Estimation of the score vector and observed information matrix in intractable models

Arnaud Doucet (University of Oxford)
Pierre E. Jacob (University of Oxford)
Sylvain Rubenthaler (Université Nice Sophia Antipolis)

November 28th, 2014

Outline

Outline

Motivation

- Derivatives of the likelihood help optimizing / sampling.
- For many models they are not available.
- One can resort to approximation techniques.

Using derivatives in sampling algorithms

Modified Adjusted Langevin Algorithm

At step t, given a point θ_t, do:

- propose

$$\theta^\star \sim q(d\theta \mid \theta_t) \equiv \mathcal{N}(\theta_t + \frac{\sigma^2}{2}\nabla_\theta \log \pi(\theta_t), \sigma^2),$$

- with probability

$$1 \wedge \frac{\pi(\theta^\star) q(\theta_t \mid \theta^\star)}{\pi(\theta_t) q(\theta^\star \mid \theta_t)}$$

  set $\theta_{t+1} = \theta^\star$, otherwise set $\theta_{t+1} = \theta_t$.

Using derivatives in sampling algorithms

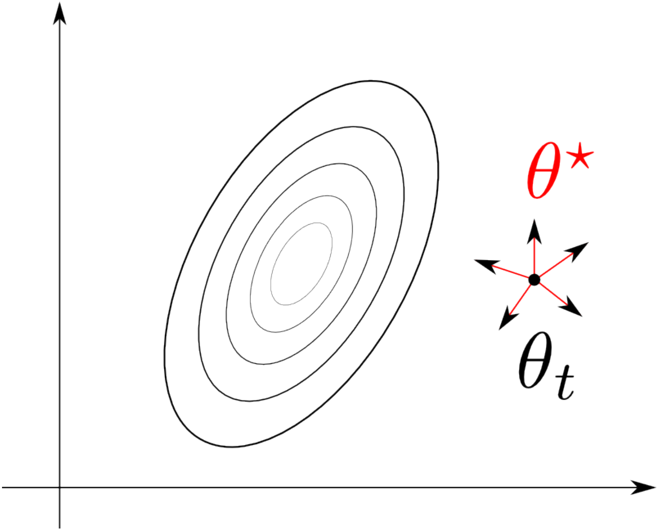

Figure : Proposal mechanism for random walk Metropolis–Hastings.

Using derivatives in sampling algorithms

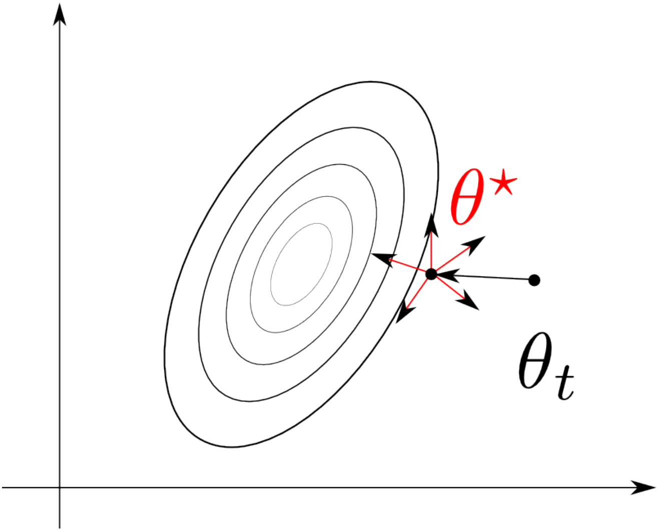

Figure : Proposal mechanism for MALA.

Using derivatives in sampling algorithms

In what sense is MALA better than MH?

Scaling with the dimension of the state space

- For Metropolis–Hastings, optimal scaling leads to

$$\sigma^2 = \mathcal{O}(d^{-1}),$$

- For MALA, optimal scaling leads to

$$\sigma^2 = \mathcal{O}(d^{-1/3}).$$

Roberts & Rosenthal, *Optimal Scaling for Various Metropolis-Hastings Algorithms*, 2001.

