# Note: Maximum Likelihood Estimation for Markov Chains

36-462, Spring 2009

29 January 2009
To accompany lecture 6

This note elaborates on some of the points made in the slides.

## Contents



## 1 Derivation of the MLE for Markov chains

To recap, the basic case we're considering is that of a Markov chain $X_1^\infty$ with $m$ states. The transition matrix, $p$, is unknown, and we impose no restrictions on it, but rather want to estimate it from data. The parameters we wish to infer are thus the $m^2$ matrix entries $p_{ij}$, which are defined as

$$p_{ij} = \Pr\left(X_{t+1} = j | X_t = i\right) \tag{1}$$

What we observe is a sample from the chain, $x_1^n \equiv x_1, x_2, \ldots x_n$. This is a realization of the random variable $X_1^n$.

The probability of this realization is

$$\begin{aligned}\Pr\left(X_1^n = x_1^n\right) &= \Pr\left(X_1 = x_1\right) \prod_{t=2}^{n} \Pr\left(X_t = x_t | X_1^{t-1} = x_1^{t-1}\right) && (2)\\ &= \Pr\left(X_1 = x_1\right) \prod_{t=2}^{n} \Pr\left(X_t = x_t | X_{t-1} = x_{t-1}\right) && (3)\end{aligned}$$

The first line just uses the definition of conditional probability, but the second line actually uses the Markov property — that the future is independent of the past, given the present.

Re-write in terms of the transition probabilities $p_{ij}$, to get the likelihood of a given transition matrix:

$$L(p) = \Pr\left(X_1 = x_1\right) \prod_{t=2}^{n} p_{x_{t-1} x_t} \tag{4}$$

Define the transition counts $N_{ij} \equiv$ number of times $i$ is followed by $j$ in $X_1^n$, and re-write the likelihood in terms of them.

$$L(p) = \Pr\left(X_1 = x_1\right) \prod_{i=1}^{k} \prod_{j=1}^{k} p_{ij}^{n_{ij}} \tag{5}$$

We want to maximize this as a function of the $p_{ij}$. This happens much as it would in a problem with IID data.

First, take the log so we're dealing with a sum, not product, and derivatives will be easier.

First, take logs to simplify optimization

$$\mathcal{L}(p) = \log L(p) = \log \Pr\left(X_1 = x_1\right) + \sum_{i,j} n_{ij} \log p_{ij} \tag{6}$$

Take the derivative:

$$\frac{\partial \mathcal{L}}{\partial p_{ij}} = \frac{n_{ij}}{p_{ij}} \tag{7}$$

Set it equal to zero at $\hat{p}_{ij}$:

$$\frac{n_{ij}}{p_{ij}} = 0 \tag{8}$$

Conclude that all estimated transition probabilities should be $\infty$.

What's gone wrong?

We have failed to really come to grips with the parameters. They can't all change arbitrarily, because the probabilities of making transitions *from* a state have to add up to 1. That is, for each $i$,

$$\sum_j p_{ij} = 1 \tag{9}$$

This means that the number of degrees of freedom for transition matrices is not $m^2$, but $m(m-1)$.

There are at lest two ways of handling this: explicitly eliminating parameters, and using Lagrange multipliers to enforce constraints.

## 1.1 Eliminating parameters

Arbitrarily pick one of the transition probabilities to express in terms of the others. Say it's the probability of going to 1, so for each $i$, $p_{i1} = 1 - \sum_{j=2}^{m} p_{ij}$. Now when we take derivatives of the likelihood, we leave out $\partial/\partial P_{i1}$, and the other terms have changed:

$$\frac{\partial \mathcal{L}}{\partial p_{ij}} = \frac{n_{ij}}{p_{ij}} - \frac{n_{i1}}{p_{i1}} \tag{10}$$

Setting this equal to zero at the MLE $\hat{p}$,

$$\frac{n_{ij}}{\hat{p}_{ij}} = \frac{n_{i1}}{\hat{p}_{i1}} \tag{11}$$

$$\frac{n_{ij}}{n_{i1}} = \frac{\hat{p}_{ij}}{\hat{p}_{i1}} \tag{12}$$

Since this holds for all $j \neq 1$, we can conclude that $\hat{p}_{ij} \propto n_{ij}$, and in fact

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{m} n_{ij}} \tag{13}$$

Clearly, the choice of $P_{i1}$ as the transition probability to eliminate in favor of the others is totally arbitrary and we get the same result for any other.

