STAT 516: Stochastic Modeling of Scientific Data **Autumn 2018**

Lecture 6: Inference on Markov Chain

Instructor: Yen-Chi Chen

These notes are partially based on those of Mathias Drton.

6.1 Introduction

We have learned a lot about properties of a Markov chain – when we know the transition probability matrix, we can do a lot of analysis. In data analysis, we often do not know the transition probability matrix so we need to estimate it. As long as we have a transition probability matrix, we can use our knowledge about Markov chain to make inference. In this lecture, we will be thinking about how to recover the transition probability matrix and make related statistical inference.

Let $X_0, \cdots, X_n$ be a set of RVs denoting our data/observations. Now we will assume that these RVs form a Markov chain. Moreover, we assume that each observation $X_i \in S = \{1, 2, \cdots, s\}$. Namely, the state space is finite.

In both Frequentist and Bayesian approaches, the likelihood function plays a key role in inference. Let $\mathbf{P} = \{p_{ij}\}$ be the transition probability matrix and $\nu_i = P(X_0 = i)$ be the initial probability. The likelihood function can be written as

$$L_n(\nu, \mathbf{P}) = \nu_{X_0} \prod_{i=1}^{n} p_{X_{i-1}, X_i}.$$

Generally, we cannot estimate the initial distribution ν unless we observed several different Markov chains. So we will make our inference condition on X_0.

Recalled that the joint PMF can be written as

$$p(x_0, \cdots, x_1; \mathbf{P}, \nu) = p(x_1, \cdots, x_n; \mathbf{P}|x_0) \times p(x_0; \nu),$$

we then use the likelihood function

$$L_n(\mathbf{P}) = p(X_1, \cdots, X_n; \mathbf{P}|x_0) = \prod_{i=1}^{s} \prod_{j=1}^{s} p_{ij}^{n_{ij}},$$

where $n_{ij} = \sum_{k=0}^{n-1} I(X_k = i, X_{k+1} = j)$ is the number of transition from state i to state j.

6.2 Frequentist Approach

With the likelihood function, we first consider the Frequentist approach to estimate $\mathbf{P}$. We find the MLE of $\mathbf{P}$.

The log-likelihood function will be

$$\ell_n(\mathbf{P}) = \log L_n(\mathbf{P}) = \sum_{i=1}^{s} \sum_{j=1}^{s} n_{ij} \log p_{ij}.$$

To find the MLE, note that there are constraints on $\mathbf{P}$:

$$\sum_{j=1}^{s} p_{ij} = 1$$

for every $i = 1, \cdots, s$. These constraints only applies to each row of $\mathbf{P}$, so we can separate the maximization of $\ell_b(\mathbf{P})$ into s different maximizations:

$$\widehat{p}_i = \mathsf{argmax}_{\sum_{j=1}^{s} p_{ij}=1} \ell_n(p_{i1}, \cdots, p_{is}) = \mathsf{argmax}_{\sum_{j=1}^{s} p_{ij}=1} \sum_{j=1}^{s} n_{ij} \log p_{ij}.$$

It turns out that this is the MLE of the multinomial distribution so

$$\widehat{p}_{ij} = \frac{n_{ij}}{n_{i+}},$$

where $n_{i+} = \sum_{j=1}^{s} n_{ij}$ is the number of observations from starting at state i.

So now we have the MLE but does this MLE has the usual properties such as statistical consistency and asymptotic normality? We cannot directly apply our conventional statistical knowledge here because each transition from state i to state j may not be independent from each other. We need some theories of MLE under the Markov chain structure.

To derive the theory of MLE for Markov chain, we first introduce the concept of *snake chain* and state two useful lemmas. Let $\{X_n\}$ be a homogeneous Markov chain with a state space S and a transition probability matrix $\mathbf{P}$. Define $Y_n = (X_n, X_{n+1})$. Then $\{Y_n\}$ is called the snake chain and it is also a homogeneous Markov chain with a state space $S_Y = \{(i_0, i_1) \in S^2 : p_{i_0,i_1} > 0\}$.

