Scand J Statist 13: 113–134, 1986

On Estimating Transition Intensities of a Markov Process with Aggregate Data of a Certain Type: "Occurrences but no Exposures"

RICHARD D. GILL

Centre for Mathematics and Computer Science, Amsterdam

ABSTRACT. In demography finite-state-space time-homogeneous Markov processes are often used, explicitly or implicitly, to model the movement of individuals between various states (e.g. studies of marital formation and dissolution or of interregional migration). However the fact that data are often only available at certain levels of aggregation, preventing a simple and exact statistical analysis, has caused much confusion and has even impeded the adoption of probabilistic modelling and statistical analysis. In this paper we consider one specific form of aggregate data and propose a new method of estimation of the underlying Markov process. Some preliminary results on the properties of this method are given and some open problems are discussed.

Key words: Markov process, aggregate data, multidimensional mathematical demography, multistate life-table, occurrence–exposure rate, fixed-point theorem, degree theory

1. Introduction

In demography finite-state-space time-homogeneous Markov processes are often used, explicitly or implicitly, to model the movement of individuals between various states (e.g. studies of marital formation and dissolution or of interregional migration). However the fact that data are often only available at certain levels of aggregation, preventing a simple and exact statistical analysis, has caused much confusion and has even impeded the adoption of probabilistic modelling and statistical analysis. In this paper we consider one specific form of aggregate data and propose a new method of estimation of the underlying Markov process. Some preliminary results on the properties of this method are given.

In this field the so-called "occurrence–exposure rate" plays a central role: this is the ratio of the number of events of a certain type, the occurrences (typically the number of direct moves from one particular state to another), to the total amount of time individuals have been at risk to this event (i.e. have occupied the first state), the exposure. The occurrence–exposure rate can be considered as an estimate of the corresponding Markov-process intensity. However data are often only available on the occurrences, aggregated over time and individuals, while the exposures are not recorded. One is usually interested in estimating the Markov process model as a means for computing the net transfers for each pair of states: the number of individuals who start in one state and finish in the other.

Let us start by summarizing some of the well-known properties of a homogeneous Markov process $\mathbf{X}=(\mathbf{X}_t: t\geqslant 0)$ with finite state space $\{1, 2, \ldots, p\}$ for some positive integer p (random variables are printed in bold type; the same symbol in ordinary (italic) type denotes a possible realization of the corresponding random variable). An early reference where much of this material can be found is Albert (1962). This process is described by an initial distribution μ, considered as a row-vector with non-negative elements μ_i, $i=1,\ldots, p$, $\Sigma\mu_i=1$, and a set of intensities Q, considered as a $p\times p$ matrix with non-negative off-diagonal elements q_{ij}, $i\neq j$, and diagonal elements $q_{ii}=-\Sigma_{j\neq i}q_{ij}\leqslant 0$. For $i\neq j$ one interprets q_{ij} by the relation: $q_{ij}\cdot h \approx \mathbb{P}(\mathbf{X}_{t+h}=j \mid X_t=i)$ for small $h>0$. The process $\mathbf{X}$ can be constructed by first selecting an initial state according to the probabilities μ, i.e. $\mu_i=\mathbb{P}(\mathbf{X}_0=i)$, staying in that state an exponentially distributed length of time with mean $-1/q_{ii}$, then jumping to a new state, say j, with

probabilities $\alpha_{ij}=-q_{ij}/q_{ii}$, etc. If $q_{ii}=0$ state i is absorbing; i.e. once state i is entered it is never left again. By convention one chooses to let the paths of $\mathbf{X}$ be right-continuous; i.e. $\mathbf{X}_t$=state at time $t+$. We define $\mathbf{X}_{0-}=\mathbf{X}_0$. Since the state-space is finite it is easy to check that this procedure really does define a process $(\mathbf{X}_t: t\geq 0)$; i.e. the number of jumps in any bounded time-interval is almost surely bounded. We shall only be concerned with the time interval $t\in[0, 1]$. The process $\mathbf{X}$ is Markov with transition matrix $P_t=\exp(Qt)$ where $(P_t)_{ij}=\mathbb{P}(\mathbf{X}_{s+t}=j\,|\,\mathbf{X}_s=i)$. Consequently the marginal distribution of $\mathbf{X}_t$ is given by the vector of probabilities μP_t. In particular we define ν to be the distribution at time 1 or the final distribution; i.e.

$$\nu=\mu e^{Q}. \tag{1}$$

Also we let l denote the row-vector of expected lengths of time spent in each state during the time interval [0, 1],

$$l_i=\mathbb{E}(\mathbf{l}_i)=\mathbb{E}\left(\int_0^1 \mathbf{I}\{\mathbf{X}_s=i\}\,\mathrm{d}s\right),$$

where $\mathbf{I}\{\ldots\}$ denotes the indicator random variable of the specified event. So we have

$$l=\int_0^1 \mu P_s\,\mathrm{d}s=\int_0^1 \mu e^{Qs}\,\mathrm{d}s. \tag{2}$$

Letting $\underline{1}$ denote a row-vector of 1's, and $^{\mathrm{T}}$ denote transpose, we obviously have

$$l\underline{1}^{\mathrm{T}}=1. \tag{3}$$

Also we have

$$lQ=\int_0^1 \mu e^{Qs}\,Q\,\mathrm{d}s=[\mu e^{Qs}]_0^1=\mu(e^{Q}-I)=\nu-\mu. \tag{4}$$

Note that $Q\underline{1}^{\mathrm{T}}=\underline{0}^{\mathrm{T}}$ so that rank $(Q)\leq p-1$. If rank $(Q)=p-1$ and moreover $\underline{1}^{\mathrm{T}}$ is linearly independent of the columns of Q (i.e. rank $(Q:\underline{1}^{\mathrm{T}})=p$) then for given μ and Q the equations in l:

$$l=\int_0^1 \mu e^{Qs}\,\mathrm{d}s \tag{5}$$

and

$$lQ=\mu(e^{Q}-I),\ l\underline{1}^{\mathrm{T}}=1 \tag{6}$$

are equivalent. (In practice one uses (6) to compute l for given μ and Q.) A necessary and sufficient condition for rank $(Q:\underline{1}^{\mathrm{T}})=p$ is that there exists at least one state to which all states have access (see Appendix I). This is also equivalent to the condition rank $(Q)=p-1$. More complex situations can be handled by appropriate decompositions of the state space, cf. Funck Jensen (1982b) and Appendix III. (We say that i has access to j if $i=j$ or if there exist states $i_0, i_1, \ldots, i_k$ with $i_0=i$, $i_k=j$ and $q_{i_{m-1}i_m}>0$ for $m=1,\ldots,k$. States i and j communicate if each has access to the other.)

Finally we denote by N the matrix with elements N_{ij}=expected number of jumps from state i to state j during the time interval [0, 1] $(i\neq j)$, $N_{ii}=-\Sigma_{j\neq i}N_{ij}$. So $N_{ij}=\mathbb{E}(\mathbf{N}_{ij})=\mathbb{E}(\Sigma_{t\in[0,1]}\mathbf{I}\{\mathbf{X}_{t-}=i, \mathbf{X}_t=j\})$ for $i\neq j$. One can show (e.g. by using Aalen (1978), Example 3 and the fact that the expectation of a martingale is constant) that for $i\neq j$, $N_{ij}=l_i q_{ij}$, which we can rewrite (taking

account of the definition of the diagonal elements of Q and N) as

$$N=\text{diag}\ (l)Q \tag{7}$$

where "diag" of a vector denotes the diagonal matrix with the corresponding elements of the vector on its diagonal. Note that by the identity (sometimes called the accounting equation)

$$\mathbf{I}\{\mathbf{X}_1=i\}=\mathbf{I}\{\mathbf{X}_0=i\}+\sum_{j\neq i}\sum_{t}\mathbf{I}\{\mathbf{X}_{t-}=j,\ \mathbf{X}_t=i\}-\sum_{j\neq i}\sum_{t}\mathbf{I}\{\mathbf{X}_{t-}=i,\ \mathbf{X}_t=j\}$$

we obtain on taking expectations the so-called flow equation

$$\nu=\mu+\underline{1}N \tag{8}$$

The statistical problem we will address is the following. For $m=1,\ldots,n$ let $\mathbf{X}^m=(\mathbf{X}_t^m: t\in[0,1])$ be processes such that conditional on $\mathbf{X}_0^m=X_0^m$, $m=1,\ldots,n$, $\mathbf{X}^m$ are independent homogeneous Markov processes on $\{1,\ldots,p\}$ with the same intensity matrix Q and with initial distributions point mass on X_0^m, $m=1,\ldots,n$. Thus we consider n individuals or particles who, starting from (and conditional on) some arbitrary initial configuration on $\{1,\ldots,p\}$, move independently from state to state in $\{1,\ldots,p\}$ during the time interval $[0,1]$ according to the description given above. Now define the random variables

$$\mathbf{N}_{ij}^n=\sum_{m,t}\mathbf{I}\{\mathbf{X}_{t-}^m=i,\ \mathbf{X}_t^m=j\}\quad i\neq j$$

=total number of moves from i to j during $[0,1]$, "occurrences"

$$\mathbf{N}_{ii}^n=-\sum_{j\neq i}\mathbf{N}_{ij}^n$$

$$\mathbf{l}_i^n=\sum_m\int_0^1\mathbf{I}\{\mathbf{X}_t^m=i\}\mathrm{d}t$$

=total time spent in state i, "exposure"

$$\boldsymbol{\mu}_i^n=\sum_m\mathbf{I}\{\mathbf{X}_0^m=i\}$$

=initial configuration

$$\boldsymbol{\nu}_i^n=\sum_m\mathbf{I}\{\mathbf{X}_1^m=i\}$$

=final configuration

where the summations are over $m=1,\ldots,n$, $t\in[0,1]$ and $j\in\{1,\ldots,p\}$. Then defining μ by $\mathbb{E}\boldsymbol{\mu}^n=n\mu$, we obtain that $\mathbb{E}\mathbf{N}^n=nN$, $\mathbb{E}\mathbf{l}^n=nl$ and $\mathbb{E}\boldsymbol{\nu}^n=n\nu$, where N, l, and ν are determined from μ and Q by formulas (1), (5) or (6), and (7). Formula (8) also holds. The statistical problem is now to estimate Q on the basis of observation of $\mathbf{N}^n$ and $\boldsymbol{\mu}^n$; i.e. given the initial configuration and the total number of moves during $[0,1]$. We assume that all other quantities, in particular $\mathbf{l}^n$, are not observed. We seek estimators which have good properties as $n\to\infty$. Note that $\boldsymbol{\mu}^n\underline{1}^{\mathrm{T}}=\boldsymbol{\nu}^n\underline{1}^{\mathrm{T}}=\mathbf{l}^n\underline{1}^{\mathrm{T}}=n$, $\mathbf{N}^n\underline{1}^{\mathrm{T}}=\underline{0}^{\mathrm{T}}$ and that $\boldsymbol{\nu}^n=\boldsymbol{\mu}^n+\underline{1}\mathbf{N}^n$.

