Steven L. Scott

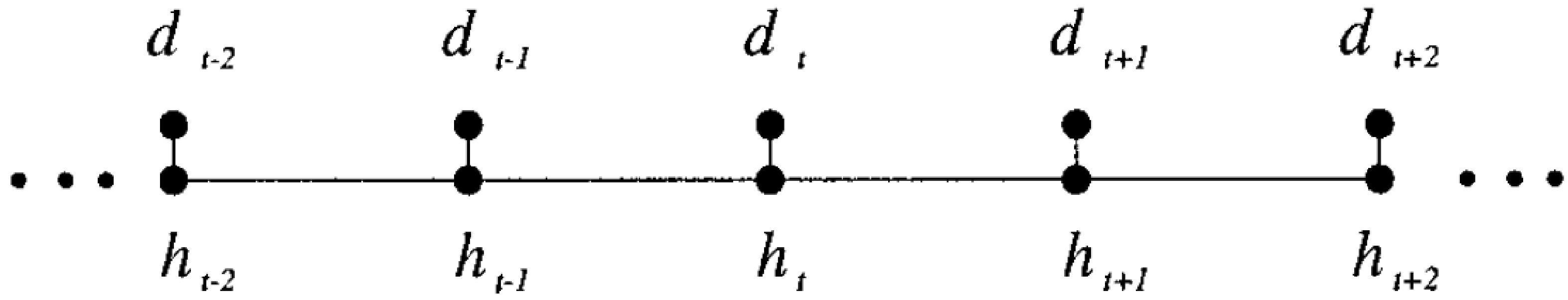
Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century

Presented by Ahmet Engin Ural

Outline

- Overview of HMM
  - Evaluating likelihoods
    - The Likelihood Recursion
    - The Forward-Backward Recursion
- Sampling HMM
  - DG and FB samplers
  - Autocovariance of samplers
  - Some issues with samplers (in general)
- Estimation
  - Marginal
  - MAP
  - Size of the state space

Hidden Markov Models



h	$(h_1 \ldots h_n)$
Q	$Q_{ij} = q(h_i, h_j)$ (stationary)
$\boldsymbol{\pi}_0$	Initial state
θ	$P_0 \ldots. P_{s-1}$

$$p(d_t \mid d_{-t}, \mathbf{h}, \theta, \mathbf{Q}, \pi_0) = P_{h_t}(d_t \mid \theta), \qquad (2)$$

Calculating the likelihood

$$p(d_1^n \mid \theta) = \sum_{\mathbf{h} \in \mathcal{S}^n} \pi_0(h_1) P_{h_1}(d_1 \mid \theta) \prod_{t=2}^{n} q(h_{t-1}, h_t) P_{h_t}(d_t \mid \theta). \tag{3}$$

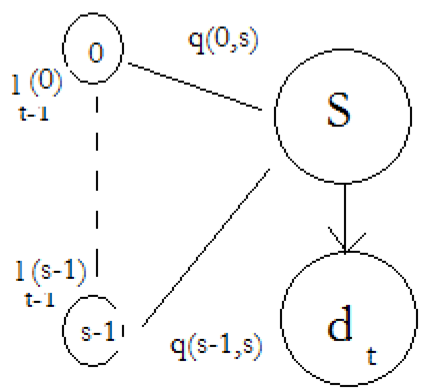
Sum over all possible hidden state sequences, the probability of the observed generated by that hidden state sequence

Calculating the likelihood

$$p(d_1^n \mid \theta) = \sum_{\mathbf{h} \in \mathcal{S}^n} \pi_0(h_1) P_{h_1}(d_1 \mid \theta) \prod_{t=2}^{n} q(h_{t-1}, h_t) P_{h_t}(d_t \mid \theta). \quad (3)$$

Instead, likelihood recursion O(S^2 n) steps

Forward variable: $$\ell_t(s) = P_s(d_t \mid \theta) \sum_{r=0}^{S-1} q(r,s) \ell_{t-1}(r).$$



Forward Backward Recursions

- Forward recursion is as likelihood recursion
- Backward variable: $\pi_t(s \mid \theta) = \ell_t(s)/\ell_t^*$

  where $\ell_t^* = \sum_{s=0}^{S-1} \ell_t(s)$
- Transition probabilities, *p(r - > s at time t | we observed until t)*

$$p_{trs} \propto p(h_{t-1} = r, h_t = s, d_t \mid d_1^{t-1}, \theta)$$

$$= \pi_{t-1}(r \mid \theta) q(r, s) P_s(d_t \mid \theta),$$

- Backward recursion

$$p'_{trs} = p(h_{t-1} = r \mid h_t = s, d_1^n, \theta) p(h_t = s \mid d_1^n, \theta)$$

Forward Backward Recursions

- Forward recursion is as likelihood recursion
- Backward variable: $\pi_t(s \mid \theta) = \ell_t(s)/\ell_t^*$

  where $\ell_t^* = \sum_{s=0}^{S-1} \ell_t(s)$
- Transition probabilities, *p(r - > s at time t | we observed until t)*

$$p_{trs} \propto p(h_{t-1} = r, h_t = s, d_t \mid d_1^{t-1}, \theta)$$
$$= \pi_{t-1}(r \mid \theta) q(r, s) P_s(d_t \mid \theta),$$

