Quasi-Newton particle Metropolis-Hastings

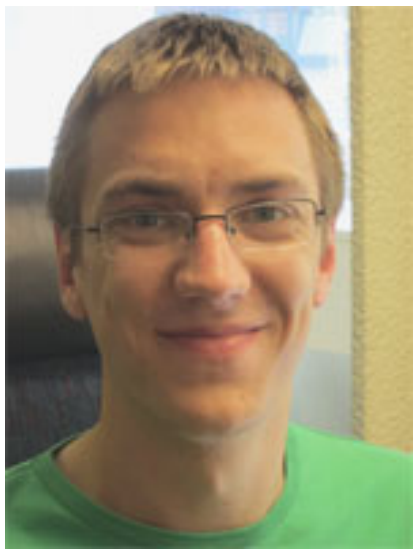
Johan Dahlin
johan.dahlin@liu.se
work.johandahlin.com

Division of Automatic Control,
Linköping University, Sweden.

LINKÖPING UNIVERSITY

This is collaborative work together with

Dr. Fredrik Lindsten (University of Cambridge, United Kingdom).
Prof. Thomas Schön (Uppsala University, Sweden).



Why are we doing this?

Bayesian parameter inference for general SSMs.
Accelerate existing posterior sampling methods.
Make it easier for the user to tune existing methods.

Why are we doing this?

Bayesian parameter inference for general SSMs.
Accelerate existing posterior sampling methods.
Make it easier for the user to tune existing methods.

How will we do this?

Construct Markov chain using quasi-Newton ideas.
Use this chain to explore the parameter posterior.
Approximate the resulting algorithm using particles.

Why are we doing this?

Bayesian parameter inference for general SSMs.
Accelerate existing posterior sampling methods.
Make it easier for the user to tune existing methods.

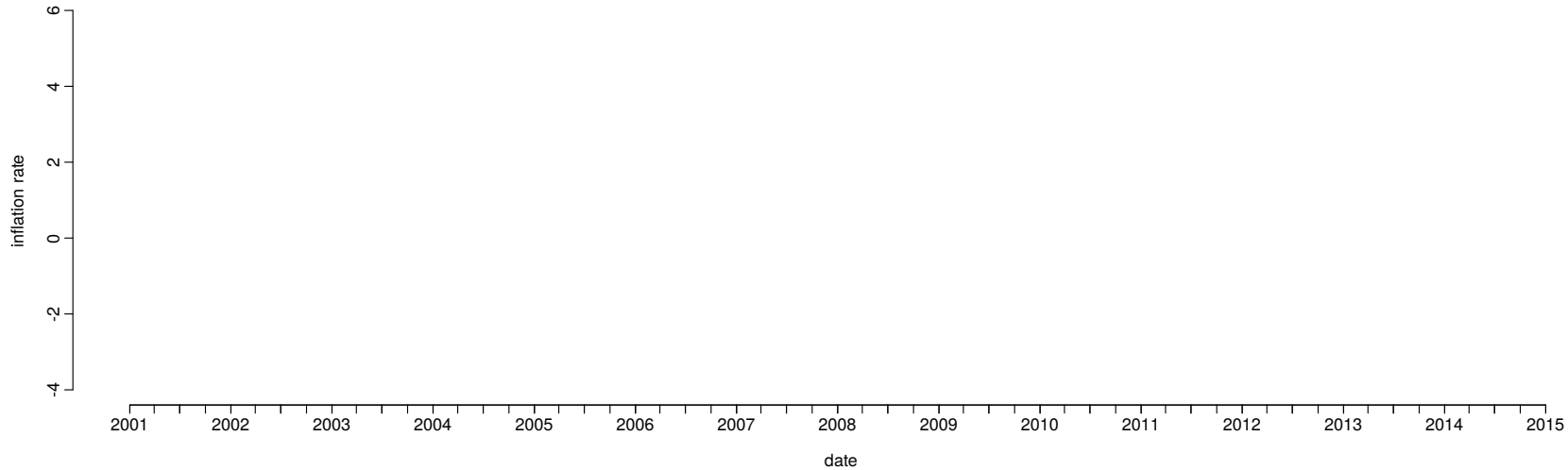
How will we do this?

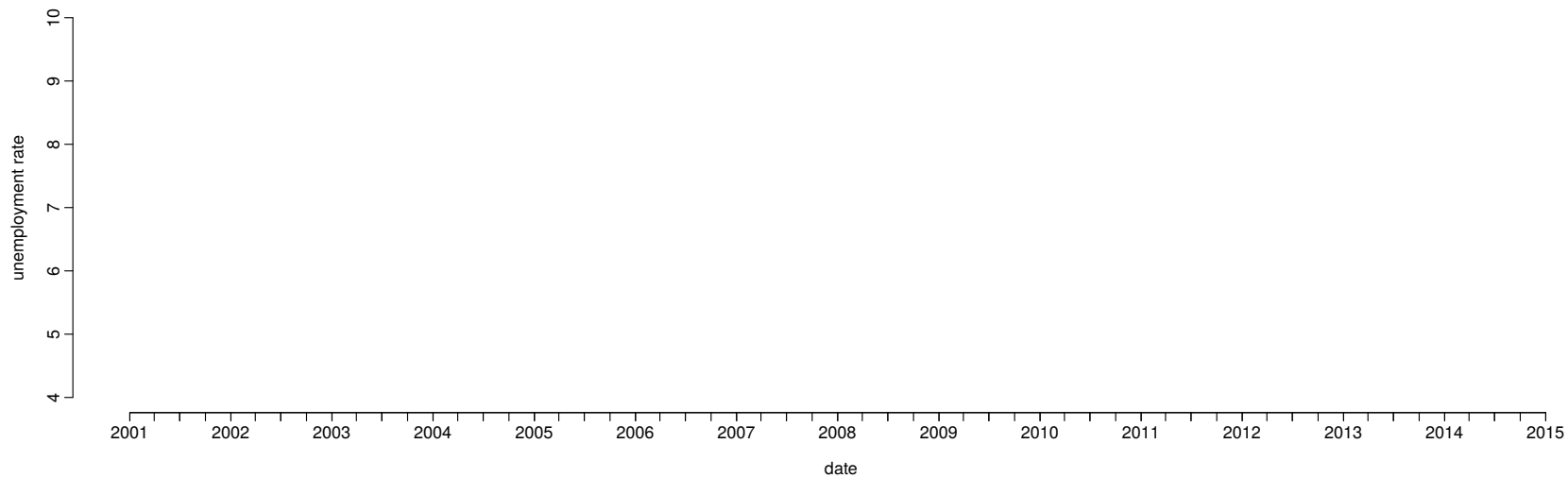
Construct Markov chain using quasi-Newton ideas.
Use this chain to explore the parameter posterior.
Approximate the resulting algorithm using particles.

What are we going to do?

Construct the algorithm step by step.
Benchmark the speed-up of the proposed algorithm.
Briefly discuss the required user interaction.

Inflation and unemployment rates in Sweden





Example: nonlinear Philips curve model

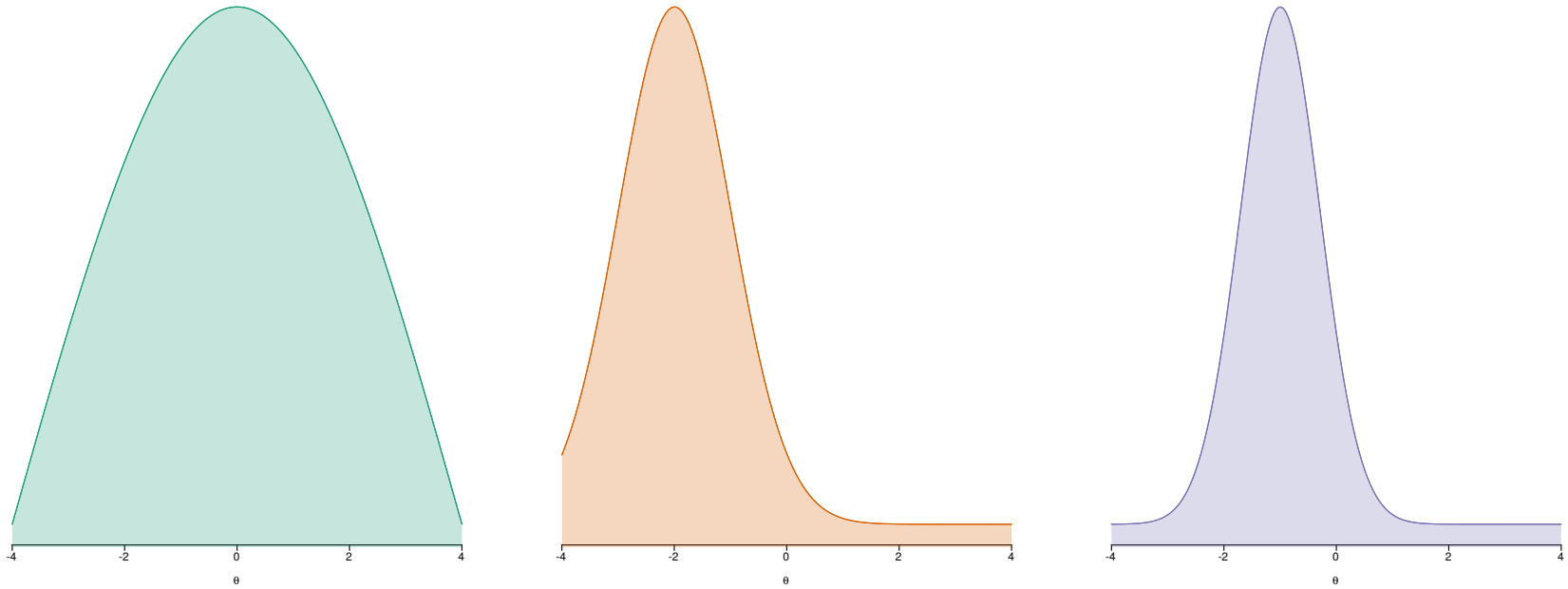
$$x_t = \phi x_{t-1} + \frac{\alpha}{1 + \exp\left(-x_{t-1}\right)} + \left[\frac{1}{1 + \exp\left(-\left|u_{t-1} - x_{t-1}\right|\right)}\right] v_t,$$

$$y_t = y_{t-1} + \beta\left(u_t - x_t\right) + \sigma_e e_t,$$

with $\theta = [\phi, \alpha, \beta, \sigma_e]$, $v_t, e_t \overset{\text{IID}}{\sim} \mathcal{N}(0, 1)$ and

