State Space and Hidden Markov Models

Kunsch, H.R., *State Space and Hidden Markov Models*. ETH-Zurich, Zurich;

Aliaksandr Hubin

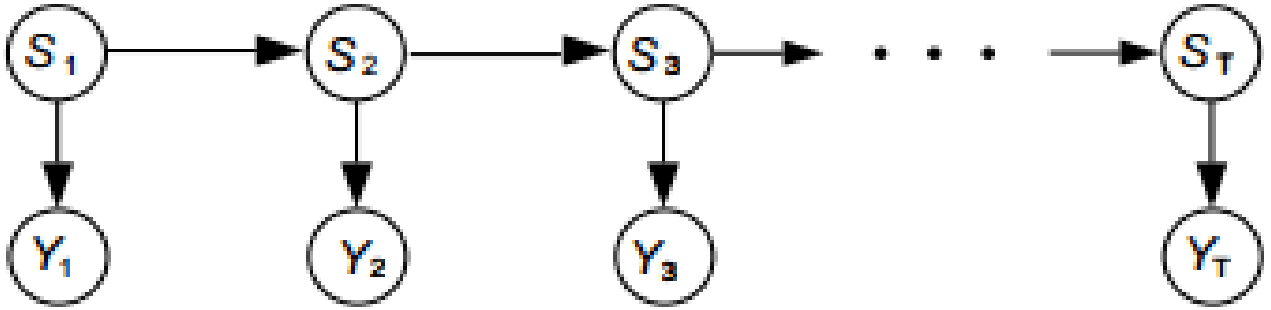
Oslo 2014

Contents

Introduction

HMM

- **Broadly speaking we will address the models where observations are noisy and incomplete functions of some underlying unobservable process, called the state process, which is assumed to have simple markovian dynamics.**



Markov chains

- Set of states: $x_i \in \{m_1, m_2, \ldots, m_N\}, i = 1,2,\ldots$
- Process moves from one state to another generating a sequence of states : $x_1, x_2, \ldots, x_k, \ldots$
- Markov chain property: probability of each subsequent state depends only on what was the previous state: $P(x_k \mid x_1, x_2, \ldots, x_{k-1}) = P(x_k \mid x_{k-1})$
- To define Markov chain, the following probabilities have to be specified:

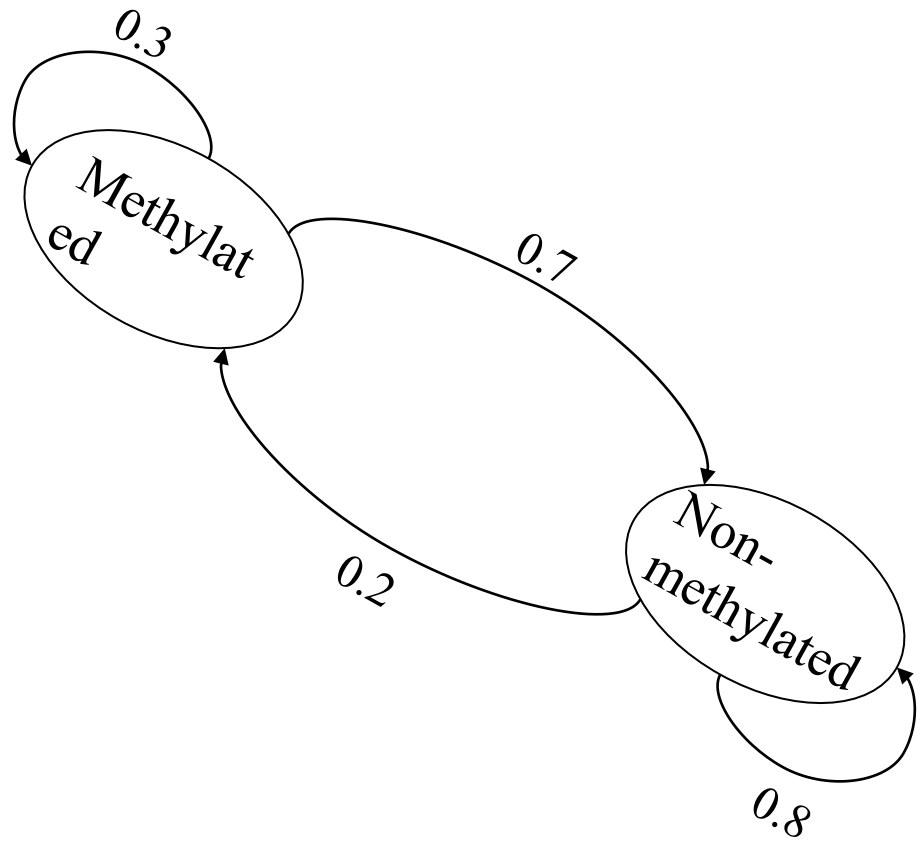
transition probabilities matrix $a_{ij} = P(m_i \mid m_j)$ and initial probabilities $\pi_i = P(m_i)$

- The output of the process is the set of states at each instant of time
- Joint probability of all states sequence:

$$\begin{aligned} P(x_1, x_2, \ldots, x_k) &= P(x_k \mid x_1, x_2, \ldots, x_{k-1}) P(x_1, x_2, \ldots, x_{k-1}) \\ &= P(x_k \mid x_{k-1}) P(x_1, x_2, \ldots, x_{k-1}) = \ldots \\ &= P(x_k \mid x_{k-1}) P(x_{k-1} \mid x_{k-2}) \ldots P(x_2 \mid x_1) P(x_1) \end{aligned}$$

Simple Example of Markov Model

- Epigenetic state process is considered
- Two states : ‘Methylated’ and ‘Non-methylated’.
- Initial probabilities: P(‘Methylated’)=0.4 , P(‘Non-methylated’)=0.6
- Transition probabilities described in the graph ->
- Inference example:
  Suppose we want to calculate a probability of a sequence of states in our example, {‘Methylated’,’ Methylated’,’ Non-methylated’, Non-methylated’}. This then corresponds to 0.4*0.3*0.7*0.8 = 6.72%



Hidden Markov Models

- The observations are represented by a probabilistic function (discrete or continuous) of a state instead of an one-to-one correspondence of a state
- The following components describe a Hidden Markov Model in the simplest case:

1. Distribution of transition probabilities $a_{ij} = P(m_i \mid m_j)$;
2. Distribution of initial state probabilities $\pi_i = P(m_i)$;
3. Distribution of observation probabilities $\vartheta_i = P(y_i|x_i)$.

