Monte Carlo methods for inference in dynamical systems

Fredrik Lindsten

Uppsala University

March 30, 2016

Identification of state-space models

Consider a nonlinear, discrete-time state-space model,

$$x_{t+1} \mid x_t \sim f_\theta(x_{t+1} \mid x_t),$$
$$y_t \mid x_t \sim g_\theta(y_t \mid x_t),$$

and $x_1 \sim \mu(x_1)$.

We observe $y_{1:T} := (y_1, \ldots, y_T)$ and wish to estimate θ.

Bayesian model: θ is a random variable with prior density $p(\theta)$.

Aim: Compute $p(\theta \mid y_{1:T})$.

Gibbs sampler for SSMs

Aim: Find $p(\theta \mid y_{1:T})$.

Alternate between updating θ and updating $x_{1:T}$.

$\Rightarrow$ Gibbs sampling

Crash course in MCMC (I/IV)

Markov kernel

Let $\{X_k\}_{k\geq 0}$ be a Markov chain on a general state space.
The *Markov transition kernel* is given by,

$$P(x_{k-1}, A) = \mathbb{P}(X_k \in A \mid X_{k-1} = x_{k-1}).$$

Stationary distribution

If π is a probability distribution such that,

$$\pi(A) = \int P(x, A)\pi(x)\, dx,$$

we say that π is a *stationary* distribution for P.

Crash course in MCMC (II/IV)

Ergodic theorem

If $\{X_k\}_{k\geq 0}$ is *ergodic* with stationary distribution π, then

$$\mathbb{E}_\pi[g(X)] = \underbrace{\int g(x)\pi(x)dx}_{\text{average over space}} = \lim_{m\to\infty} \underbrace{\frac{1}{m}\sum_{k=1}^{m} g(X_k)}_{\text{average over "time"}}$$

Markov chain Monte Carlo: We want to compute $\mathbb{E}_\pi[g(X)]$

1. Construct an ergodic Markov kernel P with stationary distribution π.
2. Simulate $\{X_k\}_{k=1}^m$ according to P for large m.
3. Approximate $\mathbb{E}_\pi[g(X)] \approx \frac{1}{m}\sum_{k=1}^m g(X_k)$

Crash course in MCMC (III/IV)

Two common methods for constructing P:

1. Metropolis-Hastings
2. Gibbs sampling

Update components X_j of the vector $X = (X_1, \ldots, X_d)$ in turn,

$$X_j \sim \pi(x_j \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d), \qquad j = 1, \ldots, d.$$

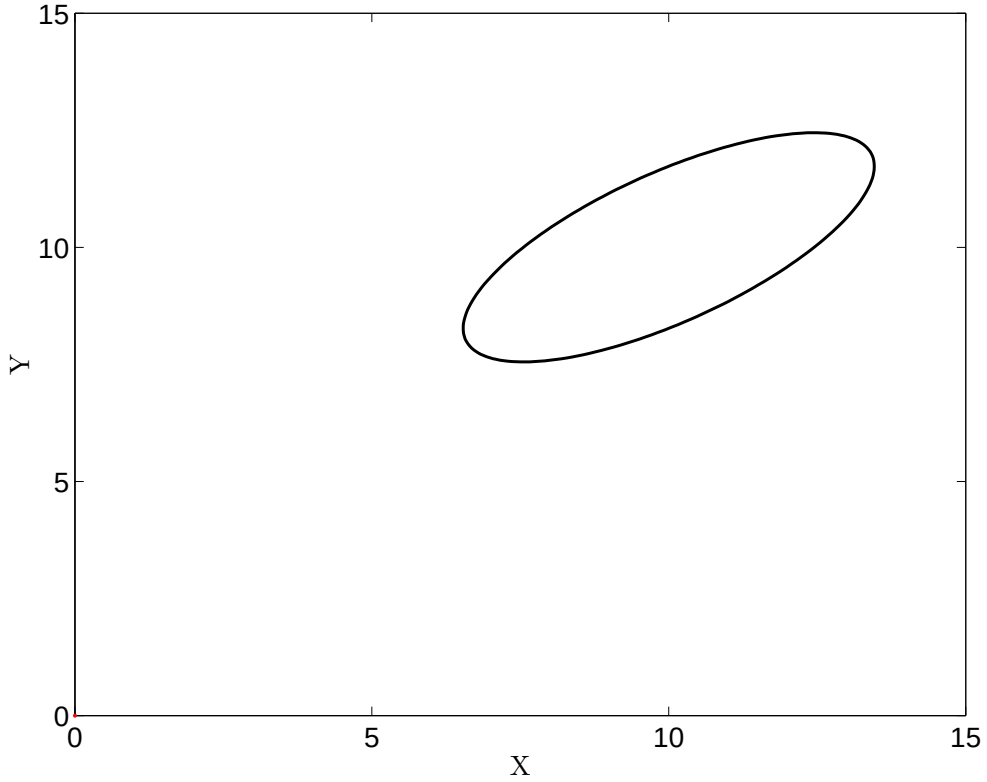
Crash course in MCMC (IV/IV)

ex)

Sample from, $\pi(x, y) = \mathcal{N}\left(\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} 10 \\ 10 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right).$

Gibbs sampler

- Draw $x' \sim \pi(x \mid y)$;
- Draw $y' \sim \pi(y \mid x')$.



Gibbs sampler for SSMs, cont'd

Aim: Find $p(\theta, x_{1:T} \mid y_{1:T})$.

Alternate between updating θ and updating $x_{1:T}$.

MCMC: Gibbs sampling for state-space models. Iterate,

- Draw $\theta[k] \sim p(\theta \mid x_{1:T}[k-1], y_{1:T})$;
- Draw $x_{1:T}[k] \sim p(x_{1:T} \mid \theta[k], y_{1:T})$.

The above procedure results in a Markov chain,

$$\{\theta[k], x_{1:T}[k]\}_{k \geq 1}$$

with stationary distribution $p(\theta, x_{1:T} \mid y_{1:T})$.

Linear Gaussian state-space model (I/II)

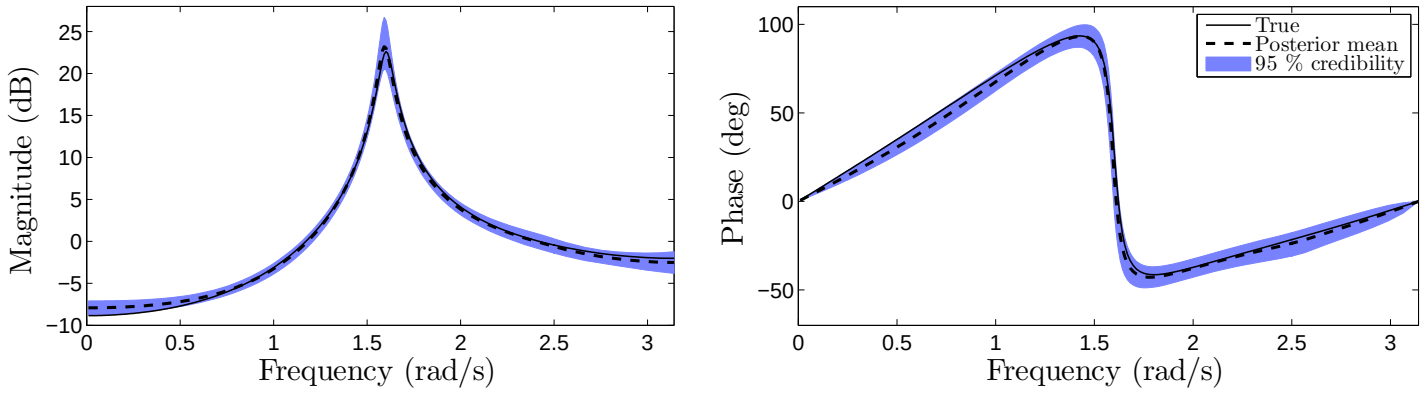
ex) Gibbs sampling for linear system identification.

$$x_{t+1} = Ax_t + v_t, \qquad v_t \sim \mathcal{N}(0, Q),$$
$$y_t = Cx_t + e_t, \qquad e_t \sim \mathcal{N}(0, R),$$

with $\theta = (A, C, Q, R)$.

