The Statistician (2002)
51, *Part* 2, *pp.* 259–265

Validity of fitting a first-order Markov chain model to data

M. H. Eggar

University of Edinburgh, UK

[Received May 2001. Revised November 2001]

Summary. The paper describes a way of testing whether, given a sequence of data, there is a Markov chain of first order that is a likely model to fit it. The test, which was suggested by geometry, is applied to a sequence of moderate length of sporting data, for which a Markov model had once been claimed.

Keywords: Eulerian paths; Markov chain model

1. Introduction

Suppose that a system can be in any one of k states $S_1, \ldots, S_k$. We observe the state $S(t)$ of the system at discrete times $t = 1, 2, \ldots, N$. In the Markov chain model (MCM) of order 1 the hypothesis is made that there are constants p_{ij} for $1 \leqslant i \leqslant k$ and $1 \leqslant j \leqslant k$, independent of t, such that if state S_i pertains at time t then state S_j pertains at time $t + 1$ with probability p_{ij}. A natural question to ask is problem 1:

> is an observed sequence $S(1), \ldots, S(N)$ compatible with the assumption of an MCM of order 1 for a suitable choice of the constants p_{ij}?

The case $k = 2$ is already interesting and models the situation where two people play a game and for each (i, j) the probability p_{ij} that player j wins a point if player i has won the previous point is assumed constant. Here $p_{11} + p_{12} = 1$ and $p_{21} + p_{22} = 1$, so the assumption reduces to the constancy of p_{11} and p_{21}. The example that motivated this paper has $k = 3$ and consists of the outcomes of the cricket test-matches between England and Australia before 1990. The data, first investigated by Colwell *et al.* (1991), are reproduced in Section 5 and consist of a sequence of length 269 in the three letters E, A and D, which represent respectively a win by England, a win by Australia and a draw.

The study of testing the fit of a Markov chain to data started around 1950 with, among others, Bartlett (1951) and was continued by Hoel (1954). For more recent accounts, contributions or applications see, for example, Bishop *et al.* (1975) and Avery and Henderson (1999). The literature in between can readily be tracked down from these references. Attention has been directed primarily towards determining whether an MCM of order s is a preferable model to an MCM of order $s - 1$, but, as Hoel (1954) (page 430) pointed out, the adequacy of the MCM assumption itself is not tested. There is also the shortcoming that the tests can be applied only for very small values of s, unless the sequence of data is very long. This is because the number k^s

Address for correspondence: M. H. Eggar, Department of Mathematics and Statistics, University of Edinburgh, James Clerk Maxwell Building, King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, UK.
E-mail: m.eggar@ed.ac.uk

 0039–0526/02/51259

of possible basic s-tuples grows rapidly with s, and so some will soon appear in data of moderate size with too small a frequency.

In this paper we approach the question of the validity of the fit of an MCM of order 1 in a different way. Our key step, which was suggested by geometry, is to convert problem 1 into (k cases of) a problem which is not concerned with Markov chains, namely problem 2:

> a set of r coloured balls has r_j balls of colour j ($1 \leqslant j \leqslant k$), where $r = r_1 + \ldots + r_k$; the balls are placed in a row by selecting one ball at a time (without replacement) from the set; test whether the resulting sequence of colours is compatible with the null hypothesis that each ball was selected at random from those balls that had not already been placed in the row.

(For example, when $k = 3$ and for r_1, r_2 and r_3 sufficiently large it would be unlikely that the row would consist of the r_1 balls of the first colour, followed by the r_2 balls of the second colour, followed by the r_3 balls of the third colour, if selected at random.)

In Section 2 we investigate precisely what information is used in estimating the parameters p_{ij} for an attempted fit of an MCM. The conversion of problem 1 into problem 2 is stated in the fifth paragraph of Section 3, and justified earlier in Section 3, using the notation set up in Section 2. Problem 2 is discussed in Section 4 first in general terms and then for the cricket data.

