We now write the log likelihood from observing $x_1, \ldots, x_n$ (as above we argue conditionally upon the initial state)

$$l(\boldsymbol{\theta}) = \sum_D n_{ij} \log p_{ij}(\boldsymbol{\theta}), \tag{2.204}$$

where as before $n_{ij}=\#\{0 \le k \le n-1 : x_k=i, x_{k+1}=j\}$. Differentiating we get the likelihood equations

$$\frac{\partial}{\partial \theta_k} L_n(\boldsymbol{\theta}) = \sum_D \frac{n_{ij}}{p_{ij}(\boldsymbol{\theta})} \frac{\partial p_{ij}(\boldsymbol{\theta})}{\partial \theta_k} = 0,\ k=1, \ldots, r. \tag{2.205}$$

Let $\boldsymbol{\theta}_0$ be the true parameter value. The following result is due to Billingsley (1961). We will not prove it here.

Theorem 2.16 Assume Conditions A.

(i) There is a consistent solution $\hat{\boldsymbol{\theta}}$ of the likelihood equations.

(ii) $\sqrt{n}\,(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \to \mathrm{N}(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0))$, where I is the **information matrix** with typical element

$$I_{uv}(\boldsymbol{\theta}_0) = \sum_{(i,j)\in D} \frac{\pi_i(\boldsymbol{\theta}_0)}{p_{ij}(\boldsymbol{\theta}_0)} \frac{\partial p_{ij}(\boldsymbol{\theta}_0)}{\partial \theta_u} \frac{\partial p_{ij}(\boldsymbol{\theta}_0)}{\partial \theta_v} \tag{2.206}$$

and $\pi_i(\boldsymbol{\theta}_0)$ is the stationary probability of state i.

(iii) $\mathbf{Var}\sqrt{n}\,(\hat{\theta}-\theta)$ can be consistently estimated by

$$\left[-\frac{n_{ij}}{n}\nabla^2 \log p_{ij}(\hat{\theta})\right]^{-1}. \tag{2.207}$$

The quantity inverted in (2.207) is called the **observed information.**

Application (Russian linguistics, continued) We estimate p by maximizing the log likelihood

$$l(p) = (n_{00}+n_{11})\log(1-p) + (n_{01}+n_{10})\log p, \tag{2.208}$$

where 1 denotes a consonant and 0 a vowel. The maximum is obtained at $\hat{p}=(n_{01}+n_{10})/n=(7{,}532+7{,}533)/20{,}000=0.753$. The second derivative of the log likelihood is

$$l''(p) = -\frac{n_{00}+n_{11}}{(1-p)^2} - \frac{n_{01}+n_{10}}{p^2} \tag{2.209}$$

so from Theorem 2.16 the asymptotic estimated standard error is $(-l''(\hat{p}))^{-\frac{1}{2}}=(\hat{p}(1-\hat{p})/n)^{\frac{1}{2}}$, yielding an asymptotic confidence interval for p of (0.747,0.759). Notice that neither $\hat{p}_{01}=0.872$ nor $\hat{p}_{10}=0.663$ fall inside this confidence interval, indicating that the simple one-parameter model is inadequate. □

It is straightforward to develop a likelihood ratio theory of testing hypotheses for Markov chains satisfying Conditions A. Here is a general result, again left without proof (see, e.g., Billingsley, 1961, Theorem 3.1).

Theorem 2.17 Assume Conditions A. Let $\hat{\boldsymbol{\theta}}$ be the mle under the parametric hypothesis H_0. Also, let $\hat{\mathbb{P}}$ be the nonparametric mle, and $\boldsymbol{\theta}_0$ the true value of $\boldsymbol{\theta}$, assuming that H_0 is true. Then

(a) $2(l(\hat{\boldsymbol{\theta}})-l(\boldsymbol{\theta}_0)) \xrightarrow{d} \chi^2(r)$

(b) $2(l(\hat{\mathbb{P}})-l(\hat{\boldsymbol{\theta}})) \xrightarrow{d} \chi^2(d(d-1)-r)$

(c) The statistics in (a) and (b) are asymptotically independent. □

Remark Under conditions similar to Conditions A it is possible to derive a result much like Theorem 2.17 for testing a parametric model against a submodel. That is, in fact, the result given by Billingsley (1961). □

Example (Testing for independence) Suppose we want to study the hypothesis that (x_k) is a sequence of iid random variables, taking values in $\{0,\ldots,K\}$ (so $d=K+1$). In terms of a parametrization this is simply $H_0: p_{ij}=\theta_j$ for all $i \in S$ and each $j \in S$. We must compute the maxima of the likelihood under the two models. We already know that $\hat{p}_{ij}=n_{ij}/n_i$. Under the independence assumption we have a multinomial distribution, with $n_{\bullet j}=\sum_i n_{ij}$ observations from the category with probability θ_j. The likelihood is

$$l(\boldsymbol{\theta}) = \sum_{j=0}^{K-1} n_{\bullet j}\theta_j + n_{\bullet K}(1-\sum_{j=0}^{K-1}\theta_j), \tag{2.210}$$

which is maximized by $\hat{\theta}_j=n_{\bullet j}/n$. Hence the log likelihood ratio statistic for testing H_0 is

$$2(l(\hat{\mathbb{P}})-l(\hat{\boldsymbol{\theta}})) = 2\sum_{i,j} n_{ij}\log\frac{n_{ij}/n_i}{n_{\bullet j}/n} \tag{2.211}$$

which asymptotically has a χ^2 distribution with $K(K+1)-K=K^2$ degrees of freedom. In the Snoqualmie Falls rain model we have $K=1$, in accordance with our earlier claim. □

Application (Sedimentology) An important aspect of geology is stratigraphy, the description of sedimental layers. An interesting question is whether or not there is memory in observed sequences of strata, or **facies.** If memory, or perhaps better preference, is present the conditional probability that facies B will be deposited on top of an observed facies A may then be different from the conditional probability given any other underlying facies.

Hiscott (1981) studied the Ordocovian Tournelle Formation in Québec and distinguished, based on field criteria, two different facies associations (or underlying transition probabilities) for six different facies:

Facies	Description
0	Thick shale
1	Interbedded graded siltstones and shale
2	Poorly sorted sandstones with dispersed clasts
3	Interbedded graded sandstones and shales
4	Amalgamated graded sandstones
5	Thick coarse sandstones

Of course, there are no $i \to i$ transitions. The transition counts for one of the facies associations are given in Table 2.4.

Table 2.4 Upward transition counts in the Tournelle Formation

Facies	0	1	2	3	4	5	
0	0	2	0	2	2	0	6
1	5	0	23	31	17	8	84
2	0	21	0	45	27	8	101
3	1	54	44	0	66	25	190
4	0	6	24	81	0	38	149
5	0	5	8	31	32	0	76
	6	88	99	190	144	79	606

Our null hypothesis is independence, i.e., no preference in the sedimentation. The log likelihood statistic works out to 405.9 with 4×6−5=19 degrees of freedom, since $K=5$ but the transition matrix is restricted to have zeros on the diagonal. Hence the P-value is 0, and we reject the hypothesis of independence. □

Remark As in the multinomial calculation in Appendix A, the log likelihood statistic is approximately a Pearson χ^2 statistic, in that

$$2\sum n_{ij}\log\frac{n_{ij}}{n_i p^0_{ij}} = \sum\frac{(n_{ij}-n_i p^0_{ij})^2}{n_i p^0_{ij}} + o_{\mathrm{P}}(1). \tag{2.212}$$

The latter form is sometimes more convenient to compute. □

Application (Russian linguistics, continued) We are testing the hypothesis $H_0: p_{10}=p_{01}$. The expected counts that we need for the χ^2 statistic are computed by multiplying the row sums (n_0,n_1)=(8,638, 11,362) with the transition matrix estimate under H_0

$$\hat{\mathbb{P}} = \begin{bmatrix} 0.247 & 0.753 \\ 0.753 & 0.247 \end{bmatrix} \tag{2.213}$$

yielding

$$\begin{bmatrix} 2131.4 & 6506.6 \\ 8558.4 & 2803.6 \end{bmatrix} \tag{2.214}$$

The χ^2 statistic for testing the one-dimensional null hypothesis given above within the general nonparametric Markov chain model is

$$\chi^2 = \sum \frac{(n_{ij}-n_i\hat{p}^0_{ij})^2}{n_i\hat{p}^0_{ij}} = 1217.7. \tag{2.215}$$

The exact likelihood ratio statistic is also 1217.7, so the approximation is excellent. The statistic has one degree of freedom, since the general nonparametric model has dimension 2 and the null hypothesis has dimension 1. Hence the null hypothesis is rejected, as was suggested by the confidence interval we derived earlier. The test used in this example can also be thought of as a test for stationary distribution (½,½), since this happens if and only if the transition matrix is doubly stochastic (Exercise **6**). □

Application (Stock market pricing) Much effort in finance theory has gone into studying the predictability of the stock market. The **efficient market hypothesis** (see Fama, 1970) implies that the deviations of the overall stock market (or, more precisely, of a portfolio containing all the stocks of a given exchange) from the mean should be independent random variables. This in turn implies that knowledge of the previous behavior of the market does little to help predict future behavior. The stock prices are said to follow a **random walk.** Empirical studies have cast some doubt over this hypothesis. There is some evidence that large deviations from the mean (in essence, highly overpriced or underpriced stocks) tend to be reverting to the mean, leading to negative correlations over long periods of time, and violating the independence assumption.

Several explanations have been proposed to this market behavior. One, proposed by Blanchard and Watson (1982), is called the **rational speculative bubbles** model. In this model, investors realize that prices exceed fundamental values, but they believe that there is a high potential for the bubble to continue to expand and yield a high return. This high return compensates precisely for the risk of a crash, showing the rationality of staying in the marked despite the overvaluation.

McQueen and Thorley (1991) applied a Markov chain model to annual stock returns. This chain took into account the behavior of the market for two successive years, each classified as above or below mean (the mean value used was a running 20-year mean—the results are not sensitive to the choice of mean value). The state space, using 0 to indicate below average prices, is $S=\{(0,0),(0,1),(1,0),(1,1)\}$, where (0,0) denotes two successive below average years. In fact, while it is a regular Markov chain on this state space, it can also be considered a **second order** Markov chain, on the state space (0,1). The term second order indicates that the dependence goes back two steps. We shall study higher order chains in more detail in the next section. The transition matrix for the chain is

$$\mathbb{P} = \begin{bmatrix} 1-p_{001} & p_{001} & 0 & 0 \\ 0 & 0 & 1-p_{011} & p_{011} \\ 1-p_{101} & p_{101} & 0 & 0 \\ 0 & 0 & 1-p_{111} & p_{111} \end{bmatrix} \tag{2.216}$$

where p_{001} is the transition probability from (0,0) to (0,1). Since the second element of the previous state must be the same as the first element of the current state we do not need four binary digits in the subscript for p. In terms of this model, the random walk hypothesis is

$$H_0\colon p_{001} = p_{011} = p_{101} = p_{111} \tag{2.217}$$

while the rational speculative bubbles hypothesis can be written

$$H_1\colon p_{001} > p_{111} \tag{2.218}$$

since the probability of state 0 should be larger following (1,1) than following (0,0). Thus, rejection of the hypothesis

$$H'_0\colon p_{001} = p_{111} \tag{2.219}$$

in the right direction can be taken as evidence in favor of the rational speculative bubbles hypothesis.

The data used by McQueen and Thorley consist of continually compounded returns for a portfolio of all New York Stock Exchange stocks for the calendar years 1947 to 1987. We consider only an equally-weighted portfolio where continually compounded inflation has been subtracted from the nominal rates. The data are given in Table 2.5. In this case we know the initial state: it is (0,0). The mle's are

$$\hat{p}_{001} = 0.750;\ \hat{p}_{011} = 0.818;\ \hat{p}_{101} = 0.5;\ \hat{p}_{111} = 0.1. \tag{2.220}$$

The likelihood ratio test of the random walk hypothesis H_0 yields a test statistic value of 14.0 on 3 degrees of freedom. Nominally, this corresponds to a P-value of 0.003. Since the numbers involved are rather small, McQueen and Thorley performed a simulation study yielding an actual P-value of 0.01. It is quite clear that the random walk hypothesis is untenable. The test of H'_0 is 8.6 on 1 degree

Table 2.5 Transition counts for NYSE portfolio 1947–1987

	0	1
(0,0)	2	6
(0,1)	2	9
(1,0)	5	5
(1,1)	9	1

of freedom, rejected at all levels using the chi-squared distribution, and receiving a P-value of 0.01 in the simulation study. Since $\hat{p}_{111} < \hat{p}_{001}$ we may want to interpret this as evidence in favor of the rational speculative bubbles hypothesis. □

2.8. Higher order chains

In the stock market example at the end of the previous section we saw how the dependence on the past can reach farther than just to the previous time. We define an rth order Markov chain (X_k) on a state space S with d elements by

$$\begin{aligned} &\mathbf{P}(X_{n+1}=x_{n+1} \mid X_n=x_n, X_{n-1}=x_{n-1}, \dots, X_0=x_0) \\ &\quad = \mathbf{P}(X_{n+1}=x_{n+1} \mid X_n=x_n,, \dots, X_{n-r+1}=x_{n-r+1}) \qquad (2.221) \\ &\quad = p(x_{n-r+1}, \dots, x_n; x_{n+1}), \end{aligned}$$

replacing subscripts by function notation for readability. There is no real novelty in an rth order chain. Let namely (Y_k) be a process with state space S^r defined by $Y_k=(X_{k-r+1}, \dots, X_k)$. Then

$$\mathbf{P}(Y_n=(a_1, \dots, a_r) \mid Y_{n-1}=(b_1, \dots, b_r)) = \begin{cases} p(b_1, \dots, b_r; a_r) & \text{if } a_i=b_{i+1},\ i=1, \dots, r-1 \\ 0 & \text{otherwise.} \end{cases} \qquad (2.222)$$

Some reflection shows that (Y_k) is a first-order Markov chain with $(d-1)d^r$ states. Hence we can use first-order chain statistical theory to test hypotheses such as the chain being lth order, where $l<r$. We can formulate the hypothesis as

$$p(a_1, \dots, a_r; a_{r+1}) = p(a_{r-l+1}, \dots, a_r; a_{r+1}). \qquad (2.223)$$

The mle under this hypothesis is

$$\hat{p}(a_1, \dots, a_r; a_{r+1}) = \frac{n(a_{r-l+1}, \dots, a_r, a_{r+1})}{n(a_{r-l+1}, \dots, a_r)} \qquad (2.224)$$

where $n(a_{r-l+1}, \dots, a_r) = \sum_l n(_{r-l+1}, \dots, a_r l)$. The χ^2 statistic is portfolio 1947–1987

ared distribution, and receiv-
e $\hat{p}_{111} < \hat{p}_{001}$ we may want to
ational speculative bubbles □

vious section we saw how the
ist to the previous time. We
ace S with d elements by

$,X_0=x_0)$

$X_{n-r+1}=x_{n-r+1})$ (2.221)

eadability. There is no real
a process with state space S^r

$))$

$, i=1,\ldots,r-1$ (2.222)

Markov chain with $(d-1)d^r$
ical theory to test hypotheses
can formulate the hypothesis

$;a_{r+1})$. (2.223)

$\frac{{}_r,a_{r+1})}{\ldots,a_r)}$ (2.224)

e χ^2 statistic is

$$\sum_{a_1,\ldots,a_{r+1}} \frac{(n(a_1,\ldots,a_{r+1})-n(a_1,\ldots,a_r)\hat{p}(a_1,\ldots,a_r;a_{r+1}))^2}{n(a_1,\ldots,a_r)\hat{p}(a_1,\ldots,a_r;a_{r+1})}, \tag{2.225}$$

which asymptotically has a χ^2 distribution with

$$(d-1)d^r-(d-1)d^l=(d-1)d^l(d^{r-l}-1) \tag{2.226}$$

degrees of freedom.

Application (Snoqualmie Falls precipitation, continued) Looking at the Snoqualmie Falls data in more detail, we obtain Table 2.6.

