote

$$\mathbf{x_t} = (x_1, x_2, \ldots, x_t)$$

then we have

$$w_t(\mathbf{x_t}) = w_{t-1}(\mathbf{x_{t\text{-}1}})\frac{\pi(x_t \mid \mathbf{x_{t\text{-}1}})}{g_t(x_t \mid \mathbf{x_{t\text{-}1}})}$$

Sequential importance sampling

- Advantages of the recursion scheme
  - Can stop generating further components of x if the partial weight is too small.
  - Can take advantage of $\pi(x_t \mid \mathbf{x_{t-1}})$ in designing $g_t(x_t \mid \mathbf{x_{t-1}})$
- However, the scheme is impractical since requires the knowledge of marginal distribution $\pi(\boldsymbol{x}_t)$.

Sequential importance sampling

- Add another layer of complexity:
- Introduce a sequence of "auxiliary distributions" $\pi_1(x_1)\pi_2(\mathbf{x_2})\pi_d(\mathbf{x})$ such that $\pi_t(\mathbf{x_t})$ is a reasonable approximation of the marginal distribution $\pi(\mathbf{x_t})$, for $t = 1,\ldots,d$ -1 and $\pi_d = \pi$.
- Note the π_d are only required to be known up to a normalizing constant.

The SIS procedure

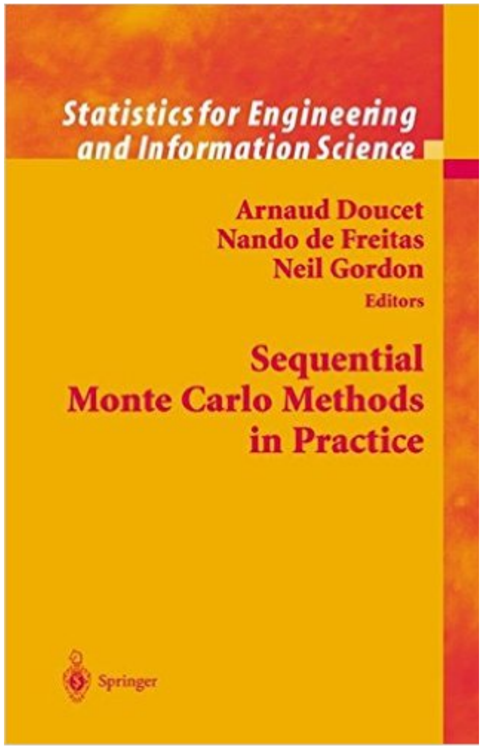
For $t = 2,\ldots,d,$

- Draw $X_t = x_t$ from $g_t(x_t \mid x_{t-1})$, and let

  $\mathbf{x_t} = (\mathbf{x_{t-1}}, x_t)$
- Compute $$u_t = \frac{\pi_t(\mathbf{x_t})}{\pi_{t-1}(\mathbf{x_{t-1}})g_t(x_t \mid \mathbf{x_{t-1}})}$$

  and let $w_t = w_{t\ \text{-}1}\, u_t$
- u_t : incremental weight.
- The key idea is to breaks a difficult task into manageable pieces.
- If w_t is getting too small, reject.

References



Statistics for Engineering and Information Science

Arnaud Doucet
Nando de Freitas
Neil Gordon
Editors

Sequential Monte Carlo Methods in Practice

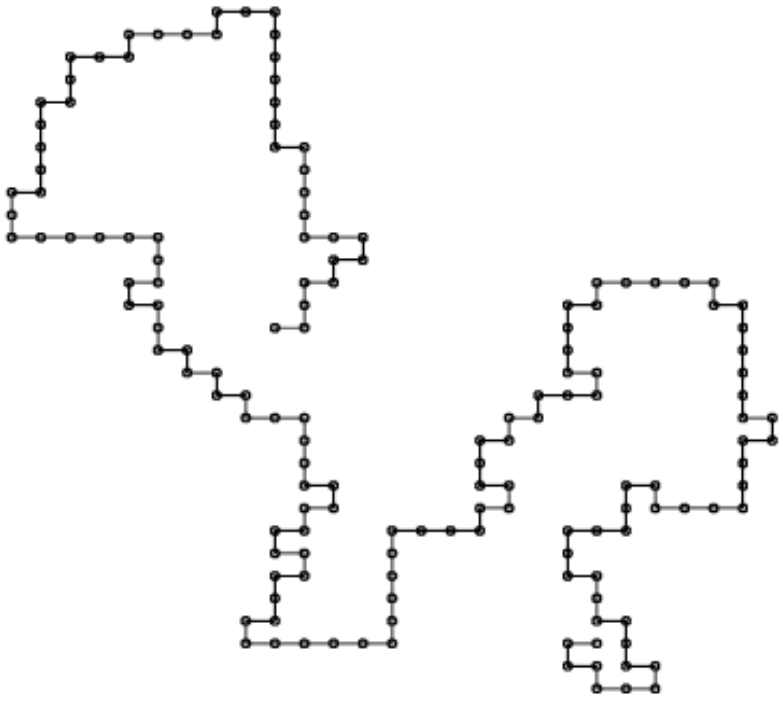
Springer

- Hammersley and Morton (1954).
- Rosenbluth and Rosenbluth (1955).
- Liu JS (2001)
- Doucet et al. (2001).

Examples of SIS

- Growing a polymer
  - Self avoid walk
- Sequential imputation for statistical missing data problem.
- More and details of these examples, see Liu 2001.

A Self-Avoiding Walk of Length N=150



Future topics

- Multigrid Monte Carlo (MGMC), density-scaling Monte Carlo, hybrid Monte Carlo (HMC), evolutionary Monte Carlo, exchange Monte Carlo.
- Cluster method, data augmentation. Parameter expansion, multicanonical sampling, umbrella sampling, simulated tempering, multi-try Metropolis, particle filtering, …