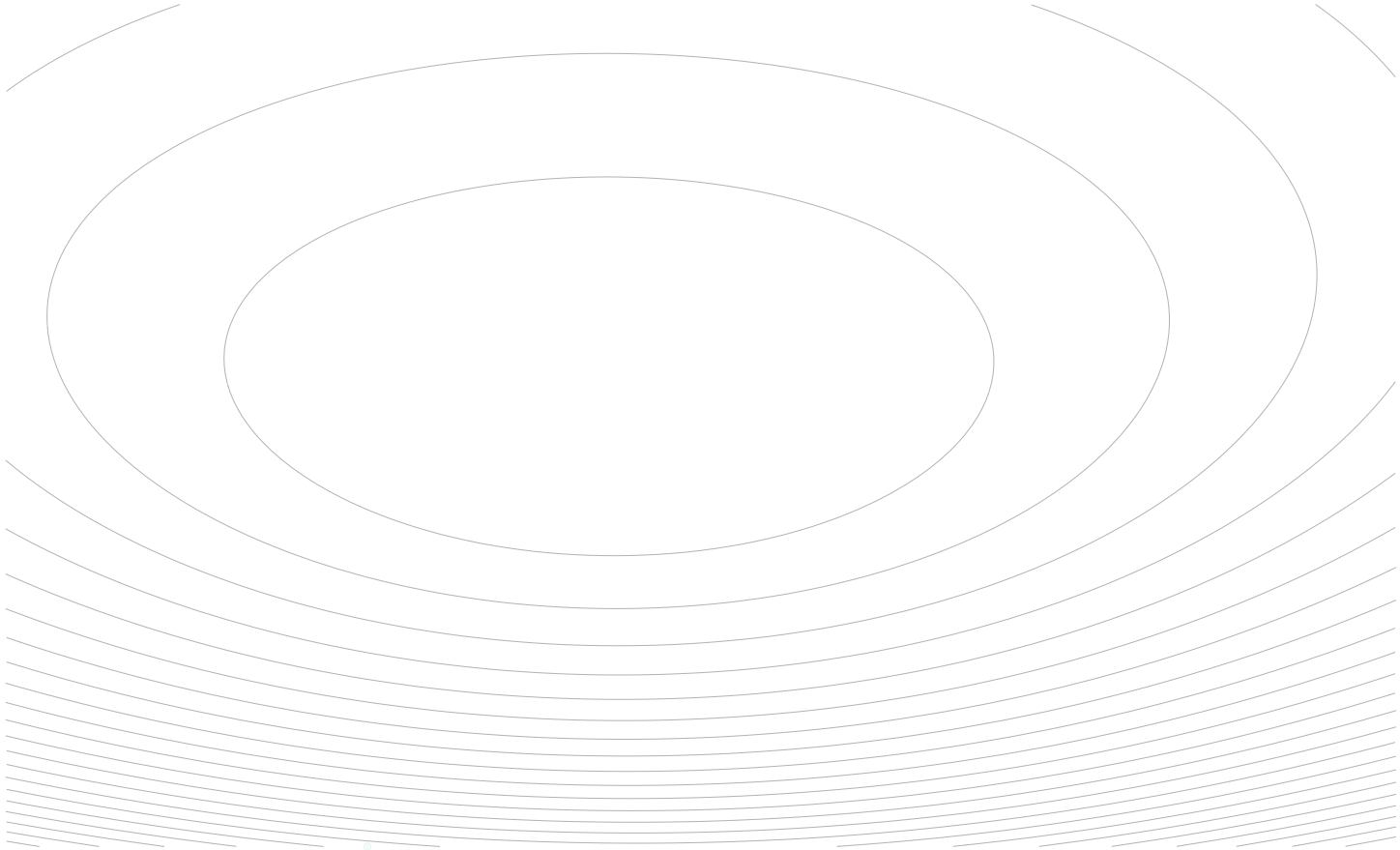
- y_t Inflation rate.
- u_t Unemployment rate.
- x_t Structural unemployment.

Bayesian parameter inference



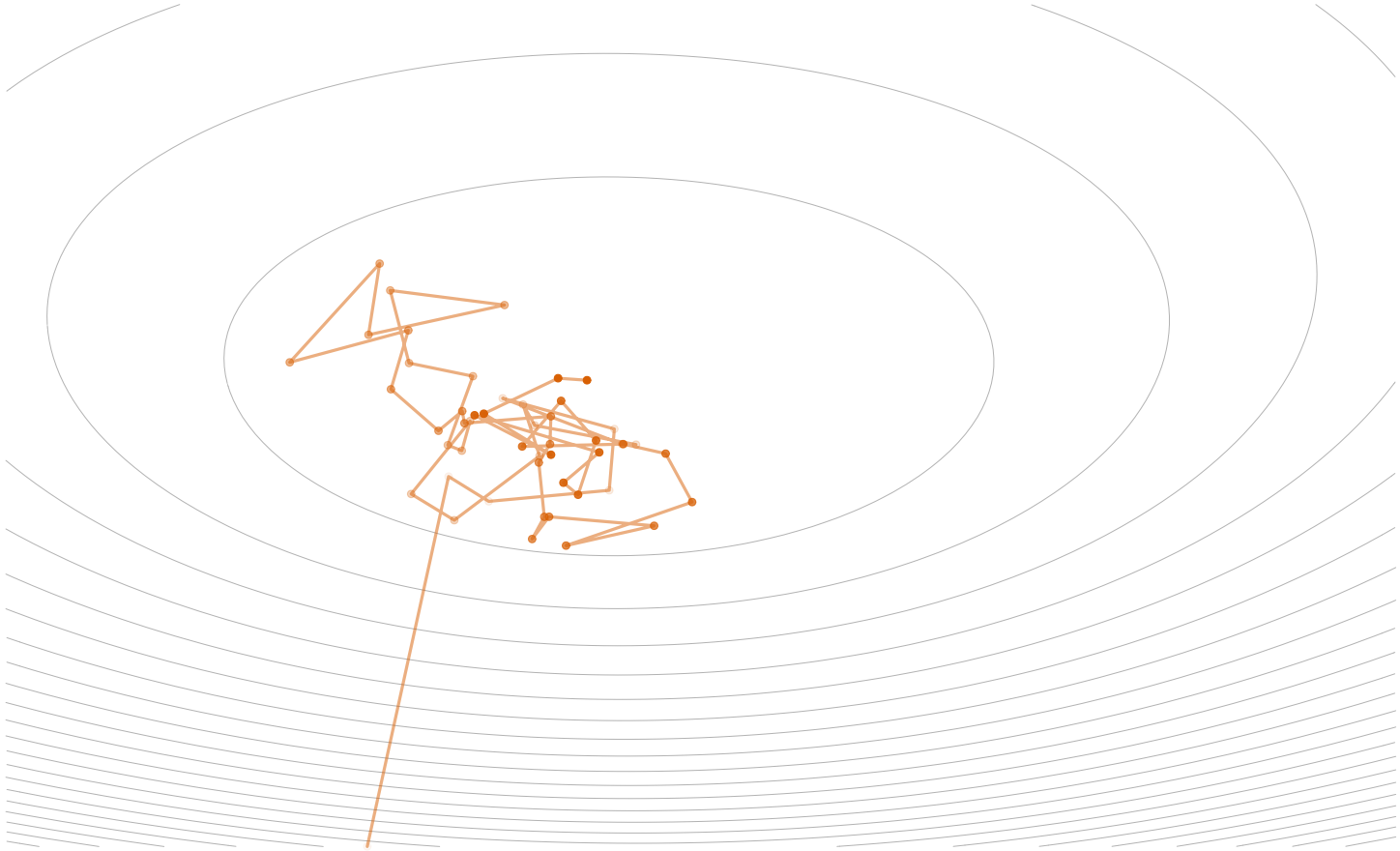
$$\pi(\theta) = p(y_{1:T}|\theta) \propto p_\theta(y_{1:T})p(\theta), \quad \pi[\varphi] = \mathbb{E}_\pi\left[\varphi(\theta)\right] = \int \varphi(\theta')\pi(\mathrm{d}\theta').$$

