Some [very biased] Perspectives on Sequential Monte Carlo and Normalising Constants

Adam M. Johansen

Collaborators Include: John Aston, Alexandre Bouchard-Côté, Pierre Del Moral, Arnaud Doucet, Pieralberto Guarniero, Anthony Lee, Fredrik Lindsten, Christian Næsseth, Thomas Schön, Yan Zhou

University of Warwick
a.m.johansen@warwick.ac.uk
www2.warwick.ac.uk/fac/sci/statistics/staff/academic/johansen/talks/

CRiSM Workshop on Estimating Normalising Constants
University of Warwick, April 20th, 2016

Outline

- Background: SMC and (Ratios of) Constants
- Some Applications and Variations on a Theme
- Some "Interesting" (Open) Questions

Essential Problem

- Given a distribution,

$$\pi(dx) = \frac{\gamma(dx)}{Z} = \frac{\gamma(x)dx}{Z},$$

- such that $\gamma(x)$ can be evaluated pointwise,
- how can we "estimate" $Z = \int \gamma(dx)$?

Importance Sampling

- Simple Identity, provided $\gamma \ll \mu$:

$$Z = \int \gamma(dx) = \int \frac{\mathrm{d}\gamma}{\mathrm{d}\mu}(x)\mu(dx) = \int \frac{\gamma(x)}{\mu(x)}\mu(x)dx$$

- So, if $X_1, X_2, \ldots \overset{\text{iid}}{\sim} \mu$, then:

unbiasedness
$$\forall N : \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{d}\gamma}{\mathrm{d}\mu}(X_i)\right] = Z$$

slln
$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{d}\gamma}{\mathrm{d}\mu}(X_i) \overset{\text{a.s.}}{\to} Z$$

clt
$$\lim_{N\to\infty}\sqrt{N}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{d}\gamma}{\mathrm{d}\mu}(X_i) - Z\right] \overset{\text{d}}{\to} W$$

where $W \sim \mathcal{N}\left(0, \mathrm{Var}\left[\frac{\mathrm{d}\gamma}{\mathrm{d}\mu}(X_1)\right]\right)$.

Sequential Importance Sampling

- Write

$$\gamma(x_{1:n}) = \gamma(x_1) \prod_{p=2}^{n} \gamma(x_p|x_{1:p-1}),$$

- define, for $p = 1, \ldots, n$

$$\gamma_p(x_{1:p}) = \gamma_1(x_1) \prod_{q=2}^{p} \gamma(x_q|x_{1:q-1}),$$

- then

$$\underbrace{\frac{\gamma(x_{1:n})}{\mu(x_{1:n})}}_{W_n(x_{1:n})} = \underbrace{\frac{\gamma_1(x_1)}{\mu_1(x_1)}}_{w_1(x_1)} \prod_{p=2}^{n} \underbrace{\frac{\gamma_p(x_{1:p})}{\gamma_{p-1}(x_{1:p-1})\mu_p(x_p|x_{1:p-1})}}_{w_p(x_{1:p})},$$

- and we can *sequentially* approximate $Z_p = \int \gamma_p(x_{1:p}) dx_{1:p}$.

Sequential Importance *Resampling* (SIR)

Given a sequence $\gamma_1(x_1), \gamma_2(x_{1:2}), \ldots$:

Initialisation, $n = 1$:

- Sample $X_1^1, \ldots, X_1^N \overset{\text{iid}}{\sim} \mu_1$
- Compute

$$W_1^i = \frac{\gamma_1(X_1^i)}{\mu_1^i(X_1^i)}$$

- Obtain $\widehat{Z}_1^N = \frac{1}{N} \sum_{i=1}^N W_1^i$

[This is *just* importance sampling.]

Iteration, $n \leftarrow n+1$:

- Resample: sample $(A_{n-1}^1, \ldots, A_{n-1}^N) \sim r(\cdot|W_{n-1}^{1:N})$[1].
- Set $B_{n,n}^i = i$ and recursively $B_{n,p}^i = A_p^{B_{n,p+1}^i}$.
- Sample $X_n^i \sim q_n(\cdot|X_1^{B_{n,1}^i}, \ldots, X_{n-1}^{B_{n,n-1}^i})$
- Compute

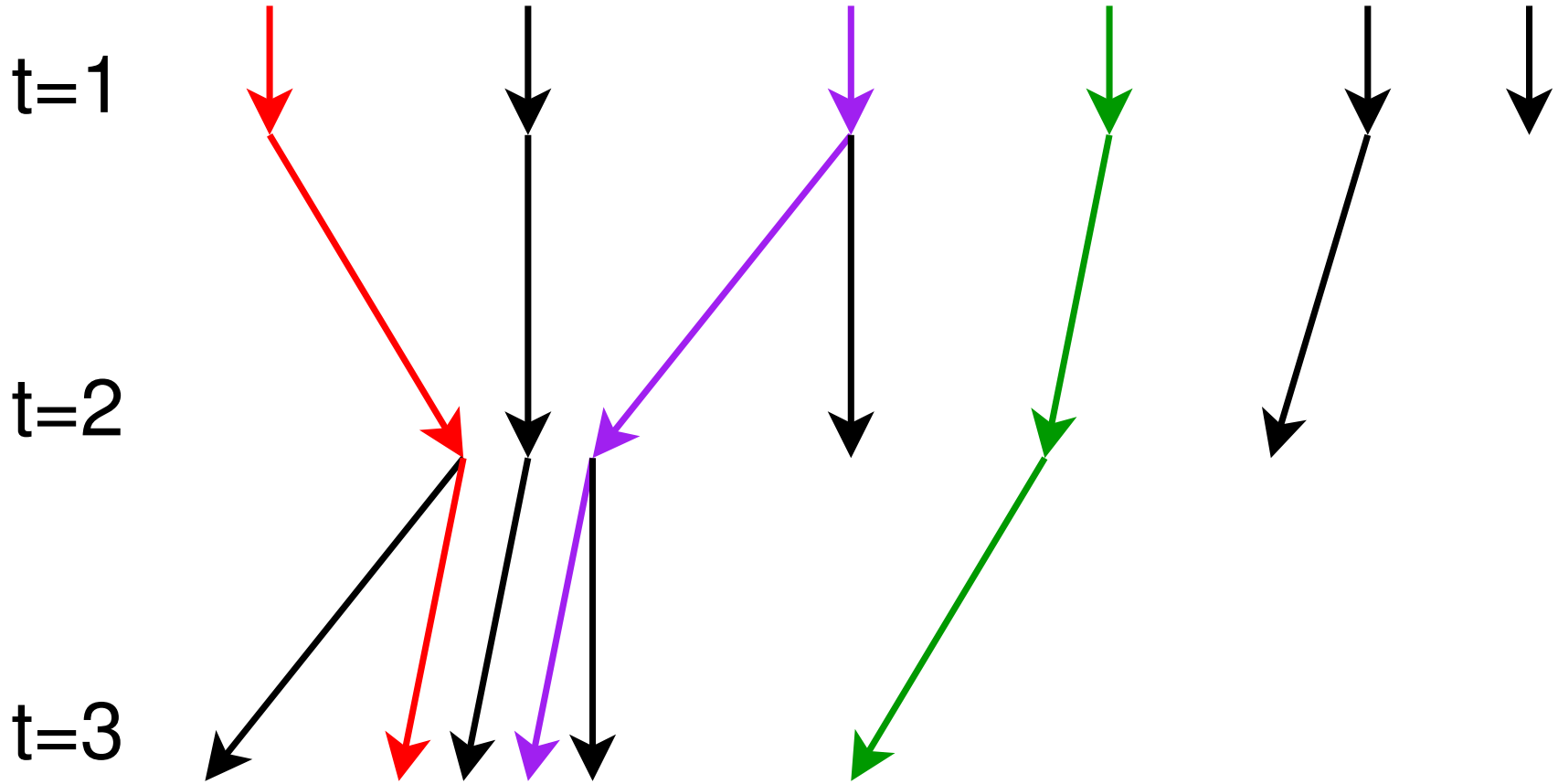
$$W_n^i = \frac{\gamma_n((X_1^{B_{n,1}^i}, \ldots, X_n^{B_{n,n}^i}))}{\gamma_{n-1}((X_1^{B_{n,1}^i}, \ldots, X_{n-1}^{B_{n,n-1}^i})) \cdot q_n(X_n^i|X_1^{B_{n,1}^i}, \ldots, X_{n-1}^{B_{n,n-1}^i})}$$

