Parameter estimation for Hidden Markov Models

Bayesian inference

Introduction to parameter estimation for HMMs

- Particle filters provide a very good approximation of $p_\theta(x_n|y_{0:n})$
  - in this case path degeneracy does not matter
  - Is this useful?
    - yes, we can track the unknown ship in the sea
    - but only when θ is known
- So how do we estimate θ?
  - this problem is known as *parameter inference* for HMMs
    - or *model calibration*, *system identification*
  - very crucial in practice
    - you cannot do filtering/prediction/smoothing without θ
    - often ad-hoc calibration methods are used

Introduction to parameter estimation

- We are interested in **principled** inferential methods or procedures
  - Bayesian
  - Maximum likelihood
- Inference can be performed either
  - on-line
  - batch (or offline)
- We need to use PFs within algorithms that are meant to perform inference for θ.

Introduction to parameter estimation

- Some algorithms
  - Likelihood methods
    - optimisation based
    - gradient based
    - expectation maximisation
  - Bayesian methods
    - **naive approach**: augment state $x_{0:n}$ with θ and do filtering
    - Pseudo marginal MCMC methods: Particle MCMC, Particle Gibbs
    - nested SMC approach: SMC^2

Bayesian inference for HMMs

θ is a random variable, with well chosen prior density $p(\theta)$

- Posterior using Bayes rule directly for θ

$$p(\theta | y_{0:n}) \propto p_\theta(y_{0:n})\, p(\theta).$$

- Posterior when augmenting with $x_{0:n}$

$$p(x_{0:n}, \theta | y_{0:n}) \propto p(\theta)\eta_\theta(x_0) \prod_{k=1}^{n} f_\theta(x_k | x_{k-1}) \prod_{k=0}^{n} g_\theta(y_k | x_k)$$

Bayesian inference for HMMs

- Off-line:
  - we are given a batch of data-points $y_{0:T}$
  - Task: compute/sample from $p\left(x_{0:T}, \theta \mid y_{0:T}\right)$
  - particle MCMC (PMCMC) algorithms
- Sequential case:
  - compute $\left\{p\left(x_{0:n}, \theta \mid y_{0:n}\right)\right\}_{n=0,\ldots,T}$ sequentially
  - use state augmentation: state is $\left(x_{0:n}, \theta\right)$
  - many techniques: from naive to state of the art SMC^2
  - <u>challenge:</u> without adding bias methods not **on-line**

Sequential Bayesian estimation

- Introducing the extended state (X_n, θ_n) with initial density $p(\theta_0)\, \mu_{\theta_0}(x_0)$
- The transition "density" is

$$f_{\theta_n}(x_n | x_{n-1})\, \delta_{\theta_{n-1}}(\theta_n)$$

  i.e. $\theta_n = \theta_{n-1}$.
- Applying a standard SMC algorithm to the Markov process $\{X_n, \theta_n\}_{n \geq 0}$:
  - parameter space would only be explored at the initialization of the algorithm.
  - successive resampling steps, after a certain time n, the approximation $\widehat{p}(d\theta_n | y_{0:n})$ will only contain a single unique value for θ.
  - implicitly requires having to approximate $p_{\theta^{(i)}}(y_{0:n})$ for all the particles $\left\{\theta^{(i)}\right\}$ approximating $p(\theta | y_{0:n})$, hence we expect estimates whose variance will increase at least linearly with n;

Sequential Bayesian estimation

- Pragmatic solutions:
  - use artificial dynamics (Liu and West 2001, Hurzeler and Kunsch 2001),
  - simple example
    $$\theta_n = \theta_{n-1} + \epsilon_n$$
    with ϵ_n being zero mean noise with small variance
    - can tune variance from the particles
  - also can use fixed lag approximations (Kitagawa 96, Kitagawa & Sato 01, Polson et al 08)
    - do not resample paths before $n - L$
      $$\left(\bar{\theta}_n^i, \bar{X}_{n-L+1:n}^i\right) = \left(\theta_n^{a_n(i)}, X_{n-L+1:n}^{a_n(i)}\right),$$
      fixed lag L is a tuning variable
  - use MCMC steps

Sequential Bayesian estimation

- Resample Move: use an MCMC kernel with invariant density $p\left(x_{0:n}, \theta \middle| y_{0:n}\right)$, i.e.