- $x_i \in R^N$ - are usually addressed as **state space models**;
- $x_i \in \{m_1, \dots, m_N\}$ – are usually addressed as **hidden markov chains**;

✓ Parameters listed above estimation might be very challenging for complicated models!

Hidden Markov Models

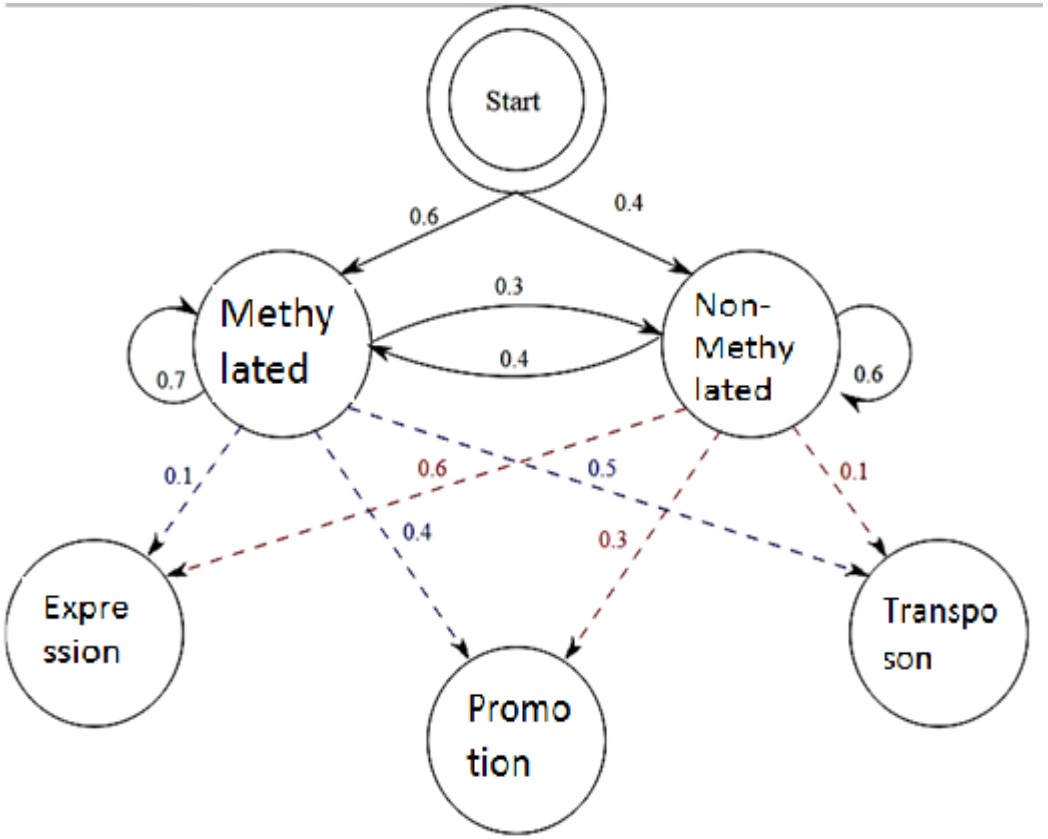
The following components describe a Hidden Markov Model in the general case:

1. Distribution of initial state probabilities $X_0 \sim a_0(x)d\mu(x);$
2. Distribution of transition probabilities $X_t|(X_{t-1} = x_{t-1}) \sim a_t(x_{t-1}, x)d\mu(x);$
3. Distribution of observation probabilities $Y_t|(X_t = x_t) \sim b_t(x_t, y)dv(y);$
4. The joint density of the process and observations then looks as follows:

