Weak law for Markov model estimation using the IDM

Erik Quaeghebeur

March 31, 2004

Contents

1 Introduction

In an as of yet still unfinished report [1] it is shown that a generalized version of the imprecise Dirichlet model [2] or IDM can be used for learning/estimating the transition probabilities of a Markov model (or MM). The estimations are based on an *observation* of a sequence of states $x_0x_1\cdots x_n$, or equivalently, of a sequence of state n *transitions* x_kx_{k+1} from the unknown MM.

We use a *transition matrix* Θ to describe a MM. We limit ourselves to the case of *regular* or *ergodic* MM's. For these there exists an l_0 for which $(\Theta^l)_{ij} > 0$, for all $l > l_0$ and for all transitions ij. In addition, they have a unique largest (in module) eigenvalue 1, for all other eigenvalues λ of Θ it holds that $|\lambda| < 1$.

In this report we are going to prove that the estimate of Θ using the model described model using the IDM [1] is consistent. This means that we can make the probability that the distance between the estimate and the real Θ exceeds some arbitrary constant arbitrarily small when n can be made arbitrarily large.

1.1 Notation

We will use the same notation here as is introduced in the aforementioned report [1]. The most important quantities are

- the (unknown) transition matrix Θ with elements $(\Theta)_{ij} = \theta_{j|i}$;
- the (unknown) l-step transition matrix Θ^l with elements $(\Theta^l)_{ij} = \theta^{(l)}_{j|i}$;
- the matrix of observed transitions N with elements n_{ij} an total sum n;
- equilibrium distribution $\boldsymbol{\rho}$, which is the left eigenvector with eigenvalue 1 of Θ: $\boldsymbol{\rho}^T\Theta = \boldsymbol{\rho}^T$;
- a parameter vector $\boldsymbol{t}^{(n)}_{\cdot|i}$ of the learning model (a row of a matrix $T^{(n)}$) with components

$$t^{(n)}_{j|i} = \frac{h_i t^{(0)}_{j|i} + n_{ij}}{h_i + n_{i+}},$$

  where h_i is the so-called initial strength for state i and n_{i+} the sum of the elements of the i-th row of N.

2 The estimate

The model consists of a product of sets of Dirichlet distributions (one set for each state) with initial parameters h_i and $\boldsymbol{t}^{(0)}_{\cdot|i}$, which takes on all values of the interior of the unit simplex.

The expectation of each row $\boldsymbol{\theta}_{\cdot|i}$ of Θ after observing N becomes the set of vectors $\boldsymbol{t}^{(n)}_{\cdot|i}$. These sets constitute an estimate of Θ.

We could now define a distance between each of these sets and their corresponding row of Θ. The total distance is just the sum of these separate distances.

We shall not be so explicit, but instead prove that any one $\boldsymbol{t}^{(n)}_{\cdot|i}$ is a consistent estimate, independent of the $\boldsymbol{t}^{(0)}_{\cdot|i}$ chosen. It is clear that this will then also hold for any reasonable choice of distance between $\boldsymbol{\theta}_{\cdot|i}$ and its estimate, the expectation under our model.

3 Putting bounds on the distance

In this section we will give an upper bound of the distance between $\boldsymbol{\theta}_{\cdot|i}$ and its estimate that is the same for any specific element of the unit simplex.

Because the total distance between Θ and its estimate is just the sum of the distances between $\boldsymbol{\theta}_{\cdot|i}$ and its estimate over all states i, it is sufficient to treat one arbitrary state i. We can consequently drop the index i further on to alleviate the notation.

We want to prove that

$$\forall \varepsilon > 0, \forall \boldsymbol{t} : \lim_{n\to\infty} P(d(\frac{h\boldsymbol{t}+\boldsymbol{n}}{h+n_+}, \boldsymbol{\theta}) > \varepsilon) = 0, \tag{1}$$

where we have set $\boldsymbol{t} = \boldsymbol{t}^{(0)}$ and d denotes a usual distance between two vectors.