Before describing our new proposal, we discuss the currently available solutions to this problem. Had $\mathbf{l}^n$ been observed too (the total exposure to the risks of making the various

possible moves), statistical theory shows that the matrix of *empirical occurrence–exposure rates* $\hat{\mathbf{Q}}^n=(\text{diag } \mathbf{l}^n)^{-1}\mathbf{N}^n$ possesses a large number of desirable properties as estimator of Q. Conditional on $\boldsymbol{\mu}^n=n\mu$ it is a maximum likelihood estimator of Q. Under conditions which ensure that the elements of $\mathbf{l}^n$ become arbitrarily large at uniform rate as $n\to\infty$ (here we consider a sequence of the situations described above, indexed by $n=1, 2, \ldots$, in which only the intensity matrix Q is kept fixed) $\hat{\mathbf{Q}}^n$ is asymptotically multivariate normally distributed about Q with all off-diagonal components asymptotically independent and with asymptotic variances which can be estimated by the corresponding elements of $(\text{diag } \mathbf{l}^n)^{-1}\hat{\mathbf{Q}}^n$. The estimator $\hat{\mathbf{Q}}^n$ also possesses asymptotic optimality properties among all estimators based on complete individual level data: i.e. where all the processes $(\mathbf{X}_t^m : t\in[0, 1])$, $m=1, \ldots, n$, are observed.

In our situation, which commonly occurs in practice, this estimator is unavailable. Also the joint distribution of $(\boldsymbol{\mu}^n, \mathbf{N}^n)$ is so intractable that a maximum likelihood estimator of Q based on data $(\boldsymbol{\mu}^n, \mathbf{N}^n)$ cannot be computed, neither directly nor by means of the EM-algorithm (cf. Dempster *et al.*, 1977), for which one would have to evaluate $\mathbb{E}_Q(\mathbf{l}^n|\boldsymbol{\mu}^n=\mu^n, \mathbf{N}^n=N^n)$. Therefore one usually takes recourse to the working approximation $\mathbf{l}^n\approx\tilde{\mathbf{l}}^n=\frac{1}{2}(\boldsymbol{\mu}^n+\boldsymbol{\nu}^n)$ and estimates Q by $\tilde{\mathbf{Q}}^n=(\text{diag } \tilde{\mathbf{l}}^n)^{-1}\mathbf{N}^n$. This estimator is generally inconsistent. Though in most situations its bias will be small compared to its standard deviation, and in any case the whole Markov process set-up is itself only a "working approximation" to reality, it is felt that it is a failure of "the statistical approach" that this very common situation does not yet have a nice statistical solution.

In practice interest often centres on the transition matrix P_1 (as a means of predicting the random variables $\Sigma_{(m)}\mathbf{I}\{\mathbf{X}_0^m=i, \mathbf{X}_1^m=j\}$) rather than on the intensity matrix Q. Within the Markov process set-up one would generally estimate P_1 by substituting an estimate of Q in the formula $P_1=\exp(Q)$. The alternative "actuarial" approach to the whole problem is to abandon the time-homogeneous Markov process model and to elevate the working approximation $\mathbf{l}^n\approx\frac{1}{2}(\boldsymbol{\mu}^n+\boldsymbol{\nu}^n)$ or $l\approx\frac{1}{2}(\mu+\nu)$ to an element of the mathematical model, denoted then as "the linear integration hypothesis". Various authors then derive, as an estimator of P_1, $\tilde{\mathbf{P}}_1^n=(I+\frac{1}{2}\tilde{\mathbf{Q}}^n)(I-\frac{1}{2}\tilde{\mathbf{Q}}^n)^{-1}$; cf. Rogers & Ledent (1976). However there are some logical inconsistencies in this derivation which are discussed in Keilman & Gill (1986). In our set-up this estimator too will typically be inconsistent though usually not disastrously so.

Our new approach is simply to use the (very old) method of moments: equate the observed variables $\boldsymbol{\mu}^n$ and $\mathbf{N}^n$ to their expected values $n\mu$ and nN and solve the resulting equations in μ and Q. This is equivalent to solving equations (5) or (6), and (7) considered for given μ and N (equal to $n^{-1}\boldsymbol{\mu}^n$ and $n^{-1}\mathbf{N}^n$ respectively), as equations in unknowns l and Q.

Various questions then arise:

(i) When, for given μ and N, do equations (5), (6) and (7) have a solution in l and Q?
(ii) When is the solution unique?
(iii) What is a good algorithm for finding a (the) solution?
(iv) What are the statistical properties of the resulting estimators?

We can prove that there always exists a solution. If all states communicate and a further simple condition is satisfied the solution is unique; however we can only verify this condition when $p=2$. When the process is hierarchial ($q_{ij}=0$ for $j<i$) it can also be shown that there is exactly one solution. We conjecture that there always exists exactly one solution.

Regarding question (iii), an obvious iteration method is based on cycling repeatedly through equations (5) or (6) and (7), first computing l for given μ and Q, then Q for given l and N. This resembles the EM-algorithm in that we compute in each cycle $\mathbb{E}_Q(\mathbf{l}^n|\boldsymbol{\mu}^n=\mu^n)$; the EM-algorithm requires one to compute $\mathbb{E}_Q(\mathbf{l}^n|\boldsymbol{\mu}^n=\mu^n, \mathbf{N}^n=N^n)$. However, this superficial resemblence does not guarantee any convergence properties of the iterations. It has been therefore a total surprise that

in every example yet considered, these iterations converge quickly, independently of the starting value, to one limiting value. No complete explanation for this has yet been found.

An alternative approach is to attempt numerical solution, in l, for given μ, ν and N, with ν defined by (8), of the equations (cf. (1), (3) and (7))

$$\nu=\mu \exp\{(\text{diag } l)^{-1}N\},\ l\underline{1}^{\mathrm{T}}=1$$

which can be shown under the "full-rank" condition rank $(N)=p-1$ to be equivalent to solving the fixed point equation of the previous method

$$l=\int_0^1 \exp\{(\text{diag } l)^{-1}Ns\}\,\mathrm{d}s.$$

In all examples we tried a standard quasi-Newton method worked excellently.

For practical purposes then questions (i) to (iii) could be considered as satisfactorily answered, though from the point of view of mathematical theory there are as many questions as answers. All the same, as regards (iv), a satisfactory mathematical-statistical theory of the proposed estimators can be given, in which their asymptotic properties can be derived and in particular their asymptotic optimality (among estimators which use only the same aggregate data) can be proved: assuming identifiability of the model.

The rest of the paper consists of two main parts, one devoted to questions (i) to (iii), the other to question (iv), i.e. to mathematical properties of equations (1) to (8), and to statistical properties of the estimator of Q which is defined as the solution to these equations when N, μ and ν are replaced by their sample analogues. An example is also given. Before proceeding with this, however, we must first put the results sketched above into perspective, in particular with regard to practical demography. A Markov process model with constant intensities is usually only considered as a rough approximation to the most realistic model. So an "exact" statistical solution to estimation of this model is not of great practical importance. The contribution we make here is however hopefully of methodological importance. We hope that it clarifies some of the controversy on the "linear integration hypothesis" by illustrating the value of keeping elements of the probabilistic model with which we describe a phenomenon distinct from questions of "numerical approximations" which might be of use when working within the model, and also from questions of data availability (which might also make certain approximations rather convenient); cf. Hoem & Funck Jensen (1982). Put differently, we hope that this contribution illustrates the value of choosing a mathematical model as a framework within which such questions can be objectively discussed. Hopefully it also illustrates that nice statistical solutions for more complicated models and more complicated data-structures (e.g. the time-inhomogeneous model with piecewise linear or piecewise quadratic intensity functions and situations with other types of aggregate data, e.g. period occurrence–exposure rates) can in principle also be obtained. In this perspective the solutions of e.g. Land & Schoen (1982) can be seen as a (possibly very good) working approximation to the solutions which a generalization of the present theory would supply.

Other types of aggregate data are handled by Kalbfleisch *et al.*, (1983) and van der Plas (1983).

2. Solving the estimating equations

As we saw in section 1, for a Markov process with initial distribution μ and intensity matrix Q the following relations hold, where ν is the final distribution or distribution at time 1, l is the

expected length of time spent in each state during [0, 1], and the matrix N contains the expected number of moves between each two states during [0, 1]:

$$\nu=\mu e^{Q} \tag{9}$$

$$l=\int_0^1 \mu e^{Qs}\,\mathrm{d}s \tag{10}$$

$$lQ=\mu(e^{Q}-I)=\nu-\mu \tag{11}$$

$$N=(\operatorname{diag} l)Q \tag{12}$$

$$\nu=\mu+\underline{1}N \tag{13}$$

$$l\underline{1}^{\mathrm{T}}=\mu\underline{1}^{\mathrm{T}}=\nu\underline{1}^{\mathrm{T}}=1;\ N\underline{1}^{\mathrm{T}}=Q\underline{1}^{\mathrm{T}}=\underline{0}^{\mathrm{T}};\qquad q_{ij},\ n_{ij}(i\neq j),\ \mu_i,\ \nu_i,\ l_i\geqslant 0. \tag{14}$$

For a vector, the symbols "$\geqslant$" and "$\gg$" will denote componentwise $\geqslant$ and $>$ respectively. (By "$>$" we mean "$\geqslant$ and not $=$".) Note that N can also be considered as an intensity matrix and as such, if $l\gg\underline{0}$, by (12) it generates the same classification of states as Q.

Our problem is now the following. Let μ be an initial distribution and let N be an intensity matrix with no redundant states, i.e. a state i with $\mu_i=n_{ij}=n_{ji}=0$ for all j; let $\nu=\mu+\underline{1}N$. Necessarily $\nu\underline{1}^{\mathrm{T}}=1$. Does there exist an intensity matrix Q satisfying (10) and (12)? First we note that if such a Q exists, then ν must also satisfy (9), by linearity and the derivation of the "flow equation" (13) in section 1. Since if a state is ever occupied it has positive probability of being occupied at any particular time>0, we must have $\nu\gg\underline{0}$ and l (defined by (10)) satisfies $l\gg\underline{0}$. Therefore we can write $Q=(\operatorname{diag} l)^{-1}N$. Thus the existence of Q implies $\nu\gg\underline{0}$ and the existence of a vector $l\gg\underline{0}$ such that, from (10),

$$l=\int_0^1 \mu\,\exp\{(\operatorname{diag} l)^{-1}Ns\}\,\mathrm{d}s \tag{15}$$

and, from (9),

$$\nu=\mu\,\exp\{(\operatorname{diag} l)^{-1}N\},\ l\underline{1}^{\mathrm{T}}=1. \tag{16}$$

We now show that (15) implies the existence of Q and, if rank $(N)=p-1$, is equivalent to (16). Now if (15) holds define $Q=(\operatorname{diag} l)^{-1}N$ and we have (10) and (12) holding trivially. On the other hand, if (15) or (16) holds, define in either case $Q=(\operatorname{diag} l)^{-1}N$ and (15) and (16) are equivalent to

$$l=\int_0^1 \mu\,\exp(Qs)\,\mathrm{d}s \tag{17}$$

and (using the identity $\nu=\mu+\underline{1}N=\mu+l(\operatorname{diag} l)^{-1}N$)

$$lQ=\mu\{\exp(Q)-I\},\ l\underline{1}^{\mathrm{T}}=1$$

respectively. But we saw in section 1 that in the presence of the rank condition rank $(Q)=$ rank $(N)=p-1$, (17) and (18) are equivalent.

