Approximation of the I-divergence between stationary and hidden Markov processes

Lorenzo Finesso[1], Angela Grassi[1], and Peter Spreij[2]

[1] ISIB-CNR
Corso Stati Uniti 4
35127 Padova, Italy
(e-mail: finesso@isib.cnr.it, angela.grassi@isib.cnr.it)

[2] Korteweg-de Vries Institute for Mathematics
Universiteit van Amsterdam
Plantage Muidergracht 24
1018TV Amsterdam, The Netherlands
(e-mail: spreij@science.uva.nl)

Abstract. We aim at the construction of a Hidden Markov Model (HMM) of assigned complexity (number of states of the underlying Markov chain) which best approximates, in Kullback-Leibler divergence rate, a given stationary process. We establish, under mild conditions, the existence of the divergence rate between a stationary process and an HMM. Since in general there is no analytic expression available for this divergence rate, we approximate it with a properly defined, and easily computable, divergence between Hankel matrices, which we propose as an approximation criterion.
Keywords. Hidden Markov Model, approximation, stochastic realization, divergence rate.

1 Introduction

The probabilistic characterization of HMMs was first given by Heller (1965). The problem analyzed was: *among all finitely valued stationary processes Y_t, characterize those that admit an HMM representation.* To some extent the results of Heller are not quite satisfactory, since the proofs are non-constructive. Even if Y_t is known to be representable as an HMM, no algorithm has been devised to produce a *realization* i.e. to construct, 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). As stated, the problem has attracted the attention of workers in the area of Stochastic Realization Theory, starting with Picci and Van Schuppen (1984), see also Anderson (1999). More recent references with related results are Vidyasagar (2004) and Vanluyten et al. (2006). While some of the issues have been clarified a constructive algorithm is still missing.

In this short paper we direct our attention to the approximation of stationary processes by HMMs and cast it as a Nonnegative Matrix Factorization (NMF) problem in terms of certain Hankel matrices. In Lee and Seung (1999)

numerical procedures for NMF have been proposed and convergence properties of some of them have been studied in Finesso and Spreij (2005); they turn out to be very close to those of the EM algorithm, although the algorithm for NMF is completely deterministic.

This paper develops and extends some preliminary ideas presented in Finesso and Spreij (2002). Proofs and more details can be found in Finesso et al. (2008).

2 Mathematical Preliminaries on HMMs

In this paper we consider discrete time Hidden Markov Models (HMM) with values in a finite set. We follow Picci (1978), Picci and van Schuppen (1984), see also Anderson (1999), for the basic definitions and notations.

Let $Y = (Y_t)_{t\in\mathbb{Z}}$ be a discrete time stationary stochastic process defined on a given probability space $\{\Omega, \mathcal{A}, P\}$ and with values in the finite set (alphabet) $\mathcal{Y}$ of cardinality m. $\mathcal{Y}^*$ will denote the set of finite strings of symbols from the alphabet $\mathcal{Y}$, with the addition of the empty string. The probability distribution of the process induces a map $p : \mathcal{Y}^* \to [0, 1]$ as follows. Let $v \in \mathcal{Y}^*$, $v = y_0 \cdots y_k$ for some k, then $p(v) = P(Y_0 = y_0, \ldots, Y_k = y_k)$. The map p represents the finite dimensional probability distributions of the process Y, sometimes referred to as *pdf*. Let Y be an HMM, with underlying Markov chain X taking values in another finite set $\mathcal{X}$ of cardinality N, the *size* of the HMM. The probability distribution of a stationary HMM is specified by the m nonnegative matrices $\{M(y), y \in \mathcal{Y}\}$ of size $N \times N$ with elements

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

We also need a probability (row) vector π of size N, such that $\pi = \pi A$, where $A := \sum_y M(y)$. The matrix A is the transition matrix of the Markov chain $(X_t)_{t\in\mathbb{Z}}$ and π is an invariant vector of A. Since the state space is finite, such an invariant vector exists, and is unique if A is irreducible.