Table 2.6 Data for second-order model

Previous days		Current day		Proportion
Second	First	Dry	Wet	wet
Wet	Wet	100	527	0.841
Dry	Wet	25	94	0.790
Wet	Dry	70	52	0.426
Dry	Dry	109	67	0.381

The 95% asymptotic joint confidence set for (p_{11},p_{01}) from the first-order model was (0.775, 0.893)×(0.272, 0.524). All the observed proportions fall inside this set, indicating that the first-order model is adequate. Note however, that if the previous two days were (dry,wet), the observed proportion 0.790 is outside the individual 95% confidence interval (0.808, 0.860) for p_{11}, showing the importance of using simultaneous rather than individual confidence bands. The χ^2 statistic for testing second order vs. first order is 2.4. Here $r=d=2$, $l=1$ so the statistic has $1\times2^1\times(2^{(2-1)}-1)=2$ degrees of freedom. There are four parameters in the second-order model, and two in the first-order model. The P-value is 0.29, and we see no reason to reject the first-order model. □

In order to test for order of a Markov chain we may use the fact that the test statistics in successively nested hypotheses are asymptotically independent. This creates a multiple decision problem, which is complicated to analyze. Suppose that a chain is really order 1. Then the probability of falsely rejecting the true order is the probability of falsely accepting order 0, and of falsely rejecting order 1 in favor of order 2. These two events are asymptotically independent. The second has asymptotic probability α, but the probability of the first depends on the sample size and on how far the p_{ij} are from the independent case (i.e., on the noncentrality parameter of the χ^2 statistic).

A different possibility is to consider the order as a parameter, and estimate it using maximum likelihood. This does not work, because we can make the likelihood arbitrarily large by having one parameter for each observation. However, if we penalize the likelihood for the number k of independent parameters, by maximizing $2\,l(\hat{\boldsymbol{\theta}})-f\,(k,n)$ for some suitable choice of f, we may be able to offset the increase in the likelihood that is due to an increase in the number of parameters, rather than to an improved fit. Many different choices of f have been suggested. We will use the **Bayes Information Criterion** (BIC), for which $f\,(k,n)=k\log n$. In other words, we look for the model that maximizes

$$\mathrm{BIC}(k) = 2\max_{\Theta_k} l(\boldsymbol{\theta}) - k\log n \tag{2.227}$$

where Θ_k is the parameter space corresponding to k parameters (not all values of k may be possible). An important point is that the sample size used must be consistent with the largest model considered. For example, if we are considering a third-order model, and we have n observations of the chain, we can only use the last $n-3$ observations for the zero order model, the last $n-2$ for the first order, etc. This is because the estimates of the third-order model do not start until the fourth observation, the first three being needed to see what state the chain is leaving. It turns out that for finite state Markov chains, BIC is a consistent estimate of the order of the chain (Katz, 1981). The rules of thumb of Jeffreys (1961, Appendix B) suggest that a difference in BIC of at least 2 log 100 = 9.2 is needed to deem the model with the smaller BIC substantially better.

Application (Snoqualmie Falls precipitation, continued) In order to apply the BIC to the Snoqualmie Falls data we need the maximum likelihood for the chains of order 0, 1, and 2. These are given in Table 2.7, with the value for the model chosen by each criterion shown in boldface.

Table 2.7 BIC for Snoqualmie Falls precipitation

Order	k	$2l(\hat{\theta}_k)$	BIC
0	1	−1259.5	−1266.5
1	2	**−1075.3**	**−1089.2**
2	4	−1073.0	−1100.8

Both the likelihood ratio test and the BIC favor the first-order model. It is usually the case that BIC only has one maximum as a function of k. Generally BIC tends to choose smaller models than the likelihood ratio test. □

High-order Markov chains have rather a lot of parameters: $(d-1)d^l$ for the full lth order chain. This makes the model unsuitable even for relatively small d, as shown in Table 2.8. One would sometimes like a model that allows for high-order dependence, although not using as many parameters as the full lth order

nsider the order as a parameter, and esti-
This does not work, because we can make
ving one parameter for each observation.
d for the number k of independent param-
for some suitable choice of f, we may be
elihood that is due to an increase in the
an improved fit. Many different choices of
the **Bayes Information Criterion** (BIC),
rds, we look for the model that maximizes

og n (2.227)

esponding to k parameters (not all values
point is that the sample size used must be
sidered. For example, if we are consider-
n observations of the chain, we can only
zero order model, the last $n-2$ for the first
ates of the third-order model do not start
three being needed to see what state the
finite state Markov chains, BIC is a con-
hain (Katz, 1981). The rules of thumb of
that a difference in BIC of at least 2 log
l with the smaller BIC substantially better.

precipitation, continued) In order
alls data we need the maximum likelihood
ese are given in Table 2.7, with the value
shown in boldface.

oqualmie Falls precipitation

$2l(\hat{\theta}_k)$	BIC
-1259.5	−1266.5
-1075.3	**−1089.2**
-1073.0	−1100.8

BIC favor the first-order model. It is usu-
aximum as a function of k. Generally BIC
e likelihood ratio test. □

r a lot of parameters: $(d-1)d^l$ for the full
l unsuitable even for relatively small d, as
etimes like a model that allows for high-
as many parameters as the full lth order

Table 2.8 Number of parameters for different order chains

	Order			
d	1	2	3	4
2	2	4	8	16
3	6	18	54	162
4	12	48	192	768
5	20	100	500	2500

model. Raftery (1985a) proposes a linear model for Markov chain transition probabilities. Let $Q=(q_{ij})$ be a transition matrix, $\boldsymbol{\lambda}$ an l-vector of parameters summing to 1. Assume that

$$\mathbf{P}(X_k=j \mid X_{k-1}=j_1,\ldots,X_{k-l}=j_l) = \sum_{i=1}^{l}\lambda_i q_{j_i j}. \tag{2.228}$$

It is clear that this defines a lth order Markov chain. If Q is ergodic then X is ergodic and has an equilibrium distribution $\boldsymbol{\pi}$, which also is the stationary probability vector for Q. This model is called the **mixture transition distribution** (MTD) model.

It is straightforward to write down the likelihood for an MTD model.

$$L(\boldsymbol{\lambda},Q) = \sum_{i,i_1,\ldots,i_l=0}^{K} n(i_1,\ldots,i_l,i)\log(\sum_{j=1}^{l}\lambda_i q_{i_j i}). \tag{2.229}$$

There is no simple closed form for the maximum likelihood estimators. The likelihood must be optimized numerically (Schimert, 1992).

Application (Wind power in Ireland) In order to design turbines for wind power generation in Ireland, data were collected on hourly wind speeds at Belmullet for the first four weeks of July, 1962. The 672 wind speed measurements were grouped into four states:

State 0:	No power produced	0–8 knots
State 1:	Less than full potential	8–16 knots
State 2:	Full capacity	16–25 knots
State 3:	Closed down due to high winds	> 25 knots

Since no transitions to other than neighboring classes were seen, transitions farther away were assumed to have probability zero in order to cut down on the number of parameters in the general Markov chain model (see Table 2.9). The likelihood ratio statistic for testing order 2 vs. 1 is 32.3 on 10 degrees of freedom (P = 0.0003), and that for testing order 3 vs. order 2 is 18.9 on 26 degrees of freedom (P = 0.84). The likelihood ratio test therefore chooses order 2. The

Table 2.9 Order selection for Irish wind power models

Order	Markov chain			MTD chain		
	k	L	BIC	k	L	BIC
0	3	−869.5	−1758.6			
1	6	−417.5	**−874.1**			
2	16	−385.2	−874.5	7	−395.6	−836.7
3	42	−366.3	−1006.1	8	−388.0	**−828.1**
4	110	−347.2	−1410.5	9	−388.0	−834.6

penalty for additional parameters in going to order 3 is simply too large. Note that the maximum likelihood for the MTD chain of order 3 is comparable to that of the general chain of order 2, so not much is lost in reducing from 16 to 8 parameters. The estimated parameter values are

$$\hat{\boldsymbol{\lambda}} = (0.629, 0.206, 0.165) \tag{2.230}$$

and

$$\hat{Q} = \begin{bmatrix} 0.837 & 0.163 & 0 & 0 \\ 0.058 & 0.854 & 0.088 & 0 \\ 0 & 0.133 & 0.847 & 0.040 \\ 0 & 0 & 0.116 & 0.884 \end{bmatrix}. \tag{2.231}$$

The estimated stationary probabilities are

$$\hat{\boldsymbol{\pi}} = (0.148, 0.416, 0.324, 0.111). \tag{2.232}$$

We see that the turbines are expected to be producing power about 3/4 of the time in July, although optimal production only occurs about 1/3 of the time. We interpret $\hat{\boldsymbol{\lambda}}$ as the relative strength of influence of past values. □

2.9. Chain-dependent models

In order to test the goodness of fit of our model of Snoqualmie Falls precipitation, we may, for example, derive the theoretical distribution of some functional of the data, and compare it (using our estimated parameters) to the empirical distribution from our data. Some examples of interesting such functionals are the number of rainy days in a week, the total amount of rainfall in a month, and the maximum rainfall in a month (a number of particular hydrologic importance).

Let $Z_1, \ldots, Z_n$ be a sequence of random variables such that the distribution of Z_k, given that $X_{k-1}=j$ and $X_k=l$, has distribution function F_{jl}. We can think of Z_{k+1} as a score associated with the kth transition. Note that this generalizes the ergodic theory in section 2.5, since we are introducing external randomness (not just looking at a non-random function of the current state).

on for Irish wind power models

n		MTD chain	
BIC	k	L	BIC
-1758.6			
–874.1			
–874.5	7	–395.6	–836.7
-1006.1	8	–388.0	**–828.1**
-1410.5	9	–388.0	–834.6

going to order 3 is simply too large. Note MTD chain of order 3 is comparable to that not much is lost in reducing from 16 to 8 values are

$$) \tag{2.230}$$

$$\begin{matrix} & 0 \\ 8 & 0 \\ 7 & 0.040 \\ 6 & 0.884 \end{matrix}\Bigg]. \tag{2.231}$$

s are

$$, 0.111). \tag{2.232}$$

d to be producing power about 3/4 of the ction only occurs about 1/3 of the time. We influence of past values. □

f our model of Snoqualmie Falls precipita- theoretical distribution of some functional our estimated parameters) to the empirical xamples of interesting such functionals are the total amount of rainfall in a month, and (a number of particular hydrologic impor-

of random variables such that the distribu- $_k=l$, has distribution function F_{jl}. We can with the kth transition. Note that this gen- 2.5, since we are introducing external ran- on-random function of the current state).

Conditional on a given sequence of transitions we assume that the Z_i are independent. Let $S_n = \sum_1^n Z_k$. S_n is called an **additive functional** of the Markov chain. Our goal is to compute the distribution of S_n, or equivalently its Laplace transform

$$\phi_n(t) = \mathbf{E}^{\pi}(\exp(-tS_n)), \tag{2.233}$$

assuming that the Markov chain is in equilibrium. Write $\phi_{jk}^{(r)}(t) = \mathbf{E}(\exp(-tS_r) \mid X_0 = j, X_r = k)$. Given that $X_0 = j$, $X_{n-1} = l$, and $X_n = k$, the random variables S_{n-1} and Z_n are conditionally independent, with Laplace transforms $\phi_{jl}^{(n-1)}(t)$ and $\phi_{lk}^{(1)}(t)$, respectively. Thus

$$\mathbf{E}(\exp(-tS_n) \mid X_1 = j, X_{n-1} = l, X_n = k) = \phi_{jl}^{(n-2)}(t)\phi_{lk}^{(1)}(t). \tag{2.234}$$

Averaging over the value of X_{n-1} we get

$$\begin{aligned} \phi_{jk}^{(n-1)}(t) &= \mathbf{E}(\exp(-tS_n) \mid X_1 = j, X_n = k) \\ &= \sum_l \mathbf{P}(X_{n-1} = l \mid X_n = k, X_0 = j) \\ &\quad \times \mathbf{E}(\exp(-tS_n) \mid X_1 = j, X_{n-1} = l, X_n = k) \end{aligned} \tag{2.235}$$

or, since $\mathbf{P}(X_{n-1} = l \mid X_n = k, X_0 = j) = p_{jl}^{(n-1)} p_{lk} / p_{jk}^{(n)}$,

$$p_{jk}^{(n)} \phi_{jk}^{(n)}(t) = \sum_l p_{jl}^{(n-1)} \phi_{jl}^{(n-1)}(t) p_{lk} \phi_{lk}^{(1)}(t). \tag{2.236}$$

Writing $\mathbf{Q}^{(n)}(t) = (p_{jk}^{(n)} \phi_{jk}^{(n)}(t))$, we see that

$$\mathbf{Q}^{(n)}(t) = \mathbf{Q}^{(n-1)}(t)\mathbf{Q}^{(1)}(t) \tag{2.237}$$

so, with $\mathbf{Q}^{(1)} \equiv \mathbf{Q}$, we must have $\mathbf{Q}^{(n)}(t) = (\mathbf{Q}(t))^n$. It now remains to average over X_0 and X_n:

$$\phi_n(t) = \mathbf{E}^{\pi}(\exp(-tS_n)) = \sum_{j,k} \pi(j) p_{jk}^{(n)} \phi_{jk}^{(n)}(t) = \boldsymbol{\pi} Q(t)^n \mathbf{1}^T. \tag{2.238}$$

Example (Semi-Markov chains) One drawback with Markov chains is the inflexibility in describing the time spent in a given state. We have seen (Exercise **2**) that this time must be geometrically distributed with parameter p_{jj}. A generalization is the semi-Markov chain, in which the process moves out of a given state according to a Markov chain with transition matrix $\mathbb{P}$, having $p_{jj} = 0$ for all j, while the time spent in state j is a random variable with distribution function $F_j(l)$, or, more generally, $F_{ij}(l)$ where i is the state the process came from. □

Application (Snoqualmie Falls precipitation, continued) Suppose that we are interested in studying the distribution of the number of days in a week with measurable precipitation. Then $Z_k=1(X_k=1)\equiv X_k$. Thus $\phi_{00}=\phi_{10}=1$ and $\phi_{01}=\phi_{11}=e^{-t}$. In Exercise **11** we show how to compute Q^n using diagonalization. However, for small values of n there is a simple recursive formula for computing the exact distribution of S_n. Let $u_k^{(n)}=\mathbf{P}^0(S_n=k)$ and $v_k^{(n)}=\mathbf{P}^1(S_n=k)$. Then

$$\begin{aligned} u_k^{(n)} &= \mathbf{P}(S_n=k,X_1=0 \mid X_0=0) + \mathbf{P}(S_n=k,X_1=1 \mid X_0=0) \\ &= \mathbf{P}(S_{n-1}=k \mid X_0=0)p_{00} + \mathbf{P}(S_{n-1}=k-1 \mid X_0=1)p_{01}. \end{aligned} \tag{2.239}$$

Performing a similar computation for $v_k^{(n)}$, we have

$$\begin{aligned} u_k^{(n)} &= u_k^{(n-1)}p_{00} + v_{k-1}^{(n-1)}p_{01} \\ v_k^{(n)} &= u_k^{(n-1)}p_{10} + v_{k-1}^{(n-1)}p_{11} \end{aligned} . \tag{2.240}$$

Finally, for a two-state chain in equilibrium, we have

$$\mathbf{P}^{\pi}(S_n=k) = \pi_0 u_k^{(n)} + \pi_1 v_k^{(n)}. \tag{2.241}$$

For the Snoqualmie Falls data the exact distribution was computed for $n=7$, using the estimated transition matrix

$$\hat{\mathbb{P}} = \begin{bmatrix} 0.602 & 0.398 \\ 0.166 & 0.834 \end{bmatrix} \tag{2.242}$$

and compared to data from 144 midwinter weeks (Table 2.10).

Table 2.10 Snoqualmie Falls weekly rainfall

# wet days	Observed frequency	Expected frequency
0	3	1.9
1	5	4.5
2	9	9.0
3	10	15.0
4	21	21.9
5	25	27.6
6	30	30.0
7	41	34.2

The χ^2 statistic for goodness of fit is 4.03 with 4 degrees of freedom, for a P-value of 0.40. Strictly speaking the χ^2 distribution is not applicable, since the observations are from successive quadruples of weeks. However, as we have

recipitation, continued) Suppose distribution of the number of days in a Then $Z_k=1(X_k=1)\equiv X_k$. Thus $\phi_{00}=\phi_{10}=1$ ow how to compute Q^n using diagonali- there is a simple recursive formula for Let $u_k^{(n)}=\mathbf{P}^0(S_n=k)$ and $v_k^{(n)}=\mathbf{P}^1(S_n=k)$.

$$) + \mathbf{P}(S_n=k,X_1=1 \mid X_0=0)$$

$$, + \mathbf{P}(S_{n-1}=k-1 \mid X_0=1)p_{01}. \qquad (2.239)$$

$^{)}$, we have

$$. \qquad (2.240)$$

ium, we have

$$. \qquad (2.241)$$

ct distribution was computed for $n=7$,

$$(2.242)$$

er weeks (Table 2.10).

nie Falls weekly rainfall

ved ncy	Expected frequency
	1.9
	4.5
	9.0
	15.0
	21.9
	27.6
	30.0
	34.2

.03 with 4 degrees of freedom, for a P- distribution is not applicable, since the ruples of weeks. However, as we have

seen, the dependence between events seven days or more apart is fairly small. More precisely, the correlation between $\sum_1^7 Z_i$ and $\sum_8^{14} Z_i$ is estimated to be 0.074 (Exercise **12**). The adequacy of the χ^2-distribution is the subject of Exercise **C5**. □

In modeling the Snoqualmie Falls precipitation we have so far concentrated on looking at the presence or absence of rainfall. By augmenting the parameter space, as in the Irish wind example in the previous section, we can take into account the amount of precipitation, at least in a rough way. From a forecasting point of view it is important to do better than that. To that end, we will follow Katz (1977) and look at a bivariate process (X_n,Z_n), where $X_n=1$(precipitation on day n). As before, we assume that X_n is a first-order Markov chain with stationary transition probabilities p_{ij} and stationary distribution $\boldsymbol{\pi}$. The amount of precipitation on the nth day is Z_n, positive precisely when $X_n=1$. We make the following assumptions:

(i) The distribution of Z_n depends on (X_{n-1},X_n).

