Introduction to Algorithmic Trading Strategies Lecture 2

Hidden Markov Trading Model

Haksun Li
haksun.li@numericalmethod.com
www.numericalmethod.com

References

- Algorithmic Trading: Hidden Markov Models on Foreign Exchange Data. Patrik Idvall, Conny Jonsson. University essay from Linköpings universitet/Matematiska institutionen; Linköpings universitet/Matematiska institutionen. 2008.
- A tutorial on hidden Markov models and selected applications in speech recognition. Rabiner, L.R. Proceedings of the IEEE, vol 77 Issue 2, Feb 1989.
- Hidden Markov Models for Time Series: An Introduction Using R. Walter Zucchini, Iain L. MacDonald. 2009.

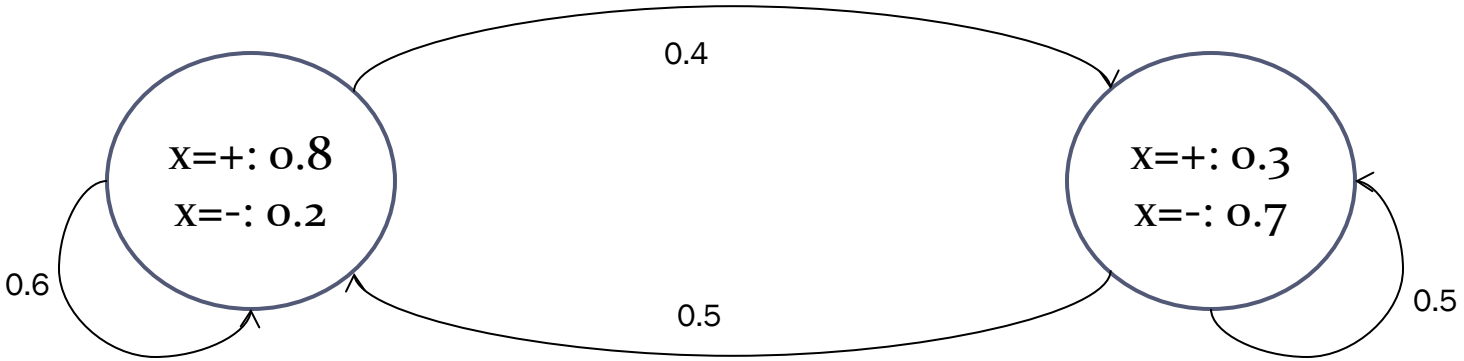
Bayes Theorem

- Bayes theorem computes the posterior probability of a hypothesis H after evidence E is observed in terms of
  - the prior probability, $P(H)$
  - the prior probability of E, $P(E)$
  - the conditional probability of $P(E|H)$
- $P(H|E) = \frac{P(E|H)}{P(E)} P(H) = \frac{P(E|H)}{P(E|H)*P(H)+P(E|\neg H)*P(\neg H)} P(H)$

Bayes Theorem Examples

- A rare event may have occurred with a high probability if the chance of the evidence is also rare. “scaled"
  - P(Jesus resurrection) = very small
  - P(apostle conversion) = very small, also
  - P(Jesus resurrection | apostle conversion)
    - ≈ P(Jesus resurrection)/ P(apostle conversion)
    - ≈ not too small and in fact quite probable
- The occurrence of a highly likely consequence does not mean that the event may have occurred. The probability needs to be “discounted” by the background probability.
  - P(Pattern | Rare) = 98%
  - P(Pattern | ¬Rare) = 5%
  - P(Rare) = 0.1%
  - P(Rare | Pattern) = ?