Hidden Markov models

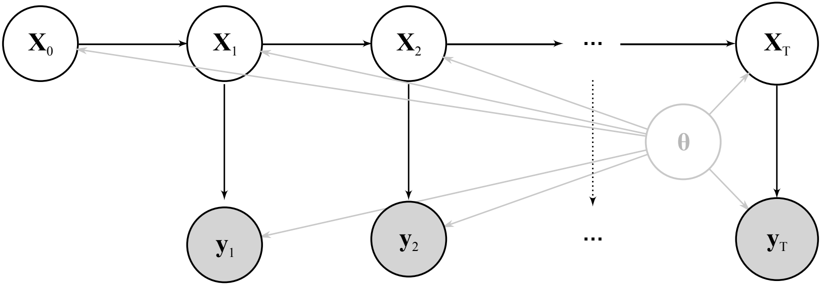

Figure : Graph representation of a general hidden Markov model.

Hidden process: initial distribution μ_θ, transition f_θ.

Observations conditional upon the hidden process, from g_θ.

Assumptions

Input:

- Parameter θ : unknown, prior distribution p.
- Initial condition $\mu_\theta(dx_0)$: can be sampled from.
- Transition $f_\theta(dx_t|x_{t-1})$: can be sampled from.
- Measurement $g_\theta(y_t|x_t)$: can be evaluated point-wise.
- Observations $y_{1:T} = (y_1, \ldots, y_T)$.

Goals:

- score: $\nabla_\theta \log \mathcal{L}(\theta; y_{1:T})$ for any θ,
- observed information matrix: $-\nabla^2_\theta \log \mathcal{L}(\theta; y_{1:T})$ for any θ.

Note: throughout the talk, the observations, and thus the likelihood, are fixed.

Why is it an intractable model?

The likelihood function does not admit a closed form expression:

$$\begin{aligned}\mathcal{L}(\theta; y_1, \ldots, y_T) &= \int_{\mathcal{X}^{T+1}} p(y_1, \ldots, y_T \mid x_0, \ldots x_T, \theta) p(dx_0, \ldots dx_T \mid \theta) \\ &= \int_{\mathcal{X}^{T+1}} \prod_{t=1}^{T} g_\theta(y_t \mid x_t)\ \mu_\theta(dx_0) \prod_{t=1}^{T} f_\theta(dx_t \mid x_{t-1}).\end{aligned}$$

Hence the likelihood can only be estimated, e.g. by standard Monte Carlo, or by particle filters.

What about the derivatives of the likelihood?

Fisher and Louis' identities

Write the score as:

$$\nabla \ell(\theta) = \int \nabla \log p(x_{0:T}, y_{1:T} \mid \theta) p(dx_{0:T} \mid y_{1:T}, \theta).$$

which is an integral, with respect to the smoothing distribution $p(dx_{0:T} \mid y_{1:T}, \theta)$, of

$$\nabla \log p(x_{0:T}, y_{1:T} \mid \theta) = \nabla \log \mu_\theta(x_0) + \sum_{t=1}^{T} \nabla \log f_\theta(x_t \mid x_{t-1}) + \sum_{t=1}^{T} \nabla \log g_\theta(y_t \mid x_t).$$

However pointwise evaluations of $\nabla \log \mu_\theta(x_0)$ and $\nabla \log f_\theta(x_t \mid x_{t-1})$ are not always available.

New kid on the block: Iterated Filtering

Perturbed model

Hidden states $\tilde{X}_t = (\tilde{\theta}_t, X_t)$.

$$\begin{cases} \tilde{\theta}_0 \sim & \mathcal{N}(\theta_0, \tau^2\Sigma) \\ X_0 \sim & \mu_{\tilde{\theta}_0}(\cdot) \end{cases} \quad \text{and} \quad \begin{cases} \tilde{\theta}_t \sim & \mathcal{N}(\tilde{\theta}_{t-1}, \sigma^2\Sigma) \\ X_t \sim & f_{\tilde{\theta}_t}(\cdot \mid X_{t-1} = x_{t-1}) \end{cases}$$

Observations $\tilde{Y}_t \sim g_{\tilde{\theta}_t}(\cdot \mid X_t)$.