Because norms are convex, we can write

$$d(\frac{h\boldsymbol{t}+\boldsymbol{n}}{h+n_+}, \boldsymbol{\theta}) \leq \frac{h}{h+n_+} d(\boldsymbol{t}, \boldsymbol{\theta}) + \frac{n_+}{h+n_+} d(\frac{\boldsymbol{n}}{n_+}, \boldsymbol{\theta}).$$

Here we assume that $n_+ > 0$, which later on we will show is acceptable. Distances between elements of the unit simplex such as $\boldsymbol{t}$ and $\boldsymbol{\theta}$ are bounded (the unit simplex for a finite number of states is a bounded set) and are smaller than the maximal distance between any two elements of the simplex, which by an appropriate scaling can be chosen to be 1. By additionally making the numerator h larger we find that

$$d(\frac{h\boldsymbol{t}+\boldsymbol{n}}{h+n_+}, \boldsymbol{\theta}) \leq \frac{h}{h+n_+} + d(\frac{\boldsymbol{n}}{n_+}, \boldsymbol{\theta}).$$

It follows that if we prove that

$$\forall \varepsilon > 0 : \lim_{n\to\infty} P(\frac{h}{h+n_+} + d(\frac{\boldsymbol{n}}{n_+}, \boldsymbol{\theta}) > \varepsilon) = 0, \tag{2}$$

then (1) is also proven.

4 Divide et impera

4.1 The splitting implication

It is useful to note that, when a_1 and a_2 are positive, the following *splitting implication* holds (the proof is immediate by looking at the contrapositive):

$$\forall \varepsilon_1, \varepsilon_2 > 0 : \begin{cases} \lim\limits_{n\to\infty} P(a_1(\boldsymbol{n}) > \varepsilon_1) = 0 \\ \lim\limits_{n\to\infty} P(a_2(\boldsymbol{n}) > \varepsilon_2) = 0 \end{cases} \Rightarrow \forall \varepsilon > 0 : \lim_{n\to\infty} P(a_1(\boldsymbol{n}) + a_2(\boldsymbol{n})) > \varepsilon) = 0.$$

Because of this implication, (2) is proven when

$$\forall \varepsilon > 0 : \lim_{n\to\infty} P(\frac{h}{h+n_+} > \varepsilon) = 0, \tag{3}$$

and

$$\forall \varepsilon > 0 : \lim_{n\to\infty} P(d(\frac{\boldsymbol{n}}{n_+}, \boldsymbol{\theta}) > \varepsilon) = 0, \tag{4}$$

are proven.

4.2 Further reductionism

When we look at (3), we see (under the assumption that $n_+ > 0$) that

$$\frac{h}{h+n_+} < \frac{h}{n_+} = \frac{h}{n_+} - \frac{h}{n\rho} + \frac{h}{n\rho} \leq \left|\frac{h}{n_+} - \frac{h}{n\rho}\right| + \frac{h}{n\rho}.$$

This means that (3) is proven when

$$\forall \varepsilon > 0 : \lim_{n\to\infty} P(\left|\frac{h}{n_+} - \frac{h}{n\rho}\right| > \varepsilon) = 0, \tag{5}$$

$$\forall \varepsilon > 0 : \lim_{n\to\infty} P(\frac{h}{n\rho} > \varepsilon) = 0 \tag{6}$$

are proven.

