Hidden Markov Modelling

M. Vidyasagar
Executive Vice President
Tata Consultancy Services
Hyderabad, India
sagar@atc.tcs.com

Forthcoming book: *(Hidden) Markov Processes: Theory and Applications to Biology*, to be published by Princeton University Press.

Outline of Talks:

We focus on a discrete-time stationary stochastic process $\{\mathtt{Y}_t\}$ taking values in a finite alphabet $\mathcal{M} := \{1, \ldots, m\}$.

We study various aspects of modelling this stochastic process using a hidden Markov model (HMM); this term is defined later.

Part I: The Complete Realization Problem

Quick Review of Markov Chains

Markov Chains:

A stochastic process $\{\mathtt{X}_t\}_{t\geq 0}$ is said to be **Markov process** if

$$\Pr\{\mathtt{X}_{t+1}|\mathtt{X}_i, i \leq t\} = \Pr\{\mathtt{X}_{t+1}|\mathtt{X}_t\},\ \forall t.$$

In words: All the information about the past is contained in the most recent measurement $\mathtt{X}_t$. We can do just as good a job of predicting $\mathtt{X}_{t+1}$ using only the most recent measurement $\mathtt{X}_t$ as we can using the entire past history.

A Markov process is said to be **stationary** if $\Pr\{\mathtt{X}_{t+1}|\mathtt{X}_t\}$ is independent of t.

A Markov chain $\{\mathtt{X}_t\}$ assuming values in a finite set $\mathcal{N} = \{1, \ldots, n\}$ can be described unambiguously by just two entities: the state transition matrix, and the initial distribution. Define

$$a_{ij} = \mathsf{Pr}\{\mathtt{X}_{t+1} = j | \mathtt{X}_t = i\}, A = [a_{ij}] \in [0, 1]^{n \times n}.$$

Thus a_{ij} is the probability that the *next* state is j, given that the *current* state is j, and A is the state transition matrix. Define

$$\pi_i = \mathsf{Pr}\{\mathtt{X}_0 = i\}.$$

If the initial distribution π satisfies $\pi A = \pi$, then it is called a **stationary** or an **invariant** distribution. In this case the process $\{\mathtt{X}_t\}$ is itself a stationary stochastic process.

Unfortunate dual use of words: A stationary Markov chain can produce a nonstationary process!

Corresponding to every A there is at least one stationary distribution.

Maximum Likelihood Estimation of Markov Chains:

Suppose $\{\mathtt{X}_t\}$ is a stationary Markov chain assuming values in a finite set $\mathcal{N} = \{1, \ldots, n\}$, but we do not know A. Instead we have a finite observation $x_t, t = 1, \ldots, l$ of the Markov chain.

Then the maximum likelihood estimate of A is given as follows: Define

$$\eta_i := \sum_{t=1}^{l-1} I_{\{x_t = i\}}, \nu_{ij} := \sum_{t=1}^{l-1} I_{\{x_t x_{t+1} = ij\}}.$$

Thus η_i is the number of times the state i occurs in the first $l-1$ observations, and ν_{ij} is the number of times the pair ij occurs in the sample path.

Then

$$\widehat{a}_{ij} := \frac{\nu_{ij}}{\eta_i}$$

is the maximum likelihood estimate of A.

Definition of a Hidden Markov Model

Historical Starting Point: Functions of a Markov Chain:

Blackwell and Koopmans (1957), Gilbert (1959): Given a stationary Markov process $\{\mathtt{X}_t\}$ over $\mathcal{N} := \{1, \ldots, n\}$, suppose $f : \mathcal{N} \to \mathcal{M}$ is some **deterministic** function. Then $\{f(\mathtt{X}_t)\}$ is stationary but not necessarily Markov.

Question: Given a stationary process $\{\mathtt{Y}_t\}$ over $\mathcal{M}$, when can it be realized as a function of a Markov chain over a finite alphabet $\mathcal{N}$, and if so, what is the smallest value of n?

Gilbert gave a simple *necessary* condition in terms of an infinite matrix associated with the stochastic process having finite rank, and conjectured that it is also a sufficient condition. However, this is now known to be untrue.

Hidden Markov Models – Definition Used Here:

The process $\{\mathtt{Y}_t\}$ is said to have a **hidden Markov model of the joint Markov process type** if there exists another process $\{\mathtt{X}_t\}$ assuming values over a finite set $\mathcal{N} := \{1, \ldots, n\}$ such that:

- The joint process $\{(\mathtt{X}_t, \mathtt{Y}_t)\}$ is Markov, and
- For each $i, j \in \mathcal{N}, u, v \in \mathcal{M}$, we have that

$$\mathsf{Pr}\{(\mathtt{X}_{t+1}, \mathtt{Y}_{t+1}) = (j, u) | (\mathtt{X}_t, \mathtt{Y}_t) = (i, v)\} = \mathsf{Pr}\{(\mathtt{X}_{t+1}, \mathtt{Y}_{t+1}) = (j, u) | \mathtt{X}_t = i\}.$$

These assumptions guarantee that (i) $\{\mathtt{X}_t\}$ is a Markov process by itself, and that (ii) $\mathtt{Y}_{t+1}$ is a random function of $\mathtt{X}_t$ (though not necessarily of $\mathtt{X}_{t+1}$).

Note: *Every* stationary stochastic process can be expressed as a function of a Markov chain with a *countable* state space. Hence the challenge is to model using a *finite* state space.

Smaller State Space: Both definitions are equivalent: If a process $\{Y_t\}$ is a function of a Markov process, then it has a 'joint Markov process' model.

However, the 'joint Markov process' definition requires fewer states. So we use the joint Markov process definition.

Motivation for Using HMMs: If a process is not Markov but has a HMM, then it still has a 'finite' description, even though it may have 'infinite memory.'

"Sum of Products Formula" for Hidden Markov Models

Some Notation:

Suppose the joint Markov model is stationary, and define Define

$$m_{ij}^{(u)} := \mathsf{Pr}\{\mathtt{X}_{t+1} = j \& \mathtt{Y}_{t+1} = u | \mathtt{X}_t = i\},$$

$$M^{(u)} = [m_{ij}^{(u)}] \in [0,1]^{n \times n}, \textbf{Important!}$$

Define

$$a_{ij} := \sum_{u \in \mathcal{M}} m_{ij}^{(u)} = \mathsf{Pr}\{\mathtt{X}_{t+1} = j | \mathtt{X}_t = i\},\ A = [a_{ij}] \in [0,1]^{n \times n}.$$

Thus A is the state transition matrix of the Markov process $\{\mathtt{X}_t\}$. Also let π denote the initial (stationary) distribution of $\mathtt{X}_0$.

The Sum of Products Formula:

For each string $u_1 \dots u_l =: \mathbf{u} \in \mathcal{M}^l$, define the frequency $f_{\mathbf{u}}$ of occurence of that string as

$$f_{\mathbf{u}} := \mathsf{Pr}\{\mathsf{Y}_{t-l+1} \dots \mathsf{Y}_t = u_1 \dots u_l\}.$$

We can compute $f_{\mathbf{u}}$ by summing over all possible state sequences:

$$f_{\mathbf{u}} = \sum_{i_0=1}^{n} \sum_{i_1=1}^{n} \cdots \sum_{i_l=1}^{n} \pi_{i_0} m_{i_0 i_1}^{(u_0)} m_{i_1 i_2}^{(u_2)} \cdots m_{i_{l-1} i_l}^{(u_l)}.$$

This can be expressed compactly as follows:

$$f_{\mathbf{u}} = \pi M^{(u_1)} \cdots M^{(u_l)} \mathbf{e}_n = \pi M^{(\mathbf{u})} \mathbf{e}_n,$$

where $\mathbf{e}_n = [1 \;\; 1 \dots 1]^t$ is a column vector of n ones.