Score estimate

Consider $V_{P,t} = \mathbb{C}ov[\tilde{\theta}_t \mid y_{1:t-1}]$ and $\tilde{\theta}_{F,t} = \mathbb{E}[\tilde{\theta}_t \mid y_{1:t}]$.

$$\sum_{t=1}^{T} {V_{P,t}}^{-1} \left(\tilde{\theta}_{F,t} - \tilde{\theta}_{F,t-1}\right) \approx \nabla \ell(\theta_0)$$

when $\tau \to 0$ and $\sigma/\tau \to 0$. Ionides, Breto, King, PNAS, 2006.

Iterated Filtering: the mystery

- Why is it valid?
- Is it related to known techniques?
- Can it be extended to estimate the second derivatives (i.e. the Hessian, i.e. the observed information matrix)?
- How does it compare to other methods such as finite difference?

Outline

Iterated Filtering

Given a log likelihood ℓ and a given point, consider a prior

$$\theta \sim \mathcal{N}(\theta_0, \sigma^2).$$

Posterior expectation when the prior variance goes to zero

First-order moments give first-order derivatives:

$$|\sigma^{-2}\left(\mathbb{E}[\theta|\,Y] - \theta_0\right) - \nabla \ell(\theta_0)| \leq C\sigma^2.$$

Phrased simply,

$$\frac{\text{posterior mean} - \text{prior mean}}{\text{prior variance}} \approx \text{score}.$$

Result from Ionides, Bhadra, Atchadé, King, *Iterated filtering*, 2011.

Extension of Iterated Filtering

Posterior variance when the prior variance goes to zero

Second-order moments give second-order derivatives:

$$|\sigma^{-4}\left(\mathbb{C}ov[\theta|\,Y] - \sigma^2\right) - \nabla^2\ell(\theta_0)| \leq C\sigma^2.$$

Phrased simply,

$$\frac{\text{posterior variance} - \text{prior variance}}{\text{prior variance}^2} \approx \text{hessian}.$$

Result from Doucet, Jacob, Rubenthaler on arXiv, 2013.

Proximity mapping

Given a real function f and a point θ_0, consider for any $\sigma^2 > 0$

$$\theta \mapsto f(\theta) \exp\left\{-\frac{1}{2\sigma^2}(\theta - \theta_0)^2\right\}$$

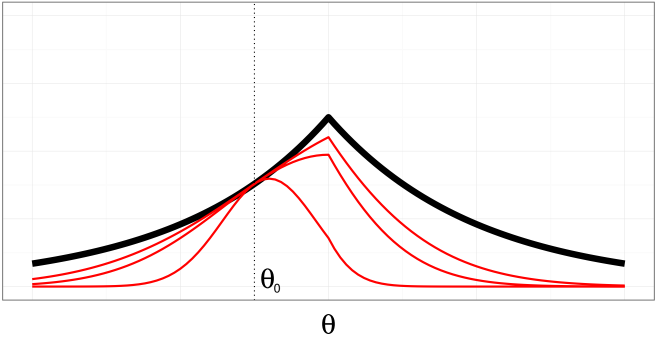

Figure : Example for $f : \theta \mapsto \exp(-|\theta|)$ and three values of σ^2.

Proximity mapping

Proximity mapping

The σ^2-proximity mapping is defined by

$$\text{prox}_f : \theta_0 \mapsto \text{argmax}_{\theta \in \mathbb{R}} \; f(\theta) \exp\left\{-\frac{1}{2\sigma^2}(\theta - \theta_0)^2\right\}.$$

Moreau approximation

The σ^2-Moreau approximation is defined by

$$f_{\sigma^2} : \theta_0 \mapsto C \sup_{\theta \in \mathbb{R}} \; f(\theta) \exp\left\{-\frac{1}{2\sigma^2}(\theta - \theta_0)^2\right\}$$

where C is a normalizing constant.