Gaussian random walk



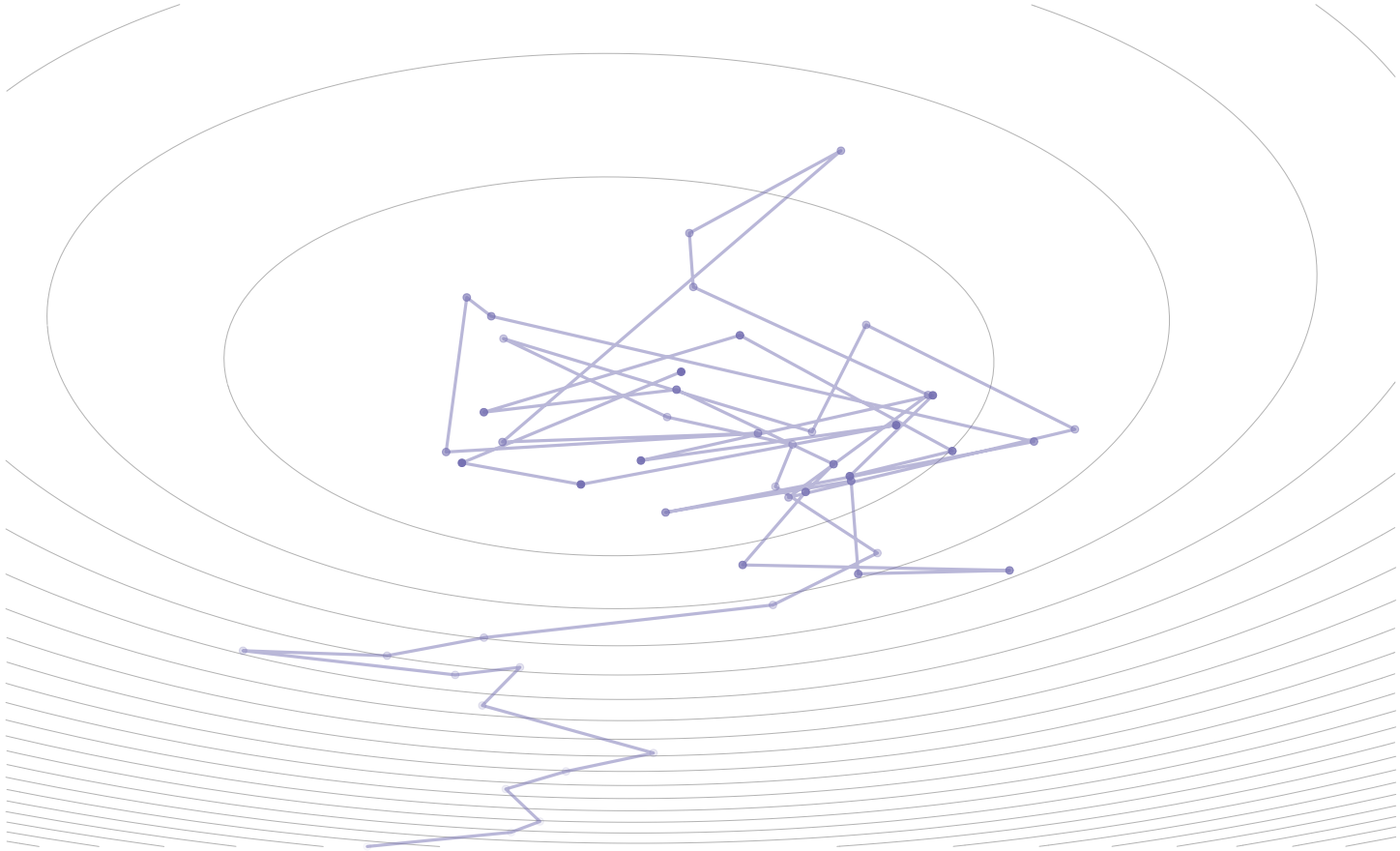
$$\theta_k|\theta_{k-1} \sim \mathcal{N}\left(\theta_k; \theta_{k-1}, \epsilon^2\mathcal{P}\right), \quad \mathcal{P} \text{ unknown pre-conditioning matrix.}$$

Noisy gradient ascent update



$$\theta_k|\theta_{k-1} \sim \mathcal{N}\left(\theta_k; \theta_{k-1} + \frac{\epsilon^2}{2}\mathcal{P}\mathcal{G}(\theta_{k-1}), \epsilon^2\mathcal{P}\right).$$

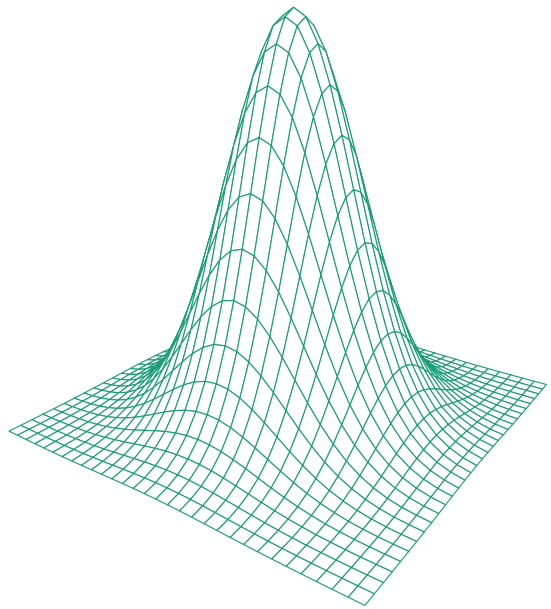
Noisy Newton update



$$\theta_k|\theta_{k-1} \sim \mathcal{N}\left(\theta_k; \theta_{k-1} + \frac{1}{2}\mathcal{H}^{-1}(\theta_{k-1})\mathcal{G}(\theta_{k-1}), \mathcal{H}^{-1}(\theta_{k-1})\right).$$

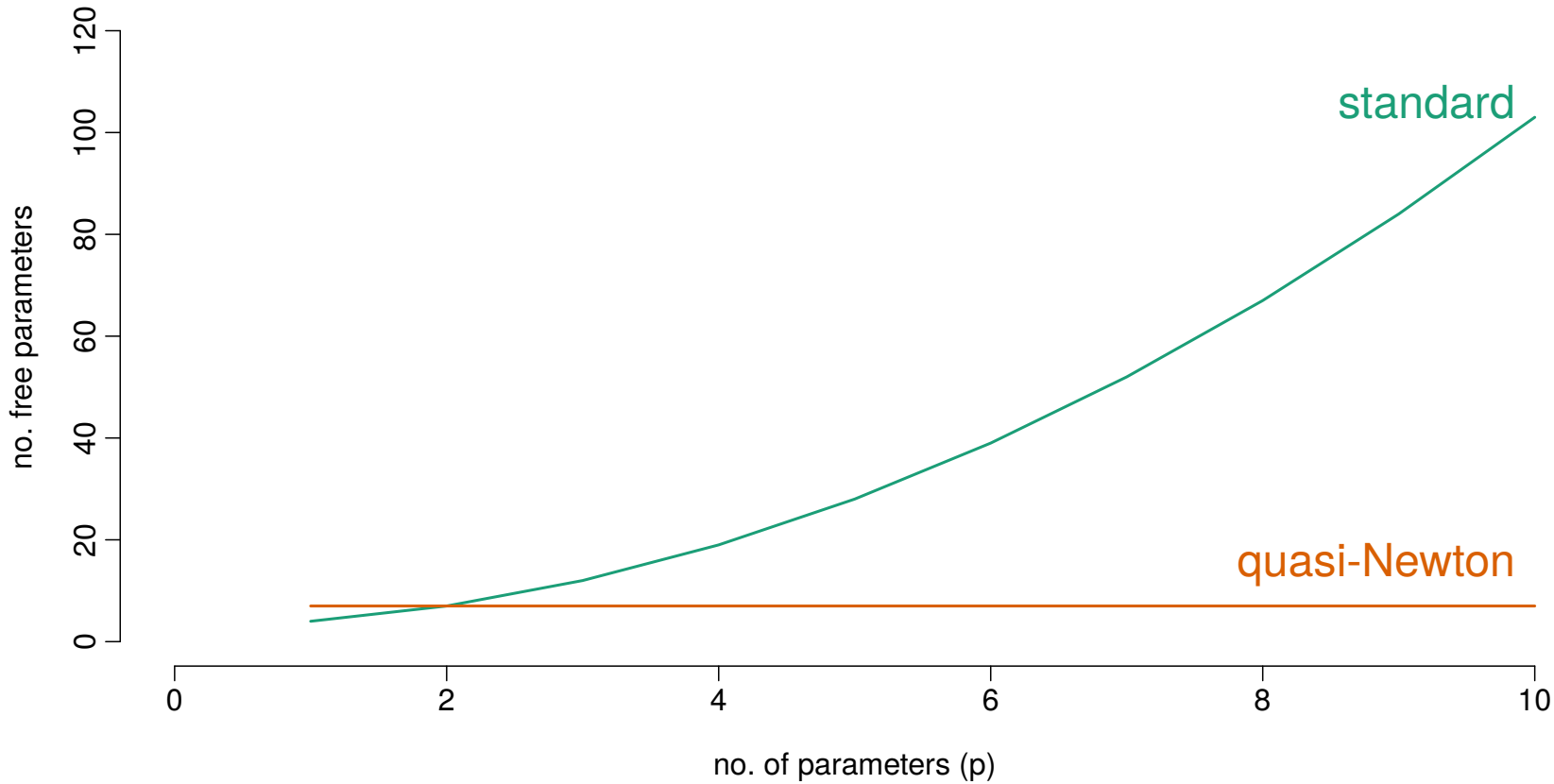
Noisy quasi-Newton update [I/II]

- Standard BFGS recursions to compute $\widehat{\mathcal{H}}^{-1}(\theta_{k-1})$.
- Makes use of $\theta_{k-M:k-1}$ and $\mathcal{G}(\theta_{k-M:k-1})$.
- Introduces a memory of length M into the Markov chain.



Noisy quasi-Newton update [II/II]

- Standard approach is use pilot runs to tune ϵ and $\mathcal{P}$.
- Number of tuning parameters.



Metropolis-Hastings for sampling from $\pi(\theta)$

(i) Propose candidate parameter: $\vartheta|\theta_{k-1} \sim q(\vartheta|\theta_{k-1})$.

(ii) Accept candidate parameter ϑ w.p.

$$1 \wedge \frac{\pi(\vartheta)}{\pi(\theta_{k-1})} \frac{q(\theta_{k-1}|\vartheta)}{q(\vartheta|\theta_{k-1})},$$

i.e. $\theta_k \leftarrow \vartheta$ and otherwise $\theta_k \leftarrow \theta_{k-1}$.