The Questions of Interest:

Suppose the complete statistics of $\{\mathrm{Y}_t\}$ are known. When is it possible to construct a hidden Markov model (HMM) for this process?

How can one construct a 'partial realization' for the process, that faithfully reproduces the statistics of the process only up to some finite order?

Can one approximate one Markov process, or a HMM, by another, simpler process, and if so, quantify the error in approximation?

We begin with the first question.

A General Necessary Condition

The 'Hankel' Matrix of a Process

Recall: $f_{\mathbf{u}}$ denotes the frequency of occurence of the string $\mathbf{u}$:

$$f_{\mathbf{u}} := \mathsf{Pr}\{\mathtt{Y}_{t-l+1} \dots \mathtt{Y}_t = u_1 \dots u_l\}.$$

Define 'flo' = first lexical order, 'llo' = last lexical order on $\mathcal{M}^l$.

Example: Let $\mathcal{M} = \{1, 2\}, l = 3$.

$$\mathcal{M}^3 \text{ in llo } = \{111, 112, 121, 122, 211, 212, 221, 222\}.$$

$$\mathcal{M}^3 \text{ in flo } = \{111, 211, 121, 221, 112, 212, 122, 222\}.$$

Define

$$F_{k,l} := [f_{\mathbf{uv}}, \mathbf{u} \in \mathcal{M}^k \text{ in flo, } \mathbf{v} \in \mathcal{M}^l \text{ in llo}] \in [0, 1]^{m^k \times m^l}.$$

Every entry in $F_{k,l}$ is the frequency of a $(k + l)$-tuple.

Example: Suppose $m = 2$. Then

$$F_{2,1} = \begin{bmatrix} f_{111} & f_{112} \\ f_{211} & f_{212} \\ f_{121} & f_{122} \\ f_{221} & f_{222} \end{bmatrix}, \; F_{1,2} = \begin{bmatrix} f_{111} & f_{112} & f_{121} & f_{122} \\ f_{211} & f_{212} & f_{221} & f_{222} \end{bmatrix}.$$

$$H_{k,l} := \begin{bmatrix} F_{0,0} & F_{0,1} & \cdots & F_{0,l} \\ F_{1,0} & F_{1,1} & \cdots & F_{1,l} \\ \vdots & \vdots & \vdots & \vdots \\ F_{k,0} & F_{k,1} & \cdots & F_{k,l} \end{bmatrix} \in [0,1]^{(1+m+\ldots+m^k)\times(1+m+\ldots+m^l)}.$$

Note: Matrices on backward diagonals all contain the same entries, but arranged differently.

Define H to be the 'infinite' matrix where we do not restrict the size of k, l. It is called the 'Hankel' matrix.

Theorem (Gilbert): If $\{\mathtt{Y}_t\}$ has a HMM, then Rank$(H) < \infty$.

Proof: Recall the 'sum of products formula':

$$f_{\mathbf{u}} = \pi M^{(u_1)} \dots M^{(u_l)} \mathbf{e}_n.$$

Define

$$U_l := [M^{(\mathbf{u})}, \mathbf{u} \in \mathcal{M}^l \text{ in flo}] \in [0,1]^{nm^l \times n},$$

$$V_l := [M^{(\mathbf{u})}, \mathbf{u} \in \mathcal{M}^l \text{ in llo}] \in [0,1]^{n \times nm^l}.$$

Then $f_{\mathbf{u}} = \pi M^{(u_1)} \dots M^{(u_l)} \mathbf{e}_n$ implies that

$$F_{k,l} = [f_{\mathbf{uv}}, \mathbf{u} \in \mathcal{M}^k \text{ in flo}, \ \mathbf{v} \in \mathcal{M}^l \text{ in llo}] = \pi U_k V_l \mathbf{e}_n,$$

$$H = \begin{bmatrix} \pi \\ \pi U_1 \\ \pi U_2 \\ \vdots \end{bmatrix} [\mathbf{e}_n | V_1 \mathbf{e}_n | V_2 \mathbf{e}_n | \dots]$$

So Rank$(H) \leq n$.

Non-Sufficiency of Finite Rank Condition:

If $\{\mathtt{Y}_t\}$ has a HMM of size n, then Rank$(H) \leq n$.

This theorem is essentially due to Gilbert (1959) though his notation was different.

He also conjectured that Rank$(H) < \infty$ was sufficient for the process to have a HMM.

Dharmadhikari (1965), Fox and Rubin (1965): The converse is *not* true: There exist processes where Rank$(H) < \infty$ but don't have a HMM.

So what can be said about the case of 'finite Hankel rank' processes?

The situation has remained pretty static for nearly forty years.

Quasi-Realizations

Consequences of the Sum of Products Formula

Recall that

$$f_{\mathbf{u}} = \pi M^{(u_1)} \dots M^{(u_l)} \mathbf{e}_n,$$

where $\mathbf{e}_n$ is the column vector of all one's. Also, if we define

$$A = \sum_{u \in \mathcal{M}} M^{(u)},$$

then A is the transition matrix of the underlying Markov process, which is stochastic, and π is a stationary distribution of A:

$$\pi \left[\sum_{u \in \mathcal{M}} M^{(u)} \right] = \pi, \left[\sum_{u \in \mathcal{M}} M^{(u)} \right] \mathbf{e}_n = \mathbf{e}_n.$$

Quasi-Realizations

A triplet $\{\theta, D^{(u)}, u \in \mathcal{M}, \phi\}$ is a **quasi-realization** of $\{\mathtt{Y}_t\}$ if

$$f_{\mathbf{u}} = \theta D^{(u_1)} \dots D^{(u_l)} \phi \; \forall \mathbf{u} \in \mathcal{M}^*,$$

and in addition

$$\theta \left[\sum_{u \in \mathcal{M}} D^{(u)} \right] = \theta, \left[\sum_{u \in \mathcal{M}} D^{(u)} \right] \phi = \phi.$$

No requirement that the vectors and matrices should be nonnegative!

The quasi-realization if **regular** if the dimension of $\theta, D^{(u)}, \phi$ is Rank(H).

The Integer k:

Suppose the process $\{\mathtt{Y}_t\}$ is 'finite Hankel rank,' i.e., has the property that Rank$(H) < \infty$. What conclusions can we draw from this property?

Lemma: Rank$(H_{k,l})$ = Rank$(F_{k,l})$.

Note that, for every $\mathbf{u} \in \mathcal{M}^*$, we have

$$f_{\mathbf{u}} = \sum_{i \in \mathcal{M}} f_{i\mathbf{u}} = \sum_{j \in \mathcal{M}} f_{\mathbf{u}j}.$$

So

$$\text{Rank}\left(\begin{bmatrix} F_{0,0} & F_{0,1} & \cdots & F_{0,l} \\ F_{1,0} & F_{1,1} & \cdots & F_{1,l} \\ \vdots & \vdots & \vdots & \vdots \\ F_{k,0} & F_{k,1} & \cdots & F_{k,l} \end{bmatrix}\right) = \text{Rank}\left(\begin{bmatrix} F_{0,l} \\ F_{1,l} \\ \vdots \\ F_{k,l} \end{bmatrix}\right) = \text{Rank}(F_{k,l}).$$

Lemma: If $\mathrm{Rank}(H) < \infty$, then there exists a smallest k such that

$$\mathrm{Rank}(F_{k,k}) = \mathrm{Rank}(H_{k,k}) = \mathrm{Rank}(H_{k+l,k+s})\ \forall l, s > 0.$$

For every k, we have either $\mathrm{Rank}(H_{k,k}) < \mathrm{Rank}(H_{k+1,k+1}) \leq \mathrm{Rank}(H)$, or else $\mathrm{Rank}(H_{k,k}) = \mathrm{Rank}(H_{k+1,k+1})$.