- Obtain

$$\widehat{Z}_n^N = \widehat{Z}_{n-1}^N \cdot \frac{1}{N} \sum_{i=1}^N W_n^i.$$

[1] such that $\mathbb{P}(A_{n-1}^i = j) \propto W_{n-1}^j$

Ancestral Trees



$a_2^1 = 1$ $a_2^4 = 3$ $a_1^1 = 1$ $a_1^4 = 3$

$b_{3,1:3}^2 = (1, 1, 2)$ $b_{3,1:3}^4 = (3, 3, 4)$ $b_{3,1:3}^6 = (4, 5, 6)$

SIR: Theoretical Justification

Under regularity conditions we still have:

unbiasedness

$$\mathbb{E}[\widehat{Z}_n^N] = Z_n$$

slln

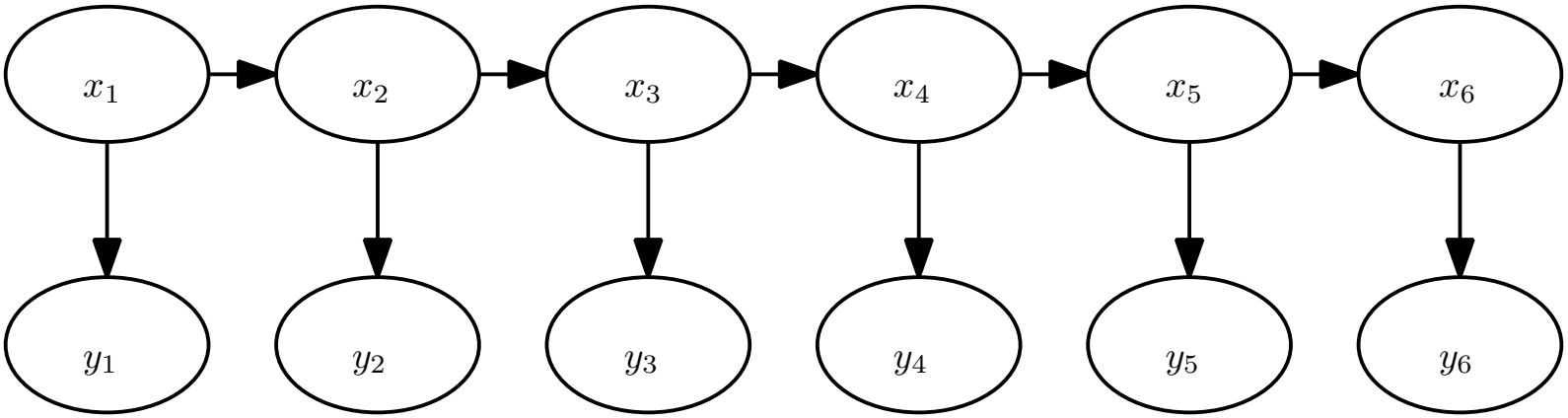
$$\lim_{N\to\infty} \widehat{Z}_n^N \stackrel{\text{a.s.}}{=} Z_n$$

clt For a normal random variable W_n of appropriate variance:

$$\lim_{N\to\infty} \sqrt{N}[\widehat{Z}_n^N - Z_n] \stackrel{\text{d}}{=} W_n$$

although establishing this becomes a little harder (cf., e.g. Del Moral (2004), Andrieu et al. 2010).

Simple Particle Filters: One Family of SIR Algorithms



- Unobserved Markov chain $\{X_n\}$ transition f.
- Observed process $\{Y_n\}$ conditional density g.
- The joint density is available:

$$p(x_{1:n}, y_{1:n}|\theta) = f_1^\theta(x_1)g^\theta(y_1|x_1)\prod_{i=2}^{n} f^\theta(x_i|x_{i-1})g^\theta(y_i|x_i).$$

- Natural SIR target distributions:

$$\pi_n^\theta(x_{1:n}) := p(x_{1:n}|y_{1:n}, \theta) \propto p(x_{1:n}, y_{1:n}|\theta) =: \gamma_n^\theta(x_{1:n})$$

$$Z_n^\theta = \int p(x_{1:n}, y_{1:n}|\theta)dx_{1:n} = p(y_{1:n}|\theta)$$

Bootstrap PFs and Similar

- Choosing

$$\pi_n^\theta(x_{1:n}) := p(x_{1:n}|y_{1:n},\theta) \propto p(x_{1:n},y_{1:n}|\theta) =: \gamma_n^\theta(x_{1:n})$$
$$Z_n^\theta = \int p(x_{1:n},y_{1:n}|\theta)dx_{1:n} = p(y_{1:n}|\theta)$$

- and $q_p(x_p|x_{1:p-1}) = f^\theta(x_p|x_{p-1})$ yields the bootstrap particle filter of Gordon et al. (1993),
- whereas $q_p(x_p|x_{1:p-1}) = p(x_p|x_{p-1},y_p,\theta)$ yields the "locally optimal" particle filter.
- Note: Many alternative particle filters are SIR algorithms with other targets. Cf. J. and Doucet (2008); Doucet and J. (2011).

