Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

R.D. Gill

On estimating transition intensities of a Markov process with aggregate data of a certain type

Department of Mathematical Statistics Report MS-R8411 September

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

ON ESTIMATING TRANSITION INTENSITIES OF A MARKOV PROCESS WITH AGGREGATE DATA OF A CERTAIN TYPE

R.D. GILL

Centre for Mathematics and Computer Science, Amsterdam

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 inter-regional migration). However the fact that data is 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.

1980 MATHEMATICS SUBJECT CLASSIFICATION: 62M05, 62Pxx.

KEY WORDS & PHRASES: Markov process, aggregate data, multidimensional mathematical demography, multistate life-table, occurence-exposure rate, fixed-point theorem, degree theory.

NOTE: This paper has been submitted to the Scandinavian Journal of Statistics.

Report MS-R8411

Centre for Mathematics and Computer Science

P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

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 is 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.

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). This process is described by an initial distribution μ, considered as a row-vector with nonnegative 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 nonnegative nondiagonal elements q_{ij}, $i \neq j$, and diagonal elements $q_{ii} = -\Sigma_{j \neq i} q_{ij} \leqslant 0$. 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 $(\mathbf{X}_t : t \geqslant 0)$; 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 $(P_t)_{ij} = \mathbb{P}(\mathbf{X}_{s+t} = j \mid \mathbf{X}_s = i)$. 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 $[0,1]$, $l_i = \mathbb{E}(\int_0^1 \mathbb{I}\{\mathbf{X}_s = i\}\mathrm{d}s)$, where $\mathbb{I}\{\cdots\}$ 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 $Q\underline{1}^{\mathrm{T}} = 0$ so that $\mathrm{rank}(Q) \leqslant p-1$. If $\mathrm{rank}(Q) = p-1$ and moreover $\underline{1}^{\mathrm{T}}$ is linearly independent of the columns of Q (i.e. $\mathrm{rank}(Q : \underline{1}^{\mathrm{T}}) = p$) then for given μ and Q the equations in l:

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

and

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

are equivalent. A necessary and sufficient condition for $\mathrm{rank}(Q : \underline{1}^{\mathrm{T}}) = p$ is that there exists at least one state to which all states have access (see Appendix). This is also equivalent to the condition $\mathrm{rank}(Q) = p-1$. We shall from now on always assume that this is the case. It appears that more complex situations can be handled by appropriate decompositions of the state space, cf. Funck Jensen (1982b). (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 $m = 1,..,k$. States i and j communicate if each has access to the other. State i is transient if it has access to a state j which does not have access to i. Otherwise it is recurrent.)

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 $N_{ij} = \mathbb{E}(\Sigma_{t \in [0,1]} \mathbf{I}\{\mathbf{X}_{t-} = i, \mathbf{X}_t = j\})$ 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

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

$$= \text{total number of moves from } i \text{ to } j \text{ during } [0,1]$$

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

$$\mathbf{I}_i^n = \sum_m \int_0^1 \mathbf{I}\{\mathbf{X}_t^m = i\} dt$$

$$= \text{total time spent in state } i$$

$$\mu_i^n = \sum_m \mathbf{I}\{\mathbf{X}_0^m = i\}$$

$$= \text{initial configuration}$$

$$\nu_i^n = \sum_m \mathbf{I}\{\mathbf{X}_1^m = i\}$$

$$= \text{final configuration}$$

where the summations are over $m = 1, \ldots, n$, $t \in [0,1]$ and $j \in \{1, \ldots, p\}$. Then defining μ by $\mathbb{E}\mu^n = n\mu$, we obtain that $\mathbb{E}\mathbf{N}^n = nN$, $\mathbb{E}\mathbf{I}^n = nl$ and $\mathbb{E}\nu^n = n\nu$, 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 μ^n; i.e. given the initial configuration and the total number of moves during [0,1]. We assume that all other quantities, in particular $\mathbf{I}^n$, are not observed. We seek estimators which have good properties as $n \to \infty$. Note that $\mu^n \underline{1}^T = \nu^n \underline{1}^T = \mathbf{I}^n \underline{1}^T = n$ and that $\nu^n = \mu^n + \underline{1}\mathbf{N}^n$.

