Approximate realization of hidden Markov chains

Lorenzo Finesso [1]
Institute of Systems Science and
Bioengineering, LADSEB-CNR
Corso Stati Uniti, 4 – 35127 Padova, Italy.
e-mail: finesso@ladseb.pd.cnr.it

Peter Spreij
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Plantage Muidergracht 24 – 1018 TV
Amsterdam, The Netherlands
email: spreij@wins.uva.nl

Abstract — **In this paper we consider the approximate realization problem for finite valued hidden Markov models i.e. stochastic processes $Y = f(X)$ where X is a finite state Markov chain and f a many-to-one function. Given the laws $p_Y(\cdot)$ of Y the weak realization problem consists in finding a Markov chain X and a function f such that, at least distributionally, $Y \sim f(X)$. The approximate realization problem consists in finding X and f such that Y and $f(X)$ are close. The approximation criterion we use is the informational divergence between properly defined nonnegative (componentwise) matrices related to the processes. To construct the realization we apply recent results on the approximate factorization of nonnegative matrices.**

I. Introduction

Let $\{Y_t, t \in \mathbb{Z}\}$ be a stationary finitely valued stochastic process that admits a representation of the form $Y_t = f(X_t)$ where $\{X_t, t \in \mathbb{Z}\}$ is a finite Markov chain and f is a many-to-one function. We call such a process a Hidden Markov Chain (HMC).

Under well known conditions on f ([9]) a HMC inherits the Markov property of X_t and becomes a finite Markov chain itself, but this case is non-generic. In general a HMC need not be a Markov chain of any finite order and will therefore exhibit long-range dependencies of some kind. Theoretical work on the specific class of HMC's has proceeded along two main lines. The early contributions, inspired by the work of Blackwell and Koopman (1957) [3], concentrated on the probabilistic aspects. The basic question was the characterization of HMC's. More specifically the problem analyzed was: *among all finitely valued stationary processes Y_t characterize those that admit a HMC representation.* This problem was solved by Heller [6] in 1965. To some extent Heller's result is not quite satisfactory since his methods are non-constructive. Even if Y_t is known to be representable as a HMC, no algorithm has been devised yet to produce, from the laws of Y_t, a Markov chain X_t and a function f such that $Y_t \sim f(X_t)$ (i.e. they have the same laws). Since then the problem has attracted the attention of workers in the area of Stochastic Realization Theory, starting with Picci and Van Schuppen ([8] and the references therein), see also [1] for recent results and a survey of older ones, and while some of the issues have been clarified a constructive algorithm is still missing. The first contributions dealing with statistical aspects were made in the late sixties. Baum and Petrie [2] studied maximum likelihood estimation of the parameters of a HMC proving consistency and asymptotic normality of the MLE. They also provided an algorithm for the numerical computation of the MLE basically inventing the EM algorithm that became popular only later thanks to the work of Dempster, Laird and Rubin.

In this paper we will direct our attention to the approximate realizations, based on recent results on the nonnegative matrix factorization problem (NMF) [7], [10] and we will also present an algorithm to numerically carry out the construction. The convergence properties of this algorithm are also studied and they turn out to be much like those of the EM algorithm [5], [11].

II. HMC's and stochastic systems

Definition. *A process Y is called a hidden Markov chain (HMC) if there is a function f and a Markov chain X such that $Y = f(X)$.*

A different (equivalent) definition is given in terms of the output of a *stochastic system*

Definition. *A pair (X, Y) of stochastic processes taking values in the finite set $\mathcal{X} \times \mathcal{Y}$ is said to be a* stationary finite stochastic system *(SFSS) if*

i) (X, Y) is jointly stationary.
ii) For all t, $x_1, \dots, x_t$, $y_1, \dots, y_t$ it holds that

$$P(Y_{t+1} = \cdot, X_{t+1} = \cdot \mid Y_1^t = y_1^t, X_1^t = x_1^t) \\ = P(Y_{t+1} = \cdot, X_{t+1} = \cdot \mid X_t = x_t)$$

The processes X and Y are called respectively the state *and the* output *of the SFSS.*

Definition. *A stochastic process Y with values in $\mathcal{Y}$ is a Hidden Markov Chain (HMC) if it has the same distribution as the output of a SFSS.*

The probability distribution of an HMC is completely specified by

[1] This work was supported the EC through the RTNs ERNSI and DYNSTOCH.

