Handout 5

STAT5710 Applied Stochastic Processes

Here we estimate transition probabilities from an observed sample path, first by maximum likelihood and then under a Dirichlet prior.

Likelihood estimation

Suppose $\{x_t\}_{t=0}^T$ is a sample path for a homogeneous Markov chain $\{X_t\}$.

Denote the transition probabilities by $p(x_{t+1}, x_t) = P(X_{t+1} = x_{t+1} \mid X_t = x_t)$ and the initial distribution by $\alpha(x_0) = P(X_0 = x_0)$.

1. Write the joint probability mass of the path as a product of the transition probabilities and the initial distribution.

$$P(X_0 = x_0, \dots, X_T = x_T) \;=$$

$$=$$

2. Take logs, then collect the terms by transition counts $N_{ij} = \#\{t : x_{t+1} = j, x_t = i\}$.

$$\ell(P, \alpha; x_0, \dots, x_T) \;=$$

$$=$$

3. Unconstrained maximization with respect to the initial distribution gives $\hat{\alpha} =$ _____. Thus we find the MLE of P by maximizing the conditional likelihood

$$\ell(P; x_1, \dots, x_T \mid x_0) \;=$$

Does $\nabla \ell(P; x_1, \dots, x_T \mid x_0) = 0$ give the solution? Why or why not?

4. The strategy is to compute the MLE on a per-row basis by maximizing the inner sum of the conditional likelihood for fixed i.

   Recall: to maximize a function $f(p)$ subject to the constraint $g(p) = 0$, set the gradient of the Lagrangian $L = f - \lambda g$ to zero, and use the constraint to pin down the multiplier λ.

   (a) Consider the ith row of P.

      i. Write the optimization objective.

      $$f(p_i) =$$

      ii. Express the constraint $\sum_j p_{ij} = 1$ in the form $g(p) = 0$.

      $$g(p_i) =$$

   (b) Now write the Lagrangian $L(p_i, \lambda) = f(p_i) - \lambda g(p)$.

   $$L(p_i, \lambda) \; = \qquad\qquad - \lambda \Big[\qquad\qquad \Big]$$

   (c) Differentiate and set equal to zero.

   $$\frac{\partial L}{\partial p_{ij}} \; = \qquad\qquad = 0 \qquad \Longrightarrow \qquad p_{ij} \; =$$

   (d) Now use the constraint to solve for the multiplier.

   $$g(p_i) \; = \; 0 \; \Leftrightarrow \qquad\qquad \Longrightarrow \qquad \lambda \; =$$

   (e) So the conditional MLE is $\widehat{p}_{ij} =$ _____

The estimator is perfectly well-defined whether the state space is finite or infinite.

5. But note that $N_{ij} = 0 \Longrightarrow \widehat{p}_{ij} =$ _____. This implies…

If the state space is countably infinite then one usually imposes some parametric structure on p_{ij}.

6. **Lazy random walk on $\mathbb{Z}$.** Here the chain has three parameters,

$$p_{ij} = \begin{cases} p_0, & j = i \\ p_1, & j = i+1 \\ p_2, & j = i-1 \end{cases}$$

and we let $I_t = X_t - X_{t-1}$ denote the tth increment.

(a) Write the joint probability mass in terms of the increments.

$$P(X_0 = x_0, \dots, X_T = x_T) \; =$$

$$=$$

(b) Now express the log-likelihood in terms of visit counts N_0, N_1, N_2.

i. The visit counts are $N_j \; =$

ii. The log-likelihood is

$$\ell(p_0, p_1, p_2, \alpha; x_0, \dots, x_T) \; =$$

(c) So by the same argument in (4), the conditional MLE is given by

$$\hat{p}_j = \qquad\qquad \text{for } j =$$

(d) "Drift" refers to directional bias in the walk and is captured by the average increment

$$\delta = E(I_t) =$$

a plug-in estimator is given by

$$\hat{\delta} \; =$$

(e) The approach generalizes directly to $\mathbb{Z}^d$, giving

$$\hat{p}_j = \qquad\qquad \text{for } j =$$

7. There are two options for intervals based on the MLE. The first is exact; the second is asymptotic.

   (a) Marginally, the distribution of transition counts is

   $$N_{ij} \sim$$

   so we can construct an interval from the tails:

   $$\hat{p}_L = \qquad\qquad \hat{p}_U =$$

   (b) Asymptotically, we have

   $$\sqrt{N_{i\cdot}}\,(\hat{p}_i - p_i) \xrightarrow{d} \qquad\Longrightarrow\quad \sqrt{N_{i\cdot}}\,(\hat{p}_{ij} - p_{ij}) \xrightarrow{d}$$

   which gives the Wald interval

   $$\hat{p}_{ij} \pm$$

8. The conditional MLE can break down in three situations:

   i.
   ii.
   iii.