Gibbs sampling: Iterate,

- Draw $\theta' \sim p(\theta \mid x_{1:T}, y_{1:T})$;
- Draw $x'_{1:T} \sim p(x_{1:T} \mid \theta', y_{1:T})$.
  - Run a Kalman filter for $t = 1, \ldots, T$.
  - Simulate backward, x'_T, x'_{T-1}, $\ldots$, x'_1.



Gibbs sampler for general SSM?

What about the general nonlinear/non-Gaussian case?

- Draw $\theta[k] \sim p(\theta \mid x_{1:T}[k-1], y_{1:T})$; **OK!**
- Draw $x_{1:T}[k] \sim p(x_{1:T} \mid \theta[k], y_{1:T})$. **Hard!**

One-at-a-time: $x_t[k] \sim p(x_t \mid \theta[k], x_{t-1}[k], x_{t+1}[k-1], y_t)$

Idea: Approximate $p(x_{1:T} \mid \theta, y_{1:T})$ using a particle filter.

The particle filter

Consider *state inference* for a fixed value of θ.

The *particle filter* approximates $p(x_t \mid \theta, y_{1:t})$, $t = 1, \ldots, T$ by

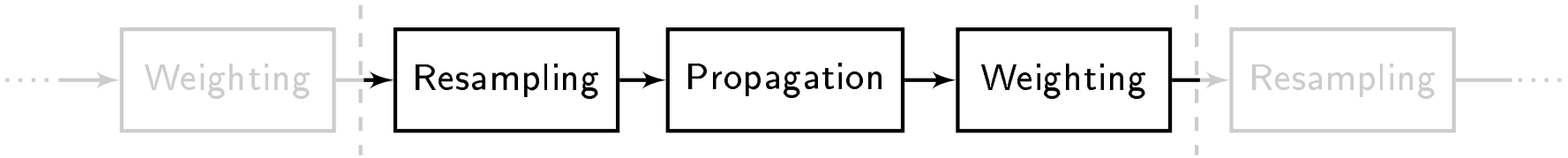
$$\widehat{p}^N(x_t \mid \theta, y_{1:t}) := \sum_{i=1}^{N} \frac{w_t^i}{\sum_\ell w_t^\ell} \delta_{x_t^i}(x_t).$$

---

t = 1, importance sampling from prior: For $i = 1, \ldots, N$,

- Sample $x_1^i \sim \mu(x_1)$;
- Compute weights, $w_1^i = g_\theta(y_1 \mid x_1^i)$.

The particle filter



$t > 1$:

- Resampling: $\{x_{t-1}^i, w_{t-1}^i\}_{i=1}^N \to \{\tilde{x}_{t-1}^i, 1/N\}_{i=1}^N$.
- Propagation: $x_t^i \sim f_\theta(x_t \mid \tilde{x}_{t-1}^i)$.
- Weighting: $w_t^i = g_\theta(y_t \mid x_t^i)$.

$$\Rightarrow \{x_t^i, w_t^i\}_{i=1}^N$$

N.B. This is the simplest PF and many extensions and improvements are avialble.

The particle filter

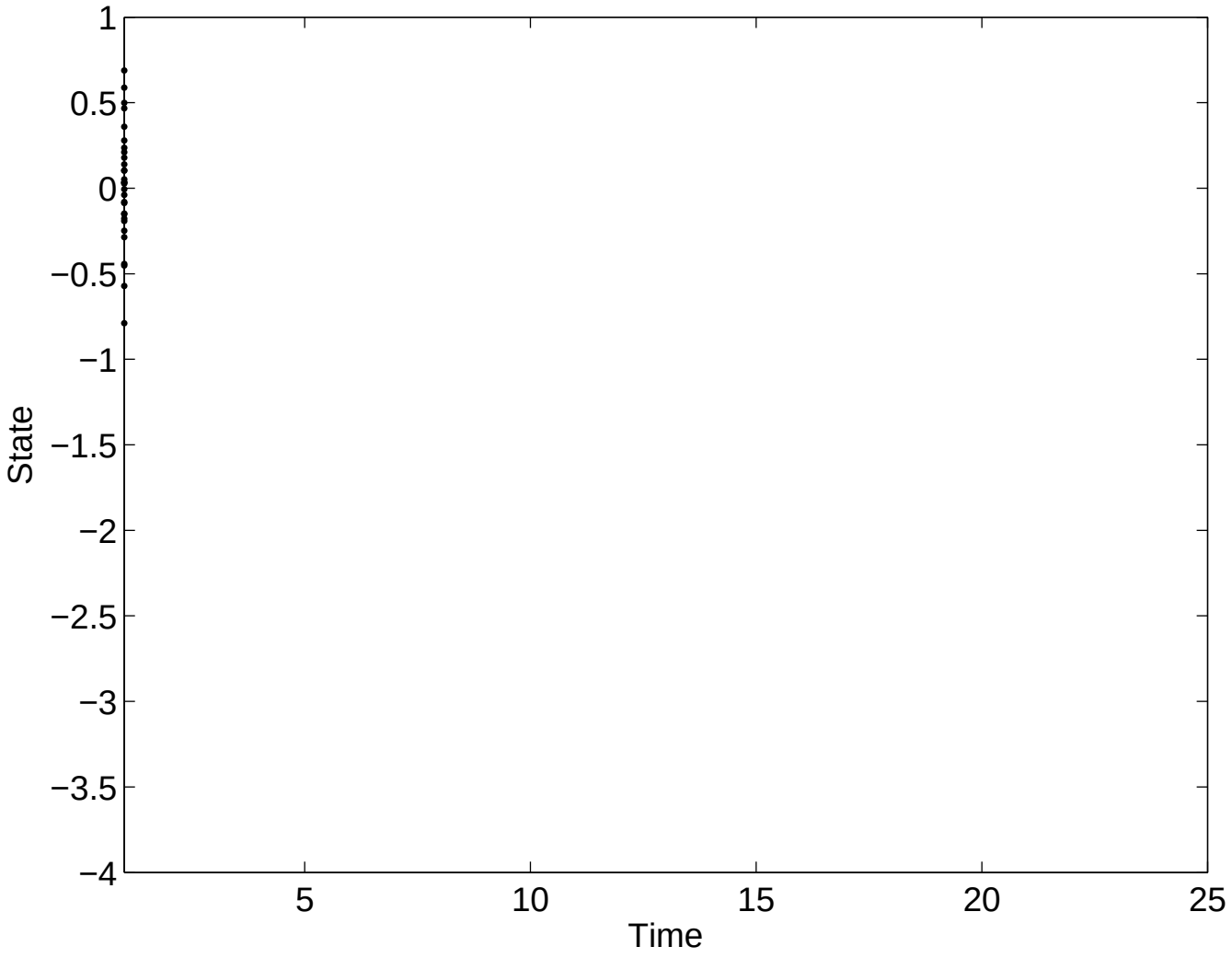
Algorithm Bootstrap particle filter (all steps for $i = 1, \ldots, N$)

1. **Initialize ($t = 1$):**
   (a) Draw $x_1^i \sim \mu(x_1)$.
   (b) Set $w_1^i = g_\theta(y_1 \mid x_1^i)$.
2. **for** $t = 2, 3, \ldots$**:**
   (a) Draw $a_t^i \sim \text{Discrete}(\{w_{t-1}^j\}_{j=1}^N)$.
   (b) Draw $x_t^i \sim f_\theta(x_t \mid x_{t-1}^{a_t^i})$.
   (c) Set $w_t^i = g_\theta(y_t \mid x_t^i)$.

Theoretical justification:

- Strongly consistent, central limit theorem at rate $\sqrt{N}$.
- Stable as $t \to \infty$ under mixing assumptions.
- Unbiased estimate of normalizing constant $p(y_{1:t} \mid \theta)$.
- ...

PF illustration



Filtering or smoothing?

