EMPIRICAL BAYES ESTIMATION FOR
MARKOV CHAINS

L. Billard†
Florida State University

and

M. R. Meshkani
College of Statistics and Informatics, Iran

†Research supported by National Institutes of Health Grant No. 1R01 GM 26851-01.
Empirical Bayes Estimation for Markov Chains

L. Billard†
Florida State University

and

M. R. Meshkani
College of Statistics and Informatics, Iran

ABSTRACT

Theoretical results for Markov chains with known transition probability matrix are extensive and well known. However, less attention has been paid to the problem of making inferences about the probabilistic structure of a chain of observations. This paper reviews briefly some recent work relating to the estimation of the transition probabilities using empirical Bayes techniques. Specifically it considers empirical Bayes estimation via the method of moments for finite state stationary chains for both a panel study and a single study situation, and for a finite state nonstationary chain for a panel study situation.

Keywords: Empirical Bayes estimation, transition probability matrix, matrix beta prior, finite state chain, stationary, nonstationary.

†Research supported by National Institutes of Health Grant No. 1 R01 GM 26851-01.
1. INTRODUCTION

Theoretical results for Markov chains with known transition probability matrix are extensive and well known. However, less attention has been paid to the problem of making inferences about the probabilistic structure of a chain of observations. This paper reviews briefly some recent work relating to the estimation of the transition probabilities using empirical Bayes techniques. Typical approaches used in the past involve the use of Bayesian, maximum likelihood and least squares procedures. See, for example, Anderson and Goodman (1957), Billingsley (1961a, 1961b), Lee et al. (1968) and Madansky (1959). The simplest type of Markov chain, namely the finite state first order stationary chain, as well as a special case of a nonstationary chain will be considered. Chains of higher order can be converted to chains of first order by a standard technique, so need not be considered separately.

An empirical Bayes approach to statistical decision problems is applicable when we repeatedly and independently encounter the same decision problem in a sequence with a fixed but unknown prior distribution for the parameters. We do not expect all decision problems in practice to be embedded in such a sequence. However, when they are, the empirical Bayes approach offers certain advantages over any approach which ignores the fact that the parameters are themselves random variables. This approach also has advantages over approaches which assume a personal prior not changing with experience [Robbins (1964)].

Stationary Markov chains are considered in Section 2 and a nonstationary Markov chain is discussed in Section 3.
2. STATIONARY MARKOV CHAINS

Suppose \( \{X_t, t \in T_0\} \) is a Markov chain with values in the finite state space \( S = \{1, \ldots, s\} \) where \( T_0 = \{0\} \cup T \) with \( T = \{1, \ldots, T\} \). Here \( T > 1 \) refers to a fixed period of time during which we observe and record the states visited by the chain. The transition probability matrix of this chain is denoted by \( \Lambda = (\Lambda_{jk}), j, k \in S \). That is, \( \Lambda_{jk} \) is the probability an individual moves from state \( j \) to state \( k \), \( j, k \in S \), at some time \( t \in T \).

Let \( N = \{1, \ldots, n\} \) and \( N_1 = \{1, \ldots, n+1\} \). It is assumed that for each \( i \in N_1 \), there are \( N \) independent and identically distributed chains observed collectively. The data \( N \) is referred to as the "past" data though in practice all sets in \( N_1 \) may in fact be observed concurrently. Preston (1971) briefly considered this case for \( s = 2 \). When \( N > 1 \), we have what is called a panel study situation. In this case, the first step in deriving the empirical Bayes estimators for \( \Lambda_{jk}, j, k \in S \), is based on the conditional distribution results of Anderson and Goodman (1957). (By "conditional" distribution we mean the distributions for given \( \Lambda \).) However, when \( N = 1 \), the Anderson and Goodman results do not hold good. Instead, we use corresponding conditional distribution results given by Whittle (1955). This latter case will be referred to as a single study situation.

An example of a panel study is a brand switching problem of the type given by Draper and Nolin (1964). Suppose there are \( s \) brands of a certain product on the market. Customers switch from one brand to another brand as time progresses. There are many quantities of interest to a producer and/or retailer such as the percentage of customers using a particular brand in the long run. The data consists of recording the brands bought over a period of time \( T \) by each of \( N \) customers at each of \( (n+1) \) different locations (stores).
When just one customer at each store is studied for brand preference, we have a single study situation. Another frequently occurring example of a single study relates to rainfall patterns for a given location. The states of the chain refer to differing amounts of precipitation. In the simplest case, $s = 2$ when states are either wet (rain) or dry (no rain) [see Gabriel and Neuman (1962)]. Observations are recorded for $T$ periods of time in each of $(n + 1)$ years.

Before considering each of these two situations in turn, let us make the following definitions.

**Definition 1**

The frequency count vector, $G(t)$, is the vector whose $j$th element $G_j(t)$, $j \in S$, is the number of individuals in the $j$th state at time $t$, $t \in T_0$.

It is noted that the initial frequency count vector $G(0)$ may be either fixed or random. We consider the case $G(0)$ fixed here. The case $G(0)$ random follows analogously [see Billard and Meshkani (1979b)].