Lemma 6.1 *The transition probability matrix of $\{Y_n\}$ has entries $q_{(i,j),(k,\ell)} = p_{k\ell} I(j = k)$. In addition, if $\{X_n\}$ is irreducible, so is $\{Y_n\}$. Moreover, if $\{X_n\}$ has a stationary distribution π, then $\{Y_n\}$ has a stationary distribution ν with $\nu_{ij} = \pi_i p_{ij}$.*

Going back to our data analysis problem, the snake chain provides a useful framework for analyzing the quantity

$$n_{ij} = \sum_{k=1}^{n} I(X_{k-1} = i, X_k = j) = \sum_{k=1}^{n} I(Y_{k-1} = (i, j))$$

can be written as summation of indicator function with snake chains.

Another problem we need to address in our MLE is that the 'sample size' for estimating p_{ij} is n_{i+} (see the denominator), which is also random. So the conventional central limit theorem cannot be applied. Here we introduce another useful lemma (sometimes it is called Anscombe Lemma) for handling this case.

Lemma 6.2 *Suppose $Y_1, Y_2, \cdots$ are IID with $\mathbb{E}(Y_1) = 0$ and $\mathsf{Var}(Y_1) = \sigma^2 < \infty$. Let $S_n = \sum_{i=1}^{n} Y_i$ and suppose that $W_1, W_2, \cdots$ are random positive integers with $W_n/n \overset{P}{\to} c$ for some constant c. Then*

$$\frac{S_{W_n}}{\sqrt{\sigma^2 W_n}} \overset{D}{\to} N(0, 1).$$

Proof: See Anscombe (1952, Proceedings of the Cambridge Philosophical Society) or Theorem 11.6.1 in Shorack (2000, Probability for Statisticians). ■

Note that there are modified version of this lemma and is often related to the Poissonization techniques. It is particularly useful when we are trying to derive the central limit theory of an integrated error of a function estimation[1].

With these two lemmas, we can show that the MLE also works in Markov chain.

Proposition 6.3 *Let $\{X_n\}$ be an irreducible homogeneous Markov chain defined on a finite state space. Then the MLE satisfies*

1. Statistical consistency: $\widehat{p}_{ij} \overset{a.s.}{\rightarrow} p_{ij}$.
2. Asymptotic normality: $\frac{\widehat{p}_{ij}-p_{ij}}{\sqrt{p_{ij}(1-p_{ij})/(n\pi_i)}} \overset{D}{\rightarrow} N(0,1)$,

where $\pi = (\pi_1, \cdots, \pi_s)$ is the stationary distribution.

Before proceeding to the proof, we first compare this asymptotic normality to the one for multinomial distribution. In multinomial case, we have

$$\frac{\widehat{p}_i - p_i}{\sqrt{p_i(1-p_i)/n}} \overset{D}{\rightarrow} N(0,1).$$

Here you see that the difference is that the sample size changes from n to $n\pi_i$. So we can say that $n\pi_i$ behaves like the *effective sample size.* This quantity makes sense because we know that the stationary distribution characterizes the long run proportion of each state. The transition probability p_{ij} is estimated by the proportion of transiting from state i to state j over the the total number of transiting from state i. So the long run proportion of state i determines how much transition from state i we will observe.

Proof: *Consistency:*
We construct a snake chain $Y_n = (X_{n+1}, X_n)$ with a state space $S_Y = \{(i_0, i_1) \in S^2 : p_{i_0 i_1} > 0\}$. Using the snake chain lemma along with the fact that $\{X_n\}$ is irreducible, $\{Y_n\}$ is also irreducible with a transition probability matrix $\mathbf{P}_Y$ with element

$$P(Y_n = (i_n, j_n) \mid Y_{n-1} = (i_{n-1}, j_{n-1})) = p_{i_n j_n} I(i_n = j_{n-1})$$

and stationary distribution $\nu(i,j) = \pi_i p_{ij}$.