- Backward recursion

$$p'_{trs} = p(h_{t-1} = r \mid h_t = s, d_1^n, \theta) p(h_t = s \mid d_1^n, \theta)$$
$$= p(h_{t-1} = r \mid h_t = s, d_1^t, \theta) \pi'_t(s \mid \theta)$$

$$\pi'_t(s \mid \theta) \equiv \Pr(h_t = s \mid d_1^n, \theta)$$

Forward Backward Recursions

- Forward recursion is as likelihood recursion
- Backward variable: $\pi_t(s \mid \theta) = \ell_t(s)/\ell_t^*$

  where $\ell_t^* = \sum_{s=0}^{S-1} \ell_t(s)$
- Transition probabilities, *p(r - > s at time t | we observed until t)*

$$p_{trs} \propto p(h_{t-1} = r, h_t = s, d_t \mid d_1^{t-1}, \theta)$$

$$= \pi_{t-1}(r \mid \theta) q(r, s) P_s(d_t \mid \theta),$$

- Backward recursion

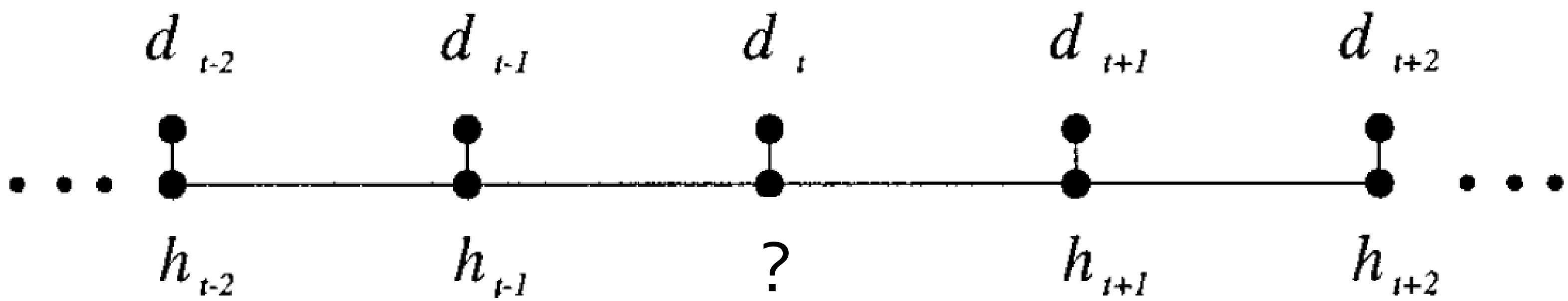
$$p'_{trs} = p(h_{t-1} = r \mid h_t = s, d_1^n, \theta) p(h_t = s \mid d_1^n, \theta)$$

$$= p(h_{t-1} = r \mid h_t = s, d_1^t, \theta) \pi'_t(s \mid \theta)$$

$$= p_{trs} \frac{\pi'_t(s \mid \theta)}{\pi_t(s \mid \theta)},$$

Sampling

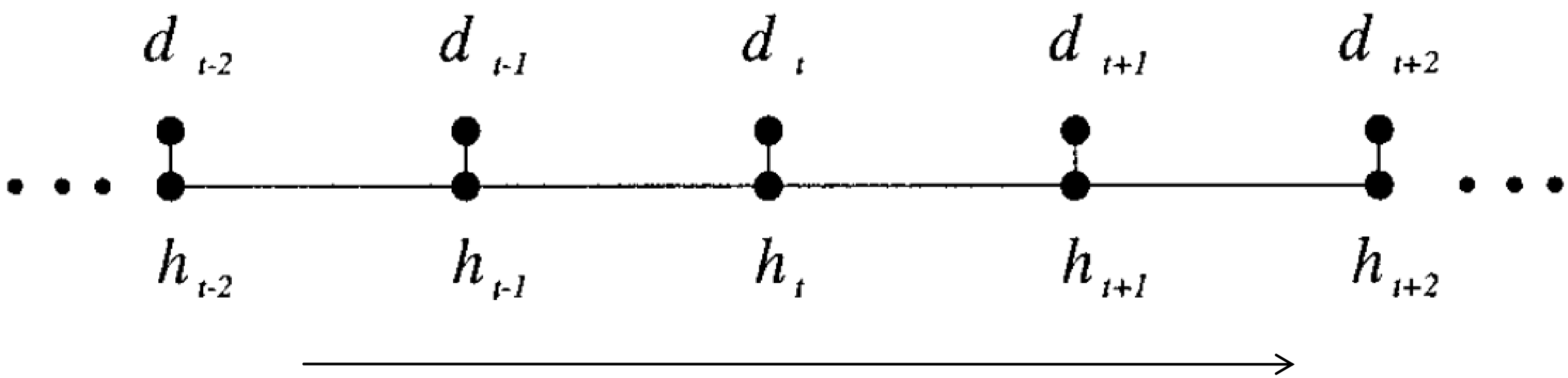
- Direct Gibbs Sampling



$$p(h_t = s \mid h_{-t}, d_1^n, \theta) \propto q(h_{t-1}, s)q(s, h_{t+1})P_s(d_t \mid \theta),$$

