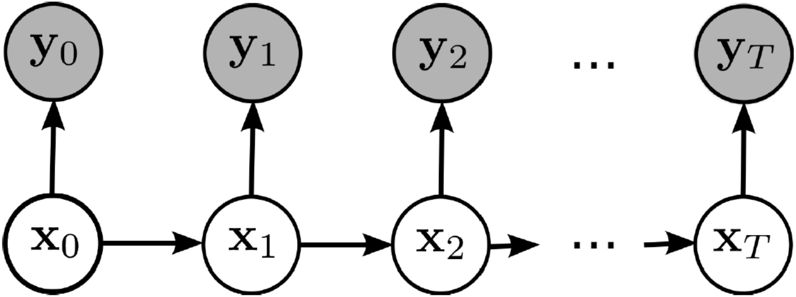
Hidden Markov models and Stochastic Filtering

Introduction

- Hidden Markov models are used in many disciplines
  - statistics, econometrics, engineering, neuroscience, medical & life sciences...
- Sequential Bayesian inference is natural for these models
  - known as (non-linear or stochastic or Bayesian) filtering

Introduction: what is a state space model?

A graphical model:



Hidden Markov Models (HMM)

- **General Model definition**: A bivariate Markov Chain $\{X_n, Y_n\}_{n\geq 0}$, where $\{X_n\}_{n\geq 0}$ is a latent part of the stochastic process and $\{Y_n\}_{n\geq 0}$ is the observed part. Each of them are defined on a general state space $\mathcal{X}$ $(\subseteq \mathbb{R}^{n_x})$ and $\mathcal{Y}$ $(\subseteq \mathbb{R}^{n_y})$ respectively.
- We will look at a particular but still very general (and natural) case:
    - all spaces $\mathcal{X}$,$\mathcal{Y}$ etc. are continuous.
    - initialisation with distribution $X_0 \sim \eta_\theta(\cdot)$
    - X_n is a discrete time Markov chain with transition density f_θ
    - $Y_n|X_n$ is i.i.d. with likelihood density g_θ
- In principle, we can extend methodology to more complex model structure cases but will not look at this here.

HMM description

- Let $X_{0:n} = (X_0, \ldots, X_n)$ and $0 \leq n \leq T$. We write the model formally as:

$$\mathbb{P}\left[X_n \in A | (X_{0:T} = x_{0:T}, Y_{0:T} = x_{0:T})\right] = \int_A f_\theta(x|x_{n-1})dx,$$
$$\mathbb{P}\left[Y_n \in B | (X_{0:T} = x_{0:T}, Y_{0:T} = x_{0:T})\right] = \int_B g_\theta(y|x_n)dy,$$

(under a canonical probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\Omega = \prod_{n=0}^{\infty} (\mathcal{X} \times \mathcal{Y})^n$ and $\mathcal{F}$ Borel σ algebra.)

HMM description

- Let $X_{0:n} = (X_0, \ldots, X_n)$ and $0 \leq n \leq T$. We write the model formally as:

$$\mathbb{P}\left[X_n \in A | (X_{0:T} = x_{0:T}, Y_{0:T} = x_{0:T})\right] = \int_A f_\theta(x|x_{n-1})dx,$$
$$\mathbb{P}\left[Y_n \in B | (X_{0:T} = x_{0:T}, Y_{0:T} = x_{0:T})\right] = \int_B g_\theta(y|x_n)dy,$$

(under a canonical probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\Omega = \prod_{n=0}^{\infty} (\mathcal{X} \times \mathcal{Y})^n$ and $\mathcal{F}$ Borel σ algebra.)

- Or more casually:

$$\begin{aligned} X_n &\sim f_\theta(\cdot|x_{n-1}), \\ Y_n &\sim g_\theta(\cdot|x_n), \end{aligned} \tag{1}$$

- for the parameter we assume $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$, Θ is open.

