4

Inference for Discrete-Time Markov Chains

Using material from chapters 2 and 3, we derive likelihood and Bayesian inference for fully and incompletely observed discrete time Markov chains. Our motivating example is precipitation at a weather station in Washington.

4.1 Inference for completely observed DTMCs

The US Weather Service maintains a large number of precipitation monitors throughout the United States. One station is located at the Snoqualmie Falls in the foothills of the Cascade Mountains in western Washington. A day is defined as wet if at least 0.01 inches of precipitation falls during a precipitation day: 8 a.m. through 8 a.m. the following calendar day. To start with, we shall ignore the amounts of rainfall, and just look at the pattern of wet and dry days. Using data from 1948 through 1983, and looking at January rainfall only, there were 325 dry and 791 wet days. Let $X_{ij} = 1_{\{\text{day } i \text{ of year } j \text{ is wet}\}}$, where $1_{\{A\}}$ is 1 if the event A occurs, and 0 otherwise. A very simple model, which we can call the Bernoulli model, is that $X_{ij} \sim \text{Bin}(1, p)$, with the X_{ij} independent, i.e., an iid model, and with p being the probability of rain at Snoqualmie Falls on a January day. The likelihood of p using this model is

$$L(p) \propto p^{791}(1-p)^{325}.$$

Letting n be the number of days observed, it is easy to see that this likelihood is maximized by $\hat{p} = \sum_{i,j} x_{ij}/n = 0.709$. A standard error for this estimate is $[p(1-p)/n]^{1/2}$ which we estimate (using $\hat{p}$ in place of p) to be 0.014.

In order to assess the fit of the binomial model for rainfall, we first try to see if the independence assumption seems reasonable. We may suspect a certain amount of persistence, i.e., stretches of like weather, in the data. This would be induced by the relatively slow movement of large weather systems through an area. In the winter, a typical front may take up to three days to pass through from the Pacific Ocean. In order to study this hypothesis, let us look at consecutive pairs of days. Figure 4.1 shows the pattern of rainfall.

If the independence model is correct, we would expect to see $36 \times 30 \times \hat{p}(1-\hat{p}) = 223 =$ dry days following wet days, since we have 36 years of data, and 30 consecutive pairs of days for each January. Table 4.1 contains the total counts, with expected counts under the independence assumption shown in parenthesis. There seems to be a lot more dry days followed by dry days, and wet days followed by wet days, than what the simple iid model predicts.

The discrepancy between observed and expected counts indicates that a Markov chain may fit the data better than the Bernoulli model.

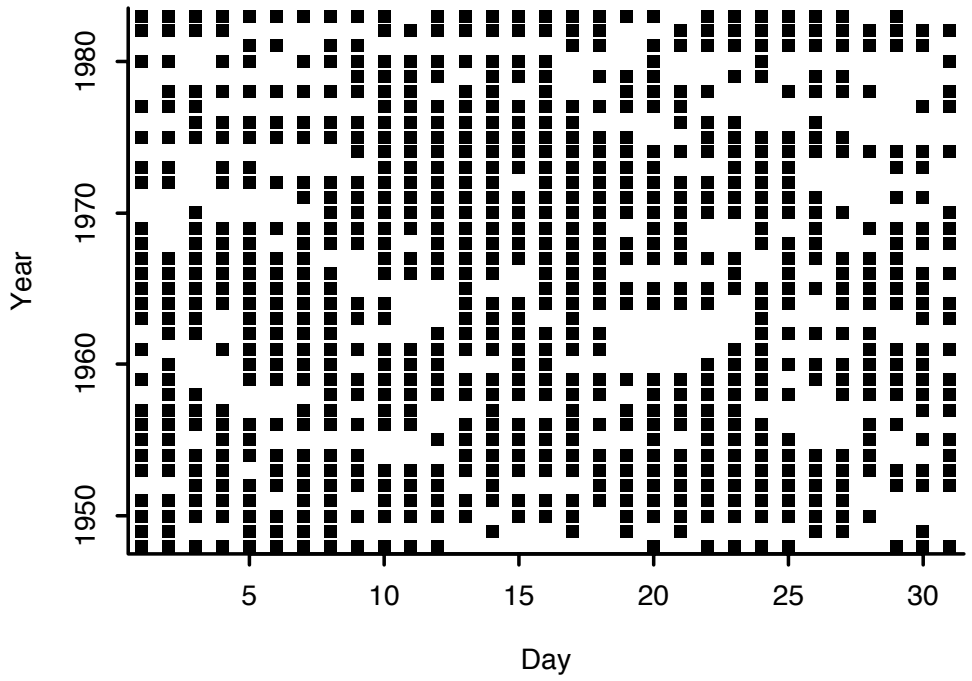

FIGURE 4.1: The pattern of January precipitation at Snoqualmie Falls. Each square is a day with measurable precipitation. Rows correspond to years, columns to days.

TABLE 4.1

	Today dry		Today wet		Total
Yesterday dry	186	(91)	123	(223)	309
Yesterday wet	128	(223)	643	(543)	771
Total	314		766		1080

4.1.1 Maximum likelihood estimation

4.1.1.1 Unrestricted transition probabilities

Our objective here is to estimate transition probabilities of a Markov chain from the observed path of the Markov chain:

$$\mathbf{y} = (y_0, y_1, \ldots, y_n).$$

We assume a finite state space $\Omega = \{1, \ldots, s\}$. The likelihood becomes

$$L(\boldsymbol{\nu}, \mathbf{P}) = \nu_{y_0} \prod_{i=1}^{n} p_{y_{i-1} y_i},$$

where $\boldsymbol{\nu}$ is an initial distribution and $\mathbf{P}$ is a transition probability matrix.

Since we observe only one initial state y_0, it is unrealistic to estimate initial distribution $\boldsymbol{\nu}$ from just one Markov chain path. Therefore, we will condition on y_0 and consider what is called a conditional likelihood:

$$L(\mathbf{P}) = \prod_{i=1}^{s} \prod_{j=1}^{s} p_{ij}^{n_{ij}},$$

where $n_{ij} = \sum_{k=0}^{n-1} 1_{\{y_k=i, y_{k+1}=j\}}$ is the number of times we observe transition $i \to j$ in $\mathbf{y}$. The log likelihood becomes

$$l(\mathbf{P}) = \sum_{i=1}^{s}\sum_{j=1}^{s} n_{ij} \ln p_{ij} = \sum_{i=1}^{s} l_i(p_{i1}, \ldots, p_{is}),$$

where $l_i(p_{i1}, \ldots, p_{is}) = \sum_{j=1}^{s} n_{ij} \ln p_{ij}$. Since $l_1(p_{11}, \ldots, p_{1s}), \ldots, l_s(p_{s1}, \ldots, p_{ss})$ do not share parameters, these functions can be maximized separately. This means that we can estimate each row of $\mathbf{P}$ separately by maximizing $l_i(p_{i1}, \ldots, p_{is})$ subject to constraint $\sum_{j=1}^{s} p_{ij} = 1$. But this is exactly the multinomial maximum likelihood, except the role of the "sample size" is played by $n_{i+} = \sum_{j=1}^{s} n_{ij}$. Hence, the mle of $\mathbf{P}$ is defined by

$$\hat{p}_{ij} = \begin{cases} \frac{n_{ij}}{n_{i+}} & \text{if } n_{i+} > 0, \\ 1_{\{i=j\}} & \text{if } n_{i+} = 0. \end{cases}$$

Proposition 4.1. *Let $\{X_n\}$ be an irreducible homogeneous Markov chain defined on a finite state space. Then maximum likelihood estimates derived from an observed Markov chain path $\mathbf{y} = (y_0, y_1, \ldots, y_n)$ have the following limiting properties as n approaches ∞:*

$$\hat{p}_{ij} \xrightarrow{a.s.} p_{ij} \text{ (consistency)}$$
$$\frac{\hat{p}_{ij} - p_{ij}}{\sqrt{p_{ij}(1-p_{ij})/n\pi_i}} \xrightarrow{D} N(0,1) \text{ (asymptotic normality)},$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_s)$ is the stationary distribution of $\{X_n\}$.