Sampling

- Forward Backward Recursion sampling

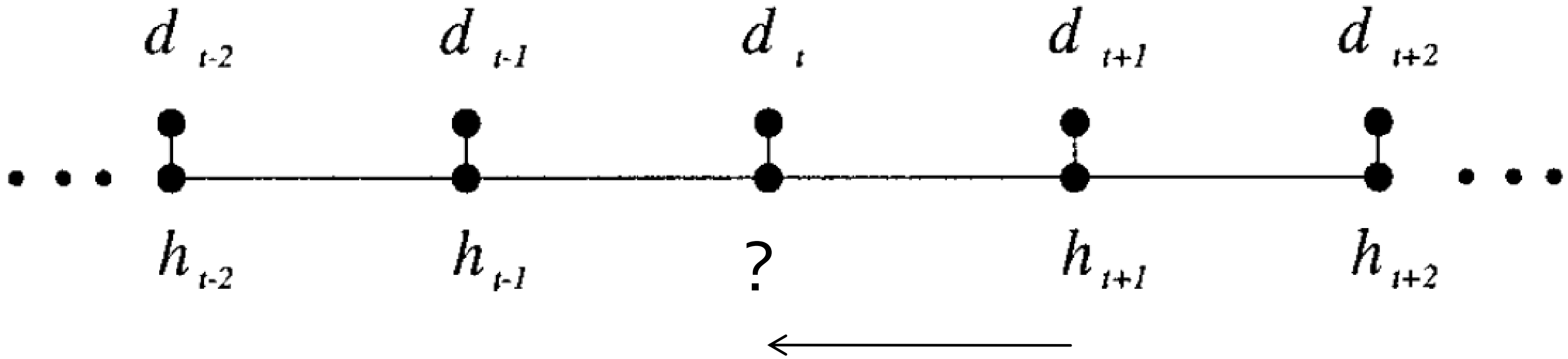


At the forward step, the transition matrices, (P) are produced;

$$p_{trs} \propto p(h_{t-1} = r, h_t = s, d_t \mid d_1^{t-1}, \theta)$$
$$= \pi_{t-1}(r \mid \theta) q(r, s) P_s(d_t \mid \theta),$$

Sampling

- Forward Backward Recursion sampling



At the backward step, the state is sampled by

$$p(h_{n-t} = r \mid h^n_{n-t+1}, d^n_1, \theta) \quad \propto p_{n-t+1, r, h_{t+1}} .$$

Autocovariance

- ***T*** is a vector that has sufficient statistics for state transitions. $T^{(\tau)}$ is the set of all such vectors iteration τ. ($\mathbf{T}_1$ is for time 1)

.

$$\mathbf{T}_1 = \begin{matrix} I(h_1, h_1) \\ I(h_1, h_2) \\ \cdot \\ \cdot \\ \cdot \\ I(h_s, h_s) \end{matrix}$$

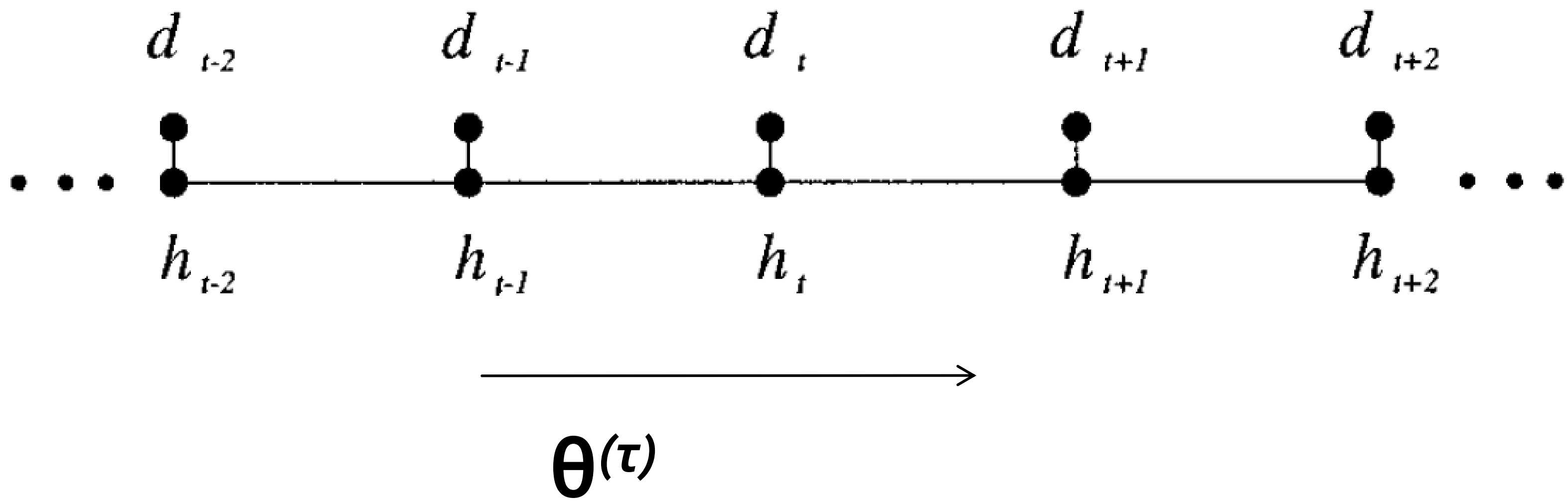
$$T = \sum_{j=1}^{m} T_j.$$

Autocovariance

- ***T*** is a vector that has sufficient statistics for state transitions. $T^{(\tau)}$ is the set of all such vectors iteration τ. ($\mathbf{T}_1$ is for time 1)
- *Let* $\theta^{(\tau)}$ be the sufficient statistics for emissions.

