About the tests: parameter estimation for HnMM

Fei Lu

Started: June 2018, Last updated: August 16, 2018

Abstract

Estimate parameter by Particle Gibbs with Ancestor Sampling (PGAS) algorithm,

1 The hidden non-Markov model

The script generates a batch of data $y_{1:T}$ from the the standard nonlinear time series model,

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + b_1 x_{t-1}/(1 + x_{t-1}^2) + 8\cos(1.2(t-1)) + d_1 v_{t-1} + v_t, \tag{1.1}$$

$$y_t = 0.05 * x_t^2 + e_t, \tag{1.2}$$

with $v_t \sim N(0, q)$ and $e_t \sim N(0, r)$.

We assume that the variances of the model noise and the measurement noise (q, r) are known (they can also be estimated often with inverse Gamma priors, but here we assume that they are known for simplicity—- Maybe they are easier to estimate?). The parameters to be estimated are

$$\theta = (a_1, a_2, b_1, d_1).$$

The PGAS algorithm generates a Markov chain which has the posterior $p(\theta, x_{1:T}|y_{1:T})$ as invariant density.

Here we consider the following estimators

- MLE-true states: MLE from true state is close to the true values of the parameters
- online EM: MLE of the noisy observation by online EM
- Bayes estimator is close to the MLE from reference trajectories

2 Likelihood, Prior and Posterior of the parameters.

In simulations, the true parameter are sampled from a prior, either Gaussian or uniform. In the Gibbs sampling, we sample parameters from the posterior

$$p(\theta|y_{1:T}, x_{1:T}) \propto p_{prior}(\theta) p_\theta(x_{1:T}|y_{1:T}) \propto p_{prior}(\theta) p_\theta(x_{1:T}),$$

where in the second step we use the fact that the parameters are from the state model (one can also replace these proportional to by equalities using Bayesian formula). Note that

$$p_\theta(x_{1:T}) = (\sqrt{2q\pi})^{-(T-2)} e^{-\sum_{t=2}^{T-1} \|x_{t+1} - f(t) - \sum_i \theta_i \phi_i(X_{t-1:t})\|^2/(2q)}.$$

This can be viewed as a distribution of θ in form of $\mathcal{N}(\mu_1, \Sigma_1)$ with

$$\Sigma_1^{-1} = \sum_{t=2}^{T-1} \Phi_t' \Sigma_V^{-1} \Phi_t, \text{ and } \mu_1 = \Sigma_1 \sum_{t=2}^{T-1} \Phi_t' \cdot (x_{t+1} - f(t)),$$

where we denoted $\Phi_t = (\phi_1(X_{t-1:t}), \ldots, \phi_4(X_{t-1:t}))$.

- *Gaussian prior.* Combing it with a Gaussian prior $\mathcal{N}(\mu_0, \Sigma_0)$, we obtain that the posterior of θ is $\mathcal{N}(\mu, \Sigma)$ with
$$\Sigma^{-1} = (\Sigma_0^{-1} + \Sigma_1^{-1}), \text{ and } \mu = \Sigma(\Sigma_0^{-1}\mu_0 + \Sigma_1^{-1}\mu_1).$$
- *Uniform prior.* When the likelihood is combined with a uniform prior $\otimes_{i=1}^4 [a_i, b_i]$, the posterior becomes difficult to sample. This is because the likelihood Gaussian can be very skewed and far from these prior intervals.

Remark: when partial of θ is known (e,g. suppose θ_3, θ_4 are known, and we only need to estimate θ_1, θ_2), the likelihood needs to be slightly changed. .

MLE by Expectation-Maximization Online EM algorithm

3 Identifiability, Stability and constraints on the parameters

About stability and parameter range:
One may consider stability of the linear parts (unclear if necessary/sufficient) [1]

$$x_{t+1} = a_1 x_t + a_2 x_{t-1}$$

to put constraints on the coefficients. Recall that for the linear parts to be stable, one needs that the roots of

$$p(z) = 1 - a_1 z - a_2 z^2$$

to be outside the unit disk. The roots are:

$$\frac{a_1 \pm \sqrt{a_1^2 + 4a_2}}{-2a_2}$$

A sufficient condition: $-a_1^2/4 < a_2 < 0, a_1$

Identifiability The likelihood inverse is well-posed. When estimating all the parameters from true states, the smallest eigenvalue of the regression matrix is about 0.04 for all four parameters.