(ii) The Z_i are conditionally independent, given the process (X_n).

It follows from these assumptions that

$$\begin{aligned}\mathbf{P}(Z_n\le x \mid X_0,Z_1,X_1,Z_2,X_2,\ldots,Z_{n-1},X_{n-1}=i)\\ = \mathbf{P}(Z_n\le x \mid X_{n-1}=i).\end{aligned} \qquad (2.243)$$

Let $F_i(x)=\mathbf{P}(Z_n\le x \mid X_{n-1}=i,X_n=1)$. Suppose that the chain is in equilibrium. Write

$$\mu_i = \mathbf{E}(Z_n \mid X_{n-1}=i,X_n=1) \qquad (2.244)$$

and

$$\sigma_i^2 = \mathbf{Var}(Z_n \mid X_{n-1}=i,X_n=1). \qquad (2.245)$$

By unconditioning

$$\mu = \mathbf{E}Z_n = \sum \mu_i \pi_i p_{i1} \qquad (2.246)$$

and

$$\begin{aligned}\rho_0 &= \mathbf{Var}Z_n = \mathbf{E}\,\mathbf{Var}(Z_n \mid X_{n-1},X_n) + \mathbf{Var}\,\mathbf{E}(Z_n \mid X_{n-1},X_n)\\ &= \mathbf{E}\sigma^2_{X_{n-1}} + \mathbf{Var}\mu_{X_{n-1}} = \sum \pi_i \sigma_i^2 p_{i1} + \sum \pi_i \mu_i^2 p_{i1} - \mu^2 \qquad (2.247)\\ &= \pi_0 p_{01}(\sigma_0^2+\mu_0^2) + \pi_1 p_{11}(\sigma_1^2+\mu_1^2) - \mu^2.\end{aligned}$$

We will now look at the total amount of rainfall in n days. Let $S_n = \sum_1^n Z_i$. Then (O'Brien 1974)

$$\frac{S_n - n\mu}{\sigma n^{1/2}} \xrightarrow{d} \mathrm{N}(0,1) \qquad (2.248)$$

where $\sigma^2=\rho_0+2\sum_1^\infty\rho_j$, and $\rho_j=\mathbf{Cov}(Z_n,Z_{n+j})$, provided that $0<\sigma^2<\infty$. To compute σ^2 we do another conditional computation:

$$\mathbf{E}Z_nZ_{n+j} = \mathbf{E}\,\mathbf{E}(Z_nZ_{n+j}\,|\,X_{n-1},X_n,X_{n+j-1},X_{n+j}) \tag{2.249}$$
$$= \mathbf{E}\mu_{X_{n-1}}\mu_{X_{n+j-1}} = \sum_{i,k}\pi_i\mu_i\mu_k p_{ik}^{(j)}.$$

This is again an additive functional. Adapting equation (2.238), let

$$Q = \begin{bmatrix} 0 & p_{01}\mu_0 \\ 0 & p_{11}\mu_1 \end{bmatrix} \tag{2.250}$$

and $\mathbf{1}=(1, 1)$. Then we can write

$$\mathbf{E}Z_nZ_{n+j} = \boldsymbol{\pi}Q\mathbb{P}^jQ\mathbf{1}^T. \tag{2.251}$$

Note that $\mu=\boldsymbol{\pi}Q\mathbf{1}^T$. Thus

$$\rho_j=\boldsymbol{\pi}Q(\mathbb{P}^j-\mathbf{1}^T\boldsymbol{\pi})Q\mathbf{1}^T \tag{2.252}$$

and

$$\sigma^2 = \rho_0 + 2\sum_{j=1}^{\infty}\boldsymbol{\pi}Q(\mathbb{P}^j-\mathbf{1}^T\boldsymbol{\pi})Q\mathbf{1}^T. \tag{2.253}$$

The following fact helps in the computation:

Lemma 2.7 $\mathbb{P}^j-\mathbf{1}^T\boldsymbol{\pi} = (\mathbb{P}-\mathbf{1}^T\boldsymbol{\pi})^j$.

Proof We prove this by induction. The case $j=1$ is trivial. Assume that the statement is true for k. Then, since $\boldsymbol{\pi}\mathbb{P}=\boldsymbol{\pi}$

$$\mathbb{P}^{k+1}-\mathbf{1}^T\boldsymbol{\pi} = (\mathbb{P}^k-\mathbf{1}^T\boldsymbol{\pi})\mathbb{P} = (\mathbb{P}-\mathbf{1}^T\boldsymbol{\pi})^k\mathbb{P}$$
$$= (\mathbb{P}-\mathbf{1}^T\boldsymbol{\pi})^{k+1} + (\mathbb{P}-\mathbf{1}^T\boldsymbol{\pi})^k\mathbf{1}^T\boldsymbol{\pi} \tag{2.254}$$
$$= (\mathbb{P}-\mathbf{1}^T\boldsymbol{\pi})^{k+1} + (\mathbb{P}^k-\mathbf{1}^T\boldsymbol{\pi})\mathbf{1}^T\boldsymbol{\pi}$$

using the induction hypothesis twice. Now, since $\mathbf{1}^T$ is a right eigenvector for $\mathbb{P}$ we have $\mathbb{P}\mathbf{1}^T\boldsymbol{\pi} = \mathbf{1}^T\boldsymbol{\pi}$ and by iterating $\mathbb{P}^k\mathbf{1}^T\boldsymbol{\pi}=\mathbf{1}^T\boldsymbol{\pi}$, so $(\mathbb{P}^k-\mathbf{1}^T\boldsymbol{\pi})\mathbf{1}^T\boldsymbol{\pi}=\mathbf{0}$, completing the induction. □

It follows that the infinite sum is $2\boldsymbol{\pi}Q(\mathbb{P}-\mathbf{1}^T\boldsymbol{\pi})^{-1}Q\mathbf{1}^T$, and some algebra shows that

$$\sigma^2 = \rho_0 + \frac{2}{1-(p_{11}-p_{10})}\mu\pi_0(p_{11}\mu_1-p_{01}\mu_0). \tag{2.255}$$

$_{n+j})$, provided that $0<\sigma^2<\infty$. To com- tation:

$$,X_n,X_{n+j-1},X_{n+j}) \qquad (2.249)$$

).

oting equation (2.238), let

$$(2.250)$$

$$(2.251)$$

$$(2.252)$$

$$Q1^T. \qquad (2.253)$$

ion:

The case $j=1$ is trivial. Assume that the π

$$(\mathbb{P}-1^T\pi)^k\mathbb{P}$$
$$(\mathbb{P}-1^T\pi)^k 1^T\pi \qquad (2.254)$$
$$(\mathbb{P}^k-1^T\pi)1^T\pi$$

ow, since 1^T is a right eigenvector for $\mathbb{P}$ $\mathbb{P}^k 1^T\pi=1^T\pi$, so $(\mathbb{P}^k-1^T\pi)1^T\pi=0$, com- □

$\mathbb{P}-1^T\pi)^{-1}Q1^T$, and some algebra shows

$$_0(p_{11}\mu_1-p_{01}\mu_0). \qquad (2.255)$$

Application (Snoqualmie Falls precipitation, continued) In order to illustrate the theory developed above we will apply it to the Snoqualmie Falls data. In order to fit the precipitation amounts, a gamma distribution was assumed. Figure 2.5 shows the observed and estimated distributions.

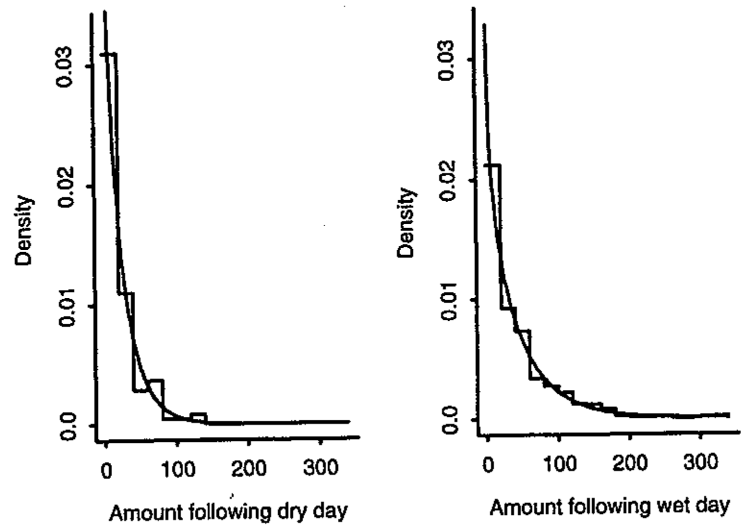


Figure 2.5. Observed and fitted densities of precipitation amounts following wet days (a) and dry days (b).

The parameters were estimated by maximum likelihood. The estimated scale parameters are different for wet days following wet days than for wet days following dry days, while the shape parameters are quite similar. Thus there is a tendency for lower amounts following a dry day. The means are $\mu_0=22.9$ and $\mu_1=43.5$, so $\mu=28.3$ (recall that μ is the overall mean, taking into account the zero rainfall on dry days). The variances are $\sigma_0^2=691.7$ and $\sigma_1^2=2351.7$, whence $\sigma^2=2339.0$. In order to check assumption (ii), the amounts of precipitation on consecutive wet days were plotted on a log–log scale (Figure 2.6). There is no evidence of dependence. In order to study the distribution of the total amounts, the exact distribution, using the fitted gamma distribution, was compared to the limiting normal distribution derived above. Figure 2.7 shows the result for $n=20$, together with a histogram of observed sums for January 6–25. The fit of the exact distribution to the observed sums is rather bad, the latter showing evidence of bimodality, perhaps corresponding to dry and wet years. In addition, the normal approximation is bad, which is not too surprising since we are only summing up 20 random variables, many of which are zero. □

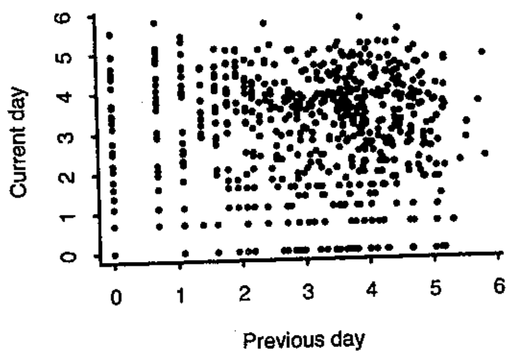


Figure 2.6. Amounts of precipitation on consecutive wet days.

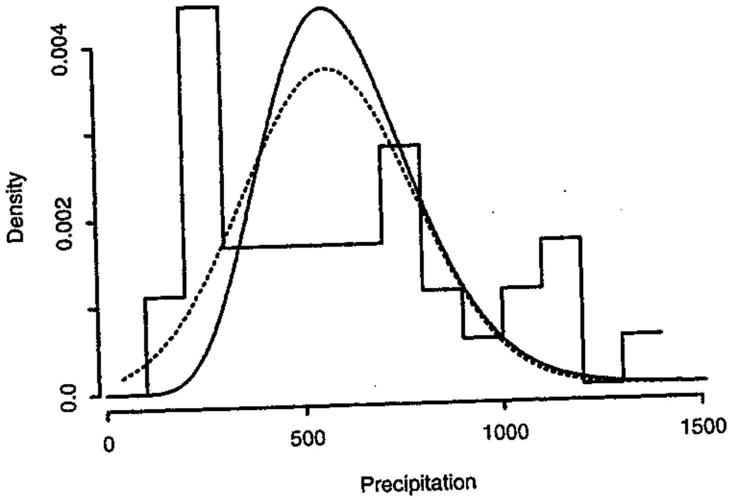
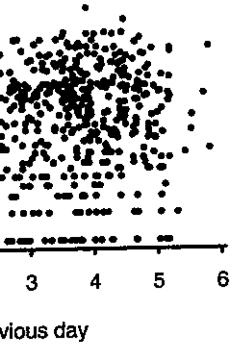
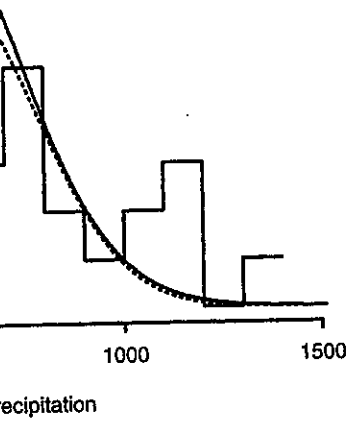


Figure 2.7. Exact theoretical (solid) and asymptotic (dotted) densities of precipitation January 6–20. The solid step function is a histogram of the 36 observed years.

One can easily derive another interesting parameter of the precipitation process, namely the distribution of maximum rainfall. Let $M_n=\max_{i\leq n}Z_i$, and $G_n(x;i)=\mathbf{P}^i(M_n\leq x)$. Also define $G_n(x)=\pi_0 G_n(x;0)+\pi_1 G_n(x;1)$. Splitting the set



n consecutive wet days.



) and asymptotic (dotted) densities of
step function is a histogram of the 36

eresting parameter of the precipitation
ximum rainfall. Let $M_n=\max_{i\le n}Z_i$, and
$\pi_0 G_n(x;0)+\pi_1 G_n(x;1)$. Splitting the set

$\{M_n\le x, X_0=0\}$ over the possible values of X_1 we get that

$$\begin{aligned} G_n(x;0) &= \frac{\mathbf{P}(M_n\le x, X_0=0)}{\mathbf{P}(X_0=0)} \\ &= \frac{\mathbf{P}(M_n\le x, X_0=0, X_1=0)}{\mathbf{P}(X_0=0)} + \frac{\mathbf{P}(M_n\le x, X_0=0, X_1=1)}{\mathbf{P}(X_0=0)} \\ &= \mathbf{P}(M_{2,n}\le x \mid X_1=0)\mathbf{P}(X_1=0 \mid X_0=0) \qquad (2.256) \\ &\quad + \mathbf{P}(M_{2,n}\le x, Z_1\le x \mid X_0=0, X_1=1)\mathbf{P}(X_1=1 \mid X_0=0) \\ &= p_{00}G_{n-1}(x;0)+p_{01}F_0(x)G_{n-1}(x;1) \end{aligned}$$

where $M_{2,n}=\max_{2\le k\le n}Z_k$. Similarly we obtain

$$G_n(x;1) = p_{10}G_{n-1}(x;0)+p_{11}F_1(x)G_{n-1}(x;1). \qquad (2.257)$$

Using the initial conditions $G_0(x;0)=G_0(x;1)=1$ these equations can be solved recursively. Assuming that the F_i are cdf's of a gamma distribution, and writing

$$F(x) = \pi_0 + \pi_0 p_{01}F_0(x) + \pi_1 p_{11}F_1(x) \qquad (2.258)$$

the following extreme value result is valid (Denzel and O'Brien 1975):

$$\lim_{n\to\infty} G_n\left[u_n + \frac{x}{nF'(u_n)}\right] = \exp(-\exp(-x)), \qquad (2.259)$$

where $1-F(u_n)=1/n$. Rewriting (2.259) we see that

$$G_n(y)\approx\exp(-\exp(-nF'(u_n)(y-u_n)). \qquad (2.260)$$

The assumption of gamma precipitation distribution is not crucial: Denzel and O'Brien (ibid.) show that the limiting behavior of M_n is the same as for iid observations from F (see Resnick, 1987, for details). Also, the assumption of starting in the stationary distribution is unnecessary: the limiting behavior is the same for all initial distributions.