When we look at (4), we see because of the triangle inequation that

$$\begin{aligned} d(\frac{\boldsymbol{n}}{n_+}, \boldsymbol{\theta}) &\leq d(\frac{\boldsymbol{n}}{n_+}, \frac{\boldsymbol{n}}{n\rho}) + d(\frac{\boldsymbol{n}}{n\rho}, \boldsymbol{\theta}) \\ &= d(\frac{\boldsymbol{n}}{n_+}, \frac{n_+}{n\rho}\frac{\boldsymbol{n}}{n_+}) + \frac{1}{n\rho} d(\boldsymbol{n}, n\rho\boldsymbol{\theta}) \\ &\leq \left|1 - \frac{n_+}{n\rho}\right| + \frac{1}{n\rho} d(\boldsymbol{n}, n\rho\boldsymbol{\theta}), \end{aligned}$$

where the last inequality follows from the fact that the distance between the origin and an element of the simplex is smaller than or equal to the maximal distance between two elements of the simplex. This means that (4) is proven when

$$\forall \varepsilon > 0 : \lim_{n\to\infty} P(\frac{1}{n\rho}\left|n\rho - n_+\right| > \varepsilon) = 0, \tag{7}$$

$$\forall \varepsilon > 0 : \lim_{n\to\infty} P(\frac{1}{n\rho} d(\boldsymbol{n}, n\rho\boldsymbol{\theta}) > \varepsilon) = 0 \tag{8}$$

are proven.

5 Bounding the probability

5.1 Investigating (6)

It is clear that

$$\forall \varepsilon > 0 : P(\frac{h}{n\rho} > \varepsilon) = \begin{cases} 0, & n > \frac{h}{\rho}\frac{1}{\varepsilon}, \\ 1, & n \leq \frac{h}{\rho}\frac{1}{\varepsilon}, \end{cases} \tag{9}$$

so (6) is trivially satisfied.

5.2 Investigating (7)

To get a bound on the probability in (7), we will follow a proof by Kemeny and Snell [3, Theorem 4.2.1]. To start, we will use Tchebycheff's theorem (where E denotes expectation) [4, 15.7]

$$\begin{aligned} P(\frac{1}{n\rho}\left|n\rho - n_+\right| > \varepsilon) &< \frac{1}{\varepsilon}E(\frac{1}{n\rho}\left|n\rho - n_+\right|) \\ &= \frac{1}{n\rho\varepsilon}E(\sqrt{[n_+ - n\rho]^2}) \\ &\leq \frac{1}{n\rho\varepsilon}\sqrt{E([n_+ - n\rho]^2)}, \end{aligned}$$

where the last inequality follows from the fact that a variance is always positive (it also follows from Jensen's inequality [5, 6.3.5]).

We suppose that the initial state is fixed, i.e., $x_0 = u$. We see that (we will again add the index denoting the state in the following derivation)

$$E([n_{i+} - n\rho_i]^2) = E([\sum_{k=1}^{n}[\delta_{x_k i} - \rho_i]]^2) = \sum_{k=1}^{n}\sum_{l=1}^{n}\left[E(\delta_{x_k i}\delta_{x_l i}) - \rho_i E(\delta_{x_k i} + \delta_{x_l i}) + \rho_i^2\right],$$

where δ is the Kronecker-delta. Working out the expectations, we get

$$E([n_{i+} - n\rho_i]^2) = \sum_{k=1}^{n}\sum_{l=1}^{n}\left[\theta_{i|u}^{(\min\{k,l\})}\theta_{i|u}^{(|k-l|)} - \rho_i[\theta_{i|u}^{(k)} + \theta_{i|u}^{(l)}] + \rho_i^2\right],$$

Kemeny and Snell have proved [3, Corollary 4.1.5] that for regular MM $\theta_{i|u}^{(k)} = \rho_i + e_\tau(k)$, with $|e_\tau(k)| \leq \tau^{k-\hat{n}}$, where $0 < \tau < 1$ and $\hat{n}$ is the first integer such that $\Theta^{\hat{n}}$ contains only positive components. This allows us to write