Before describing our new proposal, we discuss the currently available solutions to this problem. Had $\mathbb{I}^n$ 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* $\hat{Q}^n = (\text{diag } \mathbb{I}^n)^{-1}\mathbb{N}^n$ possesses a large number of desirable properties as estimator of Q. Conditional on $\mu^n = n\mu$ it is a maximum likelihood estimator of Q. Under conditions which ensure that the elements of $\mathbb{I}^n$ become arbitrarily large at a 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 components asymptotically independent and with asymptotic variances which can be estimated by the elements of $(\text{diag } \mathbb{I}^n)^{-2}\mathbb{N}^n = (\text{diag } \mathbb{I}^n)^{-1}\hat{\mathbf{Q}}^n$. 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 $(\mu^n, \mathbb{N}^n)$ is so intractable that a maximum likelihood estimator of Q based on data $(\mu^n, \mathbb{N}^n)$ cannot be computed, neither directly nor by means of the EM-algorithm (cf. Dempster, Laird & Rubin, 1977), for which one would have to evaluate $\mathbb{E}_Q(\mathbb{I}^n \mid \mu^n = \mu^n, \mathbb{N}^n = N^n)$. Therefore one usually takes recourse to the "working approximation" $\mathbb{I}^n \approx \tilde{\mathbb{I}}^n = \frac{1}{2}(\mu^n + \nu^n)$ and estimates Q by $\tilde{\mathbf{Q}}^n = (\text{diag } \tilde{\mathbb{I}}^n)^{-1}\mathbb{N}^n$. 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 setup 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 $\sum_m \mathbf{I}\{\mathbf{X}_0^m = i, \mathbf{X}_1^m = j\}$) rather than on the intensity matrix Q. Within the Markov process setup one would generally estimate P_1 by substituting an estimate of Q in the formula $P_1 = \exp(Q)$. The alternative "demographic" approach to the whole problem is to abandon the time-homogeneous Markov process model and to elevate the working approximation $\mathbb{I}^n \approx \frac{1}{2}(\mu^n + \nu^n)$ 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, $\tilde{\mathbf{P}}_1^n = (I - \frac{1}{2}\tilde{\mathbf{Q}}^n)^{-1}(I + \frac{1}{2}\tilde{\mathbf{Q}}^n)$; cf. Rogers & Ledent (1976). However there are some logical inconsistencies in this derivation which are discussed in Keilman & Gill (1984). In our setup 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 μ^n and $\mathbb{N}^n$ 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 $n^{-1}\mu^n$ and $n^{-1}\mathbb{N}^n$ 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, and when is the solution unique?

(ii) What is a good algorithm for finding a (the) solution?

(iii) What are the statistical properties of the resulting estimators?

So far we only have limited mathematical results on question (i) though practical results are very encouraging. When all states communicate we can prove that there always exists a solution. Under a further quite simple condition 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 at most one solution. We conjecture that there always exists exactly one solution. This means that question (i) is about half solved.

Regarding question (ii), an obvious iteration method is based on cycling repeatedly though 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 $\mathbb{E}_Q(\mathbb{I}^n \mid \mu^n = \mu^n)$; the EM-algorithm requires one to compute $\mathbb{E}_Q(\mathbb{I}^n \mid \mu^n = \mu^n, \mathbb{N}^n = N^n)$. However this superficial resemblance does not quarantee 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

reason 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))

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

which will be shown to be equivalent to solving the fixed point equation of the previous method

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

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

For practical purposes then questions (i) and (ii) could be considered as satisfactorily answered, though from the point of view of mathematical theory there are more questions than answers. All the same, as regards (iii), 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.

The rest of the paper consists of two parts, one devoted to questions (i) and (ii), the other to question (iii): 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. 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 & Rogers (1982) can be seen as a (possibly very good) working approximation to the solutions which a generalization of the present theory would supply.