Then we apply the Ergodic theorem to $\{Y_n\}$ and $\{X_n\}$, leading to

$$\begin{aligned}\frac{n_{ij}}{n} &= \frac{1}{n}\sum_{k=0}^{n-1} I(Y_k = (i,j)) \overset{a.s.}{\rightarrow} \pi_i p_{ij}\\ \frac{n_{i+}}{n} &= \frac{1}{n}\sum_{k=0}^{n-1} I(X_k = i) \overset{a.s.}{\rightarrow} \pi_i.\end{aligned}$$

Thus,

$$\widehat{p}_{ij} = \frac{n_{ij}/n}{n_{i+}/n} \overset{a.s.}{\rightarrow} \frac{\pi_1 p_{ij}}{\pi_i} = p_{ij}.$$

Asymptotic normality:
We will use Anscombe's lemma to prove it. Here the trick is to identify the random variable Y_m and W_n.

[1] see, e.g., section 2 of Gine et. al. (2004) "The L_1-Norm Density Estimator Process" (https://projecteuclid.org/download/pdf_1/euclid.aop/1048516534)

When we focus on the state i, you can easily see that a natural choice of W_n is $W_n = n_{i+}$ since it serves as an effective sample size.

How about Y_m? We want to construct $S_{W_n} = S_{n_{i+}} = Y_1 + \cdots + Y_{n_{i+}}$ to be related to the denominator $\widehat{p}_{ij} - p_{ij} \propto n_{ij} - n_{i+}p_{ij}$. So it turns out that a possible choice is

$$Y_m = \begin{cases} 1 - p_{ij} & \text{if } X_{\tau_m+1} = j \\ -p_{ij} & \text{if } X_{\tau_m+1} \neq j \end{cases},$$

where τ_m is the m-th visit to state i. This choice leads to

$$\begin{aligned} \mathbb{E}(Y_m) &= (1-p_{ij})p_{ij} - p_{ij}(1-p_{ij}) = 0 \\ \mathbb{E}(Y_m^2) &= (1-p_{ij})^2 p_{ij} + p_{ij}^2(1-p_{ij}) \\ &= p_{ij}(1-p_{ij}). \end{aligned}$$

Moreover,

$$S_{W_n} = Y_1 + \cdots + Y_{n_{i+}} = n_{ij}(1-p_{ij}) + (n_{i+} - n_{ij})(-p_{ij}) = n_{ij} - n_{i+}p_{ij}.$$

Thus, Ancombe's lemma implies that

$$\begin{aligned} \frac{S_{n_{i+}}}{\sqrt{p_{ij}(1-p_{ij})n_{i+}}} &= \frac{n_{ij} - n_{i+}p_{ij}}{\sqrt{p_{ij}(1-p_{ij}n_{i+})}} \\ &= \frac{\widehat{p}_{ij} - p_{ij}}{\sqrt{p_{ij}(1-p_{ij})/n_{i+}}} \overset{D}{\to} N(0,1). \end{aligned}$$

Finally, observing that $\frac{n_{i+}}{n\pi_i} \overset{a.s.}{\to} 1$ we then conclude the result.

■

With Proposition 6.3, we are able to construct a confidence interval of each p_{ij}. Also, we can do a hypothesis test on the transition probability matrix $\mathbf{P}$ to see if there are interesting structures inside.

6.3 Bayesian Approach

When we adopt a Bayesian approach in the Markov chain, we need to put a prior distribution for the transition probability matrix $\mathbf{P}$. Essentially, we need to put a prior on each p_{ij} with the constraint that $\sum_{j=1}^{s} p_{ij} = 1$.

Note that the constraints are on each row of $\mathbf{P}$ so one possibility is to put a prior on each row of $\mathbf{P}$. Here we will use the *Dirichlet prior*.

The Dirichlet distribution is a multivariate distribution over the simplex $\sum_{i=1}^{K} x_i = 1$ and $x_i \geq 0$. Its probability density function is

$$p(x_1, \cdots, x_K; \alpha_1, \cdots, \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

where $B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}$ and $\alpha = (\alpha_1, \cdots, \alpha_K)$ are the parameters of this distribution. The Dirichlet distribution generates a random vector with length K and each element of this vector is non-negative and summation of elements is 1, meaning that it generates a random probability vector. You can view it as a generalization of the Beta distribution. For $Z = (Z_1, \cdots, Z_K) \sim \mathsf{Dirch}(\alpha_1, \cdots, \alpha_K)$, $\mathbb{E}(Z_i) = \frac{\alpha_i}{\sum_{j=1}^K \alpha_j}$ and the mode of Z_i is $\frac{\alpha_i - 1}{\sum_{j=1}^K \alpha_j - K}$ so each parameter α_i determines the relative importance of category (state) i. Because it is a distribution putting probability over K categories, Dirchlet distribution is very popular in social sciences and linguistics analysis.