## 1.2 Lagrange multipliers

If you do not already know about Lagrange multipliers, I am not going to try to explain them here; the basic story is that they are a way of doing optimization under constraints without explicitly using the constraints to reduce the parameter space. Lots of applied math/math methods books have good discussions (e.g. [1]); there is also a great on-line tutorial [2].

We have $m$ constraint equations,

$$\sum_{j} p_{ij} = 1 \tag{14}$$

one for each value of $i$. This means we need $m$ Lagrange multipliers, $\lambda_1, \lambda_2, \ldots \lambda_m$. The new objective function is

$$\mathcal{L}(P) - \sum_{i=1}^{j} \lambda_i \left( \sum_{j} p_{ij} - 1 \right) \tag{15}$$

Taking derivatives with respect to $\lambda_i$ of course gives the $i^{\text{th}}$ constraint equation back. Taking derivatives with respect to $p_{ij}$,

$$\begin{aligned} 0 &= \frac{n_{ij}}{p_{ij}} - \lambda_i \\ \lambda_i &= \frac{n_{ij}}{p_{ij}} \\ p_{ij} &= \frac{n_{ij}}{\lambda_i} \end{aligned}$$

Because of the constraint equation,

$$\sum_{j=1}^{m} \frac{n_{ij}}{\lambda_i} = 1 \tag{16}$$

$$\sum_{j=1}^{m} n_{ij} = \lambda_i \tag{17}$$

leading to the same conclusion as we got by eliminating parameters outright.

# 2 Consistency of the MLE

We should really write estimates as functions of the data: $\hat{p}_{ij}(x_1^n)$. So our estimate is really a realization of a random variable:

$$\hat{P}_{ij} = \hat{p}_{ij}(X_1^n) = \frac{N_{ij}}{\sum_j N_{ij}} \tag{18}$$

We want this to converge on the true probability, $p_{ij}^0$, so let's look at the convergence of the parts.

$N_{ij}$ can be written as a sum of indicator functions:

$$N_{ij} = \sum_{t=1}^{n-1} \mathbb{I}_{x=i}(x_t)\mathbb{I}_{x=j}(x_{t+1}) \tag{19}$$

Hence $N_{ij}/(n-1)$ is the time average of an indicator function:

$$\frac{N_{ij}}{n-1} = \frac{1}{n-1}\sum_{t=1}^{n-1} \mathbb{I}_{x=i}(x_t)\mathbb{I}_{x=j}(x_{t+1}) \tag{20}$$

We know from the ergodic theorem (see notes to lecture 2) that time-averages converge to expectations:

$$\frac{N_{ij}}{n-1} \to \mathrm{E}\left(\mathbb{I}_{x=i}(x_t)\mathbb{I}_{x=j}(x_{t+1})\right) \tag{21}$$

and we know that the expectation of an indicator function is a probability:

$$\mathrm{E}\left(\mathbb{I}_{x=i}(x_t)\mathbb{I}_{x=j}(x_{t+1})\right) = \Pr\left(X_t = i, X_{t+1} = j\right) \tag{22}$$

If we say that $p_i^0$ is the true long-run probability that $X_t = i$, we have

$$\Pr\left(X_t = i, X_{t+1} = j\right) = p_i^0 p_{ij}^0 \tag{23}$$

since $p_{ij}^0$ is a conditional probability. Put this together and we get

$$\frac{N_{ij}}{n-1} \to p_i^0 p_{ij}^0 \tag{24}$$

and of course $N_{ij}/n$ converges to the same limit (which was what was given in the lecture).