Proof.

Consistency

Consider a snake chain $Y_n = (X_n, X_{n+1})$ defined on $\Phi = \{(i_0, i_1) \in \Omega^2 : p_{i_0 i_1} > 0\}$.

$$\begin{aligned} &\Pr(Y_n = (i_n, j_n) \,|\, Y_{n-1} = (i_{n-1}, j_{n-1}), \ldots, Y_0 = (i_0, j_o)) = \\ &\Pr(X_n = i_n, X_{n+1} = j_n \,|\, X_{n-1} = i_{n-1}, X_n = j_{n-1}, \ldots, X_0 = i_0, X_1 = j_o) = \\ &\Pr(X_n = i_n, X_{n+1} = j_n \,|\, X_{n-1} = i_{n-1}, X_n = j_{n-1}) = p_{i_n j_n} 1_{\{i_n = j_{n-1}\}}. \end{aligned}$$

From exercise 2.4.7, we know that $\{Y_n\}$ is a Markov chain and we have found the transition probabilities of the snake chain. We know that the snake chain is irreducible since $\{X_n\}$ is irreducible. We have also found the stationary distribution of the snake chain:

$\mu(i,j) = \pi_i p_{ij}$. The ergodic theorem, applied to $\{Y_n\}$ and $\{X_n\}$, yields

$$\frac{n_{ij}}{n} = \frac{1}{n}\sum_{k=0}^{n-1} 1_{\{Y_k=(i,j)\}} \xrightarrow{\text{a.s.}} \pi_i p_{ij},$$
$$\frac{n_{i+}}{n} = \frac{1}{n}\sum_{k=0}^{n-1} 1_{\{X_k=i\}}. \xrightarrow{\text{a.s.}} \pi_i$$

Therefore,

$$\hat{p}_{ij} = \frac{\frac{n_{ij}}{n}}{\frac{n_{i+}}{n}} \xrightarrow{\text{a.s.}} \frac{\pi_i p_{ij}}{\pi_i} = p_{ij},$$

because we know that $\pi_i > 0$ for all $i \in E$.

Asymptotic normallity

To prove asymptotic normality, we'll need a result due to Anscombe (1952):

Lemma 4.1. *Let $Y_1, Y_2, \ldots$ be iid with $E(Y) = 0$, $E\left(Y^2\right) = Var(Y_1) = \sigma < \infty$, $S_n = \sum_{i=1}^n Y_i$, and $W_1, W_2, \ldots$ be random positive integers with $W_n/n \xrightarrow{P}$ constant. Then*

$$\frac{S_{W_n}}{\sqrt{\sigma W_n}} \xrightarrow{D} N(0,1).$$

We will accept this result without a proof.

For a fixed state i, without loss of generality, assume that $X_0 = i$. Then define

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

where τ_m is the mth visit to state i.

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

Moreover, the regenerative cycles lemma **??** says that the Y_m are independent.

Let $W_n = n_{i+}$. Then we know that $W_n/n \xrightarrow{\text{a.s.}} \pi_i$. Letting $S_n = \sum_{i=1}^n Y_i$, we have

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

Therefore, Anscombe's lemma 4.1 implies that

$$\frac{S_{n_{i+}}}{\sqrt{p_{ij}(1-p_{ij})n_{i+}}} = \frac{n_{ij} - n_{i+}p_{ij}}{\sqrt{p_{ij}(1-p_{ij})n_{i+}}} = \frac{\frac{n_{ij}}{n_{i+}} - p_{ij}}{\sqrt{p_{ij}(1-p_{ij})/n_{i+}}} = \frac{\hat{p}_{ij} - p_{ij}}{\sqrt{p_{ij}(1-p_{ij})/n_{i+}}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1).$$

Observing that $n_{i+}/n\pi_i \xrightarrow{\text{a.s.}} 1$, we arrive at the desired result. □

Example: *Snoqualmie Falls precipitation* The data from the Snoqualmie US Weather Service station that can be summarized as transitional counts:

$$\mathbf{n} = \begin{pmatrix} 186 & 123 \\ 128 & 643 \end{pmatrix}.$$

The maximum likelihood estimates of transition probabilities are

$$\hat{p}_{12} = \frac{123}{186+123} = 0.398 \qquad \hat{p}_{21} = \frac{128}{128+643} = 0.166.$$

Recall that the stationary distribution of the two-state Markov chain is $[p_{21}/(p_{12}+p_{21}), p_{21}/(p_{12}+p_{21})]$. So the estimated stationary distribution is $\hat{\boldsymbol{\pi}}^T = (0.3, 0.7)^T$ and

$$\begin{aligned} \mathrm{Var}(\hat{p}_{12}) &\approx \hat{p}_{12}(1-\hat{p}_{12})/(n\hat{\pi}_1) = 0.0007 \\ \mathrm{Var}(\hat{p}_{21}) &\approx \hat{p}_{21}(1-\hat{p}_{21})/(n\hat{\pi}_2) = 0.0002 \end{aligned}$$

A 95% confidence band for p_{11} is (0.808,0.860) while one for p_{01} is (0.343,0.453). These are individual confidence bands, and the asmptotic joint coverage probability of the rectangle formed by these two intervals is, using asymptotic independence, $0.95^2 = 0.903$. To obtain a 95% joint confidence rectangle we can use individual 97.5% intervals, which yield the rectangle $(0.775, 0.893) \times (0.272, 0.524)$.

4.1.1.2 Parametric transition probabilities

In this subsection, we assume that transition probability matrix $\mathbf{P} = \mathbf{P}(\boldsymbol{\theta})$ is a function of a low dimensional parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_l)$.

Example: *Wright-Fisher model with mutation* Let $2m$ be the population size and $\{X_n\}$ be the number of A alleles in the population. We define mutation probabilities $u = \Pr(a \to A) \in (0,1)$ and $v = \Pr(A \to a) \in (0,1)$. Assuming that after sampling with replacement from the previous generation, each gene mutates to an opposite type with the corresponding probability, it is easy to show that one-step transition probabilities of $\{X_n\}$ remain their binomial form,

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

where

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

Notice that entries of the transition probabilities matrix $\mathbf{P} = \mathbf{P}(u,v)$ are defined by just two parameters, u and v.

We require the following regularity conditions:

1. $\mathscr{D} = \{i, j : p_{ij}(\boldsymbol{\theta}) > 0\}$ does not change with $\boldsymbol{\theta}$. In other words, the communication graph of the Markov chain remains the same for all $\boldsymbol{\theta}$.
2. Each $p_{ij}(\boldsymbol{\theta}) \in \mathbb{C}^3$, i.e. are three times continuously differentiable.
3. The $d \times s$ matrix $\{\partial p_{ij}(\boldsymbol{\theta})/\partial \theta_k\}$ has rank s, where $k = 1, \ldots, l$, and $d = |\mathscr{D}|$.
4. For each $\boldsymbol{\theta}$, the chain is irreducible and aperiodic.

Let $\mathbf{y} = (y_0, y_1, \ldots, y_n)$ be the observed path and n_{ij} be the number of times we observe an $i \to j$ transition. Then the log likelihood becomes:

$$l(\boldsymbol{\theta}) = \sum_{i,j \in D} n_{ij} \ln p_{ij}(\boldsymbol{\theta}).$$

The likelihood equations:

$$\frac{\partial l}{\partial \theta_k} = \sum_{i,j \in D} \frac{n_{ij}}{p_{ij}(\boldsymbol{\theta})} \frac{\partial p_{ij}(\boldsymbol{\theta})}{\partial \theta_k} = 0, \text{ for } k = 1, \ldots, l.$$

Proposition 4.2. *Let $\{X_n\}$ a Markov chain with transition probability matrix $\mathbf{P}(\boldsymbol{\theta})$ satisfying the above regularity conditions. If $\boldsymbol{\theta}_0$ is a true parameter vector, then*

1.The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ exists and it is consistent.