We can estimate $\pi[\varphi]$ using $[\theta_k]_{k=1}^K$ correlated samples obeying

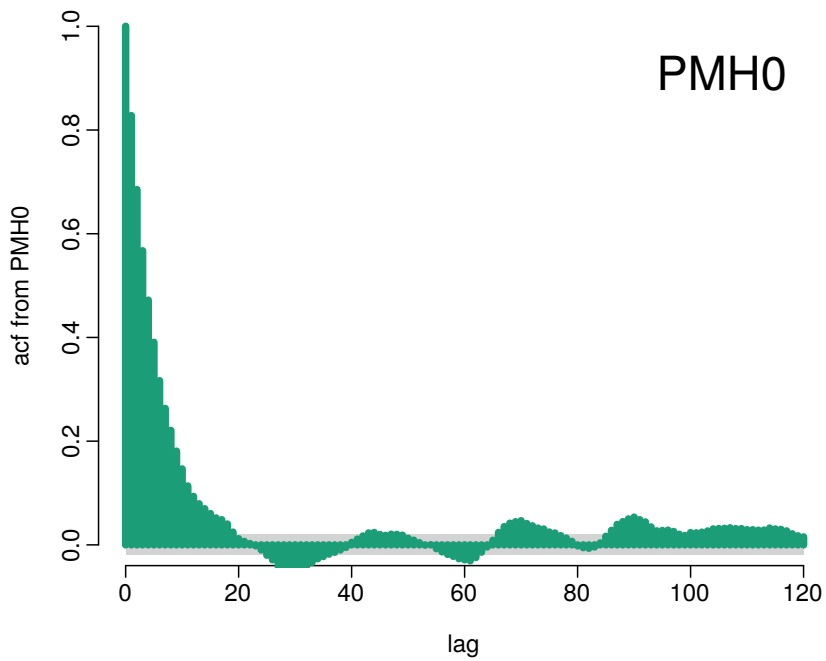
$$\sqrt{K}\Bigg[\underbrace{\frac{1}{K}\sum_{k=1}^{K}\varphi(\theta_k)}_{\triangleq\widehat{\pi}[\varphi]} - \pi[\varphi]\Bigg] \xrightarrow{\mathsf{d}} \mathcal{N}\Big(0, \mathbb{V}_\pi[\varphi] \cdot \mathsf{IACT}(\theta_{1:K})\Big)$$

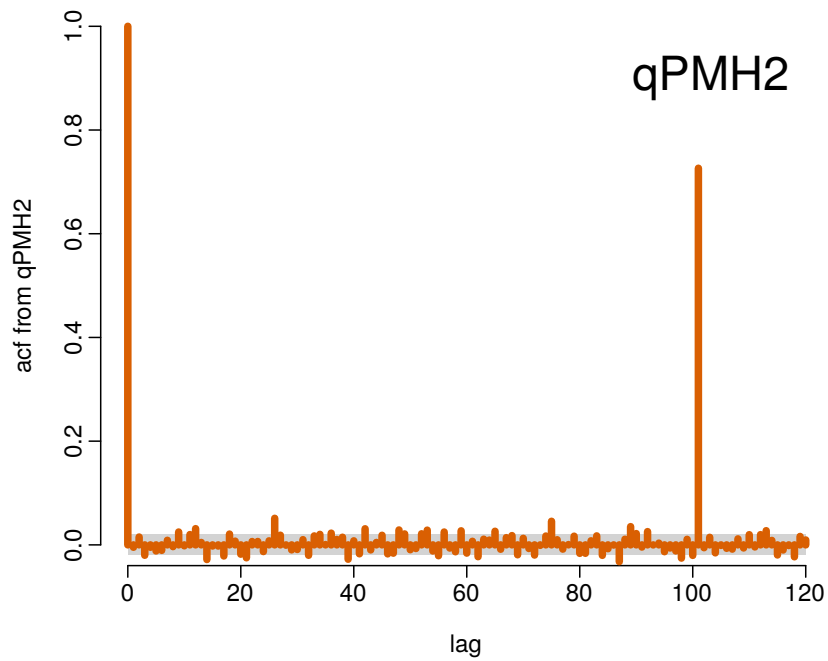
A non-standard Metropolis-Hastings algorithm

- The analysis is done on an M-fold product target

$$\pi(\theta_{k,1:M}) = \prod_{i=1}^{M} \pi(\theta_{k,i}).$$

- Proposal is centred around θ_{k-M} and reject gives $\theta_k \leftarrow \theta_{k-M}$.

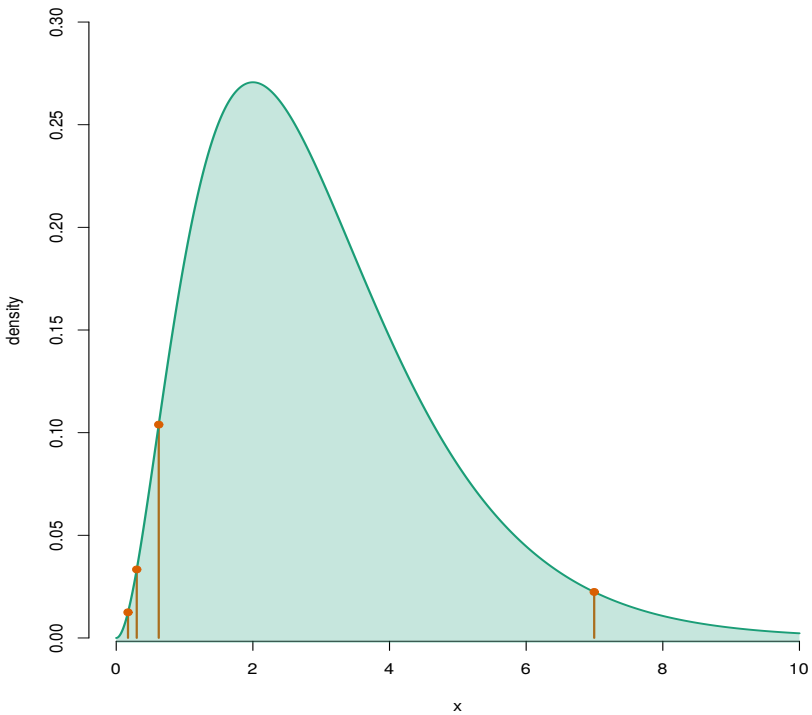


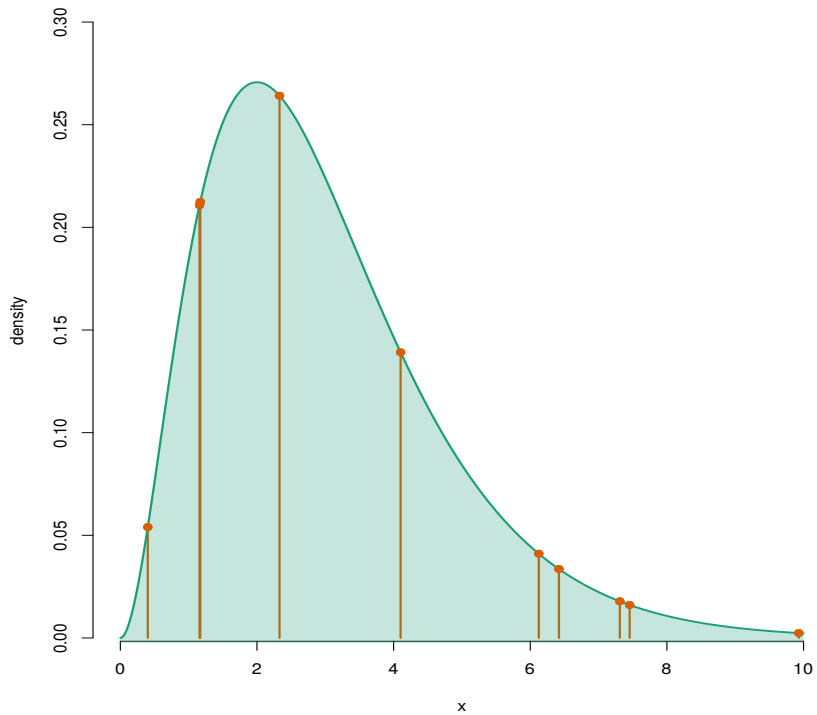


Particle Metropolis-Hastings

A particle filter/smoother is applied to estimate:

$\pi(\vartheta)$ The posterior at ϑ (for the acceptance probability).
$\mathcal{G}(\vartheta)$ The gradient of $\log \pi(\vartheta)$.
$\mathcal{H}(\vartheta)$ The Hessian of $\log \pi(\vartheta)$. (difficult)





Example: linear Gaussian model [I/IV]

We are comparing:

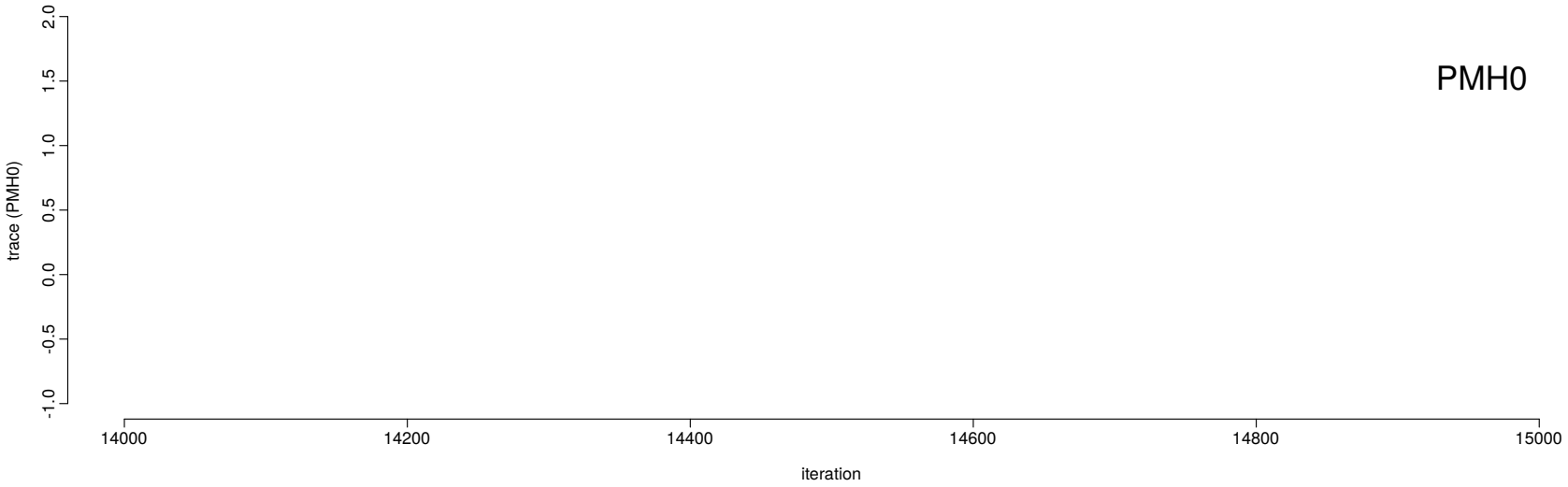
PMH0: Particle Metropolis-Hastings with random walk.
qPMH2: Particle Metropolis-Hastings with quasi-Newton.