Now think about $\sum_j N_{ij}$. If use 19, we get that

$$\sum_j N_{ij} = \sum_{t=1}^{n-1} \mathbb{I}_{x=i}(X_t) \tag{25}$$

Pulling the same trick of dividing by $n-1$,

$$\frac{1}{n-1}\sum_j N_{ij} = \frac{1}{n-1}\sum_{t=1}^{n-1} \mathbb{I}_{x=i}(X_t) \to \mathrm{E}\left(\mathbb{I}_{x=i}(X_t)\right) \tag{26}$$

So, since again the expectation of an indicator is a probability,

$$\frac{1}{n-1}\sum_j N_{ij} \to p_i^0 \tag{27}$$

Now the estimator is

$$\hat{P}_{ij} = \frac{N_{ij}}{\sum_j N_{ij}} = \frac{N_{ij}/(n-1)}{\sum_j N_{ij}/(n-1)} \tag{28}$$

The numerator converges to $p_i^0 p_{ij}^0$, the denominator to $p_i^0$; and since that is $> 0$, the ratio converges to the ratio of the limits,

$$\hat{P}_{ij} \to \frac{p_i^0 p_{ij}^0}{p_i^0} = p_{ij}^0 \tag{29}$$

as desired, and claimed.[1]

[1] If you have taken more advanced probability classes, you might wonder about the mode of convergence here. Irreducible Markov chains satisfy the Birkhoff "individual" ergodic theorem, which gives convergence almost surely. (See e.g. [3].) We need a ratio of two a.s. convergent quantities to converge, and we can get that to happen by e.g. the almost-sure version of Slutsky's Theorem [4, p. 42].

# 3 Alternate forms of the Fisher information

There are three forms of the Fisher information for Markov chains:

$$I_{uv}(\theta) = -\sum_{ij} p_i(\theta) p_{ij}(\theta) \frac{\partial^2 \log p_{ij}(\theta)}{\partial \theta_u \partial \theta_v} \tag{30}$$

$$I_{uv}(\theta) = \sum_{ij} \frac{p_i(\theta)}{p_{ij}(\theta)} \frac{\partial p_{ij}(\theta)}{\partial \theta_u} \frac{\partial p_{ij}(\theta)}{\partial \theta_v} \tag{31}$$

and

$$I_{uv}(\theta) = -\sum_{ij} p_i(\theta) p_{ij}(\theta) \frac{\partial \log p_{ij}(\theta)}{\partial \theta_u} \frac{\partial \log p_{ij}(\theta)}{\partial \theta_v} \tag{32}$$

To see that these are equivalent, start with the first, and use the chain rule for derivatives to transform it into the second.

$$\begin{aligned}
-I_{uv}(\theta) &= \sum_{ij} p_i(\theta) p_{ij}(\theta) \frac{\partial^2 \log p_{ij}(\theta)}{\partial \theta_u \partial \theta_v} && (33)\\
&= \sum_{ij} p_i(\theta) p_{ij}(\theta) \frac{\partial}{\partial \theta_u} \frac{\partial \log p_{ij}(\theta)}{\partial \theta_v} && (34)\\
&= \sum_{ij} p_i(\theta) p_{ij}(\theta) \frac{\partial}{\partial \theta_u} p_{ij}^{-1}(\theta) \frac{\partial p_{ij}(\theta)}{\partial \theta_v} && (35)\\
&= \sum_{ij} p_i(\theta) p_{ij}(\theta) \left[ p_{ij}^{-1}(\theta) \frac{\partial^2 p_{ij}(\theta)}{\partial \theta_u \partial \theta_v} - p_{ij}^{-2}(\theta) \frac{\partial p_{ij}(\theta)}{\partial \theta_u} \frac{\partial p_{ij}(\theta)}{\partial \theta_v} \right] && (36)\\
&= -\sum_{ij} \frac{p_i(\theta)}{p_{ij}(\theta)} \frac{\partial p_{ij}(\theta)}{\partial \theta_u} \frac{\partial p_{ij}(\theta)}{\partial \theta_v} + \sum_i \frac{\partial^2}{\partial \theta_u \partial \theta_v} \sum_j p_{ij}(\theta) && (37)\\
&= -\sum_{ij} \frac{p_i(\theta)}{p_{ij}(\theta)} \frac{\partial p_{ij}(\theta)}{\partial \theta_u} \frac{\partial p_{ij}(\theta)}{\partial \theta_v} + \sum_i \frac{\partial^2}{\partial \theta_u \partial \theta_v} 1 && (38)\\
&= -\sum_{ij} \frac{p_i(\theta)}{p_{ij}(\theta)} \frac{\partial p_{ij}(\theta)}{\partial \theta_u} \frac{\partial p_{ij}(\theta)}{\partial \theta_v} && (39)
\end{aligned}$$