$$\begin{aligned} &E([n_{i+} - n\rho_i]^2) \\ &= \sum_{k=1}^{n}\sum_{l=1}^{n}\left[[\rho_i + \varepsilon_\tau(\min\{k,l\})][\rho_i + \varepsilon_\tau(|k-l|)] - \rho_i[\rho_i + \varepsilon_\tau(k) + \rho_i + \varepsilon_\tau(l)] + \rho_i^2\right] \\ &\leq \frac{1}{\tau^{\hat{n}}}\sum_{k=1}^{n}\sum_{l=1}^{n}\left[\tau^{\max\{k,l\}} + \rho_i[\tau^{|k-l|} + \tau^{\min\{k,l\}}] + \rho_i[\tau^k + \tau^l]\right] \\ &= 2n\rho_i\frac{\tau}{\tau^{\hat{n}}}\frac{1-\tau^n}{1-\tau} + \frac{1}{\tau^{\hat{n}}}\left[\sum_{k=1}^{n-1}\sum_{l=k+1}^{n} + \sum_{k=l=1}^{n} + \sum_{k=2}^{n}\sum_{l=1}^{k-1}\right]\left[\tau^{\max\{k,l\}} + \rho_i[\tau^{|k-l|} + \tau^{\min\{k,l\}}]\right]. \end{aligned}$$

Because $\tau^k\frac{1-\tau^l}{1-\tau} \leq \frac{1}{1-\tau} = c \in \mathbb{R}$ for all $k \geq 0$ and because $n - 1 \leq n$ we find

$$E([n_{i+} - n\rho_i]^2) \leq 2n\rho_i\frac{c}{\tau^{\hat{n}}} + \frac{1}{\tau^{\hat{n}}}\left[nc + \sum_{k=2}^{n}[k-1]\tau^k + \rho_i[4nc + \sum_{k=1}^{n-1}[n-k]\tau^k]\right].$$

By replacing the coefficients $k - 1$ and $n - k$ by n we find

$$E([n_{i+} - n\rho_i]^2) \leq n\rho_i\frac{c}{\tau^{\hat{n}}}[\frac{2}{\rho_i} + 7] = n\rho_i O(1).$$

Combining this result with Tchebycheff's theorem, we find that

$$\forall \varepsilon > 0 : P(\frac{1}{n\rho}\left|n\rho - n_+\right| > \varepsilon) < \frac{1}{\sqrt{n\rho}}\frac{O(1)}{\varepsilon}, \tag{10}$$

so that (7) is clearly true. This result can be used to show that n_+ can be assumed to be non-zero. Consider that

$$P(\frac{1}{n\rho}\left|n\rho - n_+\right| > \varepsilon) =$$
$$P(\frac{1}{n\rho}\left|n\rho - n_+\right| > \varepsilon \mid n_+ = 0)P(n_+ = 0) + P(\frac{1}{n\rho}\left|n\rho - n_+\right| > \varepsilon \mid n_+ > 0)P(n_+ > 0).$$

Because the values of ε we're interested in are small we can assume that $P(\frac{1}{n\rho}\left|n\rho - n_+\right| > \varepsilon \mid n_+ = 0) = P(1 > \varepsilon) = 1$, which implies $P(n_+ = 0) < \frac{1}{\sqrt{n\rho}}\frac{O(1)}{\varepsilon}$.

5.3 Investigating (8)

To get a bound on the probability in (8) we will again use Tchebycheff's theorem

$$\begin{aligned} P(\frac{1}{n\rho}d(\boldsymbol{n}, n\rho\boldsymbol{\theta}) > \varepsilon) &< \frac{1}{\varepsilon}E(\frac{1}{n\rho}d(\boldsymbol{n}, n\rho\boldsymbol{\theta})) \\ &= \frac{1}{n\rho\varepsilon}E(d(\boldsymbol{n}, n\rho\boldsymbol{\theta})). \end{aligned}$$

When we take d to be the Euclidean distance, we get

$$\begin{aligned} E(d(\boldsymbol{n}, n\rho\boldsymbol{\theta})) &= E(\sqrt{\sum_j \left[n_j^2 + [n\rho\theta_j]^2 - 2n_j n\rho\theta_i\right]}) \\ &\leq \sqrt{\sum_j \left[E(n_j^2) + [n\rho\theta_j]^2 - 2E(n_j)[n\rho\theta_j]\right]}, \end{aligned}$$

where we've applied Jensen's inequality again.