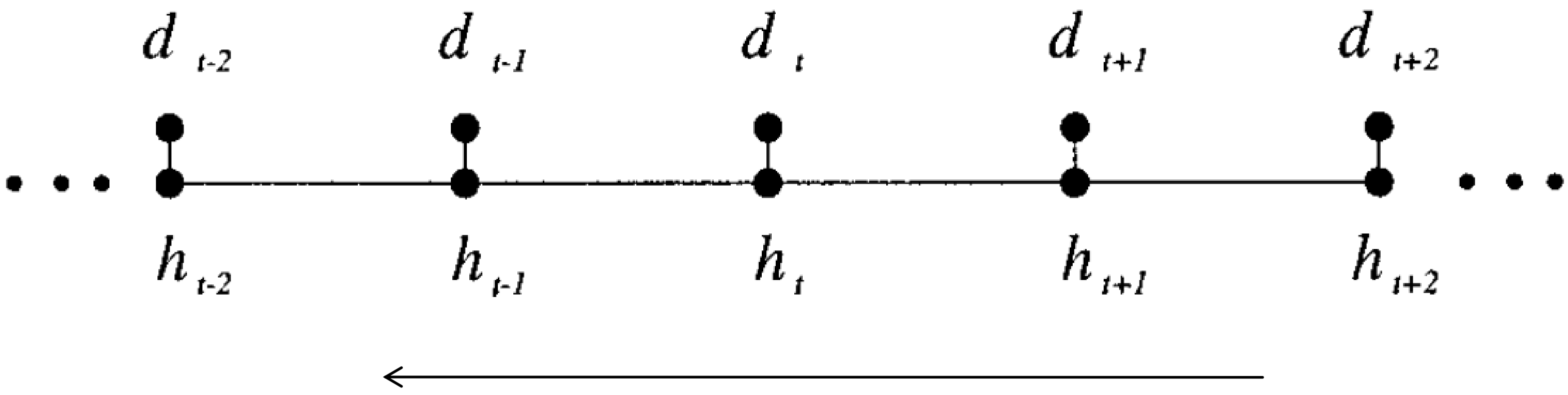
Autocovariance

- T is a vector that has sufficient statistics for state transitions. $T^{(\tau)}$ is the set of all such vectors iteration τ. ($\mathbf{T}_1$ is for time 1)
- *Let* $\theta^{(\tau)}$ be the sufficient statistics for emission probabilities.
- FB: $T^{(\tau)}$ is conditionally independent of $T^{(\tau+1)}$ given $\theta^{(\tau)}$.



Autocovariance

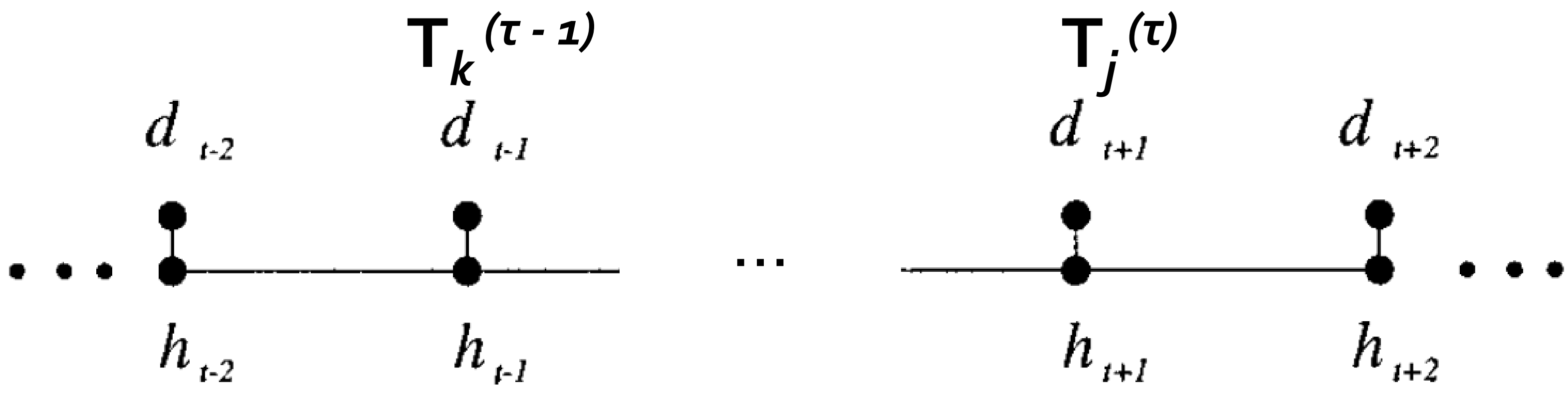
- T is a vector that has sufficient statistics for state transitions. $T^{(\tau)}$ is the set of all such vectors iteration τ. ($\mathbf{T}_1$ is for time 1)
- *Let* $\theta^{(\tau)}$ be the sufficient statistics for emission probabilities.
- FB: $T^{(\tau)}$ is conditionally independent of $T^{(\tau+1)}$ given $\theta^{(\tau)}$.



