programming algorithm due to Viterbi (1967), while for the second we iterate the Viterbi algorithm to reconstruct θ given our current estimate of $\mathbb{P}$, and use the nonparametric mle of $\mathbb{P}$ from the reconstruction in the next iteration of the Viterbi algorithm.

The basic idea of the Viterbi algorithm is as follows. At time t, suppose that we have computed the most likely sequence up to this time for both the possible values of θ_t. We now proceed to determine the most likely history up to $\theta_{t+1}=j$. If this history has $\theta_t=i$ it must include that of the previously calculated most likely histories which ends in $\theta_t=i$. In this way one moves forward through the data, maintaining and updating two possible most likely histories at each time step. At the end, we compare the likelihood of the two histories to choose between them. Formally, write for $\theta_t=0$ or 1

$$H_t(\theta_t)=\{\hat{\theta}_0,\hat{\theta}_1,\dots,\hat{\theta}_{t-1}\} \tag{2.348}$$

where $\hat{\theta}_0,\hat{\theta}_1,\dots,\hat{\theta}_{t-1}$ maximizes (for the given θ_t)

$$L_t(\theta_0,\dots,\theta_t)=\pi(\theta_0)f(x_0\mid\theta_0)\prod_{k=0}^{t-1}p_{\theta_k,\theta_{k+1}}f(x_{k+1}\mid\theta_{k+1}). \tag{2.349}$$

Here π is the stationary distribution of the chain with transition matrix $\mathbb{P}$, and f is the conditional density of x given θ. In this application we take f to be Gaussian with mean 0 or 1 and common variance σ^2. Then L_t is just the likelihood of the hidden Markov chain $(X_0,\dots,X_t)$, conditional upon the hidden sequence $\theta_0,\dots,\theta_t$. Now consider the maximum $\hat{L}_t(\theta_t)=L_t(\hat{\theta}_0,\dots,\hat{\theta}_{t-1},\theta_t)$. This satisfies the recursive relation

$$\hat{L}_{t+1}(\theta_{t+1})=\max_{\theta_t}\hat{L}_t(\theta_t)p_{\theta_t,\theta_{t+1}}f(x_{t+1}\mid\theta_{t+1}). \tag{2.350}$$

Letting $\hat{\theta}_t$ be the maximizing value in the right-hand side of (2.350) we can write a recursion for the history H_t as

$$H_{t+1}(\theta_{t+1})=\{H_t(\hat{\theta}_t),\hat{\theta}_t\}. \tag{2.351}$$

In practice it often happens that $\hat{\theta}_t$ is the same for both values of θ_{t+1}. Then the two most likely histories have merged at time t, and we no longer need to keep them in active memory.

Before we can apply the procedure to actual data, we need to estimate the noise variance σ^2. The exact value is not very critical; a simple approach is to use the threshold reconstruction of assigning all values above ½ to 1, and all below ½ to zero, and then compute the variance of the residual from the respective means. For the data in Figure 2.18 this yields $\sigma=0.067$. The reconstruction as outlined above was then applied to the data three times (each time updating the variance estimate and the transition matrix estimate based on the previous reconstruction), at which point there was no change in the updated state vector $\hat{\theta}$. The resulting transition probabilities were $p_{01}=0.64$ and $p_{10}=0.06$, and the final $\sigma^2=0.073$. The restoration is shown, together with the maximum likelihood

reconstruction, in Fi
reasonable to use the

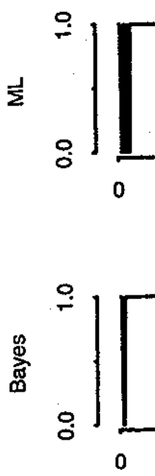

Figure 2.19. Ma
(lower panel) of the

2.13. Bibliograph

Most of the materia
been Bhattacharya
Freedman (1983), C
which is (in its owr
chains.

The material
Julian Besag, Elizat
ney (1991) are nice
on likelihood evalu
Discussion of impl
advances are in the
Carlo test examples

or the second we iterate
t estimate of $\mathbb{P}$, and use
the next iteration of the

ows. At time t, suppose
o this time for both the
most likely history up to
he previously calculated
e moves forward through
t likely histories at each
e two histories to choose

(2.348)

$_{1}f(x_{k+1} \mid \theta_{k+1})$. (2.349)

transition matrix $\mathbb{P}$, and f
ion we take f to be Gaus-
L_t is just the likelihood of
pon the hidden sequence
$\ldots, \hat{\theta}_{t-1}, \theta_t)$. This satisfies

). (2.350)

d side of (2.350) we can

(2.351)

h values of θ_{t+1}. Then the
we no longer need to keep

ta, we need to estimate the
al; a simple approach is to
lues above ½ to 1, and all
e residual from the respec-
=0.067. The reconstruction
e times (each time updating
nate based on the previous
in the updated state vector
).64 and p_{10}=0.06, and the
ith the maximum likelihood

reconstruction, in Figure 2.19. Clearly, for this quite noisy sequence it is not reasonable to use the threshold method.

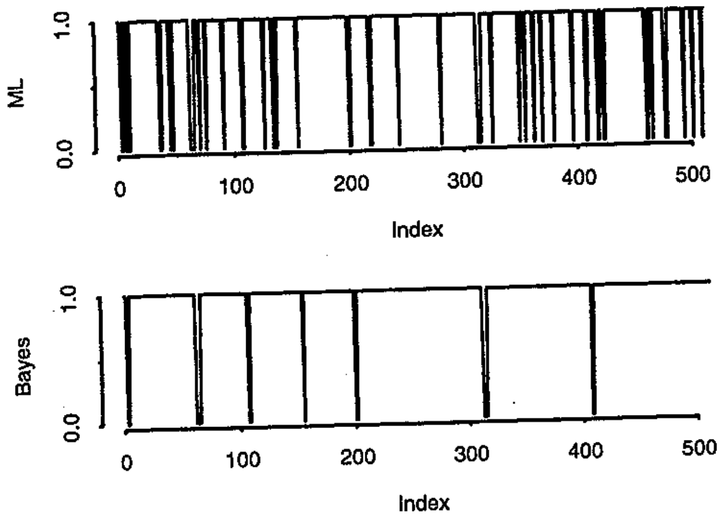

Figure 2.19. Maximum likelihood (upper panel) and Bayesian reconstruction (lower panel) of the signal in Figure 2.18.