Markov Chain

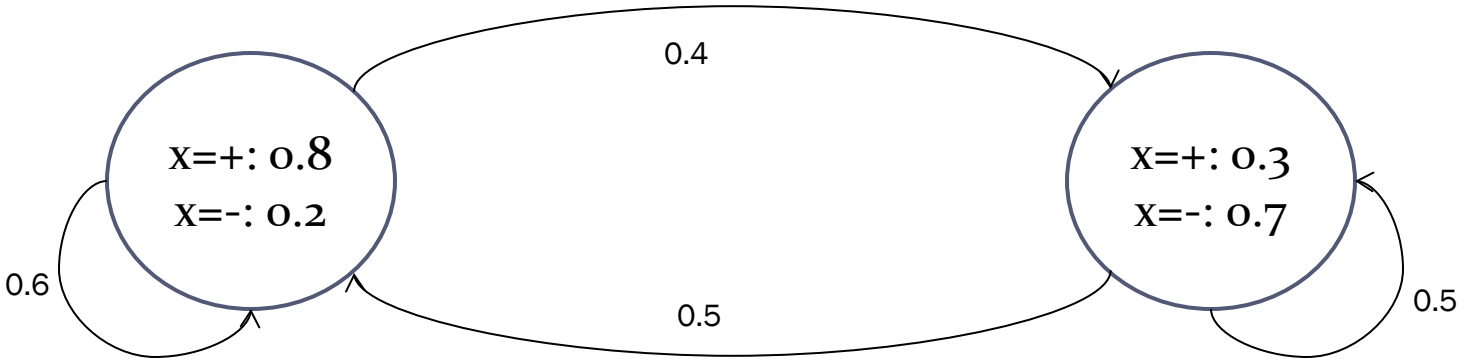


Markov Property

- The conditional probability distribution of future states of the process (conditional on both past and present states) depends only upon the present state, not on the sequence of events that preceded it.
  - $P(x_t|q_t, \cdots, q_1, x_{t-1}, \cdots, x_1) = P(x_t|q_t)$
- Consistent with the weak form of the efficient market hypothesis.

Matrix Notations

- A two-state Markov chain.

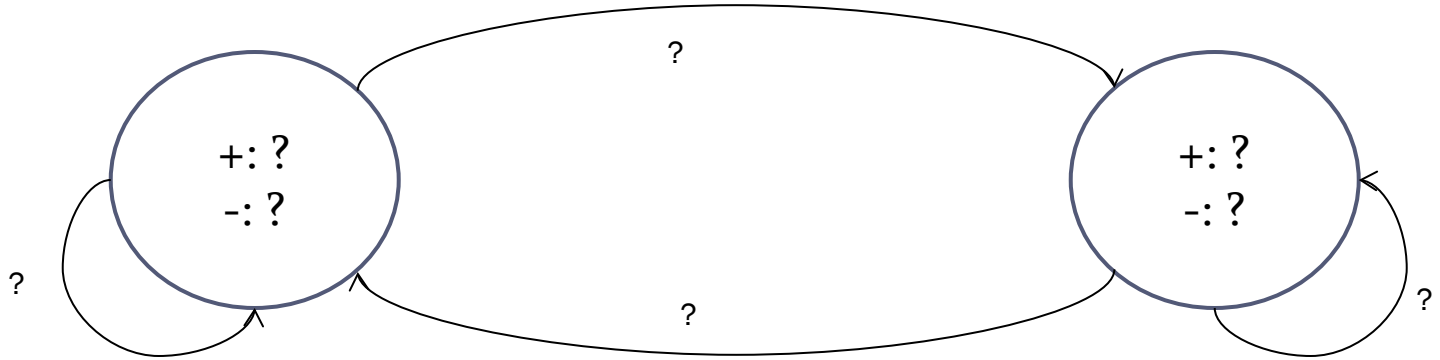


- transition probability matrix $A = \begin{bmatrix} 0.6 & 0.4 \\ 0.5 & 0.5 \end{bmatrix}$
- conditional probability matrix $P_x = \mathrm{diag}(p_1(x), \ldots, p_N(x))$
  - $P_+ = \begin{bmatrix} 0.8 & 0 \\ 0 & 0.3 \end{bmatrix}$, $P_- = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.7 \end{bmatrix}$

Examples

- What is the probability of observing the sequence $S = \{s_1, s_1, s_2\}$
  - $P(S|\text{Model}) = \text{P}(s_1, s_1, s_2|\text{Model})$
  - $= \text{P}(s_1|\text{Model}) \times \text{P}(s_1|s_1,\text{Model}) \times \text{P}(s_2|s_1,\text{Model})$
  - $= 1 \times 0.6 \times 0.6 \times 0.4$
  - $= 0.144$
- $P(X_1 = +, X_2 = +, X_3 = +) =$
  $\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\pi_i p_i(1) a_{ij} p_j(1) a_{jk} p_k(1)$
  $= \Pi P(1) A P(1) \text{A} P(1) 1'$

Hidden Markov Model



Hidden Markov Model

- Only observations are observable (duh).
- World states may not be known (hidden).
  - We want to model the hidden states as a Markov Chain.
- HMM in general does not satisfy the Markov property.

Problems

- Likelihood
  - Given the parameters, ϑ, and an observation sequence, X, compute $P(X|\vartheta)$.
- Decoding
  - Given the parameters, ϑ, and an observation sequence, X, determine the best hidden sequence Q.
- Learning
  - Given an observation sequence, X, and HMM structure, learn ϑ.