If $\mathrm{Rank}(H) < \infty$ the latter can happen only finitely many times.

Hereafter the symbol k denotes the *unique* integer referred to in the above lemma.

In particular, we have that

$$\text{Rank}(F_{k,k}) = \text{Rank}([F_{k,k} \;\; F_{k,k+1}]).$$

So there exists a matrix $C \in \mathbb{R}^{m^k \times m^{k+1}}$ such that

$$F_{k,k+1} = F_{k,k}C.$$

Partition C as

$$C = [C^{(1)} \dots C^{(m)}], C^{(u)} \in \mathbb{R}^{m^k \times m^k}.$$

Then some simple arguments show that

$$f_{\mathbf{u}} = F_{0,k}C^{(u_1)} \dots C^{(u_l)}\mathbf{e}_{m^k}.$$

Hence the triplet $(F_{0,k}, C^{(u)}, \mathbf{e}_{m^k})$ is a quasi-realization.

Actually we can say much more.

Let $r := \text{Rank}(H) = \text{Rank}(F_{k,k})$. Choose subsets I, J of cardinality r such that the $r \times r$ matrix $[f_{\mathbf{vw}}, \mathbf{v} \in I, \mathbf{w} \in J]$ has rank r. Note that the above is a submatrix of $F_{k,k}$.

Choose matrices $C^{(1)}, \ldots, C^{(m)}$ such that

$$[f_{\mathbf{v}u\mathbf{w}}, \mathbf{v} \in I, \mathbf{w} \in J] = [f_{\mathbf{vw}}, \mathbf{v} \in I, \mathbf{w} \in J]C^{(u)}.$$

Then the r-dimensional triple $(f_{\mathbf{v}}, \mathbf{v} \in I, C^{(u)}, \mathbf{e}_r)$ is a *regular* quasi-realization of the process.

Yet still more can be said.

Theorem: A finite Hankel rank process *always* has a *regular* quasi-realization. Moreover, if $\{\theta_1, D_1^{(u)}, u \in \mathcal{M}, \phi_1\}$, $\{\theta_2, D_2^{(u)}, u \in \mathcal{M}, \phi_2\}$ are two regular quasi-realizations of the same process, then there exists a nonsingular matrix T such that

$$\theta_2 = \theta_1 T^{-1}, D_2^{(u)} = T D_1^{(u)} T^{-1} \; \forall u \in \mathcal{M}, \phi_2 = T\phi_1.$$

Quasi-realizations are not 'true' realizations, but they are 'easy' to construct.

Fun Fact No. 1: Given a quasi-realization $\{\theta, D^{(u)}, u \in \mathcal{M}, \phi\}$, the problem of determining whether there exists a nonsingular matrix T such that $\theta T^{-1}, TD^{(u)}T^{-1}, T\phi$ are all positive can be solved in polynomial time.

Thus it is possible to determine, in an efficient fashion, whether a given quasi-realization is a 'true' realization in disguise.

However, there exist $\{\theta, D^{(u)}, u \in \mathcal{M}, \phi\}$ such that

$$\theta D^{(\mathbf{u})}\phi \geq 0 \; \forall \mathbf{u} \in \mathcal{M}^*,$$

and yet no T exists such that $\theta T^{-1}, TD^{(u)}T^{-1}, T\phi$ are all positive.

Fun Fact No. 2: Given a row vector θ, square matrices $D^{(u)}, u \in \mathcal{M}$ and a column vector ϕ, the problem of determining whether $\theta D^{(\mathbf{u})}\phi \geq 0$ for all strings $\mathbf{u} \in \mathcal{M}^*$ is *undecidable!*

'Nearly Necessary and Sufficient' Conditions for the Existence of Hidden Markov Models

Need to Introduce Additional Assumptions:

For an *arbitrary* stochastic process, with no additional assumptions, further progress appears difficult. So we add the assumption that the process is 'mixing' in some sense.

We begin with α-mixing (introduced here, but not really used until later).

Alpha-Mixing – A Form of Asymptotic Long-Term Independence

The formal definition of α-mixing is now given (for those who are interested).

Given a stochastic process $\{\mathtt{Y}_t\}$, define $\Sigma^0_{-\infty}$ to be the σ-algebra generated by $\mathtt{Y}_t, t \leq 0$, and let Σ^∞_l denote the σ-algebra generated by $\mathtt{Y}_t, t \geq l$. Then

$$\alpha(l) := \sup_{A \in \Sigma^0_{-\infty}, B \in \Sigma^\infty_l} |P(A \cap B) - P(A) \cdot P(B)|.$$

Interpretation: A is an event that depends only on the 'past' variables $\mathtt{Y}_t, t \leq 0$, B is an event that depends only on the 'future' variables $\mathtt{Y}_t, t \geq l$.

The difference $|P(A \cap B) - P(A) \cdot P(B)|$ quantifies the extent to which A and B are independent.

A stochastic process $\{\mathtt{Y}_t\}$ is said to be **α-mixing** if $\alpha(l) \to 0$ as $l \to \infty$.

Alpha-Mixing of Markov Chains:

A Markov process $\{\mathtt{X}_t\}$ over a finite state space is α-mixing if and only if the state transition matrix A is irreducible and aperiodic.

In other words, A has a single, simple eigenvalue at $\lambda = 1$ and all other eigenvalues have magnitude strictly less than one.

Let λ denote the **second largest** eigenvalue of A. Then $\alpha(l) = O(|\lambda|^l)$.

Irreducibility is Weaker than Alpha-Mixing!

Example:

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \pi = [0.5 \;\; 0.5]$$

The Markov 'chain' toggles between two states 1 and 2.

Let the event A be $\{\mathtt{X}_0 = 1\}$ and let the event B be $\{\mathtt{X}_l = 2\}$. Then $P(A \cap B)$ is 0 if l is even and 1 if l is odd. But $P(A) = P(B) = 0.5$. So

$$P(A \cap B) \not\to P(A) \cdot P(B) \text{ as } l \to \infty.$$

Alpha-Mixing Properties of HMMs:

Suppose a process $\{\mathtt{Y}_t\}$ has a HMM. If the underlying Markov process $\{\mathtt{X}_t\}$ is α-mixing, so is the corresponding output process $\{\mathtt{Y}_t\}$, because $\mathtt{Y}_t$ is a 'random' function of $\mathtt{X}_{t-1}$.

The converse is not true – it is possible for the output to be α-mixing even if the state process is not, provided certain 'consistency conditions' (Anderson 1999, MCSS) are satisfied – conditions are 'fragile.'

Let p denote the period of a nonprimitive, irreducible Markov chain with n states. By renumbering states, we can ensure that

$$A = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & A_1 \\ A_p & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A_{p-1} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & A_2 & \mathbf{0} \end{bmatrix}.$$

where all blocks are $(n/p) \times (n/p)$.

The matrices $M^{(u)}, u \in \mathcal{M}$ inherit the same structure. So the symbol $M_i^{(u)}$ is defined analogously.

For a string $\mathbf{u} \in \mathcal{M}^*$, define

$$M_i^{(\mathbf{u})} = M_i^{(u_1)} M_{i+1}^{(u_2)} \dots M_{i+l-1}^{(u_l)},$$

where the indices are taken mod p.