Application (Snoqualmie Falls precipitation, continued) The fit of the limiting extreme value distribution is substantially better than the fit of the limiting total amounts distribution. Using the mle's determined earlier, and noting that $u_n=114$ for these values, we compute the exact and asymptotic distributions. Figure 2.8 shows them for $n=20$. The asymptotic approximation is excellent. In addition, the observed maximal precipitation (or, more precisely, the empirical distribution function of the 36 years of maxima) for January 6–25 is shown. There is perhaps a slight skewness, with too many small and too few large maximal precipitation values, but the sample size is only 36. □

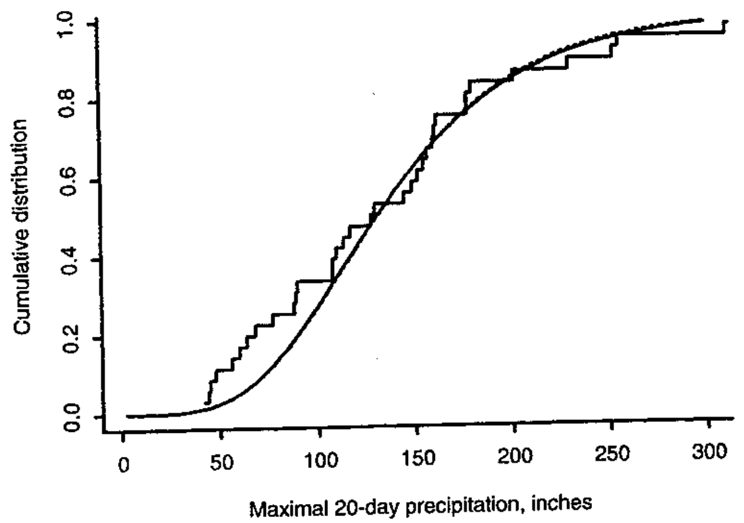


Figure 2.8. Observed (step function), exact (smooth line) and asymptotic (dotted line) distribution functions of maximum precipitation January 6–25.

2.10. Random walks and harmonic analysis

We have encountered the random walk earlier in different contexts. In this section we look at what amounts to an application to mathematics. The harmonic analysis that we shall encounter is very elementary, although it will allow us to deal with first passage problems for finite state space Markov chains in some generality. We begin, however, in the fair coin tossing setup.

Consider a fair, simple random walk on $\{0,\ldots,K\}$, so that $p_{i,i+1}=p_{i+1,i}=\frac{1}{2}$, $i=0,1,\ldots,K-1$, and all other transition probabilities are 0. An interesting problem is to find

$$f(x) = \mathbf{P}^x(\text{reach } K \text{ before } 0). \tag{2.261}$$

By conditioning on the first step we see that

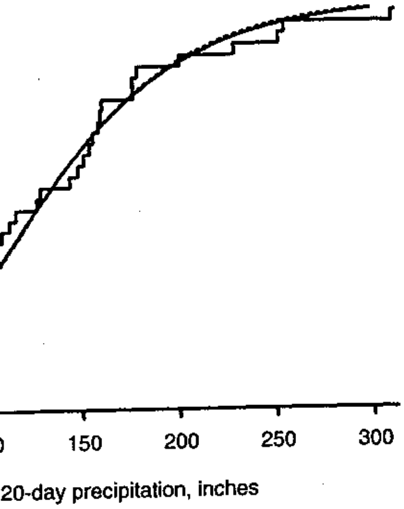
$$f(x) = \tfrac{1}{2}f(x-1) + \tfrac{1}{2}f(x+1) \quad x=1,2,\ldots,K-1. \tag{2.262}$$

The initial conditions are $f(0)=0$ and $f(K)=1$.

We call $D=\{1,\ldots,K-1\}$ the **interior** points and $B=\{0,K\}$ the **boundary** points. A function $f(x)$ on $S=D+B$ is **harmonic** if for all points in D it satisfies the averaging property

$$f(x) = \tfrac{1}{2}(f(x-1)+f(x+1)). \tag{2.263}$$

The problem of finding a harmonic function given its boundary values is called



ction), exact (smooth line) and asymptotic maximum precipitation January 6–25.

onic analysis

valk earlier in different contexts. In this sec-n application to mathematics. The harmonic very elementary, although it will allow us to or finite state space Markov chains in some ne fair coin tossing setup.

random walk on $\{0,\ldots,K\}$, so that nd all other transition probabilities are 0. An

$$\text{ore } 0). \qquad (2.261)$$

e see that

$$x+1) \quad x=1,2,\ldots,K-1. \qquad (2.262)$$

nd $f(K)=1$.

e **interior** points and $B=\{0,K\}$ the **boundary** B is **harmonic** if for all points in D it satisfies

$$+1)). \qquad (2.263)$$

c function given its boundary values is called

the **Dirichlet problem**, and the uniqueness principle for Dirichlet problems asserts that there can be no two harmonic functions having the same boundary values. To prove the uniqueness principle, we must first prove the maximum principle:

Theorem 2.18 **(Maximum principle)** A harmonic function $f(x)$ defined on S takes on its maximum M and its minimum m on the boundary.

Proof Let $M=\max_S f(x)$. If $f(x)=M$ for some $x\in D$, then $f(x-1)=f(x+1)=M$. Continuing left, eventually $f(0)=M$. The same argument works for the minimum. □

Theorem 2.19 **(Uniqueness principle)** If f and g are harmonic functions on S with $f(x)=g(x)$ for $x\in B$, then $f(x)=g(x)$ for all x.

Proof Let $h(x)=f(x)-g(x)$. Then for $x\in D$

$$\frac{h(x-1)+h(x+1)}{2}=\frac{f(x-1)+f(x+1)}{2}-\frac{g(x-1)+g(x+1)}{2}$$

$$=f(x)-g(x)=h(x), \qquad (2.264)$$

so h is harmonic. But h is 0 on B, so by Theorem 2.18 it is 0 everywhere. □

We have now reduced the problem to finding a harmonic function with the given initial conditions. Using the theory in Appendix B we see that the solution is $f(x)=x/K$. Harmonic function theory also gives simple answers to other questions about our random walk. For example, are we certain to reach the boundary eventually? Let

$$h(x)=\mathbf{P}^x(\text{never reach } B). \qquad (2.265)$$

Then $h(x)=½h(x-1)+½h(x+1)$, so h is harmonic, with boundary values $h(0)=h(K)=0$. So $h(x)$ must be identically zero.

How long does it take to hit the boundary? The answer to this question is not a harmonic function, but conditioning on the first step yields for an interior x

$$e(x)=\mathbf{E}^xT_B=1+½\mathbf{E}^{x-1}T_B+½\mathbf{E}^{x+1}T_B. \qquad (2.266)$$

The corresponding difference equation then is

$$(E-1)^2e(x)=-2\times1^x \qquad (2.267)$$

with initial conditions $e(0)=e(K)=0$. Using Appendix B again, the solution is $e(x)=(K-x)x$.

Let us now try a more complicated two-dimensional array.

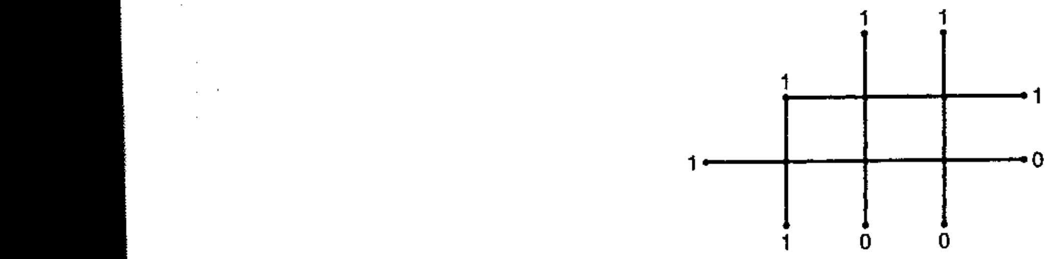


Figure 2.9. A subset of the two-dimensional lattice. Adapted from *Random Walks and Electric Networks* by P. G. Doyle and J. L. Snell, published by the American Mathematical Society.

A **lattice point** is a point with integer coordinates. In Figure 2.9 we see a subset of the two-dimensional lattice, with boundary points divided into two types, marked 0 or 1. Consider a process performing a random walk on the lattice subset. Starting in an interior point, the process moves to each neighboring point with probability ¼. We have two classes of boundary states, B_0 and B_1, and we are interested in the probability of hitting B_1 before B_0.

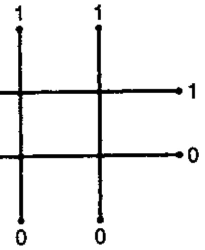
We describe the general situation as follows. Let $S=D+B$ be a finite set of lattice points, such that each point in D has four neighbors in S, and each point in B has at least one neighbor in D. D consists of the interior points, and B is the set of boundary points. Also assume that S is **connected**, i.e., that there is a route from every point to every other point. Call a function f on S harmonic if

$$f(a,b) = \tfrac{1}{4}(f(a+1,b)+f(a-1,b)+f(a,b+1)+f(a,b-1)). \qquad (2.268)$$

As before, our functions of interest, namely the hitting probability $f(x)$ is a harmonic function with boundary values: $f(x)=1(x\in B_i)$, $i=$ 0 or 1. where $B=B_0+B_1$. For a finite lattice the maximum principle and the uniqueness principle are proved just as before. The problem, then, is reduced to solving the difference equation (2.268). Using the ergodic theorem we may run many particles in a random walk on the lattice and note what proportion end up at what boundary point. This must be repeated for each initial interior point. While this approach yields an answer, it is slow and imprecise. The theory of partial difference equations is not very well developed. Let ∇^2 be the **symmetric second difference operator**, so that $\nabla^2 g(x)=g(x+1)-2g(x)+g(x-1)$. Then (2.268) can be written (using a subscript to denote which argument the operator is applied to)

$$\nabla_1^2 f(x,y) + \nabla_2^2 f(x,y) = 0 \qquad (2.269)$$

with boundary conditions $f(\mathbf{b})=1(\mathbf{b}\in B_1)$. It is natural to look for guidance to ted two-dimensional array.



imensional lattice. Adapted from *Random*
G. Doyle and J. L. Snell, published by the

coordinates. In Figure 2.9 we see a subset
boundary points divided into two types,
rforming a random walk on the lattice sub-
process moves to each neighboring point
ses of boundary states, B_0 and B_1, and we
ting B_1 before B_0.

ion as follows. Let $S=D+B$ be a finite set
nt in D has four neighbors in S, and each
n D. D consists of the interior points, and B
ssume that S is **connected**, i.e., that there is
er point. Call a function f on S harmonic if

$(a-1,b)+f(a,b+1)+f(a,b-1)).$ (2.268)

amely the hitting probability $f(x)$ is a har-
lues: $f(x)=1(x\in B_i)$, $i=0$ or 1. where
aximum principle and the uniqueness prin-
e problem, then, is reduced to solving the
ne ergodic theorem we may run many parti-
e and note what proportion end up at what
ed for each initial interior point. While this
ow and imprecise. The theory of partial
ell developed. Let ∇^2 be the **symmetric**
nat $\nabla^2 g(x)=g(x+1)-2g(x)+g(x-1)$. Then
cript to denote which argument the operator

0 (2.269)

$\in B_1$). It is natural to look for guidance to

the continuous case. The corresponding partial differential equation is **Laplace's equation**

$$\frac{\partial^2 f(x,y)}{\partial x^2}+\frac{\partial^2 f(x,y)}{\partial y^2}=0 \tag{2.270}$$

with boundary condition

$$\lim_{(x,y)\to t} f(x,y)=\phi(t),\ t\in\partial D \tag{2.271}$$

where ∂D is the boundary of D. In the continuous case any function satisfying Laplace's equation is called harmonic. The **method of relaxation** was developed to solve the continuous Dirichlet problem. It is based on a result which says that a function is harmonic if its value at (x,y) is equal to its average over any circle inside D centered on (x,y). This suggests a method of successive averaging of function values. Translated to the lattice case, the method of relaxation works as follows. Start with an arbitrary function with the right boundary values, such as

$$\begin{array}{ccccc} & & 1 & 1 & \\ & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & \end{array}$$

Pick an interior point, such as (2,2) (with (0,0) being at the lower left-hand corner of the minimum rectangle containing all the states). If the function is not equal to its values over the neighbors, adjust it. Run through all the interior points in some order. The new function becomes

$$\begin{array}{ccccc} & & 1 & 1 & \\ & 1 & 0.5 & 0.625 & 1 \\ 1 & 0.832 & 0.328 & 0.156 & 0 \\ & 1 & 0 & 0 & \end{array}$$

The new function is still not harmonic, in general, but we can repeat the procedure until it converges. After nine iterations we have

$$\begin{array}{ccccc} & & 1 & 1 & \\ & 1 & 0.823 & 0.787 & 1 \\ 1 & 0.876 & 0.506 & 0.323 & 0 \\ & 1 & 0 & 0 & \end{array}$$

Note that this can be done automatically in a spread sheet program.

There is a third way of solving Dirichlet problems. This method involves the theory of Markov chains directly. Suppose that the chain has state space S, with the boundary states B absorbing, and transition matrix $\mathbb{P}$. Call a function f **harmonic for** $\mathbb{P}$ **if**

$$f(i)=\sum_j P_{ij}f(j) \quad \text{for all } i\in D. \tag{2.272}$$

Writing f as a vector, with the absorbing states first, we have that $f^T=\mathbb{P}f^T$, so $\mathbb{P}^n f^T=f^T$ for all n. Note that

$$\mathbb{P}^n = \begin{bmatrix}\mathbb{I} & 0\\ R & Q\end{bmatrix}^n = \begin{bmatrix}\mathbb{I} & 0\\ C_n & Q^n\end{bmatrix}. \tag{2.273}$$

Here $C_{n+1}=R\mathbb{I}+QC_n$. Letting $n\to\infty$ we see that $C=R+QC$. Hence

$$\mathbb{P}^n \to \begin{bmatrix}\mathbb{I} & 0\\ C & 0\end{bmatrix}, \tag{2.274}$$

where $C=(\mathbb{I}-Q)^{-1}R$. Then

$$f^T = \begin{bmatrix}f_B^T\\ f_D^T\end{bmatrix} = \begin{bmatrix}\mathbb{I} & 0\\ C & 0\end{bmatrix}\begin{bmatrix}f_B^T\\ f_D^T\end{bmatrix}. \tag{2.275}$$

We see that

$$f_D^T=Cf_B^T=(\mathbb{I}-Q)^{-1}Rf_B^T \tag{2.276}$$

so that f_D is determined by the values of f at the boundary. Numbering the states in Figure 2.9

$$\begin{array}{ccccc}
 & & 1 & 2 & \\
 & 9 & 10 & 11 & 3\\
8 & 12 & 13 & 14 & 4\\
 & 7 & 6 & 5 &
\end{array}$$

we have

$$R = \begin{bmatrix}
¼ & 0 & 0 & 0 & 0 & 0 & 0 & 0 & ¼\\
0 & ¼ & ¼ & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & ¼ & ¼ & ¼\\
0 & 0 & 0 & 0 & 0 & ¼ & 0 & 0 & 0\\
0 & 0 & 0 & ¼ & ¼ & 0 & 0 & 0 & 0
\end{bmatrix} \tag{2.277}$$

$$Q = \begin{bmatrix}
0 & ¼ & 0 & ¼ & 0\\
¼ & 0 & 0 & 0 & ¼\\
0 & 0 & 0 & ¼ & 0\\
¼ & 0 & ¼ & 0 & ¼\\
0 & ¼ & 0 & ¼ & 0
\end{bmatrix}. \tag{2.278}$$

Since $f_B=(1,1,1,0,0,0,1,1,1)$ we get that

$$(\mathbb{I}-Q)^{-1}Rf_B^T = \begin{bmatrix}0.823\\ 0.787\\ 0.876\\ 0.506\\ 0.323\end{bmatrix}; \tag{2.279}$$

the same result as that obtained by the method of relaxation, without the need for iterations, but requiring a matrix inversion which may be difficult if the state space is large.

first, we have that $f^T=\mathbb{P}f^T$, so

(2.273)

$C=R+QC$. Hence

(2.274)

(2.275)

(2.276)

oundary. Numbering the states

3
4

(2.277)

(2.278)

(2.279)

of relaxation, without the need
hich may be difficult if the state

Application (Airplane fire escape probabilities) To assure profitability, airlines must configure their airplanes to carry the maximum number of seats. However, the maximization is constrained by passenger safety and comfort, as well as by the maximum useful load carrying capacity of the airplane. The Federal Aviation Authority (FAA) of the US requires that all commercial airplanes have a maximum evacuation time of 90 seconds for all seating configurations.