Proximity mapping

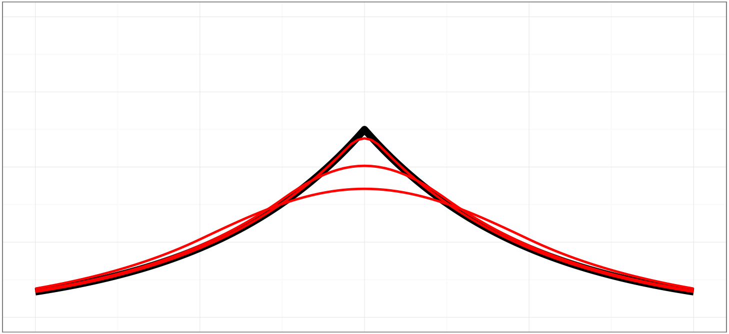

Figure : $\theta \mapsto f(\theta)$ and $\theta \mapsto f_{\sigma^2}(\theta)$ for three values of σ^2.

Proximity mapping

Property

Those objects are such that

$$\frac{\mathrm{prox}_f(\theta_0) - \theta_0}{\sigma^2} = \nabla \log f_{\sigma^2}(\theta_0) \xrightarrow[\sigma^2 \to 0]{} \nabla \log f(\theta_0)$$

Moreau (1962), Fonctions convexes duales et points proximaux dans un espace Hilbertien.

Pereyra (2013), Proximal Markov chain Monte Carlo algorithms.

Proximity mapping

Bayesian interpretation

If f is a seen as a likelihood function then

$$\theta \mapsto f(\theta) \exp\left\{-\frac{1}{2\sigma^2}(\theta - \theta_0)^2\right\}$$

is an unnormalized posterior density function based on a Normal prior with mean θ_0 and variance σ^2.

Hence

$$\frac{\text{prox}_f(\theta_0) - \theta_0}{\sigma^2} \xrightarrow[\sigma^2 \to 0]{} \nabla \log f(\theta_0)$$

can be read

$$\frac{\text{posterior mode} - \text{prior mode}}{\text{prior variance}} \approx \text{score}.$$

Stein's lemma

Stein's lemma states that

$$\theta \sim N(\theta_0, \sigma^2)$$

if and only if for any function g such that $\mathbb{E}\left[|\nabla g(\theta)|\right] < \infty$,

$$\mathbb{E}\left[(\theta - \theta_0)\, g\left(\theta\right)\right] = \sigma^2 \mathbb{E}\left[\nabla g\left(\theta\right)\right].$$

If we choose the function $g : \theta \mapsto \exp \ell\left(\theta\right)/\mathcal{Z}$ with $\mathcal{Z} = \mathbb{E}\left[\exp \ell\left(\theta\right)\right]$ and apply Stein's lemma we obtain

$$\frac{1}{\mathcal{Z}}\mathbb{E}\left[\theta \exp \ell(\theta)\right] - \theta_0 = \frac{\sigma^2}{\mathcal{Z}}\mathbb{E}\left[\nabla \ell\left(\theta\right) \exp\left(\ell\left(\theta\right)\right)\right]$$
$$\Leftrightarrow \sigma^{-2}\left(\mathbb{E}\left[\theta \mid Y\right] - \theta_0\right) = \mathbb{E}\left[\nabla \ell\left(\theta\right) \mid Y\right].$$

Notation: $\mathbb{E}[\varphi(\theta) \mid Y] := \mathbb{E}[\varphi(\theta) \exp \ell(\theta)]/\mathcal{Z}$.

Stein's lemma

For the second derivative, we consider

$$h : \theta \mapsto (\theta - \theta_0) \exp \ell(\theta) / \mathcal{Z}.$$

Then

$$\mathbb{E}\left[(\theta - \theta_0)^2 \mid Y\right] = \sigma^2 + \sigma^4 \mathbb{E}\left[\nabla^2 \ell(\theta) + \nabla \ell(\theta)^2 \mid Y\right].$$