2.$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(0, \mathbf{I}^{-1}(\boldsymbol{\theta}_0))$, where $\mathbf{I}(\boldsymbol{\theta}) = \{I_{km}(\boldsymbol{\theta})\}$ is the Fisher information matrix with

$$I_{km} = \sum_{i,j \in E} \frac{\pi_i(\boldsymbol{\theta})}{p_{ij}(\boldsymbol{\theta})} \frac{\partial p_{ij}(\boldsymbol{\theta})}{\partial \theta_k} \frac{\partial p_{ij}(\boldsymbol{\theta})}{\partial \theta_m}$$

and $\pi_i(\theta)$ is stationary probability of state i.

A proof of this proposition and many other asymptotic results for Markov processes can be found in Billingsley (1961).

4.1.2 Bayesian estimation

4.1.2.1 Unrestricted transition probabilities

We start our Bayesian analysis by putting s independent Dirichlet priors on the rows of $\mathbf{P}$:

$$(p_{i1},\ldots,p_{is}) \sim \text{Dirichlet}(\alpha_{i1},\ldots,\alpha_{is}).$$

Recall that the conditional likelihood of the Markov chain realization $\mathbf{y} = (y_0, y_1, \ldots, y_n)$ is

$$\Pr(\mathbf{y} \mid \mathbf{P}) = \prod_{i=1}^{s}\prod_{j=1}^{s} p_{ij}^{n_{ij}},$$

where $n_{ij} = \sum_{k=0}^{n-1} 1_{\{y_k=i, y_{k+1}=j\}}$. Then, the posterior distribution of model parameters becomes

$$\Pr(\mathbf{P} \mid \mathbf{y}) \propto \prod_{i=1}^{s}\prod_{j=1}^{s} p_{ij}^{n_{ij}} \prod_{i=1}^{s} \Pr(p_{i1},\ldots,p_{is}) \propto \prod_{i=1}^{s}\prod_{j=1}^{s} p_{ij}^{n_{ij}} \prod_{i=1}^{s} p_{i1}^{\alpha_{i1}-1}\cdots p_{is}^{\alpha_{is}-1} = \prod_{i=1}^{s} p_{i1}^{n_{i1}+\alpha_{i1}-1}\cdots p_{is}^{n_{is}+\alpha_{is}-1}.$$

So *a posteriori* rows of $\mathbf{P}$ are independent and

$$(p_{i1},\ldots,p_{is}) \mid \mathbf{y} \sim \text{Dirichlet}(n_{i1}+\alpha_{i1},\ldots,n_{is}+\alpha_{is}).$$

If we take posterior means as Bayesian point estimates, then

$$\hat{p}_{ij}^{B} = \text{E}(p_{ij} \mid \mathbf{y}) = \frac{n_{ij}+\alpha_{ij}}{n_{i+}+\alpha_{i+}}.$$

Note 4.1. Bayesian analysis of parametric models for Markov chains depends is analogous to the nonparametric case, except the priors need to be selected according to the particular parameterization of transition probabilities.

4.1.2.2 Hypothesis testing: likelihood ratio test, Bayes factors and predictive distribution

Suppose that H_a is a hypothesis saying that the transition probability matrix has the following parametric form $\mathbf{P} = \mathbf{P}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We would like to compare this parameterization to a nested null hypothesis H_0: $\mathbf{P} = \mathbf{P}(\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$.

Let $\hat{l} = l(\hat{\boldsymbol{\theta}})$, where l is the log likelihood and $\hat{\boldsymbol{\theta}}$ is the mle under H_a. Similarly, let $\hat{l}_0 = l(\hat{\boldsymbol{\theta}}_0)$, where $\hat{\boldsymbol{\theta}}_0$ is the mle under H_0. Then $u = 2(\hat{l} - \hat{l}_0) \geq 0$ and $u \sim \chi^2_q$, where q is the difference in the number of free parameters between H_a and H_0 (Billingsley, 1961).

Example: *Testing independence against Markov dependence*

H_a: $X_1,\ldots,X_n \mid X_0$ is a Markov chain with unknown transition probabilities p_{ij}

H_0: $X_1,\ldots,X_n \mid X_0$ are iid observations of a discrete r.v. with probability mass function $p_1,\ldots,p_s$

The null hypothesis can be restated as $X_1,\ldots,X_n|X_0$ is a Markov chain with transition probabilities $p_{ij} = p_j$ for all i. So under $\mathscr{H}$, $\hat{p}_{ij} = n_{ij}/n_{i+}$ and under $\mathscr{H}_0$, $\hat{p}_j = n_{+j}/n$. So

$$u = 2(\hat{l} - \hat{l}_0) = 2\sum_{i=1}^{s}\sum_{j=1}^{s} n_{ij} \ln\left(\frac{\hat{p_{ij}}}{\hat{p}_j}\right) = 2\sum_{i=1}^{s}\sum_{j=1}^{s} n_{ij} \ln\left(\frac{n_{ij}/n_{i+}}{n_{+j}/n}\right).$$

The number of free parameters under $\mathscr{H}$ is $s^2 - s = s(s-1)$, while it is $s-1$ under $\mathscr{H}_0$. So under the null $u \sim \chi^2_{(s-1)^2}$. We would reject $\mathscr{H}_0$ if $\Pr(u > u_{\text{obs}} \mid \mathscr{H}_0) < \alpha$ and fail to reject otherwise, where α is an arbitrary significance level.

Example: *Snoqualmie Falls continued* For the two-state precipitation Markov model, the null hypothesis of iid sampling corresponds to

$$\mathrm{H}_0 : p_{21} = 1 - p_{12} = p.$$

The maximum likelihood estimate of this probability under the null hypothesis is $\hat{p} = n_{+1}/n = 771/1080 = 0.714$. Using the maximum likelihood estimates of the unrestricted transition probabilities of the two-state Markov model, we arrive at

$$u_{obs} = 2(\hat{l} - \hat{l}_0) = 2(643\ln 0.83 + 128\ln 0.16 + 123\ln 0.398 + 186\ln 0.60 - 771\ln 0.71 - 309\ln 0.29) = 184.5$$

Under H_0, $u \sim \chi^2_1$. Therefore $\Pr(u \geq u_{obs} \mid \mathscr{H}_0) = 5 \times 10^{-42}$ - a very unlikely event, so we reject the null hypothesis of iid sampling.

Now, let us compare the unrestricted model H_a with the iid model H_0 using Bayes factors. Suppose that under H_0 $p_{21} = 1 - p_{12} = p \sim \mathrm{Beta}(\alpha, \beta)$. Then the integrated likelihood of the Snoqualmie Fall data for this model is

$$\begin{aligned}\Pr(\mathbf{n} \mid \mathrm{H}_0) &= \int_0^1 \Pr(\mathbf{n} \mid \mathrm{H}_0, p) \Pr(p) dp = \int_0^1 p^{n_{+1}+\alpha-1}(1-p)^{n-n_{+1}+\beta-1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} dp \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(n_{+1}+\alpha)\Gamma(n-n_{+1}+\beta)}{\Gamma(n+\alpha+\beta)}.\end{aligned}$$