One serious situation requiring evacuation is a fire in an engine with smoke obscuring the exit pathways. For example, in 1985 a Boeing 737 taking off from Manchester developed a fire in the left engine. The pilot first thought that the problem was a blown landing gear, thereby delaying the eventual evacuation of the aircraft. After about a dozen passengers had left the airplane, the interior was filled by thick black smoke. The cabin attendants were unable to see from the left to the right forward door. The 131 passengers were tourists, with many relatively inexperienced flyers. Many crawled over seats looking for exits. 76 passengers survived the fire.

Figure 2.10 shows the seating configuration.

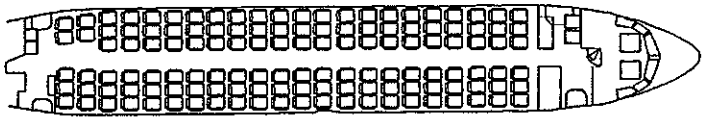


Figure 2.10. Seating arrangement for a Boeing 737.

There are 20 rows of six seats, three on either side of the aisle, and two more rows with three seats to the right and only two on the left. Two seats were unoccupied, but including infants there were 131 passengers. Only two forward exits and the right overwing escape hatch were usable. Except for the front row, escape routes were not generally to the nearest exit, indicating that the behavior was somewhat random.

Assuming the fair random walk model for passengers (this includes the possibility of trying to walk through a wall, corresponding to a zero escape probability, and assumes that individuals reaching an aisle can find an exit) we solved the corresponding harmonic function problem. Escape probabilities for an individual in a window seat on the left varied from 0.101 in row 1 to 0.25 in rows 10–14, while a right window seat occupant had probabilities ranging from 0.101 (row 1) to 0.596 (in the overwing exit row 10). Aisle seats generally had higher escape probabilities, varying from 0.467 (row 1 on left) to 0.808 (row 10 on right). The total expected number escaping the fire from this extremely simple model is 55, corresponding well to the 60–65 passengers who experienced the smoky environment and escaped. We can compute estimates of the change in escape probability if we add exits. Adding an exit amounts to increasing the number of preferred exit states. The same type of computation indicates that each additional exit (up to four additional exits) increases the expected number of survivors by about one. □

Note that a function f which is harmonic for $\mathbb{P}$ satisfies $\mathbb{P}\mathbf{f}^T=\mathbf{f}^T$, so it is, in a sense, a dual concept to the stationary distribution. More precisely, where the stationary distribution is a left eigenvector of $\mathbb{P}$, any harmonic function is a right eigenvector, both corresponding to the eigenvector 1. Recall that the function $f(x)=1$ is always a right eigenvector, and, consequently, so is any constant function. It turns out that this is the unique right eigenvector whenever $\mathbb{P}$ corresponds to an irreducible persistent chain, provided that f is either nonnegative or bounded (see Asmussen, 1987, section I.5 for some discussion on how this is useful in the classification of chains). Our application of the concept of harmonic functions to hitting probabilities deals with transient chains, where the behavior is more interesting.

Example (Electric network) Consider a network with five resistors and a unit voltage applied across it (Figure 2.11).

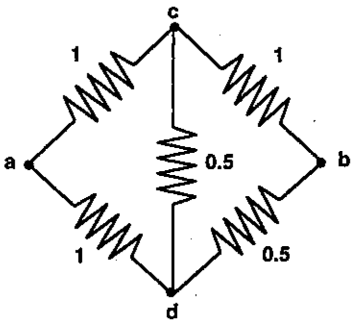


Figure 2.11. A simple Wheatstone bridge with unit voltage to be applied between a and b.

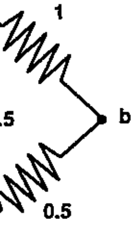
The conductance between two points x and y is the inverse of the resistance between the points. In this case $C_{ac}=C_{ad}=C_{bd}=C_{bc}=1$ while $C_{ab}=C_{cd}=2$. We are interested in determining voltages $v(x)$ at various points in this network. Two laws describe the behavior of the system:

Kirchhoff's law: The current flowing into x is the same as the current flowing out from x.

ing an exit amounts to increasing the
e type of computation indicates that
exits) increases the expected number
□

for $\mathbb{P}$ satisfies $\mathbb{P}\mathbf{f}^T=\mathbf{f}^T$, so it is, in a
tribution. More precisely, where the
r of $\mathbb{P}$, any harmonic function is a
e eigenvector 1. Recall that the func-
and, consequently, so is any constant
ique right eigenvector whenever $\mathbb{P}$
hain, provided that f is either non-
', section I.5 for some discussion on
ains). Our application of the concept
es deals with transient chains, where

der a network with five resistors and
).



idge with unit voltage to be applied

nd y is the inverse of the resistance
$=C_{bd}=C_{bc}=1$ while $C_{ab}=C_{cd}=2$. We
c) at various points in this network.
em:

x is the same as the current flowing

Ohm's law: If x and y are connected by a resistance R_{xy}, then the current flowing from x to y is

$$i_{xy}=\frac{v(x)-v(y)}{R_{xy}}. \tag{2.280}$$

The voltage at a is 1 and that at b is 0, so if $x\neq a,b$ $\sum_y i_{xy}=0$, and since $i_{xy}=-i_{yx}$ we have

$$v(x)\sum_y C_{xy}=\sum_y C_{xy}v(y) \tag{2.281}$$

or, writing $C_x=\sum_y C_{xy}$,

$$v(x)=\sum_y \frac{C_{xy}}{C_x}v(y). \tag{2.282}$$

Now notice that $\mathbb{P}=(C_{yx}/C_x)$ is a transition matrix. Making a and b absorbing states (i.e., changing the corresponding rows to have 1 on the diagonal and 0 elsewhere) yields a modified transition matrix $\mathbb{P}^*$, and v is harmonic for $\mathbb{P}^*$. We need the modification since (2.282) only holds for $x\neq a$ or b. In this case,

$$\mathbb{P}^*=\begin{bmatrix}1&0&0&0\\0&1&0&0\\1/4&1/4&0&1/2\\1/5&2/5&2/5&0\end{bmatrix} \tag{2.283}$$

so

$$(\mathbb{I}-Q)^{-1}R\mathbf{v}_B=\begin{bmatrix}5/4&5/8\\1/2&5/4\end{bmatrix}\begin{bmatrix}1/4&1/4\\1/5&2/5\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix}=\begin{bmatrix}7/16\\3/8\end{bmatrix}, \tag{2.284}$$

i.e., $v(c)=0.4375$ and $v(d)=0.375$.

Note that $\mathbb{P}$ is reversible, since $C_{yx}=C_{xy}$, so

$$p_{xy}=\frac{C_{xy}}{C_x}=\frac{C_{yx}}{C_y}\times\frac{C_y}{C_x}=p_{yx}\frac{C_y}{C_x}. \tag{2.285}$$

This reversibility is required by physical theory. In particular, therefore, we must have a stationary distribution π satisfying $\pi_y/\pi_x=C_y/C_x$, whence $\pi_y=C_y/C$ where $C=\sum_y C_y$. A lot more material on the relation between random walks and electric networks can be found in Doyle and Snell (1984). □

Application (Airplane fire escape probabilities, continued)

Another Boeing 737 fire with substantial cabin smoke occurred in 1984 in Calgary (this and the Manchester fire are the only instances of this type for the Boeing 737). Again, an engine fire was first interpreted by the pilot as a tire failure on the landing gear. In this case the 114 passengers were frequent flyers, and had all flown with this type of aircraft before. All on board the airplane survived. Here the simple random walk predicts 59 survivors, but the passengers did not act randomly. For example, the passengers in rows 1–7 all exited through the forward doors, those in rows 8–16 through the right overwing exit, and those in rows 17–22 through the right rear exit.

A weighted approach assumes that a passenger in an exit row moves to the exit with probability 5/8, and to the other three adjacent locations with probability 1/8. Passengers in a row adjacent to an exit row have equal probabilities of moving toward the aisle or climbing over seats. All other passengers have probability 5/8 of moving toward the aisle, and 1/8 of moving in any other direction. This simple model of informed behavior yields 101 expected survivors. □

2.11. Bienaymé–Galton–Watson branching processes

The simple **branching process** was first studied by Bienaymé[1] in 1845 in order to find a mathematical (rather than social or genetic) explanation for the fact that a large proportion of the family names, both among nobility and bourgeoisie, seemed to be dying out when viewed over a long period of time. It has been applied to problems of genetics, epidemiology, nuclear fission, queueing theory, and demography, among other areas, and is one of the simplest nonergodic stochastic models. It is commonly called the **Bienaymé–Galton–Watson** (BGW) process.

Let

$$Z_k = \sum_{i=1}^{Z_{k-1}} X_{k,i} \tag{2.286}$$

where for each k the $X_{k,i}$ are iid random variables with the same distribution $p_l=\mathbf{P}(X_{i,j}=l)$ called the **offspring distribution**. We interpret $\sum_1^0$ as 0. Suppose that the offspring distribution has pgf $P(s)$, with mean m and variance σ^2. Assuming that $Z_0=1$, we see that

$$P_k(s) = \mathbf{E}s^{Z_k} = P_{k-1}(P(s)). \tag{2.287}$$

By induction we see that

$$\mathbf{E}Z_k = \mathbf{E}\,\mathbf{E}(Z_k \mid Z_{k-1}) = m\mathbf{E}Z_{k-1} = \cdots = m^k. \tag{2.288}$$

Notice that $\mathbf{E}Z_k \to \infty$ where $m>1$, while it stays constant when $m=1$ and goes to zero when $m<1$. For this reason processes with $m>1$ are called **supercritical**, those with $m=1$ are **critical**, and those with $m<1$ are **subcritical**. Furthermore

[1] Bienaymé, Irénée-Jules (1796–1878). Inspector General of the Administration of Finances of France 1834–1848, subsequently independent mathematician. Much of his work had long been overlooked; see Heyde and Seneta (1977).

dicts 59 survivors, but the
he passengers in rows 1–7 all
3–16 through the right overw-
t rear exit.

nger in an exit row moves to
e adjacent locations with pro-
xit row have equal probabili-
er seats. All other passengers
nd 1/8 of moving in any other
vior yields 101 expected sur-
□

processes

y Bienaymé[1] in 1845 in order
netic) explanation for the fact
th among nobility and bour-
r a long period of time. It has
gy, nuclear fission, queueing
d is one of the simplest non-
commonly called the

(2.286)

es with the same distribution
Ve interpret $\sum_1^0$ as 0. Suppose
th mean m and variance σ^2.

(2.287)

$\cdot = m^k.$ (2.288)

nstant when $m=1$ and goes to
$n>1$ are called **supercritical**,
are **subcritical**. Furthermore

f the Administration of Finances of
. Much of his work had long been

$$\begin{aligned}\mathbf{Var}Z_k &= \mathbf{E}\,\mathbf{Var}(Z_k \mid Z_{k-1}) + \mathbf{Var}\,E(Z_k \mid Z_{k-1}) \\ &= \sigma^2\mathbf{E}Z_{k-1} + m^2\mathbf{Var}Z_{k-1} = \sigma^2 m^{k-1} + m^2\mathbf{Var}Z_{k-1} \qquad (2.289)\end{aligned}$$

so that, using the methods in Appendix B,

$$\mathbf{Var}Z_k = \sigma^2 m^{k-1}\sum_0^{k-1} m^i = \sigma^2 m^{k-1}\frac{m^k-1}{m-1}. \qquad (2.290)$$

From (2.286) it is easy to see that the process (Z_k) is a Markov chain, with transition probabilities $p_{ij} = p_j^{*i}$, where p^{*i} is the i-fold convolution of (p_j). Another way of writing the process is

$$Z_n = mZ_{n-1} + e_n, \qquad (2.291)$$

where $e_n = Z_n - mZ_{n-1}$ can be thought of as a prediction error, having conditional mean 0, given the past, and conditional variance $\sigma^2 Z_{n-1}$, given the past.

From the construction in (2.286) it is also clear that once a generation is empty, all following generations will be empty as well. Therefore we will compute the probability of eventual extinction. Since this Markov chain has infinite state space, the method used in the previous section to compute hitting probabilities does not apply.

Theorem 2.20 Suppose that $p_0>0$, $p_0+p_1<1$, and let $q=\mathbf{P}(Z_k\to 0)$. Then q is the smallest nonnegative root of the equation $P(s)=s$. Furthermore, $q=1$ iff $m\le 1$.

Proof Extinction will occur in or before the kth generation in one of the following ways: the ancestor has

0 children
1 child whose family becomes extinct in or before the $(k-1)$th generation
2 children, both of whose families become extinct in or before the $(k-1)$th generation, etc.

Let $P_k(s) = \mathbf{E}s^{Z_k}$. Since the probability of the jth of these cases is p_j, the probability $q_k = P_k(0)$ of extinction after k generations is, by conditioning on the first family size,

$$\mathbf{P}(Z_k=0) = P_k(0) = \sum_{j=0}^{\infty} p_j P_{k-1}(0)^j = P(P_{k-1}(0)). \qquad (2.292)$$

Since $p_1 \neq 1$, we note that $P(s)$ is a strictly increasing function. Thus the numbers $q_k = P(q_{k-1})$ form a strictly increasing sequence of positive numbers, all bounded by 1. This sequence therefore has a limit which we denote q, with $p_0 \le q \le 1$. This is the probability of ultimate extinction. By going to the limit in (2.292), we see that $q=P(q)$, whence

$$\frac{P(q)-P(q_k)}{q-q_k} = \frac{q-q_{k+1}}{q-q_k} < 1. \qquad (2.293)$$

Letting $k \to \infty$ we see (Figure 2.12) that $P'(q) \le 1$.

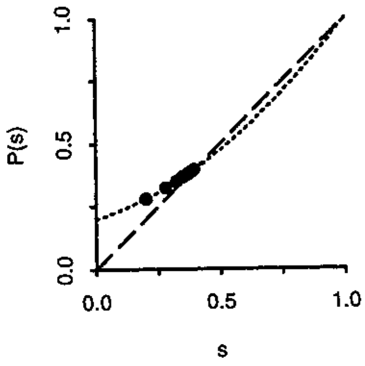


Figure 2.12. The sequence q_k (large dots) for a supercritical branching process. The dotted line is the pgf $P(s)$, and the dashed line is the line $y = s$.

Since $P'(s)$ is a power series with positive coefficients it is an increasing function, i.e., $P(s)$ is convex. If $P'(1) > 1$ we must have $q < 1$. In this case q and 1 are the only positive roots to the equation $P(s) = s$. On the other hand, if $P'(1) \le 1$, we have for $0 \le s < 1$ that $P'(s) - 1 < 0$, so that the smallest zero of the function $P(s) - s$ must be 1, whence $q = 1$.

Finally, if $p_0 = 0$, $Z_k \ge Z_{k-1}$, and extinction is impossible. If $p_1 = 1$, $Z_k = Z_0$, and extinction is again impossible. Upon noting that $P'(1) = m$, the proof is complete. □

Example (Epidemics) The problem of determining the fraction of a community that must be vaccinated in order to prevent major epidemics of a communicable disease is a crucial public health problem. In order to describe an epidemic, the population is divided into three possible health states. An individual can be susceptible to infection by a given disease agent, (s)he may have been infected by the agent and is infectious (possibly after a latent period), or (s)he is removed from the epidemic by death, by isolation, or by immunity or other natural loss of infectiousness. Initially all members of the population are susceptible to infection. The epidemic starts when one or many infectious individuals enter the population and come into contact with its members. A susceptible person is infected if (s)he has adequate contact with an infectious individual. General theory of epidemic models can be found in Bailey (1975), and their statistical inference is discussed, e.g., in Becker (1976) and Rida (1991). A BGW process can be used to approximate the infectious population during the early stages of an epidemic (Becker 1977). Clearly, since the number of susceptible individuals decreases as the epidemic progresses, it is unreasonable to assume that the offspring distribution is the same from generation to generation.

However, for early stages of epidemics in large populations the assumptions underlying the BGW process are not unrealistic.

In order for an epidemic to become serious, a large buildup of cases is needed in the early stage. Consequently we can call an epidemic **major** if the offspring mean is >1, and **minor** if it is ≤ 1 (so that the extinction probability is 1). In order to prevent major epidemics it is necessary to ensure adequate vaccination in the community so as to make the offspring mean less than one. Suppose that we select a proportion θ of the population at random for vaccination. If the vaccination is effective, the offspring distribution changes to $P^*(s)=\theta+(1-\theta)P(s)$ with mean $m^*=(1-\theta)m$, so that $m^*<1$ only if $\theta>1-1/m$. For a numerical illustration, assume that the offspring distribution is Poisson. Then the probability $1-q$ of a major epidemic is a function of m.

