Discussion

Peter Bickel

U.C.Berkeley

September 8, 2008

SAMSI Workshop

Moulines, Crisan

General HMM Model

General HMM Model

- $(X_n, Y_n), n \geq 1$
- X_n: Markov
- Y_j: independent of $(X_k, Y_k), k \neq j$, given X_j

Then,

$$X_j | (Y_1, \ldots, Y_\infty) \ \leftrightarrow Q$$

is Markov.

General HMM Model

- When is Q strongly mixing?

  i.e.

  (1) $\rho(t) \to 0$ as $t \to \infty$, where

  $\rho(t) = \sup \left\{ |Q(AB) - Q(A)Q(B)| : A \in \mathcal{B}_1^k, B \in \mathcal{B}_{t+k}^{\infty} \right\},$

  $\mathcal{B}_a^b$: sigma field generated by $(X_a, \ldots, X_b)$.

  If $\Gamma(t) = \sup_{k,u,v} \left\{ \left| Q\left[X_{k+t} \in B \middle| X_k = u\right] - Q\left[X_{k+t} \in B \middle| X_k = v\right] \right| \right\},$

  then $\Gamma(t) \geq \rho(t)$.

  (2) $\Gamma(t) \to 0$ as $t \to \infty$ (Stability, strong forgetting)

Importance

1. Validity of particle filter marginals uniformly in time as $n \to \infty$, where n is ensemble size.
2. Frequentist inference for parametric HMM models Baum,Petrie(1966),B-Ritov,B-Ritov-Ryden,Jensen -Pedersen,.up to X compact, Y arbitrary case

Questions

0. Can Douc,Moulines,Ritov(2008) and later results be used to make inference results general?
1. Bayesian inference in this context *e.g.* Bernstein-von Mises theorems
2. State space fields even Markov random fields inference ?

Particle Filters and High Dimensional Data

Main Ideas

1) Represent probability distribution P on $\mathcal{X}$ by a sample
$\mathcal{E} = \left\{ \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(B)} \right\}$
Approximate (for any g integrable) $\int g dP$ by $\frac{1}{B} \sum_{b=1}^{B} g\left(\mathbf{X}^{(b)}\right)$

Particle Filters and High Dimensional Data

2) Recursive sampling

Given $\{P_t\}$, $t = 0, 1, 2, \ldots$, generate recursively

$$
\begin{aligned}
\mathcal{E}_0 &= \left\{X_0^{(1)}, \ldots, X_0^{(B)}\right\} \text{ From } P_0 \\
\mathcal{E}_1 &= \left\{X_1^{(1)}, \ldots, X_1^{(B)}\right\} \text{ From } P_1
\end{aligned}
$$

by a Monte Carlo operation.

e.g. importance sampling from P_0, etc.

Method(simplest)

1. Given $\mathcal{E}_{t-1} = \left\{ X_{t-1}^{(1)}, \ldots, X_{t-1}^{(B)} \right\}$

$$IID \quad \approx \quad p_{t-1}\left(\cdot | Y_0^{t-1} \right)$$

Generate $\mathcal{E}_{tI} = \left\{ X_{tI}^{(1)}, \ldots, X_{tI}^{(B)} \right\}$ by,
for $j = 1, \ldots, B$,

(i) Pick $X_{t-1}^{(k_j)}$ at random from $\mathcal{E}_{t-1}$

(ii) Generate $X_{tI}^{(j)}$ according to $p_t\left(\cdot | X_{t-1}^{(k_j)} \right)$

Method(simplest)

2. Generate $\mathcal{E}_t$ so that for $j = 1, \ldots, B$
   $\mathbb{P}\left[X_t^{(j)} = X_{tI}^{(k)}\right] \propto q\left(Y_t | X_{tI}^{(k)}\right)$

   If $v_k = q_t\left(Y_t | X_{tI}^{(k)}\right),\ k = 1, \ldots, B,$
   put $X_t^{(k)} = X_{tI}^{(k)}$ with probability
   $w_b = \frac{v_b}{\sum_{k=1}^{B} v_k},\ b = 1, \ldots, B$

Method(simplest)

NB : $X_t^{(1)}, \ldots, X_t^{(B)}$ are not iid $p_t\left(\cdot|Y_0^t\right)$.