Sequential Monte Carlo Samplers: Another SIR Class

Given a sequence of targets $\pi_1, \ldots, \pi_n$ on *arbitrary* spaces, Del Moral et al. (2006) extend the space:

$$\tilde{\pi}_n(x_{1:n}) = \pi_n(x_n) \prod_{p=n-1}^{1} L_p(x_{p+1}, x_p)$$

$$\tilde{\gamma}_n(x_{1:n}) = \gamma_n(x_n) \prod_{p=n-1}^{1} L_p(x_{p+1}, x_p)$$

$$\tilde{Z}_n = \int \tilde{\gamma}_n(dx_{1:n})$$

$$= \int \gamma_n(dx_n) \prod_{p=n-1}^{1} L_p(x_{p+1}, dx_p) = \int \gamma_n(dx_n) = Z_n$$

A Simple SMC Sampler

Given $\gamma_1, \ldots, \gamma_n$, on $(E, \mathcal{E})$, for $i = 1, \ldots, N$

- Sample $X_1^i \overset{\text{iid}}{\sim} \mu_1$ compute $W_1^i = \frac{\gamma_1(X_1^i)}{\mu_1(X_1^i)}$ and $\widehat{Z}_1^N = \frac{1}{N}\sum_{i=1}^N W_1^i$
- For $p = 2, \ldots, n$
  - Resample: $A_{n-1}^{1:N} \sim r(\cdot|W_{n-1}^{1:N})$.
  - Sample: $X_n^i \sim K_n(X_{n-1}^{A_{n-1}^i}, \cdot)$, where $\pi_n K_n = \pi_n$.
  - Compute: $W_n^i = \frac{\gamma_n(X_{n-1}^{A_{n-1}^i})}{\gamma_{n-1}(X_{n-1}^{A_{n-1}^i})}$.
  - And $\widehat{Z}_n^N = \widehat{Z}_{n-1}^N \cdot \frac{1}{N}\sum_{i=1}^N W_n^i$.

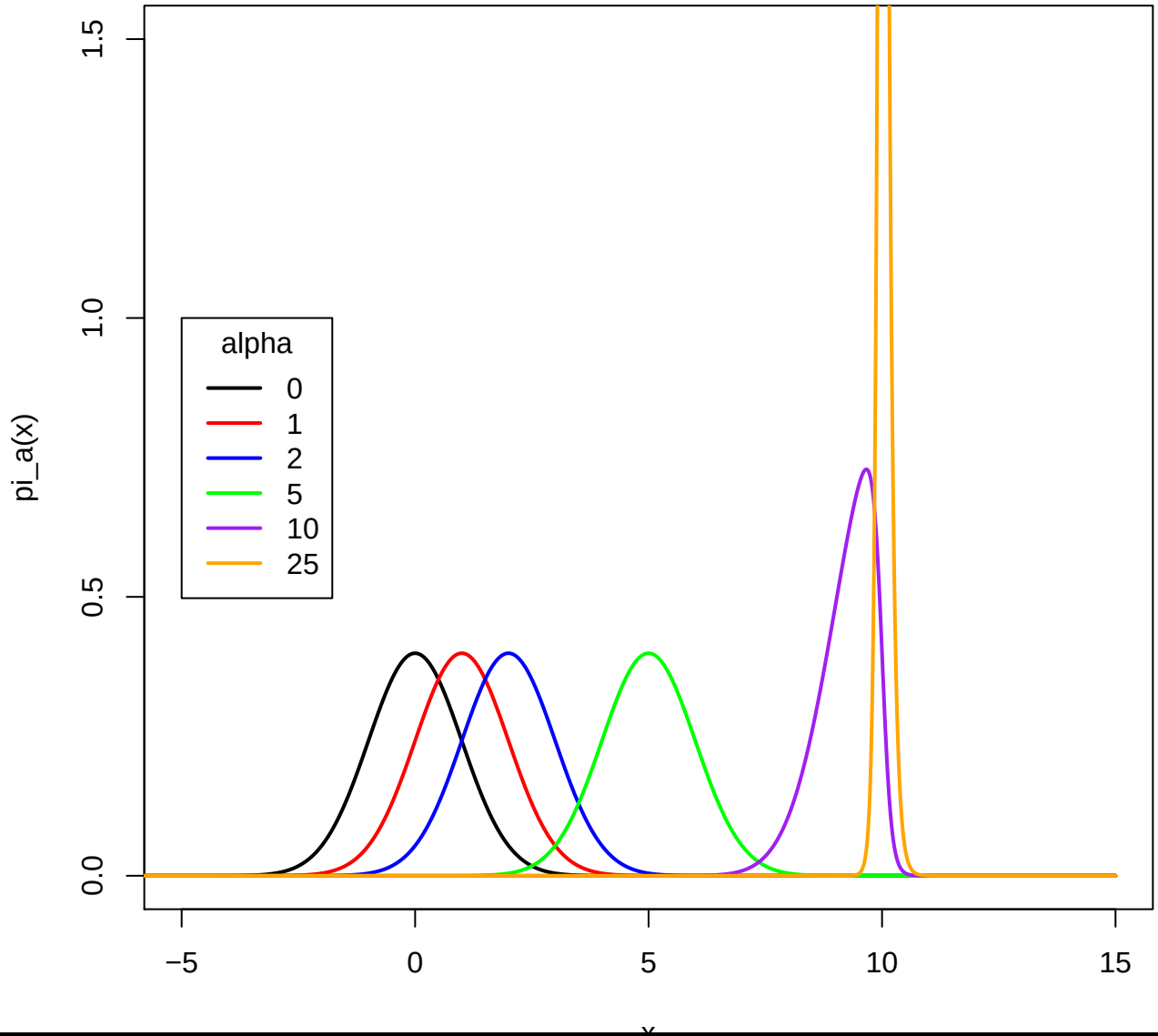
Rare Events (J., Del Moral and Doucet, 2006)

- If $Y \sim \eta$, $\mathbb{P}(Y \in A) = \eta(A) = \eta(\mathbb{I}_A)$.
- We're interested in $A = \{x : V(x) \geq \hat{V}\}$, with $\eta(A) \ll 1$.
- If $\gamma(x) = \eta(x)\mathbb{I}_A(x)$, then $Z_n = \eta(A)$.
- Let $\alpha_1 = 0 < \alpha_2 < \cdots < \alpha_T = \infty$, and define

$$\gamma_n(x) = \eta(x)(1 + \exp((\alpha_n(\widehat{V} - V(x))))^{-1}$$

- We have $\gamma_1 = \eta$ and $\gamma_T/\eta = \mathbb{I}_A$ so $Z_T = \eta(A)$.