For the rest of this section we suppose unless otherwise stated that we are given μ, N and $\nu=\mu+\underline{1}N$ satisfying rank $(N)=p-1$ and $\nu\gg\underline{0}$. Does there exist $l\gg\underline{0}$ such that (15) or (16) holds? Now let S denote the unit simplex $\{l\in\mathbb{R}^p: l\geqslant\underline{0},\ l\underline{1}^{\mathrm{T}}=1\}$ and let S^0 denote its (relative) interior $\{l\in\mathbb{R}^p: l\gg\underline{0},\ l\underline{1}^{\mathrm{T}}=1\}$. We shall give in this section a positive answer in the special case in which all states communicate—i.e. N is irreducible. In Appendices II and III we obtain a completely general (positive) result by relaxing the conditions N irreducible, rank $(N)=p-1$, in turn. It will

be useful to extend the definition of the right hand sides of (15) and (16) from $l \in S^0$ to $l \in S$. The case in which all states communicate is almost the only case in which a continuous extension is possible: in fact for there to be a continuous extension we need that each state either has access to all other states or is an absorbing state. Define functions $\hat{l}$ and $\hat{\nu}$ on S^0 by

$$\hat{l}(l)=\int_0^1 \mu \exp\{(\operatorname{diag} l)^{-1}Ns\}\, ds$$

$$\hat{\nu}(l)=\mu \exp\{(\operatorname{diag} l)^{-1}N\}.$$

We extend $\hat{l}$ and $\hat{\nu}$ to all of S by going back to the explicit construction of the process $\mathbf{X}$ in section 1. Define $\alpha_{ij}=-n_{ij}/n_{ii}$ for $i\neq j$ such that $n_{ii}<0$, $\alpha_{ij}=0$ otherwise. For $l \in S$ we say $-l_i/n_{ii}=\infty$ if $n_{ii}=0$. By an exponentially distributed random variable with mean zero or mean infinity we mean a random variable which is identically 0 or identically $+\infty$ respectively. For the following construction we suppose that each state either has access to all others or is absorbing. For $l \in S$ we define a process $\mathbf{X}$ as follows. Choose an initial state, say i, according to the distribution μ. Stay there an exponentially distributed length of time with mean $-l_i/n_{ii}$, then jump to state j with probability α_{ij}, stay there an exponentially distributed length of time with mean $-l_j/n_{jj}$, jump to state k with probability $\alpha_{jk}, \ldots$. If some l_i's are zero (all cannot be zero) the condition on the state space ensures that if one arrives in a state with $-l_i/n_{ii}=0$, then after an almost surely finite number of instantaneous jumps one arrives in a state with $-l_i/n_{ii}>0$ and stays in this state a positive length of time. It can be verified that this procedure does define a process $\mathbf{X}$ by $\mathbf{X}_t$=state at time $t+$ for all t, almost surely; see Appendix II.

For this new process we can compute the expected length of time spent in each state during [0, 1] and the final distribution over states: we denote these quantities by $\hat{l}(l)$ and $\hat{\nu}(l)$. It can be shown (see Appendix II) that this definition extends $\hat{l}$ and $\hat{\nu}$ from S^0 to S in a continuous way. For i such that $-l_i/n_{ii}=0$ we have $\hat{l}(l)_i=0$, $\hat{\nu}(l)_i=0$. Clearly $\hat{\nu}$, $\hat{l}$ map S into S and S^0 into S^0. Note that if not every state had access to all other states or was absorbing, then there would exist a proper subset of two or more states which was absorbing and communicating. If $l_i=0$ for all states in this class, then on arrival in this class one would immediately and instantaneously make an infinite number of jumps within the class, so the process $\mathbf{X}$ cannot be defined. Moreover, for $l_i>0$, as $l_i \to 0$ for all states in this class, $\hat{\nu}(l)_i$ and $\hat{l}(l)_i$ do not converge.

We now make the even stronger assumption that all states communicate, and prove under this assumption that the equation $\hat{\nu}(l)=\nu$ has a solution in S^0. Note that under this assumption, $l_i=0 \Leftrightarrow \hat{\nu}(l)_i=0$, and recall that $\nu_i>0$ for all i. We make use of a dual form of the lemma, from fixed-point theory, of Knaster, Kuratowski & Mazurkiewicz (1929) (the K-K-M lemma) which can also be found in Ch. 8, §2 of Berge (1959), in Todd (1976) or in van der Laan (1980). The dual version is due to Freidenfelds (1974, theorem 1'). For this we define the faces S_i of S by $S_i=\{l \in S : l_i=0\}$.

Lemma (Knaster, Kuratowski & Mazurkiewicz; Freidenfelds)
Let $C_1, \ldots, C_p$ be closed subsets of S such that

$$S=\bigcup_1^p C_i,$$

$S_i \subset C_i$ for all i. Then

$$\bigcap_1^p C_i \text{ 1 is non-empty.}$$

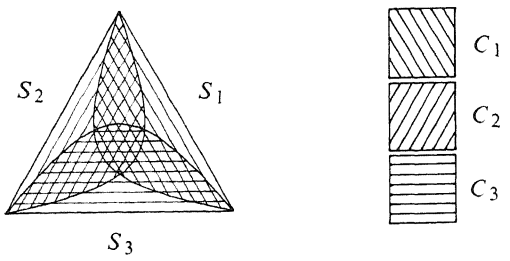

Fig. 1. The K-K-M lemma, $p=3$.

For our application we define $C_i=\{l\in S:\hat{\nu}(l)_i\leqslant\nu_i\}$. Since $\hat{\nu}:S\to S$ is continuous, C_i is closed. Since $\hat{\nu}(l)$, $\nu\in S$, for all l there exists i such that $\hat{\nu}(l)_i\leqslant\nu_i$; i.e.

$$S=\bigcup_1^p C_i.$$

Finally if $l_i=0$, then $\hat{\nu}(l)_i=0<\nu_i$, so $l\in C_i$. So $C_1,\ldots,C_p$ satisfy the conditions of the lemma and

$$\bigcap_1^p C_i$$

is non-empty. But for

$$l\in\bigcap_1^p C_i,$$

$\hat{\nu}(l)_i\leqslant\nu_i$ for all i, hence $\hat{\nu}(l)=\nu$.

In Appendices II and III we extend this result to prove finally: for any initial distribution μ and intensity matrix N with $\nu=\mu+\underline{1}N\gg\underline{0}$, there exists $l\in S^0$ satisfying (15).

We now use the methods of degree theory (cf. Ortega & Rheinboldt (1970) Chapter 6) to prove the following result under the same assumptions as above (all states communicate, $\nu\gg\underline{0}$). Define the matrix $J=J(\mu, Q)$ by

$$J_{ij}=\mathbb{E}_{\mu,Q}\left(\int_0^1 \mathbf{I}\{\mathbf{X}_t=i\}\,\mathrm{d}t\mathbf{I}\{\mathbf{X}_1=j\}\right)=\mathbb{E}_{\mu,Q}(\mathbf{l}_i\nu_j) \tag{19}$$

Then we show that if $J=J(l)=J(\mu, (\text{diag } l)^{-1}N)$ is non-singular for all $l\in S^0$, then the equation $\hat{\nu}(l)=\nu$ has a unique solution $l\in S^0$.

First we note that $-(\text{diag } l)^{-1}JQ$ is the Jacobian matrix of the transformation $\nu:S^0\subset\mathbb{R}^p\to\mathbb{R}^p$. For, denoting by $A_{i\cdot}$ the ith row of the matrix A, we have

$$\begin{aligned}\frac{\partial\hat{\nu}}{\partial l_i}&=\mu\frac{\partial e^Q}{\partial l_i}-\mu\int_0^1 e^{Qs}\frac{\partial Q}{\partial l_i}e^{Q(1-s)}\,\mathrm{d}s\\&=\int_0^1\mu e^{Qs}\left(-\frac{1}{l_i}\right)\begin{pmatrix}0\\Q_{i\cdot}\\0\end{pmatrix}e^{Q(1-s)}\,\mathrm{d}s\\&=-\frac{1}{l_i}\int_0^1\mu e^{Qs}\begin{pmatrix}0\\\{Qe^{Q(1-s)}\}_{i\cdot}\\0\end{pmatrix}\mathrm{d}s\\&=-\frac{1}{l_i}\int_0^1\mu e^{Qs}\begin{pmatrix}0\\\{e^{Q(1-s)}Q\}_{i\cdot}\\0\end{pmatrix}\mathrm{d}s.\end{aligned}$$

The second equality can be verified by substituting the power series representation for e^{Q}, e^{Qs} and $e^{Q(1-s)}$. So

$$\frac{\partial \hat{\nu}_j}{\partial l_i} = -\frac{1}{l_i}\int_0^1 (\mu e^{Qs})_i \, [e^{Q(1-s)}Q]_{ij} \, ds = -\frac{1}{l_i}(JQ)_{ij}$$

Now define

$$S^* = \left\{ x \in \mathbb{R}^{p-1} : x_i \geq 0 \, \forall i, \sum_1^{p-1} x_i \leq 1 \right\}.$$

Define $\tau_i : S^* \to S$ by

$$\tau_i(x) = (x_1, \ldots, x_{i-1}, 1 - \sum_1^{p-1} x_j, x_i, \ldots, x_{p-1}).$$

Note that τ_i^{-1} exists and $\tau_i^{-1}(l) = (l_1, \ldots, l_{i-1}, l_{i+1}, \ldots l_p)$. We can now define mappings $\hat{\nu}^{(i,j)} : S^* \to S^*$ by $\hat{\nu}^{(i,j)} = \tau_j^{-1} \circ \hat{\nu} \circ \tau_i$ (i.e. we drop the ith component of l and the jth of $\hat{\nu}(l)$). Any two such mappings are related by $\hat{\nu}^{(i,j)} = \tau_j^{-1} \circ \tau_n \circ \hat{\nu}^{(m,n)} \circ \tau_m^{-1} \circ \tau_i$ where $\tau_m^{-1} \circ \tau_i$ and $\tau_j^{-1} \circ \tau_n$ are non-singular affine maps from S^* to S^*. So if the Jacobian matrix of any $\hat{\nu}^{(i,j)}$ is singular, they all are.

