Parameter Estimation for Hidden Markov models with Intractable Likelihoods[1]

Ajay Jasra

Imperial College London

ABC in London, 5th May

[1] Joint with T. Dean, S. Singh & G. Peters

Introduction

- I will describe a theoretical analysis of a classical estimation procedure.
- We consider the asymptotic consistency and normality of a collection of estimators.
- The analysis is in the context of hidden Markov models (HMMs).
- The novelty, is that one cannot evaluate the likelihood, nor has access to an unbiased estimate of it.

- The analysis is linked to ABC: it was very nicely discussed by Christian Robert on his blog: `http://xianblog.wordpress.com/`.
- This is via the idea of maximizing an approximation of the statistical model, motivated by ABC.

HMMs

- A HMM is a pair of discrete-time processes, $\{X_k\}_{k\geq 0}$ and $\{Y_k\}_{k\geq 0}$.
- The hidden process, $\{X_k\}_{k\geq 0}$, is a Markov chain.
- The observed process $\{Y_k\}_{k\geq 0}$ takes values in $\mathbb{R}^m$.
- Given X_k the Y_k are independent of $Y_0, \dots, Y_{k-1}; X_0, \dots, X_{k-1}$.
- Many real applications: Bioinformatics, econometrics and finance.

Estimation

- Often there is a range of HMMs associated to a vector θ.
- Given $\hat{Y}_1, \dots, \hat{Y}_n$ the objective is to find θ^* that corresponds to the HMM which generated the data.
- A standard approach is maximum likelihood estimation (MLE).

- The MLE is obtained via maximizing the log-likelihood:

$$\hat{\theta}_n = \arg\sup_{\theta\in\Theta} l_n(\theta)$$

where

$$l_n(\theta) := \frac{1}{n}\log p_\theta\left(\hat{Y}_1,\dots,\hat{Y}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\log p_\theta(\hat{Y}_i|\hat{Y}_1,\dots,\hat{Y}_{i-1}).$$

- In most cases one can seldom evaluate the likelihood of the data analytically.

Example: Known Likelihood

- For example, with $k \geq 1$, $X_0 = 0$

$$\begin{aligned} Y_k &= X_k + \sigma_1 \epsilon_k \\ X_k &= X_{k-1} + \sigma_2 \nu_k \end{aligned}$$

  where ϵ_k, ν_k are i.i.d. standard normals.
- Then, $\theta = (\sigma_1, \sigma_2)$. The likelihood is Gaussian and the state-process under-goes a Gaussian Markov transition.

- In the notation to follow:

$$g_\theta(y|x) = \frac{1}{\sigma_1\sqrt{2\pi}}\exp\{-\frac{1}{2\sigma_1^2}(y-x)^2\}$$

and

$$q_\theta(x,x') = \frac{1}{\sigma_2\sqrt{2\pi}}\exp\{-\frac{1}{2\sigma_2^2}(x'-x)^2\}.$$

- Throughout the talk, the HMM is time-homogeneous.

- There are a variety of techniques for estimating the likelihood.
- For example, methods using sequential Monte Carlo (SMC) or Markov chain Monte Carlo.
- The consistency and asymptotic normality of MLEs are well-understood (see e.g. Cappé et al. (2005) and the references therein).

Intractable Likelihoods

- For some applications the conditional density of Y_k given X_k is intractable, this density:
  - cannot be evaluated analytically
  - there is no unbiased estimator.
- Throughout, I will say intractable likelihood, and this refers to g_θ.
- In this case, the standard methods cannot be applied and it is the objective to investigate new/existing ideas.

Some Existing Ideas

- One such approach is the convolution particle filter. We have found this to inaccurate in practice.
- Indirect inference. This method is likely to be very expensive.
- Approximate Bayesian computation. It provides an approximation of the true model.
- Note - we will still perform classical inference.

Approximate Bayesian Computation

- Given the data $\hat{Y}_1, \ldots, \hat{Y}_n$ one approximates the likelihood function via

$$\mathbb{P}_\theta\left(d\left(Y_1, \ldots, Y_n; \hat{Y}_1, \ldots, \hat{Y}_n\right) \leq \epsilon\right)$$

  where $d(\cdot;\cdot)$ is a metric and $\epsilon > 0$ reflects the accuracy of the approximation.
- Often one uses a summary statistic, which is discussed later on.

- We study the approximation

$$\mathbb{P}_\theta\left(Y_1 \in B^\epsilon_{\hat{Y}_1}, \ldots, Y_n \in B^\epsilon_{\hat{Y}_n}\right)$$

  where B^ϵ_y denotes the ball of radius ϵ centered around the point y (see McKinley et al. (2009)).
- That is:

$$\int \prod_{i=1}^{n} \mathbb{I}_{B^\epsilon_{\hat{Y}_i}}(y_i) g_\theta(y_i|x_i) q_\theta(x_{i-1}, x_i) \nu(dy_{1:n}) \pi_0(dx_0) \mu(dx_{1:n}).$$

- It is shown in Jasra et al. (2011) that the approximation, as ϵ falls, converges to the true model.
- This approach retains the Markovian structure.
- This facilitates simpler MCMC and SMC implementation (which we do not discuss)).
- Only requires one to sample the likelihood but not to evaluate it.