2. The transition matrix

For an observed sequence $S(1), \ldots, S(N)$ let n_{ij} denote the number of values of t such that $S(t) = S_i$ and $S(t+1) = S_j$. Denote the ith row sum $\Sigma_j n_{ij}$ of the $k \times k$ matrix (n_{ij}) by R_i and the jth column sum $\Sigma_i n_{ij}$ by C_j. Now define $\widetilde{p_{ij}} = n_{ij}/R_i$. If the observed sequence $S(1), \ldots, S(N)$ does come from an MCM of order 1, then $\widetilde{p_{ij}}$ are the maximum likelihood estimates for p_{ij} obtainable from the sequence (see Bishop *et al.* (1975), page 263), and by Markov chain theory $\widetilde{p_{ij}} \to p_{ij}$ as $N \to \infty$ with probability 1.

We observe that R_i counts the number of values of t such that $S(t) = S_i$ and $1 \leqslant t \leqslant N-1$, whereas C_i counts the number of values of t such that $S(t) = S_i$ and $2 \leqslant t \leqslant N$. Hence if $S(1) = S(N)$ we have

(a) $R_i = C_i$ for all i.

However, if $S(1) = S_a$ and $S(N) = S_b$, where $a \neq b$, then we have

(b) $C_a = R_a - 1$, $R_b = C_b - 1$ and $R_i = C_i$ for all $i \neq a, b$.

If every state appears in the sequence $S(1), \ldots, S(N)$ then

(c) it is not possible to partition the set $\{1, 2, \ldots, k\}$ into two non-empty disjoint subsets I and J such that $n_{ij} = n_{ji} = 0$ for all $i \in I$ and $j \in J$.

We now show that the information carried by the matrix (n_{ij}) is essentially equivalent to the information carried by the matrix $(\widetilde{p_{ij}})$ together with N. Obviously $\widetilde{p_{ij}}$ ($= n_{ij}/R_i$) and N ($= 1 + \Sigma n_{ij}$) are completely determined by the n_{ij}. To see the converse, first suppose that result (a) holds. It is an algebraic consequence of result (a) that the vector $(R_1, \ldots, R_k)^{\mathrm{T}}$ is an eigenvector with eigenvalue 1 of the matrix $(\widetilde{p_{ij}})^{\mathrm{T}}$. Consequently in the generic case when this eigenspace has dimension 1 $(R_1, \ldots, R_k)^{\mathrm{T}}$ is uniquely specified, given $(\widetilde{p_{ij}})$ and N, by this eigenvector property and the requirement that $\Sigma R_i = N - 1$. Hence n_{ij} ($= R_i \widetilde{p_{ij}}$) can be found. In the second scenario, when result (b) holds rather than (a), the matrix $(\widetilde{p_{ij}})^{\mathrm{T}}$ still has 1 as an eigenvalue, since $(1, 1, \ldots, 1)^{\mathrm{T}}$ is an eigenvector with eigenvalue 1 of its transpose $(\widetilde{p_{ij}})$. Let $(R'_1, \ldots, R'_k)^{\mathrm{T}}$ be the eigenvector, assumed unique, of $(\widetilde{p_{ij}})^{\mathrm{T}}$ with eigenvalue 1 and such that

$\Sigma R_i' = N - 1$. By result (b) $R_i' \doteqdot R_i$, if N is sufficiently large, and hence, as in the first scenario, each n_{ij} can be recovered (approximately) from the values of $(\widetilde{p_{ij}})$ and N.

The significance of the previous paragraph is that it enables us to keep the following exact account of what information from the observed sequence is used in estimating the parameters of the MCM that would best fit the data. Information is discarded in passing from the observed sequence to the matrix (n_{ij}), but this matrix is equivalent to $(\widetilde{p_{ij}})$ and N.

3. A reformulation of problem 1

In Section 2 we obtained the matrix (n_{ij}) from the observed sequence. We now demonstrate that given any $k \times k$ matrix $A = (a_{ij})$ of non-negative integers such that the entries satisfy result (c) and the row and column sums R_i and C_i satisfy either (a) or (b), then there is a sequence $S(1), \ldots, S(N)$, where $N = 1 + \Sigma a_{ij}$, such that the matrix (n_{ij}) for this sequence is the matrix A. To see this consider the digraph with k vertices and a_{ij} edges from vertex i to vertex j for each ordered pair (i, j). Condition (c) says that this digraph is connected and our claim is a restatement of the Eulerian path theorem for digraphs (see Wilson and Watkins (1990), theorem 6.3 on page 133, and recall that an Eulerian path is a path that traverses each edge exactly once).

