Markov process theory without time invariance?

J.F.C. Kingman

A hundred years ago A.K. Erlang was working in the Copenhagen Telephone Company applying mathematical techniques to the analysis of the congestion problems then being encountered in telephone exchanges. His first insight was that the way calls arrive irregularly at an exchange may be modelled as a Poisson process of random instants. If calls are dealt with one at a time, each call taking an exponentially distributed time independent from call to call, Erlang derived a family of differential equations for the probabilities $p_j(t)$ of there being j calls waiting at time t, namely

$$dp_j/dt = -(\alpha+\beta)\, p_j + \alpha p_{j-1} + \beta p_{j+1} \qquad (j = 1,2,\ldots) \tag{1}$$

$$dp_0/dt = -\alpha p_0 + \beta p_1 \quad .$$

This is of course what queueing theorists call the queue *M/M/*1, and is an example of the class of Markov processes known as birth and death processes. A great deal is known of such things, and indeed (1) has explicit solutions [1] *so long as* α *and* β *are constants.* But in real life the arrival rate α (and perhaps the service rate β too) will certainly change significantly over time. Erlang's derivation of (1) does not assume that the coefficients are constant, and to be realistic significant variation in arrival rate must be allowed [4].

More generally, the usual definition of a Markov process is a random process (X_t) evolving with a real time parameter t, with the property that, for $s < t$,

$$\mathbf{P}\{\, X_t \in A \mid X_u\,(u \leq s)\,\} = P(s,t;\, X_s, A)$$

for some function P. Almost all treatments of this property make the further assumption that P depends on s and t only through the time difference $t - s$. The advantage of doing so is that it leads to a rich and deep theory, but if probability is to be an applicable discipline we should acknowledge that this assumption is by no means always satisfied in applications.

There is no accepted terminology for the special case of time invariance, but in differential equation circles equations like (1) are called *autonomous* if the coefficients are constants. I suggest that we borrow this terminology, and describe a Markov process as *autonomous* if

$$P(s,t;\, x,A) = P(t\text{-}s;\, x,A) \quad . \tag{2}$$

The question in the title is whether there is a useful theory that does not assume that the Markov process is autonomous.

To start with the simplest case, suppose that the random variables X_t take only finitely many values 1,2,..,N. Then the transition probabilities are of the form

$$p_{ij}(s,t) = \mathbf{P}\,\{\, X_t = j \mid X_s = i \,\},$$

for $s \le t$, and these may be assembled into a (NxN) stochastic matrix $P(s,t)$ which then satisfies the Chapman-Kolmogorov equation

$$P(s,u) = P(s,t)\, P(t,u), \quad (s \le t \le u) \tag{3}$$

In this equation the time parameter may be discrete or continuous, depending on the application. If for instance t takes only integer values, (3) leads to an expression for $P(s,t)$ in terms of the one-step transition matrices:

$$P(s,t) = P(s,s+1)\, P(s+1,s+2)\, \ldots\, P(t-1,t).$$

In general, the terms in this matrix product will not commute, so that an analytical expression for $P(s,t)$ will not be available, but modern computers make light work of such products even when N is large.

If t is a continuous time parameter, it is natural to assume that $P(s,t)$ is near the identity matrix I when $t - s$ is small. This means that P is non-singular when $t - s$ is small, and (3) then shows that the same is true for all s,t. This allows us to express P in terms of a matrix-valued function of a single variable; for instance if $0 < s < t$,

$$P(s,t) = P(0,s)^{-1}\, P(0,t)\ .$$

But the requirement that this matrix quotient have positive elements imposes complex restrictions on the trajectory $P(0, .)$, and even the obvious question of which matrices the trajectory can reach is unsolved.

In any case, the tool likely to be useful in applications is the pair of differential equations like (1) and its dual. By analogy with the known results for the autonomous case, we may ask if there is a matrix-valued function Q of a single variable, such that

$$P(s,t) = I + (t-s)\, Q(s) + o(t-s)\ ?$$

If so, are the Kolmogorov differential equations

$$\partial P(s,t)/\partial t = P(s,t)Q(t)\ , \qquad \partial P(s,t)/\partial s = -Q(s)P(s,t) \tag{4}$$

valid? Does Q determine P uniquely?

