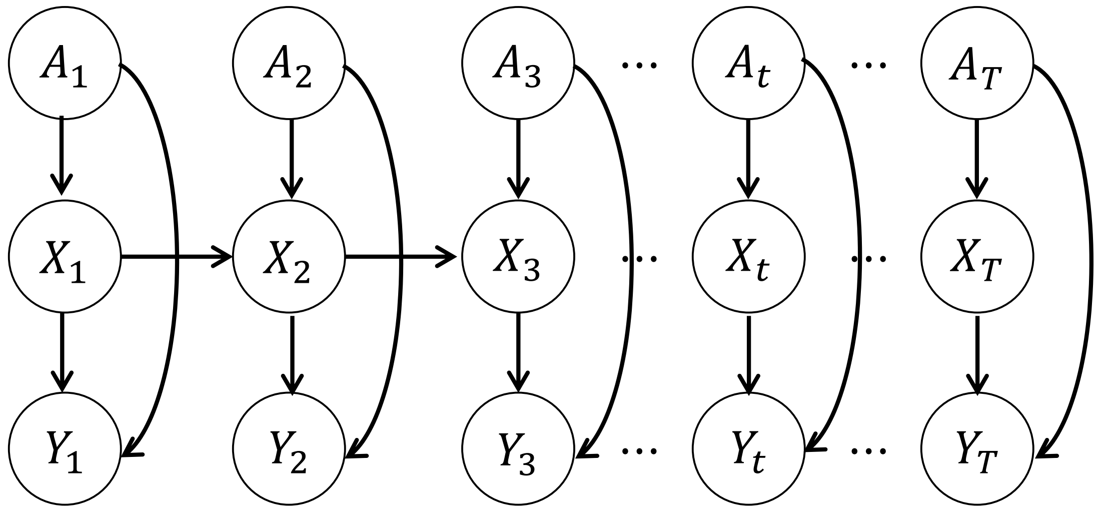
L8. Dynamic Bayesian Network

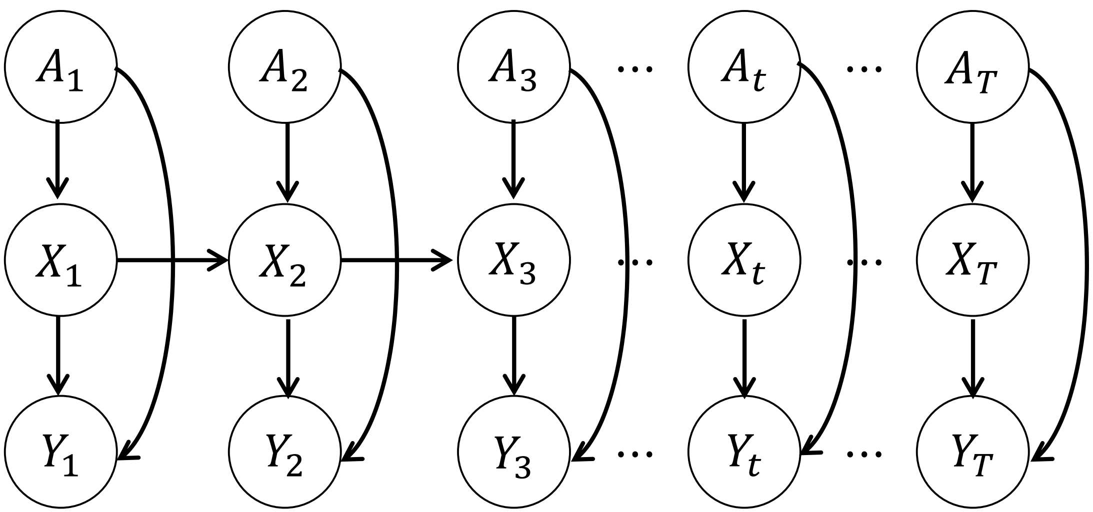
Dynamic Bayesian Network relates variables to each other over adjacent time steps.



Dynamic Bayesian Networks (temporal model)

- We are interested in reasoning about the state of the world as it evolves over time
- System state S_t is a snapshot of the relevant attributes of the system at time t
- Trajectory of states $S_1, \ldots, S_t$ represents the evolution of the target system
- $P(S_1, \ldots, S_t)$ is very complex probability space

→ we need a series of simplifying assumptions

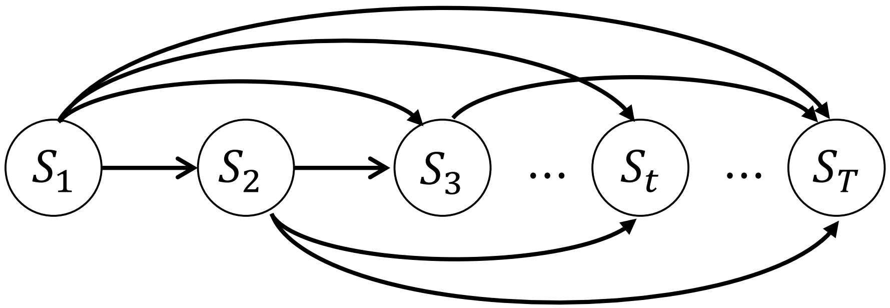


Discrete-State Markov models

Consider a distribution over trajectories sampled over a prefix of time $t = 1, \ldots, T$

$$P(S_1, S_2, \ldots, S_T) = P(S_1) \prod_{t=2}^{T} P(S_t | S_{1:t-1})$$

Cascade decomposition



Discrete-State Markov models

Consider a distribution over trajectories sampled over a prefix of time $t = 1, \dots, T$