Theorem: (Brian Anderson (1999)) The output process is α-mixing if and only if, for every string $\mathbf{u} \in \mathcal{M}^*$, we have

$$\begin{aligned} \pi_1^t M_1^{(\mathbf{u})} \mathbf{e}_{(n/p)} &= \pi_2^t M_p^{(\mathbf{u})} \mathbf{e}_{(n/p)} = \pi_3^t M_{p-1}^{(\mathbf{u})} \mathbf{e}_{(n/p)} = \ldots = \pi_p^t M_2^{(\mathbf{u})} \mathbf{e}_{(n/p)} \\ &= \frac{1}{p} \pi M^{(\mathbf{u})} \mathbf{e}_n. \end{aligned}$$

(Baby) Example: Suppose $m_{ij}^{(u)} = a_{ij} b_{ju}$, where

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}, \pi = [1/3 \;\; 1/3 \;\; 1/3]$$

Then the above consistency condition is satisfied if and only if $b_{j1} = b_{j2} = 0.5$ for all j.

Open Question: If the above consistency condition is satisfied, can be the HMM be replaced by another where the underlying Markov chain is primitive?

Mixing Property Assumed Here:

Here we use a weaker notion than α-mixing, called just 'mixing.'

Definition: The process is **mixing** if

$$\sum_{\mathbf{w} \in \mathcal{M}^l} f_{\mathbf{uwv}} \to f_{\mathbf{u}} f_{\mathbf{v}} \text{ as } l \to \infty, \ \forall \mathbf{u}, \mathbf{v} \in \mathcal{M}^k.$$

Interpretation: A kind of long-term asymptotic independence. The probability of a string beginning with $\mathbf{u} \in \mathcal{M}^k$ and ending with $\mathbf{v} \in \mathcal{M}^k$ asymptotically approaches the product of the individual probabilities as the separation between $\mathbf{u}$ and $\mathbf{v}$ becomes larger.

Comment: This is weaker than α-mixing because we are looking at only strings $\mathbf{u}, \mathbf{v}$ of **fixed** length k.

Definition: The process $\{\mathtt{Y}_t\}$ is **ultra-mixing** if $\exists\{\delta_l\} \downarrow 0$ such that

$$\left|\frac{f_{\mathbf{iu}}}{f_{\mathbf{u}}} - \frac{f_{\mathbf{iuv}}}{f_{\mathbf{uv}}}\right| \leq \delta_l, \ \forall \mathbf{i} \in \mathcal{M}^k, \mathbf{u} \in \mathcal{M}^l, \mathbf{v} \in \mathcal{M}^*.$$

Interpretation: Smooth dependence of conditional probabilities on the length of the string on which conditioning is done.

A Sidelight: Ultra-mixing is equivalent to the process being 'random Markov,' that is, being a Markov process with a memory length that is random (and independent of the process).

Theorem: If a process has finite Hankel rank, is mixing, ultra-mixing, and a technical condition is satisfied, then it has a HMM.

Moreover, (as per Anderson's theorem), if the process is mixing, the underlying Markov chain is either primitive, or else it is irreducible and satisfies a 'consistency condition.'

The 'technical condition' has to do with the cluster points of the vector of conditional probabilities $[f_{v|\mathbf{u}}], v \in \mathcal{M}, \mathbf{u} \in \mathcal{M}^*]$ (messy to state).

The countable vector of conditional probabilities $[f_{v|\mathbf{u}}], v \in \mathcal{M}, \mathbf{u} \in \mathcal{M}^*]$ completely characterizes the stochastic process and is an equivalent description instead of specifying $f_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^*$. This countable vector belongs to ℓ_∞.

The technical condition holds for an open dense subset of ℓ_∞.

A Near Converse:

Theorem: Suppose a mixing process has a HMM. Suppose that another technical condition is satisfied. Then the process has finite Hankel rank, is mixing and is also ultra-mixing.

Moreover, by Anderson's theorem, the consistency condition is satisfied.

For two HMMs of the same order (same number of states), a natural topology is obtained by the metric

$$\sum_{u \in \mathcal{M}} \| M_1^{(u)} - M_2^{(u)} \|,$$

where $M_1^{(u)}$, $M_2^{(u)}$ are the matrices of the two HMMs. This 'technical condition' holds whenever all entries of the matrix $M^{(u)}$ are positive for each $u \in \mathcal{M}$. Hence this technical condition also holds for an open dense subset of HMMs.

So, modulo two technical conditions, if a process has finite Hankel rank and is 'mixing,' it has a HMM if and only if it is also ultra-mixing.

Moreover, both technical conditions hold for an 'open dense subset' of processes in an appropriate topology.

Part II: The Partial Realization Problem

Problem Statement

Motivation:

We are able to observe a stationary stochastic process $\{\mathtt{Y}_t\}$ assuming values in a finite alphabet $\mathcal{M} := \{1, \ldots, m\}$.

We wish to construct a hidden Markov model (defined later) for this process, on the basis of a *finite length* sample path.

This involves:

• Estimating the 'true' probabilities of k-tuples $(y_1, \ldots, y_k)$ on the basis of observed frequencies, and
• Constructing a stochastic model on the basis of the statistics (true probabilities).

Process is *not* assumed to be i.i.d., so problems are non-trivial.

Problem Statements:

'Exact' Partial Realization Problem:

Suppose we are given a stationary stochastic process $\{\mathtt{Y}_t\}$ with known statistics, and an integer k. Construct a HMM that perfectly reproduces the frequencies $f_{\mathbf{u}}$ whenever $|\mathbf{u}| \leq k$.

'Inexact' Partial Realization Problem:

Suppose we have available only a finite length sample path of the process $\{\mathtt{Y}_t\}$, based on which we *estimate* the frequencies $\hat{f}_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^k$ for some integer k. What can be said about the accuracy and confidence of these estimates? Further, what can be said about the accuracy and confidence *of the parameters of the HMM*?

Exact Partial Realization Problem – Restatement:

We are given frequencies $f_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^k$ for some integer k. Find an integer n, and matrices $M^{(u)} \in [0,1]^{n\times n}$ such that $\sum_{u\in\mathcal{M}} M^{(u)} =: A$ is a stochastic matrix, and in addition, we have

$$f_{\mathbf{u}} = \pi M^{(u_1)} \dots M^{(u_k)} \mathbf{e}_n, \ \forall \mathbf{u} \in \mathcal{M}^k,$$

where π is a stationary distribution of A.

Here we are just trying to reproduce the known (or specified) frequencies $f_{\mathbf{u}}$ in terms of the current formalism.

But we wish to use the above formula to *extrapolate* and compute $f_{\mathbf{u}}$ for strings of length longer than k. In this case the computed frequencies must still 'make sense.'

Three Important Requirements

Consistency – An Important Requirement:

Suppose $\{\mathrm{Y}_t\}$ is a stochastic process over $\mathcal{M}$ and $f_{\mathbf{u}}$ denotes the frequency of the string $\mathbf{u}$.

Then the frequencies are *consistent*, that is,

$$f_{\mathbf{u}} = \sum_{v \in \mathcal{M}} f_{\mathbf{u}v} = \sum_{w \in \mathcal{M}} f_{w\mathbf{u}}.$$

So if we construct a 'partial' realization, the frequencies produced by the realization must also be consistent.

Reformulation of the Problem:

For now drop the requirement of nonnegativity. Choose a row vector θ, a column vector ϕ with n components each, and matrices $C^{(1)}, \ldots, C^{(m)}$, and define the 'pseudo-frequencies'

$$g_{\mathbf{u}} := \theta C^{(u_1)} \ldots C^{(u_l)} \phi, \ \forall \mathbf{u} \in \mathcal{M}^*$$

In order for these matrices to be a solution to the (exact) partial realization problem, the pseudo-frequencies $g_{\mathbf{u}}$ must satisfy three properties:

- **Consistency:** $g_{\mathbf{u}}$ must be consistent in the sense that

$$g_{\mathbf{u}} = \sum_{v \in \mathcal{M}} g_{\mathbf{u}v} = \sum_{w \in \mathcal{M}} g_{w\mathbf{u}}.$$

- **Reproduction:** $g_{\mathbf{u}} = f_{\mathbf{u}} \ \forall \mathbf{u} \in \mathcal{M}^k$.
- **Nonnegativity:** $g_{\mathbf{u}} \geq 0 \ \forall \mathbf{u} \in \mathcal{M}^*$.