- We provide a justification of an ABC MLE similar to MLE via asymptotic consistency.
- Our approach is based on the fact that the ABC MLE is MLE using the likelihoods of a collection of perturbed HMMs.
- This implies that the ABC MLE should inherit its behaviour from MLE.
- We also consider a noisy variant with results concerning the asymptotic behaviour of the MLE.

- We establish that the ABC MLE has an asymptotic bias. This can be made small by choosing ϵ small.
- We show that a noisy ABC MLE is asymptotically consistent and has an asymptotic Fisher information matrix less than that of MLE.
- Thus, it is shown that the noisy ABC suffers from a relative loss of information and hence statistical efficiency.
- As $\epsilon \downarrow 0$ the Fisher information of the noisy ABC MLE converges to that of the MLE.

- The key to our analysis is that

$$\mathbb{P}_\theta\left(d\left(Y_1,\ldots,Y_n;\hat{Y}_1,\ldots,\hat{Y}_n\right)\leq\epsilon\right)\propto$$

$$\int_{\mathcal{X}^{n+1}}\left[\prod_{k=1}^{n} q_\theta(x_{k-1},x_k)g_\theta^\epsilon(\hat{Y}_k|x_k)\right]\pi_0(dx_0)\mu(dx_{1:n})$$

where $\{q_\theta, g_\theta\}$ are the original densities and

$$g_\theta^\epsilon(y|x)=\frac{1}{\nu\left(B_y^\epsilon\right)}\int_{B_y^\epsilon} g_\theta(y'|x)\,\nu(dy').$$

Estimation

- Given $\epsilon > 0$ and data $\hat{Y}_1, \ldots, \hat{Y}_n$, estimate θ^* with

$$\hat{\theta}_n^{\epsilon} = \arg\sup_{\theta \in \Theta} p_{\theta}^{\epsilon}\left(\hat{Y}_1, \ldots, \hat{Y}_n\right)$$

where the function $p_{\theta}^{\epsilon}(y_1, \ldots, y_n)$ is the likelihood function of the perturbed HMM.

Results

- As $\epsilon \downarrow 0$, we show that the ABC MLE converges to the MLE.
- Moreover, under some additional assumptions, the rate of decrease of the bias is linear in ϵ.
- These results are to be expected as, clearly, we are dealing with an approximation.

Noisy ABC

- Given $\epsilon > 0$ and data $\hat{Y}_1, \dots, \hat{Y}_n$ estimate θ^* with

$$\tilde{\theta}_n^\epsilon = \arg\sup_{\theta \in \Theta} p_\theta^\epsilon \left(\hat{Y}_1 + \epsilon \hat{Z}_1, \dots, \hat{Y}_n + \epsilon \hat{Z}_n \right) \qquad (1)$$

  where
  - $p_\theta^\epsilon (y_1, \dots, y_n)$ is the likelihood function for the perturbed HMM
  - $\hat{Z}_1, \dots, \hat{Z}_n$ are i.i.d. samples from $\mathcal{U}_{B_0^1}$ [the uniform distribution on the unit ball at the origin].

- The parameter estimator in (1) is equivalent to the parameter estimator given by

$$\tilde{\theta}_n^{\epsilon} = \arg\sup_{\theta\in\Theta} \mathbb{P}_\theta\left(Y_1^{\epsilon} \in B^{\epsilon}_{\hat{Y}_1+\epsilon\hat{Z}_1}, \dots, Y_n^{\epsilon} \in B^{\epsilon}_{\hat{Y}_n+\epsilon\hat{Z}_n}\right).$$

- That is, the ABC approximation with perturbed observations. This is also described by Fearnhead & Prangle (2010).

Results

- We have proved that the noisy MLE is asymptotically consistent and normal.
- In addition, that the asymptotic variance is strictly greater than the MLE.
- As $\epsilon \downarrow 0$, we show that the variance of the noisy ABC MLE converges to the MLE quadratically.

Kernels

- In many scenarios ABC approximations are constructed via:

$$\int \left[\prod_{k=1}^{n} q_\theta(x_{k-1}, x_k) \phi\left(\frac{\hat{Y}_k - y_k}{\epsilon} \right) g_\theta(y_k | x_k) \right] \pi_0(dx_0)\, \mu(dx_{1:n}) \nu(dy_{1:n})$$

  where ϕ is a probability density.
- It is possible to prove all the previous results, under some assumptions on ϕ.
- It is also possible to prove the results when one uses summary statistics.

Summary

- In this talk I have presented a theoretical justification for an ABC approximation of HMMs with intractable likelihoods.
- Asymptotic bias of the ABC MLE has been discussed.
- A noisy ABC method has been introduced.
- The noisy ABC MLE is asymptotically consistent.
- The theory extends to kernels and summary statistics.

Acknowledgements

- Joint Work with:
  - Tom Dean (University of Cambridge)
  - Sumeetpal Singh (University of Cambridge)
  - Gareth Peters (UNSW)