PSYCHOMETRIKA—VOL. 24, NO. 2
JUNE, 1959

LEAST SQUARES ESTIMATION IN FINITE MARKOV PROCESSES

ALBERT MADANSKY

RAND CORPORATION

The usual least squares estimate of the transitional probability matrix of a finite Markov process is given for the case in which, for each point in time, only the proportions of the sample in each state are known. The purpose of this paper is to give another estimate of this matrix and to investigate the properties of this estimate. It is shown that this estimate is consistent and asymptotically more efficient than the previously considered estimate in a sense defined in this paper.

Notation

The matrix conventions and terminology used in [2, 4, and 6] will be followed. Let m_{ik} $(i = 1, 2, \cdots, a; k = 1, 2, \cdots, n)$ be the proportion observed on trial k in alternative category i of a multinomial population based on a sample of size S. Let $E(m_{ik}) = \mu_{ik}$, where $0 \leq \mu_{ik} \leq 1$ and $\sum_i \mu_{ik} = 1$ for each k. Let

$$M = \begin{bmatrix} m_{11} & \cdots & m_{1,n-1} \\ \vdots & & \vdots \\ m_{a1} & \cdots & m_{a,n-1} \end{bmatrix}.$$

Also, let t_{ij} be the transitional probability that an observation which is in alternative category j at a given trial be in alternative category i at the next trial $(i, j = 1, 2, \cdots, a)$. Define $T_i = [t_{i1} \cdots t_{ia}]$ and

$$T = \begin{bmatrix} t_{11} & \cdots & t_{1a} \\ \vdots & & \vdots \\ t_{a1} & \cdots & t_{aa} \end{bmatrix}.$$

Let $N_i = [m_{i2} \cdots m_{in}]$, and let N be the $a \times (n-1)$ matrix with N_i as row i. Also let

$$\mu = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1,n-1} \\ \vdots & & \vdots \\ \mu_{a1} & \cdots & \mu_{a,n-1} \end{bmatrix},$$

and $\nu_i = [\mu_{i2} \cdots \mu_{in}]$, and let ν be the $a \times (n-1)$ matrix with ν_i as row i.

Then $E(M) = \mu$, $E(N_i) = \nu_i$, and $E(N) = \nu$. Also, by the Markovian assumption, $\nu_i = T_i\mu$, so that $\nu = T\mu$ and $T = \nu\mu'(\mu\mu')^{-1}$ if $\mu\mu'$ is nonsingular.

One should note that since the Markov process has $a(a-1)$ parameters all these parameters cannot be estimated unless $n(a-1) \geq a(a-1)$, i.e., unless the number of independent m_{ik} is greater than or equal to the number of parameters to be estimated. It therefore should be assumed that $n \geq a$. It shall, however, be assumed that $\mu\mu'$ is nonsingular and hence, the necessary condition that $\mu\mu'$ be nonsingular, that $n - 1 \geq a$. Other assumptions will be introduced as they become necessary.

Estimation of T

Since M is not a matrix of constants, the usual proof in the theory of least squares (cf. [5], p. 55) does not apply in this case to show that

$$\bar{T}_i = N_iM'(MM')^{-1}$$

is the least squares estimate of T_i . One can work conditionally on M and find the $\bar{T}_i$ which minimizes

$$(T_iM - N_i)(T_iM - N_i)' = C_iC_i' ,$$

given M. In that case

$$\bar{T}_i = N_iM'(MM')^{-1}$$

if MM' is nonsingular, as is shown in [2], and the estimate of T, *conditional on* M, which minimizes C_iC_i' for each i is

$$\bar{T} = NM'(MM')^{-1}.$$

Even if M were a matrix of constants (or even if one works conditionally on M), $\bar{T}_i$ would still not be the "least squares" estimate of T_i . Although the elements of C_i have mean zero (this can be seen as a corollary of (2.14) of [1]), they are not uncorrelated and do not all have the same variance, σ^2.* In a situation such as this, the least squares estimate of T_i is obtained by transforming C_i to

$$\tilde{C}_i = C_i\Lambda_i^{-1/2},$$

whose elements are uncorrelated and have the same variance, $\sigma^2 = 1$, and by minimizing

$$\tilde{C}_i\tilde{C}_i' = C_i\Lambda_i^{-1}C_i' ,$$

*I am indebted to a referee for pointing out that equal variances are not required to use the method of least squares for estimation, but only necessary for the optimal properties of least squares estimates, in particular having minimum variance (as defined, e.g., in [3]) among linear unbiased estimates, to be realized. The referee also points out that cov $(m_{ik}, C_{ik}) = 0$, i.e., the error is uncorrelated with the independent variables in the relation $m_{i,k+1} = \sum_j t_{ij}\, m_{ik} - C_{ik}$. This can be seen as a direct application of Lemma 1.