Each requirement is satisfied by a different set of conditions.

Consistency:

Theorem: Suppose the vectors θ and ϕ satisfy the conditions

$$\theta \left[\sum_{u \in \mathcal{M}} C^{(u)} \right] = \theta, \left[\sum_{u \in \mathcal{M}} C^{(u)} \right] \phi = \phi.$$

Then the set of pseudo-frequencies $g_{\mathbf{u}}$ defined by

$$g_{\mathbf{u}} := \theta C^{(u_1)} \dots C^{(u_l)} \phi, \ \forall \mathbf{u} \in \mathcal{M}^*$$

is consistent.

This explains (I hope!) why we imposed these eigenvector conditions when we defined a quasi-realization.

Proof of Theorem: Let $\mathbf{u}, v, w$ be arbitrary. Write

$$C^{(\mathbf{u})} := C^{(u_1)} \dots C^{(u_l)}.$$

Then

$$\begin{aligned}
\sum_{v \in \mathcal{M}} g_{v\mathbf{u}} &= \theta \sum_{v \in \mathcal{M}} C^{(v)} C^{(\mathbf{u})} \phi \\
&= \theta \left[\sum_{v \in \mathcal{M}} C^{(v)} \right] C^{(\mathbf{u})} \phi \\
&= \theta C^{(\mathbf{u})} \phi = g_{\mathbf{u}}.
\end{aligned}$$

Similarly

$$\sum_{w \in \mathcal{M}} g_{\mathbf{u}w} = g_{\mathbf{u}}.$$

If the row and column span of the frequencies of the process is 'sufficiently rich,' then the eigenvector conditions are also necessary.

Reproduction:

Recall something that was just flashed by earlier. Suppose the process $\{Y_t\}$ has finite Hankel rank, and define the integer k such that $\text{Rank}(H) = \text{Rank}(F_{k,k})$. Then in particular

$$\text{Rank}(F_{k,k}) = \text{Rank}([F_{k,k} \;\; F_{k,k+1}]).$$

So there exists a matrix $C \in \mathbb{R}^{m^k \times m^{k+1}}$ such that

$$F_{k,k+1} = F_{k,k}C.$$

Partition C as

$$C = [C^{(1)} \dots C^{(m)}], C^{(u)} \in \mathbb{R}^{m^k \times m^k}.$$

Then it can be shown that

$$f_{\mathbf{u}} = F_{0,k}C^{(u_1)} \dots C^{(u_l)}\mathbf{e}_{m^k}.$$

Hence the triplet $(F_{0,k}, C^{(u)}, \mathbf{e}_{m^k})$ is a quasi-realization.

Define $J_l \in \{0,1\}^{m^l \times m^{l-1}}$ to be the 'block-diagonal' matrix with m^{l-1} blocks, where each block is $\mathbf{e}_m$. For example, if $m = 2$, then

$$J_3 = \left[\begin{array}{c|c|c|c} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{array}\right].$$

Then $F_{k,k} = F_{k,k+1} J_{k+1}$. This is just the 'consistency' of frequencies. But since $F_{k,k+1} = F_{k,k} C$, this implies that

$$F_{k,k} = F_{k,k} C J_{k+1}.$$

If $F_{k,k}$ is nonsingular, this implies that $C J_{k+1} = I$. We can use this 'insight' to solve the partial realization problem

Theorem: Given the integer k and a consistent set of frequencies $f_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^k$, choose *any* matrix $C \in \mathbb{R}^{m^{k-1} \times m^k}$ such that

$$F_{0,k-1} C = F_{0,k} \text{ and } C J_k = I_{m^{k-1}}.$$

Partition C as $[C^{(1)} \ldots C^{(m)}]$ where each $C^{(u)} \in \mathbb{R}^{m^{k-1} \times m^{k-1}}$. Then the triple $\theta = F_{0,k-1}, \phi = \mathbf{e}_{m^{k-1}}$ and $C^{(u)}, u \in \mathcal{M}$ defined above leads to a set of pseudo-frequencies that is *consistent* and *reproductive*.

Consistency follows from the eigenvector conditions. We started with $F_{0,k-1}C = F_{0,k}$. Because $F_{0,k}$ is a frequency vector, we have that

$$\sum_{v \in \mathcal{M}} f_{\mathbf{u}v} = f_{\mathbf{u}}, \ \forall \mathbf{u} \in \mathcal{M}^{k-1}.$$

Next,

$$\begin{aligned} \left[\sum_{w \in \mathcal{M}} C^{(w)} \right] \mathbf{e}_{m^{k-1}} &= C\mathbf{e}_{m^k} = CJ_k\mathbf{e}_{m^{k-1}} \\ &= \mathbf{e}_{m^{k-1}}. \end{aligned}$$

Hence

$$F_{0,k-1} \left[\sum_{u \in \mathcal{M}} C^{(u)} \right] = F_{0,k-1}, \left[\sum_{u \in \mathcal{M}} C^{(u)} \right] \mathbf{e}_{m^{k-1}} = \mathbf{e}_{m^{k-1}}.$$

Reproduction follows from $CJ_k = I_{m^{k-1}}$. We can show that, for every string $\mathbf{u} \in \mathcal{M}^{k-1}$, we have that the product $C^{(\mathbf{u})} = C^{(u_1)} \ldots C^{(u_{k-1})}\mathbf{e}_{m^{k-1}}$ equals a unit vector with a 1 in the row corresponding to $\mathbf{u}$, and zeros elsewhere.

Nonnegativity:

Thus far we have a way of achieving consistency and reproduction. So the only remaining question is: Is this set of pseudo-frequencies also *nonnegative?*

An easy way to ensure nonnegativity: Make sure that each $C^{(u)}$ is nonnegative.

Theorem: The only *nonnegative* solutions C to the equations

$$F_{0,k-1}C = F_{0,k} \text{ and } CJ_k = I_{m^{k-1}}$$

is

$$C = \text{BlockDiag}\{[\frac{f_{\mathbf{u}v}}{f_{\mathbf{u}}}, v \in \mathcal{M}], \mathbf{u} \in \mathcal{M}^{k-1}\}.$$

This solution leads to a model whereby the process $\{\mathtt{Y}_t\}$ is modelled as a $(k-1)$-step Markov process.

Exact Partial Realization Using a Multi-Step Markov Model:

State space: $\mathcal{M}^{k-1}$ = Set of all strings of length $k-1$.

A transition from $\mathbf{v} \in \mathcal{M}^{k-1}$ to $\mathbf{u} \in \mathcal{M}^{k-1}$ is possible if and only if the last $k-2$ symbols of $\mathbf{v}$ equal the first $k-2$ symbols of $\mathbf{u}$.

$$\mathbf{v} = v_1 v_2 \dots v_{k-1},\ \mathbf{u} = v_2 \dots v_{k-1} u_k.$$

In this case, define the transition probability

$$a_{\mathbf{vu}} := \frac{f_{\mathbf{v}u_k}}{f_{\mathbf{v}}}.$$

Output $\mathtt{Y}_t \in \mathcal{M}$ is the last symbol of $\mathtt{X}_t \in \mathcal{M}^{k-1}$.

This model exactly reproduces all k-tuple frequencies $f_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^k$.