Let $w \in \mathcal{Y}^n$ (the set of strings of length n) be given by $w = y_1 \cdots y_n$. Then $p(w)$ can be written in terms of the matrices $M(y_i)$ as

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

and for any pair of strings u and v in $\mathcal{Y}^*$, one has

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

where $e = (1, \ldots, 1)^\top$.

We recall the weak stochastic realization problem (Picci (1978)) for HMMs, which we state as follows. Let Y be an HMM with law $P_Y(\cdot)$, find matrices $M(y)$ that induce the law $P_Y(\cdot)$. A solution is inherently non-unique.

The realization problem is unsolved in general. In the present paper we propose to look for an approximate realization. The advantage of this alternative approach is that it can also be used as a procedure to approximate any given stationary distribution by that of an HMM. We formulate this approximate realization problem as a problem of optimal approximation in *divergence rate*, to be defined in the next section.

3 Divergence rate, existence and minimization

In this section we recall the definition of the divergence rate between processes, as previously given in for instance Juang and Rabiner (1985) for two HMMs. Consider a process $Y = (Y_t)_{t\in\mathbb{Z}}$ with values in $\mathcal{Y}$ under two probability measures P and Q. We interpret P and Q as the laws of the process in the path space $\mathcal{Y}^\infty$. Let $p(y_0,\dots,y_k) = P(Y_0 = y_0,\dots,Y_k = y_k)$ and $q(\cdot)$ likewise. For reasons of brevity, we write $p(Y_0^k)$ for the likelihood $p(Y_0,\dots,Y_k)$ and likewise we also write $q(Y_0^k)$.

Definition 1. Let Q and P be measures on $\mathcal{Y}^\infty$ with q and p as the corresponding families of finite dimensional distributions. Define the divergence rate of Q with respect to P as

$$D(Q\|P) := \lim_{n\to\infty} \frac{1}{n} E_Q\left[\log \frac{q(Y_0^{n-1})}{p(Y_0^{n-1})}\right] \tag{1}$$

Theorem 1. *Let Y be a process with values in $\mathcal{Y}$. Let Q be an arbitrary stationary distribution of Y on $\mathcal{Y}^\infty$ and P a stationary HMM distribution on $\mathcal{Y}^\infty$. Assume that*

(i) the distributions of all finite segments $(Y_0,\dots,Y_{n-1})$ under Q are absolutely continuous with respect to those under P.
(ii) Q admits an invariant probability measure μ^* on $\mathcal{Y}$ i.e.

$$\mu^*(y) = \sum_{y_0} Q(Y_1 = y|Y_0 = y_0)\mu^*(y_0).$$

(iii) $(Y_t)_{t\in\mathbb{Z}}$ is geometrically ergodic under Q i.e. $\exists\rho\in(0,1)$

$$|Q(Y_n = y|Y_0 = y_0) - Q(Y_n = y|Y_0 = y_0')| = O(\rho^n) \quad \forall y, y_0, y_0' \in \mathcal{Y}.$$

Then the limit in (1) exists and is finite.

Remark 1. A sufficient condition that ensures the absolute continuity condition of Theorem 1 is

$$(i') \qquad \sum_j m_{ij}(y) = P(Y_{t+1} = y|X_t = i) > 0, \qquad \forall y\in\mathcal{Y}, \forall i\in\mathcal{X}.$$

The problem we alluded to at the end of Section 2 is

Problem 1. Given Q, a stationary measure on $\mathcal{Y}^\infty$, solve $\inf_P D(Q\|P)$, where the infimum is taken over all stationary HMM distributions of size N.

This problem is well defined under the conditions of Theorem 1, since the divergence rate is then guaranteed to exist. There is however a major problem. No analytic expression is known for the divergence rate, when Q is arbitrary and P an HMM measure (except for a Markov law P, that we will treat in Remark 2). This is even the case if Q itself is an HMM measure, see Han and Marcus (2006) for some recent results. A similar observation has already been made in Blackwell (1957), where the entropy rate of an HMM was studied for the first time. This motivates an alternative approach. In the next section we will approximate the abstract minimization problem with a, in principle, numerically tractable one. For this we will need the Hankel matrix involving all finite dimensional distributions of a stationary process and that of an HMM. This is the topic of the next section.