$$P(S_1, S_2, \dots, S_T\ ) = P(S_1) \prod_{t=2}^{T} P(S_t | S_{1:t-1})$$

Cascade decomposition

1. Markov Chain:

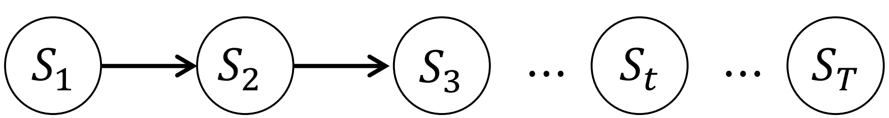
A Markov chain is defined on either discrete or continuous variables $S_{1:t}$ is one in which the following conditional independence assumption holds:

$$P(S_t | S_1, \dots, S_{t-1}) = P(S_t | S_{t-L}, \dots, S_{t-1})$$

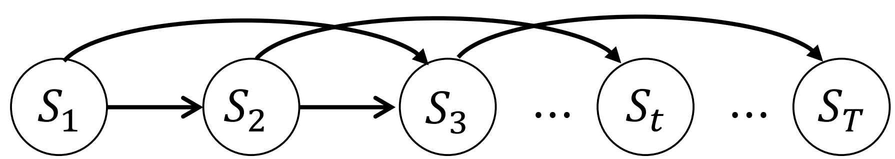
where $L \geq$ is the order of the Markov chain. First order Markov chain can be represented

$$P(S_1, S_2, \dots, S_T) = P(S_1) \prod_{t=2}^{T} P(S_t | S_{t-1})$$

the future is conditionally independent of the pas given the present: $\left(S^{t+1} \perp S^{(1:t-1)} | S^t\right)$



First Order Markov Chain



Second Order Markov Chain

Discrete-State Markov models

Consider a distribution over trajectories sampled over a prefix of time $t = 1, \ldots, T$

$$P(S_1, S_2, \ldots, S_T\ ) = P(S_1) \prod_{t=2}^{T} P(S_t | S_{1:t-1})$$

Cascade decomposition

1. Markov Chain:

A Markov chain is defined on either discrete or continuous variables $S_{1:t}$ is one in which the following conditional independence assumption holds:

$$P(S_t | S_1, \ldots, S_{t-1}) = P(S_t | S_{t-L}, \ldots, S_{t-1})$$

where $L \geq$ is the order of the Markov chain. First order Markov chain can be represented

$$P(S_1, S_2, \ldots, S_T) = P(S_0) \prod_{t=1}^{T} P(S_t | S_{t-1})$$

2. Stationary assumption

The state transition probability $P(S^{t+1} | S^t)$ is the same for all t

$$P(S^{t+1} = s' | S^t = s) = P(S' = s' | S = s) \text{ for any } t$$

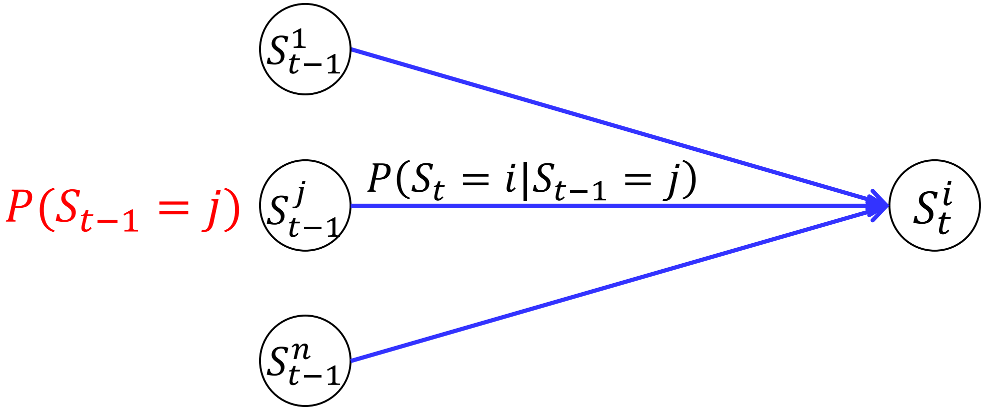
→The number of parameters are reduced substantially

Equilibrium and stationary distribution of a Markov chain

- The marginal $P(S_t)$ evolves through time. For discrete time,

$$P(S_t = i) = \sum_j P(S_t = i|S_{t-1} = j)P(S_{t-1} = j)$$

✓ $P(S_t = i)$: the frequency that we visit state i at time t, given we started with a sample from $P(S_1)$ and subsequently repeatedly drew samples from the transition $P(S_\tau|S_{\tau-1})$



- Denoting $(\mathbf{p}_t)_i = P(S_t = i)$,

$$\mathbf{p}_t = \mathbf{M}\mathbf{p}_{t-1} = \mathbf{M}^{t-1}\mathbf{p}_1$$

- If, for $t \to \infty$, $\mathbf{p}_t$ is independent of the initial distribution $\mathbf{p}_1$, then $\mathbf{p}_\infty$ is called the equilibrium distribution of the chain, that is

$$\mathbf{p}_\infty = \mathbf{M}\mathbf{p}_\infty$$

Fitting Markov Models

$$S_1 \rightarrow S_2 \rightarrow S_3 \quad \cdots \quad S_t \quad \cdots \quad S_T$$