If we assume a two-state Markov model and put independent beta priors on p_{12} and p_{21}:

$$p_{12} \sim \mathrm{Beta}(\alpha_1, \beta_1) \qquad p_{21} \sim \mathrm{Beta}(\alpha_2, \beta_2),$$

then

$$\begin{aligned}&\Pr(\mathbf{n} \mid \mathrm{H}_a) = \int_0^1 \int_0^1 \Pr(\mathbf{n} \mid \mathrm{H}_a, p_{12}, p_{21}) \Pr(p_{12}) \Pr(p_{21}) dp \\ &= \int_0^1 \int_0^1 p_{12}^{n_{12}+\alpha_1-1}(1-p_{12})^{n_{11}+\beta_1-1} p_{21}^{n_{21}+\alpha_2-1}(1-p_{12})^{n_{22}+\beta_2-1} \\ &\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)} dp_{12} dp_{21} \\ &= \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(n_{12}+\alpha_1)\Gamma(n_{11}+\beta_1)}{\Gamma(n_{1+}+\alpha_1+\beta_1)} \frac{\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)} \frac{\Gamma(n_{21}+\alpha_2)\Gamma(n_{22}+\beta_2)}{\Gamma(n_{2+}+\alpha_2+\beta_2)}.\end{aligned}$$

If we set $\alpha = \alpha_1 = \alpha_2 = \beta = \beta_1 = \beta_2 = 1$, then $\Pr(\mathbf{n} \mid \mathrm{H}_a) / \Pr(\mathbf{n} \mid \mathrm{H}_0) = 10^{38}$, providing strong evidence in favor of the Markov model. Changing prior parameters will change the Bayes factor though. For example, setting $\alpha = \alpha_1 = \alpha_2 = \beta = \beta_1 = \beta_2 = 100$, a strong prior assessment that the transition probabilities all are 1/2, will make the ratio of the two marginal likelihoods 10^{27}.

A third approach is to compute the predictive distribution of the counts of pairs of outcomes (such as RR) under the null hypothesis. First note that even under the independence distribution these counts are not independent. Rather, they follow a snake chain (see Problem 2.4.7). We compute the predictive distribution by simulating from the snake chain obtained from independent draws of Bernoulli random variables with success probability drawn randomly from the asymptotic distribution of the maximum likelihood estimator of the probability of rain or dry weather. The result of this simulation is shown in Figure 4.2

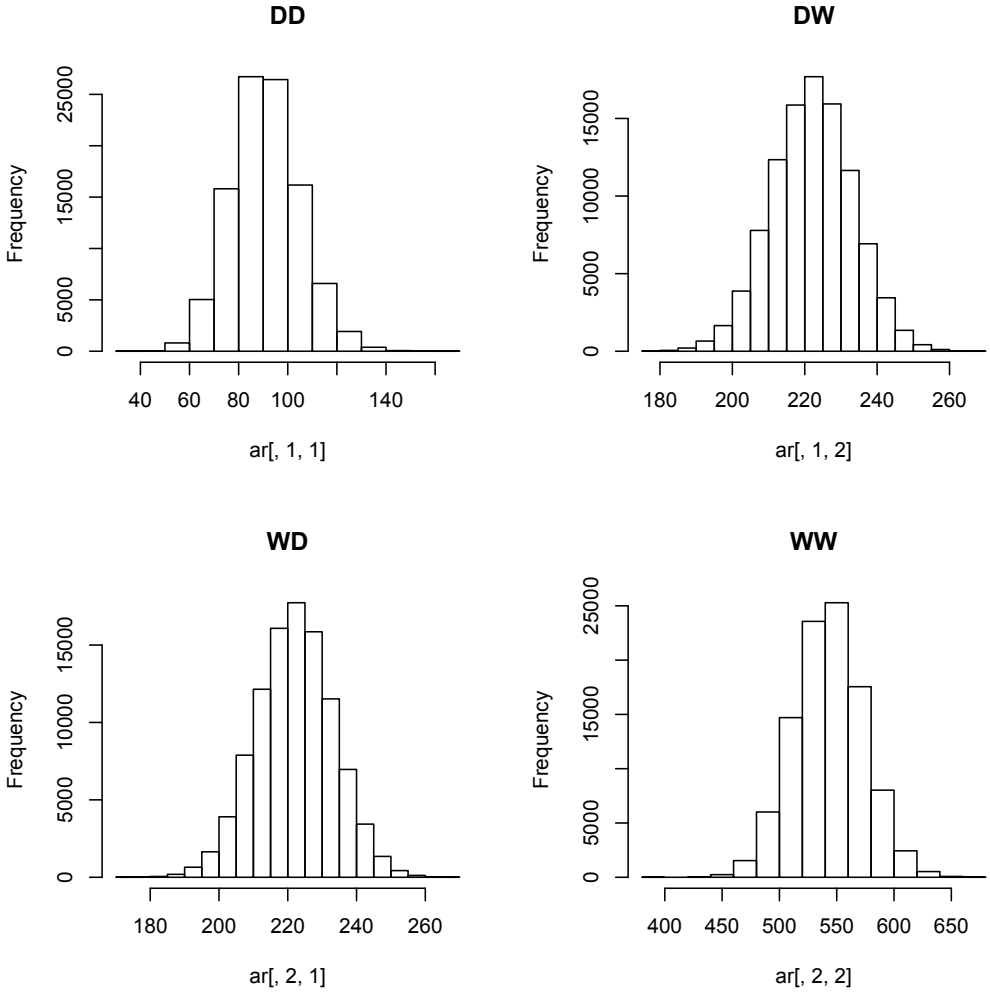

FIGURE 4.2: Predictive distribution of counts of pairs of outcomes under the null distribution of independence.

4.2 Hidden Markov chains

4.2.1 Generalities

Let us start with a Markov chain with finite state space $\Omega = \{1,\dots,s\}$, transition probability matrix $\mathbf{P}$ and initial distribution $\boldsymbol{\nu} = (\nu_1,\dots,\nu_s)$. We slightly change our notation and write

$$p(j|i) \equiv p_{ij}.$$

We also introduce new vector notation: $\mathbf{a}_{-i} = (a_1,\dots,a_{i-1},a_{i+1},\dots,a_n)$ and $\mathbf{a}_{i:j} = (a_i,\dots,a_j)$. Let $\mathbf{x} = (x_1,\dots,x_n)$ be a path of the Markov chain. Then

$$\Pr(\mathbf{x}) = \nu(x_1)\prod_{t=2}^{n} p(x_t|x_{t-1}).$$

We assume that the sequence $\mathbf{x}$ is hidden and we observe it only through a vector $\mathbf{y} = (y_1,\dots,y_n)$, where $y_t \in \mathscr{M} = \{1,\dots,m\}$. We further assume that given $\mathbf{x}$ and $\mathbf{y}_{-t}$, y_t depends only only on x_t:

$$\Pr(y_t \,|\, \mathbf{x},\mathbf{y}_{-t}) = \Pr(y_t \,|\, x_t) \equiv e(y_t|x_t).$$

As a result, we have

$$\Pr(\mathbf{y} \,|\, \mathbf{x}) = \prod_{t=1}^{n} e(y_t|x_t).$$

In summary, the hidden Markov model (HMM) parameters are:

- **Initial distribution**: $\boldsymbol{\nu} = (\nu(1),\dots,\nu(s))$.
- **Transition probabilities**: $\mathbf{P} = \{p(j|i)\}$, $i,j \in \{1,\dots,s\}$.
- **Emission probabilities**: $\mathbf{E} = \{e(k|i)\}$, $k \in \{1,\dots,m\}$, $i \in \{1,\dots,s\}$.

Example: *Occasionally dishonest casino* A casino switches between loaded and fair dice according to a Markov chain on the states space $\{L,F\}$ with initial distribution $\boldsymbol{\nu} = (0.5,0.5)$ and transition probabilities

$$\mathbf{P} = \begin{pmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{pmatrix},$$

and emission probabilities

$$\mathbf{E} = \begin{pmatrix} \frac{1}{3} & \frac{1}{4} & \frac{1}{6} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}.$$

Given a sequence of observations, e.g. 2 4 4 5 4 2 6 6 6 3 2 3 4 1 2 1 1 $\cdots$, we would like to identify which die was used for each toss in the sequence.