$$\left(X_{0:n}^{(i)}, \theta_n^{(i)}\right) \sim K_n\left(\cdot, \cdot \middle| \overline{X}_{0:n}^{i}, \overline{\theta}_n^{i}\right)$$

  where by construction K_n satisfies

$$p\left(x_{0:n}', \theta' \middle| y_{0:n}\right) = \int p\left(x_{0:n}, \theta \middle| y_{0:n}\right) K_n\left(x_{0:n}', \theta' \middle| x_{0:n}, \theta\right) d\left(x_{0:n}, \theta\right).$$

- In practice set $X_{0:n-L}^{(i)} = \overline{X}_{0:n-L}^{i}$ for some integer $L \geq 1$ and only sample $\theta_n^{(i)}$ and possibly $X_{n-L+1:n}^{(i)}$

Resample Move

- some cases we can use Gibbs step to update the parameter values

$$K_n\left(x'_{0:n},\theta'\middle|\,x_{0:n},\theta\right)=\delta_{x_{0:n}}\left(x'_{0:n}\right)p(\theta'|x_{0:n},y_{0:n}),$$

  where

$$p\left(\theta\middle|\,y_{0:n},x_{0:n}\right)=p(\theta|s_n(x_{0:n},y_{0:n}))$$

  with $s_n(x_{0:n},y_{0:n})$ fixed dimension sufficient statistic.
  - With some variation this has appeared many times: Andrieu et al 1999, Fearnhead 2002, Storvik 2002, Johannes and Polson 2007.
- Elegant, but not robust: it relies on SMC approximations of $p(s_n(x_{0:n},y_{0:n})|y_{0:n})$, and for fixed N, error increases with n.
  - <u>path degeneracy is an issue even if not easy to spot sometimes!</u>
- Challenging for high dimensions ($>5-10$)

A sequential algorithm for infering θ

At time $n = 0$, For all $i \in \{1, ..., N\}$:

- Sample $\theta_0^i \sim p(\cdot)$, $X_0^i \sim q_{\theta_0^i}(x_0 | y_0)$.
- Compute the weights $w_0\left(X_0^i | \theta_0^i\right)$ and set $W_0^i \propto w_0\left(X_0^i | \theta_0^i\right)$
- Resample $\left\{W_0^i, X_0^i, \theta_0^i\right\}$ to obtain N equally-weighted particles $\left\{\frac{1}{N}, \overline{X}_0^i, \bar{\theta}_0^i\right\}$.
- Evolve parameter $\theta_1^i \sim p(\cdot | s_0(\overline{X}_0^i, y_0), \overline{\theta}_n^i)$

At time $n \geq 1$, For all $i \in \{1, ..., N\}$:

- Sample $X_n^i \sim q_{\theta_n^i}(x_n | y_n, \overline{X}_{n-1}^i)$ and set $X_{0:n}^i \leftarrow \left(\overline{X}_{0:n-1}^i, X_n^i\right)$.
- Compute the weights $\omega_n\left(X_{n-1:n}^i | \theta_n^i\right)$ and set $W_n^i \propto \omega_n\left(X_{n-1:n}^i | \theta_n^i\right)$.
- Update sufficient statistics $S_n^i = s_n(\overline{S}_{n-1}^i, X_{0:n}^i, y_n)$
- Resample $\left\{W_n^i, X_{0:n}^i, S_n^i, \theta_n^i\right\}$ to obtain N new equally-weighted particles $\left\{\frac{1}{N}, \overline{X}_{0:n}^i, \overline{S}_n^i, \overline{\theta}_n^i\right\}$.
- Evolve parameter $\theta_{n+1}^i \sim p(\cdot | \overline{S}_n^i, \overline{\theta}_n^i, y_{0:n})$

Numerical example

- We will use again

$$X_n = \rho X_{n-1} + \tau W_n, \; Y_n = X_n + \sigma V_n \tag{1}$$

where $W_n, V_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.
- See Section 7 in Kantas et. al. 2015 (Fig. 5)
- $N = 10^4$ and 50 independent runs