Given a sequence $(S_1 = s_1, S_2 = s_2, \ldots, S_T = s_T)$, how to construct the transition matrix?

$$\theta_{i|j} = P(S_\tau = i | S_{\tau-1} = j) \propto \sum_{t=2}^{T} \mathbb{I}[S_\tau = i, S_{\tau-1} = j]$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_{1|1} \theta_{1|2} \theta_{1|3} \theta_{1|4} \theta_{1|5} \\ \theta_{2|1} \theta_{2|2} \theta_{2|3} \theta_{2|4} \theta_{2|5} \\ \theta_{3|1} \theta_{3|2} \theta_{3|3} \theta_{3|4} \theta_{3|5} \\ \theta_{4|1} \theta_{4|2} \theta_{4|3} \theta_{4|4} \theta_{4|5} \\ \theta_{5|1} \theta_{5|2} \theta_{5|3} \theta_{5|4} \theta_{5|5} \end{bmatrix}$$

State transition matrix

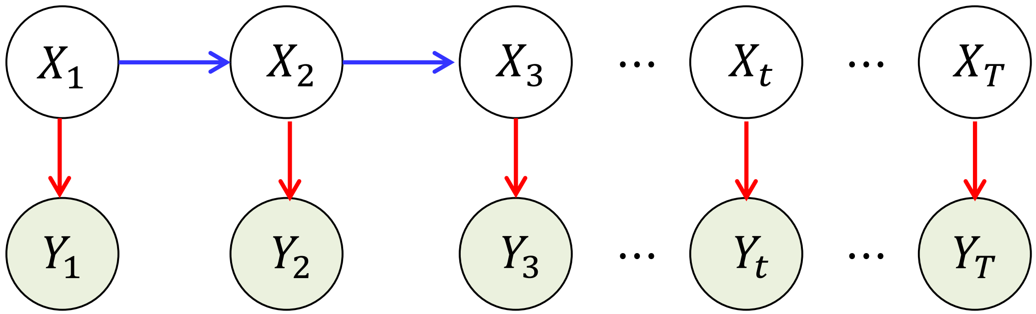
$$\sum_i \theta_{i|j} = 1$$

If we have $s_{1:t} = 1, 3, 2, 4, 1, 4, 3, 5, 1, 3, 4, 2, 1, 4, 4, 2, 4, 5, 1, 3, 3, 4, \ldots$

$$\theta_{3|1} = \frac{3}{5}$$

Definition of Hidden Markov Models

- The Hidden Markov Model (HMM) defines a Markov chain on hidden variables $X_{1:t}$
- The observed variables are dependent on the hidden variables through an emission $P(Y_t|X_t)$



- The joint distribution on the hidden variables and observations are

$$P(X_{1:t}, Y_{1:t}) = P(X_1)P(Y_1|X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})P(Y_t|X_t)$$

- Transition distribution: For a stationary HMM the transition distribution $P(X_t|X_{t-1})$ is defined as the $H \times H$ matrix

$$M_{i,j} = P(X_t = i|X_{t-1} = j)$$

- Emission distribution: For a stationary HMM and emission distribution with discrete states $Y_t \in \{1, \dots, V\}$, we define $V \times H$ matrix