An Illustrative Sequence of Targets



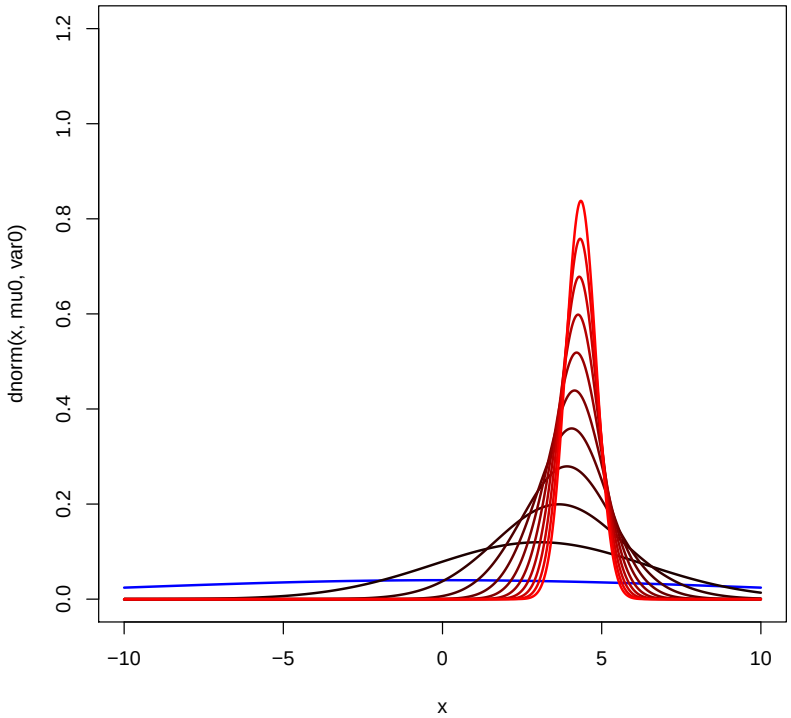
Evidence Evaluation (Chopin, 2001;Del Moral et al., 2006)

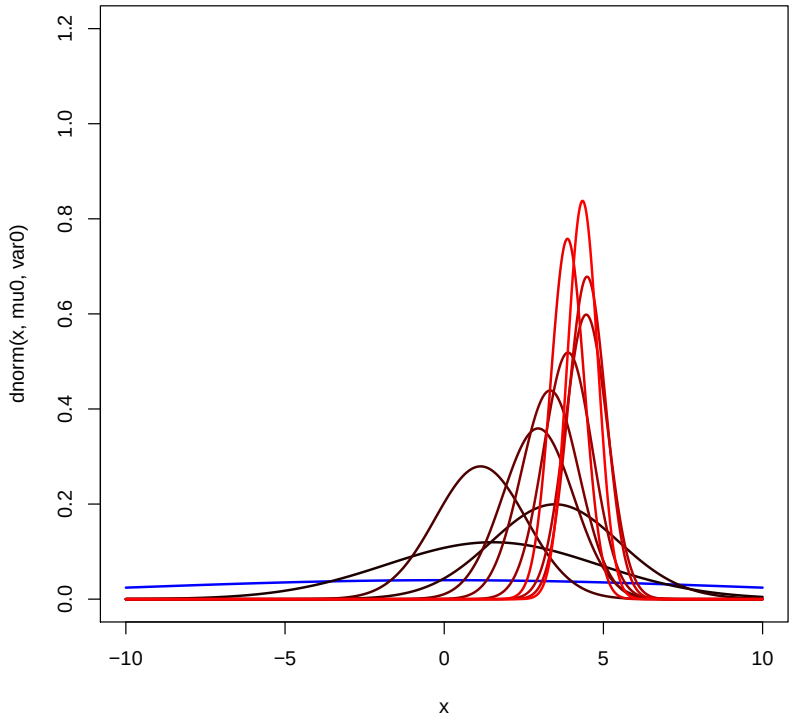
In a Bayesian context:

- Given a prior $p(\theta)$ and likelihood $l(\theta; y_{1:m})$
- One could specify:

  Data Tempering $\gamma_p(\theta) = p(\theta)l(\theta; y_{1:m_p})$ for
  $$m_1 = 0 < m_2 < \cdots < m_n = m$$
  Likelihood Tempering $\gamma_p(\theta) = p(\theta)l(\theta; y_{1:m})^{\beta_p}$ for
  $$\beta_1 = 0 < \beta_2 < \cdots < \beta_T = 1$$
  Something else?
- For such schemes $Z_T = \int p(\theta)l(\theta; y_{1:n})d\theta$.
- Specifying $(m_1, \ldots, m_T)$, $(\beta_1, \ldots, \beta_T)$ or $(\gamma_1, \ldots, \gamma_T)$ is not trivial.

Illustrative Sequences of Targets





One Adaptive Scheme (Zhou, J. & Aston, 2016)+Refs

Resample When ESS is below a threshold.

Likelihood Tempering At iteration n: Set α_n such that:

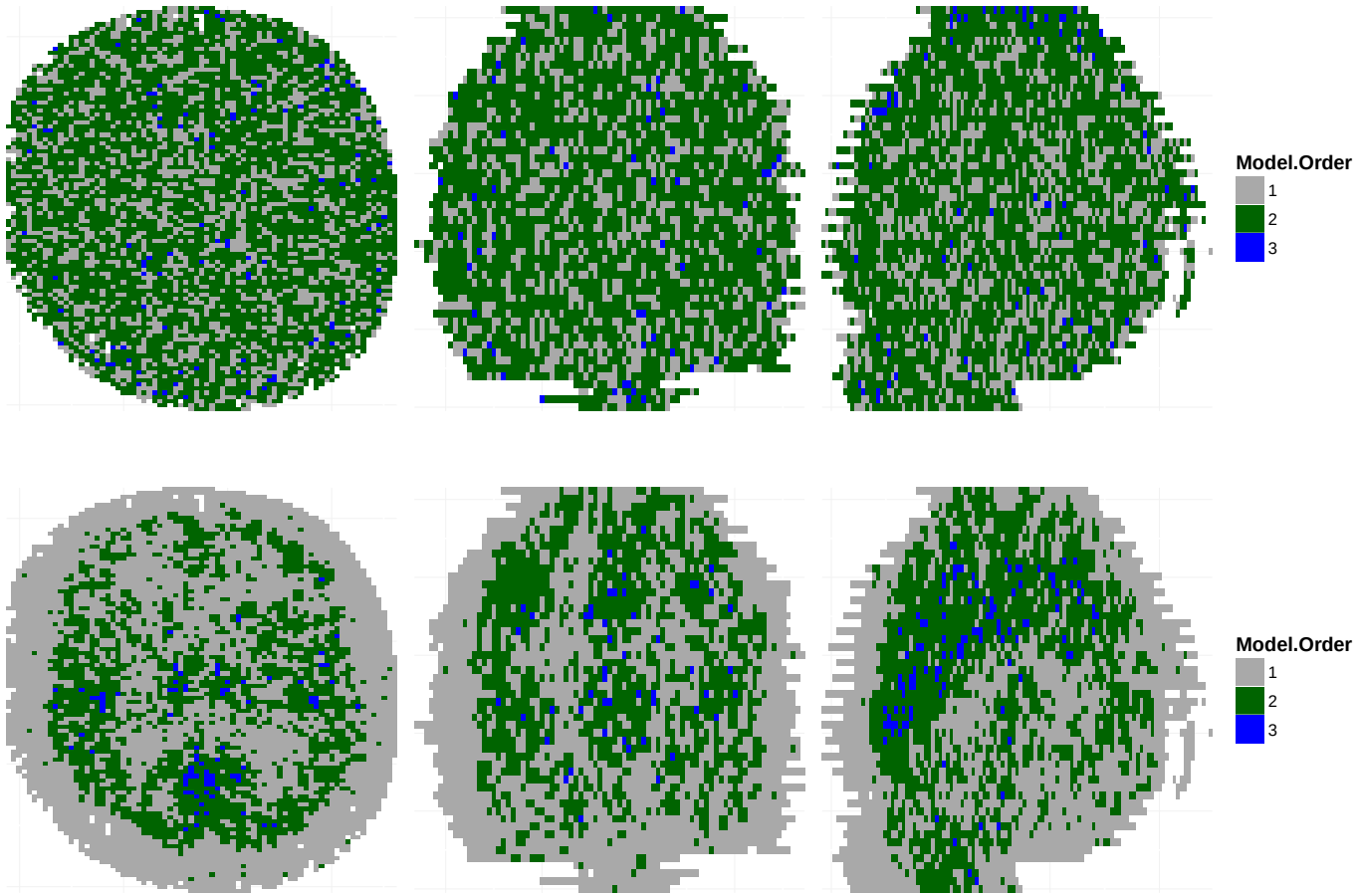
$$\frac{N(\sum_{j=1}^{N} W_{n-1}^{(j)} w_n^{(j)})^2}{\sum_{k=1}^{N} W_{n-1}^{(k)} (w_n^{(k)})^2} = \text{CESS}_\star$$

Proposals Follow (Jasra et al., 2010): adapt to keep acceptance rate about right.

Question

Are there better, practical approaches to specifying a sequence of distributions?

Results from Zhou et al. (2016)



Model selection results using AIC (above) / Evidence (below).

Iterative Adaptation (Guaniero, J. and Lee, 2015) [Poster 7]

- In filtering context

$$\pi_p(x_{1:p}) \propto p(x_{1:p}, y_{1:p}) \underbrace{p(y_{p+1:n}|x_p)}_{\psi_p^\star(x_p)}$$

  and

$$q_p(x_p|x_{1:p-1}) = p(x_p|x_{p-1}, y_{p:n})$$

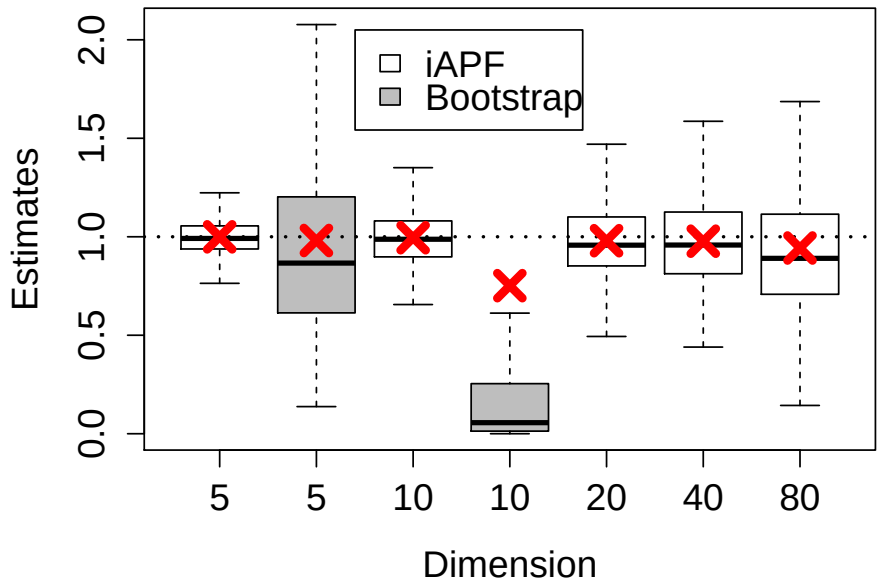
  gives better estimates of Z_T than the filtering distributions.
- The iAPF iteratively targets sequences $\pi_n^l(x_{1:n}) \propto p(x_{1:n}, y_{1:n})\psi_n^l(x_n)$ while estimating ψ_n^{l+1}.

Question

Can we approximate broader classes of transitions?
Are there better approximation schemes?