Now the Jacobian of $\hat{\nu}^{(i,j)}$ is obtained from the Jacobian of $\hat{\nu}$ by subtracting the ith row from all the other rows and then deleting the ith row and jth column (if $l_p = 1 - l_1 - \ldots - l_{p-1}$, then for i, $j < p$, $\partial \hat{\nu}_j^{(p,p)} / \partial l_i = \partial \hat{\nu}_j / \partial l_i - \partial \hat{\nu}_j / \partial l_p$). So if J is non-singular and N has rank $p-1$, then for $l \in S^0$, $-(\text{diag } l)^{-1} JQ$ has rank $p-1$. At least one row is linearly dependent on the others so subtracting such a row from all other rows and then deleting it preserves the rank. Now one column is linearly dependent on the others and may also be deleted without reducing the rank. So if J is non-singular, then for some i, j, $\hat{\nu}^{(i,j)}$ has non-singular Jacobian. Hence all $\hat{\nu}^{(i,j)}$ have non-singular Jacobian.

Next we note that the determinant of the Jacobian of $\hat{\nu}^{(i,j)}$ is a continuous function of $l \in S^0$. So if the Jacobian is non-singular everywhere, its determinant has the same sign everywhere. Consequently if J is non-singular on S^0, then the determinant of the Jacobian of $\hat{\nu}^{(i,j)}$ is non-zero and has the same sign on $E = (S^*)^0$. Pick any (i, j) and let $y = \tau_j^{-1}(\nu)$. We now consider solutions of the equation $\hat{\nu}^{(i,j)}(x) = y$, $x \in S^*$. Under the condition $\nu_i > 0$ for all i there are no solutions on the boundary of S^*. Define $H : S^* \times [0, 1] \to S^*$ by $H(x, t) = (1-t)\tau_j^{-1} \circ \tau_i(x) + t\hat{\nu}^{(i,j)}(x)$. Note that $y \in E = (S^*)^0$. Now the equation $H(x, t) = y$ also has no solutions on $\partial S^* \times [0, 1]$ since for $x \in \partial S^*$, $H(x, t) \in \partial S^*$. By continuity and compactness there also exist no solutions in

$$\left\{ x \in S^* : x_i \leq \delta \quad \text{for some } i \text{ or } \sum_1^{p-1} x_i \geq 1 - \delta \right\} \times [0, 1]$$

for some $\delta > 0$, where of course $p\delta < 1$. Let

$$C = \left\{ x \in S^* : x_i > \delta \, \forall i \quad \text{and} \quad \sum_1^{p-1} x_i < 1 - \delta \right\}.$$

We now have the following facts. The set $E \subset \mathbb{R}^{p-1}$ is open and bounded. The function $\hat{\nu}^{(i,j)} : E \to E$ is continuously differentiable on E. The set C is also open, $\bar{C} \subset E$ and $H : \bar{C} \times [0, 1] \to E$ defined as above is such that $H(x, t) = y$ has no solution on $\partial C \times [0, 1]$. By the

Homotopy invariance theorem (cf. Ortega & Rheinboldt (1970), §6.2.2, p. 156) we have deg $\{H(\cdot, t), C, y\}$ is constant for $t\in[0, 1]$. Now $H(\cdot, 0)=\tau_j^{-1}\circ\tau_i$ and $H(\cdot, 1)=\hat{\nu}^{(i,j)}$. Moreover for a continously differentiable function $F: E\to\mathbb{R}^{p-1}$ with Jacobian matrix F' which is non-singular at all solutions in C of $F(x)=y$ and which has no solutions on ∂C,

$$\deg(F, C, y)=\sum_{x\in C: F(x)=y} \operatorname{sign}\det F'(x)$$

Also $y\in C$ so $\tau_j^{-1}\circ\tau_i(x)=y$ has a unique solution and deg $\{H(\cdot, t), C, y\}=\pm 1$ for all t. Therefore $\hat{\nu}^{(i,j)}(x)=y$ also has exactly one solution in C, which is what we needed to prove.

We do not know whether the condition on J holds in any generality, and can only use this result to prove uniqueness of a solution in the case $p=2$(!). In this case, with $q_1=-q_{11}>0$ and $q_2=-q_{22}>0$, we have

$$e^{Qt}=\begin{bmatrix} \frac{q_2}{q_1+q_2}+\frac{q_1}{q_1+q_2}e^{-(q_1+q_2)t} & \frac{q_1}{q_1+q_2}-\frac{q_1}{q_1+q_2}e^{-(q_1+q_2)t} \\ \frac{q_2}{q_1+q_2}-\frac{q_2}{q_1+q_2}e^{-(q_1+q_2)t} & \frac{q_1}{q_1+q_2}+\frac{q_2}{q_1+q_2}e^{-(q_1+q_2)t} \end{bmatrix}$$

Now letting $\mathbf{U}$ denote a uniformly distributed random variable on the interval [0, 1] which is independent of the process $\mathbf{X}$, we see that the matrix J contains as elements the probabilities $\mathbb{P}(\mathbf{X}_U=i, \mathbf{X}_1=j)$. In the case $p=2$, singularity of J is equivalent to independence of the random variables $\mathbf{X}_U$ and $\mathbf{X}_1$. Now from the expression for e^{Qt} we see that $\mathbb{P}(\mathbf{X}_1=1\,|\,\mathbf{X}_u=1)$ is a strictly increasing function of $u\in[0, 1]$ and moreover this quantity is strictly larger than $\mathbb{P}(\mathbf{X}_1=1)$ for all $u>0$ (whatever μ). Hence $\mathbb{P}(\mathbf{X}_1=1\,|\,\mathbf{X}_U=1)>\mathbb{P}(\mathbf{X}_1=1)$ and $\mathbf{X}_U$ and $\mathbf{X}_1$ are not independent.

In one other case in which we can prove uniqueness of the solution by other means, J is also non-singular, though the case is not covered by the assumption above. This is the case of a *hierarchical* process, when (after a relabelling of states) we have that i does not have access to j if $i>j$. So N has under-diagonal part identically zero. In this case J also has under-diagonal zero, and positive elements on the diagonal if all n_{ii} (except for $i=p$) are non-zero. In the equation $\hat{\nu}(l)_i=\nu_i$ only $l_1, \ldots, l_i$ enter. Suppose $l_1, \ldots, l_{i-1}>0$ are such that $\hat{\nu}(l)_j=\nu_j$ for $j<i$. As l_i varies from 0 up to $1-(l_1+\ldots+l_{i-1})$, $\hat{\nu}(l)_i$ strictly increases from 0 up to some value. So either there is a unique value of l_i with $\hat{\nu}(l)_i=\nu_i$ or none at all. By an induction argument there is either one solution to $\hat{\nu}(l)=\nu$ or none. By the existence result, there is exactly one solution. These are the only presently available results on uniqueness. (Except for the following: if there is a unique solution, with non-singular J, at $(\mu, N)=(\mu_0, N_0)$, then there is a unique solution in a neighbourhood of (μ_0, N_0).) Another fixed point theorem is used by Johansen (1973, proposition 2.3) in a rather similar context: the embedding problem for stochastic matrices.

On the other major problem in this context, convergence of the iterations $l^{(k+1)}=\hat{l}\{l^{(k)}\}$, $k=1, 2, \ldots$ (starting from some initial guess $l^{(1)}$) results are very meagre. Denoting by $\partial\hat{l}/\partial l$ the matrix with (i, j)th element $\partial\hat{l}_j/\partial l_i$, it can be shown quite easily that $\partial\hat{l}/\partial l=-(\operatorname{diag} l)^{-1}(J-\operatorname{diag}\hat{l})$. Since $J(l)\underline{1}^{\mathrm{T}}=\hat{l}(l)^{\mathrm{T}}$, at a fixed point $\partial\hat{l}/\partial l$ equals the identity matrix minus a stochastic matrix. If it could be shown that the spectral radius of $\partial\hat{l}/\partial l$ is less than 1 at a fixed-point, then by the Ostrowski theorem (Ortega & Rheinboldt (1970) §10.1.3, p. 300) we would know that the iterations converge in a neighbourhood of a fixed-point. In the case $p=2$ (I am indebted to the referee for the following observations), we have just shown that this stochastic matrix has a positive determinant. Its eigenvalues are therefore 1 and $\lambda\in[0, 1]$ (real). Hence the spectral radius of $\partial\hat{l}/\partial l$ is $1-\lambda<1$ and we are guaranteed local convergence of the iterations. However, it is not clear whether or not $\partial\hat{l}/\partial l$ has this property in general. Note however that if $\mu\gg\underline{0}$ under and the elements of N are very small in absolute

value, then ν is close to μ and we expect any solution l to be close to both. For l not close to ∂S, J is close to diag $\hat{l}$ and $\partial\hat{l}/\partial l$ is therefore close to 0. So we expect local convergence in this case. Also since J is then non-singular for most l, we expect uniqueness to hold too.

3. Statistical properties of the solution of the estimating equations

In this section we will consider large sample results in the i.i.d. case in which the initial states of the component processes $\mathbf{X}_0^m$, $m=1,\ldots,n$, are independent and identically distributed with distribution μ, and hence the whole processes $\mathbf{X}^m$, $m=1,\ldots,n$, are i.i.d. This makes life easy, though one would really be more interested in conditional large sample results, conditional on $\boldsymbol{\mu}^n=\mu^n$, for some arbitrary sequence of realized initial distributions μ^n, $n=1, 2, \ldots$.

So we work in the i.i.d. case and suppose the processes are generated by a fixed $\mu=\mu_0$ and $Q=Q_0$ such that $l=l_0\in S^0$ and the matrix $J=J_0$ defined by (19) is non-singular. This implies as was shown in section 2 that the Jacobian matrix at (μ_0, N_0) for the mapping (cf. (16)).

$$\phi(l;\mu, N)=\mu \exp\{(\text{diag } l)^{-1}N\}-(\mu+\underline{1}N),\ l\underline{1}^{\mathrm{T}}=1,$$

considered as a function from $(l_1,\ldots,l_{p-1})$, to $(\phi_1,\ldots,\phi_{p-1})$ is non-singular at the solution $l=l_0$ of (16) defined by (10). Of course there may be other solutions of (16), i.e. of $\phi(l;\mu_0, N_0)=\underline{0}$; an (unverifiable) condition for uniqueness was also given in section 2. Thus by the implicit function theorem (see e.g. Ortega & Rheinboldt (1970) §5.2.4) and speaking somewhat informally there exists a neighbourhood of (μ_0, N_0) and a continuously differentiable function l^* defined on the neighbourhood such that $l=l^*(\mu, N)$, is a solution of (16), $l_0=l^*(\mu_0, N_0)$, and moreover, the derivative of l^* with respect to (μ, N) at (μ_0, N_0) is given by $-(\partial\phi/\partial l)^{-1}\{\partial\phi/\partial(\mu, N)\}_{|(\mu_0, N_0)}$. (To make this formally correct, we must first delete superfluous elements of μ, N and l – e.g. the diagonal of N, the last element of μ and l, and any "structural zeros" in N.)