HMM model parameters

- In fact previous equation is more like:

$$\begin{aligned} X_n &\sim f(\cdot|x_{n-1}, \theta), \\ Y_n &\sim g(\cdot|x_n, \theta), \end{aligned} \tag{2}$$

- θ are static model parameters, i.e. not time varying or dynamic.
- we will usually assume $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$, Θ is open.

State Space Models (SSMs)

- This class of models includes many nonlinear and non-Gaussian time series models such as

$$X_{n+1} = \psi_\theta\left(X_n, V_{n+1}\right), \;\; Y_n = \phi_\theta\left(X_n, W_n\right), \tag{3}$$

where $\{V_n\}_{n\geq 1}$ and $\{W_n\}_{n\geq 0}$ are arbitrary iid noise sequences and $(\psi_\theta, \; \phi_\theta)$ are nonlinear functions.

- θ are model parameters.
- Often these models are time discretisations of continuous time models, e.g. stochastic differential equations.

$$\begin{aligned} dX_t &= b_\theta(X_t)dt + \sigma_\theta(X_t)dB_t \\ dY_t &= h_\theta(X_t)dt + dV_t \end{aligned}$$

with V_t, B_t independent Brownian motions.

Linear Gaussian HMMs

- Linear Gaussian State Space Model

$$X_n = \alpha X_{n-1} + \sigma_v V_n,$$
$$Y_n = X_n + \sigma_w W_n,$$

where $W_n, V_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $X_0 \sim \mathcal{N}(0,1)$
In this case: $\theta = (\alpha, \sigma_v, \sigma_w)$

- Multidimensional case:

$$\begin{aligned} X_n &= \mathrm{A} X_{n-1} + \mathrm{B} W_n, \\ Y_n &= \mathrm{C} X_n + \mathrm{D} V_n, \end{aligned}$$

W_n, V_n iid zero mean Gaussian vectors. Some constraints need to be placed for $\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}$ to achieve irreducibility, identifiability etc.

Stochastic volatility

- Stochastic Volatility Model

$$\begin{aligned} X_n &= \alpha X_{n-1} + \sigma_v V_n, \\ Y_n &= \beta \exp(\tfrac{X_n}{2}) W_n, \end{aligned}$$

  where $W_n, V_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $X_0 \sim \mathcal{N}(0,1)$,
- In this case: $\theta = (\alpha, \sigma_v, \beta)$

Target tracking models

- Trajectory planning for Bearings Only Tracking. Let X_n denote the state of an arbitrary moving target, A_n the position of an observer:

$$\begin{aligned} X_n &= GX_{n-1} + HW_n, \\ Y_n &= \tan^{-1}\left(\frac{X_n(1) - A_n(1)}{X_n(3) - A_n(2)}\right) + V_n, \end{aligned}$$

W_n, V_n iid zero mean,

$$G = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}, H = \begin{bmatrix} \frac{T^2}{2} & 0 \\ T & 0 \\ 0 & \frac{T^2}{2} \\ 0 & T \end{bmatrix}.$$

Stochastic epidemic models

- SIR model for population of size N in cont. time
- State $X_t = (S_t, I_t, R_t)$

$$S_t = S_0 - \mathcal{P}^1\left(\frac{\lambda}{N}\int_0^t S(r)I(r)dr\right)$$
$$I_t = I_0 + \mathcal{P}^1\left(\frac{\lambda}{N}\int_0^t S(r)I(r)dr\right) - \mathcal{P}^2\left(\gamma\int_0^t I(r)dr\right)$$
$$R_t = R_0 + \mathcal{P}^2\left(\gamma\int_0^t I(r)dr\right)$$

  with $\mathcal{P}^1, \mathcal{P}^2$ independent standard Poisson process
- Observe

$$Y_{n\Delta} \sim \mathrm{Bin}(I_{n\Delta}, p)$$

- $\theta = (X_0, \lambda, \gamma, p)$
- Note $\mathcal{R}_0 = \frac{\lambda}{\gamma}$