This approach suggests that it is helpful (though not essential!) to think about an MCM geometrically. We think of each state $S_1, \ldots, S_k$ as a vertex and station an imp I_i at the vertex S_i for each i. The imp I_i has a sack with infinitely many balls, of which the proportion p_{ij} are of colour j for $1 \leqslant j \leqslant k$. We think of an observed sequence $S(1), \ldots, S(N)$ as a directed path with $N - 1$ edges, each of which connects two of the k vertices. The mechanism by which such a path is constructed is that to continue a path $S(1), \ldots, S(t)$ the imp at vertex $S(t)$ randomly chooses a ball from his sack to determine by its colour the next vertex on the path. In problem 1 we are seeking to analyse whether a given sequence $S(1), \ldots, S(N)$ observed is a likely outcome for some MCM.

We now obtain a reformulation of problem 1. In Section 2 we used the observed sequence to estimate the matrix (p_{ij}) and we showed that the information being used was essentially equivalent to the matrix (n_{ij}). The observed sequence will generally be one of many possible sequences that yield the same matrix (n_{ij}). We regard the n_{ij} as fixed and wish to know how likely it is for our observed sequence to occur, if an MCM of order 1 with conditional probabilities $(\widetilde{p_{ij}})$ operates. For example, consider the cricket sequence in Section 5. Construct a new sequence by replacing the first entry E by 40 consecutive Es, the first entry A by 45 consecutive As, the first entry D by 35 consecutive Ds and any later run of one of the letters by a single occurrence of that letter. The original and the new sequence yield the same matrix (n_{ij}), and hence the same $(\widetilde{p_{ij}})$, but it is intuitively clear that the original sequence is a more likely outcome under an MCM of order 1 than the new sequence is.

Thus the information being drawn from the observed sequence to estimate the parameters of the best fit MCM is equivalent to replacing the sack of imp I_i by a sack containing just n_{ij} balls of colour j, where $1 \leqslant j \leqslant k$. The set of all paths of length N that yield the same matrix (n_{ij}) is built up by the imps by choices (without replacement) from their sacks. It is important to note that any set of choices by the imps gives a sequence yielding the matrix (n_{ij}) and the choices made by the different imps are independent. Thus problem 1 is reduced to checking whether each imp is making his choices randomly.

The conversion of problem 1 to problem 2 can be summarized as follows: for each state S_i peruse the data sequence $S(1), \ldots, S(N)$ and write out the subsequence of the states that are immediately preceded by an occurrence of S_i. The sequence contains R_i entries, of which n_{ij} are the state $S_j, 1 \leqslant j \leqslant k$. If an MCM of order 1 underlies the original sequence then the

subsequence should be a random permutation. Testing for the randomness is a case of problem 2 with $r = R_i$ and $r_j = n_{ij}$. By considering $i = 1, \ldots, k$ we obtain k cases of problem 2, which are independent for the reason explained in the previous paragraph.

For example, for the cricket data in Section 5, take $S_1 = E$, $S_2 = A$ and $S_3 = D$. For $i = 1$ we obtain sequence (i):

ADEAD EAEEE EEEAE EEADD EAEAE AAEEA ADDAA EEEDA AEEEE AAAEE ADEAA DAEED EDAAA DDEAA AEEDE AEEAE EDDAE EDA.

Here $r_1 = 39$, $r_2 = 32$ and $r_3 = 17$. For $i = 2$ we obtain sequence (ii):

EEADA EDAEE AEAEE AAADD AAADA EEEEA AEADE AAAAA AADAA EDEDE EEEAA DEADA ADAAA AAEEE ADADE DDDDD DAADA EDDEA ADDAE DDADA D.

Here $r_1 = 28$, $r_2 = 44$ and $r_3 = 29$. For $i = 3$ we obtain sequence (iii):

AAEEE EADDE DADEA DADED ADDDE DADAD ADAAD DDEDA AAADE DDDAD DDEAD DEDED EEADD AEEAE DAEDE DEAA.

Here $r_1 = 21$, $r_2 = 24$ and $r_3 = 34$.