To see the equivalence between the second form and the third, note that

$$\frac{\partial \log p_{ij}(\theta)}{\partial \theta_u} \frac{\partial \log p_{ij}(\theta)}{\partial \theta_v} = \frac{1}{p_{ij}^2(\theta)} \frac{\partial p_{ij}(\theta)}{\partial \theta_u} \frac{\partial p_{ij}(\theta)}{\partial \theta_v} \tag{40}$$

and substitute into the previous equation.

# 4 Markov Chains Form Exponential Families

This section will only make sense if you already know what an exponential family is.

Let's look again at the equation for the log-likelihood, Eq. 6:

$$\mathcal{L}(P) = \log \Pr(X_1 = x_1) + \sum_{i,j} n_{ij} \log P_{ij} \tag{41}$$

This is the equation for the log-likelihood of an exponential family, in which the canonical sufficient statistics are the $n_{ij}$ and $x_1$, and the natural parameters are the $\log P_{ij}$ and the log probabilities of the initial states. If we ignore the initial distribution, or condition it away, or let it be separate from the transition probabilities, then the maximum-likelihood estimators follow straightforwardly from the usual exponential family manipulations. If, on the other hand, we connect the initial distribution to the transition matrix, say by making the initial distribution the invariant distribution, then we really have a *curved* exponential family.

If the true transition matrix is $P^0$, with corresponding invariant distribution $p_i^0$, then one can show, with the ergodic theorem, that

$$\frac{1}{n}\mathcal{L}(P) \to \sum_i p_i^0 \sum_j p_{ij}^0 \log p_{ij} \tag{42}$$

so the error in the log-likelihood introduced by ignoring the first term, $\log \Pr(X_1 = x_1)$, shrinks proportionately to zero.

# 5 Stochastic Finite Automata

We will deal only with machines where the current state and the next symbol uniquely fix the next state. (These are generally, but unfortunately, called "deterministic" by computer scientists. Other names, less confusing for us, are "resolving" and "recursively updating".)

Specifically, let's say we have $k$ states and $m$ symbols. The probability of state $i$ emitting symbol $j$ will be $Q_{ij}$. Finally, the matrix $T_{ij}$ gives us the state reached from state $i$ on symbol $j$.

Assume we observe the sequence $x_1^n$. If we knew the starting state was $s_1$, the corresponding sequence of states would be $s_1^n$, where $s_{t+1} = T_{s_t x_t}$. Thus the conditional probability of the sequence $x_1^n$ is

$$\Pr(X_1^n = x_1^n | S_1 = s_1) = \prod_{t=1}^{n} Q_{s_t x_t} \tag{43}$$

and if we write $N_{ij}$ for the number of times state $i$ emitted symbol $j$, we get

$$\Pr(X_1^n = x_1^n | S_1 = s_1) = \prod_{i=1}^{k}\prod_{j=1}^{m} Q_{ij}^{N_{ij}} \tag{44}$$

Now the argument proceeds just as for the Markov chain, leading us to conclude that the MLE is

$$\hat{Q}_{ij} = \frac{N_{ij}}{\sum_j N_{ij}} \tag{45}$$

All of this is conditional on the starting state.

# References


[1] Mary L. Boas. *Mathematical Methods in the Physical Sciences*. Wiley, New York, second edition, 1983.

[2] Dan Klien. Lagrange multipliers without permanent scarring. Online tutorial, 2001. URL `http://dbpubs.stanford.edu:8091/~klein/lagrange-multipliers.pdf`.

[3] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, Oxford, 2nd edition, 1992.

[4] Thomas S. Ferguson. *A Course in Large Sample Theory*. Texts in Statistical Science. CRC Press, Boca Raton, Florida, 1996.