$\mathbf{T}^{(\tau)}$

Autocovariance

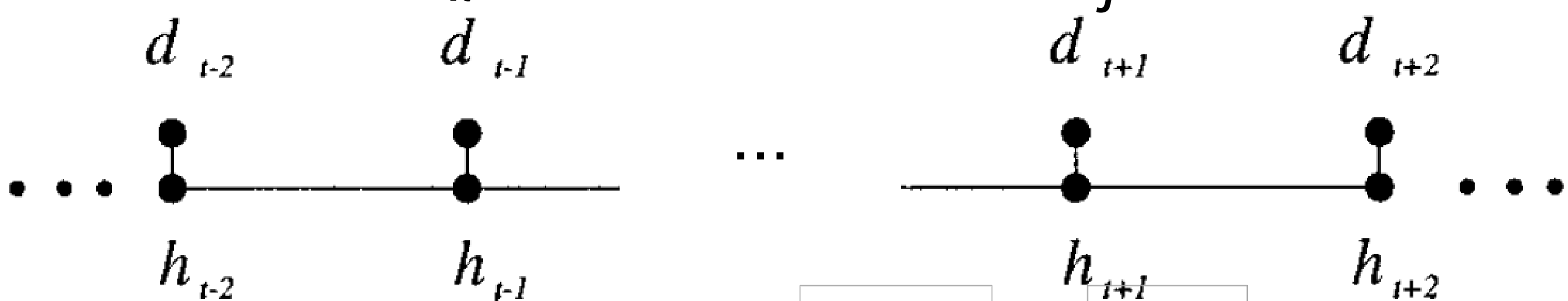
- T is a vector that has sufficient statistics for state transitions. $T^{(\tau)}$ is the set of all such vectors iteration τ. ($\mathbf{T}_1$ is for time 1)
- *Let* $\theta^{(\tau)}$ be the sufficient statistics for emission probabilities.
- FB: $T^{(\tau)}$ is conditionally independent of $T^{(\tau+1)}$ given $\theta^{(\tau)}$.
- DG: $\mathbf{T}_j^{(\tau - 1)}$ is conditionally independent of $\mathbf{T}_k^{(\tau)}$ given $\theta^{(\tau)}$.



Autocovariance

- *T* is a vector that has sufficient statistics for state transitions. $T^{(\tau)}$ is the set of all such vectors iteration τ. ($\mathbf{T}_1$ is for time 1)
- *Let* $\theta^{(\tau)}$ be the sufficient statistics for emission probabilities.
- FB: $T^{(\tau)}$ is conditionally independent of $T^{(\tau+1)}$ given $\theta^{(\tau)}$.
- DG: $\mathbf{T}_j^{(\tau - 1)}$ is conditionally independent of $\mathbf{T}_k^{(\tau)}$ given $\theta^{(\tau)}$.

$\mathbf{T}_k^{(\tau - 1)}$ $\mathbf{T}_j^{(\tau)}$



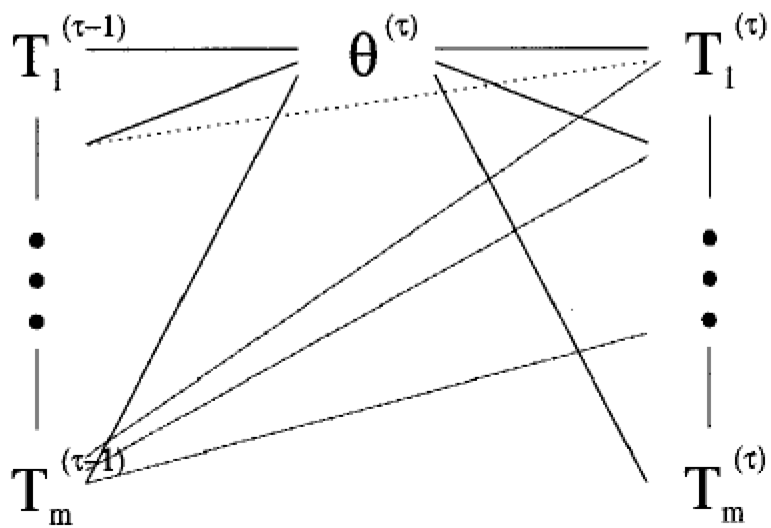
$\mathbf{T}_k^{(\tau - 1)}$ cond indep $\mathbf{T}_j^{(\tau)}$, if $\theta^{(\tau)}$, $\mathbf{T}_p^{(\tau-1)}$ $\{p > j\}$ and $\mathbf{T}_l^{(\tau)}$ $\{l < j\}$ are given.

Autocovariance

- $\boldsymbol{T}$ is a vector that has sufficient statistics for state transitions. $\boldsymbol{T}^{(\tau)}$ is the set of all such vectors iteration τ. ($\mathbf{T}_1$ is for time 1)
- *Let* $\boldsymbol{\theta}^{(\tau)}$ be the sufficient statistics for emission probabilities.
- FB: $\boldsymbol{T}^{(\tau)}$ is conditionally independent of $\boldsymbol{T}^{(\tau+1)}$ given $\boldsymbol{\theta}^{(\tau)}$.
- DG : $\mathbf{T}_j^{(\tau-1)}$ is conditionally independent of $\mathbf{T}_k^{(\tau)}$ given $\boldsymbol{\theta}^{(\tau)}$.

(a) FB

(b) DG



Autocovariance

- T is a vector that has sufficient statistics for state transitions. $T^{(\tau)}$ is the set of all such vectors iteration τ. ($\mathbf{T}_1$ is for time 1)
- *Let* $\theta^{(\tau)}$ be the sufficient statistics for emission probabilities.
- FB: $T^{(\tau)}$ is conditionally independent of $T^{(\tau+1)}$ given $\theta^{(\tau)}$.
- DG : $\mathbf{T}_j^{(\tau-1)}$ is conditionally independent of $\mathbf{T}_k^{(\tau)}$ given $\theta^{(\tau)}$.

$$
\begin{aligned}
\operatorname{cov}(T^{(\tau-1)}, T^{(\tau)}) &= E\{\operatorname{cov}(T^{(\tau-1)}, T^{(\tau)} \mid \theta^{(\tau)})\} \\
&\quad + \operatorname{cov}\{E(T^{(\tau-1)} \mid \theta^{(\tau)}), E(T^{(\tau)} \mid \theta^{(\tau)})\} \\
&= E\{\operatorname{cov}(T^{(\tau-1)}, T^{(\tau)} \mid \theta^{(\tau)})\} \\
&\quad + \operatorname{var}\{E(T^{(\tau)} \mid \theta^{(\tau)})\}, \qquad (11)
\end{aligned}
$$