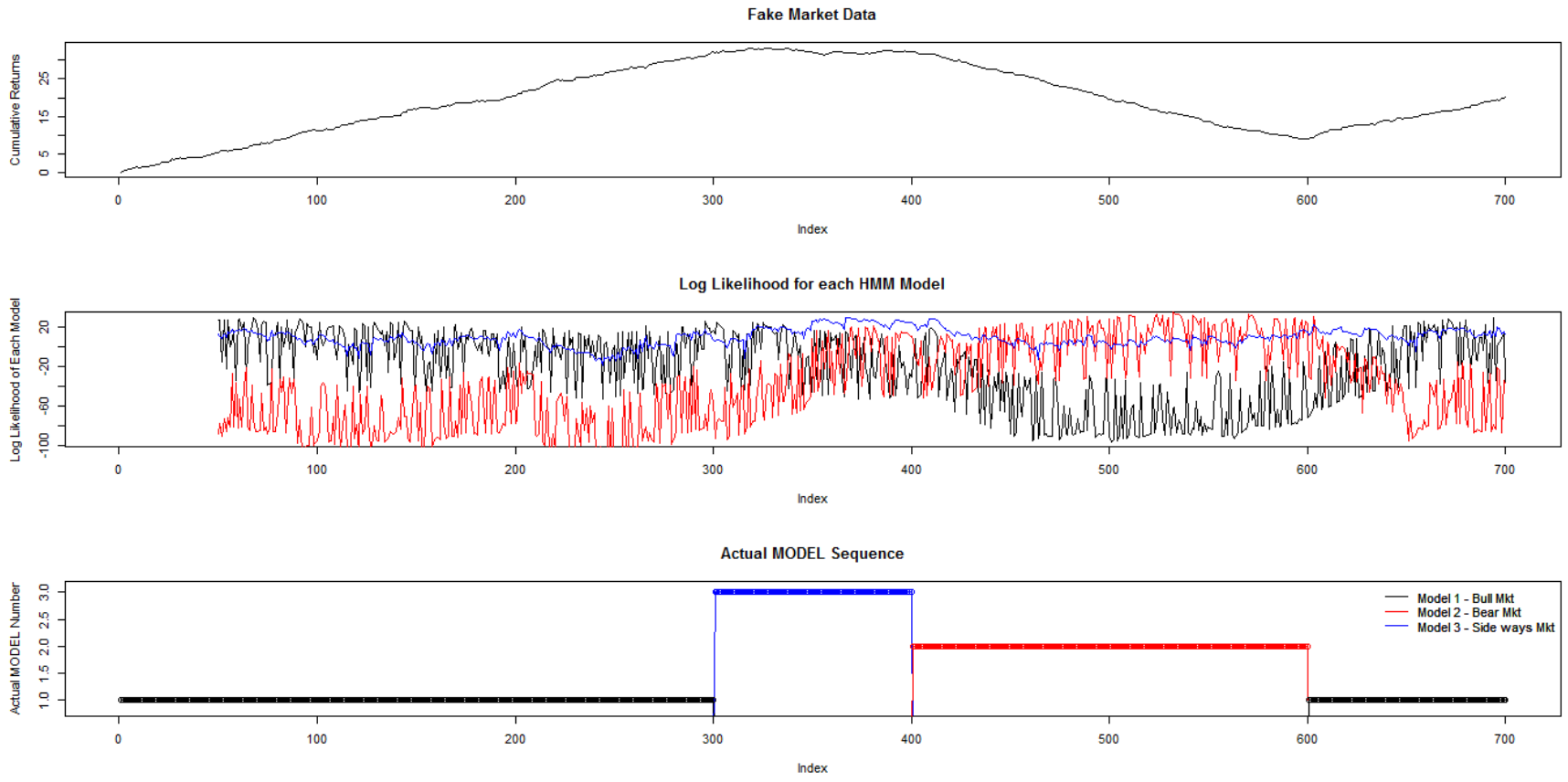
$$P(X_0, \dots, X_T, Y_1, \dots, Y_T) = a_0(x_0) \prod_{t=1}^{T} a_t(x_{t-1}, x_t) b_t(x_t, y_t) \, .$$

Simple Example of HMM

- Whether state process is considered
- Two discrete states : ‘Methylated’ and ‘Non-methylated’.
- Initial probabilities: P(‘Methylated’)=0.4 , P(‘Non-methylated’)=0.6
- Transition probabilities described in the graph ->
- Locations associated are observed with respect to their state-dependent probabilities
- Corresponding observation probabilities are represented in the graph ->



Graphical illustration of the Hidden Markov Model

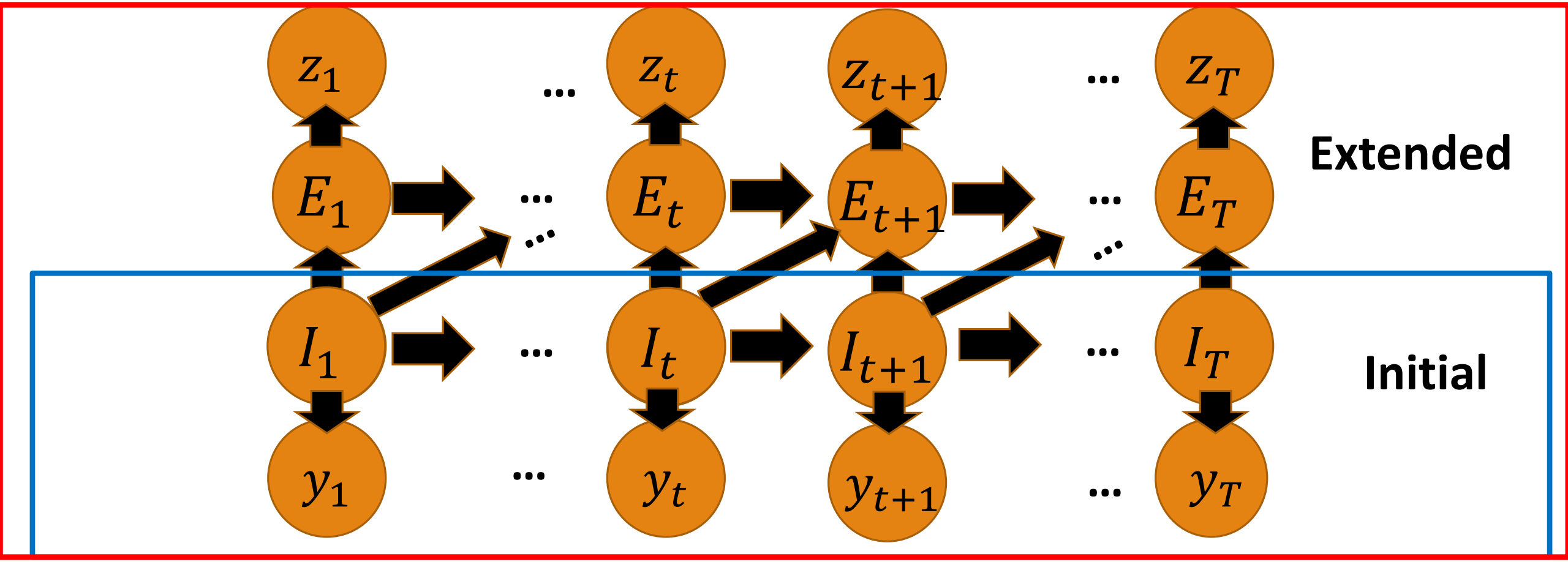


Project application

Let further:

- I_t be a binary variable indication whether location *t* is methylated
- E_t be a binary varibale indicating whether the gene, to which location t belongs is expressed
- n_t be an amount of reads for location *t*
- y_t be a binomial variable indicating the number of methylated reads at location *t*
- z_t be a quantitative measure of some phenotypic response for the expressed genes at location *t*

Project application



Project application

- $I_t = \begin{cases} 1, & - \text{ location } t \text{ is methylated} \\ 0, & - \text{ location } t \text{ is not methylated} \end{cases} \sim P(I_t) = \begin{cases} P_1, & - \text{ location } t \text{ is methylated} \\ P_2, & - \text{location } t \text{ is not methylated} \end{cases}$

- $P_{ij} = P(I_t = j | I_{t-1} = i)$ – define dynamic structures in the given neighborhood

- $y_t | I_t, n_t$ – number of metylated observations in given reads and methylation status

- $P(y_t | I_t, n_t) = P_{Binom(n_t, P(I_t))}(y_t)$ – it has binomial distribution

Extensions of the model

- Let p_t be continuous: define stochastic process $p(I_t) = \begin{cases} Beta(\beta \frac{p_{t-1}}{1-p_{t-1}}) \\ Beta(\beta \frac{q_{I_t}}{1-q_{I_t}}) \end{cases}$, giving similarities when the state do not change but a renewal in cases where the state changes.
- Link state transition probabilities to underlying genomic structure
- Look at more global structures than transition probabilities P_{ij}:

-address more complex models describing more global structures (ARIMA, n-markov chains, etc.)

-use simple markov chain model above as a first step, but then extract more global features of $\{I_t\}$

- Consider more complex spatial structures than the one-directional approach above.
- Simultaneous modelling of several tissues and/or individuals at the same time.

Parameters' estimation & Inference

1. **Viterbi algorithm** for fitting the joint distribution of the most probable sequence of states $argmax_{x_{1:T}}\{P(x_{1:T}|d_{1:T})\}$
2. **Forward algorithm** for fitting filtered probability of a given state, given data up to $P(x_t|d_1, \dots, d_t)$
3. **Forward–backward algorithm** for fitting the smoothed probability of a given state, given all data $P(x_k|d_1, \dots, d_T)$
4. **Maximal likelihood** maximization or **Bayesian** methods for parameters' vector θ estimation

❖ Where d_t is data at point *t;*

❖Note that these algorithms are linear in the number of time points.

Inference. Filtering.

1. Recursion in k for prediction density of the states:

$$f_{t+k|t}(x_{t+k}|y_1^t) = \int a_{t+k}(x, x_{t+k}) f_{t+k-1|t}(x|y_1^t) d\mu(x)$$

2. Prediction densities for the observations given the corresponding prediction densities of the states:

$$p(y_{t+k}|y_1^t) = \int b_{t+k}(x, y_{t+k}) f_{t+k}(x|y_1^t) d\mu(x)$$

3. Thus, filtering densities of the states can be computed according to the following forward recursion in t (starting with $f_{0|0}$):

$$f_{t+1|t+1}(x_{t+1}|y_1^{t+1}) = \frac{b_{t+1}(x_{t+1}, y_{t+1}) f_{t+1|t}(x_{t+1}|y_1^t)}{p(y_{t+k}|y_1^t)}$$

Interface. Smoothing.

1. Conditional on all observations sequence of states is the markov chain with the following transition densities:

$$\begin{aligned} p(x_t \mid x_{t-1}, y_1^T) &= p(x_t \mid x_{t-1}, y_t^T) \\ &= \frac{a_t(x_{t-1}, x_t) b_t(x_t, y_t) p(y_{t+1}^T \mid x_t)}{p(y_t^T \mid x_{t-1})} \end{aligned}$$

, where

$$p(y_t^T \mid x_{t-1}) = \int a_t(x_{t-1}, x_t) b_t(x_t, y_t) p(y_{t+1}^T \mid x_t) d\mu(x_t)$$

2. The backward transition densities then become:

$$p(x_t \mid x_{t+1}, y_1^T) = p(x_t \mid x_{t+1}, y_1^t) = \frac{a_{t+1}(x_t, x_{t+1}) f_{t|t}(x_t \mid y_1^t)}{f_{t+1|t}(x_{t+1} \mid y_1^t)}$$

Interface. Smoothing.