where Λ_i is the covariance matrix of C_i . Let us first determine Λ_i and Λ_i^{-1}. Use is made of the following lemma.

LEMMA 1. Cov $(m_{in}, m_{r,n-k}) = \mu_{r,n-k}\,(t_{ir}^{(k)} - \mu_{in})/S$, where $t_{ir}^{(k)}$ is element (i, r) of T^k.

PROOF. Define

$$Y_{ijs} = \begin{cases} 1 \text{ if individual } s \text{ is in category } i \text{ at time } j, \\ 0 \text{ otherwise.} \end{cases}$$

Then

$$\begin{aligned} \operatorname{cov}(m_{in}, m_{r,n-k}) &= \sum_{s=1}^{S} \operatorname{cov}(Y_{ins}, Y_{r,n-k,s})/S^2 \\ &= [\Pr\{Y_{ins} = 1, Y_{r,n-k,s} = 1\} - \mu_{in}\mu_{r,n-k}]/S \\ &= [\Pr\{Y_{ins} = 1 \mid Y_{r,n-k,s} = 1\}\mu_{r,n-k} - \mu_{in}\mu_{r,n-k}]/S \\ &= (t_{ir}^{(k)}\mu_{r,n-k} - \mu_{in}\mu_{r,n-k})/S. \qquad \text{QED.} \end{aligned}$$

When $k = 0$, $t_{ir}^{(k)}$ is defined to be $\delta_{ir} = \begin{cases} 1 & \text{if } i = r \\ 0 & \text{otherwise} \end{cases}$; the usual multinomial variances and covariances result, namely

$$\operatorname{cov}(m_{in}, m_{rn}) = -\mu_{in}\mu_{rn}/S$$

and

$$\operatorname{var} m_{in} = \mu_{in}(1 - \mu_{in})/S.$$

Let C_{ik} be the kth element of C_i , $k = 1, \cdots, n - 1$. Then

$$\begin{aligned} S \operatorname{var} C_{ik} &= S \operatorname{var}\left(\sum_{j=1}^{a} t_{ij}m_{ik} - m_{i,k+1}\right) \\ &= \mu_{i,k+1}(1 - \mu_{i,k+1}) + \sum_{j=1}^{a} t_{ij}^2\mu_{jk}(1 - \mu_{jk}) \\ &\qquad - \sum_{j \neq j'}^{a} t_{ij}t_{ij'}\mu_{jk}\mu_{j'k} - 2\sum_{j=1}^{a} t_{ij}\mu_{jk}(t_{ij} - \mu_{i,k+1}) \\ &= \mu_{i,k+1}(1 + \mu_{i,k+1}) - \sum_{j=1}^{a} t_{ij}^2\mu_{jk} - \left(\sum_{j=1}^{a} t_{ij}\mu_{jk}\right)^2 \\ &= \mu_{i,k+1} - \sum_{j=1}^{a} t_{ij}^2\mu_{jk}\,, \end{aligned}$$

and when $k > l$

$$S \operatorname{cov}(C_{ik}, C_{il}) = \mu_{i,l+1}(t_{ii}^{(k-l)} - \mu_{i,k+1}) - \sum_{j=1}^{a} t_{ij}\mu_{jl}(t_{ij}^{(k+1-l)} - \mu_{i,k+1})$$

$$- \sum_{j=1}^{a} t_{ij}\mu_{i,l+1}(t_{ji}^{(k-l-1)} - \mu_{jk})$$

$$+ \sum_{j=1}^{a} t_{ij} \sum_{r=1}^{a} t_{ir}\mu_{rl}(t_{jr}^{(k-l)} - \mu_{jk})$$

$$= -\sum_{j=1}^{a} t_{ij}[\mu_{jl}t_{ij}^{(k+1-l)} + \mu_{i,l+1}t_{ji}^{(k-l-1)}]$$