Interpretation: This model ensures that the process $\{\mathtt{Y}_t\}$ is a $(k-1)$-step Markov process, i.e.,

$$\Pr\{\mathtt{Y}_t|\mathtt{Y}_{t-1},\ldots\} = \Pr\{\mathtt{Y}_t|\mathtt{Y}_{t-1},\ldots,\mathtt{Y}_{t-k+1}\}.$$

This particular partial realization is well-known (and seems to have no name).

But our theorem gives some justification by showing that it is *the only partial realization* satisfying a nonnegativity condition.

Clearly not all stochastic processes are multi-step (have finite memory). So further work is needed!

In particular, we must permit the matrices $C^{(u)}$ to contain negative elements, while ensuring that the product $F_{0,k}C^{(\mathbf{u})}\mathbf{e}^{m^{k-1}}$ is nonnegative for all strings $\mathbf{u}$ of arbitrary length.

I don't know how to do this!

Inexact Partial Realization Problem:

Suppose we are not given the frequencies $f_{\mathbf{u}}, |\mathbf{u}| \leq k$, but only a finite length observation $y_1 \dots y_l$. Using this observation we can estimate $f_{\mathbf{u}}$; call this estimate $\widehat{f}_{\mathbf{u}}$.

If it is assumed that the process under observation is α-mixing, then it is possible to quantify the *rate* at which the estimates $\widehat{f}_{\mathbf{u}}$ converge to their true values.

Since the parameters in the HMM are just ratios of the form $f_{\mathbf{v}u_k}/f_{\mathbf{v}}$, these estimates can in turn be translated into estimates on the parameters of the HMM.

Details are omitted as they formulas are messy.

Part III: Applications of the Kullback-Leibler Divergence Rate

Topics Studied in This Part

Topics Studied in This Part:

An alternate formula for the K-L divergence rate.

An easy derivation of the (known) formula for the K-L divergence rate between two Markov chains.

Approximating a given set of frequencies using a multi-step Markov chain.

Approximating a Markov chain with 'long' memory using another Markov chain of 'shorter' memory.

An *approximation* to the K-L divergence rate between two hidden Markov processes, where the underlying state spaces could have different sizes. (No exact formula is known.)

The Kullback-Leibler Divergence Rate Between Stochastic Processes

The K-L Divergence Between Probability Vectors

Given two probability vectors $\mathbf{p}, \mathbf{q}$ on a finite set $\mathcal{N}$ of size n, the Kullback-Leibler (K-L) divergence rate between them is defined by

$$H(\mathbf{p} \parallel \mathbf{q}) := \sum_{i=1}^{n} p_i \log \left(\frac{p_i}{q_i} \right).$$

Interpretation: Given a set of i.i.d. observations from the set $\mathcal{N}$, there two competing hypotheses: (i) the data is generated by $\mathbf{p}$, (ii) the data is generated by $\mathbf{q}$.

The quantity $H(\mathbf{p} \parallel \mathbf{q})$ is the per-observation expected value of the log-likelihood ratio of misclassification when the 'truth' is that the data is generated by the distribution $\mathbf{p}$.

What do we do if the successive observations are *not independent?*

The K-L Divergence *Rate* Between Probability *Laws*

Suppose $\tilde{P}, \tilde{Q}$ are probability *laws* of two stochastic processes on the set $\mathcal{N}^\infty$, where $\mathcal{N}$ is a finite set.

The K-L divergence rate is defined as the limit, when it exists,

$$R(\tilde{P} \parallel \tilde{Q}) := \lim_{l \to \infty} \frac{1}{l} H(\tilde{P}_l \parallel \tilde{Q}_l),$$

where $\tilde{P}_l, \tilde{Q}_l$ are the l-dimensional marginal distributions of $\tilde{P}, \tilde{Q}$ respectively on $\mathcal{N}^l$.

Interpretation: Suppose we observe a sample path $\{x_i, i \geq 1\}$ from the set $\mathcal{N}$. The two competing hypotheses are: (i) The underlying law is $\tilde{P}$, (ii) the underlying law is $\tilde{Q}$.

The quantity $R(\tilde{P} \parallel \tilde{Q})$ is the expected value of the log-likelihood ratio of choosing Hyp. (ii) when the 'truth' is Hyp. (i), divided by l.

An Alternate Formulation of the K-L Divergence Rate

High School algebra: Suppose $U = \{u_1, \ldots, u_m\}, V = \{v_1, \ldots, v_n\}$ are finite sets, and P, Q are probability distributions on $U \times V$. Then

$$H(P \parallel Q) = H(P_U \parallel Q_U) + \sum_{i=1}^{m} (P_U)_i H(P_{V|u_i} \parallel Q_{V|u_i}),$$

where $P_{V|u_i}$ denotes the conditional distribution

$$(P_{V|u_i})_j = \mathsf{Pr}\{v_j | u_i\} = \frac{p_{ij}}{(P_U)_i},$$

and $Q_{V|u_i}$ is defined analogously.

Given two laws $\tilde{P}, \tilde{Q}$ on $\mathcal{N}^\infty$, define

$$\alpha_l := \frac{1}{l} H(\tilde{P}_l \parallel \tilde{Q}_l),$$

$$\beta_l := H(\tilde{P}_{l+1} \parallel \tilde{Q}_{l+1}) - H(\tilde{P}_l \parallel \tilde{Q}_l), l \geq 2.$$

Note that $R(\tilde{P} \parallel \tilde{Q})$ is the limit of α_l as $l \to \infty$ (if it exists). Clearly

$$\alpha_l = \frac{1}{l} \sum_{i=1}^{l-1} \beta_i + \frac{H(\tilde{P}_1 \parallel \tilde{Q}_1)}{l},$$

Theorem: The K-L divergence rate $R(\tilde{P} \parallel \tilde{Q})$ is the Césaro limit, if exists, of the sequence $\{\beta_l\}$.

β_l is often easier to compute than α_l!

View $\mathcal{N}^{l+1}$ as $\mathcal{N}^l \times \mathcal{N}$. Then as shown earlier

$$H(\tilde{P}_{l+1} \parallel \tilde{Q}_{l+1}) = H(\tilde{P}_l \parallel \tilde{Q}_l) + \sum_{\mathbf{u} \in \mathcal{N}^l} f_{\mathbf{u}} H(\mathbf{p}_{|\mathbf{u}} \parallel \mathbf{q}_{|\mathbf{u}}).$$

$$\beta_l = H(\tilde{P}_{l+1} \parallel \tilde{Q}_{l+1}) - H(\tilde{P}_l \parallel \tilde{Q}_l) = \sum_{\mathbf{u} \in \mathcal{N}^l} f_{\mathbf{u}} H(\mathbf{p}_{|\mathbf{u}} \parallel \mathbf{q}_{|\mathbf{u}}),$$

where

$$\mathbf{p}_{|\mathbf{u}} := [\Pr\{1|\mathbf{u}\} \ldots \Pr\{m|\mathbf{u}\}], \mathbf{q}_{|\mathbf{u}} \text{ defined analogously.}$$

Thus, even as $l \to \infty$, $\mathbf{p}_{|\mathbf{u}}, \mathbf{q}_{|\mathbf{u}}$ are probability distributions on the *fixed set* $\mathcal{N}$, though the 'tails' $\mathbf{u}$ become longer and longer.

K-L Divergence Rate Between Markov Processes

K-L Divergence Rate Between Markov Processes

Known Result (with complicated proof): Suppose $\tilde{P} \sim (\theta, A)$, $\tilde{Q} \sim (\phi, C)$ are laws of two Markov chains over a set X with n elements. (A, C are the transition matrices.) Then

$$R(\tilde{P} \parallel \tilde{Q}) = \sum_{i=1}^{n} \theta_i H(\mathbf{a}_i \parallel \mathbf{c}_i),$$

where $\mathbf{a}_i, \mathbf{c}_i \in \mathcal{S}_n$ denote the i-th rows of A, C respectively.