The smoothing densities are then computed with respect to the recursions in *t:*

$$f_{t|T}(x_t \mid y_1^T) = f_{t|t}(x_t \mid y_1^t) \int \frac{a_{t+1}(x_t, x)}{f_{t+1|t}(x \mid y_1^t)} f_{t+1|T}(x \mid y_1^T) d\mu(x)$$

and

$$f_{t|T}(x_t \mid y_1^T) = f_{t|t}(x_t \mid y_1^t) r_{t|T}(x_t, y_1^T), \quad r_{t|T}(x_t, y_1^T) = \frac{\int b_{t+1}(x, y_{t+1}) a_{t+1}(x_t, x) r_{t+1|T}(x, y_1^T) d\mu(x)}{p(y_{t+1} \mid y_1^t)}$$

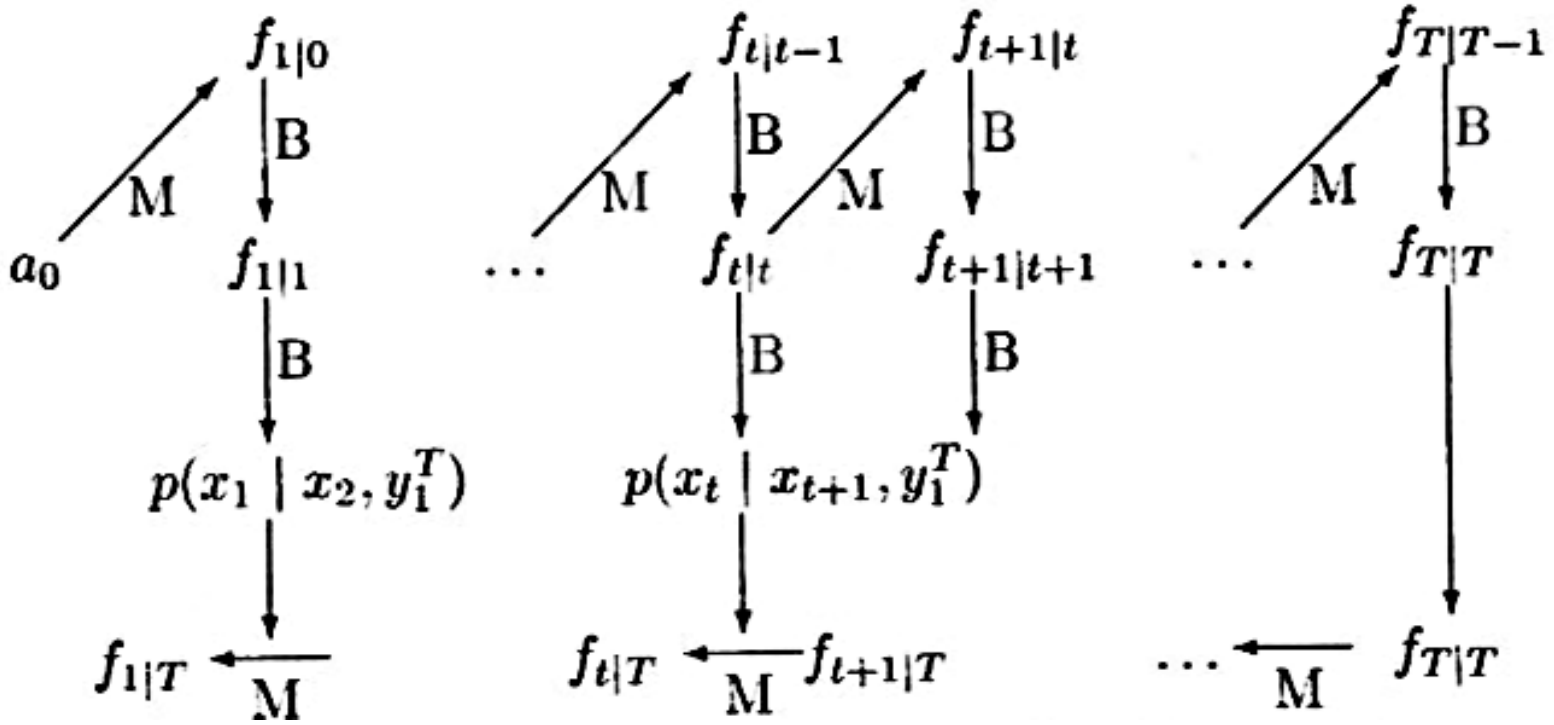
where

$$r_{t|T}(x_t, y_1^T) = \frac{p(y_{t+1}^T \mid x_t)}{p(y_{t+1}^T \mid y_1^t)}, \quad r_{T|T} \equiv 1$$

5. Note that the joint density of the subsequence of states, which might be required at ML stage:

$$p(x_s^t \mid y_1^T) = f_{t|T}(x_t \mid y_1^T) \prod_{r=s}^{t-1} p(x_r \mid x_{r+1}, y_1^T)$$

Recursion order



Inference on a function of states.

6. It can also shown that having the smoothing distribution of the sequence of states it is easy to obtain the conditional expectation of some function assigned to the sequence, which can be recursively updated when a new observation becomes available ($t>s$):

$$m_t(x_t, y_1^{t-1}) = \mathbb{E}[h(X_1^s) \mid X_t = x_t, Y_1^{t-1} = y_1^{t-1}]$$

$$m_{t+1}(x_{t+1}, y_1^t) = \frac{\int m_t(x_t, y_1^{t-1}) f_{t|t}(x_t \mid y_1^t) a_{t+1}(x_t, x_{t+1}) d\mu(x_t)}{\int f_{t|t}(x_t \mid y_1^t) a_{t+1}(x_t, x_{t+1}) d\mu(x_t)}$$

$$\mathbb{E}[h(X_1^s) \mid Y_1^t] = \int m_t(x_t, y_1^{t-1}) f_{t|t}(x_t \mid y_1^t) d\mu(x_t)$$

Posterior mode estimation

We want to get the posterior mean of the states, namely:

$$\hat{x}_0^T = \arg\max_{x_0^T} p(x_0^T \mid y_1^T)$$

Maximization of the posterior joint distribution of its states is invariant to being logorithmed:

$$\begin{aligned} \log p(x_0^T \mid y_1^T) &\propto \log p(x_0^T, y_1^T) \\ &= \log a_0(x_0) + \sum_{t=1}^{T} (\log a_t(x_{t-1}, x_t) + \log b_t(x_t, y_t)) \end{aligned}$$