Table 2.11 Probability of a major epidemic

m	1	1.05	1.1	1.4	1.8
$1-q$	0	.09	.18	.51	.73
$1-1/m$	0	.05	.09	.29	.44

The third line of Table 2.11 contains the proportion of vaccination needed to bring the new mean below one. If one accepts this model it becomes of considerable importance to be able to estimate m as accurately as possible. □

In order to estimate m, first suppose that we know all the X_{ij}. Then the likelihood would be $\sum N_k \log p_k$, where $N_k=\#\{X_{ij}=k\}$, and the mle of p_k would be N_k/N, where

$$N=\sum_{k=0}^{\infty} N_k=\sum_{0}^{n-1} Z_k = Y_{n-1}. \tag{2.294}$$

The mle of m is then

$$\hat{m}=\sum k\hat{p}_k=\sum\frac{kN_k}{N}=\frac{\#\text{ children}}{\#\text{ parents}}=\frac{Y_n-1}{Y_{n-1}}. \tag{2.295}$$

Notice that this does not depend on the (N_k), only on the generation sizes. This suggests that $\hat{m}$ may also be the mle based on only observing generation sizes. This happens to be the case (Keiding and Lauritzen, 1978).

It is clear that $\{0\}$ is an absorbing state, and it can be shown that all other states are transient. In fact, either the chain dies out, or it explodes (diverges to infinity). Unless $p_0=0$, extinction has positive probability, and therefore $\hat{m}_n$ has positive probability of converging to $1-1/Y_\infty$. The statistical theory developed in section 2.7 fails to apply, since the process is not ergodic. However, conditional on nonextinction we show below that $\hat{m}_n \to m$ with probability one. It is difficult to determine exact finite-sample properties of $\hat{m}_n$.

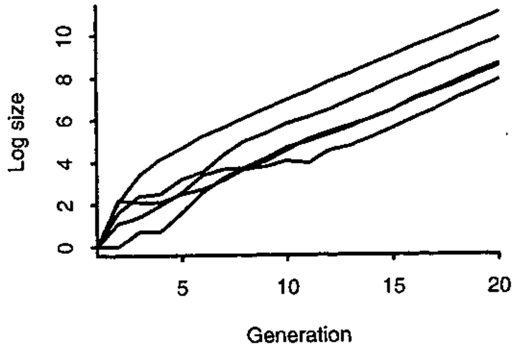


Figure 2.13. Five paths of a BGW process with Poisson offspring distribution. Adapted from P. Guttorp, *Statistical Inference for Branching Processes*, by permission of John Wiley & Sons, Inc. (1991).

The growth of a BGW process, given that it does not become extinct, is geometric. In Figure 2.13 we see the logarithms of five paths from a process with Poisson offspring distribution with $m=2$. After some initial wigglyness, they look quite linear. The figure suggests that if we rescale the generation sizes by their means, which are growing geometrically, we may end up with a limiting constant. This is not quite right: the limit turns out to be a random variable.

Theorem 2.21 Let $m>1$, and define $W_n=m^{-n}Z_n$. Then $W_n \to W$ with probability one, where $\mathbf{E}W=1$, $\mathbf{Var}W=\sigma^2/m(m-1)$, and $\mathbf{P}(W=0)=q$.

A proof of this result can be found in Guttorp (1991, Theorem 1.1). Notice that the set $\{W=0\}$ is precisely the set where the process Z_n becomes extinct.

Corollary $\hat{m}_n \to m$ on the set of nonextinction.

Proof Write

$$m^{-n}Y_n = m^{-n}\sum_{1}^{n} Z_k = \sum_{1}^{n} W_k m^{-(n-k)} \to \frac{mW}{(m-1)}. \tag{2.296}$$

Hence, whenever $W>0$ we see that

$$\hat{m}_n = (m^{-n}Y_n - m^{-n}\frac{)}{m^{-1}}(m^{-(n-1)}Y_{n-1}) \to m \tag{2.297}$$

with probability one. □

Remark Nonparametric inference for BGW processes is quite different from that for ergodic Markov chains. It can be shown (Guttorp, 1991, section 1.4) that, in essence, only the mean and the variance are consistently estimable from observing a long, non-extinct path. For example, one cannot estimate q or $P(s)$ in general. □

Application (Smallpox epidemics) We now proceed to study in more detail an example of the application of branching processes to smallpox epidemics. This disease, although now completely eradicated (only two laboratory specimens remained by 1993), had a well defined incubation time of about 12 days (rarely outside 9–15 days). Therefore the cases in the early stages of an epidemic tend to be well separated, and it is easy to determine generation sizes. We use data (Table 2.12) from Vila Guarani, a residential district in São Paulo, on *variola minor*, the less lethal form of smallpox. The epidemic was introduced by two travelers, who are not included. They were only responsible for the ancestor.

Table 2.12 Vila Guarani epidemic

Generation	0	1	2	3	4
Size	1	5	3	12	24

After the fourth generation school vacations started, and mass vaccination was introduced. Hence later data does not have similar conditions. The maximum likelihood estimate of the offspring mean is 2.10, indicating that a major epidemic was developing. One may argue, however, that although conditions were relatively constant within generations, they may not be constant between generations. In particular, the weather may be quite different almost two months from the first outbreak. The weather is important, since the disease was spread through close proximity, partly through airborne transmission, and weather affects the social behavior. To investigate this possibility, we need to introduce the idea of **random environment**. A branching process in random environment is obtained by picking an offspring distribution for each generation at random from a set of possible distributions. If there is substantial variability in the successive parent/offspring ratios $m_r=Z_r/Z_{r-1}$, compared to what would be expected from a fixed offspring distribution (constant environment) we would conclude that the environment may be random. These ratios are shown in Table 2.13. Given Z_{r-1}, the conditional variance of m_r is σ^2/Z_{r-1}. There are several different ways of estimating σ^2. We use a maximum likelihood estimator, based on computing the variance for the maximum likelihood estimator of the offspring distribution (see the Remark below). This yields $\hat{\sigma}^2=4.18$. Thus we obtain the results in Table 2.14. These ratios are consistent with the hypothesis of constant environment, in that only ratios larger than 2 in absolute value

Table 2.13 Vila Guarani estimates

r	1	2	3	4
m_r	5	0.6	4	2

Table 2.14 Vila Guarani model evaluation

r	1	2	3	4
se(m_r)	2.04	0.91	1.18	0.59
$(m_r-\hat{m})/\text{se}(m_r)$	1.42	−1.64	1.61	−0.16

would be suspicious. If we want to vaccinate enough of the population to make the epidemic minor, we need to reach $(1-1/\hat{m})$ or 52% of the population. However, since the main path of infection is through the class room, the closing of schools for vacation would already affect the offspring mean considerably. □

Remark The mle of the offspring variance (Guttorp, 1991, section 3.4) does not have a closed form, and is somewhat complicated to compute. A simpler estimate is $\tilde{\sigma}^2=(n-1)^{-1}\sum_{i=1}^n(Z_i-\hat{m}Z_{i-1})^2/Z_{i-1}$. This is generally a more variable estimate. In this case $\tilde{\sigma}^2=10.20$, considerably larger than the mle. The corresponding analysis using $\tilde{\sigma}^2$ instead of $\hat{\sigma}^2$ would show even less indication of a random environment. □

Another application of branching process theory is to a very different type of problem. Given the frequency of a particular genetic variant occurring in a population, can we figure out how long ago the mutation first arose? Here the parameter of interest is the age N of a branching process, and our observation is simply Z_N. In order to answer the question, we must specify the offspring distribution. It is not unreasonable to assume that the spread of the variant is quite similar to the population growth in general. A useful and flexible distribution, which has been found to describe human population growth adequately in different situations, is the **modified geometric** distribution:

$$p_k = bc^{k-1},\ k\geq 1 \tag{2.298}$$

$$p_0 = 1-\frac{b}{1-c}$$

with mean $m=b/(1-c)^2$ and variance $\sigma^2=b(1-b-c^2)/(1-c)^4$. We shall think of m as the mean rate of increase of the mutant type, representing a **selective advantage** if it is higher than the overall population mean, and a **disadvantage** if it is lower. It is convenient to reparametrize the distribution in terms of m and $h=(1-c)/c$, so $p_k=mh^2(1+h)^{-(k+1)}$ for $k\geq 1$, and $p_0=1-mh/(1+h)$. Then $q=1-(m-1)h$ if $m>1$. One advantage with the modified geometric distribution is that it is self-reproducing: the distribution of Z_t is of the same form, with

$$\mathbf{P}(Z_t=k) = \begin{cases} m^tM^2(1-M)^{k-1}, & k\geq 1 \\ 1-m^tM, & k=0 \end{cases} \tag{2.299}$$

where $M=(1-s_0)/(m^t-s_0)$ and $s_0=1-(m-1)h$ (Exercise **13**). Furthermore, $\mathbf{P}(Z_t=k \mid Z_t>0)=M(1-M)^{k-1}$, a regular positive geometric distribution, with mean $1/M$ and variance $(1-M)/M^2$. If we observe $Z_N=k>0$ (necessarily), the conditional log likelihood can be written

$$l(N,m,h) = \log M(N,m,h)+(k-1)\log(1-M(N,m,h)). \tag{2.300}$$

We have three parameters, but only one observation. We can reduce the parameter space by assuming that h is a known constant, corresponding to a known value of $c=p_k/p_{k-1}$, assumed to be obtained from population data. This implies that the selective advantage enters only through the parameter b. The likelihood equation becomes $M=1/k$, so the maximum likelihood estimate of N is

$$\hat{N} = \frac{\log(1+(k-1)(m-1)h)}{\log(m)} \tag{2.301}$$

provided that $m\neq 1$. If $m=1$, a similar computation yields $\hat{N}=(k-1)h$.

Application (Yanomama genetics) We apply this theory to the study of rare protein variants in South American Indian populations. A variant is considered rare if it occurs in only one tribe. Such mutations have presumably arisen after tribal differentiation, are probably descended from a single mutant, and because of the low intertribal migration have not spread to the general South American Indian population. We will look at the Yanomama tribe, living in the Andes on the border between Brazil and Columbia. The Yanomama albumin variant Yan-2 is unique to this tribe, although it is fairly frequent, and therefore must be quite old. Extensive sampling of 47 widespread Yanomama villages found 875 replicates of the variant gene in the current adult population. From demographic data for this tribe, a selectively neutral gene would have offspring well described by a modified geometric distribution with $h=1.5$. The offspring mean m is slightly above 1 (although in recent years the mean population increase has been higher, perhaps as high as 1.2). Figure 2.14 shows the likelihood (as a function of m and N) arising from this datum. The middle line is the ridge of maximum likelihood, and the outer contour lines correspond to the maximum likelihood minus 2 and 4, respectively. For $m=1.0$ the maximum likelihood estimate of the age is 1311 generations, or about 30,000 years. This is far longer than the time these tribes have been in the Americas, so it would appear unlikely that the variant would have survived only in this tribe. In fact, this value of m corresponds to a slight selective disadvantage, making the

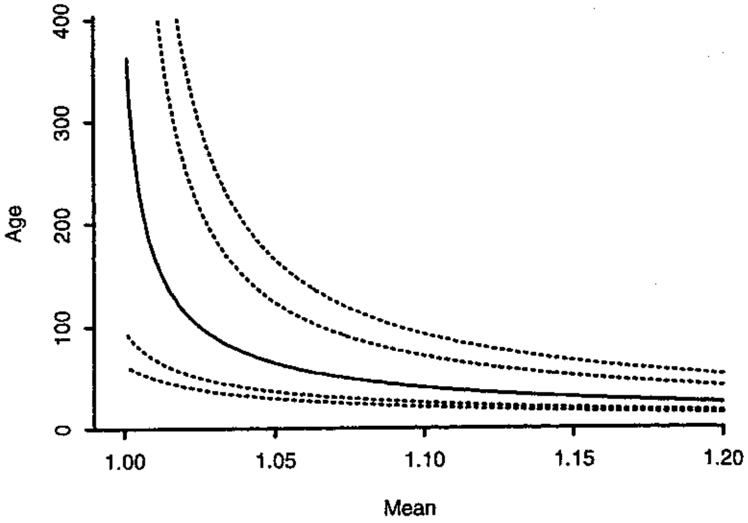


Figure 2.14. Likelihood surface for the Yanomama gene. The solid line is the ridge of maximum likelihood, while the dotted contours are 2 and 4 units of log likelihood below the maximum. Adapted from P. Guttorp, *Statistical Inference for Branching Processes*, by permission of John Wiley & Sons, Inc. (1991).

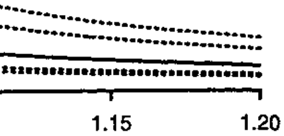
estimate even less plausible.

An increase rate of 1.02 may have been sustained by the Yanomama over long periods of time. The maximum of the likelihood then occurs at 168 generations, which is a more reasonable value, corresponding to a variant which is perhaps 3,800 years old. This supports the hypothesis that the allele is relatively old, but has arisen after tribal separation. We do not need to assume that the gene has had a selective advantage.

In order to obtain confidence intervals for these values we have to determine the asymptotic properties of the estimator. If $m>1$, given that $Z_N>0$, we have

$$\log m(\hat{N}-N) = \log(m^{-N}(1+(Z_N-1)(m-1)h)) \to \log((1-q)W \tag{2.302}$$

with probability 1 as $N \to \infty$, since $(m-1)h=1-q$. Here W, as before, is the limit of $m^{-n}Z_n$. For the positive geometric distribution one can determine the distribution of W: it is exponential with parameter $1-q$. Notice that the estimate is not consistent, even in the peculiar sense of this limiting result. We can use the result to compute confidence bands. Note first that $(1-q)W \sim \exp(1)$, and the 97.5 and 2.5 percentiles of the standard exponential distributions are log 1.0256 and log 40, respectively. Thus if $\zeta \sim \exp(1)$ we have



ama gene. The solid line is the
ntours are 2 and 4 units of log
Guttorp, *Statistical Inference for*
ey & Sons, Inc. (1991).

stained by the Yanomama over
ood then occurs at 168 genera-
ponding to a variant which is
othesis that the allele is rela-
We do not need to assume that

these values we have to deter-
. If $m>1$, given that $Z_N>0$, we

$m-1)h)) \to \log((1-q)W$(2.302)

Here W, as before, is the limit
n one can determine the distri-
$-q$. Notice that the estimate is
limiting result. We can use the
at $(1-q)W\sim\exp(1)$, and the 97.5
listributions are log 1.0256 and

$$\begin{aligned}0.95 &= \mathbf{P}(\log 1.0256 \le \zeta \le \log 40)\\ &= \mathbf{P}(\log 1.0256 \le (1-q)W \le \log 40)\\ &= \mathbf{P}(\frac{\log\log 1.0256}{\log m} \le \frac{\log((1-q)W)}{\log m} \le \frac{\log\log 40}{\log m}) \qquad (2.303)\\ &\approx \mathbf{P}(\frac{\log\log 1.0256}{\log m} \le \hat{N}-N \le \frac{\log\log 40}{\log m})\\ &= \mathbf{P}(\hat{N}-\frac{\log\log 40}{\log m} \le N \le \hat{N}-\frac{\log\log 1.0256}{\log m})\end{aligned}$$

which in this case yields the band (102, 353) at $m=1.02$. □

While the BGW branching process is a non-ergodic Markov chain, a slight modification yields an ergodic chain. Assume that at each time n there is a random number I_n of immigrants, each of which behaves like the rest of the population. This is called a **branching process with immigration**. We can write the total population size Z_n as

$$Z_n = \sum_{i=1}^{Z_{n-1}} X_{i,n} + I_n. \tag{2.304}$$

We assume that I_n is independent of the previous I_k and Z_k, and that the I_n are identically distributed with pgf $Q(s)$ and mean $\lambda<\infty$. If $m=\mathbf{E}X_{i,n}<1$ we have seen that the corresponding branching process becomes extinct with probability one; however, the immigration process ensures that 0 is no longer an absorbing state. It would therefore seem reasonable that the resulting process may be ergodic. The pgf for Z_n satisfies

$$P_n(s) = \mathbf{E}s^{Z_n} = \mathbf{E}(P(s)^{Z_{n-1}})Q(s) = P_{n-1}(P(s))Q(s). \tag{2.305}$$

If a stationary distribution exists, it must have pgf $\Pi(s)$ satisfying

$$\Pi(s) = Q(s)\Pi(P(s)). \tag{2.306}$$

Taking derivatives, we see that the stationary mean μ is

$$\begin{aligned}\mu = \Pi'(1) &= Q'(1)\Pi(P(1)) + Q(s)\Pi'(P(1))P'(1)\\ &= \lambda + \mu m\end{aligned} \tag{2.307}$$

or $\mu=\lambda/(1-m)$, provided that $m<1$.