Remark 2. The minimization problem can be solved explicitly if P runs through the set of all stationary Markov distributions. The minimizing distribution P^* in this case is such that the transition probabilities $P^*(Y_{t+1} = j|Y_t = i)$ of the approximating Markov chain coincide with the conditional probabilities $Q(Y_{t+1} = j|Y_t = i)$ and the invariant (marginal) distribution under P^* is the same as the one under Q. A similar result holds for approximation by a k-step Markov chain. Unfortunately, such appealing closed form solutions do not exist if the minimization is carried out over stationary HMM measures.

4 Hankel matrix for stationary processes

Given an integer n, we define two different orders on $\mathcal{Y}^n$: the *first lexicographical order* (*flo*) and the *last lexicographical order* (*llo*). These orders have been introduced in Anderson (1999). In the *flo* the strings are ordered lexicographically reading from right to left. In the *llo* the strings are ordered lexicographically reading from left to right (the ordinary lexicographical ordering). On $\mathcal{Y}^*$ we define two enumerations: $(u_\alpha)_{flo}$ and $(v_\beta)_{llo}$. In both cases the first element of the enumeration is the empty string. For $(u_\alpha)_{flo}$ we then proceed with the ordering of $\mathcal{Y}^1$ according to *flo*, then with the ordering of $\mathcal{Y}^2$ according to *flo*, and so on. The enumeration $(v_\beta)_{llo}$ is obtained by having the empty string followed by the ordering of $\mathcal{Y}^1$ according to *llo*, then by the ordering of $\mathcal{Y}^2$ according to *llo*, and so on. In both cases the length of a string increases monotonically with the index α or β.

Definition 2. For a stationary process with pdf $p(\cdot)$ the Hankel matrix $\mathbf{H}$ is the infinite matrix with elements $p(u_\alpha v_\beta)$, where u_α and v_β are respectively the α-th and β-th elements of the two enumerations.

Fix integers $K, L \geq 0$. Let $u_1, u_2, \ldots, u_\gamma$ with $\gamma = m^K$ be the enumeration according to the *flo* of the m^K strings of length K. Similarly let $v_1, v_2, \ldots, v_\delta$ with $\delta = m^L$ be the enumeration according to the *llo* of the m^L strings of length L. Let us denote by $\mathbf{H}_{KL}$ the (K, L) block of $\mathbf{H}$ of size $m^K \times m^L$ given by its elements $p(u_i v_j)$ with $i = 1, \ldots, \gamma$ and $j = 1, \ldots, \delta$.

Proposition 1. *Let* $\mathbf{H}$ *be the Hankel matrix of an HMM. The following factorizations hold true.*

$$\mathbf{H}_{KL} = \mathbf{\Pi}_K \mathbf{\Gamma}_L,$$

with

$$\mathbf{\Pi}_K := \begin{bmatrix} \pi M(u_1) \\ \vdots \\ \pi M(u_\gamma) \end{bmatrix}, \qquad \mathbf{\Gamma}_L := \begin{bmatrix} M(v_1)e & \cdots & M(v_\delta)e \end{bmatrix}.$$

5 Divergence rate approximation

First we define the informational divergence between two positive matrices.

Definition 3. Let $\mathbf{M}, \mathbf{N} \in \mathbb{R}_+^{m \times n}$. The informational divergence of $\mathbf{M}$ relative to $\mathbf{N}$ is

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

It follows that $D(\mathbf{M} \| \mathbf{N}) \geq 0$ with equality iff $M = N$. The divergence rate between two processes can be approximated by the informational divergence between their Hankel matrices, as we will demonstrate now.