9. So-called **Laplace smoothing** adds one to each count, $\tilde{N}_{ij} \leftarrow N_{ij} + 1$, giving

   $$\tilde{p}_{ij} = \frac{\tilde{N}_{ij}}{\tilde{N}_{i\cdot}} =$$

   (a) Now if $N_{ij} = 0$ but $N_{i\cdot} \neq 0$ we have

   $$\hat{p}_{ij} = \qquad\qquad \text{but } \tilde{p}_{ij} =$$

   (b) And if $N_{i\cdot} = 0$ we have

   $$\hat{p}_{ij} = \qquad\qquad \text{but } \tilde{p}_{ij} =$$

   (c) What does this ensure about $\tilde{P}$ and why does it matter?

Bayesian estimation

Laplace smoothing is equivalent to the posterior mean under a symmetric Dirichlet prior.

10. Take each row of P to be Dirichlet with all parameters $\alpha_{ij} = 1$:

$$(p_{i1}, \ldots, p_{ik}) \sim \text{Dirichlet}(1, \ldots, 1) \qquad \Longleftrightarrow \qquad f(p_i \mid \alpha_i) = \frac{1}{B(\alpha)} \prod_{j=1}^{k} p_{ij}^{\alpha_{ij}-1}$$

where $B(\alpha) = \prod_j \Gamma(\alpha_j) \big/ \Gamma\big[\sum_j \alpha_j\big]$.

(a) Multiply the likelihood by the prior and collect exponents.

$$f(P \mid \{x_t\}) \;\propto$$

(b) Row-wise, the posterior is therefore distributed as…

Fact. If $p \sim \text{Dir}(\alpha_1, \ldots, \alpha_k)$ then $E(p_i) = \alpha_i \big/ \sum_j \alpha_j$.

(c) So the posterior mean is

$$E_\alpha(p_{ij} \mid \{x_t\}) =$$

(d) Two special cases:

$$E_{\alpha=1}(p_{ij} \mid \{x_t\}) = \qquad \text{(Laplace smoothing)}$$

$$E_{\alpha=\alpha\mathbf{1}}(p_{ij} \mid \{x_t\}) = \qquad \text{(symmetric prior)}$$

Interestingly, the solution given by the symmetric prior with $\alpha = \sqrt{N_{i\cdot}}/k$ is identical to the minimax estimator under squared error loss.

(e) The symmetric prior smooths toward the uniform distribution, in the sense that

$$\widehat{p}_{ij} \;\longrightarrow \qquad\qquad \text{as } \alpha \to \infty$$

11. Consider instead shrinking toward some target distribution $g = (g_1, \ldots, g_k)$ via the prior, $p_i \sim \text{Dirichlet}(\alpha g)$.

    (a) Now the row-wise posterior is distributed as

    $$p_i \mid \{x_t\} \sim _____$$

    ,

    (b) Express the posterior mean as a weighted mixture of the MLE and the shrinkage target

    $$\hat{p}_{ij} =$$

    $$=$$

    (c) For an empirical Bayes estimate, use the marginal destination probabilities:

    $$\hat{g}_j =$$

    (d) The two extremes are:

    $$N_{i\cdot} \ll \alpha \implies \hat{p}_{ij} \approx \qquad\qquad N_{i\cdot} \gg \alpha \implies \hat{p}_{ij} \approx$$

    So the effect of increasing α is…

Fact. If $(p_1, \ldots, p_k) \sim \text{Dir}(\alpha)$ then $p_j \sim \text{beta}(\alpha_j, \sum_{i \neq j} \alpha_i)$.

12. So the marginal posterior of p_{ij} under an arbitrary Dirichlet prior is:

    $$p_{ij} \sim$$

    Quantiles then give pointwise credible intervals for p_{ij}.

13. A joint HPD region can be approximated by sampling from the (Dirichlet) posterior to estimate the level set

    $$R =$$

Example

14. **Server monitoring.** A server is polled once a minute and recorded as idle (I), busy (B), or down (D). Over one hour of polling, the transitions between successive minutes are

$N =$

	I	B	D	row sum
I	18	6	3	27
B	2	6	1	9
D	0	1	2	3

(a) Compute the conditional MLE.

$\hat{P} =$

	I	B	D
I			
B			
D			

(b) Give a 95% asymptotic interval for p_{IB}, the probability that an idle minute is followed by a busy one.

Now try the same for p_{DI}. What goes wrong?

(c) Compute the Laplace-smoothed estimate.

$\tilde{P} =$

	I	B	D
I			
B			
D			

(d) Now use a symmetric Dirichlet prior with $\alpha = 3$.

$\tilde{P} =$

	I	B	D
I			
B			
D			

(e) Finally shrink toward the target $g = (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$, roughly the marginal destination frequencies, with $\alpha = 3$.

$\tilde{P} =$

	I	B	D
I			
B			
D			