2. Solving the estimating equations

As we saw in Section 1, for a Markov proces 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 N is the expected number of moves between each two states during [0,1]:

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

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

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

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

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

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

We suppose rank $(Q) = p-1$ or equivalently there exists a state to which all states have access. We suppose that μ and Q are such that every state can be accessed from a state with $\mu_i > 0$; consequently $\nu_i, l_i > 0$ for all i and rank (N) = rank (Q). Note that N can also be considered as an intensity matrix and as such, since $l_i > 0$ for all i, 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 such that rank$(N) = p-1$, every state can be accessed from a state with $\mu_i > 0$, and ν defined by $\nu = \mu + \underline{1}N$ has $\nu_i > 0$ for all i. 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 defined by (13) must also satisfy (9), by linearity and the derivation of the "flow equation" (13) in Section 1. From this and the assumption that $\nu_i > 0$ for all i it follows that the elements of l defined by (10) are all positive (if an element of l is zero then the corresponding element of μe^{Qs} must be zero for all s). Therefore we can write $Q = (\mathrm{diag}\, l)^{-1}N$. Thus the existence of Q implies the existence of a vector l with $l_i > 0$ for all i such that, from (10),

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

and, from (9),

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

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

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

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

$$lQ = \mu(\exp(Q) - I), \quad l\underline{1}^{\mathrm{T}} = 1 \tag{18}$$

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.

So our problem has now become, given μ, N and $\nu = \mu + \underline{1}N$ satisfying the various rank and positivity conditions, does there exist l with $l_i > 0$ for all i such that (15) or (16) holds? Now let S denote the simplex $\{l \in \mathbb{R}^p : l_i \geq 0\ \forall i, l\underline{1}^{\mathrm{T}} = 1\}$ and let S^0 denote its (relative) interior $\{l \in \mathbb{R}^p : l_i > 0\ \forall i, l\underline{1}^{\mathrm{T}} = 1\}$. We shall give an answer in the special case in which all states communicate - i.e. N is irreducible. 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

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

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

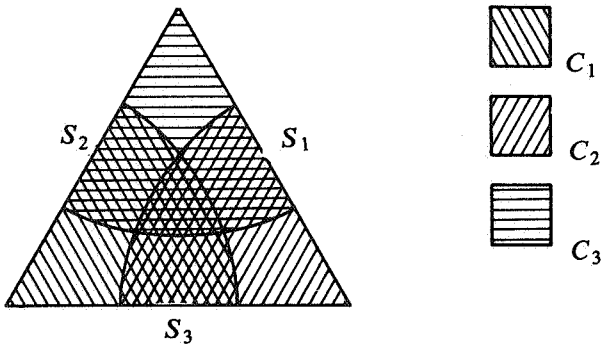
We extend $\hat{l}$ and $\hat{\nu}$ to all of S by going back to the explicit construction of the process X in Section 1. Define $\alpha_{ij} = -n_{ij}/n_{ii}$ for $i \neq j$ such that $n_{ii} < 0$, $\alpha_{ij} = 0$ otherwise. For $l \in S$ we say $-l_i/n_{ii} = \infty$ if $n_{ii} = 0$. Bij an exponentially distributed random variable with mean zero or mean infinity we mean a random variable which is identically 0 or identically $+\infty$ respectively. Suppose from now on that each state either has access to all others or is absorbing. For $l \in S$ we define a process 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 α_{jk},.... 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 $-l_i/q_{ii} = 0$,

then after an almost surely finite number of instantaneous jumps one arrives in a state with $-l_i / q_{ii} > 0$ 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. If one wants to define a process which also contains information on instantaneous jumps, one should append $\mathbf{X}_t^m$, $m = 1,2,...$ where $\mathbf{X}_t^m = m$'th state jumped into at time t, $\mathbf{X}_t^m = 0$ if there is no m'th jump at time t.

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 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 $\hat{l}(l)_i$ does 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 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). For this we define the *faces* S_i of S by $S_i = \{l \in S : l_i = 0\}$.