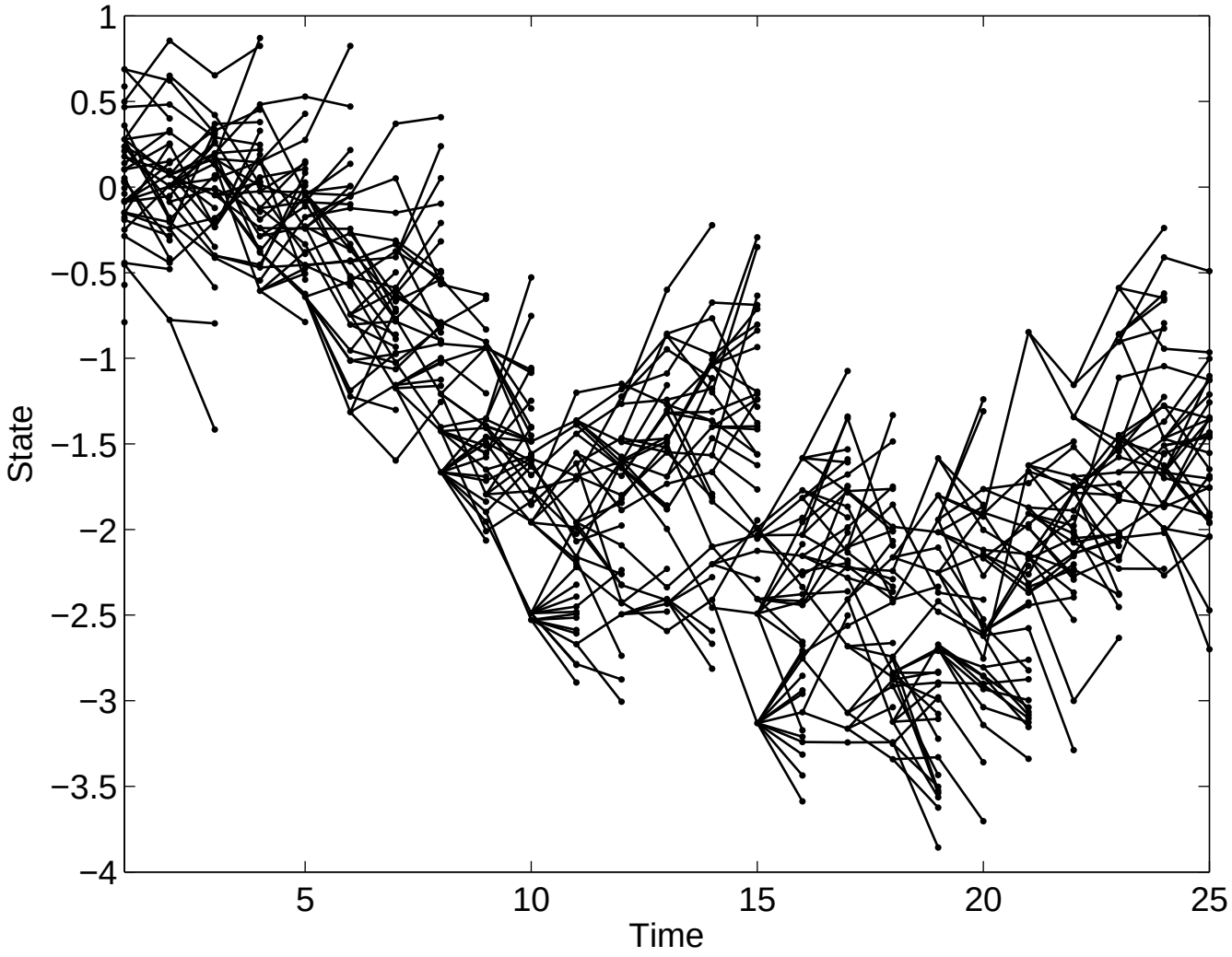
- Possible to reconstruct *state trajectories* by tracing the genealogy of the particles,

$$(\ldots,\ x_{T-2}^{a_{T-1}^{a_T^i}},\ x_{T-1}^{a_T^i},\ x_T^i) =: x_{1:T}^i$$

- At $t = T$ the particle filter provides an approximation of the *joint smoothing distribution*, $p(x_{1:T} \mid \theta, y_{1:T})$:

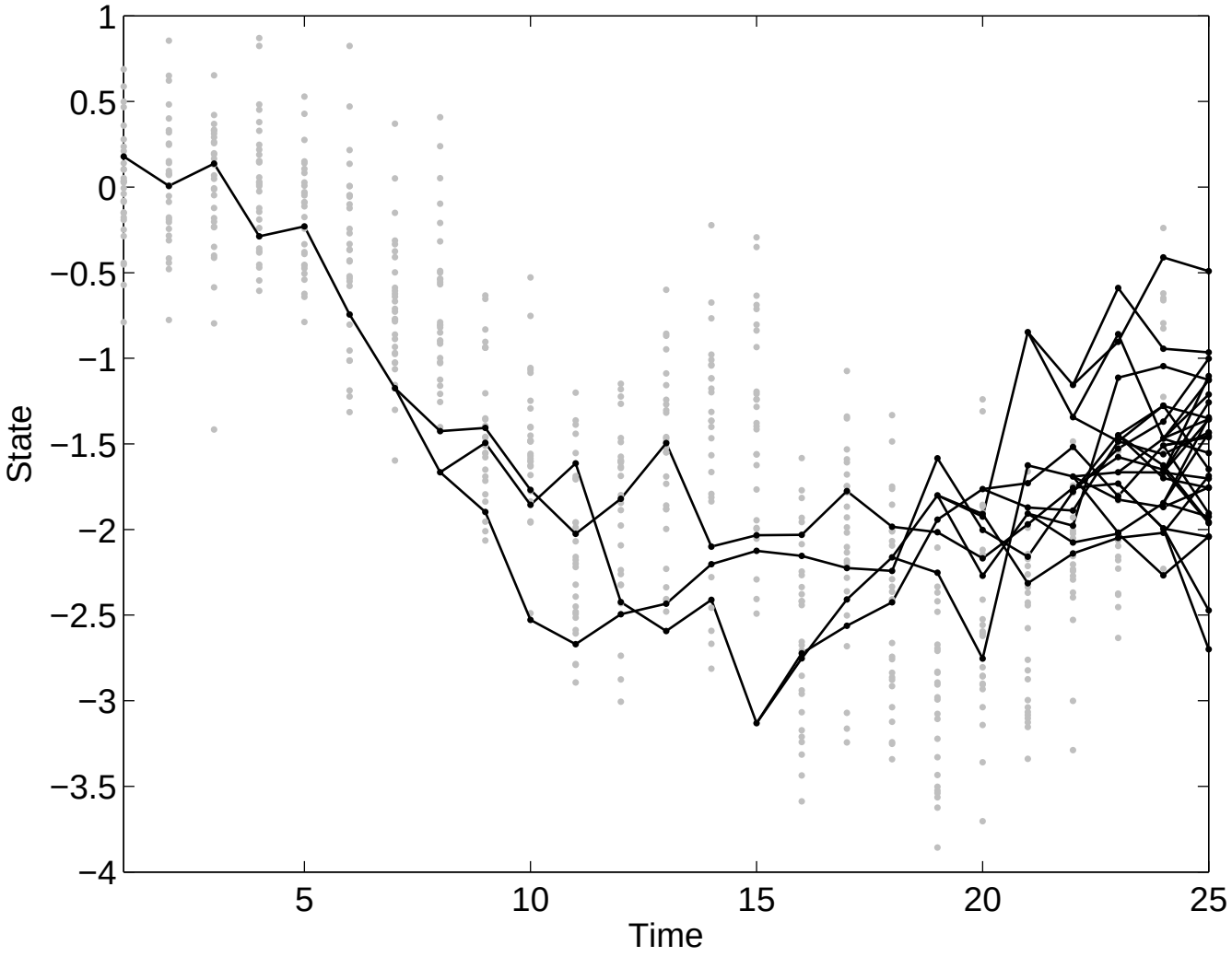
$$\widehat{p}^N(x_{1:T} \mid \theta, y_{1:T}) := \sum_{i=1}^{N} \frac{w_T^i}{\sum_\ell w_T^\ell} \delta_{x_{1:T}^i}(x_{1:T}).$$