Let Q and P be measures as in Theorem 1. Denote by $\mathbf{H}_{nn}$ and $\mathbf{H}_{nn}^P$ the (n, n) block of their Hankel matrices. A typical element of $\mathbf{H}_{nn}$ is

$$q^{(2n)}(u_i v_j) := Q(Y_0^{2n-1} = u_i v_j) \qquad \forall u_i \in \mathcal{Y}^n \text{ in } \textit{flo}, \forall v_j \in \mathcal{Y}^n \text{ in } \textit{llo},$$

and a typical element of $\mathbf{H}_{nn}^P$ has a similar expression. The informational divergence between the Hankel blocks is

$$\begin{aligned} D(\mathbf{H}_{nn} \| \mathbf{H}_{nn}^P) &= \sum_{u_i, v_j \in \mathcal{Y}^n} q^{(2n)}(u_i v_j) \log \frac{q^{(2n)}(u_i v_j)}{p^{(2n)}(u_i v_j)} \\ &= E_Q \left[\log \frac{q(Y_0^{2n-1})}{p(Y_0^{2n-1})} \right] \end{aligned}$$

which, when compared to the definition of divergence rate, provides the following

Theorem 2. *Assume that* P *and* Q *are as in Theorem 1. Then the divergence rate exists and*

$$\lim_{n \to \infty} \frac{1}{2n} D(\mathbf{H}_{nn} \| \mathbf{H}_{nn}^P) = D(Q \| P).$$

This theorem motivates the use of $\frac{1}{2n}D(\mathbf{H}_{nn}\|\mathbf{H}^P_{nn})$, for n large enough, as an approximation of the divergence rate between Q and P. In Finesso et al. (2008) an algorithm is presented that results in an approximate HMM realization of a given stationary process. It involves at each step an approximate Nonnegative Matrix Factorization problem. At the heart of the algorithm lies the factorization property of $\mathbf{H}^P_{nn}$ as in Proposition 1. The NMF problem that replaces Problem 1 is

$$\min_{\mathbf{\Pi}_n,\mathbf{\Gamma}_n} D(\mathbf{H}_{nn}\|\mathbf{\Pi}_n\mathbf{\Gamma}_n)$$

under the constraints $e^\top\mathbf{\Pi}_n e = 1$ and $\mathbf{\Gamma}_n e = e$.

References

Anderson, B.D.O. (1999): The realization problem for hidden Markov models, *Mathematics of Control, Signals, and Systems*, **12**, 80–120.

Blackwell, D. (1957): The entropy of functions of finite-state Markov chains, *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, 13–20.

Finesso, L. and Spreij, P.J.C. (2002): Approximate realization of finite hidden Markov chains, *Proceedings of the 2002 IEEE Information Theory Workshop*, Bangalore, India, 90–93.

Finesso, L. and Spreij, P.J.C. (2006): Nonnegative matrix factorization and I-divergence alternating minimization, *Linear Algebra and its Applications*, **416**, 270–287.

Finesso, L., Grassi, A. and Spreij, P. (2008): Approximation of stationary processes by Hidden Markov Models, *arXiv:math/0606591v2*.

Han, G. and Marcus, B. (2006): Analyticity of entropy rate of hidden Markov chains, *IEEE Transactions on Information Theory*, **52(12)**, 5251–5266.

Heller, A. (1965): On stochastic processes derived from Markov chains, *Ann. Math. Statist.*, **36**, 1286–1291.

Juang, B.H. and Rabiner, L.R. (1985): A probabilistic distance measure for hidden Markov models, *AT&T Technical Journal*, **64(20)**, 391–408.

Lee, D.D. and Seung, H.S. (1999): Learning the parts of objects by non-negative matrix factorization, *Nature*, **401**, 788–791.

Picci, G. (1978): On the internal structure of finite state stochastic processes, *Recent Developments in Variable Structure Systems, Economics and Biology, R.R. Mohler and A. Ruberti eds., Lecture notes in Economics and Mathematical Systems*, **162**, Springer-Verlag, Berlin, 288–304.

Picci, G. and van Schuppen, J.H. (1984): On the weak finite stochastic realization problem, *Lecture Notes in Control and Information Sciences*, **61**, 237–242, Springer, New York.

Vanluyten, B, Willems, J.C. and De Moor, B. (2006): Matrix factorization and stochastic state representations, Internal Report 06-31, ESAT-SISTA, K.U. Leuven (Leuven, Belgium).

Vidyasagar, M. (2005): The realization problem for hidden Markov models: the complete realization problem, *Proceedings of the 44th Conference on Decision and Control and the European Control Conference 2005*, Sevilla, 6632–6637.