**Definition 2**

The frequency count matrix, $F$, is the matrix whose $jk$th element $F_{jk}$, $j, k \in S$, is the number of times the state $k$ has followed the state $j$ in a sequence of states of length $T \geq 1$.

It is observed that there is a $G(t)$ and an $F$ for each $i \in N_1$, that is, $G_{-i}(t)$ and $F_{-i}$. However, we shall suppress the $i$ subscript unless it is necessary for clarification. We note that an underlying assumption of the empirical Bayes approach used implies that the $F_{-i}$ are independently and identically distributed.
Suppose the transition probability matrix $\Lambda$ has as prior distribution its natural conjugate prior, viz., the matrix beta prior distribution with parameter $\rho = (\rho_{jk}), j, k \in S$. Then, for squared error loss, it is well known that the Bayes estimate of $\Lambda$ is

$$\Lambda_{B} = (\Lambda_{B}; jk)$$

where

$$\Lambda_{B;jk} = (F_{jk} + \rho_{jk})/(F_{j+} + \rho_{j+}), j, k \in S,$$

with

$$F_{j+} = \sum_{k \in S} F_{jk} \text{ and } \rho_{j+} = \sum_{k \in S} \rho_{jk}.$$
with
\[ r_{jk} = \text{Tr}_{jk} \overline{M}_{jk} / (T + 1), \quad j, k \in S, \quad k \neq s, \]
\[ r_{js} = r_{j+} (\overline{\Sigma}_{js} + 1) / (T + 1), \quad j \in S, \]
and
\[ r_{j+} = \{[ (T + 1) / T ]^2 - c_j \} / (c_j - \overline{c}_j), \quad j \in S, \]
where
\begin{align*}
(3) \quad \overline{M}_{jk} &= n^{-1} \sum_{i \in N} M_{i;jk}, \\
(4) \quad M_{i;jk} &= F_{i;jk} / F_{i;j+}, \\
(5) \quad \overline{c}_j &= n^{-1} \sum_{i \in N} F_{i;j+}^{-1} \\
\]
\[ c_j = \{ | \hat{\Sigma}_{jj}^* | / | \hat{\Sigma}_{jj} | \}^{1/(s-1)} \]
where the matrices \( \hat{\Sigma}_{jj}^* \) and \( \hat{\Sigma}_{jj} \) have elements \( \hat{\sigma}_{jk;jh}^* \) and \( \hat{\sigma}_{jk;jh} \), \( j, h, k \in S \), respectively, given by
\[ \hat{\sigma}_{jk;jh}^* = (n - 1)^{-1} \sum_{i \in N} (M_{i;jk} - \overline{M}_{jk})(M_{i;jh} - \overline{M}_{jh}) \]
and
\[ \hat{\sigma}_{jk;jh} = \overline{M}_{jk}(\delta_{kh} - \overline{M}_{jh}) \]
with \( \delta_{kh} = 1 \) if \( k = h \) and \( 0 \) if \( k \neq h \).

Panel Study Data

For panel study situations, \( N > 1 \), we utilise the conditional distribution of \( F \) given \( \Lambda \) results given by Anderson and Goodman (1957). Thus, Billard and Meshkani (1979b) have shown that the parameters \( \rho \) satisfy...
\( \rho_{jk} = \rho_{j+} E(M_{jk} \cdot), \ j, \ k \in S, \)

\( E(G_j^x) = \sum_{h \in S} G_h(0)E_{1i} \Lambda_{hj} + \sum_{k \in S} \Lambda_{hk} \Lambda_{kj} + \ldots \)

\[ + \sum_{l_2 \in S} \ldots \sum_{l_1 \in S} \Lambda_{l_2l_1} \ldots \Lambda_{l_2-3l_2-2} \ldots \Lambda_{l_2 T-2} \Lambda_{l_2 T-1} \ldots \Lambda_{l_2 T-2} \Lambda_{l_2 T-1} \Lambda_{l_2 T-1}, \ j \in S, \]

where

\( M_{jk} = F_{jk}/F_{j+} \) and \( G_j^x = \sum_{k \in S} F_{kj}, \ j, \ k \in S, \)

and where \( E_1 \) denotes expectation with respect to the distribution of \( \Lambda. \) The equation (7) may be rewritten as

\[ \sum_{h \in S} G_h(0)(\rho_{hj}/\rho_{h+}) + \sum_{h \in S} G_h(0)\rho_{hh}(\rho_{hj} + S_{hj})/[(\rho_{h+}(\rho_{h+} + 1)] \]

\[ + \sum_{h \in S} G_h(0)[\sum_{k \in S} \rho_{hk} \rho_{kj}/(\rho_{h+}\rho_{k+})] - E(G_j^x) + \phi(\rho) = 0 \]

where \( \phi(\rho) \) equals the terms remaining in (7) for \( t > 3. \) Theoretically, it is possible to evaluate \( \phi(\rho) \) for every \( T. \) However, for large \( T, \) it does not render a neat closed form answer.