$$O_{i,j} = P(Y_t = i|X_t = j)$$

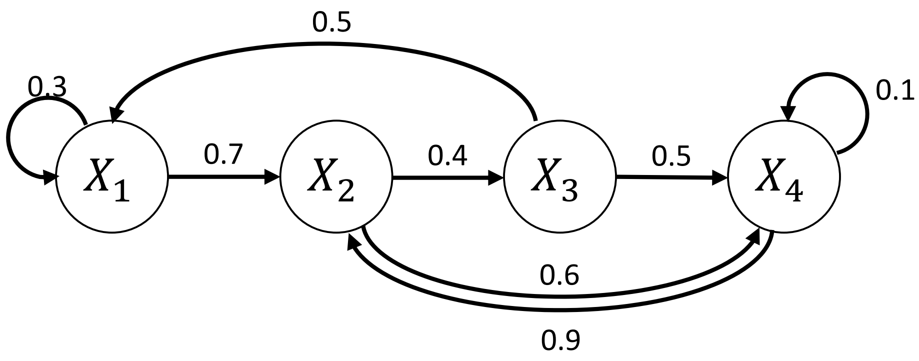
Hidden Markov Model

The state variable X_t is discrete

- The state transition model $P(X'|X)$ is usually sparse,

  → can be represented as a directed graph

$X = (X_1, X_2, X_3, X_4)$: 4 discretized states

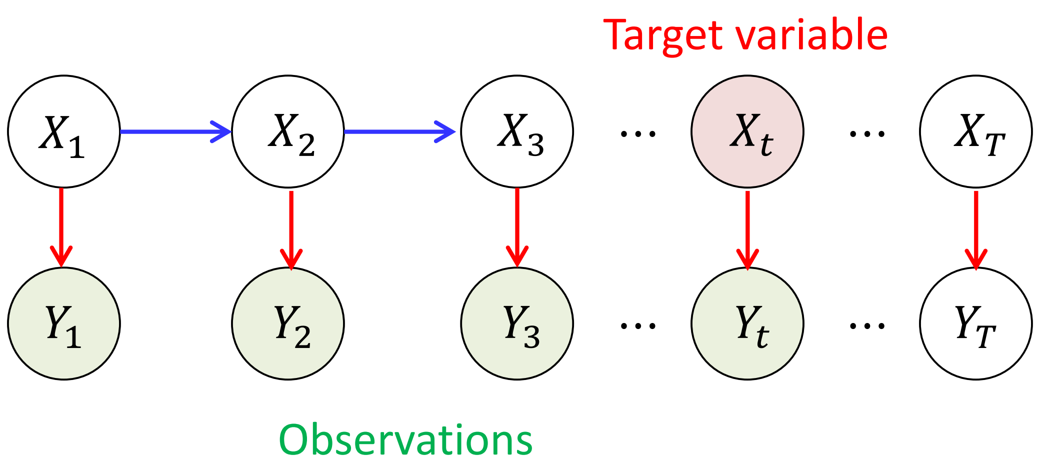


$\sum_i P(X_i'|X) = 1$

- The observation model : $P(Y|X)$ can be deterministic or random

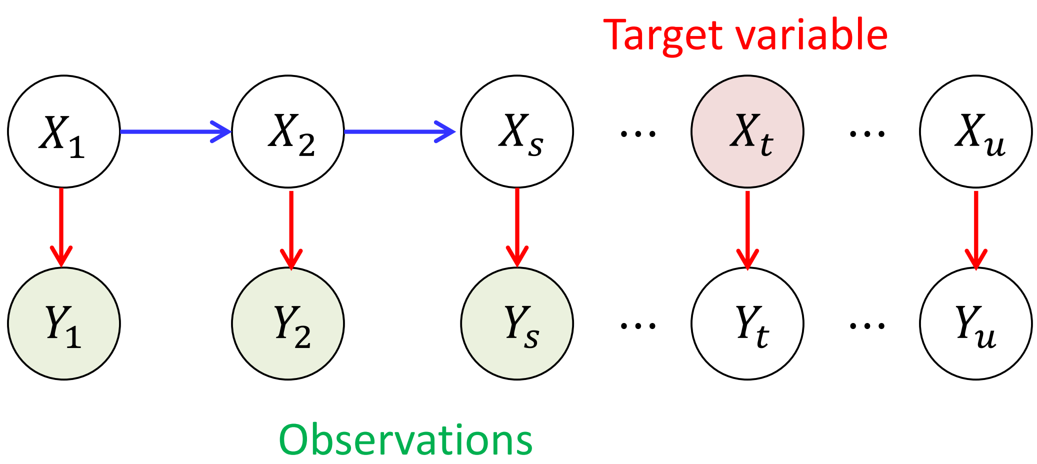
Inferencing of Hidden Markov Models

- Filtering (inferencing the present) $P(x_t|y_{1:t})$
- Prediction (inferencing the future) $P(x_t|y_{1:s})$ $t > s$
- Smoothing (inferencing the past) $P(x_t|y_{1:u})$ $t < u$
- Likelihood (inferencing the past) $P(x_{1:t})$
- Most likely hidden path (Viterbi alignment) $\underset{x_{1:t}}{\operatorname{argmax}} P(x_{1:t}|y_{1:t})$



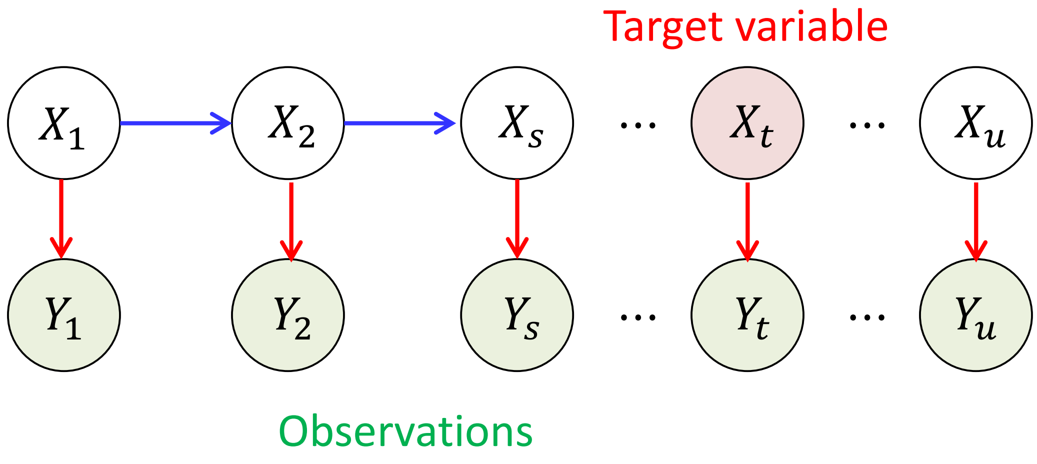
Inferencing of Hidden Markov Models

- Filtering (inferencing the present) $P(x_t|y_{1:t})$
- Prediction (inferencing the future) $P(x_t|y_{1:s})$ $t > s$
- Smoothing (inferencing the past) $P(x_t|y_{1:u})$ $t < u$
- Likelihood (inferencing the past) $P(x_{1:t})$
- Most likely hidden path (Viterbi alignment) $\underset{x_{1:t}}{\operatorname{argmax}} P(x_{1:t}|y_{1:t})$