Viterbi algorithm

Because of the special structure of the expression above the most likely value of x_0 depends on only on x_1 and after maximizing over x_0, the most likely value of x_1 depends only on x_2 and so on, which leads to the following dynamic programming algorithm:

$$
\begin{aligned}
\psi_0(x_0) &= \log a_0(x_0) \\
\psi_t(x_t) &= \max_{x_{t-1}}(\psi_{t-1}(x_{t-1}) + \log a_t(x_{t-1}, x_t)) + \log b_t(x_t, y_t) \\
\xi_{t-1}(x_t) &= \arg\max_{x_{t-1}}(\psi_{t-1}(x_{t-1}) + \log a_t(x_{t-1}, x_t)) \\
\hat{x}_T &= \arg\max_{x_T} \psi_T(x_T).
\end{aligned}
$$

After which other values of the sequence are recovered in the following way:

$$\hat{x}_{t-1} = \xi_{t-1}(\hat{x}_t)$$

Reference Probability Method

Let $\bar{P}$ define the distribution when states and observations are independent with distribution g for the observations, then the following ratio can be derived:

$$\Lambda_t = \frac{p(x_0^t, y_1^t)}{\bar{p}(x_0^t, y_1^t)} = \prod_{s=1}^{t} \frac{b_s(x_x, y_s)}{g(y_s)}$$

And for any absolutely continuous measurable transformation $\phi: x \to \mathbb{R}$, we have (easy to compute):

$$\mathbb{E}[\phi(X_t) \mid Y_1^t] = \frac{\bar{\mathbb{E}}[\Lambda_t \phi(X_t) \mid Y_1^t]}{\bar{\mathbb{E}}[\Lambda_t \mid Y_1^t]}$$

On the other hand:

$$\mathbb{E}[\phi(X_t) \mid Y_1^t = y_1^t] = \frac{\int h_{t-1}(x, \phi) f_{t-1|t-1}(x \mid y_1^{t-1}) d\mu(x)}{\int h_{t-1}(x, 1) f_{t-1|t-1}(x \mid y_1^{t-1}) d\mu(x)}$$

$$= \frac{\int \phi(x_t) b_t(x_t, y_t) \int f_{t-1|t-1}(x \mid y_1^{t-1}) a_t(x, x_t) d\mu(x)\ d\mu(x_t)}{\int b_t(x_t, y_t) \int f_{t-1|t-1}(x \mid y_1^{t-1}) a_t(x, x_t) d\mu(x)\ d\mu(x_t)}$$

Reference Probability Method

With

$$h_{t-1}(x,\phi) = \mathbb{E}[b_t(X_t,y_t)\phi(X_t) \mid X_{t-1}=x]$$
$$= \int b_t(x_t,y_t)\phi(x_t)a_t(x,x_t)d\mu(x_t).$$

From the last expression on previous slide one can easily derive filtering recursions for the states of the model:

$$f_{t|t}(x \mid y_1^t) \propto b_t(x_t,y_t)\int f_{t-1|t-1}(x \mid y_1^{t-1})a_t(x,x_t)d\mu(x).$$

Thus, $\frac{\overline{\mathbb{E}}[\Lambda_t\phi(X_t) \mid Y_1^t]}{\overline{\mathbb{E}}[\Lambda_t \mid Y_1^t]}$ gives us an easy way to compute the filtering recursions for the states,

Even for the case when the time becomes continuous (e.g. differential equation model for states)

Forgetting initial distributions

Let us define the transition operator $A^*$and Baves operator *B*:

$$A_t^* f(x) = \int f(x')a_t(x', x)d\mu(x') \quad \text{and} \quad B(f, b)(x) = \frac{f(x)b(x)}{\int f(x)b(x)d\mu(x)}$$

Then the recursion densities $f_{t|t} = B(A_t^* f_{t-1|t-1}, b_t(., y_t))$ which forget the initial distributions

if A^* and *B* are contracting for some norm of densities.

It can be shown that initial distribution of the states is forgotten exponentially fast, and thus changes in filtering distributions and **changes at fixed times disappear when the updates are made with the same observations**.

Linear and General State Space Model

Consider process for states $X_t = G_t X_{t-1} + V_t$

And for the observations $Y_t = H_t X_t + W_t$

Gaussian State Space Model

Where G_t and H_t are coefficient matrices;

And $V_t \sim N_k(0, \Sigma_t)$ and $W_t \sim N_k(0, \Omega_t)$ are independent Gaussian processes;

This can be generalized as

$$X_t = g_t(X_{t-1}, V_t)$$
$$Y_t = h_t(X_t, W_t)$$

General State Space Model

With g_t and h_t being some functions and V_t, W_t having the same meaning.

Kalman filtering and Smoothing

Then filtering densities for its hidden process look as follows:

$$
\begin{aligned}
m_{t|t} &= m_{t|t-1} + R_{t|t-1} H_t' M_t^{-1} (y_t - H_t m_{t|t-1}) \\
R_{t|t} &= R_{t|t-1} - R_{t|t-1} H_t' M_t^{-1} H_t R_{t|t-1} \\
\text{where} \qquad M_t &= \Omega_t + H_t R_{t|t-1} H_t'.
\end{aligned}
$$

Known as the so called **Kalman Filter.**

Kalman smoother correspondingly looks as follows:

$$
\begin{aligned}
m_{t|T} &= m_{t|t} + S_t (m_{t+1|T} - m_{t+1|t}) \\
R_{t|T} &= R_{t|t} - S_t (R_{t+1|t} - R_{t+1|T}) S_t' \\
\text{where} \qquad S_t &= R_{t|t} G_{t+1}' R_{t+1|t}^{-1}.
\end{aligned}
$$