Path degeneracy



fredrik.lindsten@it.uu.se

Path degeneracy



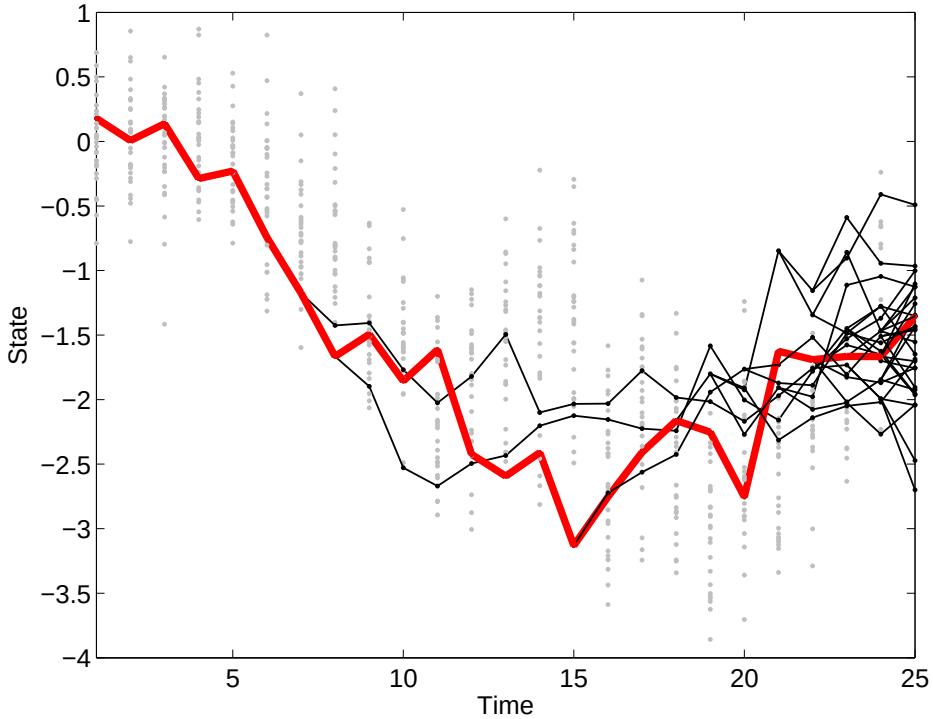
fredrik.lindsten@it.uu.se

Sampling based on the PF

Recall Gibbs sampler:

Want to sample from $p(x_{1:T} \mid \theta, y_{1:T})$ for a fixed value of θ.

With $\mathbb{P}(\boldsymbol{x}^\star_{1:T} = x^i_{1:T}) \propto w^i_T$ we get $\boldsymbol{x}^\star_{1:T} \overset{\text{approx.}}{\sim} p(x_{1:T} \mid \theta, y_{1:T})$.



Particle MCMC idea

Problems with this approach,

- Based on a PF $\Rightarrow$ approximate sample.
- $N \to \infty$ required to get a correct MCMC.
- A lot of wasted computations.

Analyze PF + MCMC together $\Rightarrow$ Particle MCMC

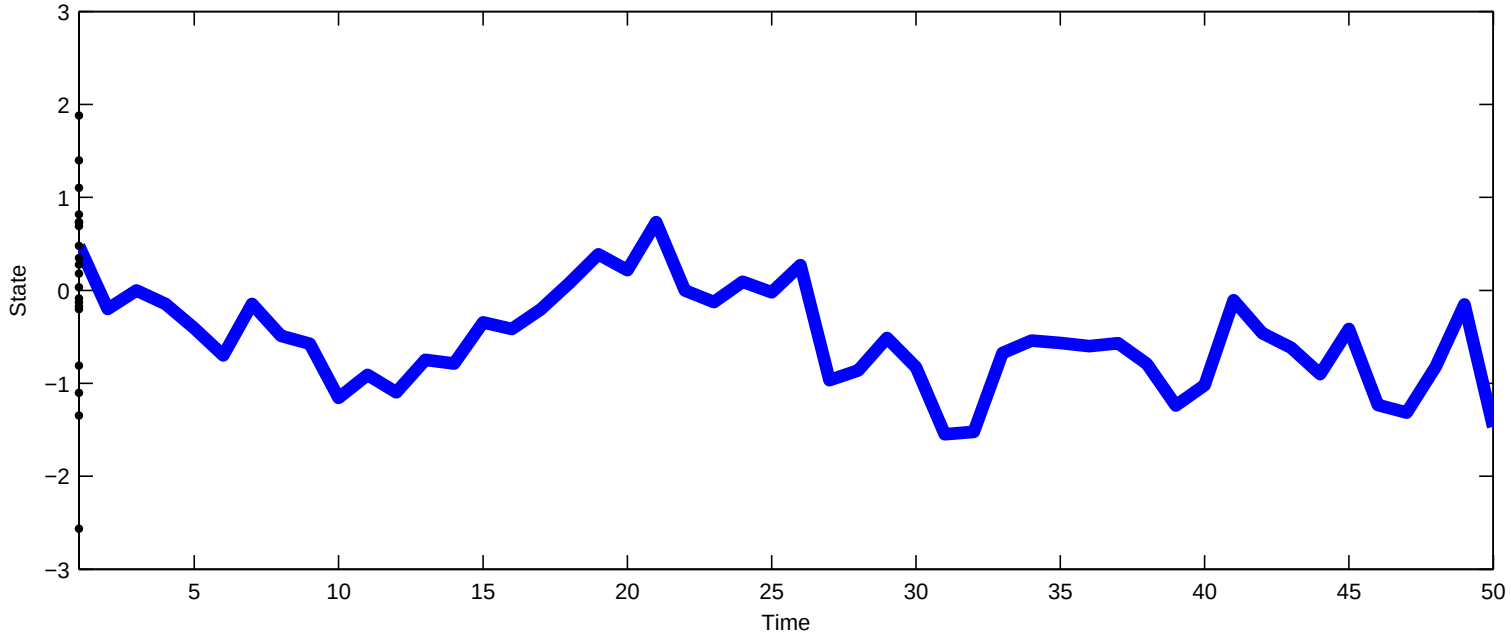
Particle Gibbs sampling

Let $\boldsymbol{x'_{1:T}} = (\boldsymbol{x'_1}, \ldots, \boldsymbol{x'_T})$ be a fixed *reference trajectory*.

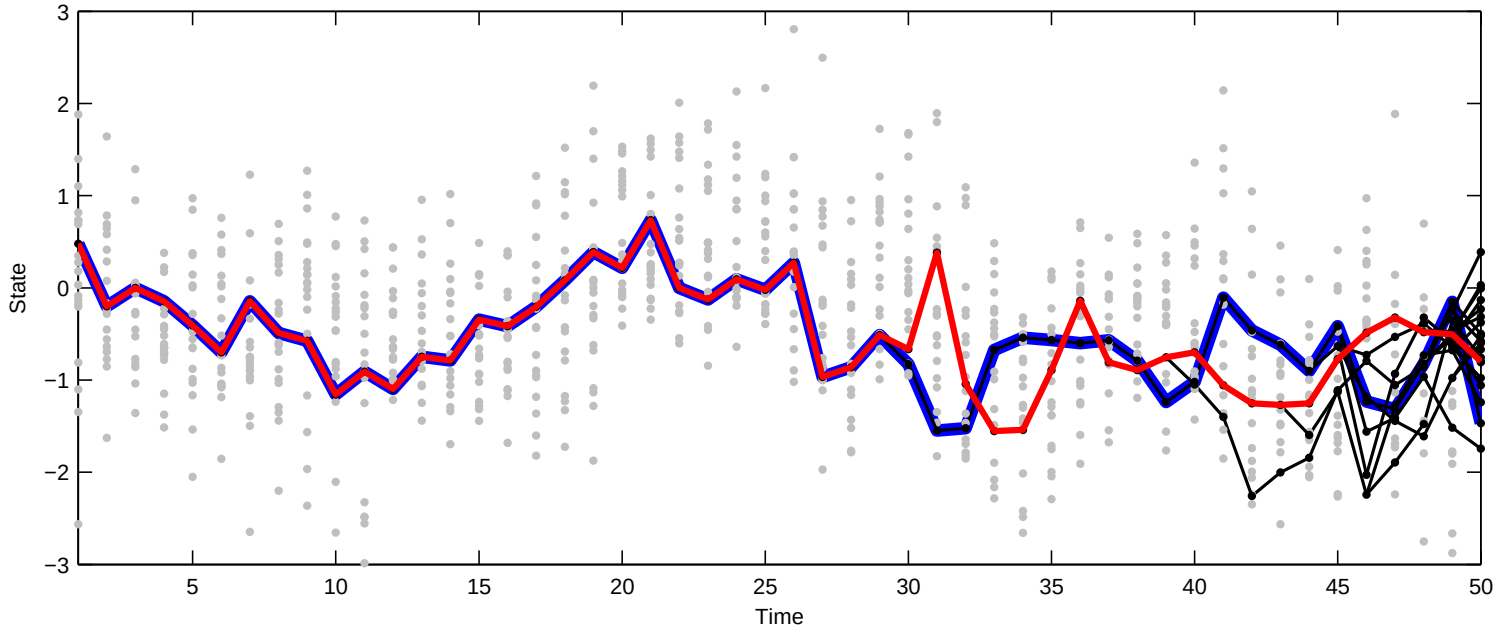
- Sample only $N-1$ particles in the standard way.
- Set the Nth particle deterministically: $x_t^N = \boldsymbol{x'_t}$ and $a_t^N = N$.