All this gives immediately by the central limit theorem and the δ-method that, if we define $\hat{\mathbf{l}}^n=l^*(n^{-1}\boldsymbol{\mu}^n, n^{-1}\mathbf{N}^n)$ for $(n^{-1}\boldsymbol{\mu}^n, n^{-1}\mathbf{N}^n)$ in the neighbourhood of (μ_0, N_0) (the probability that this is the case converges to 1 as $n\to\infty$), then $n^{1/2}(\hat{\mathbf{l}}^n-l_0)$ is asymptotically multivariate normally distributed with mean zero and with a covariance matrix which can be determined from the derivative of l^* and the covariance matrix of $n^{1/2}\{(n^{-1}\boldsymbol{\mu}^n, n^{-1}\mathbf{N}^n)-(\mu_0, N_0)\}$. Defining $\hat{\mathbf{Q}}^n=(\text{diag } \hat{\mathbf{l}}^n)^{-1}(n^{-1}\mathbf{N}^n)$, the same holds for $n^{1/2}(\hat{\mathbf{Q}}^n-Q_0)$ by a further application of the δ-method. In Gill (1984) the asymptotic distribution of $n^{1/2}\{(n^{-1}\boldsymbol{\mu}^n, n^{-1}\mathbf{N}^n)-(\mu_0, N_0)\}$ is described. See also Funck Jensen (1982a) and her references. So in principle the asymptotic covariance matrix of $n^{1/2}(\hat{\mathbf{Q}}^n-Q_0)$ is determined and can be consistently estimated by substituting $n^{-1}\boldsymbol{\mu}^n$ and $\hat{\mathbf{Q}}^n$ for μ_0, Q_0. To do this in practice will require availability of efficient matrix exponentiation and numerical integration procedures; see especially Moler & van Loan (1978). (We must also assume that the solution at (μ_0, N_0) is unique. The probability then tends to one that the solution at $(n^{-1}\boldsymbol{\mu}^n, n^{-1}\mathbf{N}^n)$ is also unique, so that $\hat{\mathbf{l}}^n$ is the estimator we actually compute.)

We now discuss asymptotic optimality of this estimator at a similar informal level. For notational convenience we shall switch over to the following general setup and first repeat the above arguments. Suppose $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are i.i.d. $\mathbb{R}^p$-valued random vectors with distribution depending on a single parameter $\theta\in\mathbb{R}^p$. Suppose we only observe

$$\bar{\mathbf{X}}_n=n^{-1}\sum_{i=1}^{n}\mathbf{X}_i.$$

Define $\mu(\theta)=\mathbb{E}_\theta(\mathbf{X}_i)$ and $\sigma^2(\theta)=\mathbb{V}\text{ar}_\theta(\mathbf{X}_i)$ (a $p\times p$ matrix) which we both suppose to exist. We shall need that $\mu(\cdot)$ and $\sigma^2(\cdot)$ are continuous, and in fact that $\mu(\cdot)$ is 1–1 and differentiable with a differentiable inverse (the implicit function theorem can sometimes be used to verify this

condition). It is then sensible to consider the method of moments estimator $\hat{\theta}_n$ defined by $\bar{\mathbf{X}}_n=\mu(\hat{\theta}_n)$. Since by the central limit theorem

$$n^{1/2}\{\bar{\mathbf{X}}_n-\mu(\theta)\}\xrightarrow{\mathscr{L}(\theta)}N\{0,\ \sigma^2(\theta)\},$$

we have by the δ-method

$$n^{1/2}(\hat{\theta}_n-\theta)\xrightarrow{\mathscr{L}(\theta)}N[0,\ \{(\partial\mu/\partial\theta)^{-1}\}^{\mathrm{T}}\sigma^2(\theta)(\partial\mu/\partial\theta)^{-1}].$$

(Here $\xrightarrow{\mathscr{L}(\theta)}$ means "converges in distribution under θ".)

In fact $\hat{\theta}_n$ is the only consistent estimator of θ which is a continuous function of $\bar{\mathbf{X}}_n$ only (and does not e.g. also depend on sample size n). Usually the maximum likelihood estimator of θ based on data $\bar{\mathbf{X}}^n$ will also depend on n: it must be asymptotically equivalent to $\hat{\theta}_n$ if it is asymptotically optimal too.

To discuss asymptotic optimality, let us for simplicity consider the case $\mu(\theta)\equiv\theta=\mu$, $p=1$. In the general case exactly the same arguments go through. So we have in $\mathbb{R}^1$ i.i.d. random variables $\mathbf{X}_i$ with

$$n^{1/2}(\bar{\mathbf{X}}_n-\mu)\xrightarrow{\mathscr{L}(\mu)}N\{0,\ \sigma^2(\mu)\}.$$

According to LeCam's (1960) theory of local asymptotic normality (cf. also LeCam (1972) and Hajek (1970, 1972), $\bar{\mathbf{X}}_n$ will have various nice asymptotic local efficiency properties as estimator of μ with data $\bar{\mathbf{X}}_n$ if the log likelihood ratio for two values of μ of order $n^{-1/2}$ apart, based on observation of $\bar{\mathbf{X}}_n$, becomes like the same log likelihood ratio based on the asymptotic distribution of $\bar{\mathbf{X}}_n$. To state this more precisely, let $p_n(x;\ \mu)$ denote the density, with respect to some fixed σ-additive measure, of the distribution of $\bar{\mathbf{X}}_n$ under μ. Then we require for asymptotic optimality that for any number h and any sequence $h_n\to h$ as $n\to\infty$, and any μ_0,

$$\log\left\{\frac{p_n(\bar{\mathbf{X}}_n;\ \mu_0+n^{-1/2}h_n)}{p_n(\bar{\mathbf{X}}_n;\ \mu_0)}\right\}\xrightarrow{\mathscr{L}(\mu_0)}N\left\{-\frac{1}{2}\frac{h^2}{\sigma^2(\mu_0)},\ \frac{h^2}{\sigma^2(\mu_0)}\right\}. \tag{20}$$

To motivate (20), let us consider equivalently for fixed μ_0 the log likelihood ratio for the same pair of parameter values based on data $\mathbf{Y}_n=n^{1/2}(\bar{\mathbf{X}}_n-\mu_0)$. Under $\mu_n=\mu_0+n^{-1/2}h_n$, $\mathbf{Y}_n$ is approximately $N\{h_n,\ \sigma^2(\mu_n)\}$ or approximately $N\{h,\ \sigma^2(\mu_0)\}$ distributed, while under μ_0, $\mathbf{Y}_n$ is approximately $N\{0,\ \sigma^2(\mu_0)\}$ distributed. Writing σ_0^2 for $\sigma^2(\mu_0)$, we would therefore expect the log likelihood ratio at the left hand side of (20) to be approximately equal to

$$\log\left[\frac{(2\pi\sigma_0^2)^{-1/2}\exp\{-(\mathbf{Y}_n-h)^2/2\sigma_0^2\}}{(2\pi\sigma_0^2)^{-1/2}\exp(-\mathbf{Y}_n^2/2\sigma_0^2)}\right]=\frac{h\mathbf{Y}_n}{\sigma_0^2}-\frac{h^2}{2\sigma_0^2}\xrightarrow{\mathscr{L}(\mu_0)}N\left(-\frac{h^2}{2\sigma_0^2},\ \frac{h^2}{\sigma_0^2}\right).$$

So (20) is not such a surprising condition. Looking at the preceding sketch of a derivation of (20), we see that we need continuity of $\sigma^2(\mu)$ as function of μ and moreover that a *local central limit theorem* should hold for $\bar{\mathbf{X}}_n$ *uniformly in* μ close to μ_0; i.e. we must be able to approximate the density of $n^{-1/2}(\bar{\mathbf{X}}_n-\mu)$ by the appropriate normal density, uniformly in μ, uniformly on arbitrarily large portions of the real line. Such uniform local central limit theorems do not hold in general, however they are available in our situation in which the $\mathbf{X}_i$s are lattice random variables and satisfy a uniformly bounded $2+\delta$ moment condition; see e.g. Petrov (1975) Ch. 7.

4. An example

We consider here a small part of the data-set given by Schoen & Nelson (1974) which has recently been used by Nour & Suchindran (1984) to illustrate the occasional breakdown of the

"actuarial formula" $\tilde{P}=(I+\frac{1}{2}\tilde{Q})(I-\frac{1}{2}\tilde{Q})^{-1}$. In fact this example, based largely on the actual marital status patterns for the whole US female population 1960, age group 20–24 (incorrectly described by Nour & Suchindran, 1984) is not a *real* data-set of the type we are interested in: however, it is a *realistic* data-set, describing a hypothetical cohort of 100 000 individuals who experience at each age of their life the same risks of marriage, divorce, etc., as experienced by the US female population in 1960. Note that the time interval [0, 1] of the previous sections now represents the *age* interval from the 20th to the 25th birthday of the hypothetical cohort. (In any case, our approach gives a means of interpolating within such life-tables, however they have been constructed.) The actual figures are summarized in Table 1; they are obtained from the multi-state life tables of Schoen & Nelson (1974). Some rounding errors have been resolved arbitrarily. The underlying model is described by Fig. 2. The reader is invited to draw his own conclusions on the relative risks of death, (re)marriage, etc., in the various states from the raw data μ^n and N^n, $n=100\,000$, before studying the various estimates of the transition matrix P in Table 2.

We present three different estimates of the transition matrix P, writing $\mu=n^{-1}\mu^n$, etc., namely the "actuarial-solution" $\tilde{P}=(I+\frac{1}{2}\tilde{Q})(I-\frac{1}{2}\tilde{Q})^{-1}$ where $\tilde{Q}=(\text{diag } \bar{l})^{-1}N$ and $\bar{l}=\frac{1}{2}(\mu+\nu)$; the "approximate statistical solution" $\tilde{\tilde{P}}=e^{\tilde{Q}}$; and the "exact statistical solution" $\hat{P}=e^{\hat{Q}}$ where $\hat{Q}=(\text{diag } \hat{l})^{-1}N$ and $\hat{l}$ is the solution (as far as we know unique, but this is not proven) of the equations $\nu=\mu e^{\hat{Q}}$, $\hat{l}\underline{1}^{\mathrm{T}}=1$. We also display $\bar{l}$ and $\hat{l}$ (or rather $\bar{l}^n=n\bar{l}$, $\hat{l}^n=n\hat{l}$)

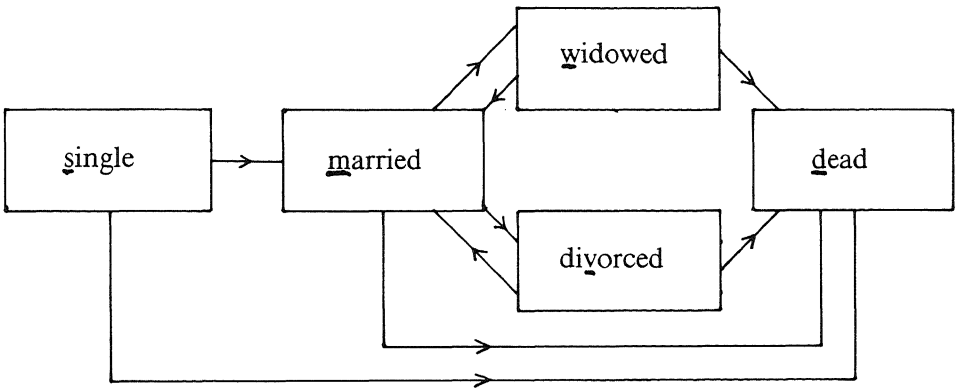

Fig. 2. Model (US females 1960 age 20–24 (t); $n=100\,000$).