Adding and subtracting terms also yields

$$\begin{aligned} \sigma^{-4}\left(\mathbb{V}[\theta \mid Y] - \sigma^2\right) &= \mathbb{E}\left[\nabla^2 \ell(\theta) \mid Y\right] \\ &\quad + \left\{\mathbb{E}\left[\nabla \ell(\theta)^2 \mid Y\right] - \left(\mathbb{E}[\nabla \ell(\theta) \mid Y]\right)^2\right\}. \end{aligned}$$

...but what we really want is

$$\nabla \ell(\theta_0), \nabla \ell^2(\theta_0)$$

and not

$$\mathbb{E}[\nabla \ell(\theta) \mid Y], \mathbb{E}\left[\nabla \ell^2(\theta) \mid Y\right].$$

Outline

Core Idea

The prior is a essentially a normal distribution $\mathcal{N}(\theta_0, \sigma^2)$, but in general has a density denoted by κ.

Posterior concentration induced by the prior

Under some assumptions, when $\sigma \to 0$:

- the posterior looks more and more like the prior,
- the shift in posterior moments is in $\mathcal{O}(\sigma^2)$.

Our arXived proof suffers from an overdose of Taylor expansions.

Details

Introduce a test function h such that $|h(u)| < c|u|^\alpha$ for some c, α.
We start by writing

$$\mathbb{E}\left\{ h(\theta - \theta_0) \middle| y \right\} = \frac{\int h(\sigma u) \exp\{\ell(\theta_0 + \sigma u) - \ell(\theta_0)\} \kappa(u) \, du}{\int \exp\{\ell(\theta_0 + \sigma u) - \ell(\theta_0)\} \kappa(u) \, du}$$

using $u = (\theta - \theta_0)/\sigma$ and then focus on the numerator

$$\int h(\sigma u) \exp\{\ell(\theta_0 + \sigma u) - \ell(\theta_0)\} \kappa(u) \, du$$

since the denominator is a particular instance of this expression with $h : u \mapsto 1$.

Details

For the numerator:

$$\int h(\sigma u) \exp\{\ell(\theta_0 + \sigma u) - \ell(\theta_0)\} \kappa(u)\, du$$

we use a Taylor expansion of ℓ around θ_0 and a Taylor expansion of exp around 0, and then take the integral with respect to κ.

Notation:

$$\ell^{(k)}(\theta).u^{\otimes k} = \sum_{1 \leq i_1, \ldots, i_k \leq d} \frac{\partial^k \ell(\theta)}{\partial \theta_{i_1} \ldots \partial \theta_{i_k}} \; u_{i_1} \ldots u_{i_k}$$

which in one dimension becomes

$$\ell^{(k)}(\theta).u^{\otimes k} = \frac{d^k f(\theta)}{d\theta^k} u^k.$$

Details

Main expansion:

$$\int h(\sigma u) \exp\left\{\ell\left(\theta_0 + \sigma u\right) - \ell(\theta_0)\right\} \kappa(u) du =$$

$$\int h(\sigma u)\kappa(u) du + \sigma \int h(\sigma u)\ell^{(1)}(\theta_0).u \; \kappa(u) du$$

$$+ \sigma^2 \int h(\sigma u) \left\{ \frac{1}{2}\ell^{(2)}(\theta_0).u^{\otimes 2} + \frac{1}{2}(\ell^{(1)}(\theta_0).u)^2 \right\} \kappa(u) du$$

$$+ \sigma^3 \int h(\sigma u) \left\{ \frac{1}{3!}(\ell^{(1)}(\theta_0).u)^3 + \frac{1}{2}(\ell^{(1)}(\theta_0).u)(\ell^{(2)}(\theta_0).u^{\otimes 2}) \right.$$

$$\left. + \frac{1}{3!}\ell^{(3)}(\theta_0).u^{\otimes 3} \right\} \kappa(u) du + \mathcal{O}(\sigma^{4+\alpha}).$$

In general, assumptions on the tails of the prior and the likelihood are used to control the remainder terms and to ensure there are $\mathcal{O}(\sigma^{4+\alpha})$.