Extended Kalman Filtering and Smoothing

Use the standard Kalman filers and smoothers for the approximation of the process below:

$$
\begin{aligned}
X_t &= g_t(m_{t-1|t-1}, 0) + \frac{\partial}{\partial x} g_t(m_{t-1|t-1}, 0)(X_{t-1} - m_{t-1|t-1}) \\
&+ \frac{\partial}{\partial v} g_t(m_{t-1|t-1}, 0) V_t \\
Y_t &= h_t(m_{t|t-1}, 0) \frac{\partial}{\partial x} h_t(m_{t-1|t-1}, 0)(X_t - m_{t|t-1}) \\
&+ \frac{\partial}{\partial w} h_t(m_{t-1|t-1}, 0) W_t.
\end{aligned}
$$

General Cases

- Use robust approximations for filtering and smoothing
- Use MCMC algorithms for filtering and smoothing

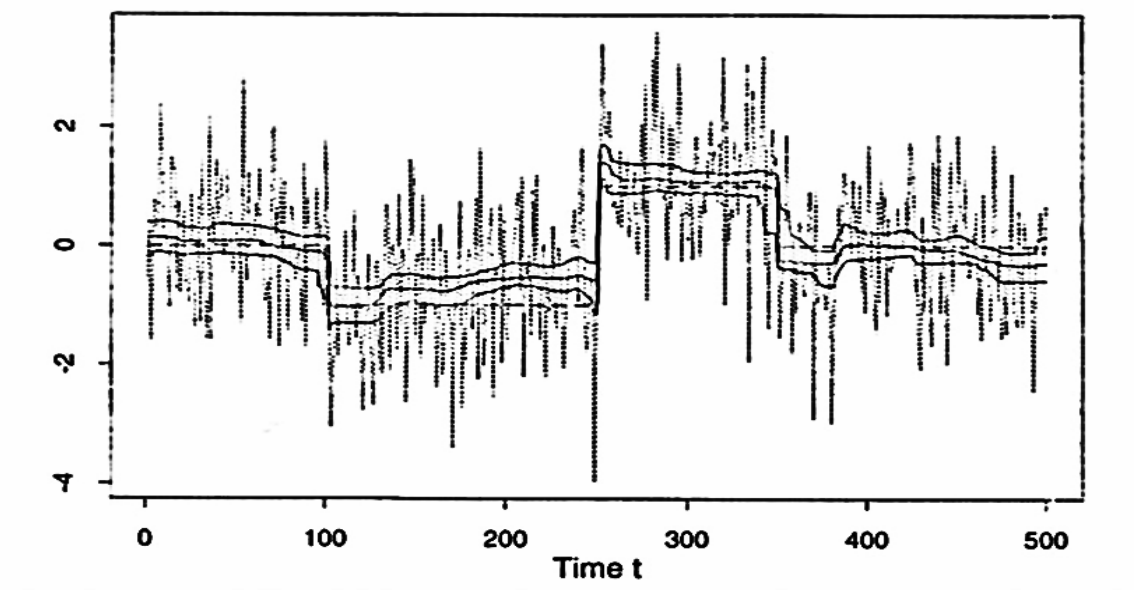


Figure 3.3 *Recursive Monte Carlo smoothing of a piecewise constant function. The true function (dashed), the observations (dotted) and the 10%-, 50%- and 90%-quantiles (solid) of the Monte Carlo smoother with* $N = 1000$ *are shown. From Hürzeler and Künsch (1998).*

Parameter estimation

Let state transitions a_t depend on the vector of parameters τ and observation densities - η;

Let $\theta = \{\tau, \eta\}$;

Then the following methods are usually addressed for estimation of θ in HMM:

- Maximal likelihood method:
  - EM(expectation-maximization)-algorithm, fast convergence to the neighborhood of the optimal value, extremely slow performance within this region;
  - Newton's algorithm, slow convergence to the neighborhood of the optimal solution, good performance within it; Methaheuristics and/or Combinations of them can be applied.
- Bayesian algorithms (Gibbs, Metropolis, etc.) can be applied after having set the priors for τ and η. Full conditionals of posteriors are then as follows:

$$p(\tau \mid x_0^T, y_1^T, \eta) = p(\tau \mid x_0^T) \propto p(\tau) a_0(x_0 \mid \tau) \prod_{t=1}^{T} a_t(x_{t-1}, x_t \mid \tau) \text{ and } p(\eta \mid x_0^T, y_1^T, \tau) = p(\eta \mid x_0^T, y_1^T) \propto p(\eta) \prod_{t=1}^{T} b_t(x_t, y_t \mid \eta).$$

Parameter estimation ML

$$P_{\mathcal{M}|X}(\mathcal{M}|\boldsymbol{X}) = \frac{1}{f_X(\boldsymbol{X})} f_{X|\mathcal{M}}(\boldsymbol{X}|\mathcal{M})\, P_{\mathcal{M}}(\mathcal{M})$$ - Posterior joint probability of interest

$$f_{X|\mathcal{M}}(\boldsymbol{X}|\mathcal{M}) = \sum_{s} f_{X|S,\mathcal{M}}(\boldsymbol{X}|\boldsymbol{s},\mathcal{M})\, P_{s|\mathcal{M}}(\boldsymbol{s}|\mathcal{M})$$ - Likelihood function to be maximized

$$f_{X|S,\mathcal{M}}(\boldsymbol{X}|\boldsymbol{s},\mathcal{M}) = f_{X|S}(\boldsymbol{x}(0)|s(0))\, f_{X|S}(\boldsymbol{x}(1)|s(1)) \cdots f_{X|S}(\boldsymbol{x}(T-1)|s(T-1))$$ -Pdf of signal given states sequence

$$P_{S|\mathcal{M}}(\boldsymbol{s}|\mathcal{M}) = \pi_{s(0)} a_{s(0)s(1)} a_{s(1)s(2)} \cdots a_{s(T-2)s(T-1)}$$ - Probability of the states' sequence

Where $\mathcal{M}$ - is a model, *s* – is a sequence of states and *X* – is a set of the corresponding Observations.