□

2.13. Bibliographic remarks

Most of the material in sections 2.1–2.5 is standard. Among my sources have been Bhattacharya and Waymire (1990), Cox and Miller (1965), Feller (1968), Freedman (1983), Grimmett and Stirzaker (1982), and Whittle (1992), each of which is (in its own way) an excellent introduction to the behavior of Markov chains.

The material in section 2.6 owes much to discussions with and papers by Julian Besag, Elizabeth Thompson, and Charlie Geyer. Geyer (1991) and Tierney (1991) are nice presentations of the procedures; the former concentrating on likelihood evaluation, and the latter on computing posterior distributions. Discussion of implementation can be found in Geyer (1992). Some recent advances are in the dissertations by Lin (1993) and Higdon (1994). The Monte Carlo test examples come from Besag and Clifford (1989).

Section 2.7 follows largely Jacobsen and Keiding (1990). The standard source is Billingsley (1961), but it can be difficult to follow at times.

The linear model for higher order chains was developed by Raftery (1985a,b; Raftery and Tavaré, 1994) in a series of papers. A dissertation by Le (1990) developed these ideas further, and the dissertation by Schimert (1992) deals with estimation problems.

Section 2.9 borrows much theory from Cox and Miller (1965) while the application is inspired by Katz (1977). The recursion (2.238) was pointed out by Julian Besag. Section 2.10 is based on the presentation by Doyle and Snell (1984). The material in section 2.11 is from my monograph (Guttorp, 1991). The application to the Amazon indian tribe follows Thompson (1976).

Finally, section 2.12 comes from two main sources, namely Zucchini and Guttorp (1991) for the precipitation model, and Fredkin and Rice (1992) for the application to neurophysiology. General material on hidden Markov models can be found in Juang and Rabiner (1991), albeit from the point of view of speech recognition, and Bayesian image analysis is described in Besag (1989).

2.14. Exercises

Theoretical exercises

1. Prove that the Markov property (2.5) is equivalent to each of the following statements:

(a) Let T_1 be a set of times later than n, and T_0 a set of times less than or equal to n. Let t_0=max T_0. Then

$$\mathbf{P}(X_k=x_k,\ k\in T_1 \mid X_l=x_l, l\in T_0) = \mathbf{P}(X_k=x_k,\ k\in T_1 \mid X_{t_0}=x_{t_0}).$$

(b) Let T_1 be a set of times later than n, and T_0 a set of times prior to n. Then

$$\mathbf{P}(X_k=x_k, k\in T_1, X_l=x_l, l\in T_0 \mid X_n=x_n)$$
$$= \mathbf{P}(X_k=x_k, k\in T_1 \mid X_n=x_n)\mathbf{P}(X_l=x_l, l\in T_0 \mid X_n=x_n).$$

2. Show that the time spent in state k upon each return to it for a Markov chain has a geometric distribution with mean $1/(1-p_{kk})$.

3. Show that $\lambda_0=1-2^{-k}$ in the Reid–Landau model of radiation damage.

4. Prove that for a fair simple random walk on the integer lattice in 2 dimensions, **0** is persistent.

5. Prove the time invariance result given by equation (2.80).

6. Show that a transition matrix $\mathbb{P}$ on a finite state space has a uniform stationary distribution iff it is doubly stochastic.

7. Derive the
a Pólya urn.

8. Verify tha
to the ergodi

9. Show that
the chain wit

10. A slight
tion of both
Hastings (19
assume for si

when $f(x)Q$
date point Y
$X_{n+1}=y$, and

(a) Show tha

(b) Show tha

11. Let X_k
order to eval
nalization: Q
the diagonal

(a) Show tha

(b) Show tha

and

(c) Using tha
(2.238) simp

(d) Deduce t

Hint: Argue

ding (1990). The standard
ollow at times.

as developed by Raftery
apers. A dissertation by Le
tation by Schimert (1992)

d Miller (1965) while the
(2.238) was pointed out by
ation by Doyle and Snell
onograph (Guttorp, 1991).
hompson (1976).

rces, namely Zucchini and
kin and Rice (1992) for the
idden Markov models can
e point of view of speech
in Besag (1989).

t to each of the following

of times less than or equal

$k \in T_1 \mid X_{t_0}=x_{t_0}$).

f times prior to n. Then

$\mid X_n=x_n$).

n to it for a Markov chain

radiation damage.

nteger lattice in 2 dimen-

(2.80).

ace has a uniform station-

7. Derive the marginal distribution of the number of red balls after n draws from a Pólya urn.

8. Verify that Theorem 2.7 holds when the process starts from any state leading to the ergodic state k.

9. Show that the Hardy–Weinberg proportions are the stationary distribution for the chain with transition matrix (2.142).

10. A slight variant of the Metropolis algorithm, which is actually a generalization of both the Gibbs sampler and the Metropolis algorithm, was proposed by Hastings (1970). Let Q be a transition matrix, not necessarily symmetric, and assume for simplicity that the target function $f(x)>0$ for all x. Define

$$\alpha(x,y) = \min\left\{\frac{f(y)Q(y,x)}{f(x)Q(x,y)},1\right\}$$

when $f(x)Q(x,y)>0$, and $\alpha(x,y)=1$ otherwise. Let $X_n=x$, and generate a candidate point $Y=y$ from the distribution $Q(x,\bullet)$. With probability $\alpha(x,y)$ we set $X_{n+1}=y$, and with the complementary probability $X_{n+1}=x$.

(a) Show that the process (X_n) is a Markov chain with stationary distribution f.

(b) Show that the Metropolis algorithm results for symmetric Q.

11. Let X_k be a 0-1 process, and consider the distribution of $S_n=\sum_{i=1}^n X_k$. In order to evaluate (2.238) we must compute $Q(t)^n$. This can be done by diagonalization: $Q(t)=E\,D\,E^{-1}$ where E contains the right eigenvectors of Q and D is the diagonal matrix of the corresponding eigenvalues.