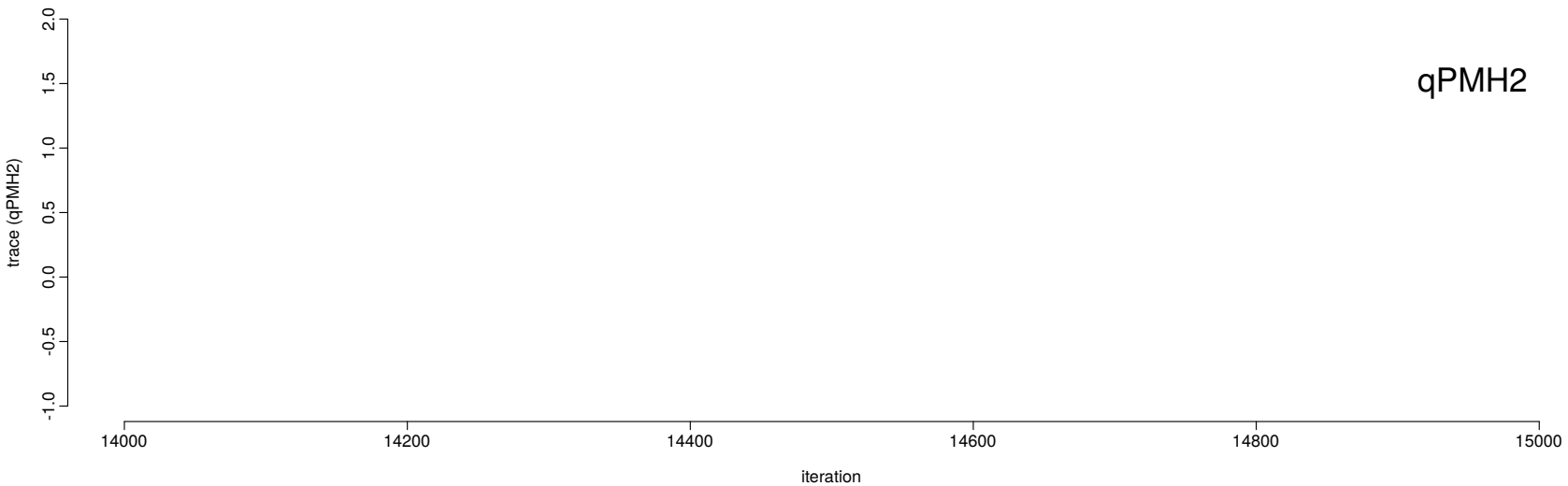
in the model given by

$$x_t = \boldsymbol{\mu} + \boldsymbol{\phi}(x_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\sigma_v} v_t,$$
$$y_t = x_t + 0.10 e_t,$$

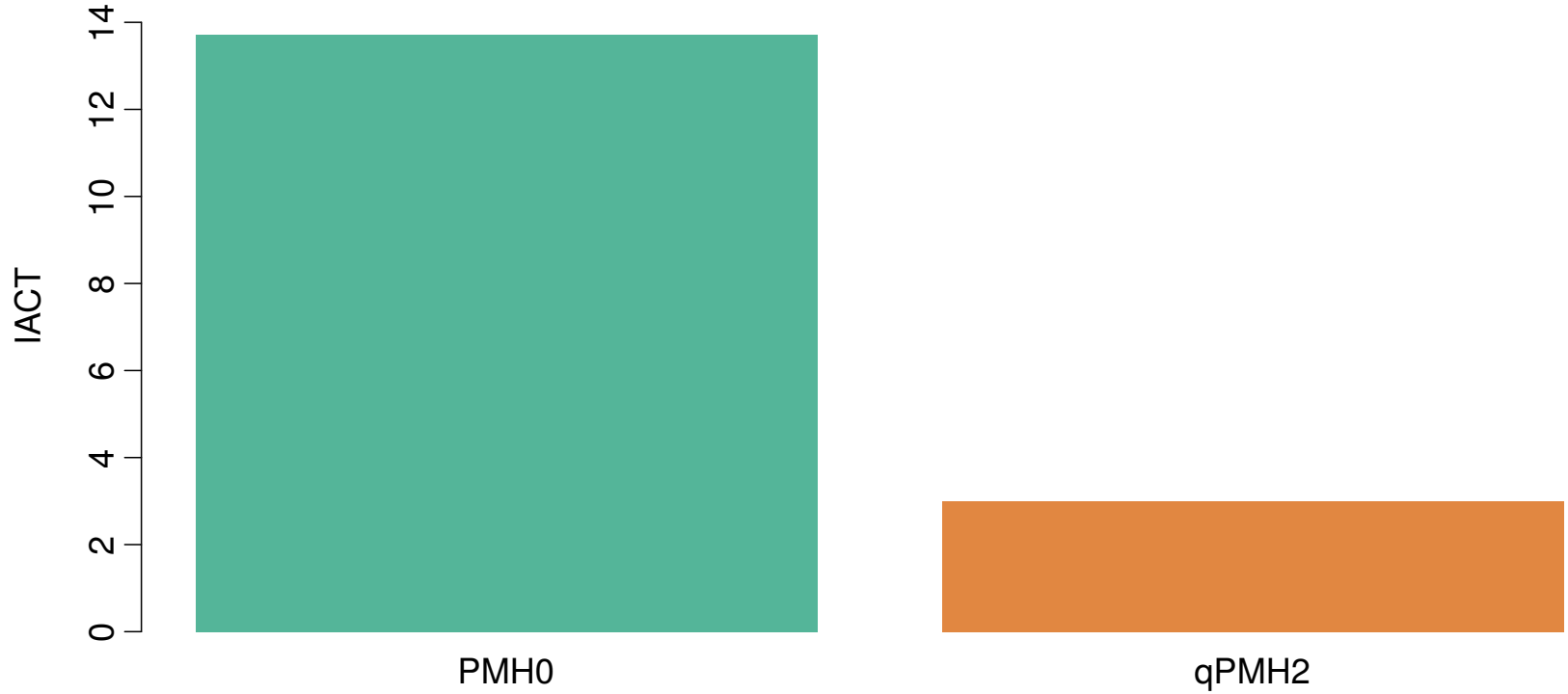
with $\theta = [\mu, \phi, \sigma_v] = [0.20, 0.80, 1.0]$ and $v_t, e_t \overset{\text{IID}}{\sim} \mathcal{N}(0, 1)$.

Example: linear Gaussian model [II/IV]

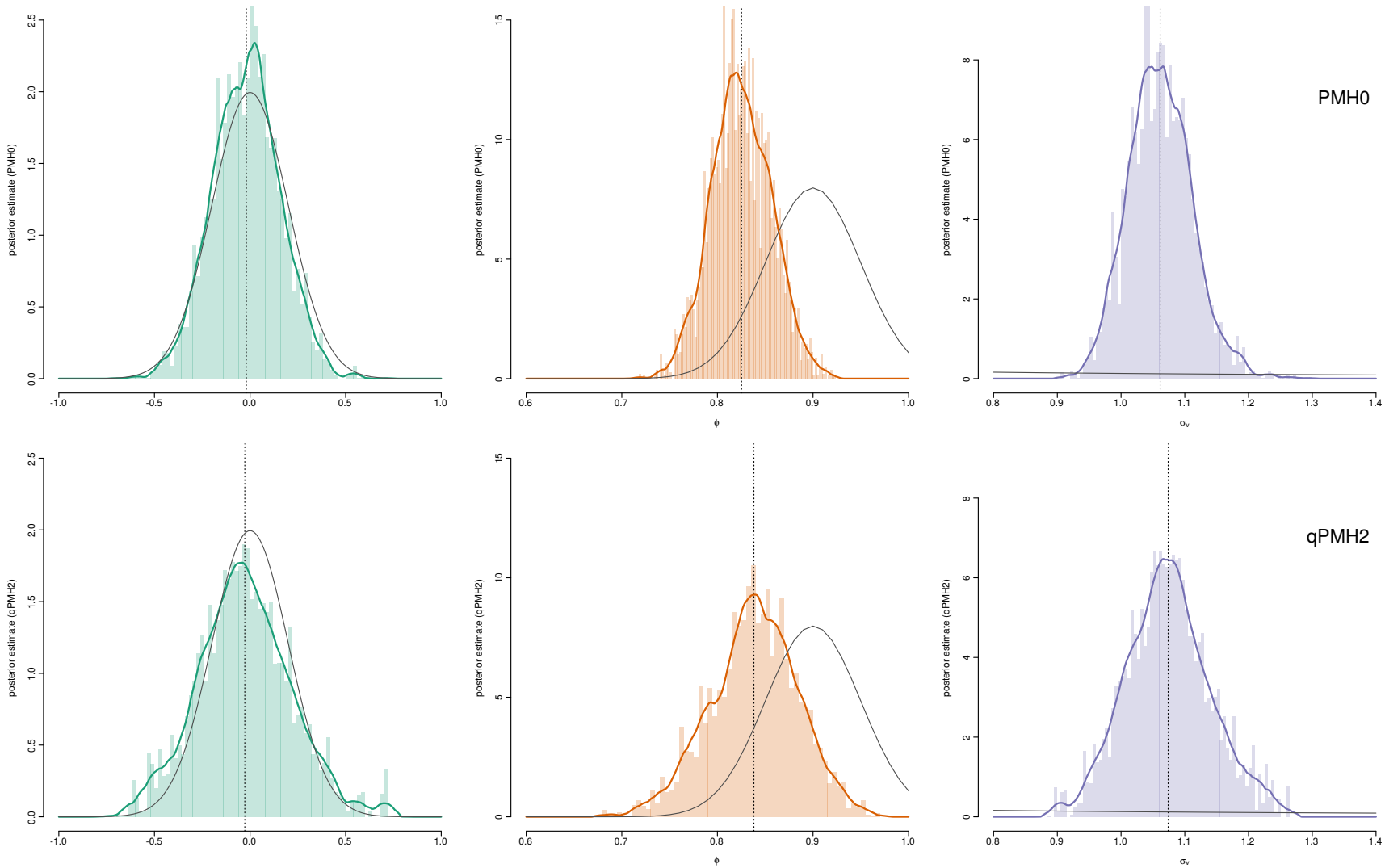




Example: linear Gaussian model [III/IV]