Details

We cut the integral into two bits:

$$\int h(\sigma u) \exp\left\{\ell\left(\theta_0 + \sigma u\right) - \ell(\theta_0)\right\} \kappa(u) du$$
$$= \int_{\sigma|u|\leq\rho} h(\sigma u) \exp\left\{\ell\left(\theta_0 + \sigma u\right) - \ell(\theta_0)\right\} \kappa(u) du$$
$$+ \int_{\sigma|u|>\rho} h(\sigma u) \exp\left\{\ell\left(\theta_0 + \sigma u\right) - \ell(\theta_0)\right\} \kappa(u) du$$

- The expansion stems from the first term, where $\sigma|u|$ is small.
- The second term ends up in the remainder in $\mathcal{O}(\sigma^{4+\alpha})$ using the assumptions.

Classic technique in Bayesian asymptotics theory.

Details

- To get the score from the expansion, choose

$$h : u \mapsto u.$$

- To get the observed information matrix from the expansion, choose

$$h : u \mapsto u^2,$$

and surprisingly (?) further assume that κ is mesokurtic, *i.e.*

$$\int u^4 \kappa(u) du = 3 \left(\int u^2 \kappa(u) du \right)^2$$

$\Rightarrow$ choose a Gaussian prior to obtain the hessian.

Outline

Hidden Markov models

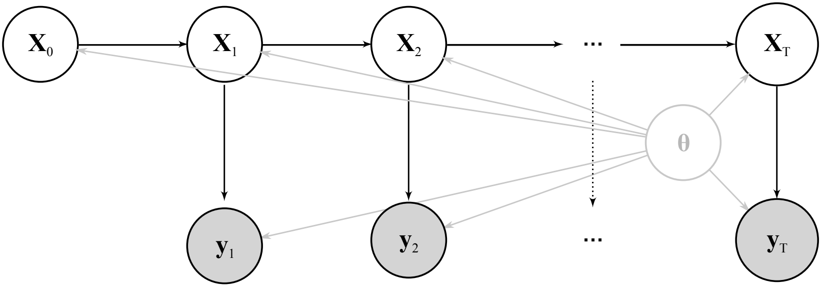

Figure : Graph representation of a general hidden Markov model.

Hidden Markov models

Direct application of the previous results

1. Prior distribution $\mathcal{N}(\theta_0, \sigma^2)$ on the parameter θ.
2. The derivative approximations involve $\mathbb{E}[\theta | Y]$ and $\mathbb{C}ov[\theta | Y]$.
3. Posterior moments for HMMs can be estimated by
   - particle MCMC,
   - SMC^2,
   - ABC

   or your favourite method.

Ionides et al. proposed another approach.

Iterated Filtering

Modification of the model: θ is time-varying.

The associated loglikelihood is

$$\begin{aligned}\bar{\ell}(\theta_{1:T}) &= \log p(y_{1:T}; \theta_{1:T}) \\ &= \log \int_{\mathcal{X}^{T+1}} \prod_{t=1}^{T} g(y_t \mid x_t, \theta_t)\ \mu(dx_1 \mid \theta_1) \prod_{t=2}^{T} f(dx_t \mid x_{t-1}, \theta_t).\end{aligned}$$

Introducing $\theta \mapsto (\theta, \theta, \ldots, \theta) := \theta^{[T]} \in \mathbb{R}^T$, we have

$$\bar{\ell}(\theta^{[T]}) = \ell(\theta)$$

and the chain rule yields

$$\frac{d\ell(\theta)}{d\theta} = \sum_{t=1}^{T} \frac{\partial \bar{\ell}(\theta^{[T]})}{\partial \theta_t}.$$

Choice of prior on $\theta_{1:T}$:

$$\widetilde{\theta}_1 = \theta_0 + V_1, \qquad V_1 \sim \tau^{-1}\kappa\left\{\tau^{-1}(\cdot)\right\}$$

$$\widetilde{\theta}_{t+1} - \theta_0 = \rho\left(\widetilde{\theta}_t - \theta_0\right) + V_{t+1}, \qquad V_{t+1} \sim \sigma^{-1}\kappa\left\{\sigma^{-1}(\cdot)\right\}$$