Going back to our data analysis problem, by choosing $K = s$, the Dirchlet distribution can be used as a prior distribution for each row of $\mathbf{P}$. For each $p_i = (p_{i1}, \cdots, p_{is})$, we use the Dirchlet prior with parameter $\alpha_i = (\alpha_{i1}, \cdots, \alpha_{is})$ such that

$$p_i \sim \mathsf{Dirch}(p_i; \alpha_i).$$

The posterior distribution can then be written as

$$\begin{aligned}\pi(\mathbf{P}|X_1, \cdots, X_n) &\propto \prod_{i=1}^s \prod_{j=1}^s p_{ij}^{n_{ij}} \prod_{k=1}^s p(p_k; \alpha_k) \\ &\propto \prod_{i=1}^s \prod_{j=1}^s p_{ij}^{n_{ij}} \left(\prod_{k=1}^s p_{k1}^{\alpha_{k1}-1} \times \cdots \times p_{ks}^{\alpha_{ks}-1} \right) \\ &= \prod_{i=1}^s \left(p_{i1}^{n_{i1}+\alpha_{i1}-1} \times \cdots \times p_{is}^{n_{is}+\alpha_{is}-1} \right),\end{aligned}$$

which is a product of each row.

So the posterior distribution of p_i given the data is

$$\pi(p_i|X_1, \cdots, X_n) = p_{i1}^{n_{i1}+\alpha_{i1}-1} \times \cdots \times p_{is}^{n_{is}+\alpha_{is}-1},$$

which is the PDF of $\mathsf{Dirch}(n_{i1} + \alpha_{i1}, \cdots, n_{is} + \alpha_{is})$. This implies that the posterior mean and MAP are

$$\widehat{p}_{ij,\pi} = \frac{n_{ij} + \alpha_{ij}}{\sum_{k=1}^s n_{ik} + \alpha_{ik}}, \quad \widehat{p}_{ij,MAP} = \frac{n_{ij} + \alpha_{ij} - 1}{\sum_{k=1}^s n_{ik} + \alpha_{ik} - K}.$$

The credible interval can be constructed using the level sets of the posterior distribution. In hypothesis test, we can use the Bayes factor and compare it with our prior on the hypotheses to decide if we can reject the null hypothesis or not.

6.4 When $\mathbf{P} = \mathbf{P}(\theta)$

In some cases, the transition probability matrix $\mathbf{P} = \mathbf{P}(\theta)$ is determined by the parameter $\theta \in \mathbb{R}^L$ (namely, each $p_{ij} = p_{ij}(\theta)$). And our goal is to estimate θ from our data. We will talk about how to make inference in this case using the Frequentist approach (you can use Bayesian as well). Here we first look at a specific example about genetic drift.

Wright-Fisher Model with Mutation. Let m be the size of the population and assume that each there are two possible alleles A and a. Let $\{X_n\}$ be the number of A alleles in the population at generation n. We assume that from one generation to the next generation, each individual is randomly mated so the transition probability from state i to state j is

$$p_{ij} = \binom{2m}{j} q_i^j (1 - q_i)^{2m-j}, \tag{6.1}$$

where $q_i = \frac{i}{2m}$ when there is no mutation. Now we model the mutation using a simple probability model: $u = P(a \to A)$ and $v = P(A \to a)$, where $\to$ here denotes the mutation. Then equation (6.1) will be modified with

$$q_i = \frac{i}{2m}(1-v) + \left(1 - \frac{i}{2m}\right)u.$$

Thus, the transition probability matrix $\mathbf{P} = \mathbf{P}(u, v)$. Our goal is to estimate parameters of interest u, v after observing generations $X_1, \cdots, X_n$.