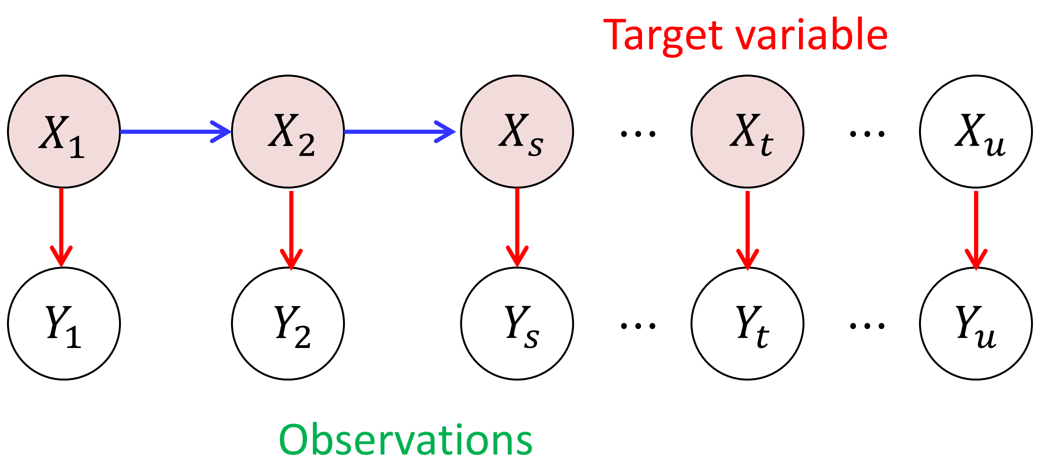
Inferencing of Hidden Markov Models

- Filtering (inferencing the present) $P(x_t|y_{1:t})$
- Prediction (inferencing the future) $P(x_t|y_{1:s})$ $t > s$
- Smoothing (inferencing the past) $P(x_t|y_{1:u})$ $t < u$
- Likelihood (inferencing the past) $P(x_{1:t})$
- Most likely hidden path (Viterbi alignment) $\underset{x_{1:t}}{\operatorname{argmax}} P(x_{1:t}|y_{1:t})$



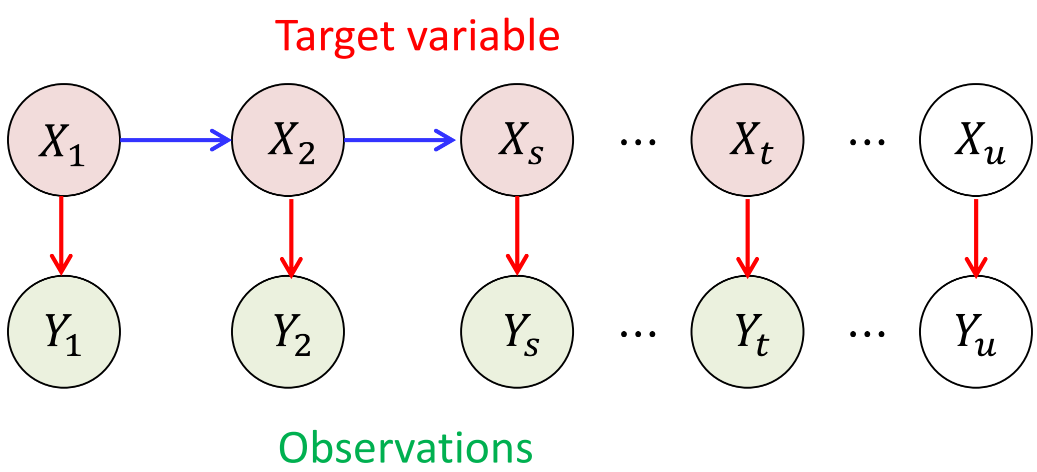
Inferencing of Hidden Markov Models

- Filtering (inferencing the present) $P(x_t|y_{1:t})$
- Prediction (inferencing the future) $P(x_t|y_{1:s})$ $t > s$
- Smoothing (inferencing the past) $P(x_t|y_{1:u})$ $t < u$
- Likelihood (inferencing the past) $P(x_{1:t})$
- Most likely hidden path (Viterbi alignment) $\underset{x_{1:t}}{\operatorname{argmax}} P(x_{1:t}|y_{1:t})$



Inferencing of Hidden Markov Models

- Filtering (inferencing the present) $P(x_t|y_{1:t})$
- Prediction (inferencing the future) $P(x_t|y_{1:s})$ $t > s$
- Smoothing (inferencing the past) $P(x_t|y_{1:u})$ $t < u$
- Likelihood (inferencing the past) $P(x_{1:t})$
- Most likely hidden path (Viterbi alignment) $\underset{x_{1:t}}{\operatorname{argmax}} P(x_{1:t}|y_{1:t})$



Inferencing of Hidden Markov Models

- Filtering (inferencing the present) $P(x_t|y_{1:t})$

$$P(x_t|y_{1:t}) = \frac{P(x_t, y_{1:t})}{P(y_{1:t})} \propto P(x_t, y_{1:t})$$

$$P(x_t, y_{1:t}) = \sum_{x_{t-1}} P\left(x_t, x_{t-1}, y_{1:t-1}, y_t\right)$$

$$\sum_{x_{t-1}} P(y_t|\cancel{y_{1:t-1}}, x_t, \cancel{x_{t-1}}) P(x_t|\, \cancel{y_{1:t-1}}, x_{t-1}) P(x_{t-1}, y_{1:t-1})$$