- MLE-true state: consistent estimators
- online EM:
- Bayes estimator is close to the MLE from reference trajectories, and tend to
  - underestimate σ_W, d_1, b_1
  - over estimate a_1, a_2

True: 0.70 -0.20 25.00 0.75 1.00

MLE : 0.74 -0.14 20.55 0.30 0.91

Bayes: 0.72 -0.14 21.06 0.36 0.92

	MLE-true states	online EM	Bayes
a_1, a_2	well-posed	TBD	TBD
a_1, a_2, σ_W	well-posed	TBD	TBD
a_1, a_2, b_1, d_1	well-posed	TBD	TBD

[1] Stability can be discussed by the standard local stability argument: find critical points and study the local, global features, Lyapunov functions, energy potentials.

3.1 Initialize the ensemble of particles

Unlike the Markov models which forgets the past, non-Markov model depend heavily on the past, therefore, the initial conditions become very important. For the NARMA, we need $t_0 = \max\{p, q\}$ step initial conditions. We may use maximum posterior of $x_{1:t_0}$ as initial conditions,

4 EM algorithms for HnMMs

Consider an HnMM with transition probability density $p(x_k|x_{1:k-1}, v_{1:q}, \theta)$ and observation pdf $p(y_k|x_k, \theta)$ for each time k. Assume that $p(x_k|x_{1:k-1}, v_{1:q}, \theta)p(y_k|x_k, \theta)$ is of the form

$$p(x_k|x_{1:k-1}, v_{1:q}, \theta)p(y_k|x_k, \theta) = h(x_{1:k}, v_{1:q}, y_k) \exp(-\psi(\theta) + \langle S(x_{k-p:k}, v_{k-q:k-1}, y_k), \phi(\theta)\rangle),$$

so that the complete-data joint distribution $p(x_{1:n}, y_{1:n}|\theta)$ belongs to an exponential family,

$$\begin{aligned}
\log p(x_{1:n}, y_{1:n}|\theta) &= \log \prod_{k=1}^{n} p(x_k|x_{k-1}, \theta)p(y_k|x_k, \theta) \\
&= \sum_{k=1}^{n} \log p(x_k|x_{k-1}, \theta)p(y_k|x_k, \theta) \\
&= \sum_{k=1}^{n} \Big(\log h(x_k, y_k) - \psi(\theta) + \langle S(x_{k-1}, x_k, y_k), \phi(\theta)\rangle \Big).
\end{aligned}$$

Given $\theta_n^{(t)}$ in iteration t at time n, notice that $\sum_{k=1}^{n} \mathbb{E}\left[h(x_k, y_k)|y_{1:n}, \theta_n^{(t)}\right]$ does not depend on θ, the intermediate quantity can be reduced to be

$$Q(\theta, \theta_n^{(t)}) = -n\psi(\theta) + \langle \sum_{k=1}^{n} \mathbb{E}\left[S(x_{k-1}, x_k, y_k)|y_{1:n}, \theta_n^{(t)}\right], \phi(\theta)\rangle.$$

One only needs to compute $\sum_{k=1}^{n} \mathbb{E}\left[S(x_{k-1}, x_k, y_k)|y_{1:n}, \theta_n^{(t)}\right]$ and the gradients of $\psi(\theta)$ and $\phi(\theta)$ for the optimization.

Example 4.1 *Consider estimating $\theta = (\theta_{1:d-1}, \sigma_1)$ in an HMM with*

$$p(y_k|x_k, \theta) = \frac{1}{\sqrt{2\pi\sigma_2}} e^{-\frac{|y_k - g(x_k)|^2}{2\sigma_2}}, \quad p(x_k|x_{k-1}, \theta) = \frac{1}{\sqrt{2\pi\sigma_1}} e^{-\frac{|x_k - \sum_{i=1}^{d-1}\theta_i f_i(x_{k-1})|^2}{2\sigma_1}},$$

the terms in the exponential family are

$$\begin{aligned}
\phi(\theta) = (\theta_{1:d-1}^2, \theta_{1:d-1}, 1)/\sigma_1, &\quad \psi(\theta) = \frac{n}{2}\log\sigma_1, \\
S(x_{k-1}, x_k, y_k) = -&\left(f_{1:d-1}^2(x_{k-1}), 2x_k f_{1:d-1}(x_{k-1}), \frac{x_k^2}{2}\right).
\end{aligned}$$