Autocovariance

- T is a vector that has sufficient statistics for state transitions. $T^{(\tau)}$ is the set of all such vectors iteration τ. ($\mathbf{T}_1$ is for time 1)
- *Let* $\theta^{(\tau)}$ be the sufficient statistics for emission probabilities.
- FB: $T^{(\tau)}$ is conditionally independent of $T^{(\tau+1)}$ given $\theta^{(\tau)}$.
- DG : $\mathbf{T}_j^{(\tau-1)}$ is conditionally independent of $\mathbf{T}_k^{(\tau)}$ given $\theta^{(\tau)}$.

$$
\begin{aligned}
\operatorname{cov}(T^{(\tau-1)}, T^{(\tau)}) &= E\{\operatorname{cov}(T^{(\tau-1)}, T^{(\tau)} \mid \theta^{(\tau)})\} \\
&\quad + \operatorname{cov}\{E(T^{(\tau-1)} \mid \theta^{(\tau)}), E(T^{(\tau)} \mid \theta^{(\tau)})\} \\
&= \boxed{E\{\operatorname{cov}(T^{(\tau-1)}, T^{(\tau)} \mid \theta^{(\tau)})\}} \quad = 0 \text{ for FB} \\
&\quad + \operatorname{var}\{E(T^{(\tau)} \mid \theta^{(\tau)})\}, \qquad (11)
\end{aligned}
$$

Autocovariance

- T is a vector that has sufficient statistics for state transitions. $T^{(\tau)}$ is the set of all such vectors iteration τ. ($\mathbf{T}_1$ is for time 1)
- *Let* $\theta^{(\tau)}$ be the sufficient statistics for emission probabilities.
- FB: $T^{(\tau)}$ is conditionally independent of $T^{(\tau+1)}$ given $\theta^{(\tau)}$.
- DG : $\mathbf{T}_j^{(\tau-1)}$ is conditionally independent of $\mathbf{T}_k^{(\tau)}$ given $\theta^{(\tau)}$.

$$\begin{aligned}\operatorname{cov}(T^{(\tau-1)}, T^{(\tau)}) &= E\{\operatorname{cov}(T^{(\tau-1)}, T^{(\tau)} \mid \theta^{(\tau)})\} \\ &\quad + \operatorname{cov}\{E(T^{(\tau-1)} \mid \theta^{(\tau)}), E(T^{(\tau)} \mid \theta^{(\tau)})\} \\ &= E\{\operatorname{cov}(T^{(\tau-1)}, T^{(\tau)} \mid \theta^{(\tau)})\} \\ &\quad \boxed{+\operatorname{var}\{E(T^{(\tau)} \mid \theta^{(\tau)})\},} \qquad (11)\end{aligned}$$