Choose σ^2 such that $\tau^2 = \sigma^2/(1-\rho^2)$. Covariance of the prior on $\theta_{1:T}$:

$$\Sigma_T = \tau^2 \begin{pmatrix} 1 & \rho & \cdots & \cdots & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \ddots & & \rho^{T-3} \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \rho^{T-2} & & & \ddots & 1 & \rho \\ \rho^{T-1} & \cdots & \cdots & \cdots & \rho & 1 \end{pmatrix}.$$

Applying the general results for this prior yields, with $|x| = \sum_{t=1}^{T} |x_i|$:

$$|\nabla \bar{\ell}(\theta_0^{[T]}) - \Sigma_T^{-1} \left(\mathbb{E}\left[\widetilde{\theta}_{1:T} \mid Y \right] - \theta_0^{[T]} \right) | \leq C\tau^2$$

Moreover we have

$$\begin{aligned} &\left| \sum_{t=1}^{T} \frac{\partial \bar{\ell}(\theta^{[T]})}{\partial \theta_t} - \sum_{t=1}^{T} \left\{ \Sigma_T^{-1} \left(\mathbb{E}\left[\widetilde{\theta}_{1:T} \mid Y \right] - \theta_0^{[T]} \right) \right\}_t \right| \\ &\leq \sum_{t=1}^{T} \left| \frac{\partial \bar{\ell}(\theta^{[T]})}{\partial \theta_t} - \left\{ \Sigma_T^{-1} \left(\mathbb{E}\left[\widetilde{\theta}_{1:T} \mid Y \right] - \theta_0^{[T]} \right) \right\}_t \right| \end{aligned}$$

and

$$\frac{d\ell(\theta)}{d\theta} = \sum_{t=1}^{T} \frac{\partial \bar{\ell}(\theta^{[T]})}{\partial \theta_t}.$$

Iterated Filtering

The estimator of the score is thus given by

$$\sum_{t=1}^{T}\left\{\Sigma_T^{-1}\left(\mathbb{E}\left[\widetilde{\theta}_{1:T}\mid Y\right]-\theta_0^{[T]}\right)\right\}_t$$

which can be reduced to

$$\begin{aligned}S_{\tau,\rho,T}\left(\theta_0\right) &= \frac{\tau^{-2}}{1+\rho}\left[(1-\rho)\left\{\sum_{t=2}^{T-1}\mathbb{E}\left(\widetilde{\theta}_t\middle| Y\right)\right\}-\left\{(1-\rho)\,T+2\rho\right\}\theta_0\right.\\ &\left.+\mathbb{E}\left(\widetilde{\theta}_1\middle| Y\right)+\mathbb{E}\left(\widetilde{\theta}_T\middle| Y\right)\right],\end{aligned}$$

given the form of Σ_T^{-1}. Note that in the quantities $\mathbb{E}(\theta_t\mid Y)$, $Y=Y_{1:T}$ is the complete dataset, thus those expectations are with respect to the smoothing distribution.

- If $\rho = 1$, then the parameters follow a random walk:

$$\widetilde{\theta}_1 = \theta_0 + \mathcal{N}(0, \tau^2) \quad \text{and} \quad \widetilde{\theta}_{t+1} = \widetilde{\theta}_t + \mathcal{N}(0, \sigma^2).$$

In this case Ionides et al. proposed the estimator

$$S_{\tau,\sigma,T} = \tau^{-2}\left(\mathbb{E}\left(\widetilde{\theta}_T \mid Y\right) - \theta_0\right)$$

as well as

$$S_{\tau,\sigma,T}^{(bis)} = \sum_{t=1}^{T} {V_{P,t}}^{-1}\left(\tilde{\theta}_{F,t} - \tilde{\theta}_{F,t-1}\right)$$

with $V_{P,t} = \mathbb{C}ov[\tilde{\theta}_t \mid y_{1:t-1}]$ and $\tilde{\theta}_{F,t} = \mathbb{E}[\theta_t \mid y_{1:t}]$.

Those expressions only involve expectations with respect to filtering distributions.