A Linear Gaussian Model: Behaviour with Dimension

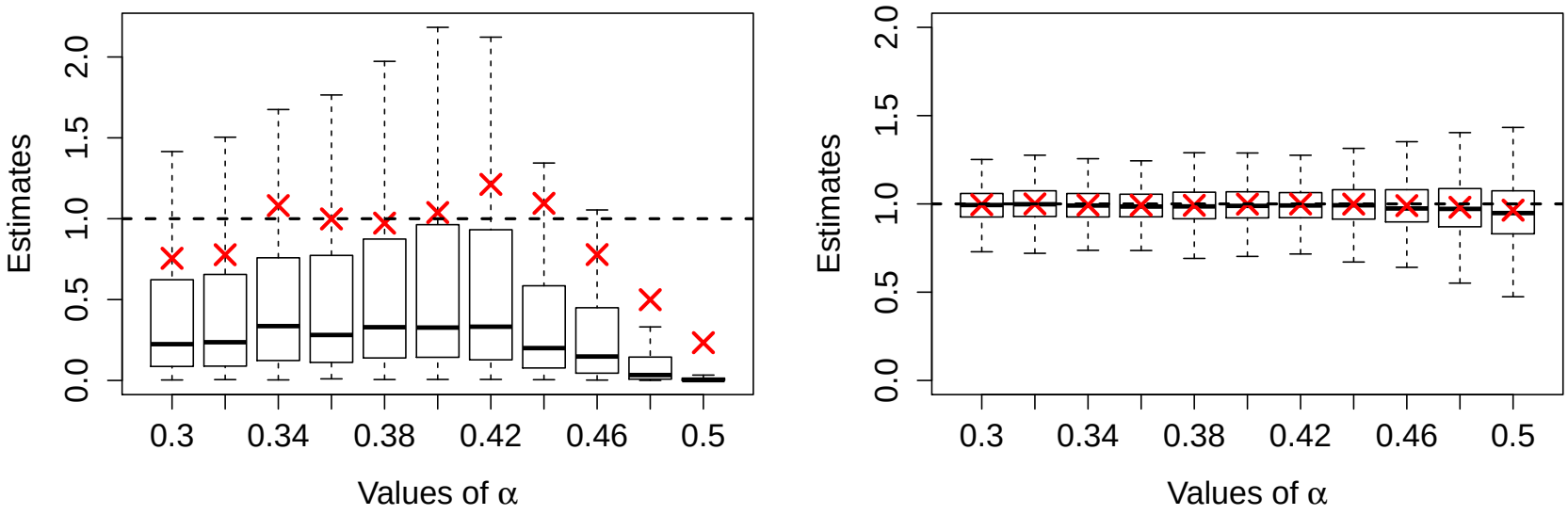
$$\mu = \mathcal{N}(\cdot; \mathbf{0}, I_d) \qquad f(x, \cdot) = \mathcal{N}(\cdot; Ax, I_d)$$
$$\text{and } g(x, \cdot) = \mathcal{N}(\cdot; x, I_d) \qquad \text{where } A_{ij} = 0.42^{|i-j|+1},$$



Box plots of $\hat{Z}/Z$ for different |X| (1000 replicates; $T = 100$).

Linear Gaussian Model: Sensitivity to Parameters

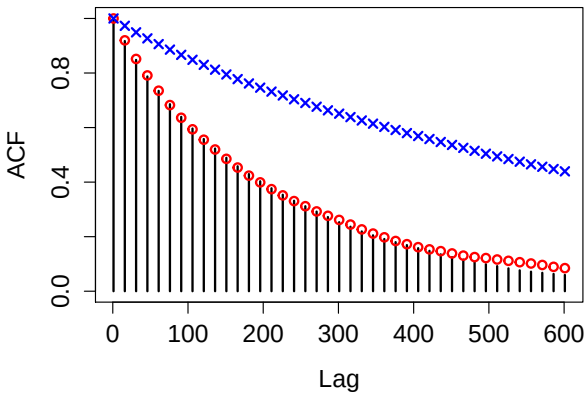
Fixing $d = 10$: Bootstrap ($N = 50,000$) / iAPF ($N_0 = 1,000$)

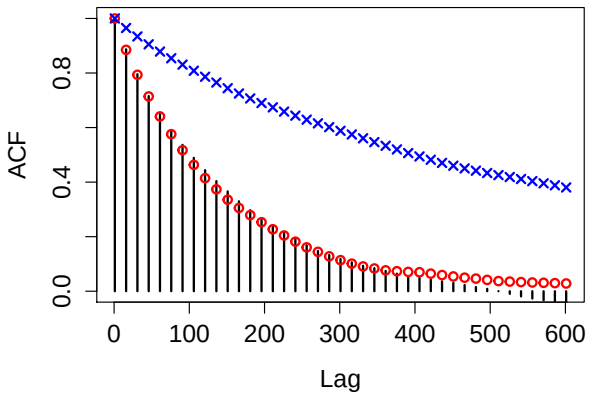


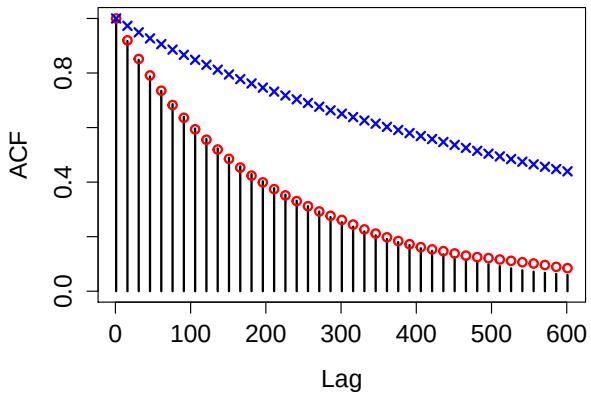
Box plots of $\frac{\hat{Z}}{Z}$ for different values of the parameter α using 1000 replicates.