$$\sum_{x_{t-1}} P(y_t|x_t) P(x_t|\, x_{t-1}) P(x_{t-1}, y_{1:t-1})$$

∵Conditional independence

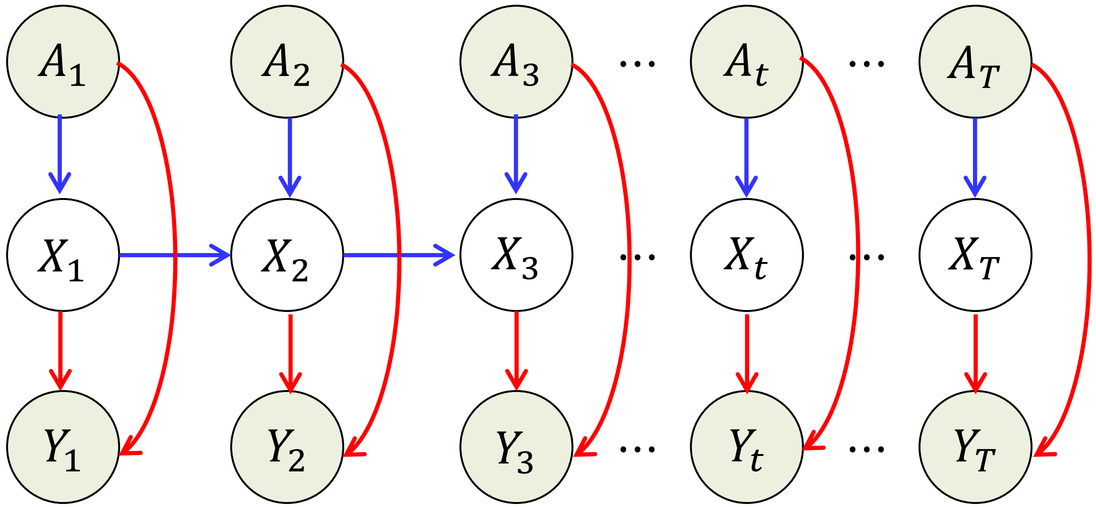
$$\underbrace{P(y_t|x_t)}_{\text{corrector}} \underbrace{\sum_{x_{t-1}} P(x_t|\, x_{t-1}) P(x_{t-1}, y_{1:t-1})}_{\text{predictor}}$$

Bayesian view

$$P(x_t|y_{1:t}) \propto \sum_{x_{t-1}} \underbrace{P(y_t|x_t) P(x_t|\, x_{t-1})}_{\text{Regarded as likelihood}} \underbrace{P(x_{t-1}|y_{1:t-1})}$$

Modified prior distribution has an effect of removing all nodes in the graph before time $t-1$

Input and output hidden Markov Model (IOHMM)



The state transition model : $P(X_t|X_{t-1}, A_t)$

The observation model : $P(Y_t|X_t, A_t)$

Continuous domain

Continuous-state Markov models

- In many practical time series applications, the data is naturally continuous (i.e., variables are not discretized), particularly for models of the physical environment

- Restrict the form of the continuous transition $p(X_t|X_{t-1})$

- A simple yet powerful class of such transitions are the linear dynamical systems

- A *deterministic linear dynamical system* defines the temporal evolution of a vector x_t according to the discrete-time update equation

$$x_t = A_t x_{t-1}$$

  where A_t is the transition matrix at time t

- If A_t is invariant with t, the process is called stationary or time-invariant

Observed linear dynamic system

- A *stochastic linear dynamical system* defines the temporal evolution of a vector x_t according to the discrete-time update equation

$$x_t = A_t x_{t-1} + \eta_t$$

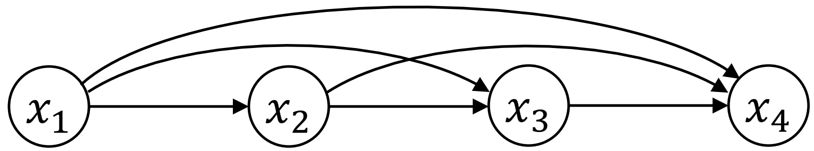
where η_t is a noise vector sampled from a Gaussian distribution

$$\eta_t \sim N(\mu_t, \Sigma_t)$$

- This is equivalent to a first-order Markov model with transition

$$p(x_t | x_{t-1}) = N(x_t | A_t x_{t-1} + \eta_t, \Sigma_t)$$

Auto-regressive models



- A scalar time-invariant Auto-Regressive (AR) model is defined by

$$x_t = \sum_{l=1}^{L} a_l x_{t-l} + \eta_t\,, \qquad \eta_t \sim N(\mu, \sigma^2)$$

where $a = (a_1, a_2, \ldots, a_L)^T$ are AR coefficients and σ^2 is innovation noise.

- As a belief network, the AR model can be written as an Lth-order Markov model:

$$p(x_{1:T}) = \prod_{t=1}^{T} p(x_t | x_{t-1}, \ldots, x_{t-L}), \qquad \text{with } x_i = 0 \text{ for } i \leq 0$$

$\hat{x}_{t-1} = (x_{t-1}, \ldots, x_{t-L})$

with $p(x_t | x_{t-1}, \ldots, x_{t-L)} = N\left(x_t \middle| \sum_{l=1}^{L} a_l x_{t-l} + \eta_t\,, \sigma^2\right) = N(x_t | a^T \hat{x}_{t-1}, \sigma^2)$

Similar to Bayesian Regression

- Heavily used in financial time series prediction, being able to capture simple trends in the data
- The AR coefficients form a compressed representation of the signal