In a hidden Markov chain context we are interested in three types of calculations:

1. Likelihood evaluation. Given model parameters $(\boldsymbol{\nu},\mathbf{P},\mathbf{E})$, we would like to be able to rapidly evaluate $\Pr(\mathbf{y})$.
2. Hidden state inference. Given $\mathbf{y}$ and $(\boldsymbol{\nu},\mathbf{P},\mathbf{E})$, we would like to reconstruct $\mathbf{x}$. This task has many names: filtering, smoothing, decoding,...
3. Parameter estimation. We would like to infer $(\boldsymbol{\nu},\mathbf{P},\mathbf{E})$ from the observed data $\mathbf{y}$.

We will consider these three types of calculations in each of the following three subsections.

4.2.2 Likelihood evaluation with forward and backward algorithms

$$\Pr(\mathbf{y}) = \sum_{\mathbf{x}} \Pr(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{x}} \Pr(\mathbf{y}\,|\,\mathbf{x}) \Pr(\mathbf{x}) = \sum_{\mathbf{x}} \prod_{t=1} e(y_t|x_t)\nu(x_1) \prod_{t=2} p(x_t|x_{t-1})$$
$$= \sum_{\mathbf{x}} \nu(x_1) e(y_1|x_1) \prod_{t=2} p(x_t|x_{t-1}) e(y_t|x_t).$$

Since there are s^n elements in this sum, we need to be clever about economizing computations.

Consider the case of $n = 3$:

$$\Pr(y_1,y_2,y_3) = \sum_{x_1\in S}\sum_{x_2\in S}\sum_{x_3\in S} \nu(x_1)e(y_1|x_1)p(x_2|x_1)e(y_2|x_2)p(x_3|x_2)e(y_3|x_3)$$
$$= \sum_{x_1\in S}\sum_{x_2\in S} \nu(x_1)e(y_1|x_1)p(x_2|x_1)e(y_2|x_2) \left[\sum_{x_3\in S} p(x_3|x_2)e(y_3|x_3)\right]$$
$$= \underbrace{\sum_{x_1\in S} \nu(x_1)e(y_1|x_1) \underbrace{\left\{\sum_{x_2\in S} p(x_2|x_1)e(y_2|x_2) \underbrace{\left[\sum_{x_3\in S} p(x_3|x_2)e(y_3|x_3)\right]}_{b_2(x_2)}\right\}}_{b_1(x_1)}}_{b_0=\Pr(\mathbf{y})}.$$

What did we gain by these manipulations? Naive summation over the hidden states results in the algorithm with computational complexity $O(ns^n)$. Recursively computing $b_t(1),\ldots,b_t(s)$, we have arrived at a dynamic programming algorithm with computational complexity $O(ns^2)$. Let us have a closer look at $b_t(i)$. For $t = 2$ and $t = 1$,

$$b_2(i) = \sum_{x_3\in S} p(x_3|x_2 = i)e(y_3|x_3) = \sum_{x_3\in S} \Pr(y_3\,|\,x_3, x_2 = i) \Pr(x_3\,|\,x_2 = i) = \sum_{x_3\in S} \Pr(y_3, x_3\,|\,x_2 = i)$$
$$= \Pr(y_3\,|\,x_2 = i),$$

$$b_1(i) = \sum_{x_2\in S}\sum_{x_3\in S} p(x_2|x_1 = i)e(y_2|x_2)p(x_3|x_2)e(y_3|x_3)$$
$$= \sum_{x_2\in S}\sum_{x_3\in S} \Pr(x_2\,|\,x_1 = i) \Pr(y_2\,|\,x_2) \Pr(x_3\,|\,x_2) \Pr(y_3\,|\,x_3)$$
$$= \sum_{x_2\in S}\sum_{x_3\in S} \Pr(y_2\,|\,x_1 = i, x_2, x_3, y_3) \Pr(y_3\,|\,x_1 = i, x_2, x_3) \Pr(x_2, x_3\,|\,x_1 = i)$$
$$= \sum_{x_2\in S}\sum_{x_3\in S} \Pr(y_2, y_3\,|\,x_1 = i, x_2, x_3) \Pr(x_2, x_3\,|\,x_1 = i)$$
$$= \sum_{x_2\in S}\sum_{x_3\in S} \Pr(x_2, x_3, y_2, y_3\,|\,x_1 = i) = \Pr(y_2, y_3\,|\,x_1 = i).$$

In general,

$$b_t(i) = \Pr(\mathbf{y}_{t+1:n}\,|\,x_t = i),$$

which we call **backward probabilities**. We are ready to formulate what is often called the backward algorithm.

Algorithm 7 Backward algorithm

Initialize $b_n(i) = 1$ for $1 \le i \le s$.
for $t = n-1$ to 1 **do**
Compute

$$b_t(i) = \sum_{j=1}^{s} p(j|i)e(y_{t+1}|j)b_{t+1}(j) \text{ for } 1 \le i \le s.$$

End with $b_0 = \sum_{j=1}^{s} \nu(j)e(y_1|j)b_1(j)$.

Proof.

$$b_t(i) = \Pr(\mathbf{y}_{t+1:n} \,|\, x_t = i) = \sum_{j=1}^{s} \Pr(x_{t+1} = j, y_{t+1}, \mathbf{y}_{t+2:n} \,|\, x_t = i)$$

$$= \sum_{j=1}^{s} \Pr(\mathbf{y}_{t+2:n} \,|\, x_t = i, x_{t+1} = j, y_{t+1}) \Pr(x_{t+1} = j, y_{t+1} \,|\, x_t = i)$$

$$= [\text{conditional independence (needs a separate proof)}] =$$

$$= \sum_{j=1}^{s} \Pr(\mathbf{y}_{t+2:n} \,|\, x_{t+1} = j) \Pr(y_{t+1} \,|\, x_{t+1} = j, x_t = i) \Pr(x_{t+1} = j \,|\, x_t = i) = \sum_{j=1}^{s} p(j|i)e(y_{t+1}|j)b_{t+1}(j)$$

□

Similarly we can move our conditioning from the front.. Again, start with $n = 3$:

$$\Pr(y_1, y_2, y_3) = \sum_{x_1 \in S} \sum_{x_2 \in S} \sum_{x_3 \in S} \nu(x_1)e(y_1|x_1)p(x_2|x_1)e(y_2|x_2)p(x_3|x_2)e(y_3|x_3)$$

$$= \underbrace{\sum_{x_3 \in S} e(y_3|x_3) \underbrace{\sum_{x_2 \in S} p(x_3|x_2) \underbrace{e(y_2|x_2) \underbrace{\sum_{x_1 \in S} p(x_2|x_1) \underbrace{\nu(x_1)e(y_1|x_1)}_{a_1(x_1)}}_{a_2(x_2)}}_{a_3(x_3)}}_{\Pr(\mathbf{y}) = \sum_{x_3} a_3(x_3)}},$$

where

$$a_t(i) = \Pr(\mathbf{y}_{1:t}, x_t = i),$$

are the **forward probabilitiies**.

Algorithm 8 Forward algorithm

Initialize $a_1(i) = \nu(i)e(y_1|i)$ for $1 \le i \le s$.
for $t = 1$ to $n-1$ **do**
Compute

$$a_{t+1}(i) = e(y_{t+1}|i) \left[\sum_{j=1}^{s} a_t(j)p(i|j) \right] \text{ for } 1 \le i \le s.$$

End with $\Pr(\mathbf{y}) = \sum_{j=1}^{s} \Pr(\mathbf{y}_{1:n}, x_n = i) = \sum_{j=1}^{s} a_n(j)$.

We have now two computationally efficient ways to compute the observed data likelihood. This likelihood calculation algorithm can be used to obtain maximum likelihood estimates of transition probabilities $\mathbf{P}$ and emission probabilities $\mathbf{E}$ by numerically maximizing the likelihood. Alternatively, we can perform Bayesian inference of these parameters with the help of MCMC.

4.2.3 Hidden state inference

Next we look at how one can estimate the hidden state. We have two different approaches, a marginal one where the state is estimated observation by observation and a simultaneous one where we estimate the most probable path.