Linear Gaussian Model: PMMH Empirical Autocorrelations

A_{11} A_{41} A_{55}





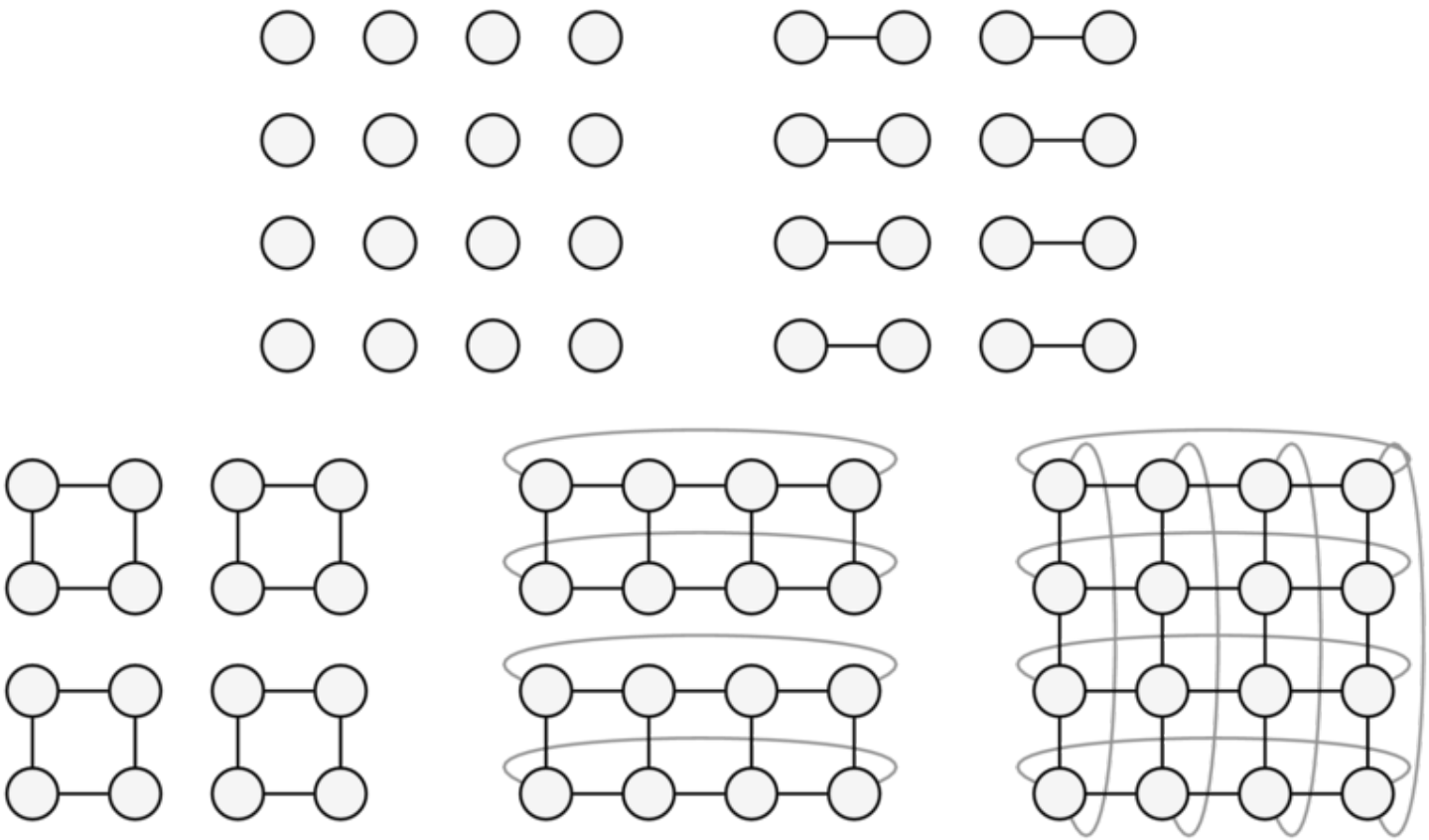


In this case:

$$d = 5$$
$$\mu = \mathcal{N}(\cdot; \mathbf{0}, I_d)$$
$$f(x, \cdot) = \mathcal{N}(\cdot; Ax, I_d)$$
$$\text{and } g(x, \cdot) = \mathcal{N}(\cdot; x, 0.25 I_d)$$

$$A = \begin{pmatrix} 0.9 & 0 & 0 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 & 0 \\ 0.1 & 0.2 & 0.6 & 0 & 0 \\ 0.4 & 0.1 & 0.1 & 0.3 & 0 \\ 0.1 & 0.2 & 0.5 & 0.2 & 0 \end{pmatrix},$$

Divide and Conquer (Lindsten, J. et al., 2014)



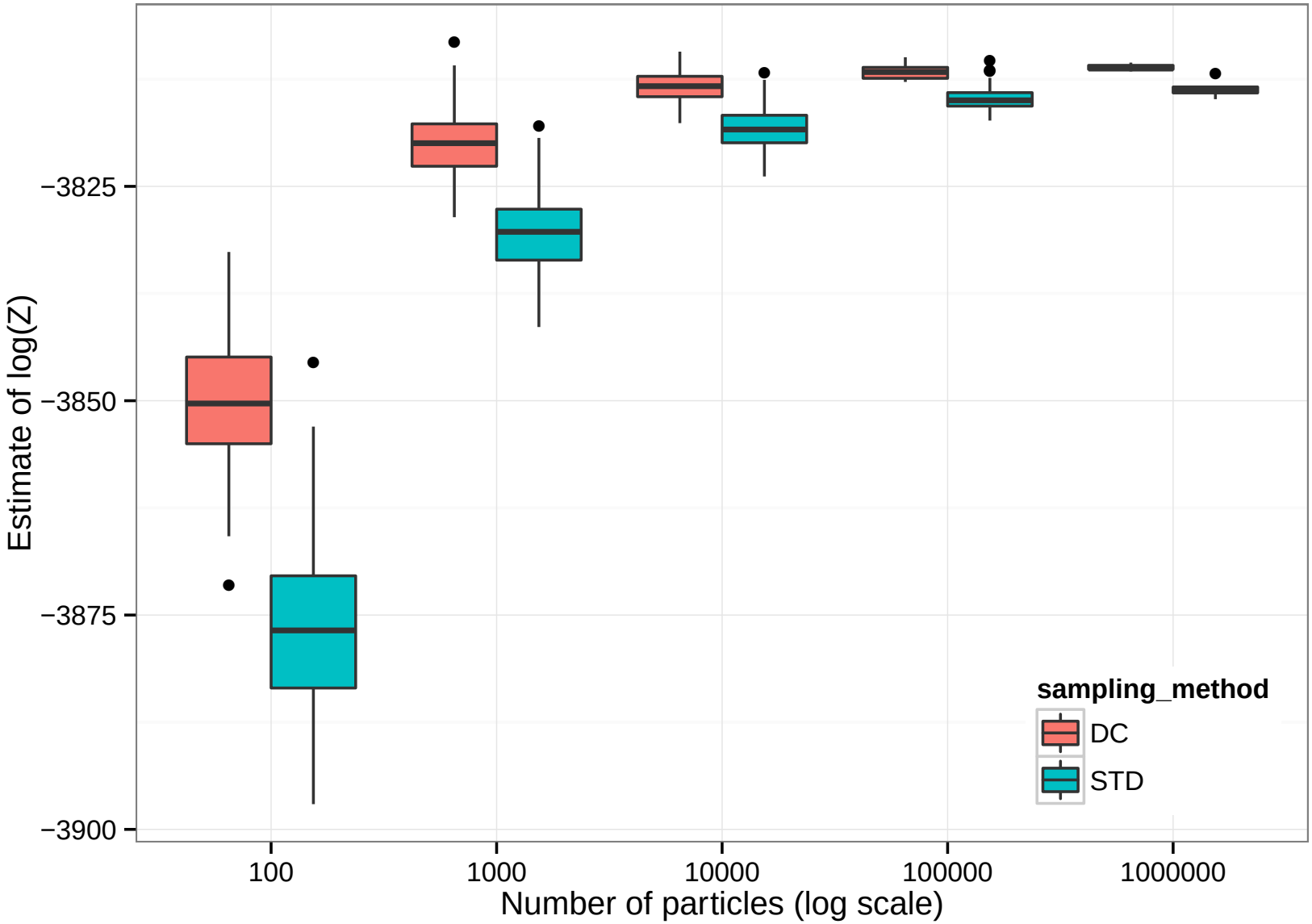
Question

Are there good ways to decompose general models?