- the matrices $\{M(y), y \in \mathcal{Y}\}$ with elements

$$m_{ij}(y) = P(Y_{t+1} = y, X_{t+1} = j \mid X_t = i).$$

- an initial (stationary) distribution vector π such that $\pi = \pi A$, where

$$A := \sum_y M(y)$$

is the transition matrix of the MC X.

Definition. *For a word $v = y_1 y_2 \cdots y_n$ define:*

$$\begin{aligned} M(v) &:= M(y_1)M(y_2)\cdots M(y_n), \\ &\quad \textit{a square matrix} \\ g(v) &:= \pi M(v), \ \textit{a row vector} \\ h(v) &:= M(v)e, \ \textit{a column vector,} \\ &\quad \textit{with } e = (1, \ldots, 1). \end{aligned}$$

It easily follows from the definitions that the probability distribution function (pdf) $p(y_1^n) = P(Y_1 = y_1, \ldots, Y_n = y_n)$ is given by

$$p(y_1^n) = \pi M(y_1) \cdots M(y_n) e,$$

and for any two words u and v

$$p(uv) = \pi M(u) M(v) e = g(u) h(v),$$

Factorization hypothesis:

$$\begin{aligned} &P(Y_{t+1} = y, X_{t+1} = j \mid X_t = i) \\ &\quad = P(Y_{t+1} = y \mid X_{t+1} = j) P(X_{t+1} = j \mid X_t = i) \end{aligned}$$

Define with $d = |\mathcal{X}|$

$$\begin{aligned} b_{iy} &:= P(Y_t = y \mid X_t = i) \\ B_y &:= \operatorname{diag}\{b_{1y}, b_{2y}, \cdots b_{dy}\}. \end{aligned}$$

The factorization hypothesis then reads

$$M(y) = A B_y$$

If $Y = f(X)$, a deterministic function of X, then $b_{iy} \in \{0, 1\}$ with $b_{iy} = 1$ iff $f(i) = y$ and the factorization hypothesis holds. Since it is always possible to represent an HMC as a deterministic function of a MC, we will assume w.l.o.g. the factorization hypothesis.

III. Realization of an HMC

Let Y be a finite valued stationary process with pdf $p_Y(\cdot)$. Heller's theorem characterizes the subset of finite valued stationary processes that are HMC's. Let C^* be the convex set of probability distributions on $\mathcal{Y}^*$. A convex subset $C \subset C^*$ is *polyhedral stable* if $C = \operatorname{conv}\{q_1(\cdot), \cdots, q_c(\cdot)\}$ and for $1 \leq i \leq c$ and $\forall y \in \mathcal{Y}$ the conditional distributions $q_i(\cdot \mid y) := \frac{q_i(y\cdot)}{q_i(y)} \in C$.

Theorem. (Heller) *$p_Y(\cdot)$ is the pdf of an HMC iff the set $C_Y := \operatorname{conv}\{p_Y(\cdot \mid u) : u \in \mathcal{Y}^*\}$ is contained in a polyhedral stable subset of C^*.*

Weak stochastic realization problem. Let Y be a HMC with pdf $p_Y(\cdot)$. Find an SFSS $(X, \hat{Y})$ such that the pdf $p_{\hat{Y}}(\cdot) = p_Y(\cdot)$.

Any such SFSS is called a *realization* of Y, and the problem reduces to finding matrices (A, B_y) that specify its distribution. The realization is inherently non-unique. The number $d = |\mathcal{X}|$ is called the *dimension* of the realization. The minimal dimension of a realization of Y is called *order* of the HMC.

The realization problem is unsolved in general. Heller's theorem doesn't help since its proof is not constructive. The following special case sheds some light on the solution.

Static realization problem. Given a joint pdf $p_0(\cdot, \cdot)$ and integer d, find, if it exists, a triple of rv's (Y^+, X, Y^-) such that

(i) the support of X has cardinality d
(ii) the marginal pdf of (Y^+, Y^-) is p_0
(iii) Y^+ and Y^- are CI given X.