Table 1. *Data*

	s	m	w	v	d	
Initial distribution						
	54177	41955	59	544	3265	μ^n
Moves						
s	−40176	40043	0	0	133	N^n
m	0	−6537	373	5971	193	
w	0	146	−148	0	2	
v	0	4009	0	−4021	12	
d	0	0	0	0	0	
Final distribution	14001	79616	284	2494	3605	ν^n

Table 2. *Solutions.*

	s	m	w	v	d	
Approximate exposures	34089	60785.5	171.5	1519.0	3435	$\tilde{l}^n$
Exact exposures	29691.2	65057.7	170.7	1641.3	3439.1	$\hat{l}^n$
Actuarial solution						
s	0.2584	0.7211	0.0015	0.0152	0.0037	$\tilde{P}$
m	0	0.9513	0.0042	0.0412	0.0033	
w	0	0.5802	0.3984	0.0123	0.0091	
v	0	1.1082	0.0024	−0.1158	0.0053	
d	0	0	0	0	1	
Approximate statistical solution						
s	0.3077	0.6687	0.0018	0.0181	0.0035	$\tilde{\tilde{P}}$
m	0	0.9594	0.0040	0.0333	0.0033	
w	0	0.5532	0.4234	0.0146	0.0089	
v	0	0.8948	0.0028	0.0975	0.0049	
d	0	0	0	0	1	
Exact statistical solution						
s	0.2584	0.7166	0.0019	0.0193	0.0038	$\hat{P}$
m	0	0.9601	0.0037	0.0331	0.0030	
w	0	0.5553	0.4216	0.0143	0.0083	
v	0	0.8810	0.0026	0.1118	0.0046	
d	0	0	0	0	1	
	s	m	w	v	d	

from which $\tilde{Q}$ and $\hat{Q}$ can be easily constructed. (Many authors give the formula $\tilde{P}=(I-\frac{1}{2}\tilde{Q})^{-1}(I+\frac{1}{2}\tilde{Q})$; fortunately the members of the product commute.)

Note that both $\tilde{P}$ and $\hat{P}$ fit the data exactly ($\tilde{\tilde{P}}$ does not) in the sense that $\mu\hat{P}=\mu\tilde{P}=\nu$. This dataset illustrates the anomaly that $\tilde{P}$ is not necessarily a stochastic matrix: it can include estimated probabilities smaller than zero or larger than one. A sufficient condition for $\tilde{P}$ to be well-behaved is $\tilde{q}_{ii}\geq -2$ for each i; this condition fails in this case. The formula for $\tilde{P}$ was derived by Rogers & Ledent (1976) under the condition (at a superficial reading of their paper) that the events of each type (each type of move) occur uniformly distributed in time over the time-interval [0, 1]. However, at a closer reading they need two strong assumptions, whose mutual consistency is not at all evident: for each type of move (from i to j, $i\neq j$), for each initial state subpopulation, moves occur uniformly distributed in time *and* the occurrence–exposure rate (computed over the whole time interval) does not depend on the initial state. Surprisingly it can be shown that these assumptions are mutually consistent and consistent with a particular μ and N if and only if $\tilde{P}$ is a stochastic matrix.

Finally we remark that, since this is only hypothetical data, an estimate of the covariance structure of the "exact" statistical estimators $\hat{\mathbf{Q}}^n$ or $\hat{\mathbf{P}}^n$ is not very meaningful. In fact we have not yet gone to the trouble of deriving explicitly the formulas for this mentioned in section 3, which will be extremely complicated. A useful practical solution is to use for $\hat{\mathbf{Q}}^n$ the estimated covariance structure for the occurrence–exposure rates *applicable when the exposures l^n are observed too.* This gives a lower bound to the asymptotic covariance matrix of the estimator actually used; i.e. our recommendation is to use the off-diagonal elements of $(\operatorname{diag}\hat{\mathbf{l}}^n)^{-2}\mathbf{N}^n$ as a lower bound to, and rough estimate of, the variances of the corresponding elements of $\hat{\mathbf{Q}}^n$.

Appendix I

Rank $(Q:\underline{1}^T)=p \Leftrightarrow$ rank $(Q)=p-1 \Leftrightarrow$ *there exists a state to which all states have access.*

References here are to Berman & Plemmons (1979) Chapter 6 "M-matrices", also some of the notation is theirs.

Suppose there exists a state to which all states have access. Consider the matrix A obtained by deleting the row and column from $-Q$ corresponding to the state i_0 in question. Then we have $A \in Z^{(p-1)\times(p-1)}$ (cf. definition on page 132). Taking x to be the column vector of $p-1$ 1's, we have that x satisfies the conditions L_{32} of theorem 2.3 (pp. 134, 136). Therefore A is a non-singular M-matrix and in particular rank $(A)=p-1$ so rank $(Q)=p-1$ too. We show that no column vector x exists with $(-Q)x=\underline{1}^T$. Let I be the (non-empty) class of states which communicate with i_0. So (after a relabelling of states) we can write

$$Q=\begin{bmatrix} F & G \\ 0 & Q_I \end{bmatrix}$$

where Q_I is the intensity matrix for the states I. Also Q_I is irreducible. Now, in obvious notation, $(-Q)x=\underline{1}^T \Rightarrow (-Q_I)x_I=\underline{1}_I^T$. So it suffices to consider the case of an irreducible intensity matrix, which we will take to be Q itself. Since $(-Q)\in Z^{p\times p}$ and $(-Q)\,\underline{1}^T=\underline{0}^T$, by exercise 4.14 (p. 155) we have that $-Q$ is a singular M-matrix of rank $p-1$ with "property C". But then by theorem 4.16 (5) (p.156), $(-Q)x \geqslant \underline{0}^T \Rightarrow (-Q)x=\underline{0}^T$. So $(-Q)x=\underline{1}^T$ is impossible.

Conversely, suppose there does not exist a state to which all other states have access. Then Q contains at least two disjoint absorbing subsets of states; i.e. we can write (after a relabelling of states)

$$Q=\begin{bmatrix} E & F & G \\ 0 & Q_I & 0 \\ 0 & 0 & Q_J \end{bmatrix}$$

Now both Q_I and Q_J are singular (row sums are zero) so rank $(Q)\leqslant p-2$. Therefore rank $(Q:\underline{1}^T)\leqslant p-1$.

More generally, suppose there exist r and no more than r disjoint absorbing subsets of states. Then we can write

$$Q=\begin{bmatrix} E & F & G & \cdot & \\ 0 & Q^{(1)} & 0 & & \\ 0 & 0 & Q^{(2)} & & \\ & & \ddots & \ddots & \\ & & & & Q^{(r)} \end{bmatrix}$$

where E has full rank (apply to $-E$ the same argument as was applied to A above) and each $Q^{(i)}$ has rank one less than its dimension. So rank $(Q)=p-r$.

Appendix II. Existence of a solution in the case rank $(N)=p-1$, $\nu \geqslant \underline{0}$

We are given $\mu \in \mathbb{R}^p$ (row vector), $N \in \mathbb{R}^{p\times p}$, and $\nu=\mu+\underline{1}N$ satisfying $\mu \geqslant \underline{0}$, $\mu\underline{1}^T=1$, $N\underline{1}^T=\underline{0}^T$, $n_{ij} \geqslant 0$ for all $i \neq j$, rank $(N)=p-1$. Recall that $S=\{l \in \mathbb{R}^p:\ \{l \geqslant \underline{0},\ l\underline{1}^T=1\}$; $S^0=\{l \in \mathbb{R}^p : l \geqslant \underline{0},\ l\underline{1}^T=1\}$.

We show that there exists $l \in S^0$ such that $\hat{l}(l)=l$ or equivalently (thanks to the rank condition) $\hat{\nu}(l)=\nu$. We build step by step on the result and method of proof given in section 2.

Case 1

If all states communicate we know $\exists l \in S^0$ s.t.

$$l = \hat{l}(l) = \hat{l}(l; \mu, N)$$

$$\nu = \hat{\nu}(l) = \hat{\nu}(l; \mu, N)$$

Case 2

Next suppose all states but one (the pth say) communicate and have access to the pth, absorbing state. Choose $\delta^{(n)} \to 0$ as $n \to \infty$, $\nu_p > \delta^{(n)} > 0$, and define $\mu^{(n)} = \mu$ and

$$N^{(n)} = N + \begin{pmatrix} & & & & 0 \\ \delta^n & 0 & \cdots & 0 & -\delta^n \end{pmatrix}$$

The problem $(\mu^{(n)}, N^{(n)})$ has $\nu^{(n)} \gneqq \underline{0}$ and all states communicating so there exists a solution $l^{(n)} \in S^0$. From this sequence we can select a subsequence (which we shall take to be $(l^{(n)})$ itself) along which $l^{(n)} \to l \in S$ and $\alpha^{(n)} = l_p^{(n)}/\delta^{(n)} \to \alpha \in [0, \infty]$. We shall first show that

$$\nu \leftarrow \nu^{(n)} = \hat{\nu}(l^{(n)}; \mu^{(n)}, N^{(n)}) \to \hat{\nu}(l, \mu, N)$$

and

$$l \leftarrow l^{(n)} = \hat{l}(l^{(n)}; \mu^{(n)}, N^{(n)}) \to \hat{l}(l; \mu, N)$$

(only the right hand convergences need to be verified; recall that $\hat{\nu}(\cdot; \mu, N)$ and $\hat{l}(\cdot; \mu, N)$ are defined on S since for this problem every state has access to all other states or is absorbing). Obviously if we knew $l \in S^0$ or that $\hat{l}(\cdot; \cdot, \cdot)$ and $\hat{\nu}(\cdot; \cdot, \cdot)$ were continuous in all three arguments jointly at a point with $l \in \partial S$ we would be ready. However, neither of these hypotheses is *a priori* true. Now define a process $\mathbf{X}_t^{(n)}$=state at time $t+$, $t \in (0, \infty)$ by constructing:

—a discrete time Markov chain on $\{1, \ldots, p\}$ with initial distribution μ and with transition probabilities

$$\begin{cases} n_{ij}/(-n_{ii}) & i \neq j,\ i < p \\ 1 & (i, j) = (p, 1) \\ 0 & \text{otherwise;} \end{cases}$$

—independently, for each i, an infinite sequence of independent exponentially distributed random variables with parameter

$$\begin{cases} (-n_{ii}) & i < p \\ 1 & i = p; \end{cases}$$

we then obtain $\mathbf{X}_t^{(n)}$ as the process whose initial state and jumps are given by the Markov chain and whose jump times, in each state i, are given by

$$\begin{cases} l_i^{(n)} & i < p \\ \alpha^{(n)} = l_p^{(n)}/\delta^{(n)} & i = p \end{cases}$$

times the random variables in the ith sequence of exponentially distributed r.v.s, taken in sequence.