Easy proof: Since both $\tilde{P}, \tilde{Q}$ correspond to Markov chains, we have that under the law $\tilde{P}$,

$$\Pr\{j|i_1 \dots i_l\} = \Pr\{j|i_l\}.$$

Hence $\mathbf{p}_{|\mathbf{u}}$ depends only on the *most recent* element u_l.

$$\mathbf{p}|_{u_1 \dots u_l} = \mathbf{p}|_{u_l}.$$

So if $u_l = i$, then $\mathbf{p}|_i = \mathbf{a}_i$, the i-th row of the transition matrix A. Similarly $\mathbf{q}|_i = \mathbf{c}_i$, the i-th row of the transition matricx C.

$$\begin{aligned}
\beta_l &= \sum_{u_1 \dots u_{l-1} \in \mathcal{N}^{l-1}} \sum_{i=1}^{n} f_{u_1 \dots u_{l-1} i} H(\mathbf{a}_i \parallel \mathbf{c}_i) \\
&= \sum_{i=1}^{n} \left[\sum_{u_1 \dots u_{l-1} \in \mathcal{N}^{l-1}} f_{u_1 \dots u_{l-1} i} \right] H(\mathbf{a}_i \parallel \mathbf{c}_i) \\
&= \sum_{i=1}^{n} \theta_i H(\mathbf{a}_i \parallel \mathbf{c}_i).
\end{aligned}$$

As a result $\beta_l = \beta_1 \; \forall l \geq 2$ (convergence in one time step!). This leads readily to the formula for the divergence between Markov processes.

For a k-step Markov process, we can show similarly that $\beta_l = \beta_k \; \forall l \geq k$.

For a Markov process, the K-L divergence rate is $H(\mathbf{f}_2 \parallel \mathbf{g}_2) - H(\mathbf{f}_1 \parallel \mathbf{g}_1)$, where $\mathbf{f}_m$ is the frequency vector of m-tuples under the law $\tilde{P}$, and $\mathbf{g}_m$ similarly for $\tilde{Q}$.

For a k-step Markov process, the K-L divergence rate is $H(\mathbf{f}_{k+1} \parallel \mathbf{g}_{k+1}) - H(\mathbf{f}_k \parallel \mathbf{g}_k)$.

A very clean interpretation with an easy proof.

Approximation Using Multi-Step Markov Processes

Approximation Using Multi-Step Markov Processes

Problem: Given a stochastic process $\{\mathtt{X}_t\}$ over a finite alphabet $\mathcal{M} = \{1, \ldots, m\}$, specifically the frequency distribution of all k-tuples, find the best possible approximation in terms of an l-step Markov process.

Precisely: Let $f_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^k$ be the set of k-tuple frequencies of a stochastic process $\{\mathtt{X}_t\}$. Let $\mathcal{P}_l$ denote the set of probability laws of l-step Markov processes, and let $P_{l,k} \in \mathcal{S}_{m^k}$ denote the set of k-tuple frequencies obtainable from l-step Markov processes.

Find the best approximation to $[f_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^k] \in \mathcal{S}_{m^k}$ from $P_{l,k}$, in terms of minimizing the K-L divergence (not rate).

Technicality: $l \leq k - 2$; otherwise a perfect match is possible.

Theorem: The best possible fit in terms of minimizing the K-L divergence is as follows: State space $= \mathcal{M}^l$, and if $\mathbf{u} \in \mathcal{M}^l, v \in \mathcal{M}$, then

$$\Pr\{\mathtt{X}_t = v | \mathtt{X}_{t-l} \ldots \mathtt{X}_{t-1} = \mathbf{u}\} = \frac{f_{\mathbf{u}v}}{f_{\mathbf{u}}}, \ \forall \mathbf{u} \in \mathcal{M}^l, \mathbf{v} \in \mathcal{M},$$

where $f_{\mathbf{u}v}, f_{\mathbf{u}}$ are the frequencies of $\mathbf{u}v$ and $\mathbf{u}$ in the stochastic process we are trying to approximate.

Interpretation: Aggregate the k-tuple frequencies into $(l+1)$-tuple frequencies. Match them *perfectly* using an l-step Markov chain.

Approximation Using Multi-Step Markov Processes - II

Problem: Given a $(k-1)$-step Markov process, find the best possible approximation in terms of an l-step Markov process, where the disparity is measured in terms of the K-L divergence rate between the two processes.

Again, problem is meaningful only if $l \leq k-2$.

Theorem: The best possible fit in terms of minimizing the K-L divergence *rate* is as before:

$$\Pr\{\mathtt{X}_t = v | \mathtt{X}_{t-l} \ldots \mathtt{X}_{t-1} = \mathbf{u}\} = \frac{f_{\mathbf{u}v}}{f_{\mathbf{u}}}, \ \forall \mathbf{u} \in \mathcal{M}^l, \mathbf{v} \in \mathcal{M},$$

where $f_{\mathbf{u}v}, f_{\mathbf{u}}$ are the frequencies of $\mathbf{u}v$ and $\mathbf{u}$ in the stochastic process we are trying to approximate.

Interpretation: Same interpretation as before. Aggregate the k-tuple frequencies into $(l+1)$-tuple frequencies. Match them *perfectly* using an l-step Markov chain.

Ergodicity Properties

If the original Markov process is ergodic (irreducible), so is the reduced order model. But the reduced order model can be ergodic (irreducible) even if the original one is not.

Similar remarks apply to primitivity (irreducibility + aperiodicity).

Filtering Equations for Hidden Markov Models

Notation: Given two HMMs $\theta \in \mathcal{S}_n, M^{(u)} \in [0,1]^{n\times n}, \psi \in \mathcal{S}_r, G^{(u)} \in [0,1]^{r\times r}$, where $\mathcal{S}_n$ denotes the n-dimensional simplex. Thus HMM1 has n states, the stationary distribution θ, and

$$m_{ij}^{(u)} = \Pr\{\mathtt{X}_{t+1} = j \& \mathtt{Y}_{t+1} = u | \mathtt{X}_t = i\}.$$

HMM2 has r states, where possibly $r \neq n$, and

$$g_{ij}^{(u)} = \Pr\{\mathtt{X}_{t+1} = j \& \mathtt{Y}_{t+1} = u | \mathtt{X}_t = i\}.$$

From 'sum of products' formula, we have

$$p_{\mathbf{u}} = \theta M^{(\mathbf{u})} \mathbf{e}_n, q_{\mathbf{u}} = \psi G^{(\mathbf{u})} \mathbf{e}_r.$$

Filtering Equations for HMMs:

Define

$$Z := [M^{(u)}\mathbf{e}_n, u \in \mathcal{M}] \in [0,1]^{n\times m},$$

$$T := [G^{(u)}\mathbf{e}_r, u \in \mathcal{M}] \in [0,1]^{r\times m}.$$

Then

$$z_{iu} = \mathsf{Pr}\{\mathtt{Y}_{t+1} = u | \mathtt{X}_t = i\}, \text{ under the law } \tilde{P},$$

and t_{iu} is defined analogously.

For $k \leq l$, let $\mathtt{Y}_k^l = (\mathtt{Y}_k, \ldots, \mathtt{Y}_l)$. Suppose $\mathbf{u} \in \mathcal{M}^l$. Define

$$\theta_{|\mathbf{u}} := [\Pr\{\mathtt{X}_l = i | \mathtt{Y}_1^l = \mathbf{u}\}, i = 1, \ldots, n] \text{ for HMM1 .}$$

$$\psi_{|\mathbf{u}} := [\Pr\{\mathtt{X}_l = i | \mathtt{Y}_1^l = \mathbf{u}\}, i = 1, \ldots, n] \text{ for HMM2 .}$$

Note that $\theta_{|\mathbf{u}} \in \mathcal{S}_n$ and $\psi_{|\mathbf{u}} \in \mathcal{S}_r$. Thus $\theta_{|\mathbf{u}}$ is the conditional distribution of $\mathtt{X}_l$ given the observation $\mathbf{u}$ under the law of HMM1; $\psi_{|\mathbf{u}}$ has a similar interpretation.