Stochastic kinetic Lotka-Volterra models

- Predator-Prey model: two species X^1 (prey) and X^2 (predator)

$$\mathbb{P}\left[X^1_{t+\delta} = x^1_t + 1, X^2_{t+\delta} = x^2_t \middle| X^1_t = x^1_t, X^2_t = x^2_t\right] = \alpha x^1_t \delta + o(\delta)$$
$$\mathbb{P}\left[X^1_{t+\delta} = x^1_t - 1, X^2_{t+\delta} = x^2_t + 1 \middle| X^1_t = x^1_t, X^2_t = x^2_t\right] = \beta x^1_t x^2_t \delta + o(\delta)$$
$$\mathbb{P}\left[X^1_{t+\delta} = x^1_t, X^2_{t+\delta} = x^2_t - 1 \middle| X^1_t = x^1_t, X^2_t = x^2_t\right] = \gamma x^2_t \delta + o(\delta)$$

  and observe

$$Y_n = X^1_{n\Delta} + V_n$$

  V_n iid zero mean Gaussian.

- $\theta = (\alpha, \beta, \gamma)$ are reaction rate constants

State Estimation problem

- Assume **for the time being that the parameter θ is *known***. We will deal with this later in the course.

Objective: Given observation sequence $y_{0:n}$ estimate $x_{0:n}$ or x_n **on-line**

- By on-line we mean:
  - n might eventually become very large, so need to compute estimate recursively
  - at time n, compute estimate of x_n sequentially as a function of y_n, previous estimates and observations.
  - all this using fixed computational and memory cost per time step.

State Estimation

- One option is using point estimation:
  - quadratic loss functions $L(\theta, Z) = \|X_n - Z\|^2$ lead to estimating the posterior mean as:

$$\hat{x}_n = \arg\min_Z \mathbb{E}\left[\|X_n - Z\|^2 \middle| Y_{0:n} \right]$$

  - similarly $L(\theta, Z) = \kappa 1_{\|X_n - Z\| \geq \epsilon}$ lead to estimating posterior mode.
  - ...
- Alternatively use Bayesian viewpoint: and try to approximate directly full posterior

$$\pi_n(\cdot) = \mathbb{P}[X_n \in \cdot | Y_{0:n}]$$

and thus quantify uncertainty.

Bayesian Filtering

Objective: Compute posterior $\pi_n(\cdot) = \mathbb{P}[X_n \in \cdot | Y_{0:n}]$ sequence on-line as y_n becomes available.

Aim:

- Approach is Bayesian as objects of interests are probability distributions on unknown variables.
- Note posterior is defined on the the **marginal space** of the hidden state.
  - we are infering $X_n | Y_{0:n}$
- Densities of interest
  - filtering density $p(x_n | y_{0:n})$
  - smoothing density $p(x_n | y_{0:T})$, $T > n$
  - prediction density $p(x_{n+p} | y_{0:n})$, $p \geq 1$

Bayesian Filtering

- filtering density $p(x_n|y_{0:n})$
  - useful for tracking hidden signals, navigation, etc.
- smoothing density $p(x_n|y_{0:T})$, $T > n$
  - useful for model calibration, backtracking (i.e. when/where did storm start?), etc.
- prediction density $p(x_{n+p}|y_{0:n})$, $p \geq 1$
  - useful for generating forecasts, prediction, trading etc.

Alert: Model choice is very critical for performance in real life application. Even perfect Bayesian filtering will not work well for bad models.

What is a filter or a particle filter?

- Often the distribution $\pi_n(\cdot)$ or its filtering density $p(x_n|y_{0:n})$ are referred to as the ***filter***
- If π_n is known at each time n, then state inference problem is solved
- **Problem**: one cannot compute it analytically most of the times
  - need numerical approximations
- Some exceptions:
  - finite spaces (integrals are sums)
  - linear Gaussian models (Kalman filter)

What is a filter or a particle filter?