Marginal approach:

Start with

$$\Pr(x_t = i \,|\, \mathbf{y}) = \frac{\Pr(\mathbf{y}, x_t = i)}{\Pr(\mathbf{y})} = \frac{\Pr(\mathbf{y}_{1:t}, \mathbf{y}_{t+1:n}, x_t = i)}{\Pr(\mathbf{y})} = \frac{\Pr(\mathbf{y}_{t+1:n} \,|\, \mathbf{y}_{1:t}, x_t = i) \Pr(\mathbf{y}_{1:t}, x_t = i)}{\Pr(\mathbf{y})}$$
$$= [\text{conditional independence (needs proof)}] = \frac{\Pr(\mathbf{y}_{t+1:n} \,|\, x_t = i) \Pr(\mathbf{y}_{1:t}, x_t = i)}{\Pr(\mathbf{y})}$$
$$= \frac{a_t(i) b_t(i)}{\sum_{i=1} a_t(i) b_t(i)}.$$

Having computed the marginal probability mass function of x_t conditional on the observed data $\mathbf{y}$, we estimate x_t with with marginal posterior mode:

$$\hat{x}_t = \arg\max_{1 \le i \le s} \Pr(x_t = i \,|\, \mathbf{y}).$$

Note 4.2. Ignoring the joint distribution of $\mathbf{x} \,|\, \mathbf{y}$ creates problems. For example, it is possible to end up with estimates

$$\hat{x}_1, \ldots, \hat{x}_t, \hat{x}_{t+1}, \ldots, \hat{x}_n$$

such that $\Pr(\hat{x}_{t+1} \,|\, \hat{x}_t) = 0$.

The most probable path approach:

Instead of reconstructing hidden states one by one, we impute them jointly by computing

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \Pr(\mathbf{x} \,|\, \mathbf{y}) = \arg\max_{\mathbf{x}} \frac{\Pr(\mathbf{x}, \mathbf{y})}{\Pr(\mathbf{y})} = \arg\max_{\mathbf{x}} \Pr(\mathbf{x}, \mathbf{y}).$$

$\mathbf{x}^*$ is called maximum *a posteriori* (MAP) estimate.

As in the likelihood evaluation, we would like to compute $\mathbf{x}^*$ without examining s^n possible hidden state configurations. This time, instead of the distributive law of multiplication we will be using

$$\max_{i,j} f(i,j) = \max_j \left[\max_i f(i,j) \right].$$

First, we observe that

$$\Pr(\mathbf{x}_{1:t+1}, \mathbf{y}_{1:t+1}) = \Pr(x_{t+1}, y_{t+1} \,|\, \mathbf{x}_{1:t}, \mathbf{y}_{1:t}) \Pr(\mathbf{x}_{1:t}, \mathbf{y}_{1:t}) = \Pr(y_{t+1} \,|\, x_{t+1}, \mathbf{x}_{1:t}, \mathbf{y}_{1:t}) \Pr(x_{t+1} \,|\, \mathbf{x}_{1:t}, \mathbf{y}_{1:t})$$
$$\times \Pr(\mathbf{x}_{1:t}, \mathbf{y}_{1:t}) = [\text{cond ind}] = \Pr(y_{t+1} \,|\, x_{t+1}) \Pr(x_{t+1} \,|\, x_t) \Pr(\mathbf{x}_{1:t}, \mathbf{y}_{1:t}) = e(y_{t+1}|x_{t+1}) p(x_{t+1}|x_t)$$
$$\times \Pr(\mathbf{x}_{1:t}, \mathbf{y}_{1:t}).$$

Define

$$d_t(i) = \max_{\mathbf{x}_{1:t-1}} \Pr(\mathbf{x}_{1:t-1}, x_t = i, \mathbf{y}_{1:t}).$$

Then,

$$d_t(i) = \max_{\mathbf{x}_{1:t-1}} \left[\Pr(\mathbf{x}_{1:t-1}, \mathbf{y}_{1:t-1}) e(y_t|i) p(i|x_{t-1}) \right] = e(y_t|i) \max_{\mathbf{x}_{1:t-1}} \left[\Pr(\mathbf{x}_{1:t-1}, \mathbf{y}_{1:t-1}) p(i|x_{t-1}) \right]$$
$$= e(y_t|i) \max_{x_{t-1}} \left\{ \max_{\mathbf{x}_{1:t-2}} \left[\Pr(\mathbf{x}_{1:t-1}, \mathbf{y}_{1:t-1}) p(i|x_{t-1}) \right] \right\} = e(y_t|i) \max_{x_{t-1}} \left\{ p(i|x_{t-1}) \max_{\mathbf{x}_{1:t-2}} \left[\Pr(\mathbf{x}_{1:t-1}, \mathbf{y}_{1:t-1}) \right] \right\}$$
$$= e(y_t|i) \max_{x_{t-1}} \left[p(i|x_{t-1}) d_{t-1}(j) \right].$$

This suggests that we compute $d_t(i)$ recursively (Viterbi, 1967):

Algorithm 9 Viterbi algorithm

Initialize $d_1(i) = \nu(i)e(y_1|i)$ for $1 \le i \le s$.
for $t = 2$ to n **do**
Compute
$$d_t(i) = \max_{1\le j\le s} [d_{t-1}(j)p(i|j)]\, e(y_t|i) \text{ for } 1 \le i \le s,$$
$$f_t(i) = \arg\max_{1\le j\le s} [d_{t-1}(j)p(i|j)] \text{ for } 1 \le i \le s.$$

End with $p^* = \max_{1\le j\le s} d_n(j)$ and $x_n^* = \arg\max_{1\le j\le s} d_n(j)$.
for $t = n-1$ to 1 **do**
$x_t^* = f_{t+1}(x_{t+1}^*)$ (*backtracking*)
return $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$.

Note 4.3. Backtracking follows from the following backward recursion

$$x_t^* = \arg\max_j p(x_{t+1}^*|j)d_t(j)$$

with initial condition $x_n^* = \arg\max_j d_n(j)$. The logic behind this recursive relationship is similar to the recursion for $d_t(i)$ and realization of the following fact:

$$(i^*, j^*) = \arg\max_{i,j} f(i,j) \qquad \Rightarrow \qquad j^* = \arg\max_j \left[\max_i f(i,j)\right].$$

Note 4.4. For large n, one often runs into round-off errors while running the Viterbi algorithm, because $d_t(i)$s become very small. To avoid this problem, we move all calculations to the log scale by defining

$$\tilde{d}_t(i) = \max_{\mathbf{x}_{1:t-1}} \ln \Pr(\mathbf{x}_{1:t-1}, x_t = i, \mathbf{y}_{1:t}),$$

replacing the initialization of the Viterbi algorithm with

$$\tilde{d}_1(i) = \ln \nu(i) + \ln e(y_1|i) \text{ for } 1 \le i \le s$$

and the recursion step with

$$\tilde{d}_t(i) = \max_{1\le j\le s} \left[\tilde{d}_t(j) + \ln p(i|j)\right] + \ln e(y_t|i) \text{ for } 1 \le i \le s.$$

4.2.4 Parameter estimation

American mathematician Leonard Baum and information theorist Lloyd Welch independently developed the algorithm for maximizing the likelihood of a hidden Markov model. According to Welch, they "then joined forces." Baum and his co-workers proved that each iteration of the algorithm increased the likelihood and published the algorithm and corresponding theoretical results in a series of papers. Welch was not a co-author on any of these papers.

We now consider estimating the parameters of the model. If we could observe the hidden state the estimation would be straightforward.

4.2.4.1 The Baum-Welch algorithm

Baum and Petrie (1966) proposed an EM-type algorithm to do the estimation when the hidden state is unobserved.