$$+ \sum_{r=1}^{a} t_{ir}\mu_{rl} \sum_{j=1}^{a} t_{ij}t_{jr}^{(k-l)} + \mu_{i,l+1}t_{ii}^{(k-l)} = 0,$$

since

$$\sum_{j=1}^{a} t_{ij}t_{jr}^{(k-l)} = t_{ir}^{(k+1-l)}.$$

Hence Λ_i is a diagonal matrix with diagonal elements

$$\mu_{i,k+1} - \sum_{j=1}^{a} t_{ij}^2\mu_{jk}.$$

Now Λ_i is an unknown matrix, but each μ is consistently estimated by its corresponding m. Also, since

$$\operatorname*{plim}_{S\to\infty} N = \nu \quad \text{and} \quad \operatorname*{plim}_{S\to\infty} M = \mu,$$

it follows that

$$\operatorname*{plim}_{S\to\infty} \bar{T} = \nu\mu'(\mu\mu')^{-1} = T,$$

so that elements of $\bar{T}$ are consistent estimates of corresponding elements of T. Therefore, $\hat{\Lambda}_i$, a consistent estimate of Λ_i, can be formed by substituting the m's and the elements of $\bar{T}$ for the corresponding μ's and t's in the above equations for the elements of Λ_i.

Consider the estimate

$$\tilde{T}_i = N_i\hat{\Lambda}_i^{-1}M'(M\hat{\Lambda}_i^{-1}M)^{-1}$$

(if $M\hat{\Lambda}_i^{-1}M'$ is nonsingular) which would be the least squares estimate of T_i if $\hat{\Lambda}_i^{-1} = \Lambda_i^{-1}$ and if Λ_i were not a function of the t's. It is easy to see that $\tilde{T}_i$ is also a consistent estimate of T_i.

If R and Q are symmetric matrices then $R \geq Q$ if and only if $x(R - Q)x' \geq 0$ for all vectors x. Let us adapt the definition of minimum variance linear unbiased estimate of a vector-valued parameter given in ([3], section 2.5) as follows. It may be said that of two consistent estimators,

say u and v, of a vector-valued parameter θ, u is asympotically minimum variance if the asymptotic covariance matrix of $\sqrt{S}\ (u - \theta)$ is less than or equal to the asymptotic covariance matrix of $\sqrt{S}\ (v - \theta)$ in the sense defined above.

THEOREM. *Let $\tilde{\Sigma}_i$ be the asymptotic covariance matrix of $\sqrt{S}\ (\tilde{T}_i - T_i)$ and let Σ_i be the asymptotic covariance matrix $\sqrt{S}\ (\bar{T}_i - T_i)$. Then $\tilde{\Sigma}_i \leq \Sigma_i$.*

PROOF. We know that

$$m_{kl} = \mu_{kl} + O(1/\sqrt{S}),$$

since for fixed l, m_{kl} is the maximum likelihood estimate of μ_{kl}. Therefore

$$\bar{t}_{ij} = t_{ij} + O(1/\sqrt{S})$$

and also

$$\tilde{t}_{ij} = t_{ij} + O(1/\sqrt{S})$$

for all i and j. Then write

$$N_i = \nu_i + X/\sqrt{S}, \qquad M = \mu + Y/\sqrt{S}, \quad \text{and} \quad \hat{\Lambda}_i^{-1} = \Lambda_i^{-1} + Z/\sqrt{S}.$$

To terms of order $1/\sqrt{S}$,

$$MM' = \mu\mu' + (\mu Y' + Y\mu')/\sqrt{S},$$

so that

$$(MM')^{-1} = [I - (\mu\mu')^{-1}(\mu Y' + Y\mu')/\sqrt{S}](\mu\mu')^{-1}$$

to terms of order $1/\sqrt{S}$. Similarly,

$$(M\hat{\Lambda}_i^{-1}M')^{-1} = [I - (\mu\Lambda_i^{-1}\mu')^{-1}(\mu\Lambda_i^{-1}Y' + \mu Z\mu' + Y\Lambda_i^{-1}\mu')](\mu\Lambda_i^{-1}\mu')^{-1}$$

to terms of order $1/\sqrt{S}$. Therefore, asymptotically,

$$\sqrt{S}\ (\bar{T}_i - T_i) = [(\nu_i Y' + X\mu') - T_i(\mu Y' + Y\mu')](\mu\mu')^{-1}$$

and

$$\sqrt{S}\ (\tilde{T}_i - T_i) = [(\nu_i Z\mu' + X\Lambda_i^{-1}\mu' + \nu_i\Lambda_i^{-1}Y') - T_i(\mu\Lambda_i^{-1}Y' + \mu Z\mu' + Y\Lambda_i^{-1}\mu')](\mu\Lambda_i^{-1}\mu')^{-1}.$$