The equation (6) and (9) are then solved for the \( \rho. \) The empirical Bayes estimate for \( \rho \) is then found by replacing theoretical components in those solutions by their sample counterparts. Thus, when \( T = 2, \) we find that the empirical Bayes estimate for \( \Lambda \) by the method of moments is

\[ \Lambda_{EB} = (\Lambda_{EB;jk}) \]

where

\[ \Lambda_{EB;jk} = (F_{jk} + r_{jk})/(F_{j+} + r_{j+}), \ j, \ k \in S, \]
with
\[ r_{jk} = M_{jk} x_j / (1 - x_j), \]
\[ r_{j*} = x_j / (1 - x_j), \]

where \( M_{jk} \) is as given in (3) and where \( x = (x_j), j \in S, \) satisfies

(10) \[ C x = b \]

with
\[ C = (c_{jk}) \text{ and } b = (b_j), j, k \in S, \]

where
\[ c_{jk} = G_h(0) (M_{hj} - \delta_{hj}), \]
\[ b_j = G^*_j - \sum_{h \in S} G_h(0) (M_{hj} + M_{hh} \delta_{hj} + \sum_{k \in S} M_{hk} M_{kj}), \]

and
\[ G^*_j = \frac{1}{n} \sum_{i \in \mathcal{N}} \left( \sum_{k \in S} F_{i;kj} \right). \]

When solving (10) for \( x \) we may use a generalised inverse of \( C \) [see, for example, Searle (1971)]. More complete details are provided in Billard and Meshkani (1979b).
3. A NON-STATIONARY MARKOV CHAIN

The previous section was concerned with stationary Markov chains. However, there are many situations, especially in the medical and social sciences, where the transition probability matrix is not constant over time. For example, consider a process regarding changes in people's behavior when a definite decision about an issue (such as voting on a proposal or electing a candidate) must be made by a due date. It is reasonable to expect that the transition probability matrix of the penultimate week might be different from those of earlier weeks. In another example, Howard (1971) has considered a learning process for a student who becomes discouraged with time. We note that the transition probability matrix is now a function of time, viz., \( \Lambda(t) \), \( t \in T \).

Definition 3

The frequency count matrix at time \( t \), \( F(t) \), is the matrix whose \( jk \)th element \( F_{jk}(t) \), \( j, k \in S \), is the number of individuals moving from state \( j \) at time \( (t - 1) \) to state \( k \) at time \( t \).

As in Section 2, there is an \( F(t) \) for each sequence \( i \in N_1 \), \( F_i(t) \). The \( F_i(t) \) are assumed to be independently and identically distributed. Except where necessary to avoid confusion the subscript \( i \) will not be used.

Let us assume a squared error loss function and that \( \Lambda(t) \) has as its prior distribution the matrix beta prior distribution with parameter \( \rho(t) = (\rho_{jk}(t)), j, k \in S \). Then, it is well known that the Bayes estimate of \( \Lambda(t) \) is given by

\[
\Lambda_B(t) = (\Lambda_B; jk(t))
\]

where
\[ \Lambda_{B;jk}(t) = \frac{F_{jk}(t) + \rho_{jk}(t)}{F_{j+}(t) + \rho_{j+}(t)}, \quad j, k \in S, \]

with

\[ F_{j+}(t) = \sum_{k \in S} F_{jk}(t) \text{ and } \rho_{j+}(t) = \sum_{k \in S} \rho_{jk}(t). \]

The conditional distribution of \( F(t) \) for a given \( \Lambda(t) \) is given by Anderson and Goodman (1957). As before, we remove the conditionality and hence we can show that an empirical Bayes estimate of \( \Lambda(t) \) is given by

\[ \Lambda_{EB}(t) = (\Lambda_{EB};jk(t)) \]

where

\[ \Lambda_{EB;jk}(t) = \frac{F_{jk}(t) + r_{jk}(t)}{F_{j+}(t) + r_{j+}(t)}, \quad j, k \in S, \]

with

\[ r_{jk}(t) = r_{j+}(t)\overline{M}_{jk}(t) \]

and

\[ r_{j+}(t) = \left( |D^*(t)|/|C^*(t)| \right)^{1/(s-1)} \]

where

\[ \overline{M}_{jk}(t) = n^{-1} \sum_{i \in N} \frac{F_{i;jk}(t)}{F_{i;j+}(t)} \]

and \( D^*(t) \) and \( C^*(t) \) are the matrices obtained from the first \( s - 1 \) rows and columns of the matrices \( D(t) \) and \( C(t) \) whose elements are, respectively,

\[ d_{j;kh}(t) = \delta_{kh} \overline{M}_{jk}(t) - \overline{Z}_{j;kh}(t) \]

and

\[ c_{j;kh}(t) = \overline{Z}_{j;kh}(t) - \overline{M}_{jk}(t) \cdot \overline{M}_{jh}(t) \]

where

\[ \overline{Z}_{j;kh} = n^{-1} \sum_{i \in N} \{ F_{i;jk}(t) \left[ F_{i;jh}(t) - \delta_{kh}/(F_{i;j+}(t) - 1) \right] \}, \quad j, k, h \in S. \]

Further details may be found in Billard and Meshkani (1979c).
REFERENCES