dc-smc(t)

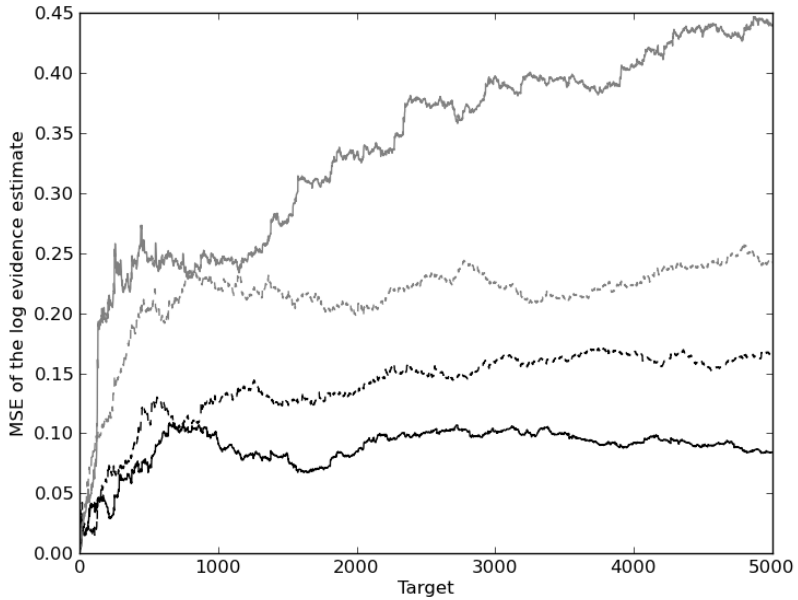
1. For $c \in \mathcal{C}(t)$:
   1.1 $(\{X_c^i, W_c^i\}_{i=1}^N, \widehat{Z}_c^N) \leftarrow$ dc-smc(c).
   1.2 Resample $\{\mathbf{x}_c^i, \mathbf{w}_c^i\}_{i=1}^N$ to obtain the equally weighted particle system $\{\hat{\mathbf{x}}_c^i, 1\}_{i=1}^N$.
2. For particle $i = 1 : N$:
   2.1 If $\tilde{\mathcal{X}}_t \neq \emptyset$, simulate $\widetilde{\mathbf{x}}_t^i \sim q_t(\cdot \mid \hat{\mathbf{x}}_{c_1}^i, \ldots, \hat{\mathbf{x}}_{c_C}^i)$, where $(c_1, c_2, \ldots, c_C) = \mathcal{C}(t)$;
   else $\widetilde{\mathbf{x}}_t^i \leftarrow \emptyset$.
   2.2 Set $\mathbf{x}_t^i = (\hat{\mathbf{x}}_{c_1}^i, \ldots, \hat{\mathbf{x}}_{c_C}^i, \widetilde{\mathbf{x}}_t^i)$.
   2.3 Compute $\mathbf{w}_t^i = \frac{\gamma_t(\mathrm{x}_t^i)}{\prod_{c \in \mathcal{C}(t)} \gamma_c(\hat{\mathrm{x}}_c^i)} \frac{1}{q_t(\widetilde{\mathrm{x}}_t^i \mid \hat{\mathrm{x}}_{c_1}^i, \ldots, \hat{\mathrm{x}}_{c_C}^i)}$.
3. Compute $\widehat{Z}_t^N = \left\{ \frac{1}{N} \sum_{i=1}^N \mathbf{w}_t^i \right\} \prod_{c \in \mathcal{C}(t)} \widehat{Z}_c^N$.
4. Return $(\{\mathbf{x}_t^i, \mathbf{w}_t^i\}_{i=1}^N, \widehat{Z}_t^N)$.

Multilevel modelling of NYC maths test data



How Wrong is it to Approximate? (Everitt, J. et al., 2015) [Poster 4]

What if W_p^i can't be evaluated exactly or K_n isn't exactly π_n-invariant?



MSE of $\log(\widehat{Z}_n^N)$, weights:
exact (black solid),
unbiased random (black dashed),
biased random (grey solid)
biased random — perfect mixing (grey dashed).

Question

Can we actually say anything practically relevant?

Inconclusive Conclusions

- SMC provides good estimates of many (normalizing) constants.
- There are many questions left to answer:
  - How best to choose sequences of distributions...
    - "Filtering"
    - Generally: Tempering / Data-tempering / not tempering?
    - Model-decompositions? Tempering to Independence?
  - To what extent can we ever justify using approximate approximations within SMC?
  - How best can we approximate optimal lookahead functions?
  - How much can be done adaptively?

References

[1] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo. *Journal of the Royal Statistical Society B*, 72(3):269–342, 2010.

[2] N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–551, 2002.

[3] P. Del Moral. *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Probability and Its Applications. Springer Verlag, New York, 2004.

[4] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo methods for Bayesian Computation. In *Bayesian Statistics 8*. Oxford University Press, 2006.

[5] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fiteen years later. In D. Crisan and B. Rozovsky, editors, *The Oxford Handbook of Nonlinear Filtering*, pages 656–704. Oxford University Press, 2011.

[6] R. G. Everitt, A. M. Johansen, E. Rowing, and M. Evdemon-Hogan. Bayesian model selection with un-normalised likelihoods. *Statistics and Computing*, 2016. In press.

[7] N. J. Gordon, S. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, April 1993.

[8] P. Guarniero, A. M. Johansen, and A. Lee. The iterated auxiliary particle filter. ArXiv mathematics e-print 1511.06286, ArXiv Mathematics e-prints, 2015.

[9] A. Jasra, D. A. Stephens, A. Doucet, and T. Tsagaris. Inference for Lévy-Driven Stochastic Volatility Models via Adaptive Sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38(1):1–22, Dec. 2010.

[10] A. M. Johansen and A. Doucet. A note on the auxiliary particle filter. *Statistics and Probability Letters*, 78 (12):1498–1504, September 2008. URL `http://dx.doi.org/10.1016/j.spl.2008.01.032`.

[11] A. M. Johansen, P. Del Moral, and A. Doucet. Sequential Monte Carlo samplers for rare events. In *Proceedings of the 6th International Workshop on Rare Event Simulation*, pages 256–267, Bamberg, Germany, October 2006.

[12] F. Lindsten, A. M. Johansen, C. A. Næsseth, B. Kirkpatrick, T. Schön, J. A. D. Aston, and A. Bouchard-Côté. Divide and conquer with sequential Monte Carlo samplers. Technical Report 1406.4993, ArXiv Mathematics e-prints, 2014.

[13] Y. Zhou, A. M. Johansen, and J. A. D. Aston. Towards automatic model comparison: An adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 2015. In press.