- Most successful methods use simulation:
  - approximate $\pi_n(\cdot)$ using a Monte Carlo approach

- We need to define a procedure that generates samples $\{X_n^i\}_{i=1}^N$ for each time n and approximate π_n as

$$\hat{\pi}_n(\cdot) = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_n^i}(\cdot)$$

where $\delta_{X_n^i}(dx)$ is a Dirac point measure centred at X_n^i.

- **Particle filters** achieve this
  - using importance sampling with resampling to get $\hat{\Pi}_n(\cdot) = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_{\mathbf{0}:n}^i}(\cdot)$

Bayesian Filtering and the path space

- In the most general case the object of interest is the whole path $X_{0:n} | \, Y_{0:n}$ and the **joint filtering distribution**

$$\mathbb{P}\left[X_{0:n} \in \cdot | \, Y_{0:n}\right]$$

  - often statistians call this *smoothing distribution*
  - Marginal can be straight-forwardly derived from path space and marginalisation.
- Similarly some densities of interest in **path space**
  - joint filtering density $p(x_{0:n}|y_{0:n})$
  - joint smoothing density $p(x_{0:n}|y_{0:T})$, $T > n$
  - joint prediction density $p(x_{0:n+p}|y_{0:n})$, $p \geq 1$

Bayesian Filtering

- Why do we care on the path space?
  - some applications require it
  - useful for analysis of Monte Carlo numerical schemes as typically algorithms operate on a path space.
- Clarification:
  - we will use terms like joint or path filtering density for $p(x_{0:n}|y_{0:n})$ to distinguish with **marginal** $p(x_n|y_{0:n})$.

Bayesian Filtering

Next part:

- Look at filtering recursions
  - these will be used to approximate filter on-line
  - will look separately at marginal and path space case

Bayesian Filtering on the marginals

- From the model we know initial condition:
  - initial distribution $\eta(\cdot)$ or initial state x_0 (so $\eta = \delta_{x_0}$).
- At time n say we are given
  - observed data $y_{0:n}$,
  - the **filter** at time $n-1$, $p(x_{n-1}|y_{0:n-1})$
- inference about the state X_n may done recursively in two steps
  - prediction: Chapman Kolmogorov
  - update: Bayes rule

Filtering: Recursive Formulation

- Prediction

$$
\begin{aligned}
p_\theta\left(x_n \mid y_{0:n-1}\right) &= \int f_\theta\left(x_n \mid x_{n-1}\right) p_\theta\left(x_{n-1} \mid y_{0:n-1}\right) dx_{n-1} \\
&= \int p_\theta\left(x_n, x_{n-1} \mid y_{0:n-1}\right) dx_{n-1}
\end{aligned}
$$

Filtering: Recursive Formulation

- Prediction

$$
\begin{aligned}
p_\theta\left(x_n \mid y_{0:n-1}\right) &= \int f_\theta\left(x_n \mid x_{n-1}\right) p_\theta\left(x_{n-1} \mid y_{0:n-1}\right) dx_{n-1} \\
&= \int p_\theta\left(x_n, x_{n-1} \mid y_{0:n-1}\right) dx_{n-1}
\end{aligned}
$$

- Update

$$
\begin{aligned}
p_\theta\left(x_n \mid y_{0:n}\right) &= \frac{p_\theta\left(x_n \mid y_{0:n-1}\right) g_\theta\left(y_n \mid x_n\right)}{p_\theta\left(y_n \mid y_{0:n-1}\right)} \\
&= \frac{p_\theta\left(x_n \mid y_{0:n-1}\right) g_\theta\left(y_n \mid x_n\right)}{\int p_\theta\left(x_n \mid y_{0:n-1}\right) g_\theta\left(y_n \mid x_n\right) dx_n} \\
&= \frac{p_\theta(x_n, y_n \mid y_{0:n-1})}{p_\theta(y_n \mid y_{0:n-1})}
\end{aligned}
$$

Bayesian filtering

- Update procedure is Bayesian
  - Prior is predictor $p_\theta\left(x_n | y_{0:n-1}\right)$
  - Likelihood is $g_\theta\left(y_n | x_n\right)$
  - Evidence $p_\theta\left(y_n | y_{0:n-1}\right)$
- computation can be done analytically
  - If model is linear and Gaussian (one very special case); this is the **Kalman filter.**
  - if $\mathcal{X}, \mathcal{Y}$ are discrete state spaces
    - integrals are sums, densities (row) vectors and kernels matrices.