The following result characterizes the joint pdf's admitting a realization

Theorem. *The static realization problem has a solution iff the positive matrix $P_0 = ||p_0(i, j)||$ can be factored as*

$$P_0 = GH,$$

where G and H are positive matrices with inner size d.

The smallest d for which the solution exists is called the *positive rank* of P_0.

Definition. Compound sequence matrices (c.s.m.) *Let $p(\cdot)$ be a given pdf, and $v_1 \cdots v_n$, $v_1' \cdots v_n'$ words from $\mathcal{Y}^*$. The c.s.m. P is*

$$P = P(v_1 \cdots v_n, v_1' \cdots v_n') = ||p(v_i v_j')||_{i,j}$$

The rank *of $p(\cdot)$ is defined as the maximum of the ranks of all possible c.s.m. if such maximum exists or $+\infty$ otherwise.*

If $p(\cdot)$ is the *pdf* of an HMC which admits a representation of dimension d then since $p(v_i v_j') = g(v_i) h(v_j')$ one has

$$P = GH$$

where G, H are $n \times d$ and $d \times n$ matrices.

Let $k = |\mathcal{Y}|$, then

Lemma. *The rank of an HMC is a lower bound to its order. Moreover the distribution of Y is completely determined by the $k^d \times k^d$ c.s.m. with elements $p(vv')$, where v and v' exhaust all words of length d.*

IV. Approximate realization of HMC's

The following considerations prompt the interest for the approximate realization problem for HMC's.

(i) There is no algorithm to construct an exact realization of a given HMC of known order.

(ii) The order of the HMC may be too large.

(iii) We might be interested in approximating a general pdf with an HMC of given dimension.

Our aim will be the *construction* of an approximate realization of *assigned (low)* dimension. To this end we apply recent results on the problem of approximate nonnegative matrix factorization ([7]) which we briefly recall.

Definition. *Let M and N be nonnegative (componentwise) matrices of the same size, their informational divergence is defined as*

$$D(M||N) = \sum_{ij} M_{ij} \log \frac{M_{ij}}{N_{ij}} - M_{ij} + N_{ij}.$$

It holds that $D(M||N) \geq 0$ with equality iff $M = N$.

Approximate nonnegative matrix factorization. Given P, a positive $n \times m$ matrix, and integer $k < n, m$, find positive matrices G and H, of sizes $n \times k$, $k \times m$ respectively, such that

$$D(P||GH) \text{ is minimized.}$$

If $k = 1$, then the minimizing G and H are proportional to the row sums and column sums of P. The problem has no closed form solution for $k > 1$, but a recursive algorithm for the construction of the optimal G and H has been developed ([7]). We are now ready to formulate the main problem.

Approximate realization of an HMC. Given a *pdf* $p(\cdot)$, and an integer d, find an HMC Y, of dimension d, at minimal divergence distance from $p(\cdot)$.

Notice that, in general, $p(\cdot)$ is not required to be an HMC. The following result motivates our approach to the construction of approximate realizations

Lemma. *Let P^n and P^n_Y be the complete c.s.m.'s corresponding to n observations from $p(\cdot)$ and $p_Y(\cdot)$, then*

$$D(p(\cdot)||p_Y(\cdot)) = \lim_{n\to\infty} \frac{1}{n} D(P^n||P^n_Y)$$

The c.s.m. P_Y of an HMC factorizes as $P_Y = GH$ for some nonnegative matrices G and H. This observation and the previous lemma suggest a natural approximation scheme. From $p(\cdot)$ construct the c.s.m. P of size $n \times n$ and then find matrices G and H of sizes $n \times d$ and $d \times n$ respectively to

$$\text{minimize } D(P||GH)$$

From these G and H construct a pair (A, B) representing the optimal approximate HMC.