Lemma *(Knaster, Kuratowski & Mazurkiewicz): Let $C_1, \ldots, C_p$ be closed subsets of S such that $S = \bigcup_1^p C_i$, $\bigcap_{i \in I} S_i \subset \bigcup_{j \notin I} C_j$ for all $I \subset \{1, \ldots, p\}$. Then $\bigcap_1^p C_i$ is nonempty.*



The K-K-M lemma, $p = 3$

For our application we define $C_i = \{l \in S : \hat{\nu}(l)_i \geq \nu_i\}$. 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 $\hat{\nu}(l)_i \geq \nu_i$; i.e. $S = \bigcup_1^p C_i$. Finally if $l_i = 0, i \in I$, then $\hat{\nu}(l)_i = 0 < \nu_i, i \in I$, so $l \notin \bigcup_{i \in I} C_i$. Therefore $l \in \bigcup_{j \notin I} C_j$. So $C_1, \ldots, C_p$ satisfy the conditions of the lemma and $\bigcap_1^p C_i$ is nonempty. But for $l \in \bigcap_1^p C_i$, $\hat{\nu}(l)_i \geq \nu_i$ for all i, hence $\hat{\nu}(l) = \nu$.

An even simpler proof is obtained by reversing the inequality in the definition of C_i and appealing to Freidenfelds' (1974, Theorem 1') version of the K-K-M lemma.

If $\nu_i = 0$ for $i \in I$, we can apply the same reasoning on the lower-dimensional simplex $\bigcap_{i \in I} S_i$, to show that under the other assumptions we have made, the equation $\hat{\nu}(l) = \nu$ still has a solution.

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, $\nu_i > 0$ for all i). Define the matrix $J = J(\mu, Q)$ by

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

Then we show that if $J = J(\mu,(\operatorname{diag} l)^{-1}N)$ is nonsingular 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 $-(\operatorname{diag} l)^{-1}JQ$ is the Jacobian matrix of the transformation $\hat{\nu}: S^0 \subset \mathbb{R}^p \to \mathbb{R}^p$. For, denoting by $A_{i\cdot}$ the i'th row of the matrix A, we have

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

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

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

Now define $S^* = \{x \in \mathbb{R}^{p-1}: \ x_i \geqslant 0 \ \forall i, \ \Sigma_1^{p-1} x_i \leqslant 1\}$. Define $\tau_i: S^* \to S$ by $\tau_i(x) = (x_i, \ldots, x_{i-1}, 1-\Sigma_1^{p-1}x_j, x_i, \ldots, x_{p-1})$. Note that τ_i^{-1} exists and $\tau_i^{-1}(l) = (l_i, \ldots, l_{i-1}, l_{i+1}, \ldots, l_p)$. 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 i'th component of l and the j'th 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 nonsingular linear 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 i'th row from all the other rows and then deleting the i'th row and j'th column (if $l_p = 1 - l_1 - \cdots - l_{p-1}$, then for $i,j < p$, $\partial \hat{\nu}_j^{(p,p)} / \partial l_i = \partial \hat{\nu}_j / \partial l_i - \partial \hat{\nu}_p / \partial l_i$). So if J is nonsingular and N has rank $p-1$, then for $l \in S^0$, $-(\operatorname{diag} l)^{-1}JQ$ 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 nonsingular, then for some i,j, $\hat{\nu}^{(i,j)}$ has nonsingular Jacobian. Hence all $\hat{\nu}^{(i,j)}$ have nonsingular 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 nonsingular everywhere, its determinant has the same sign everywhere. Consequently if J is nonsingular 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 $y = \tau_i^{-1}(\nu)$. 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 $\{x \in S^*: x_i \leqslant \delta$ for some i or $\Sigma_1^{p-1} x_i \geqslant 1-\delta\} \times [0,1]$ for some $\delta > 0$, where of course $p\delta < 1$. Let $C = \{x \in S^*: x_i > \delta \ \forall i$ and $\Sigma_1^{p-1} x_i < 1-\delta\}$. 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.56) 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 $F:E\to\mathbb{R}^p$ with Jacobian matrix F' which is nonsingular at all solutions in C of $F(x) = y$ and which has no solutions on ∂C,

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

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 existence and 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 assumptions 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 nonzero. 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 $1-(l_1 + \cdots + l_{i-1})$, $\hat{\nu}(l)_i$ 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.