C. Andrieu, A. Doucet and R. Holenstein, Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B, 72:269-342, 2010.*

Particle Gibbs



Particle Gibbs



- Sampling $\boldsymbol{x}^{\star}_{1:T}$ with $\mathbb{P}(\boldsymbol{x}^{\star}_{1:T} = x^{i}_{1:T}) \propto w^{i}_{T}$ stochastically "maps" $\boldsymbol{x}'_{1:T}$ into $\boldsymbol{x}^{\star}_{1:T}$.
- Implicitly defines a Markov kernel (the PG kernel) on X^{T}, the *space of state trajectories*.

Particle Gibbs

Theorem (Stationary distribution)

The joint smoothing distribution $p(x_{1:T} \mid \theta, y_{1:T})$ is a stationary distribution of the PG kernel for any $N \geq 1$.

C. Andrieu, A. Doucet and R. Holenstein, Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B, 72:269-342, 2010.*

Theorem (Uniform ergodicity)

The PG kernel is uniformly ergodic under weak conditions.

Additionally, under strong mixing conditions the convergence rate does not degenerate as $T \to \infty$ provided that $N \propto T$.

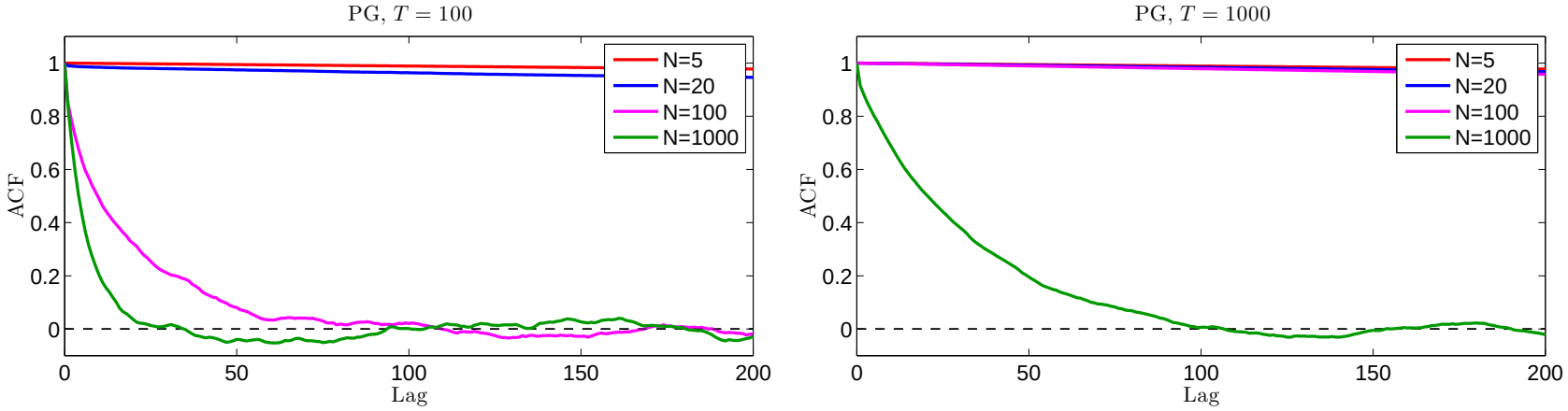
F. Lindsten, R. Douc, and E. Moulines, Uniform ergodicity of the Particle Gibbs sampler. *Scandinavian Journal of Statistics, 42(3): 775-797*, 2015.

ex) Particle Gibbs

Stochastic volatility model,

$$x_{t+1} = 0.9x_t + v_t, \qquad v_t \sim \mathcal{N}(0, \theta),$$
$$y_t = e_t \exp\left(\tfrac{1}{2}x_t\right), \qquad e_t \sim \mathcal{N}(0, 1).$$

Consider the ACF of $\theta[k] - \mathbb{E}[\theta \mid y_{1:T}]$.



Ancestor Sampling

Particle Gibbs:

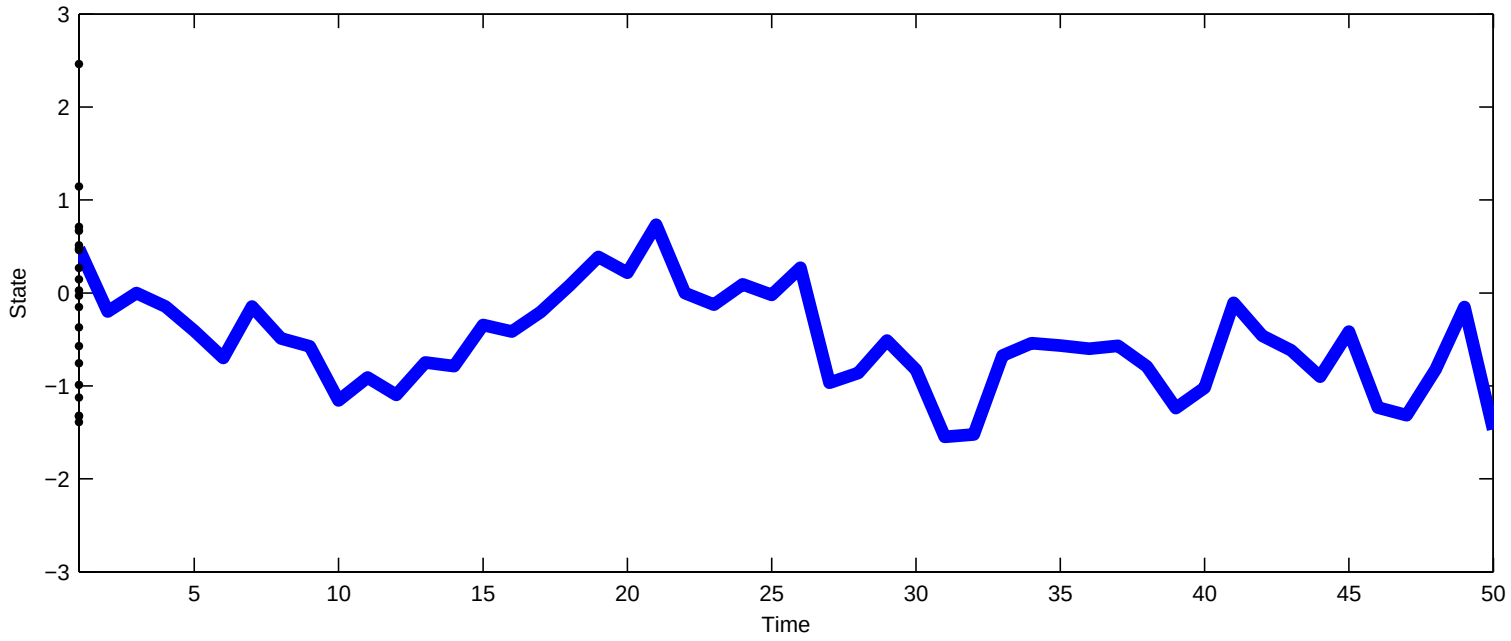
Let $\boldsymbol{x'_{1:T}} = (\boldsymbol{x'_1}, \ldots, \boldsymbol{x'_T})$ be a fixed *reference trajectory*.