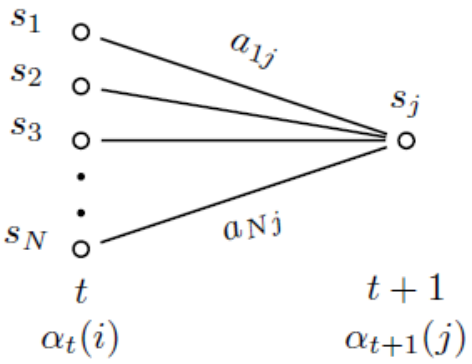
Likelihood Solutions

Likelihood By Enumeration

- $P(X|\vartheta) = \sum_{\{q\}'s} P(X, Q|\vartheta)$
- $= \sum_{\{q\}'s} P(X|Q, \vartheta) \times P(X|\vartheta)$
- $P(X|Q, \vartheta) = \prod_{t=1}^{T} P(x_t|q_t, \vartheta)$
- $P(Q|\vartheta) = \pi_{q_1} \times a_{q_1 q_2} \times a_{q_2 q_3} \times \cdots \times a_{q_{T-1} q_T}$
- But… this is not computationally feasible due to the need to enumerate all possible (finite) state sequences.

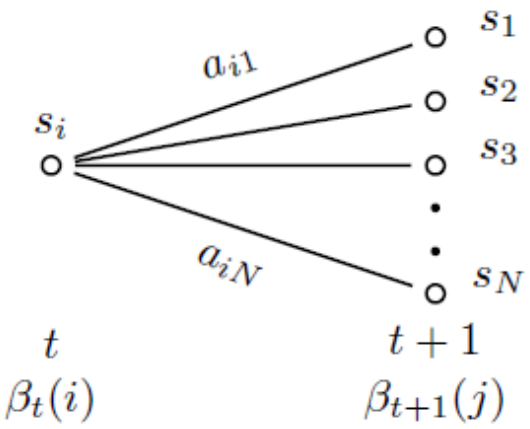
Forward Procedure

- $\alpha_t(i) = P(x_1, x_2, \cdots, x_t, q_t = i|\vartheta)$
  - the probability of the partial observation sequence until time t and the system in state s_i at time t.
- Initialization
  - $\alpha_1(i) = \pi_i p_i(x_1)$
    - p_i: the conditional distribution of x in s_i
- Induction
  - $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] p_j(x_{t+1})$
- Termination
  - $P(X|\vartheta) = \sum_{i=1}^{N} \alpha_T(i)$, the likelihood



Backward Procedure

- $\beta_t(i) = P(x_{t+1}, x_{t+2}, \cdots, x_T | q_t = i, \vartheta)$
  - the probability of the system in state i at time t, and the partial observations from then onward till time t
- Initialization
  - $\beta_T(i) = 1$
- Induction
  - $\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, p_j(x_{t+1}) \beta_{t+1}(j)$



Decoding Solutions

Decoding Solutions

- Given the observations and model, the probability of the system in state i is:
- $\gamma_t(i) = P(q_t = i|X, \vartheta)$
- $= \frac{P(q_t=i,X|\vartheta)}{P(X|\vartheta)}$
- $= \frac{\alpha_t(i)\beta_t(i)}{P(X|\vartheta)}$
- $= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}$

Maximizing The Expected Number Of States

- $q_t = \text{argmax}_{1 \le i \le N}[\gamma_t(i)]$
- This determines the most likely state at every instant, t, without regard to the probability of occurrence of sequences of states.

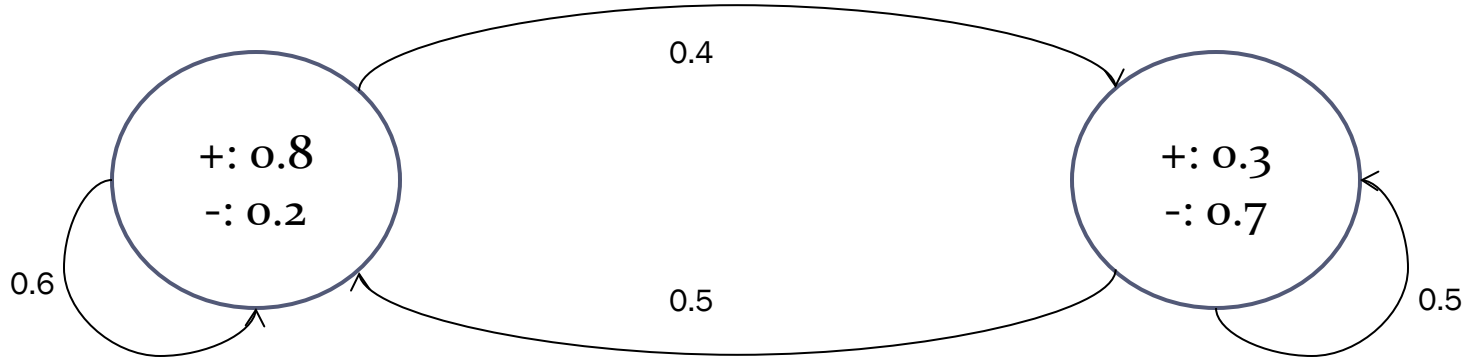
Viterbi Algorithm

- The maximal probability of the system travelling these states stopping at state i and generating these observations:
- $\delta_t(i) = \max[P(q_1, q_2, \cdots, q_t = i, x_0, \cdots, x_t|\vartheta)]$