Bayesian Filtering on the path space

- Given observed data $y_{0:n}$, inference about the states $X_{0:n}$ may be based on the following posterior density:

$$p_\theta\left(x_{0:n} \mid y_{0:n}\right) = \frac{p_\theta\left(x_{0:n}, y_{0:n}\right)}{p_\theta\left(y_{0:n}\right)} \tag{4}$$

where

$$p_\theta\left(x_{0:n}, y_{0:n}\right) = \eta_\theta\left(x_0\right) \prod_{k=1}^{n} f_\theta\left(x_k \mid x_{k-1}\right) \prod_{k=0}^{n} g_\theta\left(y_k \mid x_k\right) \tag{5}$$

and the *marginal likelihood*, $p_\theta\left(y_{0:n}\right)$, is given by

$$p_\theta\left(y_{0:n}\right) = \int p_\theta\left(x_{0:n}, y_{0:n}\right) dx_{0:n}. \tag{6}$$

Filtering recursion in the path space

- Prediction

$$p_\theta\left(x_{0:n}\middle|\, y_{0:n-1}\right) = f_\theta\left(x_n\middle|\, x_{n-1}\right) p_\theta\left(x_{0:n-1}\middle|\, y_{0:n-1}\right)$$

- Update

$$\begin{aligned} p_\theta\left(x_{0:n}\middle|\, y_{0:n}\right) &= \frac{p_\theta\left(x_{0:n}\middle|\, y_{0:n-1}\right) g_\theta\left(y_n\middle|\, x_n\right)}{\int p_\theta\left(x_{0:n}\middle|\, y_{0:n-1}\right) g_\theta\left(y_n\middle|\, x_n\right) dx_{0:n}} \\ &= \frac{p_\theta\left(x_{0:n}\middle|\, y_{0:n-1}\right) g_\theta\left(y_n\middle|\, x_n\right)}{p_\theta\left(y_n\middle|\, y_{0:n-1}\right)} \end{aligned}$$

Filtering Recursion

- In recursive form:

$$p_\theta\left(x_{0:n}\middle|\, y_{0:n}\right) = p_\theta\left(x_{0:n-1}\middle|\, y_{0:n-1}\right) f_\theta\left(x_n\middle|\, x_{n-1}\right) g_\theta\left(y_n\middle|\, x_n\right) \frac{p_\theta(y_{0:n-1})}{p_\theta(y_{0:n})}$$

- Note

$$p_\theta\left(y_n\middle|\, y_{0:n-1}\right) = \frac{p_\theta(y_{0:n})}{p_\theta(y_{0:n-1})}$$

is the normalising constant in denominator.

The marginal and recursive likelihoods

- Recall some definitions:
  - the marginal likelihood: $p_\theta(y_{0:n})$
  - the predictive or recursive likelihood: $p_\theta(y_n | y_{0:n-1})$
- Both are very important quantity for parameter inference.
- Important identity:

$$p_\theta(y_{0:n}) = \prod_{k=0}^{n} p_\theta(y_k | y_{0:k-1})$$

Discussion

- In the heart of filtering lies the problem of numerical integration
  - The most common solution is to use simulation
    - particle or ensemble methods
  - takes advantage of more computational power available, more suitable for higher dimensions
  - other alternatives: quadrature type methods, Gaussian approximations, projection and assumed density/moments methods, spectral or other numerical methods for (S)PDE