These are the only presently available results on existence and uniqueness. * We have hope that a way will be found, using the same tools, to obtain better results in the future. 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 $l^{(k+1)} = \hat{l}(l^{(k)})$, $k = 1,2,\cdots$ (starting from some initial quess $l^{(1)}$) results are even more meagre. Denoting by $\frac{\partial \hat{l}}{\partial l}$ the matrix with (i,j)'th element $\frac{\partial \hat{l}_j}{\partial l_i}$, it can be shown quite easily that $\frac{\partial \hat{l}}{\partial l} = -(\text{diag } l)^{-1}(J - \text{diag } \hat{l})$. Since $J(l)\underline{1}^T = \hat{l}(l)^T$, at a fixed-point $\frac{\partial \hat{l}}{\partial l}$ equals the identity matrix minus a stochastic matrix. If it could be shown that the spectral radius of $\frac{\partial \hat{l}}{\partial l}$ 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. However it is not clear whether or not $\frac{\partial \hat{l}}{\partial l}$ has this property.

---

* A further result is: if there is a unique solution, with nonsingular J, at $(\mu,N) = (\mu_0,N_0)$, then there is a unique solution in a neighbourhood of (μ_0,N_0). In particular this applies to $(\mu_0,N_0) = (\nu_0,0)$.

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 $\mu^n = \mu^n$, for some arbitrary sequence of realized initial distributions μ^n, $n = 1,2,..$

To start with 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 nonsingular. 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 nonsingular at the solution $l = l_0$ of (16) defined by (10). Of course there may be other solutions of (16) , i.e. of $\phi(l;\mu_0,N_0) = 0$; 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 $-(\frac{\partial\phi}{\partial l})^{-1}(\frac{\partial\phi}{\partial(\mu,N)})|_{(\mu_0,N_0)}$. (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 $\hat{\mathbf{l}}^n = l^*(n^{-1}\mu^n, n^{-1}\mathbf{N}^n)$ for $(n^{-1}\mu^n, n^{-1}\mathbf{N}^n)$ in the neighbourhood of (μ_0,N_0) (the probability of this event converges to 1 as $n\to\infty$), then $n^{\frac{1}{2}}(\hat{\mathbf{l}}^n - l_0)$ 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 $n^{\frac{1}{2}}((n^{-1}\mu^n, n^{-1}\mathbf{N}^n) - (\mu_0,N_0))$. Defining $\hat{\mathbf{Q}}^n = (\text{diag}\hat{\mathbf{l}}^n)^{-1}(n^{-1}\mathbf{N}^n)$, the same holds for $n^{\frac{1}{2}}(\hat{\mathbf{Q}}^n - Q_0)$ by a further application of the δ-method. In Gill (1984) the asymptotic distribution of $n^{\frac{1}{2}}((n^{-1}\mu^n, n^{-1}\mathbf{N}^n) - (\mu_0,N_0))$ is described. See also Funck Jensen (1982a) and her references. So in principle the asymptotic covariance matrix of $n^{\frac{1}{2}}(\hat{\mathbf{Q}}^n - Q_0)$ is determined and can be consistently estimated by substituting $n^{-1}\mu^n$ 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 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 $\mathbf{X}_1,\mathbf{X}_2,...$ 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 $\bar{\mathbf{X}}_n = n^{-1}\Sigma_{i=1}^n\mathbf{X}_i$. 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 $n^{\frac{1}{2}}(\bar{\mathbf{X}}_n - \mu(\theta)) \overset{\mathcal{D}(\theta)}{\to} N(0,\sigma^2(\theta))$, we have by the δ-method $n^{\frac{1}{2}}(\hat{\theta}_n - \theta) \overset{\mathcal{D}(\theta)}{\to} N(0,((\frac{\partial\mu}{\partial\theta})^{-1})^{\mathrm{T}}\,\sigma^2(\theta)(\frac{\partial\mu}{\partial\theta})^{-1})$. (Here $\overset{\mathcal{D}(\theta)}{\to}$ 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 argumemts go through. So we have in $\mathbb{R}^1$ i.i.d. random variables $\mathbf{X}_i$ with

$$n^{\frac{1}{2}}(\bar{\mathbf{X}}_n - \mu) \overset{\mathcal{D}(\mu)}{\to} N(0,\sigma^2(\mu))$$