Remark The equation (2.306) has a solution under the assumption that $\mathbf{E}\log(\max(0,I_1))<\infty$. When $m=1$ the resulting Markov chain can be either null persistent or transient. Asmussen and Hering (1984) give detailed proofs.

Application (Traffic theory) Fürth (1918) counted the number of pedestrians at five-second intervals passing a certain building. Thinking of each pedestrian as arriving as an immigrant, and having offspring 0 or 1 depending on whether or not the pedestrian leaves the observation area between two consecutive observations, we have $P(s)=1-m+ms$ and, assuming Poisson distributed input (some motivation for this will be given in the next chapter), $Q(s)=\exp(\lambda(s-1))$. The stationary distribution satisfies (2.306), i.e.,

$$\Pi(s) = \exp(\lambda(s-1))\Pi(1-m+ms). \tag{2.308}$$

It is easy to check that $\Pi(s)=\exp(\mu(s-1))$ satisfies (2.308), i.e., that the stationary distribution is Poisson with parameter μ.

In order to find the mle for the parameters m and λ, note that

$$p_{ij} = \mathbf{P}(Z_n=j \mid Z_{n-1}=i) = e^{-\lambda}\sum_{k=0}^{\min(i,j)} \frac{\lambda^{j-k}}{(j-k)!}\left[\begin{matrix} i \\ k \end{matrix}\right] m^k(1-m)^{i-k}. \tag{2.309}$$

It helps to reparametrize in terms of μ and m. Then

$$\log L(\mu,m) = \sum n_{ij}\log p_{ij} \tag{2.310}$$

where

$$p_{ij} = e^{-\mu(1-m)}\sum_{k=0}^{\min(i,j)} \frac{\mu^{j-k}}{(j-k)!}\left[\begin{matrix} i \\ k \end{matrix}\right] m^k(1-m)^{i+j-2k}. \tag{2.311}$$

To compute the mle we may of course maximize the likelihood function numerically, but it is instructive to manipulate the likelihood equations a little further. Note that

$$\frac{\partial p_{ij}}{\partial\mu} = -(1-m)p_{ij} + \frac{j}{\mu}p_{ij} - \frac{r_{ij}}{\mu}, \tag{2.312}$$

where

$$r_{ij} = ie^{-\mu(1-m)}\sum_{k=1}^{\min(i,j)} \frac{\mu^{j-k}}{(j-k)!}\left[\begin{matrix} i-1 \\ k-1 \end{matrix}\right] m^k(1-m)^{i+j-2k}. \tag{2.313}$$

Furthermore,

$$\frac{\partial p_{ij}}{\partial m} = \mu p_{ij} - \frac{i+j}{1-m}p_{ij} + \left[\frac{1}{m}+\frac{2}{1-m}\right]r_{ij}. \tag{2.314}$$

Hence the likelihood equations are

$$\frac{\partial \log L(\mu,m)}{\partial\mu} = \sum n_{ij}\{-(1-m)+\frac{i}{\mu}-\frac{r_{ij}}{\mu p_{ij}}\} = 0 \tag{2.315}$$

$$\frac{\partial \log L(\mu,m)}{\partial m} = \sum n_{ij}\{\mu-\frac{i+j}{1-m}+\left[\frac{1}{m}+\frac{2}{1-m}\right]\frac{r_{ij}}{p_{ij}}\} = 0. \tag{2.316}$$

Let $R=\sum n_{ij}r_{ij}/p_{ij}$, $N_1=\sum in_{ij}$, and $N_2=\sum jn_{ij}$. Notice that R is a function of μ and m, while N_1 and N_2 are not. Then the equations can be written

$$-(1-m)n+\frac{N_1-R}{\mu}=0$$

$$\mu n-\frac{N_1+N_2-2R}{1-m}+\frac{R}{m}=0. \qquad (2.317)$$

Solving the first equation for μ we have $\mu=(N_1-r)/n\,(1-m)$, and using this in the second we get $m=R/N_2$. Substituting this back into the expression for μ we see that

$$\hat{\mu}=\frac{(N_1-R)N_2}{n(N_2-R)}=N_2/n \qquad (2.318)$$

or the mean of the observed path, provided that we ignore edge effects, so $N_1=N_2$. The equation $m=R(\hat{\mu},m)/N_2$ must be solved numerically, and numerical optimization of $\log L(\hat{\mu},m)$ (the **profile likelihood** for m) is just as easy.

The transition counts from Fürth's data are given in Table 2.15.

Table 2.15 Pedestrian counts

	0	1	2	3	4	5	6	7
0	67	24	6	1				
1	23	83	46	11	2			
2	7	41	58	25	5			
3	1	16	18	23	7	5		
4		1	7	8	7	2		1
5			1	1	5		1	
6				1				
7						1		

For these data $N_1=N_2=804$, while $n=505$ so $\hat{\mu}=804/505=1.59$. Numerical optimization yields $\hat{m}=0.69$.

The number $P=1-m$ is called the **probability aftereffect**. It measures the probability that an individual in the system will leave it during the time between observations, but can also be thought of as a measure of dependence. If v is the average speed of a pedestrian, τ the time between observations, and b the width of the building, we see that $P=v\tau/b$. We can estimate P from the data. so the average speed of a pedestrian, using $\tau=5$ seconds and $b=20$ meters (this value is just a guess, as Fürth does not say how wide the building is), can be estimated as $\hat{v}=(1-\hat{m})b/\tau=1.25$ m/s. We can assess the variability of this estimate by producing a likelihood interval for m. Figure 2.15 depicts the likelihood surface. We see that (0.63, 0.74) is a confidence interval for m, so the

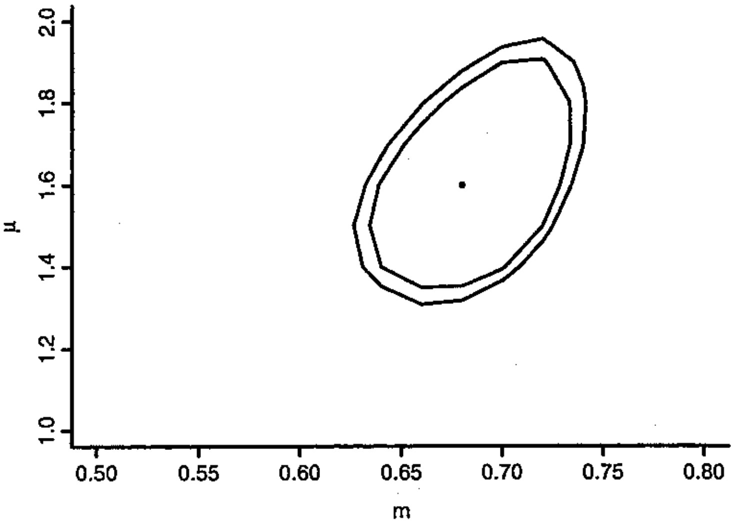


Figure 2.15. Likelihood surface for the pedestrian data. Contours are at 2.3 (approximately 90% coverage) and 3.0 (95% coverage) units of log likelihood below the maximum.

average pedestrian speed confidence interval (using the assumed value of b) is (1.0, 1.5).

In order to assess the fit of this model, Venkataraman (1982) suggests a time series approach, which has been extended and applied to these data by Mills and Seneta (1989). Since we have that

$$\mathbf{E}(Z_k \mid Z_1^{k-1}) = \lambda + mZ_{k-1} \tag{2.319}$$

we look at the fitted residuals

$$\hat{\varepsilon}_k = Z_k - \hat{\lambda} - \hat{m}Z_{k-1}. \tag{2.320}$$

Consider the jth-order autocorrelation function of the $\hat{\varepsilon}_k$, $k=1,\dots,n$, namely

$$\hat{r}_j = \frac{\sum_{k=j+1}^{n} \hat{\varepsilon}_k \hat{\varepsilon}_{k-j}}{\sum_{k=1}^{n} \hat{\varepsilon}_k^2} \equiv \hat{\psi}_j / \hat{\psi}_0. \tag{2.321}$$

Then Venkataraman shows that

$$N(\hat{r}_2 - \hat{m}\hat{r}_1)^2 \hat{\nu}^2 \xrightarrow{d} \chi_1^2 \tag{2.322}$$

where

$$\hat{\nu}^2 = \frac{(1/N)\hat{\psi}_0}{\hat{r}_2+\hat{m}^2\hat{r}_1-2\hat{m}\sum\hat{\varepsilon}_k^2\hat{\varepsilon}_{k-1}\hat{\varepsilon}_{k-2}/\hat{\psi}_0}. \tag{2.323}$$

Mills and Seneta work this out to be 18.83 for the Fürth data, and the model thus is a poor fit. □

2.12. Hidden Markov models

A criticism frequently raised regarding the Markov chain model of precipitation has been its inability to produce realistic weather data. In particular, the lengths of wet and dry periods do not correspond to observed data. We show an example in Figure 2.16.

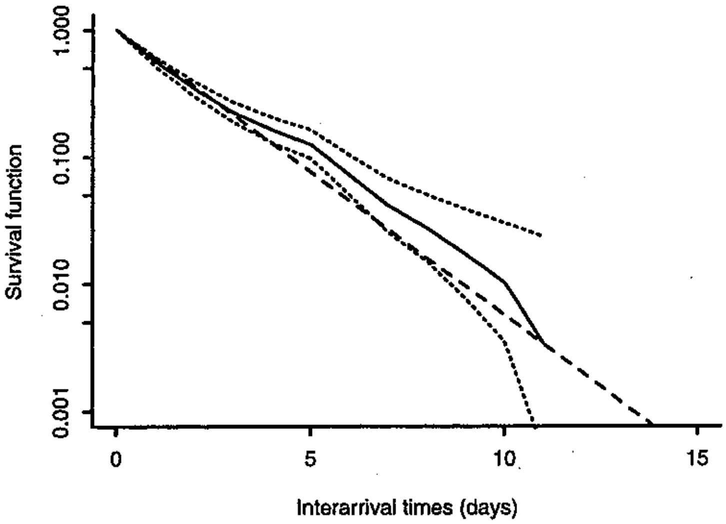


Figure 2.16. Estimated survival function (solid line) for the Snoqualmie Falls dry periods, January–March, together with the Markov chain survival function (short-dash line). The dotted line is an asymptotic 95% pointwise confidence band.

Here the estimated **survival function** (one minus the distribution function) for dry periods is plotted, together with the survival function for the estimated geometric distribution that obtains in the Markov chain situation (see Exercise 2). We see that the Markov chain survival function lies below the observed one, indicating a tendency towards shorter dry spells than what are observed. The dotted lines are asymptotic 95% pointwise confidence bands. The estimation of the survival function has to be done with care, since dry periods at the beginning and the end of the month are incompletely observed. Here we have used methods from survival analysis (Cox and Oakes 1984, ch. 4). Another possibility is to look back and forward in time to get the complete length of dry periods straddling January 1 or 31, assuming that the probability model is stationary over periods longer than exactly one month.

Another criticism of the Markov chain model is that it contains virtually no scientific knowledge (it is a statistical model, in the sense of section 1.2). The dependence observed in actual precipitation data is presumed to be a function of weather systems that pass through an area. It appears reasonable that dependence between rainfall from different weather systems is much smaller than that between rainfall on successive days of the same weather system.

A third problem is that the Markov chain approach has not been very successful when applied to more than one station in a region. While it is straightforward to develop a chain with an N-dimensional state space, representing all possible outcomes of rainfall, the large number (2^N) of parameters does not seem warranted for such a model.

In order to alleviate the problems discussed above, we shall build a model for precipitation at k stations by introducing unobserved states (thought of as somehow summarizing ''climate'') which account for different distributions of rainfall over the stations. The weather states are assumed to follow a Markov chain, while the pattern of occurrence/non-occurrence of precipitation over the network at any given time, given the weather states, is conditionally independent of the pattern at any other time. This type of model is often called a **hidden Markov model** and is a special case of the general **state space model** approach.

Let $C(t)$ denote the weather state at day t (throughout this section t will denote the discrete time variable, while n will denote the station number). We assume that $C(t)$ is a Markov chain with stationary transition probabilities

$$\gamma_{ij} = \mathbf{P}(C(t)=j \mid C(t-1)=i), \quad i,j=1,\ldots,M \tag{2.324}$$

and equilibrium probabilities

$$\delta = (\delta_1,\ldots,\delta_M) \tag{2.325}$$

so that, writing $\Gamma=(\gamma_{ij})$ for the transition matrix of the weather process we have

$$\delta\Gamma = \delta. \tag{2.326}$$

Let $X_n(t)=1$(rain at site n on day t), $n=1,\ldots,N$, $t=0,1,\ldots$ Write

$$\mathbf{X}(t) = (X_1(t),\ldots,X_N(t)) \tag{2.327}$$

and

$$Y(t) = (2^{N-1},2^{N-2},\ldots,2^0)\mathbf{X}(t)^T \tag{2.328}$$

since dry periods at the begin-
y observed. Here we have used
1984, ch. 4). Another possibil-
complete length of dry periods
probability model is stationary

odel is that it contains virtually
el, in the sense of section 1.2).
data is presumed to be a func-
area. It appears reasonable that
eather systems is much smaller
the same weather system.

approach has not been very suc-
in a region. While it is straight-
nal state space, representing all
er (2^N) of parameters does not

d above, we shall build a model
nobserved states (thought of as
unt for different distributions of
re assumed to follow a Markov
rrence of precipitation over the
states, is conditionally indepen-
of model is often called a **hid-**
the general **state space model**

t (throughout this section t will
denote the station number). We
ary transition probabilities

$$\ldots,M \qquad (2.324)$$

$$(2.325)$$

of the weather process we have

$$(2.326)$$

$\ldots,N,\ t=0,1,\ldots$ Write

$$(2.327)$$

$$(2.328)$$

so that $Y(t)$, which can take on values $l=0,1,\ldots,L\equiv 2^N-1$, is the decimal representation of the binary number $\mathbf{X}(t)$, ordered from all stations dry ($Y=0$, $\mathbf{X}=(0,\ldots,0)$) to all stations wet ($Y=N$, $\mathbf{X}=(1,\ldots,1)$). We assume that, given the weather state, the conditional probability of rain is given by

$$\mathbf{P}(Y(t)=l \mid C(t)=m)=\pi_{lm}. \qquad (2.329)$$

The matrix of conditional distributions $\Pi=(\pi_{lm})$ has all columns summing to one. For a square matrix A we write $A_{l\bullet}$ for the lth row vector, and $A_{\bullet l}$ for the vector of elements of the lth column vector.

We assume now that $C(t)$ is in equilibrium, as is therefore $\mathbf{X}(t)$. In the following example we see that in general $\mathbf{X}(t)$ is not a Markov chain.