Since $\nu_i = T_i\mu$, the above equations reduce to

$$\sqrt{S}\ (\bar{T}_i - T_i) = (X - T_i Y)\mu'(\mu\mu')^{-1}$$

and

$$\sqrt{S}\ (\tilde{T}_i - T_i) = (X - T_i Y)\Lambda_i^{-1}\mu'(\mu\Lambda_i^{-1}\mu')^{-1}.$$

But the covariance matrix of $X - T_i Y$ is Λ_i, so that

$$\bar{\Sigma}_i = (\mu\mu')^{-1}(\mu\Lambda_i\mu')(\mu\mu')^{-1}$$

and

$$\tilde{\Sigma}_i = (\mu\Lambda_i^{-1}\mu')^{-1}.$$

By using the Schwarz inequality for matrices given in ([3], section 2.5), it follows immediately that

$$\bar{\Sigma}_i \geq \tilde{\Sigma}_i .$$

Thus $\tilde{T}_i$ is asymptotically more efficient, in the sense defined above, than is $\bar{T}_i$. In fact, a modification of the above proof will show that $\tilde{T}_i$ is asymptotically at least as efficient as any other estimator of the form $N_iPM'(MPM')^{-1}$, where P is a positive definite matrix.

One should note that $\tilde{T}$ is obtained by first computing $\bar{T}$ and then modifying it. One can just as well, via the same procedure as above, modify $\tilde{T}$ to obtain another estimate of T, T^*, say. However, the asymptotic efficiency of T_i^* will be the same as that of $\tilde{T}_i$ for each i, since T_i^* will be of the form $N_iPM'(MPM')^{-1}$, where P is $\hat{\Lambda}_i^{-1}$ based on $\tilde{T}$, i.e.,

$$P = \Lambda_i^{-1} + O(1/\sqrt{S}).$$

It has been shown in [2] that $\bar{T}$ has the desirable property that $e\bar{T} = e$, where e is the a-dimensional vector $(1\ 1\ \cdots\ 1)$. That this is not true of $\tilde{T}$ can be seen from the example of the next section.

Discussion

Let us consider Miller's example [cf. 6]. In his case $a = 2$, $n = 20$, and

$$\bar{T} = \begin{bmatrix} .92 & .39 \\ .08 & .61 \end{bmatrix}.$$

Using this estimate and N and M, the diagonal elements of $\hat{\Lambda}_1$ are found to be

[.20 −.04 −.03 .23 .09 .09 −.01 .16 .19 .02
−.01 .16 .19 .22 .15 −.05 .09 .19 .12]

and the diagonal elements of $\hat{\Lambda}_2$ are

[.11 .28 .25 .05 .12 .12 .22 .08 .02 .16
.22 .08 .02 −.04 −.01 .19 .12 .02 .06].

Note that some of these elements are negative. However, since these elements are estimates of variance (which are non-negative quantities) these "estimates" were replaced by .001 (zero cannot be used since $\hat{\Lambda}_i$ must be nonsingular). It is then found that

$$\tilde{T} = \begin{bmatrix} .81 & .25 \\ -.02 & .49 \end{bmatrix}.$$

This example illustrates that $e\tilde{\bar{T}} \neq e$ in general. It also shows that elements of $\tilde{\bar{T}}$ may be inadmissible in that they can turn out to be negative. However, as Goodman points out [2], $\bar{T}$ also has this latter fault.