(a) Show that the eigenvalues solve

$$\lambda^2-\lambda(p_{11}e^{-t}+p_{00})+\det(\mathbb{P})e^{-t}=0.$$

(b) Show that the eigenvectors are given by

$$\mathbf{q}_0(t)=(p_{01}e^{-t},\lambda_1(t)+p_{00})$$

and

$$\mathbf{q}_1(t)=(p_{01}e^{-t},\lambda_2(t)+p_{00}).$$

(c) Using that $\lambda_0(t)+\lambda_1(t)=p_{00}+p_{11}e^{-t}$ and $\lambda_0(t)\lambda_1(t)=(p_{00}-p_{01})e^{-t}$, show that (2.238) simplifies to

$$\phi_n(t)=\mathbf{E}^{\pi}(e^{-tS_n})=\frac{\lambda_1^{n+1}-\lambda_2^{n+1}-(p_{11}-p_{01})(\lambda_1^n-\lambda_2^n)}{\lambda_1-\lambda_2}.$$

(d) Deduce that

$$S_n\sim\mathrm{AsN}\left[\frac{n(p_{01}^2+p_{10}^2)}{(p_{01}+p_{10})^2},\frac{np_{01}p_{10}(p_{00}+p_{11})}{(p_{01}+p_{10})^3}\right].$$

Hint: Argue that $\log\phi_1^{(n)}(t)\approx n\log\lambda_1(t)$ and do a Taylor expansion of the eigenvalue.

12. Let X_n be a 0-1 chain. Derive the correlation between $\sum_k^l X_i$ and $\sum_m^n X_j$, where $l<m$. Apply the result to the Snoqualmie Falls data with $k=1$, $l=7$, $m=8$, and $n=14$.

13. Let $p_k=bc^{k-1}$, $k\geq 1$ and $p_0=1-\sum_{k\geq 1} p_k$.

(a) Write $P(s)$ in the form

$$P(s) = \frac{\alpha+\beta s}{\gamma+\delta s},\ \alpha\delta-\beta\gamma\neq 0.$$

This is called a **fractional linear** generating function.

(b) Consider a BGW branching process with offspring distribution given by P(s). Show by induction that $P_k(s)$ is a fractional linear generating function.

14. Suppose that $p_{ij}>0$ for all $i,j\in S$, where S is a finite set, so that the chain X_n has a limiting distribution. Show that

$$|p_{ij}^{(n)}-\pi_i| \leq (1-d\delta)^n$$

where $d=|S|$ and $\delta=\min p_{ij}$.

Hint: Divide the terms in the equation $\sum_j (p_{ij}-p_{kj})=0$ into those with positive and negative values. The sum of the positive terms is bounded by $1-n\delta$. Now bound $\max_i p_{ij}^{(n+1)}-\min_i p_{ij}^{(n+1)}$ using Chapman–Kolmogorov.

15. Using the parametric result in Theorem 2.16 and the parametrization $p_{ij}(\boldsymbol{\theta})=\theta_{ij}$ verify the nonparametric result in Theorem 2.14.

16. Prove or disprove the following statement: The Markov property is equivalent to the property that for any class of (measurable) sets $A_1,\ldots,A_n$ we have that

$$\mathbf{P}(X_n\in A_n \mid X_1\in A_1,\ldots,X_{n-1}\in A_{n-1}) = \mathbf{P}(X_n\in A_n \mid X_{n-1}\in A_{n-1}).$$

17. Let S_n be a simple random walk. Show that $|S_n|$ is a Markov chain, and determine its transition probabilities.

Hint: What can you say about the distribution of S_n given the values of $|S_n|,\ldots,|S_1|$?

18. Show that all two-state chains are reversible.

19. Let (X_k) be a K-state ergodic Markov chain with transition matrix P, such that all $P_{ij}>0$. Let $p=(p_1,\ldots,p_K)$ be a probability distribution. Develop a likelihood ratio test for the hypothesis that p is a stationary distribution for (X_k).

Remark: You may not be able to get an explicit solution (except if K=2), but you should develop a system of equations that the constrained estimates have to satisfy.

20. It is not very difficult (at least in principle) to allow for a Markov chain with continuous state space. Let $X_1, X_2, \ldots$ be continuous random variables such that the density of X_k given that $X_{k-1}=x$ is given by $f\,(x,\bullet)$. Similar to the discrete state case, we say that the process is a Markov chain if the conditional density of X_n, given $X_{n-1}, \ldots, X_1$, depends only on the value of X_{n-1}.

(a) Let $f_n(x,\bullet)$ denote the conditional density of X_n, given that $X_0=x$. Show that the Chapman–Kolmogorov equation holds in the form

$$f_{n+m}(x,y) = \int_{\mathbb{R}} f_m(x,z) f_n(z,y)\,dz.$$

(b) Show that the stationary distribution (when it exists) has density $\pi(y)$ satisfying

$$\pi(y) = \int_{\mathbb{R}} \pi(x) f(x,y)\,dx.$$

21. The hidden Markov model originated in work in engineering, where a system described by the **state variable** $\mathbf{X}_k$ of dimension s was assumed to develop according to the **state equation**

$$\mathbf{X}_k = \mathbf{X}_{k-1}\mathbf{A} + \varepsilon_k$$

where ε_k is assumed $\mathrm{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbb{I}_{s\times s})$. However, the system cannot be observed directly. Instead, one observes a related random vector $\mathbf{Y}_k$, of dimension o, satisfying the **observation equation**

$$\mathbf{Y}_k = \mathbf{X}_k\mathbf{B} + \delta_k$$

where δ_k is $\mathrm{N}(\mathbf{0}, \sigma_\delta^2 \mathbb{I}_{o\times o})$. The parameters $\mathbf{A}$, $\mathbf{B}$, σ_ε^2 and σ_δ^2 were originally taken as known, although modern applications allow for estimation of these parameters. The **Kalman filter** (Kalman, 1960) estimates the value of $\mathbf{X}_k$, given observations $\mathbf{Y}_1^k$. Show that $\mathbf{m}_k=\mathbf{E}(\mathbf{X}_k \mid \mathbf{Y}_1^k)$ and $\Gamma_k=\mathbf{Var}(\mathbf{X}_k \mid \mathbf{Y}_1^k)$ satisfy the recursions

$$\mathbf{m}_{k+1} = \mathbf{m}_k\mathbf{A} + (Y_k - \mathbf{m}_k\mathbf{B})\mathbf{K}_k$$

and

$$\Gamma_{k+1} = \mathbf{A}^T\Gamma_k(\mathbb{I} - B(\mathbf{B}^T\Gamma_k\mathbf{B} + \sigma_\delta^2\mathbb{I})^{-1}\mathbf{B}^T)\mathbf{A} + \sigma_\varepsilon^2\mathbb{I}$$

where the **gain** $\mathbf{K}_k$ is given by

$$\mathbf{K}_k = (\mathbf{B}\Gamma_k\mathbf{B}^T + \sigma_\delta^2\mathbb{I})^{-1}\mathbf{B}^T\Gamma_k\mathbf{A}.$$

Hint: Use the formula for conditional expectation of jointly Gaussian random variables, i. e., normal regression.