Iterated Filtering

- If $\rho = 0$, then the parameters are i.i.d:

$$\widetilde{\theta}_1 = \theta_0 + \mathcal{N}(0, \tau^2) \quad \text{and} \quad \widetilde{\theta}_{t+1} = \widetilde{\theta}_0 + \mathcal{N}(0, \tau^2).$$

  In this case the expression of the score estimator reduces to

$$S_{\tau,T} = \tau^{-2} \sum_{t=1}^{T} \left(\mathbb{E}\left(\widetilde{\theta}_t \mid Y \right) - \theta_0 \right)$$

which involves smoothing distributions.

- There's only one parameter τ^2 to choose for the prior.
- However smoothing for general hidden Markov models is difficult, and typically resorts to "fixed lag approximations".

Iterated Smoothing

Only for the case $\rho = 0$ are we able to obtain simple expressions for the observed information matrix. We propose the following estimator:

$$I_{\tau,T}\left(\theta_0\right) = -\tau^{-4}\left\{\sum_{s=1}^{T}\sum_{t=1}^{T}\mathbb{C}ov\left(\widetilde{\theta}_s, \widetilde{\theta}_t \middle| Y\right) \; - \tau^2 T\right\}.$$

for which we can show that

$$\left|I_{\tau,T} - (-\nabla^2 \ell(\theta_0))\right| \leq C\tau^2.$$

Numerical results

Linear Gaussian state space model where the ground truth is available through the Kalman filter.

$$X_0 \sim \mathcal{N}(0,1) \quad \text{and } X_t = \rho X_{t-1} + \mathcal{N}(0, V)$$
$$Y_t = \eta X_t + \mathcal{N}(0, W).$$

Generate $T = 100$ observations and set $\rho = 0.9$, $V = 0.7, \eta = 0.9$ and $W = 0.1, 0.2, 0.4, 0.9$.

240 independent runs, matching the computational costs between methods in terms of number of calls to the transition kernel.

Numerical results

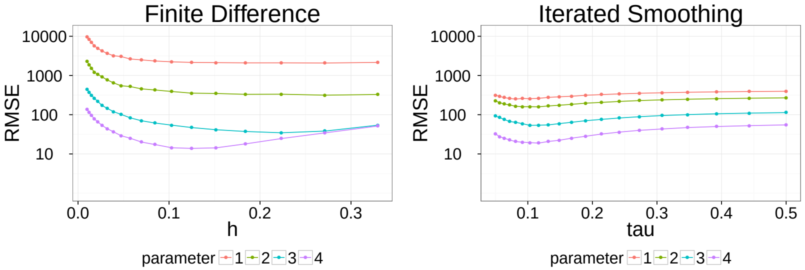

Figure : 240 runs for Iterated Smoothing and Finite Difference.

Numerical results

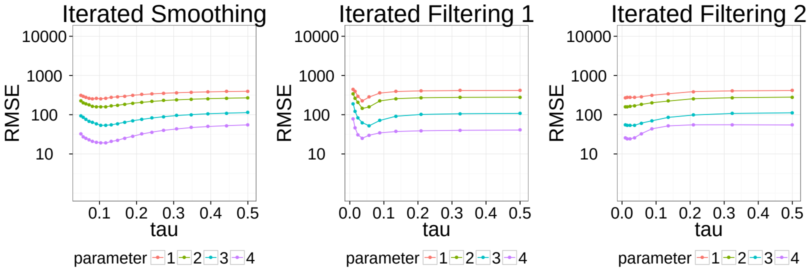

Figure : 240 runs for Iterated Smoothing and Iterated Filtering.

Bibliography

Main references:

- Inference for nonlinear dynamical systems, Ionides, Breto, King, PNAS, 2006.
- Iterated filtering, Ionides, Bhadra, Atchadé, King, Annals of Statistics, 2011.
- Efficient iterated filtering, Lindström, Ionides, Frydendall, Madsen, 16th IFAC Symposium on System Identification.
- Derivative-Free Estimation of the Score Vector and Observed Information Matrix, Doucet, Jacob, Rubenthaler, 2013 (on arXiv).