Numerical example: sequential learning variance

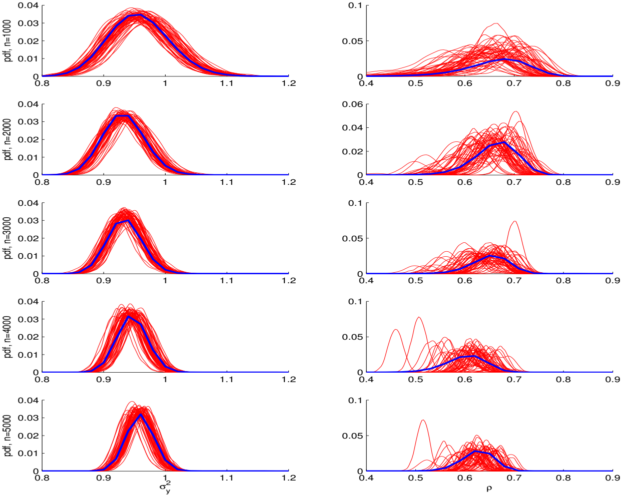

Figure: Particle method with MCMC, $\theta = (\rho, \sigma^2)$;

Particle learning of θ with MCMC steps

- Pros:
  - elegant, no bias introduced
  - in some models can work reasonably in practice especially with informative priors and short n
- Cons:
  - method suffers from path degeneracy and often this is not easy to detect without running multiple runs
  - for reasonable results requires very high N
  - exploration in θ reduces with n
  - Monte Carlo variance increases with n
    - $\mathbb{V}ar\, \hat{p}_{\theta_{\mathbf{0}:n}}(y_{0:n})$ empirically seems be superlinear with n
- State of the art is SMC^2:
  - Use PMCMC targetting $p(x_{0:n}\theta|y_{0:n})$ for the dynamics of θ_n

Bayesian Inference & MCMC: a brief reminder

- Parameter θ is a random variable and Y is some dataset
- Bayes rule: posterior$\propto$ likelihood $\times$ prior

$$p(\theta|Y) \propto p(Y|\theta)p(\theta)$$

- Markov chain Monte Carlo (MCMC): Obtain samples of θ using and appropriate ergodic Markov chain $\{\theta\,(k)\}_{k\geq 0}$ with stationary distribution $p(\theta|Y)$

MCMC with Metropolis Hastings

Sample $\theta(0) \sim p(\cdot)$. At iteration $k \geq 1$

- Sample proposal $\theta' \sim q(\cdot|\theta(k-1))$
- Compute acceptance ratio

$$\alpha(\theta, \theta') = 1 \wedge \frac{p(Y|\theta')p(\theta')q(\theta(k-1)|\theta')}{p(Y|\theta(k-1))p(\theta(k-1))q(\theta'|\theta(k-1))}$$

- With probability $\alpha(\theta, \theta')$ accept proposal setting $\theta(k) = \theta'$, otherwise reject sample and set $\theta(k) = \theta(k-1)$

Bayesian inference for HMMs

- Off-line case: given a batch of data $y_{0:T}$
  - likelihood is $p_\theta(y_{0:T})$
  - Choose a suitable prior density $p(\theta)$ for θ
- Approximate $p(\theta | y_{0:T})$ which is given by

$$p(\theta | y_{0:T}) \propto p_\theta(y_{0:T})\, p(\theta). \tag{2}$$

Naive Metropolis Hastings for HMMs

Sample $\theta^0 \sim p(\cdot)$. At iteration $k > 0$

- Sample proposal $\theta' \sim q(\cdot|\theta)$, where $\theta = \theta(k-1)$.
- Compute acceptance ratio

$$\alpha(\theta, \theta') = 1 \wedge \frac{p_{\theta'}\left(y_{0:T}\right) p\left(\theta'\right) q(\theta|\theta')}{p_{\theta}\left(y_{0:T}\right) p\left(\theta\right) q(\theta'|\theta)}$$

- with probability $\alpha(\theta, \theta')$ accept proposal setting $\theta(k) = \theta'$, otherwise reject sample and set $\theta\left(k\right) = \theta(k-1)$.

Metropolis Hastings for HMMs

- Hard to implement directly as $p_{\theta'}\left(y_{0:T}\right)$ is intractable
- Could use Monte Carlo approximations
  - but hard to justify as a sampler targetting directly $p\left(\theta \mid y_{0:n}\right)$
- Consider instead the joint posterior density

$$p\left(x_{0:T}, \theta \mid y_{0:T}\right) \propto p_\theta\left(x_{0:T}, y_{0:T}\right) p(\theta)$$

- Could use $p_\theta\left(x_{0:T}, y_{0:T}\right) = \eta_\theta\left(x_0\right) \prod_{k=1}^{T} f_\theta\left(x_k \mid x_{k-1}\right) \prod_{k=0}^{T} g_\theta\left(y_k \mid x_k\right)$ to design sampler
  - but mixing could deteriote rapidly with T
  - difficult to find a proposal to break conditional dependencies.