22. Let (X_k) be a random walk, but suppose that it is observed with independent measurement error.

(a) Write this process in state space form.

(b) Derive the Kalman filter for it.

Computing exercises

C1. Given a subroutine that generates multinomial random vectors from input values for sample size and probability vector, how would you build a subroutine that generates a Markov chain with given (finite-dimensional) transition matrix $\mathbb{P}$?

C2. (a) How long a stretch of data from an ergodic chain do you need to estimate $\mathbb{P}$ accurately? You choose $\mathbb{P}$ and what you mean by ''accurately''.

(b) How long a stretch do you need to estimate the stationary distribution accurately?

C3. (a) Let (X_k) be a Markov chain with transition matrix

$$P = \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.3 & 0.1 \\ 0.5 & 0.1 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.5 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.2 & 0.5 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.4 & 0.1 \end{bmatrix}.$$

Starting from state 1, find $\mathbf{E}^1(T_5)$.

(b) Suppose that you do not know P, but have a realization $(x_k, k \le 100)$ of (X_k). Find a way of estimating the quantity in (a). Perform the estimation on a simulated realization.

(c) The expected value in part (a) is a certain function of

$$F_{15}(t;P) = \mathbf{P}^1\{T_5 \le t\}.$$

It is complicated to determine this distribution analytically, even when P is known. If P is unknown, it is necessary to apply simulation methods. The **bootstrap** (Efron, 1978) can be modified to do this. Let $\hat{P}$ be the estimated transition matrix. Generate a path from $\hat{P}$, and use that to estimate the transition matrix the usual way, yielding a matrix $\hat{P}^*$. The bootstrap idea is to say that the distribution of

$$\sqrt{n}\,(F_{15}(t;\hat{P}) - F_{15}(t;P))$$

is well approximated by that of

$$\sqrt{n}\,(F_{15}(t;\hat{P}^*) - F_{15}(t;\hat{P})).$$

In practice, one would use repeated samples from $\hat{P}$ to get realizations of the hitting time, and then use the empirical distribution function of these hitting times to estimate $F_{15}(t;P)$ as well as whatever function of it you are interested in. The approximation argument above is the theoretical basis for this procedure. Compute a bootstrap estimate using the sample generated in part (b).

Use a bootstrap sample of size 100. Compare the mean of the resulting estimate to the answer in (b). Can you use the bootstrap distribution to estimate the variance of the estimated mean?

C4. An electrical network consists of nine nodes, labeled a to i. A battery is connected between nodes a and i such that the potential difference $v(a)-v(i)$ is 1. [Without loss of generality $v(a)=1$, $v(i)=0$.] The network consists of the following eighteen wires, with the given resistances:

10-ohm resistance: *cf*

5-ohm resistance: *ab*, *bc*, *bh*, *ce*, *hi* and *fi*

2-ohm resistance: *ac*, *bd*, *cd*, *de*, *df*, *eg* and *gi*

1-ohm resistance: *ad*, *ef*, *eh* and *fg*.

Find the transition matrix for the corresponding Markov random walk on the nodes a to i, and hence find the potentials of the nodes in each of the following ways:

(a) Use the method of relaxation to solve the equation $v = \mathbb{P}v$.

(b) Use the exact method (2.276) to find the solution; be careful about how you put in the boundary conditions, and remember to make a and i absorbing states.

(c) Simulate the random walk to find $v(e)$. See how large a sample you need to obtain an accurate estimate. See if you can think of a way of estimating all of v without having to do a lot of separate simulations.

C5. Simulate the χ^2 statistic on p. 76 (based on the Snoqualmie Falls Markov model) and compare its distribution under this model to the χ^2-distribution suggested by standard (iid) goodness-of-fit theory.

C6. Generate a data set of 100 observations from a mixture of three exponentials with means 48.7, 5.83, and 0.65 with weights 0.0032, 0.1038, and 0.8930, respectively. From these data, using a MCMC method (Gibbs sampler, Metropolis algorithm, or otherwise), estimate the parameters of the underlying distribution. Compare the results to a regular optimization of the likelihood. How many iterations do you need to estimate the likelihood accurately?

Data exercises

D1. Data set 1 contains the Snoqualmie Falls precipitation occurrence data. Extract the January precipitation (the first 31 numbers on each line), and recode it to 0 or 1, where 1 denotes measurable precipitation.

(a) Test the hypothesis that the years are identically distributed.

(b) Current research in precipitation modeling (Woolhiser, 1992) indicates that the El Niño-Southern Oscillation phenomenon (an unusual pattern of surface pressure and sea surface temperature in the south Pacific) may have a profound effect on precipitation occurrence in North America. Table 2.18 contains the Southern Oscillation Index (SOI), defined as the difference in mean January sea level pressure between Tahiti and Darwin, for each of the years of the Snoqualmie Falls data. Assess the relationship between transition probabilities and the Southern Oscillation Index for January.

Table 2.18 January SOI 1948–1983

38	29	55	79	25	49	57	33	68	56
9	26	45	38	80	62	36	35	19	74
53	16	22	50	52	38	88	33	69	37
38	36	51	50	64	–20				

D2. Gabriel and Neumann (1962) analyzed precipitation data from the rainy season (December through February) in Tel Aviv, Palestine/Israel, from 1923 to 1950. Table 2.19 contains the data.