Now define

$$\mathbf{x}_{|\mathbf{u}} := \theta M^{(\mathbf{u})}, \mathbf{y}_{|\mathbf{u}} := \psi G^{(u)}.$$

Then it follows from the definition that

$$(\mathbf{x}_{|\mathbf{u}})_i = \mathrm{Pr}\{\mathtt{Y}_1^l = \mathbf{u} \& \mathtt{X}_l = i\} \text{ under HMM1},$$

and $\mathbf{y}_{|\mathbf{u}}$ has a similar interpretation. Note that these vectors satisfy a simple recursion relationship

$$\mathbf{x}_{|\mathbf{uv}} = \mathbf{x}_{|\mathbf{u}} M^{(\mathbf{v})}, \mathbf{y}_{|\mathbf{uv}} = \mathbf{y}_{|\mathbf{u}} G^{(\mathbf{v})}.$$

By Bayes' rule, we have

$$\mathrm{Pr}\{\mathtt{X}_l = i | \mathtt{Y}_1^l = \mathbf{u}\} = \frac{\mathrm{Pr}\{\mathtt{X}_l = i \& \mathtt{Y}_1^l = \mathbf{u}\}}{\mathrm{Pr}\{\mathtt{Y}_1^l = \mathbf{u}\}}.$$

But by the 'sum of products' formula, we have

$$\mathrm{Pr}\{\mathtt{Y}_1^l = \mathbf{u}\} = \begin{cases} \theta M^{(\mathbf{u})} \mathbf{e}_n = \mathbf{x}_{|\mathbf{u}} \mathbf{e}_n \text{ for HMM 1,} \\ \psi G^{(\mathbf{u})} \mathbf{e}_r = \mathbf{y}_{|\mathbf{u}} \mathbf{e}_r \text{ for HMM2} \end{cases}.$$

Therefore

$$\theta_{|\mathbf{u}} = \frac{1}{\mathbf{x}_{|\mathbf{u}} \mathbf{e}_n} \mathbf{x}_{|\mathbf{u}}, \psi_{|\mathbf{u}} = \frac{1}{\mathbf{y}_{|\mathbf{u}} \mathbf{e}_r} \mathbf{y}_{|\mathbf{u}}.$$

The conditional probabilities $\theta_{|\mathbf{u}}$ and $\psi_{|\mathbf{u}}$ *do not* satisfy any simple equation. However, the 'unnormalized conditional probabilities' $\mathbf{x}_{|\mathbf{u}}$ $\mathbf{y}_{|\mathbf{u}}$ satisfy the linear recursion relations

$$\mathbf{x}_{|\mathbf{uv}} = \mathbf{x}_{|\mathbf{u}} M^{(\mathbf{v})}, \mathbf{y}_{|\mathbf{uv}} = \mathbf{y}_{|\mathbf{u}} G^{(\mathbf{v})}.$$

This is a well-known phenomenon in filtering theory.

Estimating the K-L Divergence Rate Between Hidden Markov Models

To compute the K-L divergence rate between HMM1 and HMM2, we need to compute $\Pr\{\mathtt{Y}_{l+1}|\mathtt{Y}_1^l = \mathbf{u}\}$.

In a HMM, the outputs are conditionally independent, given the state. Therefore

$$\begin{aligned}\Pr\{\mathtt{Y}_{l+1} = w|\mathtt{Y}_1^l = \mathbf{u}\} &= \sum_i \Pr\{\mathtt{Y}_{l+1} = w|\mathtt{X}_l = i\} \\ &\cdot \Pr\{\mathtt{X}_l = i|\mathtt{Y}_1^l = \mathbf{u}\},\end{aligned}$$

or more compactly,

$$\mathbf{p}_{|\mathbf{u}} = \theta_{|\mathbf{u}} Z, \mathbf{q}_{|\mathbf{u}} = \psi_{|\mathbf{u}} T.$$

The Alignment Distance – Motivation:

To compute the K-L divergence rate between two HMMs, we need to compute $H(\mathbf{p}_{|\mathbf{u}} \,\|\, \mathbf{q}_{|\mathbf{u}}) = H(\theta_{|\mathbf{u}} Z \,\|\, \psi_{|\mathbf{u}} T)$ as $|\mathbf{u}| \to \infty$.

Difficulty: $\theta_{|\mathbf{u}}, \psi_{|\mathbf{u}}$ don't satisfy any simple recursions.

Remedy: Replace $\theta_{|\mathbf{u}}, \psi_{|\mathbf{u}}$ by $\mathbf{x}_{|\mathbf{u}}, \mathbf{y}_{|\mathbf{u}}$ which *do* satisfy a simple linear recursion.

Problem: $\mathbf{x}_{|\mathbf{u}}, \mathbf{y}_{|\mathbf{u}}$ are not normalized!

Remedy: Find a way of measuring the disparity between *unnormalized* 'probability vectors'!

The Alignment Distance – Definition:

Suppose $\mathbf{x}, \mathbf{y} > \mathbf{0}$. Then

$$d(\mathbf{x}, \mathbf{y}) := \log \left[\frac{\max_i (x_i/y_i)}{\min_j (x_j/y_j)} \right]. = \log \max_i \frac{x_i}{y_i} + \log \max_j \frac{y_j}{x_j}.$$

$d(\mathbf{x}, \mathbf{y})$ is invariant under scaling the two vectors, because

$$d(\alpha\mathbf{x}, \beta\mathbf{y}) = d(\mathbf{x}, \mathbf{y}),\ \forall \alpha, \beta > 0$$

Also $d(\mathbf{x}, \mathbf{y}) > 0$ unless $\mathbf{x} = \lambda\mathbf{y}$ for some constant λ.

Birkhoff Contraction Coefficient:

Suppose $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_+, > \mathbf{0}$ and $T \in \mathbb{R}^{n \times m}_+$, and every column of T contains at least one positive element. Then

$$d(\mathbf{x}T, \mathbf{y}T) \leq d(\mathbf{x}, \mathbf{y}).$$

Suppose $T \in \mathbb{R}^{n \times n}_+$, and define the **Birkhoff contraction coefficient**

$$\tau(T) := \sup_{\mathbf{x},\mathbf{y},\mathbf{x} \neq \lambda \mathbf{y}} \frac{d(\mathbf{x}T, \mathbf{y}T)}{d(\mathbf{x}, \mathbf{y})}.$$

Then $\tau(T) \leq 1$ for all T. Moreover, the Birkhoff contraction coefficient is *submultiplicative*, that is, we have $\tau(TS) \leq \tau(T)\tau(S)$.

Estimating the K-L Divergence Rate Between HMMs:

Using the Birkhoff contraction rate, we can shown that $\mathbf{p}_{|\mathbf{uv}}$ is closely approximated by $\mathbf{p}_{|\mathbf{u}}$ provided $|\mathbf{u}|$ is sufficiently long, and similarly for $\mathbf{q}_{|\mathbf{uv}}$.

This observation plus some high school algebra allow us to obtain *geometrically convergent estimates* for the K-L divergence rate between two HMMs.

Details in MV, CDC 2007.

Open Problems:

The biggest open problem is finding the K-L divergence rate between two *hidden* Markov models.

Other interesting question: What else can be done with these 'closed form formulae' for the K-L divergence rate between Markov processes?

Thank You!