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 $n^{-\frac{1}{2}}$ 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,

$$\log\left(\frac{p_n(\bar{\mathbf{X}}_n;\mu_0+n^{-\frac{1}{2}}h_n)}{p_n(\bar{\mathbf{X}}_n;\mu_0)}\right) \overset{\mathcal{D}(\mu_0)}{\to} N\left(-\tfrac{1}{2}\frac{h^2}{\sigma^2(\mu_0)}, \frac{h^2}{\sigma^2(\mu_0)}\right). \tag{20}$$

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 $\mathbf{Y}_n = n^{\frac{1}{2}}(\bar{\mathbf{X}}_n - \mu_0)$. Under $\mu_n = \mu_0 + n^{-\frac{1}{2}}h_n$, $\mathbf{Y}_n$ 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

$$\log\left(\frac{(2\pi\sigma_0^2)^{-\frac{1}{2}}\exp\left(-(\mathbf{Y}_n-h)^2/2\sigma_0^2\right)}{(2\pi\sigma_0^2)^{-\frac{1}{2}}\exp\left(-\mathbf{Y}_n^2/2\sigma_0^2\right)}\right) = \frac{1}{2\sigma_0^2}(\mathbf{Y}_n^2-(\mathbf{Y}_n-h)^2)$$

$$= \frac{h\mathbf{Y}_n}{\sigma_0^2} - \frac{h^2}{2\sigma_0^2}$$

$$\overset{\mathcal{D}(\mu_0)}{\to} N\left(-\frac{h^2}{2\sigma_0^2}, \frac{h^2}{\sigma_0^2}\right).$$

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 $n^{\frac{1}{2}}(\bar{\mathbf{X}}_n - \mu)$ 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.

Acknowledgement. This paper owes much to many stimulating discussions with Jan Hoem, Nico Keilman, Frans Willekens, and many colleagues at CWI.

References

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

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. Roy. 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, Univ. 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, Univ. of

Stockholm.

GILL, R.D. (1984), A note on two papers in central limit theory, Report MS-R8410, Centrum voor Wiskunde en Informatica, Amsterdam (to appear in *Proc. 44th Session I.S.I., Madrid; Bull. Int. Stat. Inst.* **50**, 3).

HAJEK, J. (1970), A characterization of limiting distributions of regular estimates, *Zeit. Wahrsch. u. 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; p. 155-264 in: *Multidimensional mathematical demography,* Land, K.C., & Rogers, A. (eds.), Academic Press, New York.

JOHANSEN, S. (1973), The bang-bang problem for stochastic matrices, *Zeit. Wahrsch. u. verw. Geb.* **26,** 191-195.

KEILMAN, N. & GILL, R.D. (1984), New light on the linear integration hypothesis (in preparation).

KNASTER, B., KURATOWSKI, C., & MAZURKIEWICZ, S. (1929), Ein Beweis des Fixpunktsatzes für n-dimensionale Simplexen, *Fund. Math.* **14** 132-137.

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

LAND, K.C., & ROGERS, A. (1982), Statistical methods for Markov generated increment-decrement life tables with polynomial gross flow functions; p. 265-346 in: *Multidimensional mathematical demography,* Land, K.C., & Rogers, A. (eds.), Academic Press, New York.

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 review* **20** 801-836.

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.

ROGERS, A., & LEDENT, J. (1976), *Increment-decrement life tables: a comment,* Demography **13** 287-290.

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

Appendix

rank$(Q:\underline{1}^{\mathrm{T}}) = p \Leftrightarrow$ *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 (p. 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 $(-Q)x = \underline{1}^{\mathrm{T}}$. Without loss of generality we can take i_0 to be recurrent. 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, $(-Q)x = \underline{1}^{\mathrm{T}} \Rightarrow (-Q_I)x_I = \underline{1}_I^{\mathrm{T}}$. 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 $(-Q)\underline{1}^{\mathrm{T}} = 0$, 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), $(-Q)x \geqslant 0 \Rightarrow (-Q)x = 0$. So $(-Q)x = \underline{1}^{\mathrm{T}}$ 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$(Q:\underline{1}^{\mathrm{T}}) \leqslant p-1$.