Problem 1 is now reduced to k independent cases of problem 2. Choose a test T for problem 2. Suppose that test T would reject the null hypothesis of randomness in the ith case with probability c_i, where $1 \leqslant i \leqslant k$. To combine the results of the k independent tests into a single criterion, we may regard c_i as a value of a variate uniformly distributed on the interval [0, 1]. By the multiplicative rule for probabilities of independent events, the probability of observing results that are less likely than those observed is

$$b = 1 - \int \ldots \int \mathrm{d}x_1 \ldots \mathrm{d}x_k,$$

where the integral is evaluated over the region determined by the inequalities $x_1 \ldots x_n \geqslant c_1 \ldots c_n$, $0 \leqslant x_i \leqslant 1$ for each i. For example, when $k = 3$, we have

$$b = c\{1 - \log(c) + \tfrac{1}{2}\log(c)^2\},$$

where $c = c_1c_2c_3$. Test T would then lead us to reject at the 5% level of significance the null hypothesis that an MCM fitted the observed sequence $S(1), \ldots, S(N)$ if $b < 0.05$.

The situation of problem 2 can be simulated and such a simulation run a large number of times to estimate the probability distribution for the value of any statistic that test T investigates. Of course the situation of problem 1 could also be simulated, but the resulting sequences tend to show a large amount of variability, so there is a real gain in passing from problem 1 to problem 2.

4. Investigation of problem 2

There is no definitive procedure for testing the null hypothesis of randomness. Indeed, to quote Kendall and Stuart (1976) (page 365), 'there is no limit to the number of tests which can be set up for this purpose'. The choice of test T will depend on the alternative hypothesis, but an element of arbitrariness in the choice of test T is still unavoidable.

Table 1. Observed frequencies for sequence (i)

Result	*Frequencies for the following thirds:*			*Total*
	First	*Middle*	*Last*	
E	16	12	11	39
A	9	13	10	32
D	4	5	8	17
Total	29	30	29	88

We have arrived at problem 2 from problem 1, where we were testing for the constancy of the p_{ij} over time. One aspect, that test T could be chosen to test, of randomness for the corresponding sequences of problem 2 would be whether the relative frequency of each state in these sequences changes significantly along the sequence. For the cricket data a suitable test T would be to count the number of each of E, A and D appearing in the first, middle and last thirds of the sequences. Here then the alternative hypothesis is that the proportions change between the first, middle and final thirds of the sequence. For sequence (i) Table 1 is the table of observed frequencies.

If the sequence is random, the expected frequencies are $29 \times 39/88$ etc. We could now calculate the value of $\Sigma(f - e)^2/e$ for Table 1, where f is the observed frequency and e is the expected frequency of each entry, and the sum is taken over all the entries in Table 1. For Table 1 the value is 3.41. We could simulate the situation of problem 2 with $r_1 = 39$, $r_2 = 32$ and $r_3 = 17$ sufficiently many (say 1000) times to estimate the probability that a value exceeding 3.41 is obtained. As an approximation we compare 3.41 with values of a χ^2_4-variate and find that a value exceeding 3.41 for such a variate occurs with probability 0.5. Similarly for sequences (ii) and (iii) values of 13.85 and 7.83 are obtained and the corresponding probabilities are 0.008 and 0.1. By the penultimate paragraph of Section 3 the combination of the three independent tests yields a probability $b = 0.016$, and so at the 5% level of significance we would reject the null hypothesis that all the sequences (i), (ii) and (iii) are random.

We remark that in the above test the choice made in deciding to divide the sequences into three equal parts (and so the choice of the alternative hypothesis) is somewhat arbitrary. It is a suitable choice in view of the length of the sequences and numbers involved, but there is no reason to have the same number of divisions as there are states. With the cricket data we might, instead, have posed as the alternative hypothesis that the proportions were different before and after World War 2 and divided the data thus into two pieces. Indeed we mention in passing that there are tests (see Pettitt (1979)) to locate a binary changepoint, if we suspected one but did not know where it was.