Table 2.19 Tel Aviv precipitation

Previous days		Current day		
Second	First	Wet	Dry	Total
Wet	Wet	439	249	688
Dry	Wet	248	192	350
Wet	Dry	87	261	348
Dry	Dry	263	788	1051

We may contemplate a 3-parameter submodel of the general second-order Markov chain, based on the idea of relatively short-term fronts passing through the area. Then the only long-term influence should be in cases where it rained the previous day. This suggests the following transition matrix:

$$\mathbb{P} = \begin{pmatrix} p_{ww} & 1-p_{ww} \\ p_{dw} & 1-p_{dw} \\ p_d & 1-p_d \\ p_d & 1-p_d \end{pmatrix}.$$

Evaluate this model using the BIC.

D3. A common assumption in sociology is that the social classes of successive generations in a family follow a Markov chain. Thus, the occupation of a son is assumed to depend only on his father's occupation and not on his grandfather's. Suppose that such a model is appropriate, and that the transition matrix is given by

$$\mathbb{P} = \begin{pmatrix} 0.4 & 0.5 & 0.1 \\ 0.05 & 0.7 & 0.25 \\ 0.05 & 0.5 & 0.45 \end{pmatrix}.$$

Here the social classes are numbered 1 to 3, with 1 the highest. Father's class are the rows, son's class the columns. A sociologist has some data that is supposed to illustrate this model. However, he got these data from a colleague, who did not say whether the counts had father's class along rows or columns. The counts are

$$\begin{bmatrix} 7 & 6 & 5 \\ 9 & 207 & 60 \\ 2 & 64 & 60 \end{bmatrix}.$$

Try to determine the more likely labeling.
Hint: How are the counts from a chain run backwards related to those from the same chain run forwards?

D4. In the early 1950s, the Washington Public Opinion Laboratory in Seattle carried out "Project Revere" which was intended to study the diffusion of information (in particular, since the project was funded by the US Air Force, information contained in leaflets dropped from the air). A subexperiment took place in a village with 210 housewives. Forty-two of these were told a slogan about a particular brand of coffee. Each housewife was asked to pass the information on. As an incentive, participants were told that they would get a free pound of coffee if they knew the slogan when an interviewer called 48 hours later. It was possible to trace the route by which each hearer had obtained the slogan, so that they could be classified by generations. The data are given in Table 2.20.

Table 2.20 Spread of slogan

Generation	1	2	3	4	5
Size	69	53	14	2	4

Using a branching process approach, estimate the mean number of offspring in the first generation. How about the final generation?

D5. The data in Table 2.21 show the state of health (1 = self-care, 2 = intermediate care, 3 = intensive care) for 47 patients suffering from chronic bronchial asthma during five different asthmatic seasons (Jain, 1986). For each season we give the observed transitions between the three health states. Test the hypothesis of stationarity, i.e., that the transition matrix does not depend on season.

D6. Table 2.22 contains second-order transition counts between high (1) and low (0) return weeks, relative to mean weekly returns in the past, for weekly nominal returns on a value-weighted portfolio of stocks (McQueen and Thorley, 1991) from April, 1975, through December, 1987. Test the hypothesis that the portfolio returns perform a random walk.

Table 2.21 Severity of asthma by season

	State of health		
Season	1	2	3
	19	1	2
Winter	2	9	4
	1	2	7
	15	2	2
Trees	3	10	4
	1	1	9
	17	1	2
Grass	3	10	5
	1	1	7
	13	2	3
Ragweed	3	12	3
	1	1	9
	12	2	3
Fall	6	6	7
	1	2	8

Table 2.22 High and low return weeks for a stock portfolio

Previous		0	1
0	0	63	75
0	1	74	90
1	0	75	88
1	1	89	109

D7. The data in data set 2 are bi-daily counts of the number E_k of emerging blowflies (*Lucilia cuprina*) and the number D_k of deaths in a cage maintained in a laboratory experiment by the Australian entomologist A. J. Nicholson. The flies were supplied with ample amounts of water and sugar, but only limited amounts of meat, which is necessary for egg production. Let $\mathbf{A}_t = (A_{1,t}, \ldots, A_{m,t})^T$ be the age distribution vector, so $A_{x,t}$ is the number of individuals aged x at time t. Let $p_{k,t}$ be the proportion of individuals aged k at time t that survive to age $k+1$, and let P_t be the survival matrix having $p_{k,t}$ on the sub-diagonal, and zeroes elsewhere. A model for the population dynamics

(Brillinger et al., 1980) is given by

$$\mathbf{A}_{t+1} = P_t(H_t) + \mathbf{B}_{t+1} + \boldsymbol{\varepsilon}_t$$

where $\boldsymbol{\varepsilon}_t$ is an age- and time-dependent noise sequence and H_t denotes the history of the population sizes N_t up to time t.

(a) Write this model, using $\mathbf{A}_t$ as the state, in a state space model like that in Exercise 21 (except it will have random time-dependent coefficients).

(b) Deduce that the state vector can be estimated by

$$a_{1,t+1} = E_{t+1}$$

$$a_{k,t+1} = \frac{p_{k-1,t}\, a_{k-1,t}}{\sum_{l \geq 1} p_{l,t}\, a_{l,t}} (N_{t+1} - E_{t+1}), \quad k > 1$$

either using a Kalman filter approach, or by using an intuitive argument.

(c) For the model

$$p_{i,t} = (1-\alpha_i)(1-\beta N_t)$$

estimate the parameters α_i and β.
Hint: One possibility is to compare observed and predicted deaths for the next time interval, in a (weighted) least squares fashion. Another is to try a likelihood approach.

D8. The data in data set 3 consists of four years of hourly average wind directions at the Koeberg weather station in South Africa. The period covered is 1 May 1985 through 30 April 1989. The average is a vector average, categorized into 16 directions from N, NNE,... to NNW. Analyze the data using a hidden Markov model, with 2 hidden states, and conditionally upon the state a multinomial observation with 16 categories. Does the fitted model separate out different weather patterns? Also estimate the underlying states, and look for seasonal behavior in the sequence of states.

D9. The data in Table 2.23 are numbers of movements by a fetal lamb observed by ultrasound and counted in successive 5-second intervals (Wittman et al., 1984; given by Leroux and Puterman, 1992). Changes in activity may be due to physical changes in the uterus, or to the development of the central nervous system. Assume that there is an underlying unobserved binary Markov chain Z_i, such that if Z_i=k, the observed counts have a Poisson distribution with parameter λ_k, k=0,1. Fit this model using maximum likelihood.