We suggest in this case the estimate

$$\tilde{\bar{T}} = \begin{bmatrix} .81 & .25 \\ .19 & .75 \end{bmatrix}.$$

To compare $\tilde{\bar{T}}$ with $\bar{T}$, it suffices to compare the covariance matrices of $\tilde{\bar{T}}_1 = \tilde{T}_1$ and $\bar{T}_1$ (since $\tilde{\bar{T}}_2 = e - \tilde{T}_1$ and $\bar{T}_2 = e - \bar{T}_1$). An estimate of $\tilde{\Sigma}_1$ is $(M\hat{\Lambda}_1^{-1}M')^{-1}$, and $\bar{\Sigma}_1$ is estimated by

$$(MM')^{-1}(M\hat{\Lambda}_1 M')(MM')^{-1}.$$

After re-estimating Λ_1 by using $\tilde{\bar{T}}$ (and finding that all the diagonal elements of this estimate are positive), we used this estimate and found that

$$\text{est } \tilde{\Sigma}_1 = \begin{bmatrix} .049 & -.123 \\ -.123 & .517 \end{bmatrix}$$

and

$$\text{est } \bar{\Sigma}_1 = \begin{bmatrix} .059 & -.149 \\ -.149 & .616 \end{bmatrix}.$$

The easiest way to see that est $\bar{\Sigma}_1 \geq$ est $\tilde{\Sigma}_1$, where $\geq$ is as defined above, is to note that all principal minors of est $\bar{\Sigma}_1$ − est $\tilde{\Sigma}_1$ are non-negative.

Goodman also states in [2] that if the observed transitional proportions are available, they would clearly be more appropriate in the estimation of transitional probabilities. In [1] Anderson and Goodman give estimates of T and the asymptotic variances and covariances of the elements of their estimate when the observed transitional proportions are available. Calling their estimates $\hat{t}_{ij}$, the asymptotic variances and covariances may be written as

$$\text{var } \hat{t}_{ij} = t_{ij}(1 - t_{ij})/\phi_i , \qquad \text{cov } (\hat{t}_{ij} , \hat{t}_{i'j'}) = -\delta_{jj'} t_{ij} t_{i'j'}/\phi_j ,$$

where

$$\delta_{ik} = \begin{cases} 1 & i = k \\ 0 & i \neq k, \end{cases}$$

and

$$\phi_i = \sum_{j=1}^{a} \sum_{k=2}^{n} \mu_{ij} t_{ij}^{[k-1]} = \sum_{k=2}^{n} \mu_{ik} .$$

We can estimate μ_{ik} by m_{ik} and so, in the example under consideration,

we can estimate $\hat{\Sigma}_1$, the asymptotic covariance matrix of the estimate of T_1 (and hence $T_2 = e - T_1$) when the observed transitional proportions are available, in order to see the improvement when this information is available. It is found that est $\phi_1 = 15.3$ and est $\phi_2 = 3.7$ so that

$$\text{est } \hat{\Sigma}_1 = \begin{bmatrix} .010 & .000 \\ .000 & .051 \end{bmatrix},$$

using $\bar{\bar{T}}$ to estimate T. Hence est $\hat{\Sigma}_1 \leq$ est $\tilde{\Sigma}_1$ in this case.

With regard to the amount of extra computation involved in computing $\tilde{T}$, one can see from the form of the estimate that once $\bar{T}$ is given, the modification process will take about as much time as the computation of $\bar{T}$ itself. Since we have no idea of the relative decrease in the variances of the estimates, we cannot discuss the trade-off between the doubled computation time and the reduction in variance. This trade-off is, however, an important practical factor in determining whether $\tilde{T}$ or $\bar{T}$ is used.

REFERENCES

[1] Anderson, T. W. and Goodman, L. A. Statistical inference about Markov chains. *Ann. math. Statist.*, 1957, **28**, 89–110.

[2] Goodman, L. A. A further note on "Finite Markov processes in psychology." *Psychometrika*, 1953, **18**, 245–248.

[3] Grenander, U. and Rosenblatt, M. *Statistical analysis of stationary time series.* New York: Wiley, 1957.

[4] Kao, R. C. Note on Miller's "Finite Markov processes in psychology." *Psychometrika*, 1953, **18**, 241–243.

[5] Kempthorne, O. *The design and analysis of experiments.* New York: Wiley, 1952.

[6] Miller, G. A. Finite Markov processes in psychology. *Psychometrika*, 1952, **17**, 149–167.

Manuscript received 6/13/58

Revised manuscript received 11/8/58