This scenario–the parameters of a distribution is controlled by another set of parameters– is very common in real data analysis. When the parametric family is from an exponential family, this model is also known as the *curved exponential family*. Here we briefly mention a few results about the case of $\mathbf{P} = \mathbf{P}(\theta)$. Note that in this case, the log-likelihood function is

$$\ell_n(\theta) = \sum_{i,j=1}^{s} n_{ij} \log p_{ij}(\theta)$$

and the score equations (score function = 0) are

$$\frac{\partial \ell_n(\theta)}{\partial \theta_k} = \sum_{i,j=1}^{s} \frac{n_{ij}}{p_{ij}(\theta)} \frac{\partial p_{ij}(\theta)}{\partial \theta_k} = 0$$

for $k = 1, \cdots, L$ and the Fisher's information matrix $I_1(\theta) = \{I_{km}(\theta)\}$ is

$$I_{km}(\theta) = \sum_{i,j=1}^{s} \frac{\pi_i(\theta)}{p_{ij}(\theta)} \frac{\partial p_{ij}(\theta)}{\partial \theta_k} \frac{\partial p_{ij}(\theta)}{\partial \theta_m},$$

where $\pi_i(\theta)$ is the stationary distribution of state i.

Proposition 6.4 *Let $\{X\}_n$ be a Markov chain with a transition probability matrix $\mathbf{P}(\theta)$ satisfying*

1. *$S_Y = \{(i,j) : p_{ij} > 0\}$ does not change with θ.*
2. *Each $p_{ij}(\theta)$ is at least three times continuously differentiable.*
3. *For each $k = 1, \cdots, L$, the $|S_Y| \times L$ matrix $\left\{\frac{\partial p_{ij}(\theta)}{\partial \theta_k}\right\}$ has rank L.*
4. *For each θ, the chain is irreducible and aperiodic.*

Let θ_0 be the true value of the parameter of interest. Then

1. Statistical consistency: $\widehat{\theta}_{MLE} \overset{a.s.}{\to} \theta_0$.
2. Asymptotic normality: $\sqrt{n}(\widehat{\theta} - \theta_0) \overset{D}{\to} N(0, I_1^{-1}(\theta_0))$.
3. Variance estimation: *Let* $\widehat{I}_1(\theta) = \frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta^T}\sum_{i,j=1}^{s} n_{ij}\log p_{ij}(\theta)$. *Then* $\widehat{I}_1\left(\widehat{\theta}\right) \overset{P}{\to} I_1(\theta_0)$.

Proof: Sections I.2 and I.5 in Billingsley (1961, Statistical Inference for Markov Processes) ■

With Proposition 6.4, we know that the MLE will be a good estimator of θ_0 and we can use the MLE with the variance estimator to construct an asymptotically valid confidence interval.

Note that if we are using a Bayesian approach for the problem of $\mathbf{P} = \mathbf{P}(\theta)$, we need to find a prior distribution of θ. In the case of Wright-Fisher model with mutation, a possible choice of the prior will be the Beta distribution (for both u and v) since it is a conjugate prior fro the binomial distribution.

6.5 Testing the Independence VS Markov Dependence

When we observe a sequence of RVs $X_0, \cdots, X_n$. Sometimes we are not sure if they are IID or they form a Markov chain. We can do a hypothesis test to examine which assumption is more plausible. Determining if the sequence is IID is very important because if they are IID, we have a lot more statistical tools to analyze the underlying population.

In this case, the null and the alternative hypotheses are

$$H_0 : X_1, \cdots, X_n \mid X_0 \text{ are IID.}$$
$$H_1 : X_1, \cdots, X_n \mid X_0 \text{ form a Markov chain with a t.p.m. of non-identical rows.}$$

Under H_0, the observations are like IID from a PMF $p_i = P(X_1 = i)$ whereas under H_1, the underlying probability model is described by the transition probability matrix $\mathbf{P} = \{p_{ij}\}$.