Training Auto-regressive model

- Maximum likelihood training of the AR coefficients is straightforward based on

$$
\begin{aligned}
\log p(x_{1:T}) &= \log \prod_{t=1}^{T} p(x_t | x_{t-1}, \ldots, x_{t-L}) \\
&= \sum_{t=1}^{T} \log(x_t | \hat{x}_{t-1}) \\
&= -\frac{1}{2\sigma^2} \sum_{t=1}^{T} (x_t - a^T \hat{x}_{t-1})^2 - \frac{T}{2} \log(2\pi\sigma^2)
\end{aligned}
$$

- Differentiating w.r.t. a and equating to zero we arrive at

$$\sum_{t=1}^{T} (x_t - a^T \hat{x}_{t-1}) \hat{x}_{t-1} = 0$$

$$\rightarrow a = \left[\sum_t \hat{x}_{t-1} \hat{x}_{t-1}^T\right]^{-1} \sum_t x_t \hat{x}_{t-1}$$

x_t : target output (scalar)

- Similarly,

$$\sigma^2 = \frac{1}{T} \sum_{t=1}^{T} (x_t - a^T \hat{x}_{t-1})^2$$

Time-varying Auto-regressive model

- Learning the AR coefficients as a problem in inference in a latent linear dynamical system (LDS):

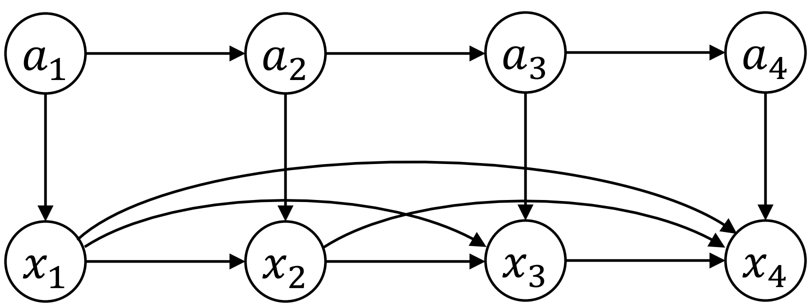
$$x_t = \hat{x}_{t-1}^T a_t + \eta_t, \qquad \eta_t \sim N(0, \sigma^2)$$

which can be viewed as the emission distribution of a latent LDS in which the hidden variable is a_t and the time dependent emission matrix is given by $\hat{x}_{t-1}^T$

- By placing a simple latent transition

$$a_t = a_{t-1} + \eta_t^a, \qquad \eta_t^a \sim N(0, \sigma_a^2 \mathbf{I})$$

which encourages the AR coefficients to change slowly with time



Time-varying Auto-regressive model

- Learning the AR coefficients as a problem in inference in a latent linear dynamical system (LDS):

$$x_t = \hat{x}_{t-1}^T a_t + \eta_t, \qquad \eta_t \sim N(0, \sigma^2)$$

which can be viewed as the emission distribution of a latent LDS in which the hidden variable is a_t and the time dependent emission matrix is given by $\hat{x}_{t-1}^T$

- By placing a simple latent transition

$$a_t = a_{t-1} + \eta_t^a, \qquad \eta_t^a \sim N(0, \sigma_a^2 \mathbf{I})$$

which encourages the AR coefficients to change slowly with time

- The joint distribution between the observation $x_{1:T}$ and the coefficients $\boldsymbol{a}_{1:t}$

$$p(a_{1:T} | x_{1:T}) \propto p(x_{1:T}, a_{1:T}) = \prod_{t=2}^{T} p(x_t | a_t, \hat{x}_{t-1}) p(a_t | a_{t-1})$$

then we can compute

$$a_{1:T}^* = \mathrm{argmax}_{a_{1:T}}\, p(a_{1:T} | x_{1:T})$$

from which the MAP estimates for the AR coefficients can be determined

Time-varying variance Auto-regressive model

- For some applications, particularly in finance, the variance can change with time due to volatility

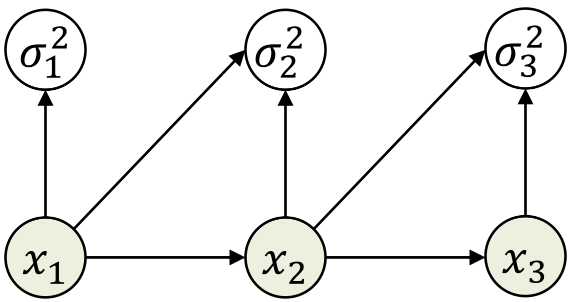
$$x_t = \sum_{l=1}^{L} a_l x_{t-l} + \eta_t, \qquad \eta_t \sim N(\mu, \underline{\sigma_t^2})$$

Time varying variance

$$\bar{x}_t = \sum_{l=1}^{L} a_l x_{t-l}$$

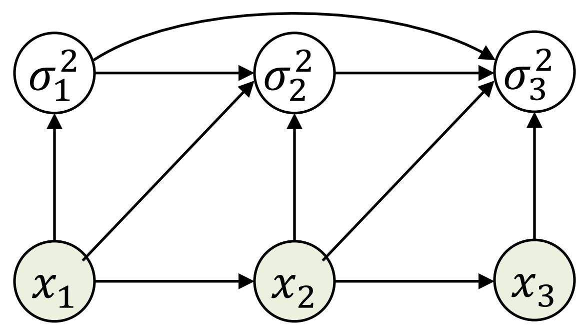
- The estimated time varying variance of noise can be computed

