

Applications of large deviations to hidden Markov chains estimation

Fabiola Del Greco M.¹

Institute of Genetic Medicine,
EURAC research, e-mail: fabiola.delgreco@eurac.edu

Abstract: Consider a Hidden Markov model where the observations are generated by an underlying Markov chain plus a perturbation. In this paper we apply large deviation theory to get an approximate confidence interval for the stationary distribution for the underlying Markov chain.

Keywords: Hidden Markov chain, large deviations theory, information and statistical modelling, confidence intervals.

1. Introduction

Hidden Markov models (HMMs) describe the relationship between two stochastic processes: an observed one and an underlying hidden (unobserved) process. These models are used for two purposes. The first one is to make inferences or predictions about an unobserved process based on the observed one. A second reason is to explain variation in the observed process on variation in a postulated hidden one. For these reasons, the HMMs constitute a very popular class of discrete time stochastic processes, with many applications (i.e. biology, speech recognition, finance, ...). For a general reference on these models, see (CMR05).

More precisely, suppose a phenomena is driven by a discrete time finite state space Markov chain $\mathbf{X}$. Due to measurements errors we can observe only perturbed values. Suppose we have a set of observations $\mathbf{Y} := \{Y_i, i \in \mathbb{N}\}$, which take values in $\mathbb{R}^d$, and the following relationship holds

$$Y_i = X_i + \varepsilon_i, \quad \forall i \geq 1,$$

where the processes $\mathbf{X} := \{X_j, j \in \mathbb{N}\}$ and $\varepsilon := \{\varepsilon_j, j \in \mathbb{N}\}$ satisfy the following assumptions.

The process $\mathbf{X}$ is a homogeneous Markov chain with a finite known state space $\Omega := \{x_1, \dots, x_m\}$, with $x_i \in \mathbb{R}^d$. Let P^* be the true transition matrix of $\mathbf{X}$, whose entries are $P_{i,j}^* := \mathbb{P}(X_1 = x_j \mid X_0 = x_i)$. Of course P^* is unknown and we assume it is irriducible and aperiodic. This assumption implies the existence and uniqueness of the stationary distribution, which we denote by $\pi^* = (\pi_1^*, \pi_2^*, \dots, \pi_m^*)$.

In the literature we found three different methods for computing confidence intervals (CIs) of Hidden Markov chains parameters, namely likelihood profiling, bootstrapping and CIs based on finite-differences approximation of the Hessian. The problem of interval estimate for parameters of a HMM has been studied in (VRM00). Their conclusion is that likelihood profiling and bootstrapping provide similar results, whereas the

¹Address of correspondence: Viale Druso n. 1, 39100, Bolzano, Italy

finite-differences intervals are mostly too small. There is no detailed analysis of the true coverage probability CIs in the context of HMMs. We propose a CI for the stationary distribution of the underlying Markov chain using a large deviations approach.

Nowadays, the theory of large deviations is one of the most rapidly expanding topics in many areas, such as statistics, probability, engineering and physic. It has been applied to a wide range of problems in which detailed information on rare events is required. Of course, one could be interested not only in the probability of rare events but also in the characteristic behavior of the system as the rare event occurs.

2. Study framework

The Markov chain $\mathbf{X} := \{X_i\}_{i \geq 1}$ is observed through a perturbation sequence $\varepsilon := \{\varepsilon_j\}_{j \geq 1}$ of random variables. Assume that the variables ε_j are, conditionally to X_i , independent. Moreover, given $\{X_n = x_j\}$, ε_n has distribution $\mathbb{Q}_j$, with $j \in \{1, 2, \dots, m\}$, and is independent from $\{\varepsilon_l, X_l, l \in \{1, 2, \dots, n-1\}\}$.

For any subset $C \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$ let $C - x := \{y \in \mathbb{R}^d : y + x \in C\}$. Let $\mathcal{U}$ the collection of partitions $U := (U_1, U_2, \dots, U_m)$ with $U_i \subset \mathbb{R}^d$, satisfying the following properties. The sets $U_1, U_2, \dots, U_m$ are disjoint, with non-empty interior set (i.e. the largest open set contained in U_i is not the empty set), $\bigcup_{j=1}^m U_j = \mathbb{R}^d$ and the matrix $Q_U = [q_{i,j}^{(U)}]_{1 \leq i,j \leq m}$ has full rank, where $q_{i,j}^{(U)} := \mathbb{Q}_j(U_i - x_j)$. We assume that the measures $\mathbb{Q}_j$, with $j \in \{1, 2, \dots, m\}$, make $\mathcal{U}$ non-empty.

For a vector $\mathbf{x} \in \mathbb{R}^d$ we use the notation $\mathbf{x} \gg 0$ to indicate that each coordinate of $\mathbf{x}$ is nonnegative. For any vector $\mathbf{u}$ set $\text{diag}(\mathbf{u})$ to be the diagonal matrix whose (i, i) element is u_i , and let I_m be the $m \times m$ identity matrix.