Let $\boldsymbol{\theta} = (\mathbf{P}, \mathbf{E}, \boldsymbol{\nu})$. For HMMs, the complete data consists of both hidden and observed states, $(\mathbf{x}, \mathbf{y})$. Therefore, the complete data likelihood is

$$\Pr(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \nu(x_1) \prod_{t=2}^{n} p(x_t|x_{t-1}) \prod_{t=1}^{n} e(y_t|x_t).$$

Taking the expectation of the above function with respect to the distribution of complete data conditional on the observed data we arrive at the expected complete data log likelihood:

$$\begin{aligned}
&\mathrm{E}[\ln \Pr(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) \,|\, \mathbf{y}; \boldsymbol{\theta}_k] = \sum_{\mathbf{x}} \ln \Pr(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) \Pr(\mathbf{x} \,|\, \mathbf{y}; \boldsymbol{\theta}_k) \\
&= \sum_{\mathbf{x}} \left[\ln \nu(x_1) + \sum_{t=2}^{n} \ln p(x_t|x_{t-1}) + \sum_{t=1}^{n} \ln e(y_t|x_t) \right] \Pr(\mathbf{x} \,|\, \mathbf{y}; \boldsymbol{\theta}_k) \\
&= \underbrace{\sum_{\mathbf{x}} \ln \nu(x_1) \Pr(\mathbf{x} \,|\, \mathbf{y}; \boldsymbol{\theta}_k)}_{F_1(\boldsymbol{\nu})} + \underbrace{\sum_{\mathbf{x}} \sum_{t=2}^{n} \ln p(x_t|x_{t-1}) \Pr(\mathbf{x} \,|\, \mathbf{y}; \boldsymbol{\theta}_k)}_{F_2(\mathbf{P})} + \underbrace{\sum_{\mathbf{x}} \sum_{t=1}^{n} \ln e(y_t|x_t) \Pr(\mathbf{x} \,|\, \mathbf{y}; \boldsymbol{\theta}_k)}_{F_3(\mathbf{E})}.
\end{aligned}$$

The additive form of the expected complete data log likelihood allows us to maximize $F_1(\boldsymbol{\nu})$, $F_2(\mathbf{P})$, and $F_3(\mathbf{E})$ separately.

Updating $\boldsymbol{\nu}$*:*

$$\begin{aligned}
&F_1(\boldsymbol{\nu}) = \sum_{\mathbf{x}} \ln \nu(x_1) \Pr(\mathbf{x} \,|\, \mathbf{y}; \boldsymbol{\theta}_k) = \sum_{x_1} \ln \nu(x_1) \sum_{\mathbf{x}_{2:n}} \Pr(\mathbf{x} \,|\, \mathbf{y}; \boldsymbol{\theta}_k) = \sum_{x_1} \ln \nu(x_1) \Pr(x_1 \,|\, \mathbf{y}; \boldsymbol{\theta}_k) \\
&= \sum_{i} \ln \nu(i) \gamma_1(i, \boldsymbol{\theta}_k),
\end{aligned}$$

where

$$\gamma_t(i, \boldsymbol{\theta}_k) \equiv \Pr(x_t = i \,|\, \mathbf{y}; \boldsymbol{\theta}_k) = \frac{a_t(i) b_t(i)}{\sum_{i=1} a_t(i) b_t(i)}.$$

Maximizing $F_1(\boldsymbol{\nu})$ via Lagrange multipliers, we obtain

$$\nu_{k+1}(i) = \gamma_1(i, \boldsymbol{\theta}_k) = \text{imputed frequency of state i at time 1.}$$

Updating $\mathbf{P}$*:*

$$\begin{aligned}
&F_2(\mathbf{P}) = \sum_{\mathbf{x}} \left[\sum_{t=2}^{n} \ln p(x_t|x_{t-1}) \right] \Pr(\mathbf{x} \,|\, \mathbf{y}; \boldsymbol{\theta}_k) = \sum_{t=2}^{n} \sum_{x_t} \sum_{x_{t-1}} \ln p(x_t|x_{t-1}) \sum_{\mathbf{x}_{1:t-2}, \mathbf{x}_{t+1:n}} \Pr(\mathbf{x} \,|\, \mathbf{y}; \boldsymbol{\theta}_k) \\
&= \sum_{t=2}^{n} \sum_{i} \sum_{j} \ln p(j|i) \Pr(x_{t-1} = i, x_t = j \,|\, \mathbf{y}; \boldsymbol{\theta}_k).
\end{aligned}$$

Define

$$g_t(i, j; \boldsymbol{\theta}_k) = \sum_{t=2}^{n} \sum_{i} \sum_{j} \ln p(j|i) \Pr(x_{t-1} = i, x_t = j \,|\, \mathbf{y}; \boldsymbol{\theta}_k)$$

and assume for the moment that we can calculate these quantities. Then

$$F_2(\mathbf{P}) = \sum_{t=2}^{n} \sum_{i} \sum_{j} \ln p(j|i) g_t(i, j; \boldsymbol{\theta}_k) = \sum_{i} \sum_{j} \left[\sum_{t=2}^{n} g_t(i, j; \boldsymbol{\theta}_k) \right] \ln p(j|i).$$

This maximization problem is analogous to the fully observed Markov chain log likelihood maximization except here $n_{ij} = \sum_{t=2}^{n} g_t(i,j;\boldsymbol{\theta}_k)$ and $n_{i+} = \sum_j \sum_{t=2}^{n} g_t(i,j;\boldsymbol{\theta}_k)$. Therefore, maximization using Lagrange multipliers results in

$$p_{k+1}(j|i) = \frac{\sum_{t=2}^{n} g_t(i,j;\boldsymbol{\theta}_k)}{\sum_j \sum_{t=2}^{n} g_t(i,j;\boldsymbol{\theta}_k)} = \frac{\sum_{t=2}^{n} g_t(i,j;\boldsymbol{\theta}_k)}{\sum_{t=2}^{n} \sum_j \Pr(x_{t-1}=i, x_t=j \,|\, \mathbf{y};\boldsymbol{\theta}_k)} = \frac{\sum_{t=2}^{n} g_t(i,j;\boldsymbol{\theta}_k)}{\sum_{t=2}^{n} \Pr(x_{t-1}=i \,|\, \mathbf{y};\boldsymbol{\theta}_k)}$$
$$= \frac{\sum_{t=2}^{n} g_t(i,j;\boldsymbol{\theta}_k)}{\sum_{t=2}^{n} \gamma_{t-1}(i;\boldsymbol{\theta}_k)} = \frac{\text{expected number of } i \to j \text{ transitions}}{\text{expected number of transitions from } i}.$$

Now, let us return to the problem of calculating $g_t(i,j;\boldsymbol{\theta}_k)$:

$$g_t(i,j;\boldsymbol{\theta}_k) = \Pr(x_{t-1}=i, x_t=j \,|\, \mathbf{y};\boldsymbol{\theta}_k) = \frac{\Pr(x_{t-1}=i, x_t=j, \mathbf{y};\boldsymbol{\theta}_k)}{\Pr(\mathbf{y};\boldsymbol{\theta}_k)}$$
$$= \frac{\Pr(\mathbf{y}_{t+1:n} \,|\, x_{t-1}=i, x_t=j, \mathbf{y}_{1:t};\boldsymbol{\theta}_k) \Pr(x_{t-1}=i, x_t=j, \mathbf{y}_{1:t};\boldsymbol{\theta}_k)}{\Pr(\mathbf{y};\boldsymbol{\theta}_k)} = [\text{cond ind}]$$
$$= \frac{\Pr(\mathbf{y}_{t+1:n} \,|\, x_t=j;\boldsymbol{\theta}_k) \Pr(y_t \,|\, x_{t-1}=i, x_t=j, \mathbf{y}_{1:t-1};\boldsymbol{\theta}_k) \Pr(x_{t-1}=i, x_t=j, \mathbf{y}_{1:t-1};\boldsymbol{\theta}_k)}{\Pr(\mathbf{y};\boldsymbol{\theta}_k)} = [\text{cond ind}]$$
$$= \frac{\Pr(\mathbf{y}_{t+1:n} \,|\, x_t=j;\boldsymbol{\theta}_k) \Pr(y_t \,|\, x_t=j;\boldsymbol{\theta}_k) \Pr(x_t \,|\, x_{t-1}=i, \mathbf{y}_{1:t-1};\boldsymbol{\theta}_k) \Pr(x_{t-1}=i, \mathbf{y}_{1:t-1};\boldsymbol{\theta}_k)}{\Pr(\mathbf{y};\boldsymbol{\theta}_k)} = [\text{cond ind}]$$
$$= \frac{\Pr(\mathbf{y}_{t+1:n} \,|\, x_t=j;\boldsymbol{\theta}_k) \Pr(y_t \,|\, x_t=j;\boldsymbol{\theta}_k) \Pr(x_t \,|\, x_{t-1}=i;\boldsymbol{\theta}_k) \Pr(x_{t-1}=i, \mathbf{y}_{1:t-1};\boldsymbol{\theta}_k)}{\Pr(\mathbf{y};\boldsymbol{\theta}_k)}$$
$$= \frac{b_t(j,\boldsymbol{\theta}_k) e_k(y_t|j) p_k(j|i) a_{t-1}(i;\boldsymbol{\theta}_k)}{\sum_i \sum_j b_t(j,\boldsymbol{\theta}_k) e_k(y_t|j) p_k(j|i) a_{t-1}(i;\boldsymbol{\theta}_k)}.$$