D10. Christchurch aerodrome is one of two international airports in New Zealand. The runways at Christchurch are prone to fog occurrence, and fog forecasting is difficult, particularly because of the sheltering of the area by the Southern Alps. Renwick (1989) reports hourly data on weather type for 1979–1986, a total of 70,128 observations. Table 2.24 show the transition frequencies for observations between 2 hours before and 3 hours after sunrise. The

Table 2.23 Fetal lamb movements

0	0	0	0	0	1	0	1	0	0	0	0	0	0	1
0	1	0	0	0	0	2	2	0	0	0	0	1	0	0
1	1	0	0	1	1	1	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	7	3	2	3	2	4
0	0	0	0	1	0	0	0	0	0	0	0	1	0	2
0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
2	1	0	0	1	0	0	0	1	0	1	1	0	0	0
1	0	0	1	0	0	0	1	2	0	0	0	1	0	1
1	0	1	0	0	2	0	1	2	1	1	2	1	0	1
1	0	0	1	1	0	0	0	1	1	1	0	4	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2.24 Fog and mist occurrence at Christchurch aerodrome

		Next		
		Clear	Mist	Fog
Now	Clear	13,650	245	30
	Mist	480	427	71
	Fog	11	171	198

overall distribution of weather type has clear weather 95.1% of the time, mist 3.4% of the time, and fog 1.5% of the time. Test whether the hours around sunrise are reasonably described by these stationary probabilities.

APPENDIX A
Some statistical theory

A.1. Multinomial likelihood

Let **X** be a random r-vector having the multinomial distribution with r categories, n draws and success probabilities $\mathbf{p}=(p_1,\ldots,p_r)$, so that

$$\mathrm{P}(\mathbf{X}=\mathbf{x}) = \frac{n!}{x_1!\cdots x_r!}p_1^{x_1}\cdots p_r^{x_r} \tag{A.1}$$

provided that $\mathbf{x}\mathbf{1}^T=n$. We write this $\mathbf{X}\sim\mathrm{Mult}_r(n;\mathbf{p})$. Having observed the outcome **x**, we want to estimate **p**. The **likelihood** $L(\mathbf{p})$ of the parameter **p** (given the outcome **x**) is the probability[1] of the observed value as a function of the unknown parameter. The **maximum likelihood estimator** (mle) $\hat{\mathbf{p}}$ of **p** is the value of **p** that maximizes $L(\mathbf{p})$. Think of it as the value of **p** that best explains the observation **x**. To perform the maximization, first note that we must have $\mathbf{p}\mathbf{1}^T=1$. It helps to take logarithms, and thus to maximize

$$l(\mathbf{p}) = \log L(\mathbf{p}) = \log\frac{n!}{x_1!\cdots x_r!} + \sum_{i=1}^{r}x_i\log p_i \tag{A.2}$$

The function $l(\mathbf{p})$ is called the **log likelihood function**. Notice that the first term on the right hand side of (A.2) does not depend on **p**, so that when we maximize $l(\mathbf{p})$ we can ignore it. Generally, the likelihood is only defined up to additive constants (with respect to the unknown parameters).

Since we want to maximize $l(\mathbf{p})$ over all **p** that sum to 1, we introduce a Lagrange multiplier λ, and maximize

$$l^*(\mathbf{p},\lambda) = \sum_{i=1}^{r}x_i\log p_i + \lambda(1-\sum_{i=1}^{r}p_i). \tag{A.3}$$

Taking partial derivatives we get

$$\frac{\partial}{\partial p_i}l^*(\mathbf{p},\lambda) = \frac{x_i}{p_i}-\lambda$$
$$\frac{\partial}{\partial\lambda}l^*(\mathbf{p},\lambda) = 1-\sum p_i \tag{A.4}$$

[1] If **X** has a continuous distribution, the likelihood is defined as the probability density of the observation as a function of the parameter.

and setting them equal to zero yields the equations

$$p_i = \frac{x_i}{\lambda}, \quad \sum p_i = 1 \tag{A.5}$$

whence $\sum x_i/\lambda = 1$, or $\lambda = \sum x_i = n$, so $\hat{p}_i = x_i/n$.

Example (Social mobility) Mosteller (1968) quotes some data on occupational mobility in Denmark. The data are counts of fathers and sons in five different occupational categories. We show the distribution of the 2,391 sons in Table A.1.

Table A.1 Danish social mobility data

Category	1	2	3	4	5
Count	79	263	658	829	562

We estimate the probability p_i that a randomly chosen son from the population studied is in occupational category i, yielding

$$\hat{p}_1 = 0.03 \quad \hat{p}_2 = 0.11 \quad \hat{p}_3 = 0.275 \quad \hat{p}_4 = 0.35 \quad \hat{p}_5 = 0.235 \tag{A.6}$$

□

So far we have derived an estimate of **p**, i.e., a function $\hat{\mathbf{p}}(\mathbf{x})$ of the data **x**. When discussing properties of an estimation procedure, it is customary to think of the **estimate** $\hat{\mathbf{p}}(\mathbf{x})$ as an instance of the random variable $\hat{\mathbf{p}}(\mathbf{X})$, called the **estimator.** Note that $X_i \sim \mathrm{Bin}(n, p_i)$, so by the law of large numbers $\hat{p}_i = X_i/n \to p_i$ with probability 1. Furthermore, by the central limit theorem,

$$n^{1/2} \frac{\hat{p}_i - p_i}{(p_i(1-p_i))^{1/2}} \to \mathrm{N}(0,1) \quad \text{in distribution}\,. \tag{A.7}$$

A.2. The parametric case

Sometimes we are interested in **p** that are a function of a lower-dimensional parameter θ. Let θ be one-dimensional for simplicity. We want to maximize $l(\theta) = l(\mathbf{p}(\theta))$. Presumably, $\mathbf{p}(\theta)$ is parametrized to sum to one, so we do not need to constrain the maximization. By the chain rule we must solve

$$\frac{dl(\theta)}{d\theta} = \sum_{i=1}^{r} \frac{\partial l(\mathbf{p}(\theta))}{\partial p_i(\theta)} \frac{dp_i(\theta)}{d\theta} = 0 \tag{A.8}$$

or, equivalently,

$$\sum_{i=1}^{r} \frac{x_i}{p_i(\theta)} \frac{dp_i(\theta)}{d\theta} = 0. \tag{A.9}$$