Example: linear Gaussian model [IV/IV]



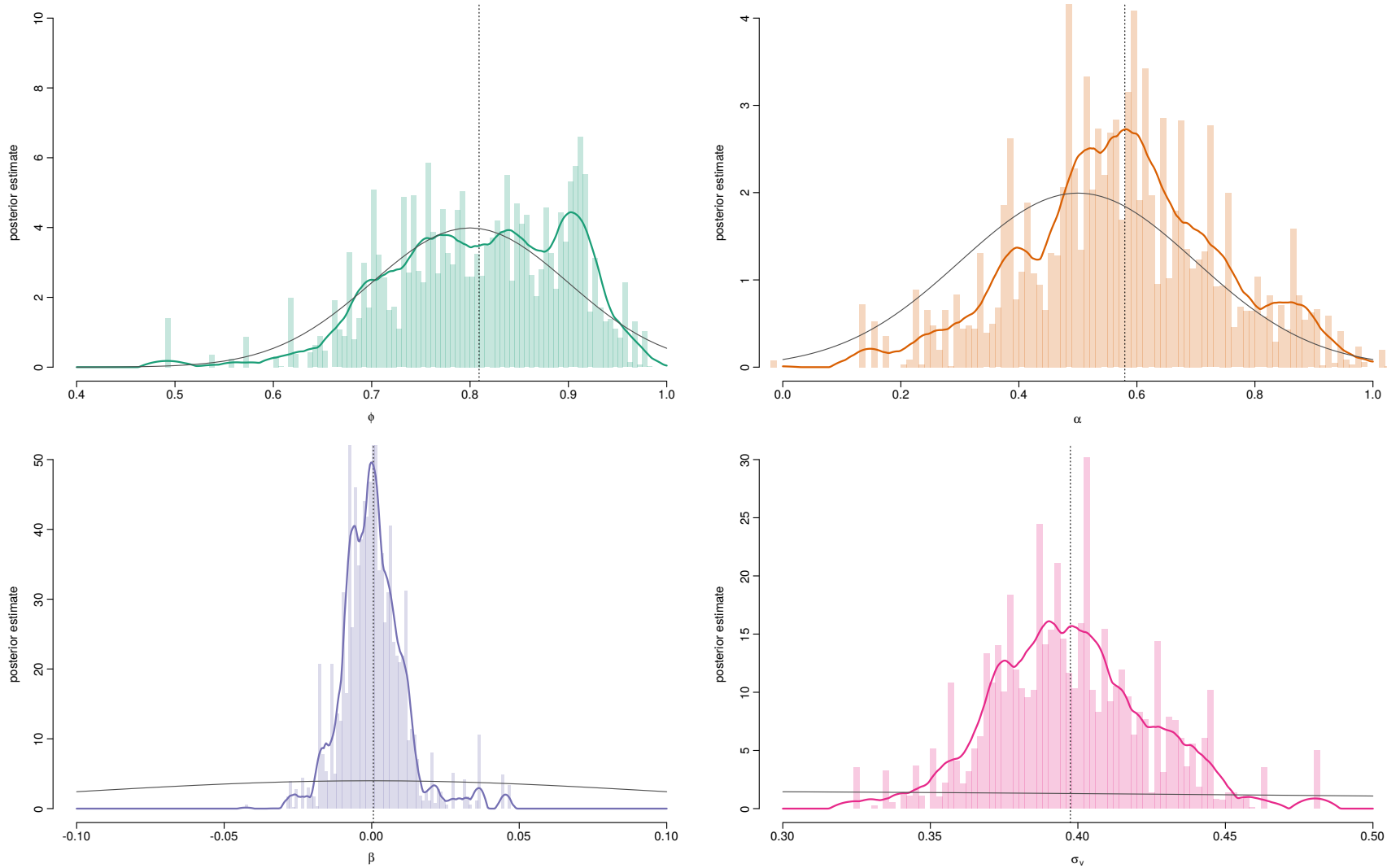
Example: nonlinear Philips curve model [I/III]

$$x_t = \phi x_{t-1} + \frac{\alpha}{1 + \exp\big(-x_{t-1}\big)} + \left[\frac{1}{1 + \exp\Big(-\big|u_{t-1} - x_{t-1}\big|\Big)}\right] v_t,$$

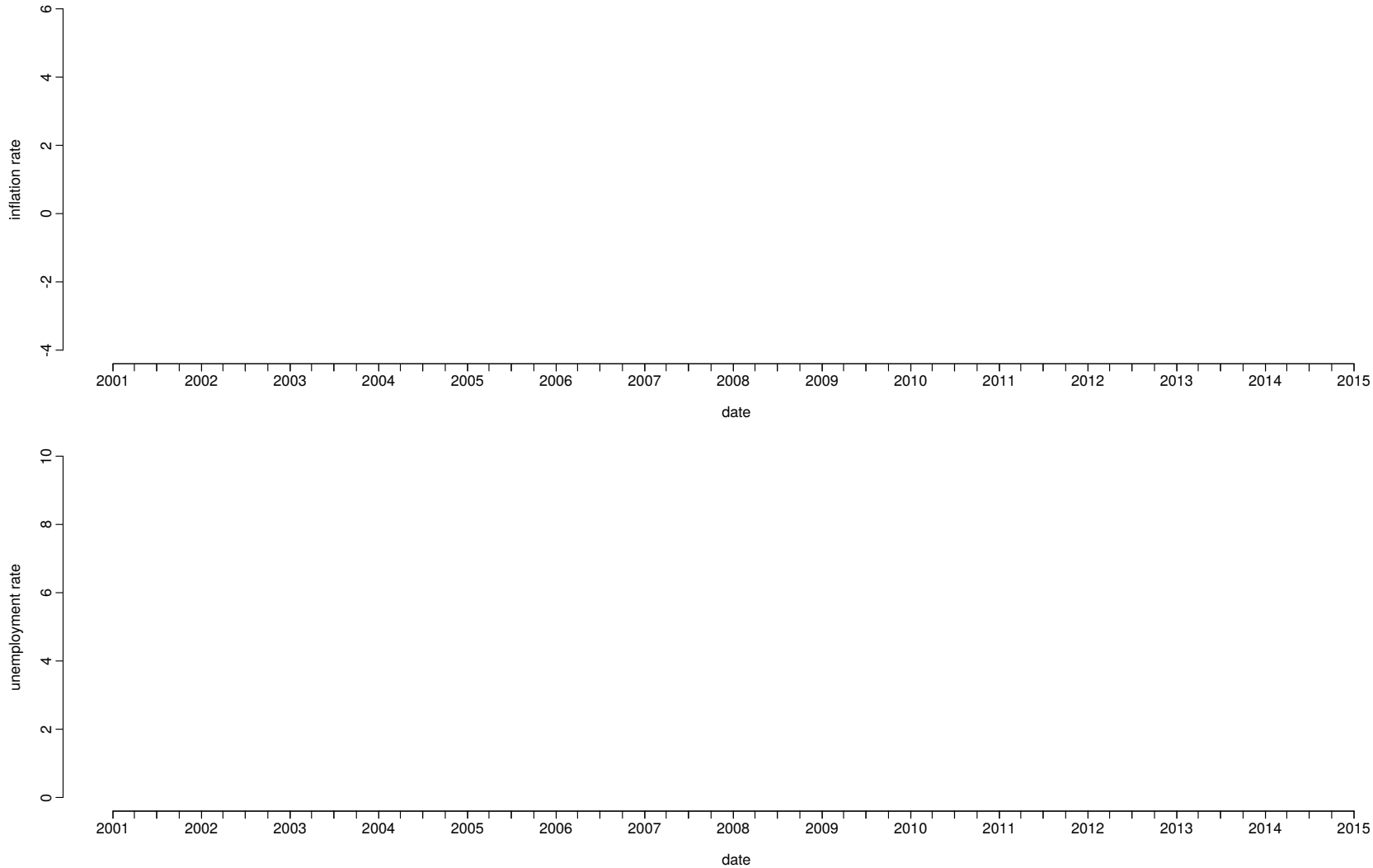
$$y_t = y_{t-1} + \beta\big(u_t - x_t\big) + \sigma_e e_t,$$

with $\theta = [\phi, \alpha, \beta, \sigma_e]$ and $v_t, e_t \overset{\text{IID}}{\sim} \mathcal{N}(0, 1)$.

Example: nonlinear Philips curve model [II/III]



Example: nonlinear Philips curve model [III/III]



Why did we do this?

Bayesian parameter inference for general SSMs.
Accelerate existing posterior sampling methods.
Make it easier for the user to tune existing methods.

Why did we do this?

Bayesian parameter inference for general SSMs.
Accelerate existing posterior sampling methods.
Make it easier for the user to tune existing methods.

How did we do it?

Construct Markov chain using quasi-Newton ideas.
Use this chain to explore the parameter posterior.
Approximate the resulting algorithm using particles.