Ideal Marginal Metropolis-Hastings sampler

- The ideal MMH sampler would utilize the following proposal density:

$$q\left(\left(x'_{0:T},\theta'\right)\middle|\left(x_{0:T},\theta\right)\right)=q\left(\theta'\middle|\theta\right)p\left(x'_{0:T}\middle|y_{0:T},\theta'\right)\qquad(3)$$

- The acceptance probability is

$$\begin{aligned}
&1\wedge\frac{p\left(x'_{0:T},\theta'\middle|y_{0:T}\right)q\left(\left(x_{0:T},\theta\right)\middle|\left(x'_{0:T},\theta'\right)\right)}{p\left(x_{0:T},\theta\middle|y_{0:T}\right)q\left(\left(x'_{0:T},\theta'\right)\middle|\left(x_{0:T},\theta\right)\right)}\\
=&1\wedge\frac{p\left(x'_{0:T},\theta'\middle|y_{0:T}\right)q\left(\theta\middle|\theta'\right)p\left(x_{0:T}\middle|y_{0:T},\theta\right)}{p\left(x_{0:T},\theta\middle|y_{0:T}\right)q\left(\theta'\middle|\theta\right)p\left(x'_{0:T}\middle|y_{0:T},\theta'\right)}\\
=&1\wedge\frac{p_{\theta'}\left(y_{0:T}\right)p\left(\theta'\right)q(\theta|\theta')}{p_{\theta}\left(y_{0:T}\right)p\left(\theta\right)q(\theta'|\theta)}
\end{aligned}$$

Particle marginal Metropolis Hastings for HMMs

- Remarkably the acceptance ratio is the same as naive M-H
- Problem: We cannot sample exactly from $p\left(x'_{0:T} \mid y_{0:T}, \theta'\right)$ and we cannot compute $p_{\theta'}\left(y_{0:T}\right)$
- Use Particle Filters and particle approximations
  - particle marginal M-H (PMMH) Sampler
  - data augmentation with PF variables
  - instead of marginalising $x_{0:T}$ to get θ we will marginalise all the variables in the PF

Particle Marginal Metropolis-Hastings sampler

- Consider performing standard MCMC on

$$p\left(\left\{ \left\{ x_n^i, o_n(i) \right\}_{i=1}^{N} \right\}_{n=1}^{T}, \theta \middle| y_{0:T} \right)$$

  i.e. the joint density of the parameter θ and all the simulated variables in the SMC algorithm
- Use proposal

$$q\left(\theta' \middle| \theta \right) \hat{p}\left(\left\{ \left\{ x_n^i, o_n(i) \right\}_{i=1}^{N} \right\}_{n=1}^{T} \middle| y_{0:T}, \theta' \right)$$

- Take an approach that uses appropriate auxiliary variables.
  - $\left\{ \left\{ X_n^i, O_n(i) \right\}_{i=1}^{N} \right\}_{n=1}^{T}$ are included as auxiliary variables and then integrated out.
  - often called pseudo-marginal approach (Andrieu and Roberts 2009)

Particle Marginal Metropolis-Hastings sampler

- The algorithm a valid based on unbiasedness of likelihood

$$\mathbb{E}_N[\hat{p}_{\theta'}(y_{0:T})] = p_{\theta'}(y_{0:T})$$

- This is for any N
- Satisfies detailed balance with

$$p\left(\left\{ \left\{ X_n^i, O_n(i) \right\}_{i=1}^{N} \right\}_{n=1}^{T}, \theta \middle| y_{0:n} \right)$$

- Details: Andrieu, Doucet and Holenstein 2010 particle MCMC paper

Particle Marginal Metropolis-Hastings (PMMH) sampler

At iteration $k = 0$,

- Set $\theta(0) \sim p(\cdot)$.
- Run a PF targeting $p\left(x_{0:T} \middle| y_{0:T}, \theta(0)\right)$, sample $X_{0:T}(0) \sim \widehat{p}\left(\cdot \middle| y_{0:T}, \theta(0)\right)$, and compute estimate $\widehat{Z}_T\left(\theta(0)\right) = \hat{p}_{\theta(0)}\left(y_{0:T}\right)$