Viterbi Algorithm

- Initialization
  - $\delta_1(i) = \pi_i p_i(x_1)$
- Recursion
  - $\delta_t(j) = \max_i \left[\delta_{t-1}(i) a_{ij}\right] p_j(x_t)$
    - the probability of the most probable state sequence for the first t observations, ending in state j
  - $\psi_t(j) = \text{argmax}\left[\delta_{t-1}(i) a_{ij}\right]$
    - the state chosen at t
- Termination
  - $P^* = \max[\delta_T(i)]$
  - $q^* = \text{argmax}[\delta_T(i)]$

Viterbi Algorithm Example

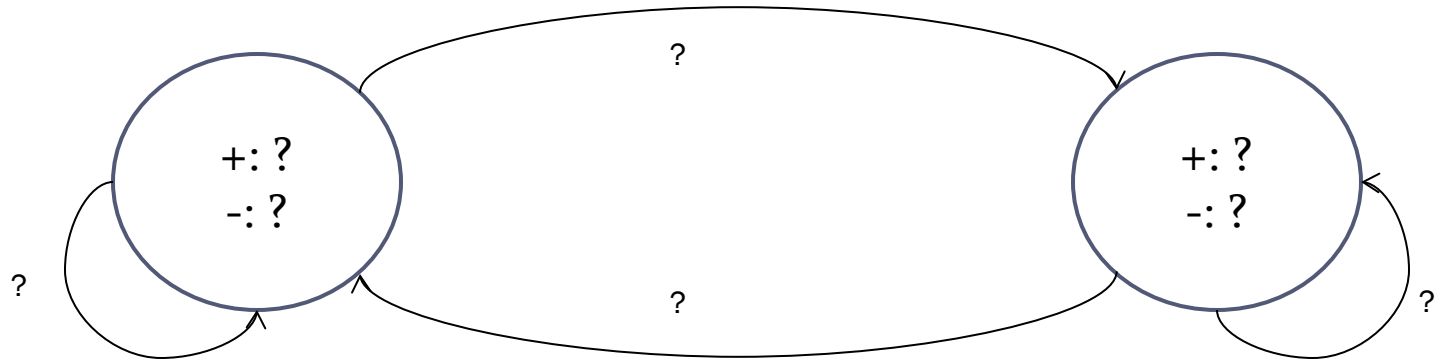


- $+, +, -, +$
- $\delta_1(U) = \pi_U p_U(+) = 0.5 * 0.8 = 0.4$
- $\delta_1(D) = \pi_D p_D(+) = 0.5 * 0.3 = 0.15$
- $\delta_2(U) = \max\{\delta_1(U)a_{UU}, \delta_1(D)a_{DU}\} p_U(+) =$
  $\max\{0.4 * 0.6, 0.15 * 0.5\} * 0.8 = 0.24 * 0.8 = 0.192$
- $\delta_2(D) = \max\{\delta_1(U)a_{UD}, \delta_1(D)a_{DD}\} p_D(+)$

Learning Solutions

Rabiner Model

- Discrete probability masses for observations.

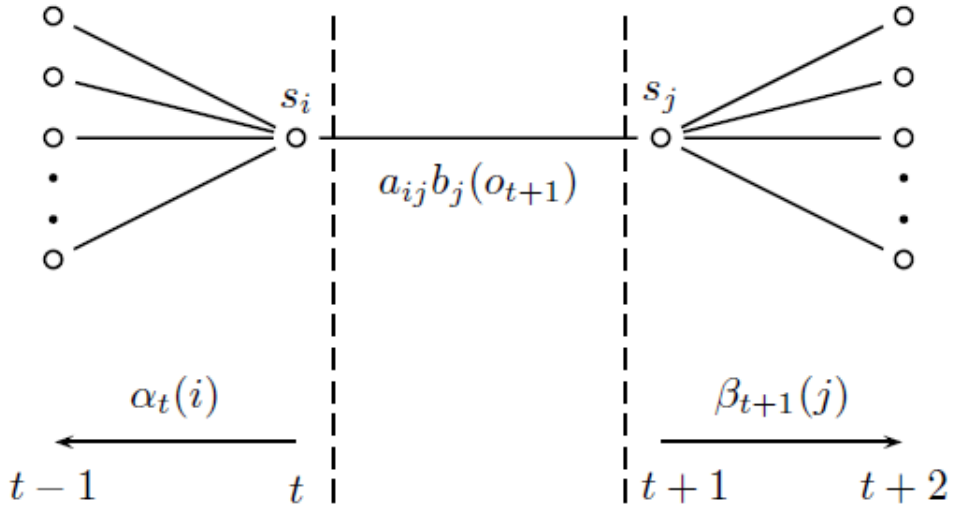


As A Maximization Problem

- Our objective is to find ϑ that maximizes $P(X|\vartheta)$.
- For any given ϑ, we can compute $P(X|\vartheta)$.
- Then solve a maximization problem.
- Algorithm: Nelder-Mead.