For each n this results in a homogeneous Markov process with parameters $(\mu^{(n)}, Q^{(n)}) = \{\mu, (\text{diag } l^{(n)})^{-1} N^{(n)}\}$, expected exposures $l^{(n)}$ and expected occurrences $N^{(n)}$.

We define

$$\mathbf{X}_t^{(\infty)}=\lim_{h\downarrow 0}\lim_{n\to\infty}\mathbf{X}_{t+h}^{(n)}$$

which we claim exists for all $t\in[0, 1]$ almost surely. After checking that, we check that $\{\mathbf{X}_t^{(\infty)}: t\in[0, 1]\}$ is the process $(\mathbf{X}_t: t\in[0, 1])$ by means of which $\hat{l}(l; \mu, N)$ and $\hat{\nu}(l; \mu, N)$ are defined; i.e. homogeneous Markov with parameter $(\mu, Q)=\{\mu, (\text{diag } l)^{-1}N\}$ where $l\in\partial S$ is allowed. Then finally we check $l\in\partial S$ is impossible.

Note first that since all states communicate, almost surely the Markov *chain* visits a state i with $l_i>0$ infinitely often. Suppose first there exists such a state with $i<p$. Since the partial sums of the ith sequence of exponentials converge almost surely to infinity, it follows that the processes $\mathbf{X}_t^{(n)}$, $n\leqslant\infty$ are indeed well defined. If $l_i=0$ for all $i<p$ then $l_p=1$ and $\alpha=\infty$ and again the processes are well defined, in particular for $n=\infty$. So we have by almost sure convergence of the bounded random variables

$$\mathbf{l}_j^{(n)}=\int_0^1 \mathbf{I}\{\mathbf{X}_t^{(n)}=j\}\,\mathrm{d}t \qquad n\leqslant\infty$$

and

$$\boldsymbol{\nu}_j^{(n)}=\mathbf{I}\{\mathbf{X}_1^{(n)}=j\} \qquad n\leqslant\infty$$

that

$$l=\mathbb{E}\{\mathbf{l}^{(\infty)}\}$$

and

$$\nu=\mathbb{E}\{\boldsymbol{\nu}^{(\infty)}\}.$$

Now if $\alpha=\infty$ the processes $\mathbf{X}_t^{(\infty)}$ and $\mathbf{X}_t$, $t\in[0, 1]$, are the same (from which follows the required result $l=\hat{l}(l; \mu, N)$). However if $\alpha<\infty$ there will be (for $\mathrm{X}_t^{(\infty)}$) with positive probability a positive number of jumps in the time interval $[0, 1]$ from state p back to state 1. Now this number of jumps for the process $\mathbf{X}_t^{(n)}$ converges almost surely to the same number for $\mathbf{X}_t^{(\infty)}$. Its expectation for each $n<\infty$ is $\delta^{(n)}\to 0$ as $n\to\infty$. Hence the number converges in probability to zero as $n\to\infty$; hence the number of jumps for $\mathbf{X}_t^{(\infty)}$ is almost surely zero. Thus we do indeed have $\alpha=\infty$ and hence

$$l=\hat{l}(l; \mu, N)$$

$$\nu=\hat{\nu}(l; \mu, N).$$

Finally we show that $l\in S^0$. Suppose first that $l_i=0$ for some $i<p$. Then we would have $\hat{\nu}(l)_i=0$, a contradiction. On the other hand since $\alpha>0$ we must have $\hat{l}(l)_p>0$, so $l_p=0$ is impossible too.

We have now finished with the case that all states but one communicate and the exceptional state is absorbing and accessible from the others.

Case 3

Consider next the case in which $\{1, \ldots, p\}$ is partitioned into non-empty subsets $\mathscr{C}_0$ and $\mathscr{C}_1$ where $\mathscr{C}_1$ is a communicating class of at least two states; and no state in $\mathscr{C}_1$ has access to a state in $\mathscr{C}_0$. We make no use of the conditions rank $(N)=p-1$ till the very last step.

Let M^* be the $p\times(r+1)$ matrix which collapses all states in $\mathscr{C}_1$ to a single state (here r is the

number of states in $\mathscr{C}_0$); so if $\mathscr{C}_0$ consists of the states $1,\ldots,r$ we have

$$m_{ij}^{*}=\begin{cases}1 & \text{if } i=j \text{ or } i>r,\ j=r+1\\ 0 & \text{otherwise}\end{cases}$$

We also denote by a * all quantities for the collapsed process; e.g. $\mu^{*}=\mu M^{*}$, $N^{*}=M^{*\mathrm{T}}NM^{*}$, etc. Suppose $l^{*}\gg\underline{0}^{*}$ is a solution for the collapsed process; i.e.

$$l^{*}=\hat{l}^{*}(l^{*};\mu^{*},N^{*})$$

$$\nu^{*}=\hat{\nu}^{*}(l^{*};\mu^{*},N^{*}).$$

Consider the problem of finding $l\in S$ such that

$$lM^{*}=l^{*};$$

$$\hat{\nu}(l;\mu,N)=\nu.$$

Note that $\hat{l}(l;\mu,N)$ and $\hat{\nu}(l;\mu,N)$ *are* defined for all $l\in S$ (not just S^{0}) with $lM^{*}=l^{*}$ by the same construction as before since $l^{*}\gg\underline{0}^{*}$ implies that for each $l\in S$ there exists $i\in\mathscr{C}_1$ such that $l_i>0$; therefore once in $\mathscr{C}_1$ one always reaches (with probability one) infinitely often a state with $l_i>0$. Moreover, $\hat{l}$ and $\hat{\nu}$ are continuous functions of $l\in S$, l^{*} fixed.

Now we apply the K-K-M lemma just as before to the lower dimensional simplex $\{l\in S,\ lM^{*}=l^{*}$ fixed$\}$. Since $\{\hat{\nu}(l)\}^{*}=\nu^{*}$ only depends on l through l^{*}, exactly the same argument goes through, giving an $l\in S^{0}$ such that $\hat{\nu}(l;\mu,N)=\nu$. Under the full rank condition rank $(N)=p-1$ this l also satisfies $\hat{l}(l;\mu,N)=l$

Case 4

Next we consider the case $\{1,\ldots,p\}=\mathscr{C}_0\cup\mathscr{C}_1\cup\{p\}$, a partition of the state space into three classes, such that $\mathscr{C}_0$ and $\mathscr{C}_1$ are as before (only $\mathscr{C}_1$ has at least *one* state, not at least *two* states), state p is absorbing and accessible from $\mathscr{C}_1$ (otherwise we would have rank $N<p-1$). Now we combine the proofs of the two previous cases to show that: let $l^{*}\gg\underline{0}^{*}$ satisfy

$$l^{*}=\hat{l}^{*}(l^{*};\mu^{*},N^{*})$$

$$\nu^{*}=\hat{\nu}^{*}(l^{*};\mu^{*},N^{*})$$

where * denotes the problem with all states in $\mathscr{C}_1\cup\{p\}$ collapsed to a single state. Then there exists $l\gg\underline{0}$ (with $lM^{*}=l^{*}$) such that

$$l=\hat{l}(l;\mu,N)$$

$$\nu=\hat{\nu}(l;\mu,N)$$

Case 5

Finally we suppose that we can partition the state space into communicating classes $\mathscr{C}_0,\mathscr{C}_1,\ldots,\mathscr{C}_r$ such that each $\mathscr{C}_i$ does not have access to $\mathscr{C}_j$, $j<i$, but does have access to some $\mathscr{C}_j$, $j>i$ (except that $\mathscr{C}_r$ is absorbing). Under the rank condition rank $(N)=p-1$ such a decomposition is always possible (see e.g. Funck Jensen, 1982b). First we solve for $\mathscr{C}_0$ with $\mathscr{C}_1,\ldots,\mathscr{C}_r$ collapsed into a single state (using case 1 if $r=0$; case 2 if $r>0$); then (supposing now $r\geqslant 1$ and, if $r=1$, $\mathscr{C}_1$ has more than one state) we solve for $\mathscr{C}_1$ with $\mathscr{C}_0$ already solved and $\mathscr{C}_2,\ldots,\mathscr{C}_r$ collapsed into a single state (using case 3 if $r=1$ and case 4 if $r>1$); etc.

Appendix III. Existence of a solution in the general case $\nu \gg \underline{0}$

We now extend the previous result to the case rank $(N)<p-1$, retaining the assumption $\nu \gg \underline{0}$. We note that rank $(Q)=p-r$ (for an intensity matrix Q) if and only if there exist r, and not more than r, disjoint absorbing subsets of states; $1 \leqslant r \leqslant p$. Clearly many of the previous arguments go through. The real problem arises at an early stage of the argument: we no longer have for $l \in S^0$ the equivalence of

$$l=\hat{l}(l; \mu, N)$$

with

$$\nu=\hat{\nu}(l; \mu, N).$$

In particular if there are several absorbing states then $\hat{l}(l)$ and $\hat{\nu}(l)$ do not vary with the values of l_i, i an absorbing state. We can only show that $l=\hat{l}(l)$ implies $\nu=\hat{\nu}(l)$, not the reverse implication. (This result is almost trivial: if $l \in S^0$ satisfies $l=\hat{l}(l; \mu, N)$ then we know that $\mu=\mathbb{E}_{\mu, Q}(\boldsymbol{\mu})$, $N=\mathbb{E}_{\mu, Q}(\mathbf{N})$ where $Q=(\operatorname{diag} l)^{-1}N$. Hence by linearity and the fact $\nu=\mu+\underline{1}\, N$ we find

$$\nu=\mathbb{E}_{\mu, Q}(\boldsymbol{\mu}+\underline{1}\mathbf{N})=\mathbb{E}_{\mu, Q}(\boldsymbol{\nu})=\hat{\nu}(\mathbf{l}).)$$

We shall prove the following theorem:

Theorem

For any μ, N with $\nu=\mu+\underline{1}N \gg \underline{0}$ there exists $l \in S^0$ such that $l=\hat{l}(l; \mu, N)$. Any such l also satisfies $\nu=\hat{\nu}(l; \mu, N)$.

First we state and prove a lemma which shows how $l=\mathbb{E}_{\mu, Q}(\mathbf{l})$ may be computed for general Q (i.e. not necessarily of rank $p-1$) by solving linear equations, analogously to the result (for rank $(Q)=p-1$) "$l=\mathbb{E}_{\mu, Q}(\mathbf{l})$ if and only if $lQ=\mu(e^Q-I)=\nu-\mu$, $l\underline{1}^{\mathrm{T}}=1$".