There are respects in which an MCM might not be an applicable model, which our test T cannot possibly detect. For example in the case of the cricket data there might be a dependence on other factors, such as the venue, or there might be a tendency for a team to slacken its effort after an ashes series has been won. Sequences which were periodic (i.e. consisted of a short string of states repeated several times, and hence non-random) would certainly not be detected by such a test as T. The literature contains various tests for randomness, which we might apply according to our alternative hypothesis. One classical test uses autovariance coefficients with various lag times (see Chatfield (1996), pages 18–25).

5. The cricket data

The sequence of wins and draws in the first 269 cricket test-matches between England and Australia is as follows, where E denotes a win by England, A denotes a win by Australia and D denotes a draw:

AEAED	AADAA	EEADE	DEEAA	EEEEE	EEAEE	EEAAE	DEDEE
AAEEA	EEAAA	ADADD	DEAAA	ADDAA	EEEAE	AEDDE	DAEAA
AEAAD	DAEEE	EDDEA	AAAAA	AADDA	AAEAD	DDDEE	EEEAE
ADDAE	AEEEA	EDDAE	EAAAD	DAEAA	DDAAA	DAAAA	AAEDD
DDEAE	EEDDA	EEDAA	DAADA	EADDE	ADDDD	ADDDD	EADAD
DDEDD	EDDEE	ADEAA	ADAAE	ADDDA	DEEED	EEAEE	EAAAD
ADEEE	DDAAE	DEADD	EEEDD	EADAA	DAAD		

These data were first investigated by Colwell *et al.* (1991). They came to the conclusion that there was a remarkably good fit to an MCM of order 1, but there was a fallacy in the reasoning, as identified by Bendall and Eggar (2001), that made such a conclusion inevitable. Indeed there are intuitive reasons to doubt the applicability of an MCM. For example, we would expect that the result 'draw' would often depend not just on the result of the previous match but more on the weather (although some further factor, such as the country of the venue, might influence both). Also one could argue that the latter part of the cricket data is more likely to contain draws than the early part owing to changes in the rules of test-matches. Bendall and Eggar (2001) tested their doubts in a crude way concentrating on draws and came to the same conclusion as in the present paper, that there is good reason to doubt the applicability of an MCM of order 1. Yet another approach to problem 1 would be to test the distribution of the lengths of runs and gaps, which for an MCM can easily be derived, but the present data bear out the common experience that tests based on runs or gaps tend to be weak.

6. Conclusion

The more parameters there are in a probability model, the more likely it generally is that we can choose them so that the model fits given data. We might therefore expect difficulty in showing that a sequence of data of moderate length was not compatible with a first-order MCM with k states for some choice of transitional probabilities. We have demonstrated that by converting such a problem (problem 1) into k independent cases of a simpler problem (problem 2) it may be possible to show that a Markov model is unlikely to be applicable to the data. The main thrust of this paper has been to elucidate and justify the conversion step. We do, however, encounter the difficulty that there is no definitive test to resolve problem 2, though there is no lack of possible tests in the literature depending on our alternative hypothesis.

Acknowledgements

The author is grateful to the referees and colleagues for their helpful comments.

References

Avery, P. J. and Henderson, D. A. (1999) Fitting Markov chain models to discrete state series such as DNA sequences. *Appl. Statist.*, **48**, 53–61.
Bartlett, M. S. (1951) The frequency goodness of fit test for probability chains. *Proc. Camb. Phil. Soc.*, **47**, 86–95.
Bendall, S. and Eggar, M. H. (2001) Markov chains in cricket revisited. *Math. Gaz.*, **85**, 101–103.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice.* Cambridge: Massachusetts Institute of Technology Press.
Chatfield, C. (1996) *The Analysis of Time Series: an Introduction*, 5th edn. London: Chapman and Hall.
Colwell, D., Jones, B. and Gillett, J. (1991) A Markov chain in cricket. *Math. Gaz.*, **75**, 183–185.
Hoel, P. J. (1954) A test for Markov chains. *Biometrika*, **41**, 430–433.
Kendall, M. and Stuart, A. (1976) *The Advanced Theory of Statistics*, vol. 3, *Design and Analysis, and Time Series.* London: Griffin.
Pettitt, A. N. (1979) A non-parametric approach to the change-point problem. *Appl. Statist.*, **28**, 126–135.
Wilson, R. J. and Watkins, J. J. (1990) *Graphs: an Introductory Approach.* New York: Wiley.