Baum-Welch

- the probability of being in state i at time t, and state j at time $(t+1)$, given the model and the observation sequence
- $\xi_t(i,j) = P(q_t = i, q_{t+1} = j | X, \vartheta)$



Xi

- $\xi_t(i,j) = P(q_t = i, q_{t+1} = j | X, \vartheta)$
- $= \frac{P(q_t = i, q_{t+1} = j, X|\vartheta)}{P(X|\lambda)}$
- $= \frac{\alpha_t(i) a_{ij} p_j(x_{t+1}) \beta_{t+1}(j)}{P(X|\vartheta)}$
- $\gamma_t(i) = P(q_t = i \,|X, \vartheta)$
- $= \sum_{j=1}^{N} \xi_t(i,j)$

Estimation Equation

- By summing up over time,
- $\gamma_t(i)$ ~ the number of times state i is visited
- $\xi_t(i,j)$ ~ the number of times the system goes from state i to state j
- Thus, the parameters λ are:
  - $\hat{\pi}_i = \gamma_1(i)$, initial state probabilities
  - $\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$, transition probabilities
  - $\hat{p}_j(v_k) = \frac{\sum_{t=1, x_t=v_k}^{T-1} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(j)}$, conditional probabilities

Conditional Probabilities

- Our formulation so far assumes discrete conditional probabilities.
- The formulations that take other probability density functions are similar.
  - But the computations are more complicated, and the solutions may not even be analytical, e.g., t-distribution.

Heavy Tail Distributions

- t-distribution
- Gaussian Mixture Model
  - a weighted sum of Normal distributions

Trading Ideas

- Compute the next state.
- Compute the expected return.
- Long (short) when expected return > (<) 0.
- Long (short) when expected return > (<) c.
  - c = the transaction costs
- Any other ideas?

Experiment Setup

- EURUSD daily prices from 2003 to 2006.
- 6 unknown factors.
- Λ is estimated on a rolling basis.
- Evaluations:
  - Hypothesis testing
  - Sharpe ratio
  - VaR
  - Max drawdown
  - alpha

Best Discrete Case

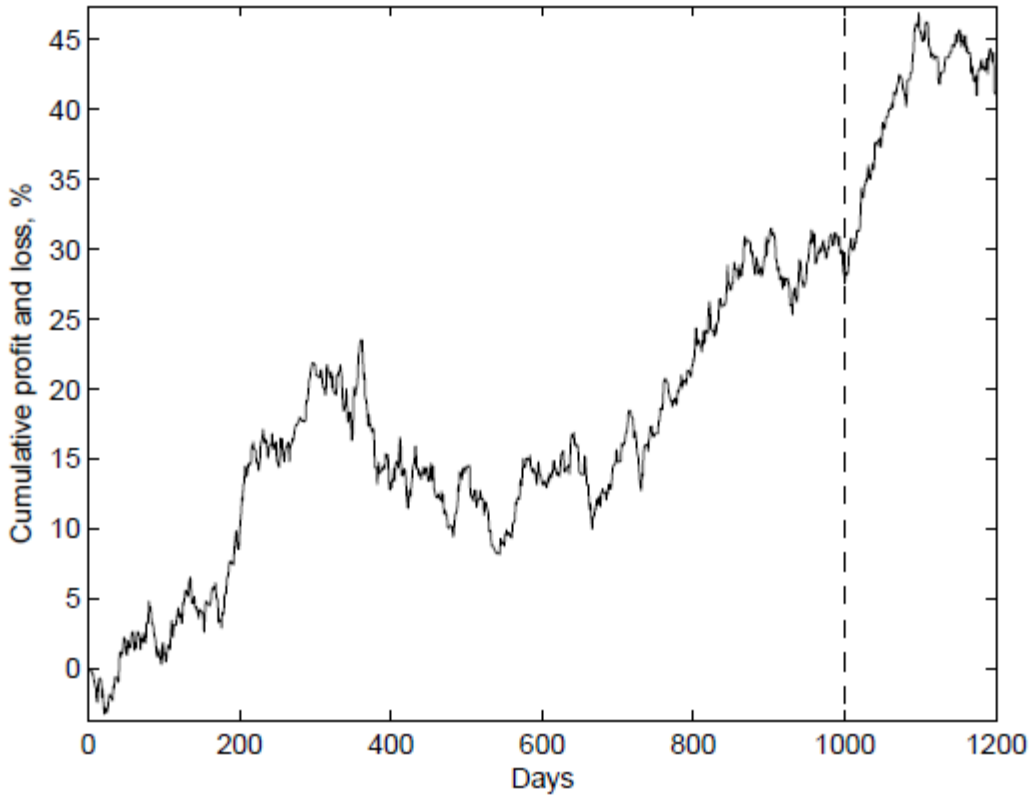


Figure 4.2: Using only the cross as input data with a 30 days window and 3 states.