In fact, if

$\mathcal{E}_{t-1} \leftrightarrow \tilde{p}_{t-1}\left(\cdot|Y_1^{t-1}\right)$

I :

$$\begin{aligned}\mathcal{E}_{tI} &\leftrightarrow \tilde{p}_{tI} \\ \tilde{p}_{tI}(x) &= \frac{1}{B}\sum_{b=1}^{B} p_t\left(x|X_{t-1}^{(b)}\right)\end{aligned}$$

II :

$$\begin{aligned}\mathcal{E}_{t} &\leftrightarrow \tilde{p}_{t} \\ \tilde{p}_{t}(x) &= \frac{\sum_{b=1}^{B} \delta_{X_{tI}^{(b)}}(x) q_t\left(Y_t|X_{tI}^{(b)}\right)}{\sum_{j=1}^{B} q_t\left(Y_t|X_{tI}^{(j)}\right)}\end{aligned}$$

Collapse

Observed empirically that

if $\mathcal{X} = \mathbb{R}^d$, d large, even if $B > d$,

after one (or a few steps), if $w_{(1)} \leq \cdots \leq w_{(B)}$ are ordered weights,

$w_{(B)} \approx 1$.

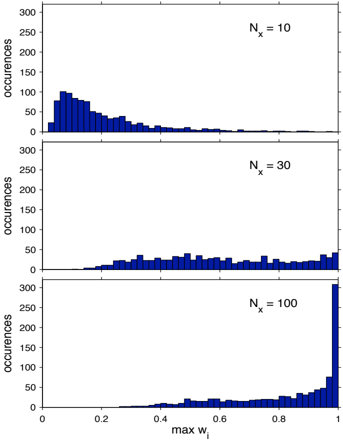
Prototype

(i) $\mathbf{X}_{d\times 1} \sim \mathcal{N}(\mathbf{0}, J_d)$

(ii) $\mathbf{Y} = \mathbf{X} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \perp \mathbf{X}, \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, J_d)$

$$\begin{aligned} \mathbf{X} &\leftrightarrow \mathcal{E}_{1I} \\ \mathbf{Y}|\mathbf{X} &\leftrightarrow \prod_{j=1}^{d} \varphi(y_j - x_j) = q_t(\mathbf{y}|\mathbf{x}) \end{aligned}$$

Prototype



Histogram of $\max w_i$ for $N_x = 10, 30, 100$ and $N_e = 10^3$ from the particle-filter simulations described in text

$$[N_e = 10^3,\ \mathbf{x}^t \sim N(0, \mathbf{I}),\ N_y = N_x,\ \mathbf{H} = \mathbf{I} \text{ and } \boldsymbol{\varepsilon} \sim N(0, \mathbf{I})]$$

$$N_x \equiv d,\ N_e \equiv B,\ \frac{B}{d} = 100,\ 33,\ 10,\ \frac{\log_e B}{d} = 0, 69,\ 0.23,\ 0.069$$

High Dimensional Data

A. In prototype situation, if $\frac{\log B}{d} \to 0$,

$\mathbb{E}(w_{(B)}) = 1 - \frac{2}{\sqrt{5}}\sqrt{\frac{\log B}{d}}(1 + o(1))$.

In this case, for g bounded, continuous,

$$\frac{1}{B}\sum_{b=1}^{B} g\left(\mathbf{X}^{(b)}\right) \Rightarrow g\left(\mathbf{X}^{(k)}\right)$$

where $k = \underset{j}{\operatorname{argmin}} \left|\mathbf{X}^{(j)}\right|, \mathbf{X}^{(j)} \in \mathcal{E}_{1I}, \mathbf{X}^{(j)} \;\sim\; \tilde{p}_{1I}$

High Dimensional Data

B. If $\frac{\log B}{d} \to \infty$ and g is bounded,

$$\left| \frac{1}{B} \sum_{b=1}^{B} g\left(\mathbf{X}^{(b)}\right) - \mathbb{E} g\left(\mathbf{X} | \mathbf{Y}\right) \right| \to_p 0$$

where $\mathbb{E}$ is under correct $P(\mathbf{X}|\mathbf{Y})$.

Question

How does one deal with situations where,

(i) $\mathbf{X}$ and $\mathbf{Y}$ are very high dimensional

(ii) There is no parametric model

But

(iii) The distributions live mainly on low dimensional structures