State-space models

Hidden Markov models

Dr. Jarad Niemi

Iowa State University

September 20, 2017

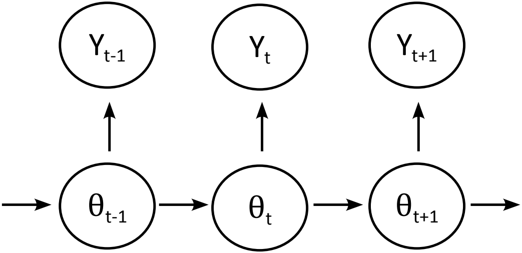
Structure

Observation equation:

$$Y_t \quad = f_t(\theta_t, v_t) \qquad\qquad Y_t \quad \sim p_t(y_t|\theta_t, \dots)$$

State transition (evolution) equation:

$$\theta_t \quad = g_t(\theta_{t-1}, w_t) \qquad\qquad \theta_t \quad \sim p_t(\theta_t|\theta_{t-1}, \dots)$$

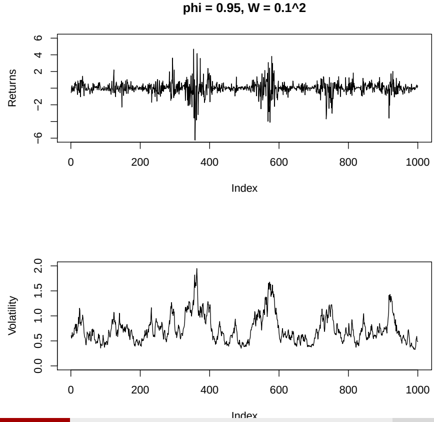


Notation and terminology

Observation equation:	$Y_t = f_t(\theta_t, v_t)$
Observations:	Y_t
Observation (measurement) error:	v_t
State transition (evolution) equation:	$\theta_t = g_t(\theta_{t-1}, w_t)$
Latent (unobserved) state:	θ_t
Evolution noise	w_t

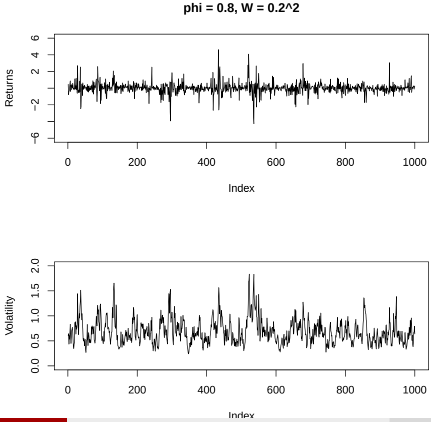
Stochastic volatility

$$\begin{aligned} y_t &\sim N(0, \sigma_t^2) \\ \log \sigma_t &\sim N(\mu + \phi[\log \sigma_{t-1} - \mu], W) \end{aligned}$$



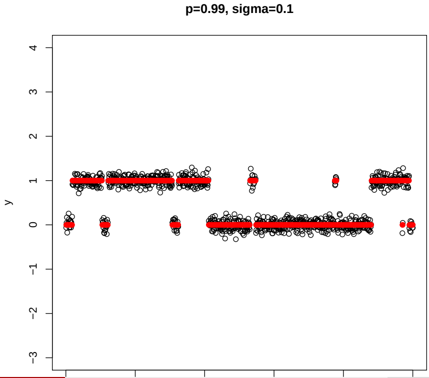
Stochastic volatility

$$\begin{aligned} y_t &\sim N(0, \sigma_t^2) \\ \log \sigma_t &\sim N(\mu + \phi(\log \sigma_{t-1} - \mu), W) \end{aligned}$$



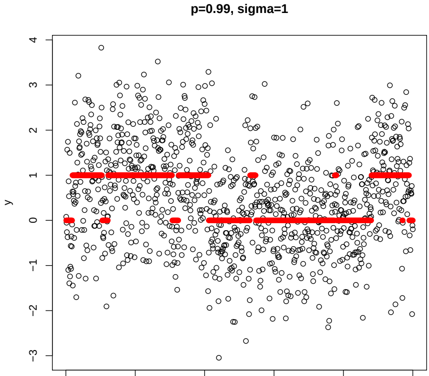
Markov switching model

$$
\begin{aligned}
y_t &\sim N(\theta_t, \sigma^2) \\
\theta_t &\sim p\delta_{\theta_{t-1}} + (1-p)\delta_{1-\theta_{t-1}} \\
\theta_0 &= 0
\end{aligned}
$$



Markov switching model

$$
\begin{aligned}
y_t &\sim N(\theta_t, \sigma^2) \\
\theta_t &\sim p\delta_{\theta_{t-1}} + (1-p)\delta_{1-\theta_{t-1}} \\
\theta_0 &= 0
\end{aligned}
$$



Goals:

- Filtering
- Smoothing
- Forecasting

What do we know?

- $p(y_t|\theta_t)$ for all t
- $p(\theta_t|\theta_{t-1})$ for all t
- $p(\theta_0)$

In principle, we could have subscripts for the distributions/densities, e.g.

- $p_t(y_t|\theta_t)$ for all t
- $p_t(\theta_t|\theta_{t-1})$ for all t

to indicate that the form of the distribution/density has changed. But, most in most models the form stays the same and only the state changes with time.

For simplicity, we will assume a time-homogeneous process and therefore remove the subscript.

Filtering

Goal: $p(\theta_t|y_{1:t})$ where $y_{1:t} = (y_1, y_2, \ldots, y_t)$ (filtered distribution)

Recursive procedure:

- Assume $p(\theta_{t-1}|y_{1:t-1})$
- Prior for θ_t

$$\begin{aligned} p(\theta_t|y_{1:t-1}) &= \int p(\theta_t, \theta_{t-1}|y_{1:t-1}) d\theta_{t-1} \\ &= \int p(\theta_t|\theta_{t-1}, y_{1:t-1}) p(\theta_{t-1}|y_{1:t-1}) d\theta_{t-1} \\ &= \int p(\theta_t|\theta_{t-1}) p(\theta_{t-1}|y_{1:t-1}) d\theta_{t-1} \end{aligned}$$

- One-step ahead predictive distribution for y_t

$$\begin{aligned} p(y_t|y_{1:t-1}) &= \int p(y_t, \theta_t|y_{1:t-1}) d\theta_t \\ &= \int p(y_t|\theta_t, y_{1:t-1}) p(\theta_t|y_{1:t-1}) d\theta_t \\ &= \int p(y_t|\theta_t) p(\theta_t|y_{1:t-1}) d\theta_t \end{aligned}$$

- Filtered distribution for θ_t

$$p(\theta_t|y_{1:t}) = \frac{p(y_t|\theta_t, y_{1:t-1}) p(\theta_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} = \frac{p(y_t|\theta_t) p(\theta_t|y_{1:t-1})}{p(y_t|y_{1:t-1})}$$

What do we know now?

- $p(y_t|\theta_t)$ for all t
- $p(\theta_t|\theta_{t-1})$ for all t
- $p(\theta_0)$
- $p(\theta_t|y_{1:t-1})$ for all t
- $p(y_t|y_{1:t-1})$ for all t

Smoothing

Goal: $p(\theta_t|y_{1:T})$ for $t < T$

- Backward transition probability $p(\theta_t|\theta_{t+1}, y_{1:t})$

$$\begin{aligned} p(\theta_t|\theta_{t+1}, y_{1:T}) &= p(\theta_t|\theta_{t+1}, y_{1:t}) \\ &= \frac{p(\theta_{t+1}|\theta_t, y_{1:t})p(\theta_t|y_{1:t})}{p(\theta_{t+1}|y_{1:t})} \\ &= \frac{p(\theta_{t+1}|\theta_t)p(\theta_t|y_{1:t})}{p(\theta_{t+1}|y_{1:t})} \end{aligned}$$

- Recursive smoothing distributions $p(\theta_t|y_{1:T})$ starting from $p(\theta_T|y_{1:T})$

$$\begin{aligned} p(\theta_t|y_{1:T}) &= \int p(\theta_t, \theta_{t+1}|y_{1:T})d\theta_{t+1} \\ &= \int p(\theta_{t+1}|y_{1:T})p(\theta_t|\theta_{t+1}, y_{1:T})d\theta_{t+1} \\ &= \int p(\theta_{t+1}|y_{1:T})\frac{p(\theta_{t+1}|\theta_t)p(\theta_t|y_{1:t})}{p(\theta_{t+1}|y_{1:t})}d\theta_{t+1} \\ &= p(\theta_t|y_{1:t})\int \frac{p(\theta_{t+1}|\theta_t)}{p(\theta_{t+1}|y_{1:t})}p(\theta_{t+1}|y_{1:T})d\theta_{t+1} \end{aligned}$$

Forecasting

Goal: $p(y_{t+k}, \theta_{t+k}|y_{1:t})$

$$p(y_{t+k}, \theta_{t+k}|y_{1:t}) = p(y_{t+k}|\theta_{t+k})p(\theta_{t+k}|y_{1:t})$$

Recursively, given $p(\theta_{t+(k-1)}|y_{1:t})$

$$\begin{aligned} p(\theta_{t+k}|y_{1:t}) &= \int p(\theta_{t+k}, \theta_{t+(k-1)}|y_{1:t})\, d\theta_{t+(k-1)} \\ &= \int p(\theta_{t+k}|\theta_{t+(k-1)}, y_{1:t})p(\theta_{t+(k-1)}|y_{1:t})d\theta_{t+(k-1)} \\ &= \int p(\theta_{t+k}|\theta_{t+(k-1)})p(\theta_{t+(k-1)}|y_{1:t})d\theta_{t+(k-1)} \end{aligned}$$

Filtering in a Markov switching model

$$\begin{array}{rcl} y_t & \sim & N(\theta_t, \sigma^2) \\ \theta_t & \sim & p\delta_{\theta_{t-1}} + (1-p)\delta_{1-\theta_{t-1}} \\ \theta_0 & = & 0 \end{array}$$

- Note: $p(\theta_t = 1) = 1 - p(\theta_t = 0)$ for all t
- Suppose $q = p(\theta_{t-1} = 1|y_{1:t-1})$. What is $p(\theta_t = 1|y_{1:t-1})$?

$$p(\theta_t = 1|y_{1:t-1}) = \sum_{k=0}^{1} p(\theta_t = 1|\theta_{t-1} = k)p(\theta_{t-1} = k|y_{1:t-1}) = (1-p)(1-q) + pq = p_1$$

- What is $p(\theta_t = 1|y_{1:t-1})$?

$$p(\theta_t = 0|y_{1:t-1}) = \sum_{k=0}^{1} p(\theta_t = 0|\theta_{t-1} = k)p(\theta_{t-1} = k|y_{1:t-1}) = p(1-q) + (1-p)q = p_0$$

- What is $p(y_t|y_{1:t-1})$?

$$p(y_t|y_{1:t-1}) = \sum_{k=0}^{1} p(y_t|\theta_t = k)p(\theta_t = k|y_{1:t-1}) = p_0 N(y_t; 0, \sigma^2) + p_1 N(y_t; 1, \sigma^2)$$

- What is $p(\theta_t = 1|y_{1:t})$?

$$p(\theta_t = 1|y_{1:t}) = \frac{p(y_t|\theta_t = 1)p(\theta_t = 1|y_{1:t-1})}{p(y_t|y_{1:t-1})} = \frac{p_1 N(y_t; 1, \sigma^2)}{p_0 N(y_t; 0, \sigma^2) + p_1 N(y_t; 1, \sigma^2)}$$

Hidden Markov model

Definition

A hidden Markov model (HMM) is a state-space model whose state is finite.

(Note: this is not a universal definition.)

So let

- $\pi_t^{t'}$ be the probability distribution for the state at time t given information up to time t', e.g. $\pi_{t,i}^{t'} = P(\theta_t = i | y_{1:t'})$.
- P be the transition probability matrix, e.g. P_{ij} is the probability of moving from state i to state j in 1 time step.
- $p(y_t|\theta_t)$ be the observation density or mass function.

Inference in a hidden Markov model

Assume π_0^0 is given.

- What is forecast distribution at time t given only π_0^0, i.e. π_t^0? Recursively, we have

$$\pi_t^0 = \pi_{t-1}^0 P.$$

  Alternatively, we have

$$\pi_t^0 = \pi_0 P^t \qquad P^t = P^{t-1} P \quad \text{and} \quad P^1 = P$$

- What is the filtered distribution at time t, i.e. $\pi_{t,i}^t$? Find this recursively via

$$\pi_{t,i}^t \propto p(y_t|\theta_t = i)\pi_{t-1}^{t-1} \cdot P_{\cdot,i}$$

Although smoothing can be useful, it is often of more use in Bayesian analyses to perform backward sampling.

Joint posterior

The joint distribution for $\theta = (\theta_0, \theta_1, \ldots, \theta_T)$ can be decomposed as

$$\begin{aligned} p(\theta|y) &= p(\theta_0, \theta_1, \ldots, \theta_T | y_{1:T}) \\ &= p(\theta_T|y_{1:T}) \prod_{t=T}^{1} p(\theta_{t-1}|\theta_t, y_{1:T}) \\ &= p(\theta_T|y_{1:T}) \prod_{t=T}^{1} p(\theta_{t-1}|\theta_t, y_{1:t-1}) \end{aligned}$$

where

$$\begin{aligned} p(\theta_{t-1}|\theta_t, y_{1:t-1}) &= \frac{p(\theta_t|\theta_{t-1}, y_{1:t-1})p(\theta_{t-1}|y_{1:t-1})}{p(\theta_t|y_{1:t-1})} \\ &= \frac{p(\theta_t|\theta_{t-1})p(\theta_{t-1}|y_{1:t-1})}{p(\theta_t|y_{1:t-1})} \\ &\propto p(\theta_t|\theta_{t-1})p(\theta_{t-1}|y_{1:t-1}) \end{aligned}$$

Backward sampling

The joint distribution for θ can be decomposed as

$$p(\theta|y) = p(\theta_T|y_{1:T}) \prod_{t=1}^{T} p(\theta_{t-1}|\theta_t, y_{1:t-1})$$

and

$$p(\theta_{t-1}|\theta_t, y_{1:t-1}) \propto p(\theta_t|\theta_{t-1})p(\theta_{t-1}|y_{1:t-1}).$$

Suppose we have all the filtered distributions, i.e. π_t^t for $t = 0, \ldots, T$.

An algorithm to obtain a joint sample for θ is

1. Sample $\theta_T \sim p(\theta_T|y_{1:T})$ which is a discrete distribution with $P(\theta_T = i|y_{1:T}) = \pi_{T,i}^T$.
2. For $t = T, \ldots, 1$, sample θ_{t-1} from a discrete distribution with

$$P(\theta_{t-1} = i|\theta_t, y_{1:t-1}) \propto P_{i,\theta_t}\pi_{T-1,i}^{T-1} = \frac{P_{i,\theta_t}\pi_{T-1,i}^{T-1}}{\sum_{i'=1}^{S} P_{i',\theta_t}\pi_{T-1,i'}^{T-1}}.$$

Markov model

Consider a Markov model where the states are observed directly, but the transition probability matrix Ψ is unknown. If the sequence of states are $y_{1:t} = (y_1, \ldots, y_t)$, we are interested in the posterior

$$p(\Psi | y_{1:t}).$$

Since this is a row stochastic matrix Ψ, we have

$$\sum_{j=1}^{S} \Psi_{ij} = 1 \quad \forall\, i.$$

So what priors are reasonable for Ψ?

Priors for row stochastic matrices

One option is a set of independent Dirichlet distributions for each row, i.e. let $\Psi_{i\cdot}$ be the ith row of Ψ, then

$$\Psi_{i\cdot} \sim Dir(A_i)$$

where A_i is a vector of length S and A is the matrix with rows A_i.

Do we want more structure here?

- sparsity (many zero elements)
- similarity between rows

Dirichlet distribution

The Dirichlet distribution (named after Peter Gustav Lejeune Dirichlet), i.e. $P \sim Dir(a)$, is a probability distribution for a probability vector of length H. The probability density function for the Dirichlet distribution is

$$p(P; a) = \frac{\Gamma(a_1 + \cdots + a_H)}{\Gamma(a_1) \cdots \Gamma(a_H)} \prod_{h=1}^{H} p_h^{a_h - 1}$$

where $p_h \geq 0$, $\sum_{h=1}^{H} p_h = 1$, and $a_h > 0$.

Letting $a_0 = \sum_{h=1}^{H} a_h$, then some moments are

- $E[p_h] = \frac{a_h}{a_0}$,
- $V[p_h] = \frac{a_h(a_0 - a_h)}{a_0^2(a_0+1)}$,
- $Cov(p_h, p_k) = -\frac{a_h a_k}{a_0^2(a_0+1)}$, and
- mode$(p_h) = \frac{a_h - 1}{a_0 - H}$ for $a_h > 1$.

A special case is $H = 2$ which is the beta distribution.

Conjugate prior for multinomial distribution

The Dirichlet distribution is the natural conjugate prior for the multinomial distribution. If

$$Y \sim Mult(n, \pi) \qquad \text{and} \qquad \pi \sim Dir(a)$$

then

$$\pi | y \sim Dir(a + y).$$

Some possible default priors are

- $a = 1$ which is the uniform density over π,
- $a = 1/2$ is Jeffreys prior for the multinomial,
- $a = 1/S$ and
- $a = 0$, an improper prior that is uniform on $\log(\pi_h)$. The resulting posterior is proper if $y_h > 0$ for all h.

Dirichlet priors for Markov models

Let A be the hyperparameter with rows A_i such that

$$\Psi_i \stackrel{ind}{\sim} Dir(A_i)$$

and C be the count matrix of observed transitions, i.e. C_i is the count vector of transitions from i to all states and C_{ij} is the count of transitions from i to j.

The posterior distribution $p(\Psi|y_t)$ is fully conjugate with $A' = A + C$ such that

$$\Psi_i|y \stackrel{ind}{\sim} Dir(A_i') \stackrel{d}{=} Dir(A_i + C_i)$$

where A_i' is the ith row of A'.

Inference for HMM with unknown transition matrix Ψ

Suppose we have a HMM with unknown transition matrix Ψ. How can we perform posterior inference?

If we assume $\Psi_i \stackrel{ind}{\sim} Dir(A)$, then a Gibbs sampling approach is

1. Sample $\theta_{1:t}|\Psi, y \sim \prod_{t=1}^{T} p(\theta_{t-1}|\theta_t, y_{1:t}, \Psi)$.
2. For $i = 1, \ldots, S$, sample $\Psi_i|\theta, y \stackrel{ind}{\sim} Dir(A_i + C_i)$ where C_i is the count vector of transitions from i to all states.