Best Continuous Case

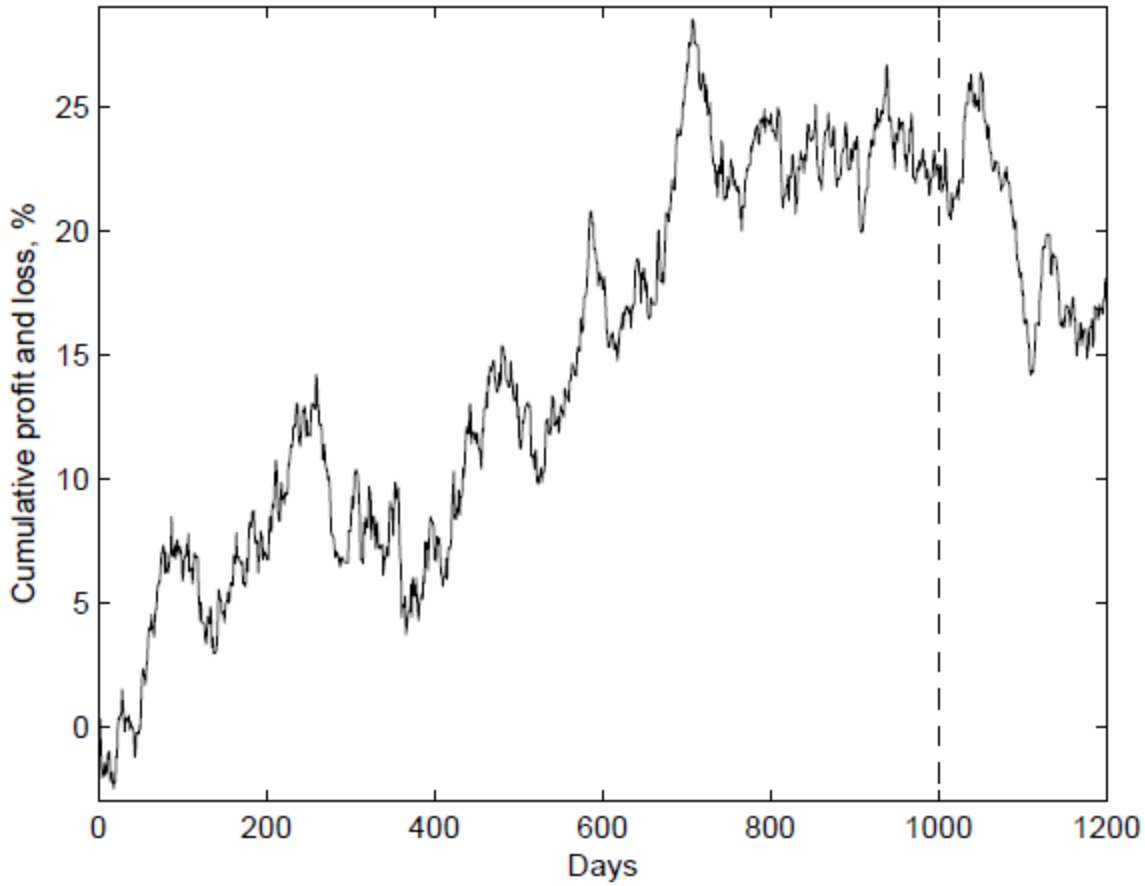


Figure 4.8: Using only the cross as input data with a 20 days window, 4 states and 2 mixture components.

Results

- More data (the 6 factors) do not always help (esp. for the discrete case).
- Parameters unstable.

TODOs

- How can we improve the HMM model(s)? Ideas?

Maximum Likelihood

- One way to estimate parameters for a model.



- Which is the most likely model/dice/number of faces to generate the following observations?
  - 1,2,1,2,1,1,3,4,1,1,2,4,2,4,1,2
  - 1,2,3,4,5,6,4,5,6,3,5,2,4,6,2
  - 1,1,1,1,1,1,1,1,1,1
- Do you think you get the right model?
  - P(1,1,1,1,1,1,1,1,1,1 | 12-faced-dice) = ?

Likelihood Function

- Probability: a function of outcomes given a fixed parameter value.
  - What is the probability of getting 10 Heads flipping a fair coin?
- Likelihood: a function of parameter value given an outcome.
  - What is the likelihood that the coin is fair when it landed Heads 10 times in a roll?

Maximum Likelihood Estimate

- Intuition: we want to find a model (parameter value) such that the probability of observing the outcome is maximized, i.e., most likely.
  - We want to find a ϑ that $p(X|\vartheta)$ is the biggest.
  - $L(\vartheta; X) = p(X|\vartheta)$
- We find ϑ such that $L(\vartheta; X)$ is maximized given the observation.

Example Using the Normal Distribution

- We want to estimate the mean of a sample of size N drawn from a Normal distribution.
  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
- $\vartheta = \{\mu, \sigma\}$
- $L_N(\vartheta; X) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$

Log-Likelihood

- $\log L_N(\vartheta; X) = \sum_{i=1}^{N} \left\{ \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x_i-\mu)^2}{2\sigma^2} \right\}$
- Maximizing the log-likelihood for μ is equivalent to maximizing the following.
  - $-\sum_{i=1}^{N} \{(x_i - \mu)^2\}$
- First order condition w.r.t.,μ
  - $\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- Likewise, for variance, we have
  - $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$

Marginal Likelihood

- For the set of hidden states, $\{Z_t\}$, we write
  - $L(\vartheta; X) = p(X|\vartheta) = \sum_Z p(X, Z|\vartheta)$
- Assume we know the conditional distribution of Z, we could instead maximize the following.
  - $\max_{\vartheta} \mathrm{E}_Z[L(\vartheta|X, Z)]$, or
  - $\max_{\vartheta} \mathrm{E}_Z[\log L(\vartheta|X, Z)]$
- The expectation is a weighted sum of the (log-) likelihoods weighted by the probability of the hidden states.

The Q-Function

- Where do we get the conditional distribution of $\{Z_t\}$ from?
- Suppose we somehow have an (initial) estimation of the parameters, ϑ_0. Then the model has no unknowns. We can compute the distribution of $\{Z_t\}$.
- $Q\left(\vartheta|\vartheta^{(t)}\right) = \underset{Z|X,\vartheta}{\mathrm{E}}[\log L(\vartheta|X,Z)]$

EM Intuition

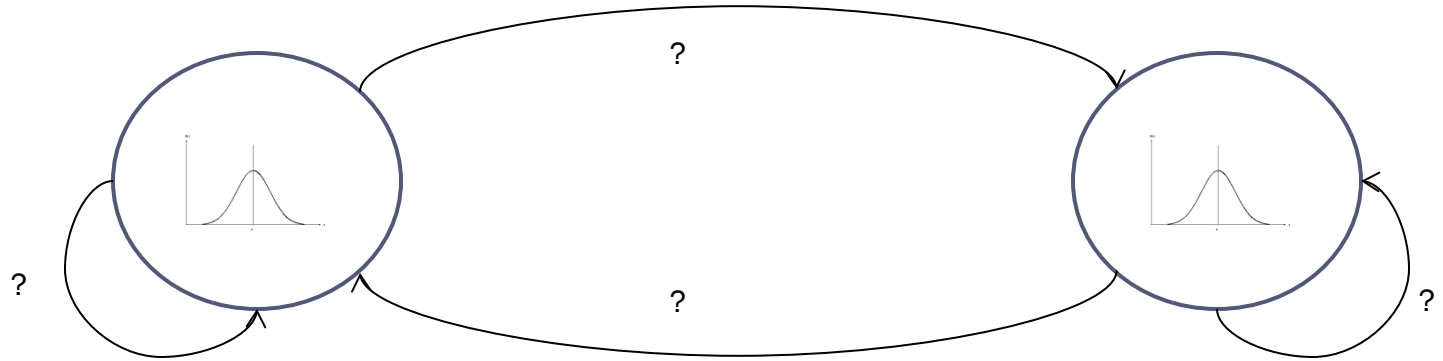
- Suppose we know ϑ, we know completely about the model; we can find Z.
- Suppose we know Z, we can estimate ϑ, by, e.g., maximum likelihood.
- What do we do if we don't know both ϑ and Z?

Expectation-Maximization Algorithm

- Expectation step (E-step): compute the expected value of the log-likelihood function, w.r.t., the conditional distribution of Z under X and $\vartheta^{(t)}$.
  - $Q\left(\vartheta|\vartheta^{(t)}\right) = \underset{Z|X,\vartheta^{(t)}}{\mathrm{E}}[\log L(\vartheta|X,Z)]$
- Maximization step (M-step): find the parameters, ϑ, that maximize the Q-value.
  - $\vartheta^{(t+1)} = \underset{\vartheta}{\operatorname{argmax}}\, Q\left(\vartheta|\vartheta^{(t)}\right)$

Mixture HMM

- Continuous probabilities for observations.