We can rewrite the above as follows,

$$E(d(\boldsymbol{n}, n\rho\boldsymbol{\theta}))^2 \leq \sum_j \left[V(n_j) + E(n_j)^2 + [n\rho\theta_j]^2 - 2E(n_j)[n\rho\theta_j]\right].$$

From Martin [6, Theorems 6.1.2, 6.1.3] we get expressions for the expectation and variance of the number of observed transitions ij,

$$\begin{aligned} E(n_j) &= n\rho\theta_j + a_j, \\ V(n_j) &= n\rho\theta_j[1 - \rho\theta_j + 2a_j] + b_j, \end{aligned}$$

where $a_j \in O(|1 - \lambda^n|)$ and $b_j \in O(|1 - \lambda^n| + n\,|\lambda|^n)$, with λ an eigenvalue of Θ with module smaller than 1. Using the above expressions, we get

$$E(d(\boldsymbol{n}, n\rho\boldsymbol{\theta}))^2 \leq n\rho\sum_j \theta_j[1 - \rho\theta_j + 2a_j] + \sum_j [a_j^2 + b_j].$$

From which follows

$$E(d(\boldsymbol{n}, n\rho\boldsymbol{\theta})) \leq \sqrt{n\rho}O(\sqrt{|1-\lambda^n|}) < \sqrt{n\rho}O(1).$$

Combining this result with Tchebycheff's theorem, we find that

$$\forall\varepsilon > 0 : P(\frac{1}{n\rho}d(\boldsymbol{n}, n\rho\boldsymbol{\theta}) > \varepsilon) < \frac{1}{\sqrt{n\rho}}\frac{O(1)}{\varepsilon}, \tag{11}$$

so that (8) is clearly true.

5.4 Investigating (5)

Equation (10) implies

$$P(n_+\left|\frac{1}{n_+} - \frac{1}{n\rho}\right| > \frac{\varepsilon}{h}) < \frac{1}{\sqrt{n\rho}}\frac{O(1)}{\varepsilon/h},$$

which in turn implies

$$\forall\varepsilon > 0 : P(\left|\frac{h}{n_+} - \frac{h}{n\rho}\right| > \varepsilon) < \frac{1}{\sqrt{n\rho}}\frac{O(1)}{\varepsilon}, \tag{12}$$

because we can assume $n_+ \geq 1$. It is immediately clear that (5) is true.

6 Combining the bounds

In the last section we have proved (1). We can say more however. Equations (9), (10), (11) and (12) allow us to specify how fast the probability that the estimates differs from the true transition probabilities diminishes. If we combine them, we get

$$\forall\varepsilon > \frac{h}{n\rho} : P(d(T^{(n)}, \Theta) > \varepsilon) < \frac{1}{\sqrt{n\rho}}\frac{O(1)}{\varepsilon}. \tag{13}$$

References

[1] Erik Quaeghebeur. IDM-variants for inference in markov models. Technical report, Onderzoeksgroep SYSTeMS. 1

[2] Peter Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society* B, 58(1):3–57, 1996. 1

[3] John G. Kemeny and J. Laurie Snell. *Finite Markov Chains*. Springer-Verlag, New York, 1976. 4

[4] Harald Cramér. *Mathematical methods in statistics*. Princeton University press, Princeton, 1966. 4

[5] Robert B. Ash and Catherine A. Doléans-Dade. *Probability and Measure Theory*. Academic Press, San Diego, 2000. 4

[6] J. J. Martin. *Bayesian Decision Problems and Markov Chains*. Number 13 in Publications in Operations Research. John Wiley & Sons, New York, 1966. 5