Why did we do this?

Bayesian parameter inference for general SSMs.
Accelerate existing posterior sampling methods.
Make it easier for the user to tune existing methods.

How did we do it?

Construct Markov chain using quasi-Newton ideas.
Use this chain to explore the parameter posterior.
Approximate the resulting algorithm using particles.

What did we achieve?

Indications of better computational efficiency.
Does not (in principal) require pilot runs.
No. free parameters is independent of $\dim(\theta)$.

LINKÖPING UNIVERSITY

Thank you for listening

Comments, suggestions and/or questions?

Johan Dahlin
johan.dahlin@liu.se
work.johandahlin.com

Extended version and tutorial (soon) on arXiv!

LINKÖPING UNIVERSITY

References

Dahlin, J., Lindsten, F. and Schön T.B., **Particle Metropolis-Hastings using gradient and Hessian information**. Statistics and Computing 25(1), pp. 81-92, 2015.

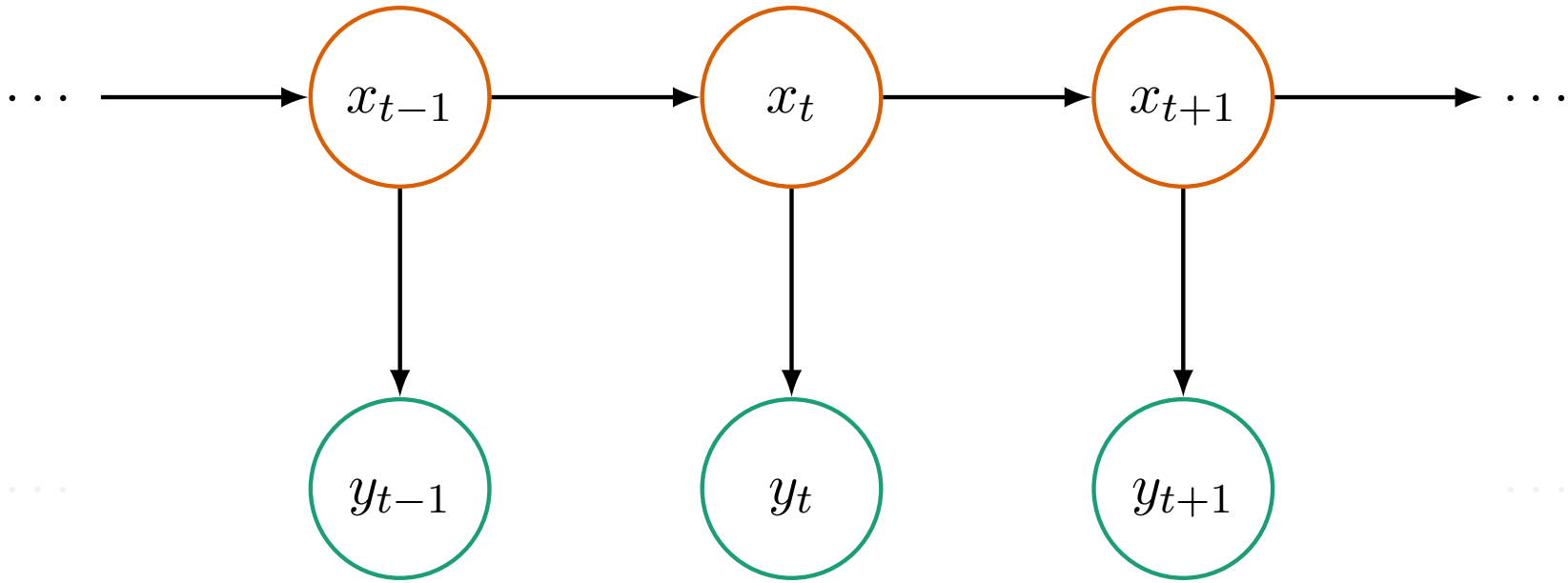
Dahlin, J., **Sequential Monte Carlo for inference in nonlinear state space models**. Licentiate's Thesis, Linköping University, 2014.

Ninness, B. and Henriksen, S., **Bayesian System Identification via Markov Chain Monte Carlo Techniques**. Automatica 46(1), pp. 40-51, 2010.

Zhang, Y. and Sutton, C.A., **Quasi-Newton methods for Markov chain Monte Carlo**. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger (editors), Advances in Neural Information Processing Systems 24, pp. 2393-2401, 2011.

State space models: structure

Markov chain $[X_{0:T}, Y_{1:T}]$ with $X_t \in \mathcal{X} = \mathbb{R}$, $Y_t \in \mathcal{Y} = \mathbb{R}$ and $t \in \mathbb{N}$.



$$x_0 \sim \mu_\theta(x_0) \qquad x_{t+1}|x_t \sim f_\theta(x_{t+1}|x_t), \qquad y_t|x_t \sim g_\theta(y_t|x_t).$$

Example: stochastic volatility with α-stable observations ($\theta = [\mu, \phi, \sigma_v, \alpha]$):

$$x_{t+1}|x_t \sim \mathcal{N}\Big(x_{t+1}; \mu + \phi(x_t - \mu), \sigma_v\Big), \qquad y_t|x_t \sim \mathcal{A}\Big(y_t; \alpha, \exp(x_t)\Big).$$

Particle filtering [I/II]

An instance of sequential Monte Carlo (SMC) samplers.

Estimates $\mathbb{E}\big[\varphi(x_t)|y_{1:t}\big]$ and $p_\theta(y_{1:T})$.

Computational cost of order $\mathcal{O}(NT)$ (with $N \sim T$).

Well-understood statistical properties.

(unbiasedness, large deviation inequalities, CLTs)

References:

A. Doucet and A. Johansen, **A tutorial on particle filtering and smoothing**. In D. Crisan and B. Rozovsky (editors), The Oxford Handbook of Nonlinear Filtering. Oxford University Press, 2011.

O. Cappé, S.J. Godsill and E. Moulines, **An overview of existing methods and recent advances in sequential Monte Carlo**. In Proceedings of the IEEE 95(5), 2007.

Particle filtering [II/II]

By iterating:

Resampling: $\mathbb{P}(a_t^{(i)} = j) = \widetilde{w}_{t-1}^{(j)}$, for $i, j = 1, \ldots, N$.

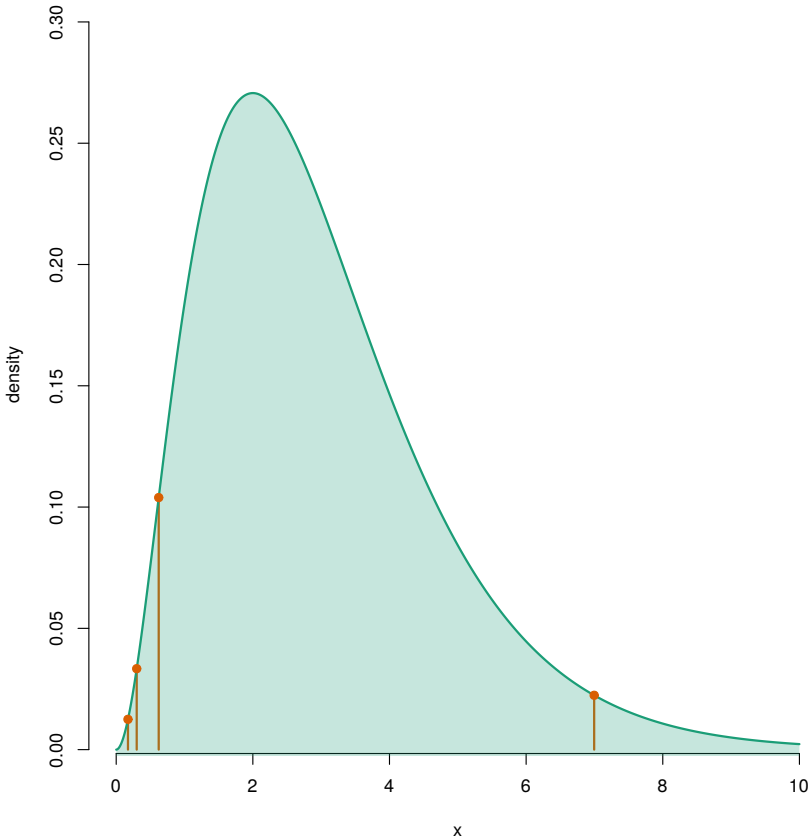
Propagation: $x_t^{(i)} \sim f_\theta\left(x_t \middle| x_{t-1}^{a_t^{(i)}}\right)$, for $i = 1, \ldots, N$.

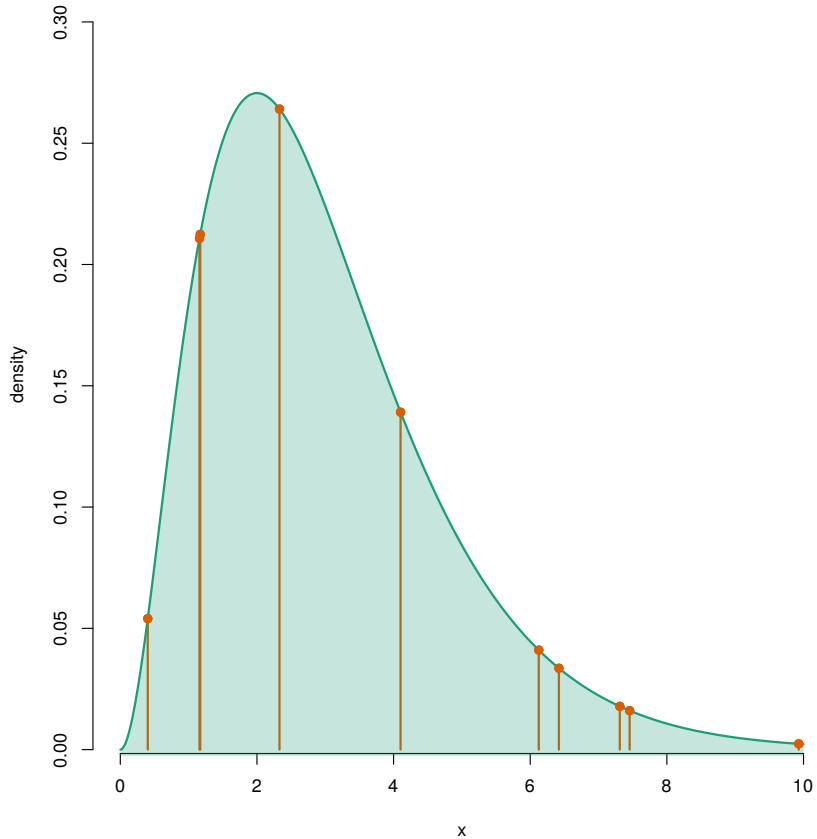
Weighting: $w_t^{(i)} = g_\theta\left(y_t \middle| x_t^{(i)}\right)$, for $i = 1, \ldots, N$.