Matrix Notation

- Likelihood: $L_T(\vartheta; X) = \Pi P(x_1)AP(x_2)AP(x_3) \dots AP(x_T)1'$
- Forward probabilities:
  $\alpha_t = \Pi P(x_1)AP(x_2)AP(x_3) \dots AP(x_t) = \Pi P(x_1) \prod_{i=2}^{t} AP(x_i)$
  - Each entry is the joint probability of seeing all the observations up to time t and ending up in state j:
    $\alpha_t(j) = Pr\left(X_1{}^t = x_1{}^t, q_t = j\right)$
  - Induction: $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] p_j(x_{t+1})$
- Backward probabilities:
  $\beta_t{}' = AP(x_{t+1})AP(x_{t+2}) \dots AP(x_T)1'$
  - Each entry is the conditional probability of seeing all the future observations given starting out from state j:
    $\beta_t(j) = Pr\left(X_1{}^t = x_1{}^t | q_t = j\right)$
  - Induction: $\beta_t{}' = AP(x_{t+1})\beta_{t+1}{}'$
- $\lambda_t(i) = P(X, q_t = i|\vartheta) = \alpha_t(i)\beta_t(i)$
  - Likelihood: $\sum_{i=1}^{N} \lambda_t(i) = \alpha_t \beta_t{}' = L_T$

EM for HMM

- $u_j(t) = 1$ if $q_t = j$
- $\nu_{jk} = 1$ if $q_{t-1} = j$ and $q_t = k$
- $\log Pr\left(x_1{}^T, q^T\right) = \log\left(\pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1},q_t} \prod_{t=1}^{T} p_{q_t}(x_t)\right)$
  - $= \log\left(\pi_{q_1}\right) + \sum_{t=2}^{T} \log a_{q_{t-1},q_t} + \sum_{t=1}^{T} p_{q_t}(x_t)$
  - $= \sum_{j=1}^{N} u_j(1) \log\left(\pi_j\right) + \sum_{t=2}^{T} \sum_{j=1}^{N} \sum_{k=1}^{N} \nu_{jk} \log a_{j,k} + \sum_{t=1}^{T} \sum_{j=1}^{T} u_j(t) \log p_j(x_t)$
- Term 1: $\sum_{j=1}^{N} u_j(1) \log\left(\pi_j\right)$
- Term 2: $\sum_{t=2}^{T} \sum_{j=1}^{N} \sum_{k=1}^{N} \nu_{jk} \log a_{j,k}$
- Term 3: $\sum_{t=1}^{T} \sum_{j=1}^{T} u_j(t) \log p_j(x_t)$

E Step

- Given the current $\vartheta = \{\pi, A, \lambda\}$, we can estimate $u_j(t)$ and ν_{jk} from the forward and backward probabilities.
- $\hat{u}_j(t) = Pr\left(q_t = j \mid {x_1}^T\right) = \alpha_t(j)\beta_t(j)/L_T$
- $\hat{\nu}_{jk}(t) = Pr\left(q_{t-1} = j, q_t = k \mid {x_1}^T\right) = \alpha_{t-1}(j)a_{jk}p_k(x_t)\beta_t(k)/L_T$

M Step

- Term 1: max $\sum_{j=1}^{N} \hat{u}_j(1) \log(\pi_j)$ w.r.t each π_j
  - $\hat{\pi}_j = \hat{u}_j(1) / \sum_{j=1}^{N} \hat{u}_j(1) = \hat{u}_j(1)$
- Term 2: max $\sum_{t=2}^{T} \sum_{j=1}^{N} \sum_{k=1}^{N} \hat{v}_{jk} \log a_{j,k}$ w.r.t. each $\hat{v}_{jk}$
  - $\hat{v}_{jk} = f_{jk} / \sum_{j=1}^{N} f_{jk}$, where $f_{jk} = \sum_{t=2}^{T} \hat{v}_{jk}(t)$
- Term 3: max $\sum_{t=1}^{T} \sum_{j=1}^{T} \hat{u}_j(t) \log p_j(x_t)$ w.r.t. the parameters λ for each conditional probability distribution in each state $p_j(x)$.

Poisson-HMM

- $p_j(x) = e^{-\lambda_j} {\lambda_j}^x / x!$
- $\max \sum_{t=1}^{T} \sum_{j=1}^{T} \hat{u}_j(t) \log p_j(x)$
- Each of the j term can be individually maximized.
  - $\sum_{t=1}^{T} \hat{u}_j(t) \log p_j(x)$
  - $\sum_{t=1}^{T} \hat{u}_j(t) \left[-\lambda_j + x \log \lambda_j - x!\right]$
- $0 = \sum_{t=1}^{T} \hat{u}_j(t) \left[-1 + x/\lambda_j\right]$
- $\hat{\lambda}_j = \sum_{t=1}^{T} \hat{u}_j(t) x_t \,/\, \sum_{t=1}^{T} \hat{u}_j(t)$

Normal-HMM

- $\hat{\mu}_j = \sum_{t=1}^{T} \hat{u}_j(t) x_t \,/\, \sum_{t=1}^{T} \hat{u}_j(t)$
- $\hat{\sigma}_j{}^2 = \sum_{t=1}^{T} \hat{u}_j(t) \left(x_t - \hat{u}_j\right)^2 \,/\, \sum_{t=1}^{T} \hat{u}_j(t)$