Auto Regressive Conditional Heteroscedasticity (ARCH)



$$\sigma_t^2 = \sigma_0 + \sum_{i=1}^{Q} \alpha_i (x_{t-i} - \bar{x}_{t-i})^2$$

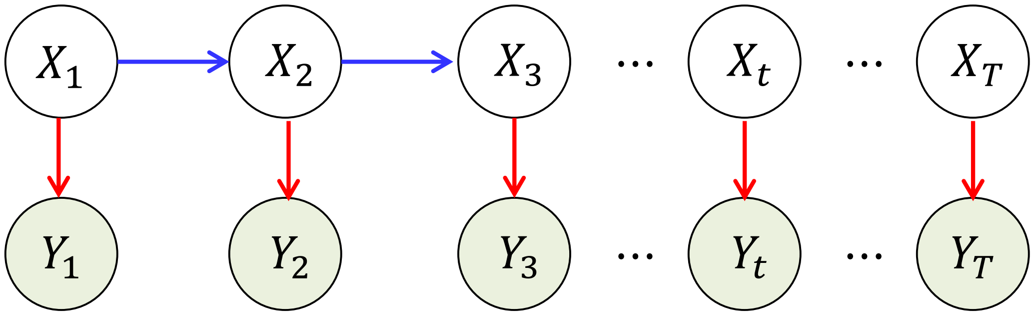
Generalized ARCH model (GARCH)



$$\sigma_t^2 = \sigma_0 + \sum_{i=1}^{Q} \alpha_i (x_{t-i} - \bar{x}_{t-i})^2 + \sum_{i=1}^{Q} \beta_i \sigma_{t-i}^2$$

Linear Gaussian State Space Model

- The latent LDS defines a stochastic linear dynamical system in a latent space on a sequence of states $x_{1:T}$
- Observations $y_{1:T}$ are used to infer the hidden states that tracks or explains the system evolution



Transition model : $x_t = A_t x_{t-1} + \eta_t^x,$ $\quad \eta_t^x \sim N(\bar{x}_t, \Sigma_t^x)$

Emission model : $y_t = B_t x_t + \eta_t^y,$ $\quad \eta_t^y \sim N(\bar{y}_t, \Sigma_t^y)$

A_t: transition matrix

B_t: emission matrix

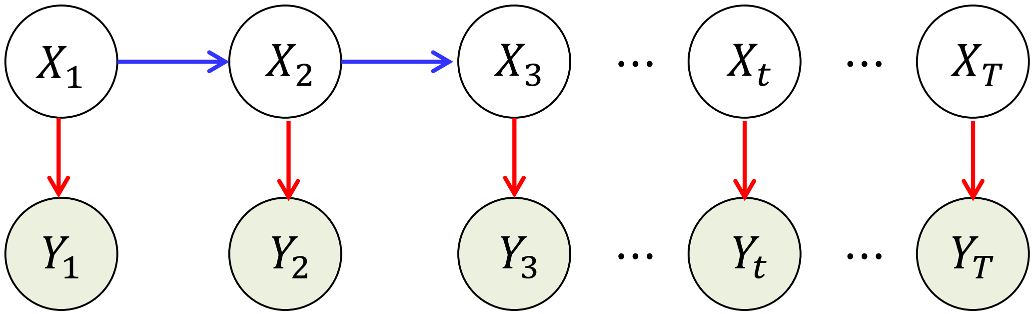
η_t^x transition noise vector with a hidden bias $\bar{x}_t$

η_t^y emission noise vector with a hidden bias $\bar{y}_t$

Linear Gaussian State Space Model

- The latent LDS defines a stochastic linear dynamical system in a latent space on a sequence of states $x_{1:T}$

- Observations $y_{1:T}$ are used to infer the hidden states that tracks or explains the system evolution



Transition model : $p(x_t|x_{t-1}) = N(x_t|A_t x_{t-1} + \bar{x}_t, \Sigma_t^x), \quad p(x_1) = N(x_1|\mu_\pi, \Sigma_\pi)$

Emission model : $p(y_t|x_t) = N\left(y_t \middle| B_t x_t + \bar{y}_t, \Sigma_t^y\right)$

- The first order Markov model is then defined as

$$p(x_{1:T}, y_{1:T}) = p(x_1)p(y_1|x_1)\prod_{t=2}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

Kalman Filter

- Recall the filtering recursion for HMM:

$$P(x_t|y_{1:t}) \propto \sum_{x_{t-1}} P(y_t|x_t)P(x_t|\, x_{t-1})P(x_{t-1}|y_{1:t-1})$$

- For linear Gaussian State-space model, the recursion becomes

$$P(x_t|y_{1:t}) \ \propto \int_{x_{t-1}} P(y_t|x_t)P(x_t|x_{t-1})P(x_{t-1}|y_{1:t-1}) \ \ \text{for} \ \ t > 1$$

- Since the product of two Gaussians is another Gaussian, and the integral of a Gaussian is another Gaussian, $P(x_t|y_{1:t})$ is Gaussian:

$$P(x_t|y_{1:t}) = N(x_t|f_t, F_t)$$

- Thus the recursion is for computing the mean f_t and the variance F_t for $P(x_t|y_{1:t})$ using f_{t-1} and the variance F_{t-1} for $P(x_{t-1}|y_{1:t-1})$

$$\begin{matrix} f_{t-1} \\ F_{t-1} \end{matrix} \xrightarrow{\text{Update}} \begin{matrix} f_t \\ F_t \end{matrix}$$

$P(x_{t-1}|y_{1:t-1}) = N(x_{t-1}|f_{t-1}, F_{t-1})$

$P(x_t|y_{1:t}) = N(x_t|f_t, F_t)$

Supplements