same for FB and DG **<=**

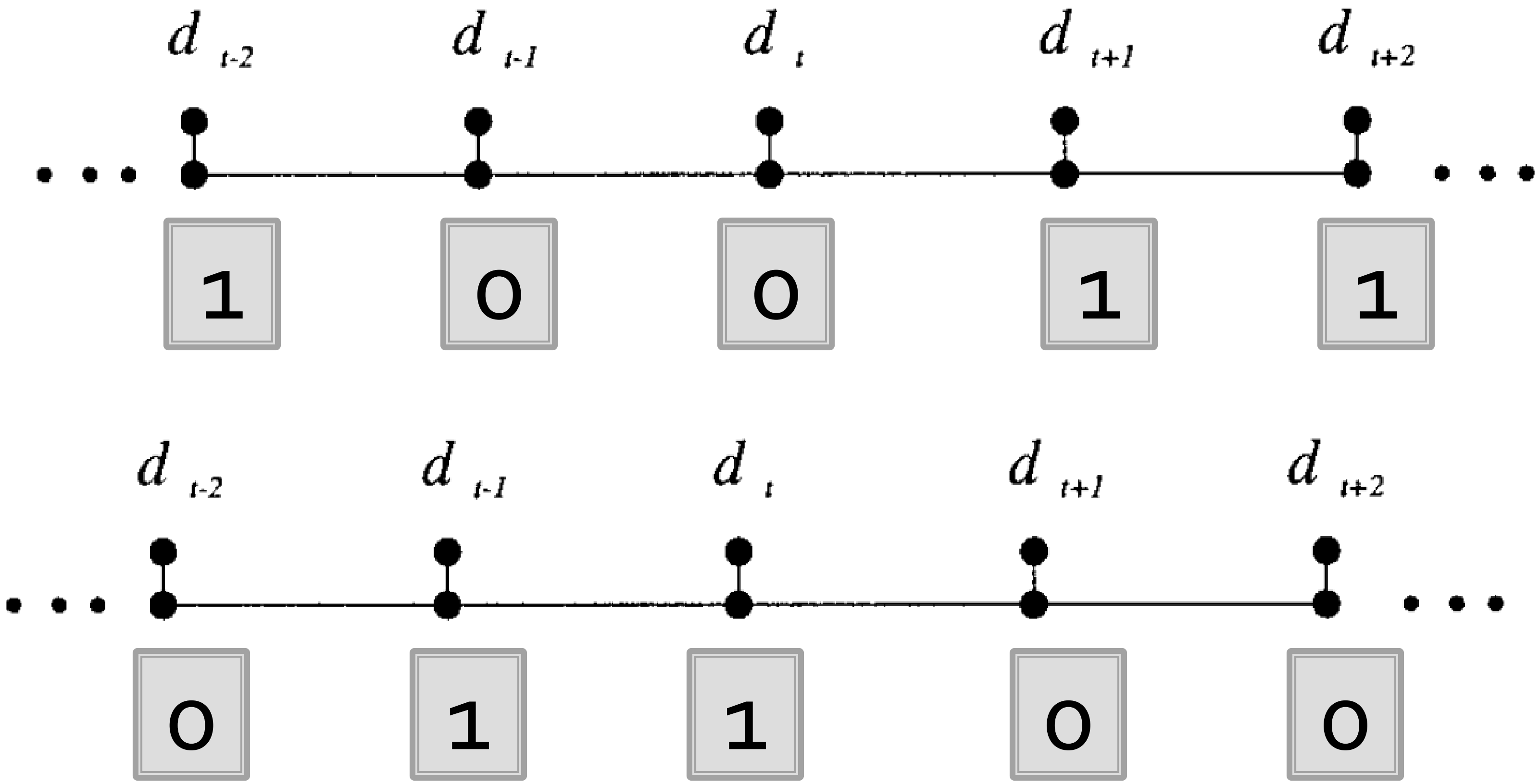
Autocovariance

- $\boldsymbol{T}$ is a vector that has sufficient statistics for state transitions. $\boldsymbol{T}^{(\tau)}$ is the set of all such vectors iteration τ. ($\mathbf{T}_1$ is for time 1)
- *Let* $\boldsymbol{\theta}^{(\tau)}$ be the sufficient statistics for emission probabilities.
- FB: $\boldsymbol{T}^{(\tau)}$ is conditionally independent of $\boldsymbol{T}^{(\tau+1)}$ given $\boldsymbol{\theta}^{(\tau)}$.
- DG : $\mathbf{T}_j^{(\tau-1)}$ is conditionally independent of $\mathbf{T}_k^{(\tau)}$ given $\boldsymbol{\theta}^{(\tau)}$.

$$\mathrm{cov}_{\mathrm{DG}}(T^{(\tau-1)}, T^{(\tau)}) = \mathrm{cov}_{\mathrm{FB}}\left(T^{(\tau-1)}, T^{(\tau)}\right) + E_{\mathrm{DG}}\left\{\mathrm{cov}_{\mathrm{DG}}\left(T^{(\tau-1)}, T^{(\tau)} \mid \theta^{(\tau)}\right)\right\}.$$

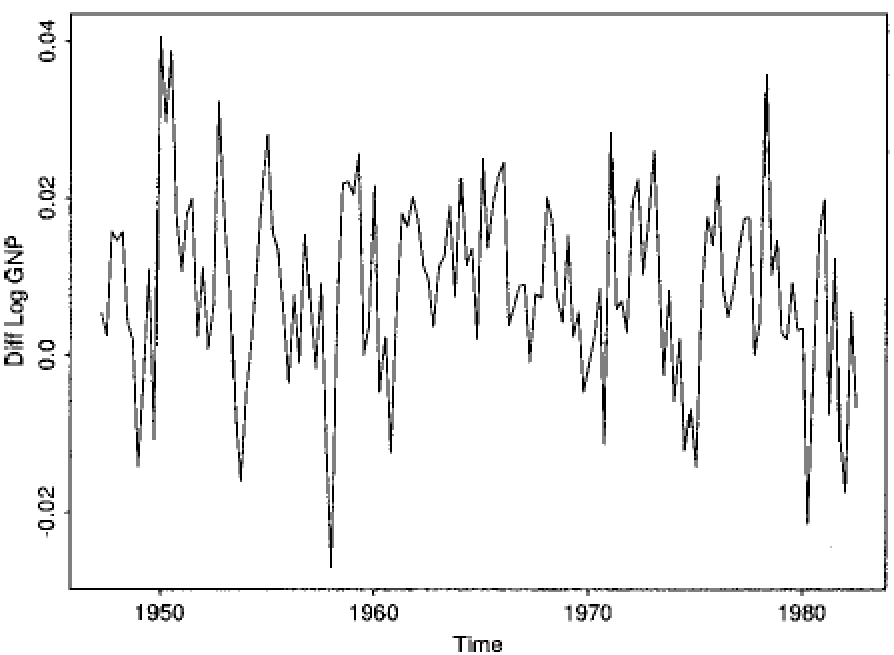
Some issues

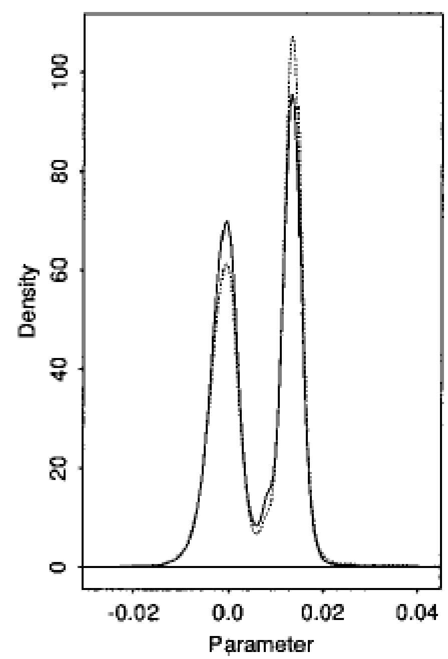
- Label switching

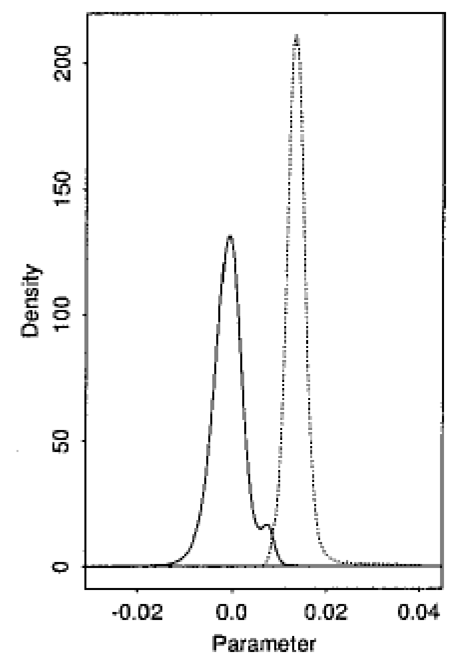


Some issues

- Label switching
  - Implications
  - Solution: constraints







Some issues

- Label switching
- Collapsed states
  - May be evidence for over parameterizations
  - Priors