Parameter estimation ML

Thus we are to maximize the likelihood of the model to estimate the parameters of interest:

$$f_{X|\mathcal{M}}(\boldsymbol{X}|\mathcal{M}) = \sum_{\boldsymbol{s}} f_{X|S,\mathcal{M}}(X|s,\mathcal{M})\, P_{s|\mathcal{M}}(s|\mathcal{M})$$

$$= \sum_{\boldsymbol{s}} \pi_{s(0)} f_{X|S}\left(\boldsymbol{x}(0)\,|s(0)\right) a_{s(0)s(1)} f_{X|S}\left(\boldsymbol{x}(1)\,|s(1)\right)$$

$$\cdots a_{s(T-2)s(T-1)} f_{X|S}\left(x(T-1)\,|s(T-1)\right)$$

Note that in general case the difficulty of these calculations is exponential $O(N^T)$, however in the efficient implementation the structure of the sequence of states yields into getting the polynomial difficulty algorithm.

Spatial-Temporal Model

- Consider positions spatial-temporal positions *t* and their neighborhoods *L(t)* on $\{1 \dots T_1\} \times \{1 \dots T_2\}$, with the following markov property $P(x(t)|x(u), u \neq t) = P(x(t)|x(u), u \in L(t))$
- Then given the prior for *x*: $P(x_C) = \frac{1}{Z}\prod_{c \in C} e^{-\Phi_c(x_c)}$ for any class *C* of non-empty complete subsets of *L*
- The posterior for the state looks as: $p(x_t \mid x_s; s \neq t) \propto p(x) \propto \prod_{C \in \mathcal{C}, C \ni t} \exp(-\Phi_C(x_C))$
- The joint distribution of the observation within given neighborhood: $p(y \mid x) = \prod_{t \in L} b_t(x_t, y_t).$
- The joint density for the neighborhood: $p(x, y) = p(x)p(y \mid x) = \frac{1}{Z} \prod_C \exp(-\Phi_C(x_C)) \prod_{t \in L} b_t(x_t, y_t)$

Spatial-Temporal Model. Issues

1. Marginal probabilities $p(x_t)$ or $p(x_V)$ for $V \subset L$ and conditional probabilities $p(x_t \mid y)$ or $p(x_V \mid y)$ cannot be computed in closed form. The number of operations is exponential in the size of L and no effective reduction is known. This makes the computations prohibitive in almost all cases of interest.
2. One cannot simulate from $p(x)$ or $p(x \mid y)$ in a fixed number of steps, but has to use MCMC methods.
3. The X_t's may exhibit long-range dependence which means that the effect of how the field is defined at the boundary does not disappear even in the limit if L increases in all directions.
4. The joint distribution is not equal to the product of the conditional distributions for each x_t given the rest

$$\prod_{t \in L} p(x_t \mid x_s, s \sim t) \neq p(x).$$

Spatial-Temporal Model. Issues

5. The conditional distributions $p(x_t \mid x_s; s \sim t)$ cannot be chosen arbitrarily. In order to obtain a valid model, one has to use the Gibbs representation. Only the potential can be chosen arbitrarily.
6. If the potential contains unknown parameters τ, then the normalization Z also depends on τ, but $Z(\tau)$ cannot be evaluated in closed form. This makes maximum likelihood difficult even in the fully observed case when the X_t's are available.
7. Fully Bayesian inference on (x, τ, η) (where τ and η are the unknown parameters in the potential and b_t respectively) is difficult because

$$p(\tau \mid x, y, \eta) = p(\tau \mid x) \propto p(x \mid \tau)p(\tau)$$

also needs the normalization $Z(\tau)$.

8. Computation of the posterior mode (MAP)

$$\arg\max_x p(x \mid y) = \arg\max_x (-\sum_C \Phi_C(x_C) + \sum_t \log(b_t(x_t, y_t)))$$

is made difficult because there are typically many local maxima.

Spatial-Temporal Model. Treatments

1. Use pseudo likelihood: $\prod_{t \in L} p(x_t \mid x_s, s \sim t; \tau)$
2. Use mean field approximation for the normalizing constant:

$$Z \approx \sum_x \exp(-\sum_t \sum_{C \ni t} \Phi_C(x_t \prod_{s \in C, s \neq t} \mathbb{E}[X_s]))$$

$$= \prod_t \sum_{x_t} \exp(\sum_{C \ni t} \Phi_C(x_t \prod_{s \in C, s \neq t} \mathbb{E}[X_s])).$$

3. Compute iterated conditional mode: $\hat{x}_t = \arg\max_{x_t} p(x_t \mid \hat{x}_s, s \sim t; y)$
4. Apply EM to pseudo likelihood: $\arg\max_{\tau, \eta} \sum_t \log p(y_t \mid \hat{x}_s, s \sim t; \tau, \eta)$
5. Use methaheuristics to compute the posterior mode and/or maximize pseudo likelihood
6. Estimate real likelihood by means of MCMC.
7. Etc.

References

1. Kunsch, H.R., *State Space and Hidden Markov Models*. ETH-Zurich, Zurich;

2. Vaseghi, S. V., *Hidden Markov Models.* Wiley & Sons Ltd.

The End