To calculate the maximizer of the intermediate quantity, note that it is quadratic in θ,

$$Q(\theta, \theta_n^{(t)}) = C_h - n\psi(\theta) + \frac{1}{2\sigma_1}\sum_{k=1}^{n} \mathbb{E}\left[|X_k - \sum_{i=1}^{d-1}\theta_i f_i(X_{k-1})|^2 \Big| y_{1:n}, \theta_n^{(t)} \right],$$

its maximizer can be analytically calculated:

$$\theta_n^{(t+1)} = A^{-1}b, \quad \sigma_n^{(t+1)} = \frac{1}{n}\sum_{k=1}^{n}|X_k - \sum_{i=1}^{d-1}\theta_i f_i(X_{k-1})|^2$$

where $A \in \mathbb{R}^{(d-1)\times(d-1)}$ *is a matrix with entries* $A_{ij} = \sum_{k=1}^{n}\mathbb{E}\left[f_i(X_{k-1})f_j(X_{k-1})|y_{1:n}, \theta_n^{(t)}\right]$*, and* $b \in \mathbb{R}^{d-1}$ *is a vector with entries* $b_i = \sum_{k=1}^{n}\mathbb{E}\left[f_i(X_{k-1})X_k|y_{1:n}, \theta_n^{(t)}\right]$.

This algorithm depends on the approximation of the conditional expectations. Unfortunately, the conditional expectations are likely to be approximated poorly, especially those with respect to the empirical smoothing densities $\hat{p}(x_{k-1:k}|y_{1:n}, \theta)$. This is due to the degeneracy of particles filters: the empirical smoothing densities are often approximated by only a single particle when $k < n$. Various methods have been proposed to improving the smoothing density, including backward sampling algorithms such as [?] and block sampling [?] or lookahead strategies [?] which do not require backward sampling.

On-line EM algorithm The above algorithm is offline, and the conditional expectations in Q have to be re-computed for each n. That is, at time n, one has to compute n conditional expectations of the sufficient statistics $\{S(X_{k-1}, X_k, Y_k)\}_{k=1}^{n}$ under distribution $p(x_{1:n}|y_{1:n}, \theta_{n-1})$, and at time $n+1$, one has to re-compute these conditional expectations with respect to the new distribution $p(x_{1:n+1}|y_{1:n+1}, \theta_n)$.

The online EM algorithm aims to avoid the repeated computation of these conditional expectations, and more importantly, to run through the data only once, so as to save computational cost when the data size is large. At each time n, only the expectation $\mathbb{E}\left[S(X_{n-1}, X_n, Y_n), \phi(\theta)\middle|y_{1:n}, \theta_{n-1}\right]$ is computed, and all the previous expectations are reused. More precisely, suppose that at time n, with an estimator θ_n and approximate function $\widehat{Q}_{n-1}(\theta)$ from previous time, then the intermediate function Q is updated by

$$\widehat{Q}_n(\theta) = \widehat{Q}_{n-1}(\theta) + \gamma_n\left(\langle\mathbb{E}\left[S(X_{n-1}, X_n, Y_n)\middle|y_{1:n}, \theta_{n-1}\right], \phi(\theta)\rangle - \widehat{Q}_{n-1}(\theta)\right),$$

where γ_n is a sequence of numbers satisfying:

$$\gamma_n > 0, \quad \sum_{n=1}^{\infty}\gamma_n = \infty, \text{ and } \sum_{n=1}^{\infty}\gamma_n^2 < \infty.$$

For example, the commonly used numbers are $\gamma_n = \gamma_0 n^{-\alpha}$ with $\alpha \in (\frac{1}{2}, 1]$ and suitable γ_0.

Especially, when $Q(\theta) = -\psi(\theta) + \langle\bar{s}, \phi(\theta)\rangle$ has a unique maximizer that can be analytically represented as a function $\theta^* = \bar{\theta}(\bar{s})$, the online EM algorithm reduced to be simply

$$\begin{aligned} &\text{E-step:} \qquad && \widehat{s}_n = \widehat{s}_{n-1} + \gamma_n(\bar{s}_n - \widehat{s}_{n-1}), \\ &\text{M-step:} \qquad && \widehat{\theta}_n = \bar{\theta}(\widehat{s}_n), \end{aligned} \tag{4.1}$$

where $\bar{s}_n = \mathbb{E}\left[S(X_{n-1}, X_n, Y_n)\middle|y_{1:n}, \widehat{\theta}_{n-1}\right]$.