At iteration $k \geq 1$

- Sample a proposal $\theta' \sim q\left(\theta \middle| \theta(k-1)\right)$.
- Run a PF targeting $p\left(x'_{0:T} \middle| y_{0:T}, \theta'\right)$, sample $X'_{0:T} \sim \widehat{p}\left(\cdot \middle| y_{0:T}, \theta'\right)$, and compute estimate $\widehat{Z}_T\left(\theta'\right) = \hat{p}_{\theta'}\left(y_{0:T}\right)$.
- Set $\theta(k) = \theta'$, $X_{0:T}(k) = X'_{0:T}$, and store $\widehat{Z}_T\left(\theta(k)\right) = \widehat{Z}_T(\theta')$ with probability

$$1 \wedge \frac{\widehat{Z}_T\left(\theta'\right) p(\theta') q(\theta(k-1)|\theta')}{\widehat{Z}_T\left(\theta(k-1)\right) p(\theta(k-1)) q(\theta'|\theta(k-1))},$$

otherwise set $\theta(k) = \theta(k-1)$, $X_{0:T}(k) = X_{0:T}(k-1)$, $\widehat{Z}_T\left(\theta(k)\right) = \widehat{Z}_T(\theta(k-1))$.

Discussion on PMMH sampler

- The remarkable feature of this algorithm is that the invariant distribution of the Markov chain $\{X_{0:T}(k), \theta(k)\}$ is $p\left(x_{0:T}, \theta \middle| y_{0:T}\right)$ whatever being N.
  - SMC approximations do not introduce any bias.
  - minimal tuning required compared to usual MCMC.
- Unbiasedness of the likelihood is a key requirement
  - so **cannot use adaptive resampling** for PMCMC

Discussion on PMMH sampler

- Often PMCMC chains are “sticky”
  - spend too long on a certain θ
  - this is due to over-estimating $p_\theta(y_{0:T})$ with $\hat{p}_\theta(y_{0:T})$
  - then we need a few samples of $q(\theta' | \theta)$ to get a high enough $\hat{p}_{\theta'}(y_{0:T})$ to move to new state θ
- The higher N the better the mixing properties of the algorithm.
  - the less sticky the chain
  - tradeoff with added computational cost should be balanced
- Under good mixing assumptions the variance of the acceptance rate of the PMMH sampler is proportional to T/N
  - N should roughly increase linearly with T, so computational cost $\mathcal{O}(T^2)$

Numerical example: PMCMC vs sequential approach

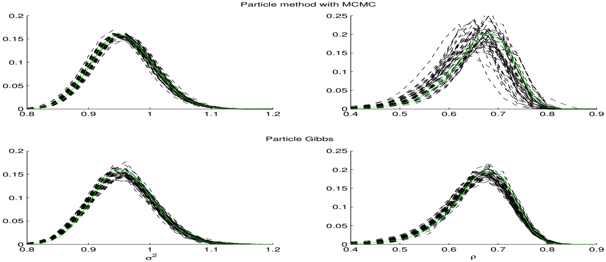

Figure: Estimated marginal posterior densities for $\theta = (\rho, \sigma^2)$ with $T = 10^3$ over 50 runs (black-dotted) versus ground truth (green). Top: Particle method with MCMC, $N = 7.5 \times 10^4$. Bottom: Particle Gibbs with 3000 iterations and $N = 50$.

Discussion

- Online Bayesian estimation notoriously hard
  - no truly online solution available
  - current state of the art: SMC^2
- Offline Bayesian inference for HMMs
  - particle MCMC
  - open challenges: long data, large dimensions, faster samplers for given models

Reading list

- Particle MCMC by Andrieu, Doucet and Holenstein
  - http://www.stats.ox.ac.uk/~doucet/andrieu_doucet_holenstein_pmcmc_mcqmc.pdf
  - http://www.stats.ox.ac.uk/~doucet/andrieu_doucet_holenstein_PMCMC.pdf
- Review on parameter estimation:
  - https://arxiv.org/pdf/1412.8695.pdf