Let $\mu := (\mu_1, \mu_2, \dots, \mu_m)$ such that $\mu_k \in (0, 1)$ and $\sum_{k=1}^m \mu_k = 1$ and define

$$\mathcal{H}_U := \left\{ \mathbf{x} : \det[P^* \text{diag}(\mathbf{x} Q_U) - I_m] = 0 \right\}, \quad J_U(\mu) := \sup_{\mathbf{x} \in \mathcal{H}_U : \mathbf{x} \gg 0} \sum_{k=1}^m \mu_k \log x_k,$$

and $B_r(\mathbf{x}) := \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| \geq r\}$, where $\|\cdot\|$ stands for the Euclidean norm in $\mathbb{R}^d$.

Set $\overline{B}_r = B_r(Q_U \pi^*)$. We assume that there exists a partition U such that $\inf_{\mu \in \overline{B}_r^c} J(\mu) - m + 1$ has a known lower bound H_r which is strictly positive for $r > 0$. Denote by $\hat{d}_i^{(n)} := \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{\{Y_j \in U_i\}}$ where $\mathbb{1}_A$ is the indicator function of the event A and let $\hat{d}^{(n)}(U) := (\hat{d}_1^{(n)}, \hat{d}_2^{(n)}, \dots, \hat{d}_m^{(n)})$.

Theorem 1. (Confidence Interval) Fix $\alpha > 0$ and choose the smallest r such that $e^{-nH_r} \leq \alpha$. Then, the set $A_r = Q_U^{-1}(B_r(\hat{d}^{(n)}(U)))$ is an approximate $(1 - \alpha)$ -confidence interval.

Our approach relies on the fact that $\hat{d}_n(U)$ converges to $Q_U \pi^*$ and does it quite fast. The large deviations principle makes this statement rigorous. Hence we use H_r to lower estimate the rate function.

3. Construction of the confidence interval

Definition 2. (Large deviations principle) A rate function is a function which is non-negative and lower semicontinuous. A sequence of random variables Z_n , $n \in \mathbb{N}$, satisfies

a large deviations principle with rate function $I(\cdot)$ if we have

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(Z_n \in F) &\leq - \inf_{x \in F} I(x), & \text{for any closed set } F, \text{ and} \\ \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(Z_n \in A) &\geq - \inf_{x \in A} I(x), & \text{for any open set } A. \end{aligned}$$

Proposition 3. Define

$$\Lambda_j^{(U)}(\lambda) := \mathbb{E} \left[\exp \left\{ \sum_{k=1}^m \lambda_k \mathbb{1}_{\{\varepsilon_1 \in U_k - x_j\}} \right\} \right].$$

For any Borel set $U \subset \mathbb{R}^d$, $\widehat{d}^{(n)}(U)$ satisfies a large deviations principle with rate function $I_U(z) := \sup_{\lambda \in \mathbb{R}^m} \{ \langle \lambda, z \rangle - \log \rho(P_\lambda) \}$, where $\langle \cdot \rangle$ is the usual inner product, P_λ is the matrix whose (i, j) entry is $P_{i,j}^* \Lambda_j^{(U)}(\lambda)$ and for any irreducible matrix A the scalar $\rho(A)$ is the so called Perron–Frobenius eigenvalue, with the properties $\rho(A) \in (0, \infty)$ and $|\theta| \leq \rho(A)$ for any eigenvalue θ of A .

Proposition 4. $\widehat{d}^{(n)}(U)$ converges a.s. to $Q_U \pi^*$.

Denote by $\mathcal{N}(Q_U)$ the set of linear combinations, with non-negative coefficients, of the rows of Q_U . More precisely

$$\mathcal{N}(Q_U) := \{ \mathbf{x} Q_U : \mathbf{x} \gg 0 \}.$$

For any $\mathbf{t} := (t_1, t_2, \dots, t_m)$, with $\mathbf{t} \gg 0$, define

$$L_k(\mathbf{t}) := \frac{t_k}{\sum_{i=1}^m t_i P_{i,k}^*}, \quad \text{for } k \in \{1, \dots, m\},$$

and $L(\mathbf{t}) := (L_1(\mathbf{t}), L_2(\mathbf{t}), \dots, L_m(\mathbf{t}))$. Notice that for any positive scalar a we have $L(a\mathbf{t}) = L(\mathbf{t})$. Denote by $\mathcal{M}_1$ the set of probability measures on Ω . We consider $\mathcal{M}_1$ as a subset of $\mathbb{R}^m$, that is the m -dimensional simplex. Let $q_U^{(-1)}(i, j)$ the (i, j) -th entry of the matrix Q_U^{-1} . Denote by $\mathcal{K}$ the image $L(\mathcal{M}_1) \subset \mathbb{R}^m$. Define the map $G_U : \mathcal{M}_1 \rightarrow \mathbb{R}$

$$G_U(\mu) := \sup_{\mathbf{z} \in \mathcal{K} \cap \mathcal{N}(Q_U)} \sum_{k=1}^m \mu_k \log \sum_{i=1}^m z_i q_U^{(-1)}(i, k).$$

Proposition 5. For $\mu \notin \mathcal{M}_1$ we have $I(\mu) = \infty$, and if $\mu \in \mathcal{M}_1$ then $I_U(\mu) \geq G_U(\mu) - (m-1)$.