- Sample only $N-1$ particles in the standard way.
- Set the Nth particle deterministically: $x_t^N = \boldsymbol{x'_t}$.
- ~~Set $a_t^N = N$~~.
- Sample $a_t^N \in \{1, \ldots, N\}$ with

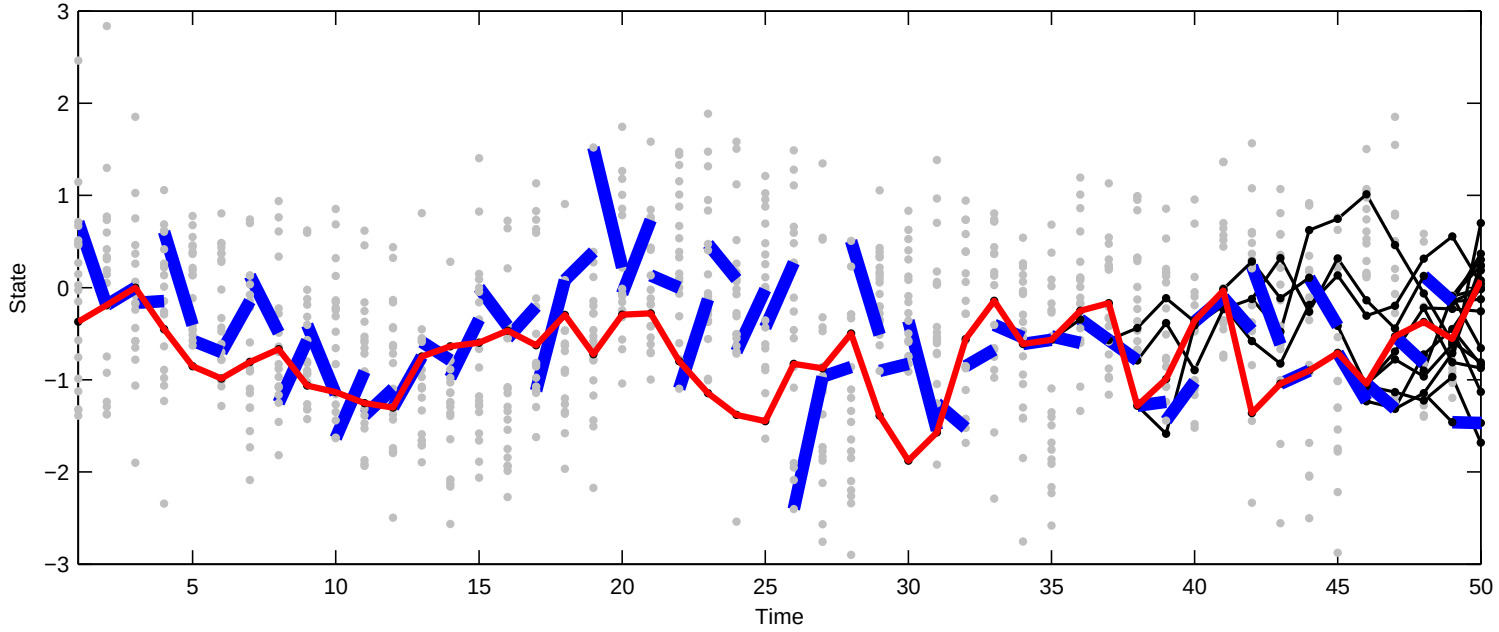
$$\mathbb{P}(a_t^N = j) \propto w_{t-1}^j f_\theta(\boldsymbol{x'_t} \mid x_{t-1}^j).$$

F. Lindsten, M. I. Jordan and T. B. Schön, Particle Gibbs with Ancestor sampling, *Journal of Machine Learning Research, 15: 2145-2184*, 2014.