Updating **E***:*

$$F_3(\mathbf{E}) = \sum_{\mathbf{x}} \sum_{t=1}^{n} \ln e(y_t|x_t) \Pr(\mathbf{x} \,|\, \mathbf{y};\boldsymbol{\theta}_k) = \sum_{t=1}^{n} \sum_{x_t} e(y_t|x_t) \sum_{\mathbf{x}_{-t}} \Pr(\mathbf{x}_{1:n} \,|\, \mathbf{y};\boldsymbol{\theta}_k)$$
$$\sum_{t=1}^{n} \sum_{x_t} e(y_t|x_t) \Pr(\mathbf{x}_t \,|\, \mathbf{y};\boldsymbol{\theta}_k) = \sum_{x_t} \sum_{t=1}^{n} e(y_t|x_t) \gamma_t(i;\boldsymbol{\theta}_k).$$

Again, using Lagrange multipliers, we can show that

$$e_{k+1}(l|i) = \frac{\sum_{t=1}^{n} \gamma_t(i;\boldsymbol{\theta}_k) 1_{\{y_t=l\}}}{\sum_{t=1}^{n} \gamma_t(i;\boldsymbol{\theta}_k)} = \frac{\text{expected number of times state } i \text{ emits value } l}{\text{expected number of times } x_n \text{ is in state } i}.$$

Note 4.5. Backward and forward algorithms also run into roundoff errors. Unfortunately, it is impossible to transfer these calculations onto a log scale, because the backward and forward algorithms involve both products and sums. The computations can be stabilized by rescaling. For example, at each step we can rescale forward probabilities $a_t(i)$ by a factor $c_t = [\sum_i a_t(i)]^{-1}$ to get $\tilde{a}_t(i) = c_t a_t(i)$. It turns out that $\tilde{a}_t = \prod_{\tau=1}^{t} c_\tau a_t(i)$ and $\ln \Pr(\mathbf{y}) = -\sum_{\tau=1}^{n} \ln c_\tau$. In the Baum-Welch algorithm, these dummy scaling factors cancel out so we can use $\tilde{a}_t(i)$ instead of $a_t(i)$. See (Rabiner, 1989) for a more detailed discussion of the implementation issues.

4.2.4.2 Bayesian data augmentation

From a Bayesian point of view, we are interested in the posterior distribution

$$\Pr(\boldsymbol{\theta} \,|\, \mathbf{y}) \propto \Pr(\mathbf{y} \,|\, \boldsymbol{\theta}) \Pr(\boldsymbol{\theta}),$$

Algorithm 10 Baum-Welch Algorithm

Start with $\boldsymbol{\theta}_0 = (\boldsymbol{\nu}_o, \mathbf{P}_0, \mathbf{E}_0)$.
repeat
Run forward and backward algorithms to calculate $\{a_t(i;\boldsymbol{\theta}_k)\}$ and $\{b_t(i;\boldsymbol{\theta}_k)\}$. Use these partial likelihood matrices to update HMM parameters:

$$\begin{aligned}
\nu_{k+1}(i) &= \gamma_1(i,\boldsymbol{\theta}_k),\\
p_{k+1}(j|i) &= \frac{\sum_{t=2}^{n} g_t(i,j;\boldsymbol{\theta}_k)}{\sum_j \sum_{t=2}^{n} g_t(i,j;\boldsymbol{\theta}_k)},\\
e_{k+1}(l|i) &= \frac{\sum_{t=1}^{n} \gamma_t(i;\boldsymbol{\theta}_k) 1_{\{y_t=l\}}}{\sum_{t=1}^{n} \gamma_t(i;\boldsymbol{\theta}_k)}
\end{aligned}$$

until $|\ln \Pr(\mathbf{y};\boldsymbol{\theta}_k) - \ln \Pr(\mathbf{y};\boldsymbol{\theta}_{k-1})| < \varepsilon$ for a predefined $\varepsilon > 0$.
return $\boldsymbol{\theta}_k = (\boldsymbol{\nu}_k, \mathbf{P}_k, \mathbf{E}_k)$.

which we can, in principle, approximate via MCMC. If, however, we augment our state space with the hidden states $\mathbf{x}$, our interest shifts to the augmented posterior

$$\Pr(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}) \propto \Pr(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}) \Pr(\boldsymbol{\theta}).$$

Our MCMC sampler will alternate between sampling from $\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}$ and $\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y}$. It should be clear that placing independent Dirichlet priors on rows of $\mathbf{P}$ and $\mathbf{E}$ and initial distribution $\boldsymbol{\nu}$ makes the full conditional of these vectors also Dirichlet. Therefore sampling from $\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}$ is straightforward. Sampling $\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y}$ can also be accomplished exactly.

Stochastic backward recursion:

We omit dependence on $\boldsymbol{\theta}$ to simplify notation. We first notice that

$$\Pr(\mathbf{x} \,|\, \mathbf{y}) = \Pr(x_n \,|\, \mathbf{y}) \prod_{t=1}^{n-1} \Pr(x_{n-t} \,|\, \mathbf{x}_{n-t+1:n}, \mathbf{y}) = [\text{cond ind}] = \Pr(x_n \,|\, \mathbf{y}) \prod_{t=1}^{n-1} \Pr(x_{n-t} \,|\, x_{n-t+1}, \mathbf{y}).$$

This factorization invites the following sampling scheme:

$$\begin{aligned}
&\text{sample } x_n^* \mid \mathbf{y},\\
&\text{sample } x_{n-1}^* \mid x_n, \mathbf{y},\\
&\text{sample } x_{n-2}^* \mid x_{n-1}, \mathbf{y},\\
&\qquad\vdots\\
&\text{sample } x_1^* \mid x_2, \mathbf{y}.
\end{aligned}$$

The resulting $\mathbf{x}^*$ is joint sample from the distribution of hidden states conditional on the observed data. Since we know how to calculate the marginal distribution $\Pr(x_n \,|\, \mathbf{y})$, initializing the sampler is easy. Next, we need to understand how to compute $\Pr(x_t \,|\, x_{t+1}, \mathbf{y})$.

$$\Pr(x_{t-1} = i \,|\, x_t = j, \mathbf{y}) = \frac{\Pr(x_{t-1} = i, x_t = j \,|\, \mathbf{y})}{\Pr(x_t = j \,|\, \mathbf{y})} = \frac{g_t(i,j)}{\gamma_t(j)}.$$

Since $g_t(i,j)$ and $\gamma_t(j)$ are computable via the forward-backward algorithm, we now have an exact method for sampling hidden states conditional on the observed data. See Scott (2002) for more details.