Lemma

Let μ and Q be a given initial distribution and intensity matrix respectively. Let $\nu=\mu e^Q$. Suppose rank $(Q)=p-r$ so there exist r disjoint absorbing subsets of states, not more, $1 \leqslant r \leqslant p$. Let Q^* be the intensity matrix obtained by collapsing each of these subsets to single states, and Q^{**} be that obtained when these states are further collapsed to one single state. Define μ^*, μ^{**}, etc. analogously. Let M^*, M^{**} be the matrices which perform these successive collapsing operations (so $\mu^*=\mu M^*$, $\mu^{**}=\mu^* M^{**}$, etc.). Then $l=\mathbb{E}_{\mu, Q}(\mathbf{l})$ iff

Step 1: $l^{**}Q^{**}=\nu^{**}-\mu^{**}$, $l^{**}\underline{1}^{**\mathrm{T}}=1$ (defines l^{**})
Step 2: $m^{**}Q^{**}=l^{**}-\mu^{**}$, $m^{**}\underline{1}^{**\mathrm{T}}=\frac{1}{2}$ (defines m^{**} given l^{**})

$l^*=\mu^*+m^*Q^*$ where m^* is any solution of $m^{**}=m^*M^{**}$ (defines l^* given m^{**})

Step 3: $lQ=\nu-\mu$ (defines l given l^*).

Proof of the lemma. Step 1: Since rank $(Q^{**})=\dim(Q^{**})-1$ this is already proved (for the ** process all states have access to a single state).

Step 2: Let

$$m_t=\mathbb{E}_{\mu, Q}\left(\int_{t=0}^{1}\int_{s=0}^{t}\mathbf{I}\{\mathbf{X}_s=t\}\,\mathrm{d}s\,\mathrm{d}t\right):$$

so

$$m=\int_{t=0}^{1}\int_{s=0}^{t}\phi_s\,\mathrm{d}s\,\mathrm{d}t$$

where $(\phi_s)_i=\mathbb{P}_{\mu,Q}(\mathbf{X}_s=i)=(\mu e^{Qs})_i$.

We have

$$m\underline{1}^{\mathrm{T}}=\int_{t=0}^{1}\int_{s=0}^{t}\mathrm{d}s\,\mathrm{d}t=\tfrac{1}{2}$$

and

$$mQ=\int_{t=0}^{1}\int_{s=0}^{t}\mu e^{Qs}Q\,\mathrm{d}s\,\mathrm{d}t$$

$$=\int_{s=0}^{1}(1-s)\mu e^{Qs}Q\,\mathrm{d}s$$

$$=\{(1-s)\mu e^{Qs}\}_0^1+\int_0^1\mu e^{Qs}\,\mathrm{d}s$$

$$=-\mu+l=l-\mu.$$

In particular this general result applies to the two collapsed processes $\mathbf{X}_t^*$ and $\mathbf{X}_t^{**}$ yielding

$$m^*Q^*=l^*-\mu^*,\ m^*\underline{1}^{*\mathrm{T}}=\tfrac{1}{2}$$

and

$$m^{**}Q^{**}=l^{**}-\mu^{**},\ m^{**}\underline{1}^{**\mathrm{T}}=\tfrac{1}{2}.$$

Since rank $(Q^{**}:\underline{1}^{**\mathrm{T}})=\dim(Q^{**})$ the second pair of equations here defines m^{**} uniquely given l^{**} and μ^{**}. Next we note that each row of Q^* corresponding to one of the r absorbing states of the process is identically zero, so m^*Q^* only depends on m^* via the components it has in common with m^{**}. So given m^{**}, we can compute $l^*=\mu^*+m^*Q^*$.

Step 3: Write $l=(l^0\ l^1\ldots l^r)$ partitioned according to the r absorbing subsets of states (superscript $1,\ldots,r$) and the remaining states (superscript 0). Partition Q, etc., similarly. Each of the absorbing subsets of states does not contain two or more disjoint absorbing subsubsets or equivalently each state in the subset has access to one particular state in the subset. Since l^* is given, we already know the elements of l^0 and the value of $l^i\underline{1}^{i\mathrm{T}}$, $i=1,\ldots r$. Now

$$Q=\begin{bmatrix} Q^{00} & Q^{01} & \cdots & Q^{0r} \\ 0 & Q^{11} & & 0 \\ & & \ddots & \\ 0 & 0 & & Q^{rr} \end{bmatrix}$$

So from the equation $lQ=\nu-\mu$ we obtain, for $i=1,\ldots,r$,

$$l^0Q^{0i}+l^iQ^{ii}=\nu^i-\mu^i.$$

Since l^0 and $l^i\underline{1}^{i\mathrm{T}}$ is given, and

$$\text{rank }(Q^{ii}:\underline{1}^{i\mathrm{T}})=\dim(Q^{ii}),$$

we can solve for l^i.

Proof of the theorem. Given rank $(N)=p-r$ partition the state space $\{1,\ldots,p\}$ into $\mathscr{C}_0$, $\mathscr{C}_1,\ldots,\mathscr{C}_r$ where each $\mathscr{C}_i$, $i\geqslant 1$, is a communicating, absorbing subset of states ($\mathscr{C}_0$ may be empty). From any state in $\mathscr{C}_0$ one has access to some state in

$$\bigcup_1^r \mathscr{C}_i.$$

Denote by a * and a ** the systems obtained when each $\mathscr{C}_i$ is collapsed into an absorbing state, and when these absorbing states are further collapsed into a single absorbing state, respectively. Since rank $(N^{**})=\dim(N^{**})-1$ and $\nu^{**}\geqslant\underline{0}^{**}$ there exists a solution $l^{**}\geqslant 0^{**}$ to the ** problem. For the * problem we can now compute $Q^*=(\operatorname{diag} l^*)^{-1}N^*$ since each row of N^* corresponding to one of the absorbing states (for which the corresponding component of l^* is unknown) is zero. For each i separately for which $\mathscr{C}_i$ consists of two or more states, we now apply the result of "Case 3", Appendix II, taking, for $\mathscr{C}_0$ and $\mathscr{C}_1$ *there*, $\mathscr{C}_0$ together with the collapsed $\mathscr{C}_j$, $j\neq i$, and $\mathscr{C}_i$, respectively. This shows the existence of an $l\geqslant\underline{0}$ for *this* new problem. Now we can piece together the r solutions to obtain a "solution" l to the whole problem; this is a solution in the sense that it satisfies $\hat{\nu}(l;\mu,N)=\nu$. Define $Q=(\operatorname{diag} l)^{-1}N$. We now verify that l satisfies the conditions of steps 1, 2 and 3 of the previous lemma. Note to begin with that we do have $\nu=\mu e^Q$, as required by the lemma. Now l^{**} and l^* do satisfy the relations in "step 1" and "step 2" of the lemma, by the very construction of l. From the equality $\mu e^Q=\nu=\mu+\underline{1}N=\mu+lQ$ we obtain the condition of "step 3", $lQ=\nu-\mu$, and the result is proved.

Acknowledgement

This paper owes much to many stimulating discussions with Jan Hoem, Nico Keilman, Frans Willekens, and many colleagues at CWI. I am also very grateful to the referee for his careful reading of the paper and valuable comments.

References

Aalen, O. O. (1978). Nonparametric estimation for a family of counting processes. *Ann. Statist.* **6,** 701–725.

Albert, A. (1962). Estimating the infinitesimal generator of a continuous time, finite state Markov process. *Ann. Math. Statist.* **33,** 727–753.

Berge, C. (1959). *Espaces topologiques, functions multivoques.* Dunod, Paris; translated (1963) as *Topological spaces,* Macmillan, New York.

Berman, A. & Plemmons, R. J. (1979). *Nonnegative matrices in the mathematical sciences.* Academic Press, New York.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. (B)* **39,** 1–38.

Freidenfelds, J. (1974). A set intersection theorem and applications. *Math. Prog.* **7,** 199–211.

Funck Jensen, U. (1982a). An elementary derivation of moment formulas for numbers of transitions in time-continuous Markov chains. Research Rep. 7, Section of Demography, University of Stockholm.

Funck Jensen, U. (1982b). The Feller–Kolmogorov differential equation and the state hierarchy present in models in demography and related fields. Research Rep. 9, Section of Demography, University of Stockholm.

Gill, R. D. (1984). A note on two papers in central limit theory. Report MS-R8410, Centrum voor Wiskunde en Informatica, Amsterdam; also in *Proc. 44th Session ISI, Madrid; Bull. Int. Stat. Inst.* **50** (3), 239–243.

Hajek, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrs. verw. Geb.* **14,** 323–330.

Hajek, J. (1972). Local asymptotic minimax and admissability in estimation. *Proc. 6th Berkeley Symp. Math. Stat. Prob.* **1,** 175–194.

Hoem, J. & Funck Jensen, U. (1982). Multistate life table methodology: a probabilistic critique; In *Multidimensional mathematical demography*. (ed. K.C. Land & A. Rogers) pp. 155–264. Academic Press, New York.

Johansen, S. (1973). The bang-bang problem for stochastic matrices. *Z. Wahrs. verw. Geb.* **26,** 191–195.

Kalbfleisch, J. D., Lawless, J. F. & Vollmer, W. M. (1983). Estimation in Markov models from aggregate data. *Biometrics* **39,** 907–919.

Keilman, N. & Gill, R. D. (1986). On the estimation of multidimensional demographic models with population registration data (in preparation).

Knaster, B., Kuratowski, C. & Mazurkiewicz, S. (1929). Ein Beweis des Fixpunktsatzes für n-dimensionale Simplexen. *Fund. Math.* **14,** 132–137.

van der Laan, G. (1980). Simplicial fixed point algorithms. *MC. Tract* **129,** Mathematical Centre, Amsterdam.

Land, K. C. & Schoen, R. (1982). Statistical methods for Markov generated increment–decrement life tables with polynomial gross flow functions; In *Multidimensional mathematical demography* (ed. K. C. Land & A. Rogers), pp. 265–346.

LeCam, L. (1960). Locally asymptotically normal families of distributions. *Univ. Calif. Publ. Statist.* **3,** 37–98.

LeCam, L. (1972). Limits of experiments, *Proc. 6th Berkeley Symp. Math. Stat. Prob.* **1,** 245–261.

Moler, C. & van Loan, C. (1978). Nineteen dubious ways to compute the exponential of a matrix, SIAM *Rev.* **20,** 801–836.

Nour, E.-S. & Suchindran, C. M. (1984). The construction of multi-state life tables: comments on the article by Willekens *et al. Population Studies* **38,** 325–328.

Ortega, J. M. & Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables.* Academic Press, New York.

Petrov, V. V. (1975). *Sums of independent random variables.* Springer-Verlag, Berlin.

van der Plas, A. P. (1983). On the estimation of the parameters of Markov probability models using macro data. *Ann. Statist.* **11,** 78–85.

Rogers, A. & Ledent, J. (1976). Increment–decrement life tables: a comment. *Demography* **13,** 287–290.

Schoen, R. & Nelson, V. E. (1974). Marriage, divorce and mortality: a life table analysis. *Demography* **11,** 267–290.

Todd, M. J. (1978). The computation of fixed points and applications. *Lecture Notes in Economics and Mathematical Systems* **124,** Springer-Verlag, Berlin.

Received September 1984; in final form November 1985.

Richard D. Gill, Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands.