Bayesian Econometrics

Introduction to Sequential Monte Carlo

Andrés Ramírez Hassan

Universidad Eafit
Departamento de Economía

May 3, 2021

Outline

Setting

Signal modeled as Markovian, nonlinear, non-Gaussian, state–space models.
The unobserved signal (hidden states) $\{\theta_t : t \in \mathcal{N}\}$, $\theta_t \in \Theta$ is modeled as a Markov process of initial distribution $\pi(\theta_0)$, and transition equation $\pi(\theta_t|\theta_{t-1})$.
The observations $\{y_t : t \in \mathcal{N}^+\}$, $y_t \in \mathcal{Y}$, are assumed to be conditionally independent, given the process $\{\theta_t : t \in \mathcal{N}\}$, with marginal distribution $f(y_t|\theta_t)$.

- $\pi(\theta_0)$
- $\pi(\theta_t|\theta_{t-1})$, $t \geq 1$
- $f(y_t|\theta_t)$, $t \geq 1$

Objectives

- The updating step given by the filtering distribution, $\pi(\theta_t|y_{1:t}) = \frac{f(y_t|\theta_t)\pi(\theta_t|y_{1:t-1})}{\int f(y_t|\theta_t)\pi(\theta_t|y_{1:t-1})d\theta_t}$ where $\pi(\theta_t|y_{1:t-1}) = \int \pi(\theta_t|\theta_{t-1})\pi(\theta_{t-1}|y_{1:t-1})d\theta_{t-1}$ is the prediction step.
- The joint posterior distribution, $\pi(\theta_{0:t}|y_{1:t}) = \frac{f(y_{1:t}|\theta_{0:t})\pi(\theta_{0:t})}{\int f(y_{1:t}|\theta_{0:t})\pi(\theta_{0:t})d\theta_{0:t}}$ where $\pi(\theta_{0:t+1}|y_{1:t+1}) = \pi(\theta_{0:t}|y_{1:t})\frac{f(y_{t+1}|\theta_{t+1})\pi(\theta_{t+1}|\theta_t)}{f(y_{t+1}|y_{1:t})}$.
- $I(g_t) = E_{\pi(\theta_{0:t}|y_{1:t})}[g_t(\theta_{0:t})] = \int g_t(\theta_{0:t})\pi(\theta_{0:t}|y_{1:t})d\theta_{0:t}$

Examples

- Kalman filter: Gaussian linear model.
- HMM filter: Partially observed finite state-space Markov chains model.

Why not MCMC or IS?

- No well suited for iterative problems
  - On–line prediction
  - Storage restrictions

Importance Sampling

Importance Sampling

$$I(g_t) = \int g_t(\theta_{0:t})\pi(\theta_{0:t}|y_{1:t})d\theta_{0:t}$$
$$\frac{\int g_t(\theta_{0:t})w(\theta_{0:t})q(\theta_{0:t}|y_{1:t})d\theta_{0:t}}{\int w(\theta_{0:t})q(\theta_{0:t}|y_{1:t})d\theta_{0:t}}$$

where $w(\theta_{0:t}) = \frac{\pi(\theta_{0:t}|y_{1:t})}{q(\theta_{0:t}|y_{1:t})}$.

Importance Sampling

Importance Sampling

Simulate N $i.i.d$ particles $\left\{\theta_{0:t}^{(i)}\right\}$ according to $q(\theta_{0:t}|y_{1:t})$. A possible Monte Carlo estimate of $I(f_t)$ is

$$\hat{I}_N(f_t) = \frac{\frac{1}{N}\sum_{i=1}^{N} g_t(\theta_{0:t}^{(i)}) w(\theta_{0:t}^{(i)})}{\frac{1}{N}\sum_{i=1}^{N} w(\theta_{0:t}^{(i)})} = \sum_{i=1}^{N} g_t(\theta_{0:t}^{(i)}) \tilde{w}(\theta_{0:t}^{(i)})$$

where $\tilde{w}_t^{(i)} = \frac{w(\theta_{0:t}^{(i)})}{\sum_{j=1}^{N} w(\theta_{0:t}^{(j)})}$.

Sequential Importance Sampling

SIS

Setting $q(\theta_{0:t}|y_{1:t}) = q(\theta_{0:t-1}|y_{1:t-1})q(\theta_t|\theta_{0:t-1}, y_{1:t})$, iterating $q(\theta_{0:t}|y_{1:t}) = q(\theta_0)\prod_{l=1}^{t} q(\theta_l|\theta_{0:l-1}, y_{1:l})$. Then,

$$\tilde{w}_t^{(i)} \propto \tilde{w}_{t-1}^{(i)} \frac{f(y_t|\theta_t^{(i)})\pi(\theta_t^{(i)}|\theta_{t-1}^{(i)})}{q(\theta_t^{(i)}|\theta_{0:t-1}^{(i)}, y_{1:t})}$$

Important case $q(\theta_{0:t}|y_{1:t}) = \pi(\theta_{0:t}) = \pi(\theta_0)\prod_{l=1}^{t} \pi(\theta_l|\theta_{l-1})$. So,

$$\tilde{w}_t^{(i)} \propto \tilde{w}_{t-1}^{(i)} f(y_t|\theta_t^{(i)})$$

Sequential Importance Sampling

SIS shortcomings

- It is a constrained version of IS.
- IS is usually inefficient in high-dimensional spaces.
- As $t \to \infty$, $\tilde{w}_t^{(i)} \to 0$ for all i, except one.

Origins

Sampling Importance Resampling

- Random variable via a weighted bootstrap (Smith and Gelfand, 1992)
- Sampling Importance Resampling (SIR). Comment by D. Rubin in Tanner and Wong (1987).

Origins

Random variable via a weighted bootstrap (Smith and Gelfand, 1992)

- Given $\pi(\theta|y) = \frac{h(\theta|y)}{\int h(\theta|y)d\theta}$ and a proposal distribution $q(\theta)$, such that we obtain $\theta^{(i)}$ draws from $q(\theta)$, $i = 1, 2, \ldots, N$. Then we define $w(\theta)^{(i)} = \frac{h(\theta|y)}{q(\theta)}$, and $\tilde{w}^{(i)} = \frac{w(\theta)^{(i)}}{\sum_{j=1}^{N} w(\theta)^{(j)}}$.

  Then we draw $\theta^{*(i)}$, N times, from the discrete distribution over $\left\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}\right\}$ with replacement and $\tilde{w}^{(i)}$ as weights.

Origins

Random variable via a weighted bootstrap (Smith and Gelfand, 1992)

- Observe that

$$P(\theta^* \leq a) = \sum_{i=1}^{N} \tilde{w}^{(i)} 1_{(-\infty,a)}(\theta^{(i)})$$
$$= \frac{\frac{1}{N}\sum_{i=1}^{N} w^{(i)} 1_{(-\infty,a)}(\theta^{(i)})}{\frac{1}{N}\sum_{i=1}^{N} w^{(i)}}$$
$$\to \frac{E_q \frac{h(\theta|y)}{q(\theta)} 1_{(-\infty,a)}(\theta)}{E_q \frac{h(\theta|y)}{q(\theta)}} = \int_{-\infty}^{a} \pi(\theta|y) d\theta$$

- A consistent estimator for $\int h(\theta|y)d\theta$ is $N^{-1}\sum_{i=1}^{N} w^{(i)}$.

The Bootstrap filter

Gordon et al. (1993). See Doucet et al. (2001), page 11.

1. Initialization, $t = 0$
   - For $i = 1, 2, \ldots, N$, sample $\theta_0^{(i)} \sim \pi(\theta_0)$, and set $t = 1$.
2. Importance sampling step
   - For $i = 1, 2, \ldots, N$, sample $\tilde{\theta}_t^{(i)} \sim \pi(\theta_t | \theta_{t-1}^{(i)})$, and set $\tilde{\theta}_{0:t}^{(i)} = (\theta_{0:t-1}^{(i)}, \tilde{\theta}_t^{(i)})$.
   - For $i = 1, 2, \ldots, N$, evaluate the importance weights, $w_t^{(i)} = f(y_t | \tilde{\theta}_t^{(i)})$ ((Crassidis and Junkins, 2011, p. 286) suggest $w_t^{(i)} = w_{t-1}^{(i)} \times f(y_t | \tilde{\theta}_t^{(i)})$).
   - Normalize the importance weights.
3. Selection step
   - Resample with replacement N particles $(\theta_{0:t}^{(i)}, i = 1, 2, \ldots, N)$ from the set $(\tilde{\theta}_{0:t}^{(i)}, i = 1, 2, \ldots, N)$ according to the importance weights.
   - Set $t \leftarrow t + 1$, and go to step 2.

The Bootstrap filter

Advantages

- It is very quick and easy to implement.
- It is modular, that is, when changing the problem one need only change the expressions for the importance distribution and the importance weights,
- It can be straightforwardly implemented on a parallel algorithm.
- Allows easily carrying out sequential inference for very complex models.

The Bootstrap filter

Example: (Crassidis and Junkins, 2011, p. 285)

- $\theta_t = 0.5\theta_{t-1} + 25\frac{\theta_{t-1}}{1+\theta_{t-1}^2} + 8cos(1.2t) + v_t$
- $y_t = \frac{\theta_t^2}{20} + u_t$
- $\theta_0 \sim \mathcal{N}(0, \sqrt{10})$, $v_t \sim \mathcal{N}(0, \sqrt{10})$ and $u_t \sim \mathcal{N}(0, \sqrt{1})$

The Bootstrap filter

Implementation

- The proposal (importance function) is the stated transition density ("the prior distribution").
- We can resample when the p.d.f. of the importance weights is degenerate. This can be monitored using the effective sample size.

The Bootstrap filter

Disadvantages

1. The resampling step introduces extra MC variability.
2. The use of the state transition density as importance distribution can often lead to poor performance, which is manifested in a lack of robustness with respect to the values taken by the observed sequence, for example when outliers occur in the data or on the contrary when the variance of the observation noise is small.
3. This procedure is not well suited to sample from $\pi(\theta_{0:t}|y_{1:t})$. This is because most of the particles come from the same ancestor.

The Bootstrap filter

To handle extra MC variability

1. Optimal kernel $q_t(\theta_t|\theta_{t-1}, y_t) = \frac{\pi_t(\theta_t|\theta_{t-1})f(y_t|\theta_t)}{\int \pi_t(\theta_t|\theta_{t-1})f(y_t|\theta_t)d\theta_t}$. This generates that the conditional variance of the weights is zero, given the past history of the particles. Unfortunately, it is intractable in most cases. See Auxiliary Particle Filter, Algorithm 4 in page 908, Cappé et al. (2007).
2. Residual sampling (Liu and Chen, 1995)
3. Systematic resampling (Carpenter et al., 1999)

The Bootstrap filter

To obtain $\pi(\theta_{0:t}|y_{1:t})$

- $\pi(\theta_{0:t}|y_{1:t})$ is very relevant to obtain good estimates of static parameters.
- Fixed–lag approximation (see (Kantas et al., 2009), section 2.3.1 in page 5), Backward smoothing recursions (see Cappé et al. (2007) Algorithm 5, page 914.), Generalized two–filter smoothing (see (Kantas et al., 2009), section 2.3.3 in page 6) can be good solutions to obtain $\pi(\theta_{0:t}|y_{1:t})$.

The Bootstrap filter

To obtain $\pi(\theta_{0:t}|y_{1:t})$

- Maximum a posteriori (MAP)
  $\arg\max_{\theta_{0:T}} \theta_{0:T|0:T}(\theta_{0:T|y_{1:T}}) =$
  $\arg\max_{\theta_{0:T}} \pi_0(\theta_0) \prod_{t=1}^{T} \pi(\theta_t|\theta_{t-1}) \prod_{t=1}^{T} f(y_t|\theta_t)$
  This can be done using Algorithm 6, page 916 in Cappé et al. (2007).
- MCMC steps within SMC. See section 4.2.4, page 15, Kantas et al. (2009). However, it suffers from the standard degeneracy problem.

Estimation of static parameters

Estimation of static parameters (Cappé et al., 2007, Kantas et al., 2009, 2015)

- Batch methods: Data is available for estimation
  - Particle MCMC methods (Andrieu et al., 2010).
    - Particle Independent M–H. Section 2.4.1, (Andrieu et al., 2010).
    - Particle Marginal M–H. Section 2.4.2, (Andrieu et al., 2010) and Section 4.1.1, Kantas et al. (2009).
    - Particle Gibbs sampler. Section 2.4.3, (Andrieu et al., 2010).
  - Particle approximation to likelihood
    a) Expectation-Maximization
    b) Gradient based methods:
    $\theta_{k+1} = \theta_k + \gamma_{k+1} \nabla_\theta l_T(\theta)|_{\theta=\theta_k}$. γ_k is a sequence of small positive real numbers.

Estimation of static parameters

Estimation of static parameters (Cappé et al., 2007, Kantas et al., 2009, 2015)

- On–line estimation
  - Adding MCMC steps among the particles (Gilks and Berzuini, 2001).
  - Treat static parameters as dynamic introducing negligible shocks
  - Bootstrap filter with parameter regeneration: Requires sufficient statistics. This can be done using Algorithm 7, page 919 in Cappé et al. (2007).

Estimation of static parameters

Estimation of static parameters (Kantas et al., 2009, 2015)

1. Gradient methods is preferable if the step–size sequence γ_{k+1} is replace by $-\gamma_{k+1}\Gamma_k^{-1}$ where Γ_k is Hessian of $l_T(\theta)$, which can be computed using SMC techniques. Then, the rate of convergence is quadratic, and faster than the EM which converges linearly. In addition, the gradient algorithm can be implemented even when the M–step cannot be solved in close–form.
2. EM can be preferable if the M–step can be computed analytically. In addition, the EM is numerically more stable and typically computationally cheaper for high dimensional spaces.
3. Both algorithms are locally optimal.

Estimation of static parameters

Estimation of static parameters (Kantas et al., 2009, 2015)

1. MCMC steps among particles are not robust. This is because these algorithms are based on the SMC approximation of $\pi(\theta_{0:t}|y_{0:t})$ whose dimension increases with t. So, they suffer from the standard degeneracy problem. However, these methods cannot completely rule out. For small time horizons, low dimensional parameter space (typically not more than 5-10), informative priors and large number of particles, they can perform well (Kantas et al., 2009).

Nonfiltering Uses of SMC

Population Monte Carlo (Cappé et al., 2004)

- The main objective of PMC is to draw samples of size T of the targeting distribution (π). That is, the support of π is $\mathcal{R}^T$. In this setting the targeting distribution is static.
- PMC borrows from MCMC algorithms for the construction of the proposals, from IS for the construction of appropriate estimators.
- Extending regular importance sampling techniques to cases where the importance distributions for $\theta_t^{(i)}$ may depend on both the sample index t and the iteration index i, thus possibly on past samples, does not modify their validity.

Nonfiltering Uses of SMC

Population Monte Carlo (Cappé et al., 2004)

- The main concern is the proposal distribution.
- PMC produces $i.i.d$ chains. This is a huge advantage over MCMC methods because the latter have acceptance rate which decreases approximately as a power of T.
- The MCMC environment is hasher for adaptive schemes, because adaptivity cancels the Markovian nature of the sequence and thus calls for more elaborate ergodicity results.

Nonfiltering Uses of SMC

Population Monte Carlo (Cappé et al., 2004)

- PMC methods ergodicity is not an issue because the validity is obtained via importance sampling justifications.
- PMC does not require stopping rules as MCMC.
- PMC is very appealing in models with latent variables, such as models where data augmenting is a good strategy, for instance: probit and multinomial models.

Nonfiltering Uses of SMC

Algorithm (Cappé et al., 2004)

- For $i = 1, 2, \ldots, N$
  - For $t = 1, 2, \ldots, T$
    a) Select the generating distribution q_{ti}.
    b) Generate $\theta_t^{(i)} \sim q_{ti}(\theta)$, and compute $w_t^{(i)} = \pi(\theta_t^{(i)}) / q(\theta_t^{(i)})$.
    c) Normalize $w_t^{(i)}$, that is, obtain $\tilde{w}_t^{(i)}$.
    d) Resample T values from $\theta_t^{(i)}$'s with replacement, using $\tilde{w}_t^{(i)}$, to create the sample $(\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_T^{(i)})$.

Nonfiltering Uses of SMC

Algorithm (Cappé et al., 2004)

- A central feature of PMC is that the proposal can be individualized at each step of the algorithm while preserving the validity of the method.
- They can be picked according to the performance of the previous $q_t^{(i-1)}$'s and, in particular, they can depend on the previous sample $(\theta_1^{(i-1)}, \theta_2^{(i-1)}, \ldots, \theta_T^{(i-1)})$. For instance, the $q_t^{(i)}$'s are random walk proposals centered at the $\theta_t^{(i-1)}$'s, with various possible scales chosen from earlier performances.

References I

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Cappé, O., Godsill, S. J., and Moulines, E. (2007). An overview of existing methods and recent advances in sequential monte carlo. *Proceedings of the IEEE*, 95(5):899–924.

Cappé, O., Guillin, A., Marin, J. M., and Robert, C. P. (2004). Population monte carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929.

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation*, 146(1):2–7.

Crassidis, J. L. and Junkins, J. L. (2011). *Optimal estimation of dynamic systems*. CRC press.

References II

Doucet, A., De Freitas, N., and Gordon, N. (2001). An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, chapter 1, pages 3–14. Springer New York, New York.

Gilks, W. R. and Berzuini, C. (2001). Following a moving target—monte carlo inference for dynamic bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):127–146.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113.

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J., Chopin, N., et al. (2015). On particle methods for parameter estimation in state-space models. *Statistical science*, 30(3):328–351.

References III

Kantas, N., Doucet, A., Singh, S. S., and Maciejowski, J. M. (2009). An overview of sequential monte carlo methods for parameter estimation in general state–space models. *IFAC Proceedings Volumes*, 42(10):774–785.

Liu, J. S. and Chen, R. (1995). Blind deconvolution via sequential imputations. *Journal of the American Statistical Association*, 90(430):567–576.

Smith, A. F. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling-–resampling perspective. *The American Statistician*, 46(2):84–88.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.