An interesting fact is – even under H_0, we can still write the model using the transition probability matrix but with the requirement that *every row is the same*. Namely, we can use the quantity $\mathbf{P} = \{p_{ij}\}$, where $p_{ij} = P(X_n = j \mid X_{n-1} = i)$ to describe the case of both H_0 and H_1. In this scenario,

$$H_0 : p_{ij} = p_j \text{ for all } i, j \in S. \tag{6.2}$$

Note that identifying a common quantity that both H_0 and H_1 is very important in hypothesis testing because we can then transform the problem of testing a 'statement/property' into testing some parameters of interest.

There are multiple ways to test equation (6.2). We use likelihood ratio test (LRT) and the Bayes factor in this case.

LRT:
Recall that the likelihood ratio test would use the MLE under H_0 and $H_0 \cup H_1$ to perform the test. Under H_0, $\widehat{p}_{ij,H_0} = \widehat{p}_j = \frac{n_{+j}}{n}$ while under H_1, $\widehat{p}_{ij} = \frac{n_{ij}}{n_{i+}}$. Thus, the test statistic is

$$\begin{aligned} T_n &= 2(\ell_n(\widehat{\mathbf{P}}_{MLE}) - \ell_n(\widehat{\mathbf{P}}_{MLE,H_0})) \\ &= 2\sum_{i,j=1}^{s} n_{ij} \log\left(\frac{\widehat{p}_{ij}}{\widehat{p}_j}\right) \\ &= 2\sum_{i,j=1}^{s} n_{ij} \log\left(\frac{n_{ij} \cdot n}{n_{i+} \cdot n_{+j}}\right). \end{aligned}$$

To compute the p-value using LRT, we use the χ^2 distribution as a reference distribution of T_n. Now here comes the question: what is the degree of freedom in this case? First, we think about H_1. There are totally s^2 parameters in $\mathbf{P}$. Each row has to sum to 1 so there are s constraints. Thus, under H_1, there are $s^2 - s$ degrees of freedom. What about H_0? Under H_0, all rows are identical so there are at most s parameters. But all these s numbers have to sum to 1, leading to one constrain so there are totally $s - 1$ degrees of freedom. Thus, the remaining degrees of freedom is $(s^2 - s) - (s - 1) = (s - 1)^2$. So $T_n \sim \chi^2_{(s-1)^2}$.

Bayes Factor:
To use the Bayes factor, we need to put priors on the parameters $p_i = (p_{i1}, \cdots, p_{is})$. As we have mentioned, the Dirichlet piror seems to be a good choice. Under H_1, although we can use different priors on different rows, here we use the same prior for every row. Moreover, we would use the same prior for both H_0 and H_1. This choice reflects the fact that we do not have any different believes of $\mathbf{P}$ under the two hypotheses. So our prior is

$$p_i \sim \mathsf{Dirichlet}(\alpha_1, \cdots, \alpha_s)$$

for some given hyperparameters $\alpha_1, \cdots, \alpha_s$. The Bayes factor is

$$\begin{aligned}
\mathsf{BF} &= \frac{\int p(\mathsf{Data}|\mathbf{P})\pi(\mathbf{P}|H_0)d\mathbf{P}}{\int p(\mathsf{Data}|\mathbf{P})\pi(\mathbf{P}|H_1)d\mathbf{P}} \\
&= \frac{\frac{1}{B(\alpha)} \int \prod_{j=1}^{s} p_j^{n_{+j}} p_j^{\alpha_j - 1} dp_j}{\frac{1}{B^s(\alpha)} \int \prod_{i,j=1}^{s} p_{ij}^{n_{ij}} p_{ij}^{\alpha_j - 1} dp_{ij}} \\
&= \frac{\frac{1}{B(\alpha)} \prod_{j=1}^{s} \int p_j^{n_{+j}+\alpha_j - 1} dp_j}{\frac{1}{B^s(\alpha)} \prod_{i,j=1}^{s} \int p_{ij}^{n_{ij}-\alpha_j - 1} dp_{ij}} \\
&= \frac{B(n_{+1}+\alpha_1, \cdots, n_{+s}+\alpha_s)/B(\alpha_1, \cdots, \alpha_s)}{\prod_{i=1}^{s} [B(n_{i1}+\alpha_1, \cdots, n_{is}+\alpha_s)/B(\alpha_1, \cdots, \alpha_s)]}.
\end{aligned}$$