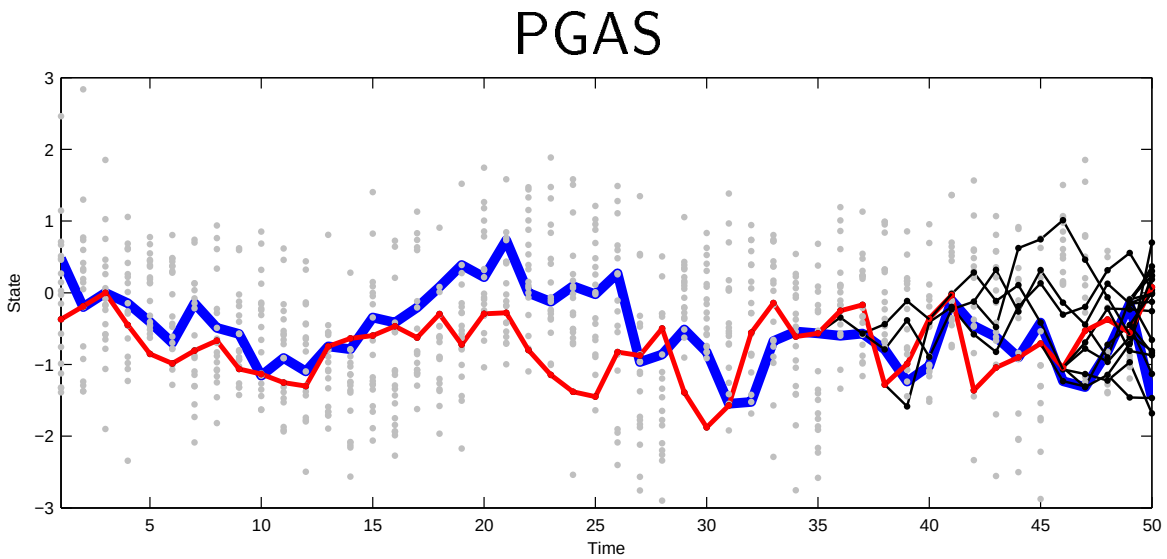
Particle Gibbs with Ancestor Sampling

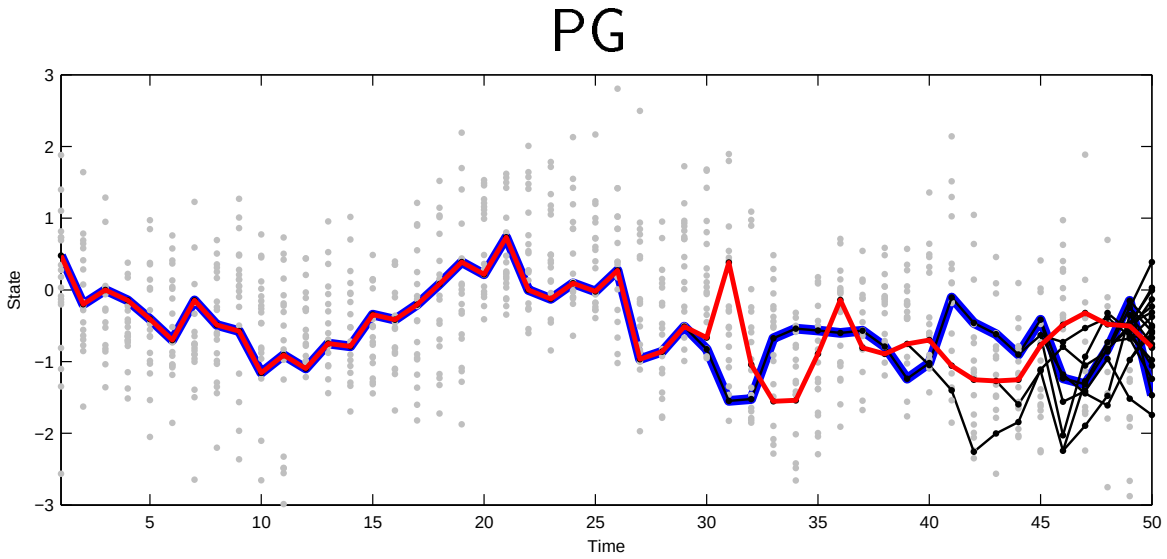


Particle Gibbs with Ancestor Sampling



PGAS vs. PG


PGAS
State
Time
3
2
1
0
−1
−2
−3
5
10
15
20
25
30
35
40
45
50

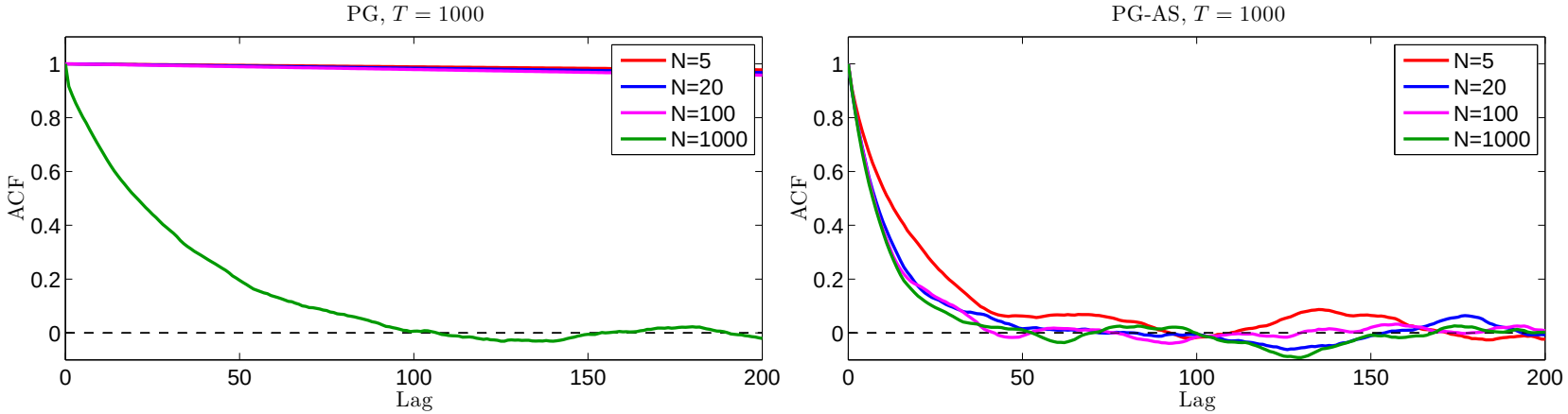

PG
State
Time
3
2
1
0
−1
−2
−3
5
10
15
20
25
30
35
40
45
50

ex) Stochastic volatility model, cont'd

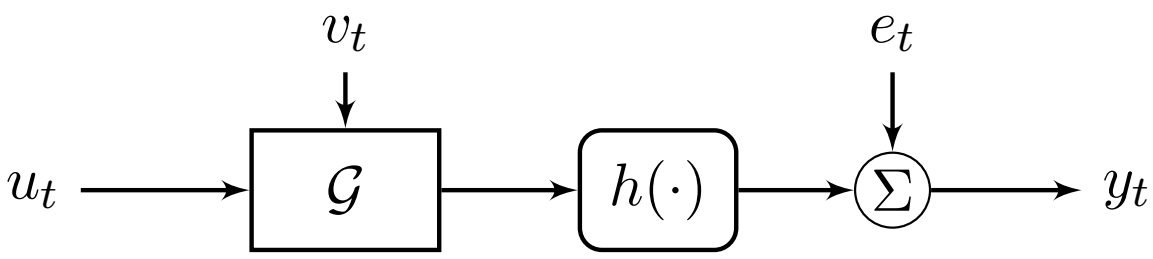
Stochastic volatility model,

$$x_{t+1} = 0.9x_t + v_t, \qquad v_t \sim \mathcal{N}(0, \theta),$$
$$y_t = e_t \exp\left(\tfrac{1}{2}x_t\right), \qquad e_t \sim \mathcal{N}(0, 1).$$

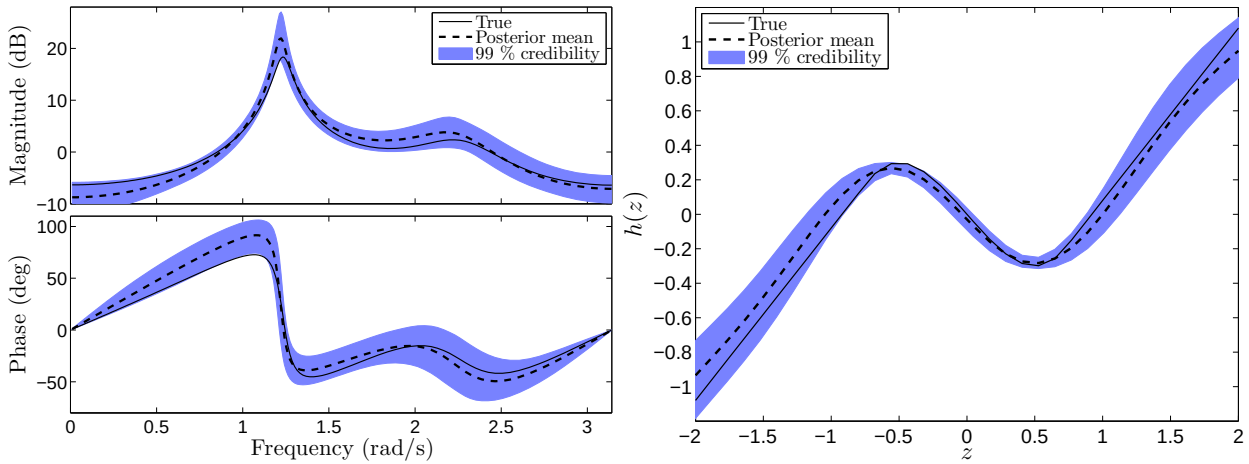
Consider the ACF of $\theta[k] - \mathbb{E}[\theta \mid y_{1:T}]$.



ex) Wiener system identification



Semi-parametric model: State-space model for $\mathcal{G}$, Gaussian process model for $h(\cdot)$.

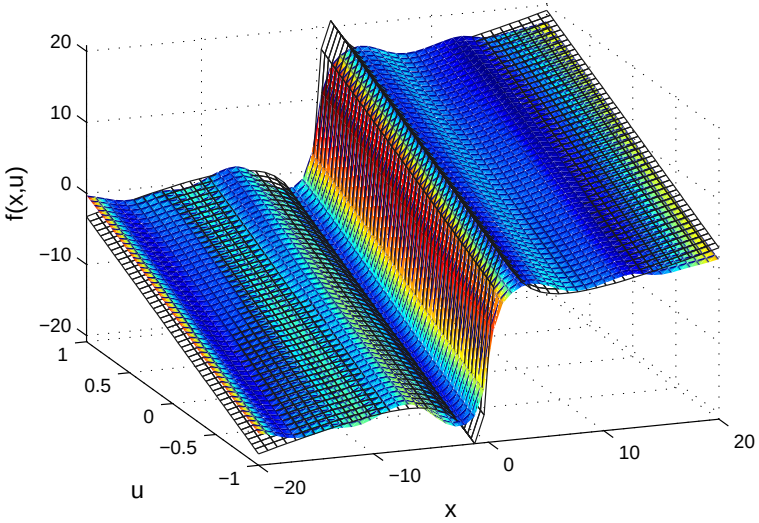


F. Lindsten, T. B. Schön and M. I. Jordan, Bayesian semiparametric Wiener system identification, *Automatica, 49(7): 2053-2063*, 2013.

ex) Gaussian process state-space models

A GP-SSM is a flexible nonparametric dynamical systems model,

$$
\begin{aligned}
f(\cdot) &\sim \mathcal{GP}\big(m_\theta(x_t), k_\theta(x_t, x_t')\big), \\
x_{t+1} \mid f, x_t &\sim \mathcal{N}(f(x_t), Q), \\
y_t \mid x_t &\sim g_\theta(y_t \mid x_t).
\end{aligned}
$$



Idea: Marginalize out $f(\cdot)$ and infer $x_{1:T}$ (and θ) directly.

Marginalization of f introduces non-Markovian dependencies in $\{x_t\}$.