This equation may or may not have a solution, but it is possible to write down regularity conditions under which it, for n large enough, will have a solution with probability approaching 1 (see Rao, 1973, sec. 5e, for details). Under regularity conditions Rao also proves consistency of $\hat{\theta}$, as well as asymptotic normality in the sense that

$$\left[-l_n''(\hat{\theta})\right]^{½}(\hat{\theta}-\theta)\xrightarrow{d}\mathrm{N}(0,1). \tag{A.10}$$

Example (Identical twins) Human twins are either identical or nonidentical. The identical twins arise from the splitting of a fertilized egg, while nonidentical twins occur when two ova are fertilized simultaneously. Assuming that the probability of twins being identical is α, and the probability of male children is π=0.516, how can we estimate α from data on numbers and genders of twins, but without knowledge of what twins are identical and what are nonidentical?

First note that two twins can both be male either by being identical or by being nonidentical. The former case has probability $\alpha\pi$, while the latter has probability $(1-\alpha)\pi^2$, for a total probability of $\pi(\alpha+\pi(1-\alpha))$. Likewise the probability of female twins is $(1-\pi)(\alpha\pi+1-\pi)$, and that of mixed twins is $2\pi(1-\pi)(1-\alpha)$, since such twins can only be nonidentical. Hence, writing n_1 for the number of male twins, n_2 for the number of female twins, and n_3 for the number of mixed twins, the log likelihood is

$$l(\alpha)\propto n_1\log(\alpha+\pi(1-\alpha))+n_2\log(\alpha\pi+1-\pi)+n_3\log(1-\alpha) \tag{A.11}$$

Differentiating this and setting the derivative equal to zero yields a quadratic equation in α. □

A.3. Likelihood ratio tests

Suppose we are interested in testing a fully specified hypothesis of the from $\mathbf{p}=\mathbf{p}_0=(p_{10},\ldots,p_{r0})$. General testing theory suggests the usefulness of the **likelihood ratio**

$$\Lambda=\frac{L(\hat{\mathbf{p}})}{L(\mathbf{p}_0}=\prod_{i=1}^{r}\left(\frac{\hat{p}_i}{p_{i0}}\right)^{x_i}. \tag{A.12}$$

As before, it is helpful to take logarithms (and this time multiply by two), to get the **log likelihood ratio statistic**

$$\lambda=2\log\Lambda=2\sum_{i=1}^{r}x_i\log\left(\frac{x_i}{np_{i0}}\right). \tag{A.13}$$

The **likelihood ratio test** rejects for large values of λ. One can show that under the null hypothesis $\mathbf{p}=\mathbf{p}_0$, the statistic λ is approximately distributed as $\chi^2(r-1)$.

but it is possible to write down
ge enough, will have a solution
73, sec. 5e, for details). Under
ncy of θ, as well as asymptotic

(A.10)

twins are either identical or
the splitting of a fertilized egg,
va are fertilized simultaneously.
dentical is α, and the probability
nate α from data on numbers and
what twins are identical and what

le either by being identical or by
bility $\alpha\pi$, while the latter has pro-
$+\pi(1-\alpha)$). Likewise the probabil-
of mixed twins is $2\pi(1-\pi)(1-\alpha)$,
nce, writing n_1 for the number of
, and n_3 for the number of mixed

$(\alpha\pi+1-\pi) + n_3\log(1-\alpha)$ (A.11)

equal to zero yields a quadratic
□

specified hypothesis of the from
y suggests the usefulness of the

(A.12)

l this time multiply by two), to get

(A.13)

ues of λ. One can show that under
roximately distributed as $\chi^2(r-1)$.

For large n, we can write

$$\lambda \approx \sum_{i=1}^{r} \frac{(x_i-np_{i0})^2}{np_{i0}} \equiv \chi^2. \tag{A.14}$$

Since X_i~Bin(n,p_{i0}) under the null hypothesis, the quantity np_{i0} is simply the expected value of X_i. The quantity χ^2 is a measure of how far the observations deviate from their expected values, and was introduced by Karl Pearson[1] (1900). To see the validity of the approximation, we do a Taylor expansion of the logarithm,

$$\begin{aligned}\log\hat{p}_i - \log p_{i0} &\approx \frac{\hat{p}_i-p_{i0}}{p_{i0}} - \frac{(\hat{p}_i-p_{i0})^2}{2p_{i0}^2}\\ &= \frac{X_i-np_{i0}}{np_{i0}} - \frac{(X_i-np_{i0})^2}{2n^2p_{i0}^2}.\end{aligned} \tag{A.15}$$

Thus

$$\begin{aligned}2\sum X_i\log\left[\frac{\hat{p}_i}{p_{i0}}\right] &\approx 2\sum_{i=1}^{r}\left[\frac{(X_i-np_{i0})^2}{np_{i0}} - \frac{(X_i-np_{i0})^3}{2n^2p_{i0}^2}\right]\\ &\quad + 2\sum_{1}^{r}\left[(X_i-np_{i0}) - \frac{(X_i-np_{i0})^2}{2np_{i0}}\right].\end{aligned} \tag{A.16}$$

Some algebra shows that this simplifies to

$$\sum_{i=1}^{r}\frac{(X_i-np_{i0})^2}{np_{i0}} - \sum_{i=1}^{r}\frac{(X_i-np_{i0})^3}{n^2p_{i0}^2}. \tag{A.17}$$

The last term is bounded by

$$\sum_{i=1}^{r}\frac{(X_i-np_{i0})^2}{np_{i0}}\max_{i}\frac{|\hat{p}_i-p_{i0}|}{p_{i0}}. \tag{A.18}$$

But $\hat{p}_i \rightarrow p_{i0}$ in probability under the null hypothesis, so the last term in (A.17) is negligible compared to the first, and the two statistics λ and χ^2 are approximately the same.