Estimating the hidden states

- Stating $\pi_t(s)$ is usually sufficient.

Estimating the hidden states

- Stating $\pi_t(s)$ is usually sufficient.
- Overall configuration may be needed;
  - Marginal Distributions
  - MAP estimates

Estimating the hidden states

- Stating $\pi_t(s)$ is usually sufficient.
- Overall configuration may be needed;
  - Marginal Distributions
    - Averaging over all runs (1 – m) with indicator function:

$$\tilde{\pi}'_t(s) = 1/m \sum_{j=1}^{m} I(h_t^{(j)} = s).$$

Estimating the hidden states

- Stating $\pi_t(s)$ is usually sufficient.
- Overall configuration may be needed;
  - Marginal Distributions
    - Averaging over all runs (1 – m) with indicator function
    - Averaging over all runs (1 – m) probabilities

      (Rao-Blackwellized estimate)

$$\hat{\pi}'_t(s) = 1/m \sum_{j=1}^{m} \pi'_t(s \mid \theta^{(j)})$$

Estimating the hidden states

- Stating $\pi_t(s)$ is usually sufficient.
- Overall configuration may be needed;
  - Marginal Distributions
    - Averaging over all runs (1 – m) with indicator function
    - Averaging over all runs (1 – m) probabilities
  - MAP estimate (L = max p (h, d | θ)

$$L_1(s) = \pi_0(s) P_s(d_1 \mid \theta)$$

$$L_t(s) = \max_r [L_{t-1}(r) q(r, s)] P_s(d_t \mid \theta).$$

Estimating the hidden states

- Stating $\pi_t(s)$ is usually sufficient.
- Overall configuration may be needed;
  - Marginal Distributions
    - Averaging over all runs (1 – m) with indicator function
    - Averaging over all runs (1 – m) probabilities
  - MAP estimate: to find $\hat{h}$

$$\hat{h}_t = \arg\max_{r \in \mathcal{S}} L_t(r) q(r, \hat{h}_{t+1})$$

converges when it is same for all s in h_{t+1}.

Size of the state space

- Calculating $p(S \mid D)$

Size of the state space

- Calculating p(S | D)

$$p(S \mid d_1^n) = \int p(S \mid d_1^n, \theta) p(\theta \mid d_1^n)\, d\theta$$

$$\approx 1/m \sum_{j=1}^{m} p(S \mid d_1^n, \theta^{(j)}),$$

$$p(S \mid d_1^n, \theta^{(j)}) \propto p(d_1^n \mid \theta_S^{(j)}, S) p(S)$$

Size of the state space

- Calculating p(S | D)
- Schwartz criterion C(S):

$$C(S) = \log \ell - k_S \log(n)/2,$$

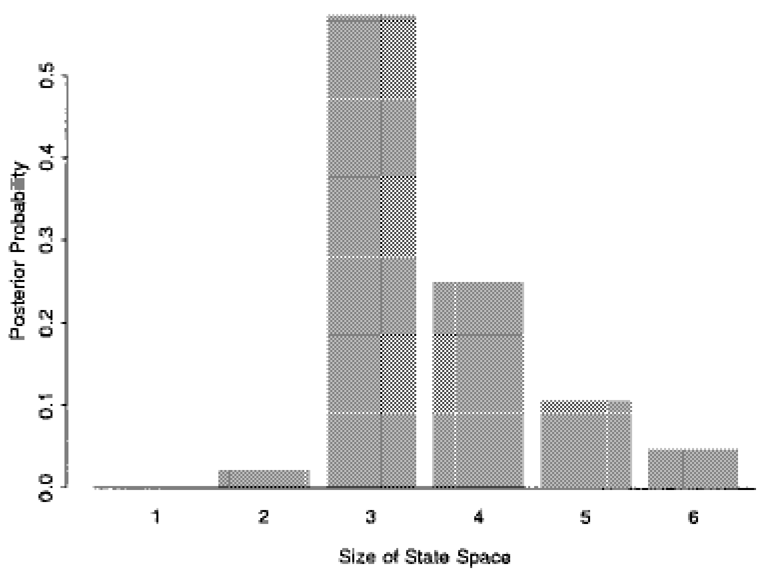
Size of the state space

- Calculating p(S | D)
- Schwartz criterion C(S):
- Bayesian Information Criterion BIC:
  - p(S | D) – 2 C(S)

Size of the state space

- Calculating p(S | D)
- Schwartz criterion C(S):
- Bayesian Information Criterion BIC

(a)



(b)

S	maximized log-posterior	k_S	$C(S)$	BIC
1	-174.3	1	-177.0	354.0
2	-150.7	4	-161.6	323.2
3	-140.7	9	-165.3	330.6
4	-139.2	16	-183.1	366.2
5	-139.5	25	-208.0	416.0
6	-139.8	36	-238.4	476.8

Thank you