R. Frigola, F. Lindsten, T. B. Schön and C. E. Rasmussen, Bayesian Inference and Learning in Gaussian Process State-Space Models with Particle MCMC, *Conference on Neural Information Processing Systems (NIPS)*, 2013.

ex) Epidemiological forecasting

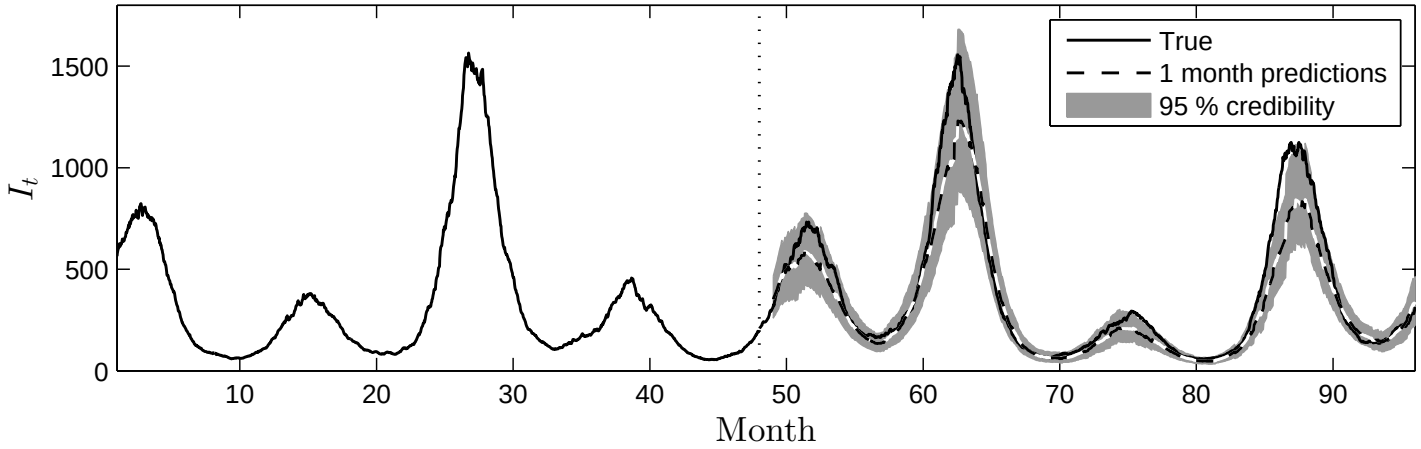


Figure: Disease activity over an eight year period. First half used for inference, second half for one-month predictions.

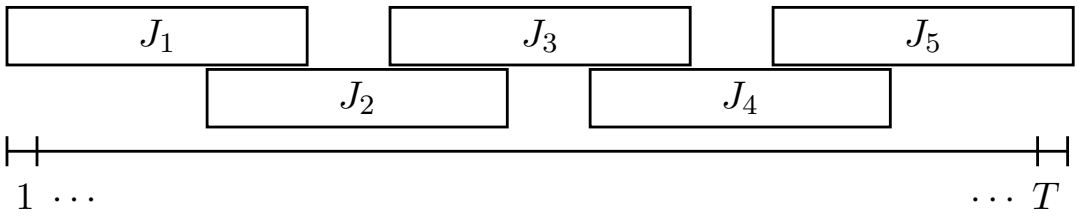
Inference (and prediction) for SIR model with environmental noise and seasonal fluctuations

$$
\begin{aligned}
S_{t+dt} &= S_t + \mu\mathcal{P}dt - \mu S_t dt - (1 + Fv_t)\,\beta_t S_t \mathcal{P}^{-1} I_t dt, \\
I_{t+dt} &= I_t - (\gamma + \mu) I_t dt + (1 + Fv_t)\,\beta_t S_t \mathcal{P}^{-1} I_t dt, \\
R_{t+dt} &= R_t + \gamma I_t dt - \mu R_t dt,
\end{aligned}
$$

Applications of PGAS

- Oceanography [Marcos et al., JGR, 2015]
- Econometrics [Nonejad, Economics Letters, 2015]
- Synaptic plasticity models [Linderman et al., NIPS, 2014]
- Probabilistic programming [Meent et al., AISTATS, 2015]
- Power disaggregation [Valera et al., NIPS, 2015]
- ...

Ongoing: Blocking



- Update *blocks of states* using PG(AS).
- Stable as $T \to \infty$ using fixed N.
- Opens up for parallelisation.
- Possible to do apply blocking also over components of the state vector (spatial blocking).

S. S. Singh, F. Lindsten, and E. Moulines, Blocking Strategies and Stability of Particle Gibbs Samplers, *arXiv preprint*, *arXiv:1509.08362*, 2015.

Ongoing: Spatio-temporal models

Spatio-temporal model:

- State x_t spatially structured
- High dimensionality prevents conventional methods!
- Using multiple *Nested Particle Filters* we can handle dimensions $\sim 100 - 1\,000$.

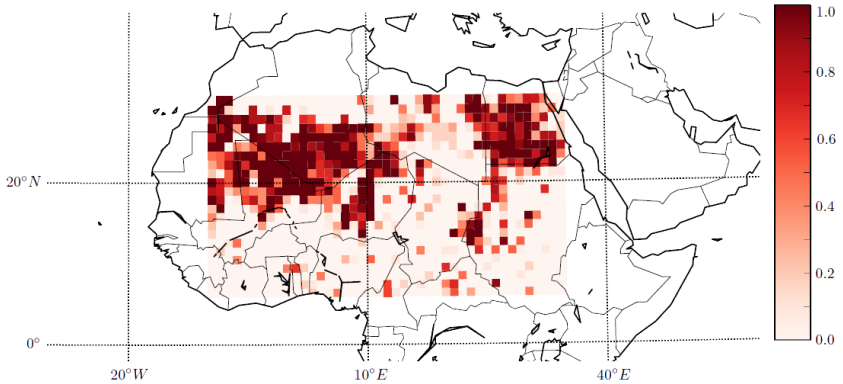
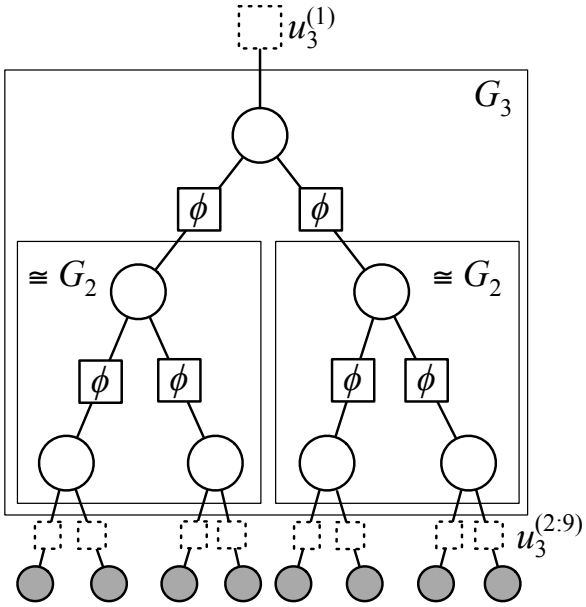


Figure: Drought probability in Sahel region in 1989.

C. A. Naesseth, F. Lindsten, and T. B. Schön, Nested Sequential Monte Carlo Methods. *International Conference on Machine Learning (ICML)*, 2015.

Ongoing: Probabilistic graphical models



- Solving *static* problems using *sequential* methods.
- Divide-and-Conquer PF – extending particle filtering from *chains* to *trees*
- In particular: *hierarchical Bayesian models*
- Applications in genetics, linguistics, social science, . . .

F. Lindsten, A. M. Johansen, C. A. Naesseth, B. Kirkpatrick, T. B. Schön, J. Aston, and A. Bouchard-Côté, Divide-and-Conquer with Sequential Monte Carlo, *arXiv preprint*, *arXiv:1406.4993v2*, 2015.

S. Linderman, C. H. Stock, and R. P. Adams.
A framework for studying synaptic plasticity with neural spike train data.
In *Advances in Neural Information Processing Systems (NIPS) 27*. 2014.

M. Marcos, F. M. Calafat, A. Berihuete, and S. Dangendorf.
Long-term variations in global sea level extremes.
Journal of Geophysical Research, 120(12):8115–8134, 2015.

N. Nonejad.
Flexible model comparison of unobserved components models using particle Gibbs with ancestor sampling.
Economics Letters, 133:35–39, 2015.

I. Valera, F. Ruiz, L. Svensson, and F. Perez-Cruz.
Infinite factorial dynamical model.
In *Advances in Neural Information Processing Systems (NIPS) 28*. 2015.

J.-W. van de Meent, Y. Hongseok, V. Mansinghka, and F. Wood.
Particle Gibbs with ancestor sampling for probabilistic programs.
In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.