Construction of (A,B) – Preliminaries

Definition. *Given a pdf $p(\cdot)$, a value $s \in \mathcal{Y}$, and words $u_1, \dots, u_n, v_1, \dots, v_n$, the* pinned c.s.m. is

$$P(s) = P(u_1, \dots, u_n, s, v_1, \dots, v_n).$$

Fact: If Y is an HMC, with c.s.m. $P_Y = GH$ the *pinned* c.s.m. $P_Y(s)$ is expressed as

$$P_Y(s) = GAB_sH$$

The rough idea is to solve, for each of the pinned c.s.m. $P(s)$, $s \in \mathcal{Y}$, corresponding to $p(\cdot)$, the minimization in (A, B_s)

$$\text{minimize } D(P(s)||GAB_sH)$$

Properly gluing the solutions will produce the optimal (A, B) pair.

Notation. If Y is an HMC of the form $Y = f(X)$, then we have the rank factorization

$$B_s = e_s e_s^\top,$$

with e_s a matrix with elements in $\{0, 1\}$. Then

$$P_Y(s) = G\,K_s = G\,A_s H_s,$$

where $A_s = Ae_s, H_s = e_s^\top H$. Postmultiplication by e_s selects the columns of A corresponding to the states i for which $f(i) = s$.

Three step algorithm to produce A

Suppose that we seek an approximate realization of dimension d such that $Y = f(X)$, for a fixed function f.

1. Minimize $D(P||GH)$ over G and H subject to the constraint $He = e$. Decompose the minimizing H^* into 'row blocks' $H^*_s = e_s^\top H$, $(s = 1, \dots, k)$.
2. Take the minimizing G^* from the previous step and minimize $D(P(s)||GK)$ over K for each $s \in \mathcal{Y}$, call the minimizer K^*_s.
3. Minimize $D(K^*_s||A_sH^*_s)$ over A_s for each s under the constraint that $Ae = e$, where $A := (A_1, \dots, A_k)$.

The resulting minimizer A^* is taken as the transition matrix of the underlying Markov chain.

Final remarks

1. There is an EM-type alternating algorithm for the first step that displays monotone convergence to a local optimum. For the other two steps one can apply results form Csiszar and Tusnady to have convergence to a global optimum.

2. If the process Y is approximated by a Markov chain, the resulting matrix A^* has elements $A^*_{ij} = P(Y_1 = j|Y_0 = i)$.

3. Alternative algorithms may also be considered, in particular one that 'reverses' the order of the first two steps of the given algorithm.

References

[1] B.D.O. Anderson (1999), The Realization Problem for Hidden Markov Models, *Mathematics of Control, Signals, and Systems* **12**, 80-120.

[2] Baum L.E. and Petrie T. (1966), Statistical inference for probabilistic functions of finite Markov chains, *Ann. of Mathem. Statist.*, **37**, pp. 1554-1563

[3] Blackwell and Koopman (1957), On the identifiability problem for functions of finite Markov chains, *Ann. of Mathem. Statist.*, **28**, 1011-1015.

[4] Carlyle J.W. (1969), Stochastic finite-state system theory, in *Systems Theory*, L. Zadeh and L. Polak eds., McGraw-Hill, New York, Chapter 10.

[5] I. Csiszár and G. Tusnády (1984), Information geometry and alternating minimization procedures, *Statistics & Decisons, supplement issue* **1**, 205-237

[6] Heller A. (1965), On stochastic processes derived from Markov chains, *Ann. Mathem. Stat.*, **36**, pp. 1286-1291

[7] D.D. Lee and H.S. Sebastian Seung (1999), Learning the parts of objects by non-negative matrix factorization, *Nature* **401**, pp 788-791.

[8] G. Picci, J.M. van den Hof, and J.H. van Schuppen (1998), Primes in several classes of the positive matrices, *Linear Algebra and its Applications*, **277**, 149-185.

[9] P.J.C. Spreij (2001), *On the Markov property of a finite hidden Markov chain*, Statistics and Probability Letters, Vol 52/3, pp 279–288.

[10] J.A. O'Sullivan (1998), Alternating minimization algorithms: From Blahut-Arimoto to Expectation-Maximization, in *A. Vardy, ed., Codes, Curves and Signals, Common Threads in Communications*, Kluwer Academic, Boston, pp 173-192.

[11] J.A. O'Sullivan (2000), Properties of the information value decomposition, *Proceedings ISIT 2000, Sorrento, Italy*, pp 491.