Proof. Notice that

$$\begin{aligned} \Lambda_j^{(U)}(\lambda) &= \mathbb{E} \left[\exp \left\{ \sum_{k=1}^m \lambda_k \mathbb{1}_{\{\varepsilon_1 \in U_k - x_j\}} \right\} \right] \\ &= \sum_{k=1}^m e^{\lambda_k + m-1} \mathbb{P}(\varepsilon_1 \in U_k - x_j) = e^{m-1} \sum_{k=1}^m e^{\lambda_k} q_{k,j}^{(U)}. \end{aligned}$$

Fix $\mathbf{t}$ such that $L(\mathbf{t}) \in \mathcal{N}(Q_U)$. For a vector $\mathbf{x} \gg 0$ denote by $\log(\mathbf{x})$ the vector whose j -th entry is $\log x_j$. The vector $\lambda^* := \log[e^{1-m}(L(\mathbf{t})Q_U^{-1})]$ satisfies $\Lambda_j(\lambda^*) = \frac{t_j}{\sum_{i=1}^m t_i P_{i,j}^*}$.

Now the argument is standard. In fact $\mathbf{t}$ satisfies $\mathbf{t}P_{\lambda^*} = \mathbf{t}$, which together with the fact $\mathbf{t} \gg 0$, implies $\rho(P_{\lambda^*}) = 1$. Hence

$$\begin{aligned} I(\mu) &= \sup_{\lambda \in \mathbb{R}^m} \{ \langle \lambda, z \rangle - \log \rho(P_\lambda) \} \geq \langle \lambda^*, \mu \rangle - \log \rho(P_{\lambda^*}) \\ &= \sum_{k=1}^m \mu_k \log \sum_{i=1}^m e^{1-m} L_i(\mathbf{t}) q_U^{(-1)}(i, j). \end{aligned}$$

It remains to take the suprema over the set of $\mathbf{t}$ satisfying $L(\mathbf{t}) \in \mathcal{N}(Q)$ to get our result. $\square$

Proposition 6. $G_U = J_U$.

Proof. Notice that $\mathbf{x}Q_U \in \mathcal{K}$ if and only if $\sum_{i=1}^m x_i Q_{i,k}^{(U)} = \frac{t_k}{\sum_{j=1}^m t_j P_{j,k}^*}$, which can be reformulated as $\mathbf{t}P^* \text{diag}(\mathbf{x}Q_U) = \mathbf{t}$. This is verified if and only if the matrix $P^* \text{diag}(\mathbf{x}Q_U)$ as eigenvalue 1, and is equivalent to $\det[P^* \text{diag}(\mathbf{x}Q_U) - I_m] = 0$. $\square$

Hence

$$\mathbb{P}(\hat{d}^{(n)}(U) \in \overline{B}_r^c) \leq e^{-n \inf_{\mu \in \overline{B}_r^c} (J_U(\mu) - (m-1) + o(1))}.$$

We get

$$\mathbb{P}(\pi^* \notin A_r) = \mathbb{P}(\hat{d}^{(n)}(U) \in \overline{B}_r^c) \leq e^{-n \sup_{U \in \mathcal{U}} \inf_{\mu \in \overline{B}_r^c} (J_U(\mu) - (m-1) + o(1))} \leq e^{-n(H_r + o(1))},$$

which proves Theorem 1.

4. Conclusions

We end this paper by considering few of the properties of the confidence interval. The coverage of this set is quite good, in fact, for any $p > 0$, there exists a constant $c(p)$ such that

$$\mathbb{P}(\pi^* \notin A_r) \leq \alpha + \frac{c(p)}{n^p}.$$

The advantage of using the large deviations approach stands in the fact that for fixed r , the probability that A_r does not contain π^* decreases exponentially fast as n increases. On the other hand, the measure of the confidence interval is less than $\|Q_U^{-1}\|r$, where $\|\cdot\|$ here denotes the usual matrix norm, and r is the ray of B_r defined in the introduction.

References

- [CMR05] O. CAPPÉ, E. MOULINES and T. RYDÉN, Inference in Hidden Markov Models, Springer, (2005).
- [DZ98] A. DEMBO and O. ZEITOUNI, Large deviations techniques and applications, Springer, (1998).
- [VRM00] I. VISSER, M. RAIJMAKERS and P. MOLENAAR, Confidence intervals for Hidden Markov models parameters, British Journal of Mathematical and Statistical Psychology, (2000), **53**, 317 – 327.