Example (Fairness of dice) One of the most extensive dice experiments, performed by an English mathematician named Weldon and his wife, involved 315,672 throws of dice. These were rolled, twelve at a time, down a fixed slope of cardboard. There were 106,602 instances of the outcome 5 or 6. If the dice were true the probability of 5 or 6 should be 1/3. The expected number

[1] Pearson, Karl (1857–1936), English biometrician. A disciple of Francis Galton. Put biological statistics on a mathematical basis.

of such outcomes would be 315,672/3=105,224. Pearson's χ^2-statistic therefore is

$$\frac{(106{,}602-105{,}224)^2}{105{,}224}+\frac{(209{,}070-210{,}448)^2}{210{,}448}=27.1 \qquad (A.19)$$

which is highly significant on one degree of freedom. The likelihood ratio statistic is 27.0, virtually identical. □

A.4. Sufficiency

A statistic $T(X_1,\ldots,X_n)$ is called **sufficient** for θ if the conditional density (or probability function) of the data, given the value of $T(X_1,\ldots,X_n)$, does not depend on θ. In other words, if you tell me the value of $T(X_1,\ldots,X_n)$, I don't need to know anything else about the sample in order to infer something about the value of θ.

Example (Binomial case) Let X_1,X_2,X_3 be iid Bin(1, θ), and let $T(\mathbf{X})=(X_1+2X_2+3X_3)/6$. Do we need to know more about the sample than just the value of T in order to make a good guess as to the value of θ? To check that, suppose that the sample is $(X_1,X_2,X_3)=(1,1,0)$, so that the observed value of T is ½. Then

$$\begin{aligned}\mathrm{P}(\mathbf{X}=(1,1,0)\mid T(\mathbf{X})=\tfrac12) &= \frac{\mathrm{P}(\mathbf{X}=(1,1,0)\cap T(\mathbf{X})=\tfrac12)}{\mathrm{P}(T(\mathbf{X})=\tfrac12)}\\ &= \frac{\mathrm{P}(\mathbf{X}=(1,1,0)}{\mathrm{P}(T(\mathbf{X})=\tfrac12)}=\frac{\theta^2(1-\theta)}{\mathrm{P}(\mathbf{X}=(1,1,0)\cup\mathbf{X}=(0,0,1))}\\ &= \frac{\theta^2(1-\theta)}{\theta^2(1-\theta)+\theta(1-\theta)^2}=\theta\end{aligned} \qquad (A.20)$$

Since this probability depends on θ, we need more information about the sample than just the fact that $T(\mathbf{x})=\frac12$. When $T=\frac12$ there are two possible explanations, either $\mathbf{X}=(1,1,0)$, which would be likely if θ were large, or $\mathbf{X}=(0,0,1)$, which would be likely if θ were small. Thus, the information that $T=\frac12$ does not tell is much about the actual value of θ. □

The method used in this example is fine when we want to show that a statistic is not sufficient, but it is not all that helpful in trying to figure out reasonable candidates for sufficient statistics. A criterion for doing this is the following, due to Fisher[1] and Neyman.

[1] Fisher, Ronald Ayles (1890–1962). English statistician and geneticist. Introduced the likelihood approach to statistical inference. Invented the analysis of variance and the randomization approach to experimental design.

Proposition A.1 **(Fisher–Neyman factorization criterion)** A statistic $T(\mathbf{X})$ is sufficient if and only if the density (or probability function) can be factored as

$$f(\mathbf{x};\theta) = g(T(\mathbf{x});\theta)\,h(\mathbf{x}) \tag{A.21}$$

where g only depends on the x_i's through $T(\mathbf{x})$, while h does not depend on θ.

Example **(Binomial case, continued)** The density of $\mathbf{X}$ is

$$\begin{aligned}\mathbf{P}(\mathbf{X}{=}\mathbf{x}) &= \theta^{x_1}(1-\theta)^{1-x_1}\theta^{x_2}(1-\theta)^{1-x_2}\theta^{x_3}(1-\theta)^{1-x_3} \\ &= \left[\frac{\theta}{1-\theta}\right]^{\sum_{i=1}^{3}x_i}(1-\theta)^3.\end{aligned} \tag{A.22}$$

Using for g the entire expression on the right-hand side, and letting $h(\mathbf{x}){=}1$, we see that g depends on the data $\mathbf{x}$ only through their sum $\sum x_i$, which therefore is a sufficient statistic. To check back with the definition, notice that $\sum_{i=1}^{3} X_i \sim \text{Bin}(3,\theta)$, so that

$$\mathbf{P}(\mathbf{X}{=}\mathbf{x}\,|\,\sum X_i{=}\sum x_i) = \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\left[\begin{matrix}3\\ \sum x_i\end{matrix}\right]\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}} = \frac{1}{\left[\begin{matrix}3\\ \sum x_i\end{matrix}\right]}. \tag{A.23}$$

Since the right-hand side is independent of θ, regardless of the value of the x's, we see that $\sum X_i$ is indeed a sufficient statistic. Notice that, in fact, the conditional distribution is uniform over the set of possible outcomes with the given value of the sufficient statistic. □

Example **(The normal case)** Suppose now that $X_1,\ldots,X_n$ are iid $\text{N}(m,\sigma^2)$. First assume that σ^2 is a known number, so that we are only interested in estimating m. Then

$$\begin{aligned}f(\mathbf{x};m) &= (\sigma\sqrt{2\pi})^{-n}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2) \\ &= \exp(\frac{m\sum x_i}{\sigma^2}-\frac{m^2}{2\sigma^2})\exp(-\frac{\sum x_i^2}{2\sigma^2}+n\log(\sigma\sqrt{2\pi})).\end{aligned} \tag{A.24}$$

The first exponential function is g, the second is h. We see that $\sum X_i$ is a sufficient statistic.

Now assume instead that m is known, and σ is the parameter of interest. Then we write the density

$$f(\mathbf{x};\sigma^2) = (\sigma\sqrt{2\pi})^{-n}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2). \tag{A.25}$$

In this case $\sum(X_i-m)^2$ is a sufficient statistic. We have $h(\mathbf{x})=1$.

Finally, if both parameters are unknown, we write the density

$$f(\mathbf{x};m,\sigma^2) = \exp(\frac{m\sum x_i}{\sigma^2}-\frac{\sum x_i^2}{2\sigma^2}-\frac{m^2}{2\sigma^2}+n\log(\sigma\sqrt{2\pi})), \tag{A.26}$$

from which it follows that $(\sum X_i,\sum X_i^2)$ together are sufficient for (m,σ^2). □