Example (A hidden Markov model which is not a Markov chain) Let $N=2$, $M=2$, and assume that weather state 1 means that the two sites each have probability p of rain, independently of each other, while weather state 2 means that the two sites independently have probability q of rain. Then

$$\Pi=\begin{bmatrix}(1-p)^2 & (1-q)^2\\ p(1-p) & q(1-q)\\ p(1-p) & q(1-q)\\ p^2 & q^2\end{bmatrix}. \qquad (2.330)$$

Suppose now that

$$\Gamma=\frac{1}{3}\begin{bmatrix}1 & 2\\ 2 & 1\end{bmatrix} \qquad (2.331)$$

so that $\delta=\frac{1}{2}\mathbf{1}$ where $\mathbf{1}$ is a vector of ones. Then one can compute the conditional probabilities

$$\mathbf{P}(\mathbf{X}(t)=(1,1) \mid \mathbf{X}(t-1)=(1,1),\mathbf{X}(t-2)=(1,1))$$
$$=\frac{1}{3}(p^2+q^2)\left[1+\frac{3p^2q^2}{p^4+4p^2q^2+q^4}\right] \qquad (2.332)$$

and

$$\mathbf{P}(\mathbf{X}(t)=(1,1)\,|\,\mathbf{X}(t-1)=(1,1),\mathbf{X}(t-2)=(0,0))$$

$$= \frac{1}{3}\left\{p^2+q^2 + \frac{3p^2q^2((1-p)^2+(1-q)^2)}{p^2(1-p)^2+2p^2(1-q)^2+2(1-p)^2q^2+q^2(1-q)^2}\right\}. \tag{2.333}$$

Setting, for example, $p=0.9$ and $q=0.1$, we see that

$$\mathbf{P}(\mathbf{X}(t)=(1,1)\,|\,\mathbf{X}(t-1)=(1,1),\mathbf{X}(t-2)=(1,1)) = 0.286 \tag{2.334}$$

while

$$\mathbf{P}(\mathbf{X}(t)=(1,1)\,|\,\mathbf{X}(t-1)=(1,1),\mathbf{X}(t-2)=(0,0)) = 0.278. \tag{2.335}$$

Hence **X** is not a Markov chain. □

The likelihood of observations $y_1,\ldots,y_t$ can be written using $\lambda(k)=\mathrm{diag}(\Pi_{k\cdot})$, a diagonal matrix of length M with diagonal elements consisting of the kth row of Π, as

$$L_t(\Pi,\Gamma) = \mathbf{P}(Y(1)=y_1,\ldots,Y(t)=y_t)$$

$$= \delta\lambda(y_1)\Gamma\lambda(y_2)\Gamma\lambda(y_3)\cdots\lambda(y_t)\mathbf{1}^T. \tag{2.336}$$

Application (Snoqualmie Falls precipitation, continued) We fit the Snoqualmie Falls precipitation data using a two-state version of the general model described above. Since we have only one site, $Y=X$. In order to have sufficient amounts of data, we use January–March. The model is given by

$$\Pi = \begin{bmatrix} \pi_{01} & \pi_{02} \\ \pi_{11} & \pi_{12} \end{bmatrix} = \begin{bmatrix} 1-\pi_1 & 1-\pi_2 \\ \pi_1 & \pi_2 \end{bmatrix} \tag{2.337}$$

and

$$\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{12} & \gamma_{22} \end{bmatrix} = \begin{bmatrix} 1-\gamma_1 & \gamma_1 \\ \gamma_2 & 1-\gamma_2 \end{bmatrix}. \tag{2.338}$$

The model is determined by the four parameters $(\pi_1,\pi_2,\gamma_1,\gamma_2)$. The steady-state probabilities are

$$\boldsymbol{\delta} = (\delta_1,\delta_2) = \left[\frac{\gamma_2}{\gamma_1+\gamma_2},\frac{\gamma_1}{\gamma_1+\gamma_2}\right]. \tag{2.339}$$

The following
zation of the likelih
March: $\hat{\gamma}_1=0.326$, $\hat{\gamma}_2$
$\delta=(0.303,0.697)$. S
state, which occurs a
state, occurring 2/3 c
tion, we would ha
significant improven
-1790.2, while that
Markov model havi
therefore rejects the
with a P-value of 0.
with the hidden Mar

In order to con
and that obtained fo
expression for it. No
from 1 to 0 in the X-

$$G(k) =$$

which can easily be
ure 2.17 shows the s
hidden Markov mod
ment over the Mark
confidence band fo
pointwise bands, and
val out of ten fail t
independent surviva
Markov model (alth

The advantage of th
more clearly illustra

Application (A
the general model w
namely Omaha, Neb
Nebraska (site C) u
on the assumption c
three sites. Given a
conditionally indepe
each site varies wit
The matrix Π has en

$$\pi_{lm} = \prod_{i=}$$

The following parameter estimates were obtained by numerical maximization of the likelihood for the Snoqualmie Falls data for January through March: $\hat{\gamma}_1$=0.326, $\hat{\gamma}_2$=0.142, $\hat{\pi}_1$=0.059 and $\hat{\pi}_2$=0.941. From (2.339) we see that δ=(0.303, 0.697). Since $\hat{\pi}_1$ is nearly zero, this corresponds to a mostly dry state, which occurs about 1/3 of the time, while $\hat{\pi}_2$ corresponds to a mostly wet state, occurring 2/3 of the time. If a Markov chain were an appropriate description, we would have π_1=0 and π_2=1. The hidden Markov model is a significant improvement. The likelihood under the hidden Markov model is −1790.2, while that under the Markov chain model (the constrained hidden Markov model having π_1=0 and π_2=1) is −1796.64. The likelihood ratio test therefore rejects the Markov chain model in favor of the hidden Markov model with a P-value of 0.002. Using BIC, however, the two models are comparable, with the hidden Markov model coming out slightly worse.

In order to compare the survival function for dry periods to that observed, and that obtained for the Markov chain, we need to calculate the theoretical expression for it. Notice that a dry period starts whenever there is a transition from 1 to 0 in the X-process. Denoting the survival function by $G(k)$, we have

$$G(k) = \mathbf{P}(X(1)=0,X(2)=0,\ldots,X(k)=0 \mid X(0)=1,X(1)=0) \qquad (2.340)$$

which can easily be expressed in terms of the likelihood function (2.336). Figure 2.17 shows the survival functions of Figure 2.16, and in addition that of the hidden Markov model. We see that the hidden Markov model is an improvement over the Markov chain model, although still falling short of the observed confidence band for k=5. Remember, though, that the confidence bands are pointwise bands, and that one would therefore not be surprised to see one interval out of ten fail to cover. On the other hand, the bands are appropriate for independent survival times. This assumption is not strictly met for the hidden Markov model (although it does hold for the Markov chain model). □

The advantage of the hidden Markov model over the Markov chain model is more clearly illustrated for a network of sites.

Application (A Great Plains network) A simple two-state version of the general model was fitted to the sequence of wet and dry days at three sites, namely Omaha, Nebraska (site A), Des Moines, Iowa (site B) and Grand Island, Nebraska (site C) using records for the period 1949–1984. Our model is based on the assumption of two weather states which are temporally common to all three sites. Given a particular weather state, the event of rain at any given site is conditionally independent of rain at any other site. The probability of rain at each site varies with the weather state, and can be different from site to site. The matrix Π has entries

$$\pi_{lm} = \prod_{i=1}^{3} \theta_{im}^{x_i^l}(1-\theta_{im})^{1-x_i^l},\ 0 \le l \le 7,\ \ m=1,2 \qquad (2.341)$$

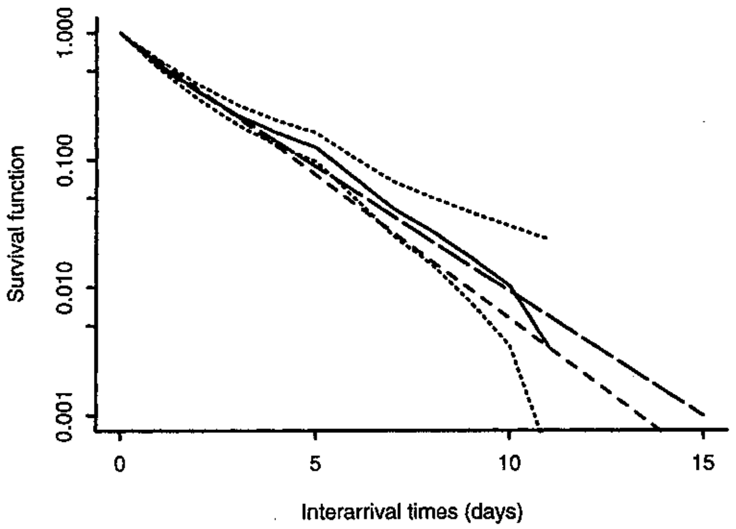


Figure 2.17. Estimated survival function (solid line) for Snoqualmie Falls dry periods in January, together with the survival function for the fitted Markov chain (short-dash line) and for the hidden Markov model (long-dash line). The dotted lines form an asymptotic 95% pointwise confidence band.

where (x_1^l, x_2^l, x_3^l) is the binary representation of l. For example, l=5 corrsponds to the pattern (1, 0, 1) of rain at sites 1 and 3, and no rain at site 2. This model has eight parameters, namely $\boldsymbol{\theta}=(\theta_{11},\theta_{21},\theta_{31},\theta_{12},\theta_{22},\theta_{32})$ and $\boldsymbol{\gamma}=(\gamma_1,\gamma_2)$.

To allow for seasonal changes the year was divided into six seasons, as defined in Table 2.16. Any rain occurring on February 29 was added to that of March 1. The parameter estimates, computed by numerical maximization of the likelihood are also given in the table.

The two weather states which result can again be described as "mostly wet" and "mostly dry", i.e., $\hat{\theta}_{l1}$ are fairly close to one and $\hat{\theta}_{l2}$ close to zero. The estimates vary quite smoothly with respect to changes in season, as does the estimated unconditional probability of being in a given state.

Table 2.17 gives the observed frequencies for season 1 of various events of interest together with the frequencies derived from the fitted model. For comparison we also give the fitted pattern from the Markov chain model, obtained by fitting an 8-state chain and computing the expected pattern under the stationary distribution. Thus, for example, there were 866 days on which it rained at both sites A and B in season 1, while the hidden Markov model predicted 862

Table 2.16 Parameter estimates for Great Plains network

Season	Days	$\hat{\gamma}_1$	$\hat{\gamma}_2$
1	001–060	0.352	0.335
2	061–120	0.358	0.350
3	181–180	0.333	0.381
4	181–240	0.389	0.354
5	241–300	0.389	0.248
6	301–365	0.359	0.269

Season	$\hat{\theta}_{11}$	$\hat{\theta}_{21}$	$\hat{\theta}_{31}$	$\hat{\theta}_{12}$	$\hat{\theta}_{22}$	$\hat{\theta}_{23}$
1	0.927	0.874	0.762	0.039	0.211	0.139
2	0.944	0.875	0.785	0.041	0.196	0.163
3	0.908	0.806	0.782	0.042	0.185	0.208
4	0.851	0.745	0.720	0.010	0.141	0.187
5	0.880	0.804	0.775	0.030	0.126	0.062
6	0.891	0.865	0.707	0.015	0.175	0.066

Table 2.17 Observed and fitted frequencies of rainfall patterns

	Dry	A	B	C	AB	AC	BC	ABC
Obs	718	1020	1154	957	866	752	728	657
HMM	725	1019	1153	956	862	749	728	657
MC	722	942	1076	1031	789	750	727	655

and the Markov chain predicted 789. The hidden Markov model has preserved the spatial dependence structure of the records very well, while the Markov chain model fails badly at that task. □

An important feature of the hidden Markov model is that it is possible to estimate the underlying Markov chain. In the precipitation model above this implies that it is possible to relate the hidden weather states to atmospheric observations (this has been carried out, with promising results, by Hughes, 1993). We shall see how it is done in a different setting below.

Application (Neurophysiology) Cell membranes contain several types of ion channels which allow selected ions to pass between the outside and the inside of the cell. These channels, corresponding to the action of a single protein molecule in the cell membrane, can be open (allowing passage of the selected ion) or closed at any one time. Some channels respond to changes in electric potential, while others are chemically activated. Among the latter are acetylcholine activated channels at so-called post-synaptic membranes of neurons, the basic cells of the nervous system. A more detailed description of neurons is in section 3.8, where we discuss the modeling of neural activity.

Patch clamp recordings provide the principal source of information about channel activity. A glass micro-pipette is positioned at the cell membrane and sealed to it by suction. Currents through the channels in the tip of the pipette are measured directly. The technique was developed by Sakmann and Neher, who were awarded the 1991 Nobel prize in medicine for it.

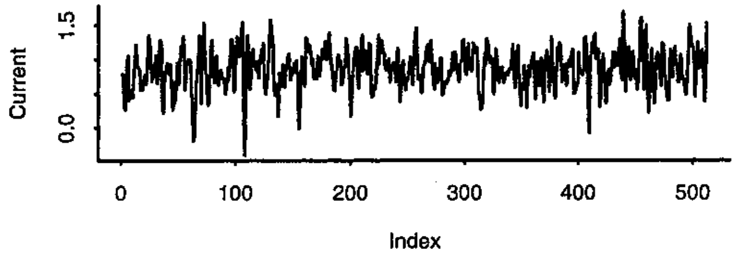


Figure 2.18. Current measured in a single channel of an acetylcholine receptor in a rat.

Figure 2.18 depicts a single channel recording from an acetylcholine receptor. Disregarding the noise, the current appears to move between two levels, corresponding to the open and closed states. The sampling interval is of the order of 10^{-4} seconds, while openings typically last some microseconds.

A very simple model for these observations are independent random variables, $X_1,\ldots,X_n$, distributed $N(\theta_i,\sigma^2)$ where θ_i is 0 or 1, in appropriate units. The likelihood for each n-digit binary sequence $\theta^{(i)}$, $i=1,\ldots,2^n$, is (ignoring irrelevant constants)

$$L(\theta^{(i)}) = \sigma^{-n}\exp(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\theta_j^{(i)})^2). \tag{2.342}$$

Clearly, the mle of θ must be $\hat{\theta}_j=1(x_j>½)$. This method of estimating the channel status is called the **threshold method**.

In order to analyze data such as those in Figure 2.18 we must restore the underlying quantized signal, i.e., the sequence of zeros and ones. If we, as is commonly done, use the threshold method, we tend to get a very ragged reconstruction, since this method does not take into account the fact that nearby signals are more likely to be similar than different.

To avoid the raggedness we need somehow to force the estimate to be smoother. Similarly to the approach in section 2.7, where we estimated the order of a Markov chain, we ensure smoothness by penalizing the likelihood for ragged reconstructions. Technically we shall use what is called **Bayesian statistics**. The way it works is that we assume that $\boldsymbol{\theta}$ is a realization of a stochastic process $\Theta_1,\Theta_2,\ldots$; in this case a Markov chain. Recall Bayes' theorem

$$\mathbf{P}(\boldsymbol{\Theta}=\boldsymbol{\theta} \mid \mathbf{X}=\mathbf{x}) = \frac{\mathbf{P}(\mathbf{X}=\mathbf{x} \mid \boldsymbol{\Theta}=\boldsymbol{\theta})\mathbf{P}(\boldsymbol{\Theta}=\boldsymbol{\theta})}{\sum\limits_{\boldsymbol{\theta}}\mathbf{P}(\mathbf{X}=\mathbf{x} \mid \boldsymbol{\Theta}=\boldsymbol{\theta})\mathbf{P}(\boldsymbol{\Theta}=\boldsymbol{\theta})} \tag{2.343}$$

where the sum on the right-hand side is taken over all the 2^n possible binary n-digit numbers $\boldsymbol{\theta}$. We call $\mathbf{P}(\boldsymbol{\Theta}=\boldsymbol{\theta})$ the **prior** distribution of $\boldsymbol{\Theta}$ (it is obtained prior to observing $\mathbf{X}$), and $\mathbf{P}(\boldsymbol{\Theta}=\boldsymbol{\theta} \mid \mathbf{X}=\mathbf{x})$ the **posterior** distribution. One way of writing (2.343), with $\pi(\boldsymbol{\theta} \mid \mathbf{x})=\log(\mathbf{P}(\mathbf{X}=\mathbf{x} \mid \boldsymbol{\Theta}=\boldsymbol{\theta}))$ and $\pi(\boldsymbol{\theta})=\log \mathbf{P}(\boldsymbol{\Theta}=\boldsymbol{\theta})$, is

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = c(\mathbf{x}) + l_{\mathbf{x}}(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta}) \tag{2.344}$$

where $l_{\mathbf{x}}$ is the log likelihood function. Thus, maximizing $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ is equivalent to maximizing the penalized log likelihood $l_{\mathbf{x}}(\boldsymbol{\theta})+\pi(\boldsymbol{\theta})$. The effect is to discount outcomes that are unlikely under the prior distribution. One way to think about this is that the threshold method maximizes $\pi(\theta_i \mid x_i)$ for each i, while the Bayesian approach maximizes $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$, simultaneously for all i. The two methods are the same whenever the θ_i are conditionally independent given the data, but if there is dependence (smoothness) the results, as we shall see shortly, can be quite different.

If we assume that the x_i are independent, conditionally upon the θ_i, and that the θ_i are a realization of a Markov chain, we must maximize

$$-n\log\sigma^2 + \sum_{i=1}^{n}\log p_{\theta_{i-1},\theta_i} - \frac{1}{2\sigma^2}\sum(x_i-\theta_i)^2. \tag{2.345}$$

In the case where $p_{01}=p_{10}=p<\frac{1}{2}$, this is equivalent to minimizing

$$J\log\frac{1-p}{p} + n\log\sigma^2 + \frac{1}{2\sigma^2}\sum(x_i-\theta_i)^2 + n\log p \tag{2.346}$$

where $J=\sum_1^n 1(\theta_i\neq\theta_{i-1})$ is the number of jumps in the sequence $\boldsymbol{\theta}$. Equivalently we can write (2.346)

$$\log\frac{1-p}{p}\sum(\theta_i-\theta_{i-1})^2 + n\log\sigma^2 + \frac{1}{2\sigma^2}\sum(x_i-\theta_i)^2. \tag{2.347}$$

We see explicitly how the penalty term (the first term in (2.347)) penalizes adjacent dissimilar θ-values.

A naive minimization of (2.344) entails evaluating $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ for all the 2^n possible values of $\boldsymbol{\theta}$, once the transition matrix $\mathbb{P}$ is known. There are two problems with this: 2^n is a gigantic number for a typical data set, and the transition matrix is unknown. In order to avoid the first problem we employ a dynamic