We obtain the particle system

$$\left[x_{0:T}^{(i)}, w_{0:T}^{(i)}\right]_{i=1}^{N}.$$

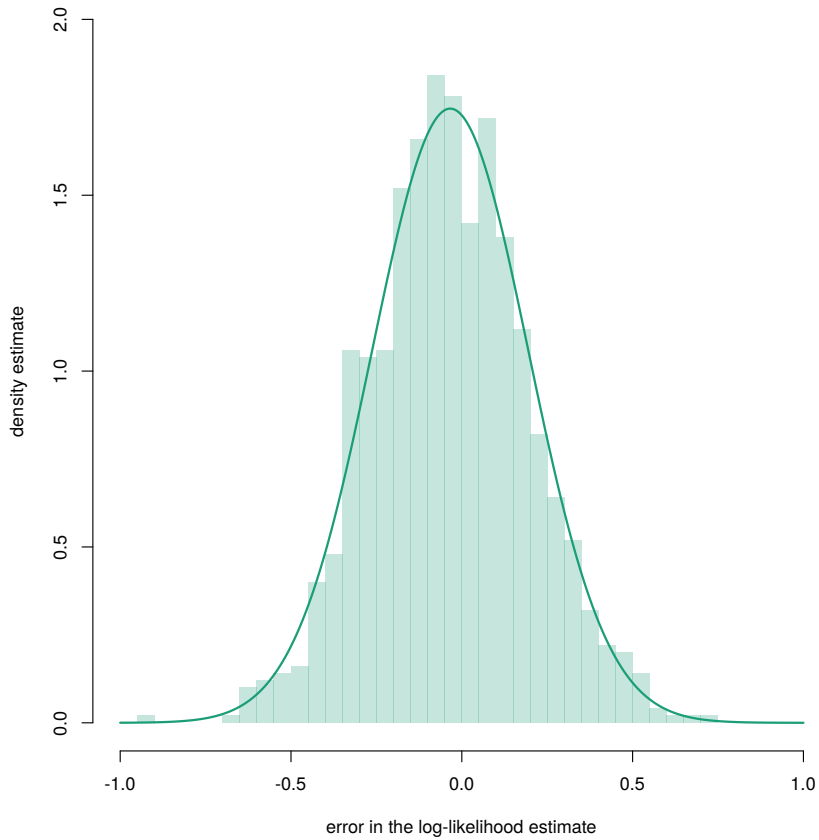
Particle filtering: state estimation

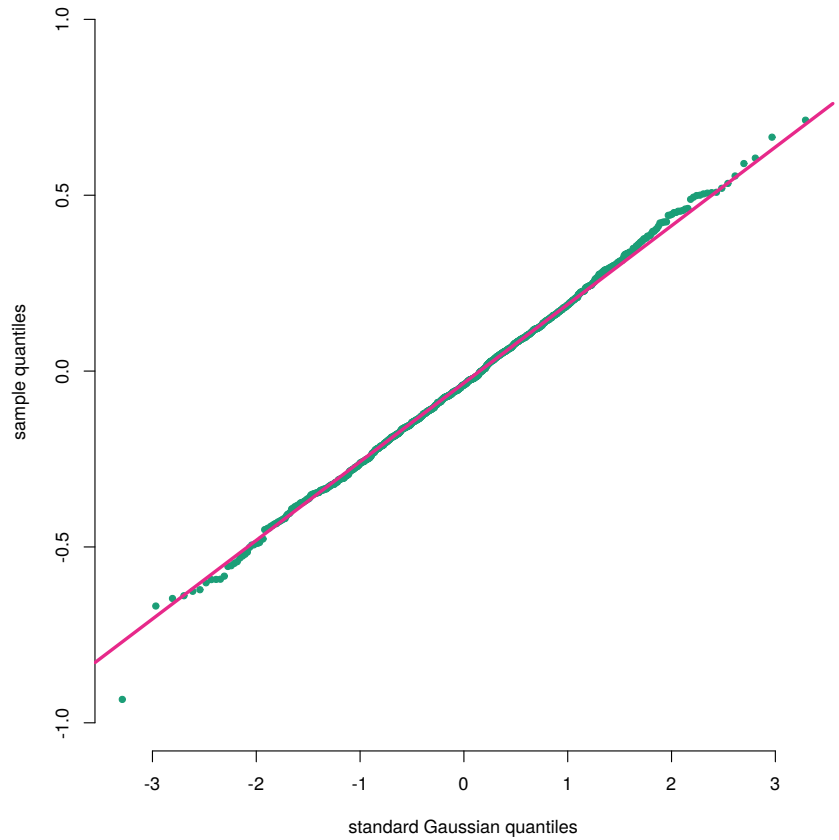




$$\widehat{\varphi}_t^N \triangleq \widehat{\mathbb{E}}\Big[\varphi(x_t)|y_{1:t}\Big] = \sum_{i=1}^{N} w_t^{(i)} \varphi\big(x_t^{(i)}\big), \qquad \sqrt{N}\Big(\varphi_t - \widehat{\varphi}_t^N\Big) \xrightarrow{d} \mathcal{N}\Big(0, \sigma_t^2(\varphi)\Big).$$

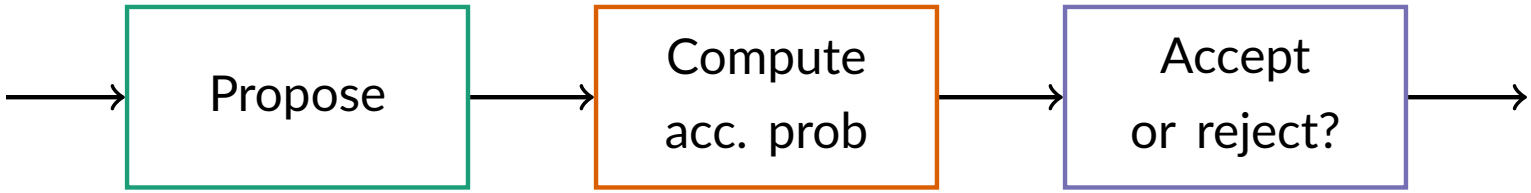
Particle filtering: likelihood estimation





$$\underbrace{\log \widehat{p}_\theta(y_{1:T})}_{\triangleq \widehat{\ell}(\theta)} = \sum_{t=1}^{T} \log\left(\sum_{i=1}^{N} w_t^{(i)}\right) - T\log N, \quad \sqrt{N}\left(\ell(\theta) - \widehat{\ell}(\theta) + \frac{\sigma^2_{\widehat{\pi}}}{2N}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\widehat{\pi}}\right).$$

Particle Metropolis-Hastings [I/III]



- Propose: $\theta' \sim q(\theta'|\theta_{k-1}, u')$ and $u' \sim \mathrm{PF}(\theta')$.
- Compute $\widehat{p}_{\theta'}(y_{1:T}|u')$ and the acceptance probability:

$$\alpha(\theta', \theta_{k-1}) = 1 \wedge \frac{p(\theta')}{p(\theta_{k-1})} \frac{\widehat{p}_{\theta'}(y_{1:T}|u')}{\widehat{p}_{\theta_{k-1}}(y_{1:T}|u_{k-1})} \frac{q(\theta_{k-1}|\theta', u')}{q(\theta'|\theta_{k-1}, u_{k-1})}.$$

- Accept or reject? $\theta' \to \theta_k$ and $u' \to u_k$ w.p. $\alpha(\theta', \theta_{k-1})$.

Particle Metropolis-Hastings [II/III]

The target distribution is given by the parameter proposal

$$\pi(\theta) = \frac{p_\theta(y_{1:T})p(\theta)}{p(y_{1:T})}.$$

An unbiased estimator of the likelihood is given by

$$\mathbb{E}_m\left[\widehat{p}_\theta(y_{1:T}|u)\right] = \int \widehat{p}_\theta(y_{1:T}|u)m_\theta(u)\,\mathrm{d}u = p_\theta(y_{1:T}).$$

An extended target is given by

$$\pi(\theta, u) = \frac{\widehat{p}_\theta(y_{1:T}|u)m_\theta(u)p(\theta)}{p(y_{1:T})} = \frac{\widehat{p}_\theta(y_{1:T}|u)m_\theta(u)\pi(\theta)}{p_\theta(y_{1:T})}.$$

Particle Metropolis-Hastings [III/III]

$$
\begin{aligned}
\int \pi(\theta, u) \, \mathrm{d}u &= \int \frac{\widehat{p}_\theta(y_{1:T}|u) m_\theta(u) \pi(\theta)}{p_\theta(y_{1:T})} \, \mathrm{d}u \\
&= \frac{\pi(\theta)}{p_\theta(y_{1:T})} \underbrace{\int \widehat{p}_\theta(y_{1:T}|u) m_\theta(u) \, \mathrm{d}u}_{=p_\theta(y_{1:T})}, \\
&= \pi(\theta).
\end{aligned}
$$

That is, the marginal is the desired target distribution and the Markov chain is kept invariant.