The simplest non-trivial case is that in which N=2 and return from 2 to 1 is forbidden. This has

$$p_{11}(s,t) = p(s,t),\ p_{12}(s,t) = 1 - p(s,t),\ p_{21}(s,t) = 0,\ p_{22}(s,t) = 1$$

for some function p, and (3) shows that $p(s,u) = p(s,t)p(t,u)$, so that

$$p(s,t) = f(t)/f(s)$$

for some continuous decreasing function f. Any such function f will generate a solution of (3), and (4) will hold if f is differentiable, with

$$q_{11}(t) = - q_{12}(t) = f'(t)/f(t), \quad q_{21}(t) = q_{22}(t) = 0.$$

But f may not be differentiable, in which case the Kolmogorov differential equations hold, if at all, only in a weak sense. This example illustrates an important truth; the autonomous theory is a theory of generalised exponentials, while the general theory is about monotone functions.

This can be made precise by a geometrical representation. Define a convex compact subset of $\mathbf{R}^N$ as the image

$$\Phi_{st} = \{P(s,t)x \; ; \; 0 \leq x \leq 1\}$$

of the unit cube. Then (3) shows that the set Φ_{st} decreases with t, and the intersection Φ_s is again convex and compact, with dimension d_s between 1 and N. This dimension increases with s, and thus is a constant d in $1 \leq d \leq N$. This geometry can be linked with results of Iosifescu and Cohn on the tail σ-field of the Markov process, and it turns out that this is atomic with exactly d atoms. Simple martingale arguments relate the atoms to asymptotic behaviour of the process, and in particular $d = 1$ corresponds to *weak ergodicity*:

$$p_{ij}(s,t) - p_{hj}(s,t) \; \rightarrow 0 \quad (t \rightarrow \infty).$$

An important, and rather neglected, paper by G.S. Goodman contains the key to understanding the finite matrix solutions of (3). He takes determinants of (3), and concludes that there is a continuous decreasing function g such that

$$\det \{P(s,t)\} = g(t)/g(s)$$

for all $s<t$. He then transforms the time scale by taking a new time variable

$$t^*(t) = - \log g(t),$$

so that the determinant becomes simply $\exp(s^* - t^*)$. Finally he proves the inequality that the matrix $P = P(s,t)$ always satisfies the inequality

$$p_{11}p_{22}...p_{NN} \; \geq \; \det (P) > 0 \; . \qquad (5)$$

This enables him to show that the p_{ij} satisfy Lipschitz conditions, and that the Kolmogorov equations (4) hold almost everywhere.

The 'almost everywhere' suggests integrating the differential equations, writing for instance the forward equation as

$$P(s,t) = I + \int P(s,u) \, Q(u) \, \mathrm{d}u,$$

or better

$$P(s,t) = I + \int P(s,u) \, Q(\mathrm{d}u) \qquad (6)$$

where the integral extends over (s,t) and the matrix-valued measure Q in (6) is defined by its density $Q(.)$. The advantage of the second form is that it allows us to transform the time scale back to the original one, the measure Q transforming in the usual way. If g is not differentiable, Q may no longer have a density, but it is still a non-atomic, locally finite matrix-valued measure. Its off-diagonal elements are positive measures, but its row sums are zero, so that the diagonal elements are purely negative.

Conversely, if we are given locally finite, non-atomic positive measures measures q_{ij} $(i \neq j)$, we assemble these into a matrix Q, choosing the diagonal elements to make the row sums zero, and consider the integral equation (6). This may be solved by a simple iterative procedure, and the result is a solution of the Chapman-Kolmogorov equation (3). Thus Goodman's analysis, though he did not realise this, gives a complete description of the general Markov process in continuous time, with finitely many states.

The idea that the transition intensities are naturally described by measures should be no surprise. In (1) the intensity measure from j to j+1 simply gives the expected number of arrivals in any time interval. In principle this can be any non-atomic measure; Erlang could have considered the case when calls arrive at the instants of a Cantor set.

Of course, the interesting problems arise when there are infinitely many states, as in Erlang's example. In the autonomous case, the infinitesimal generator is no longer a matrix of intensities, but in general an unbounded operator on a suitable Banach space. However, the construction of a solution of (3) from intensity measures still goes through [5], so long as the measures $-q_{ii}$ are dominated by some locally finite measure.

This condition is satisfied by *M/M/*1, but it is unnecessarily restrictive. More generally, (6) has an iterative solution which will satisfy (3) but may have row sums strictly less than 1, corresponding to the possibility that the random process makes infinitely many jumps in finite time. This is a phenomenon well understood in the countable case, and it should not be too difficult to give conditions for it to occur, for instance in the general birth and death process.

All the above discussion is relatively straightforward, with no really deep results, but there are deeper questions worth studying. One of the most tantalising results in the autonomous theory (with a countable infinity of states) is what is sometimes called the *Lévy dichotomy*. Suppose that $p_{ij}(t)$ are the transition probabilities of an autonomous Markov process. Then Lévy asserted, and Ornstein proved, that each function p_{ij} is either always zero or never zero. It is easy to show that if $p_{ij}(s) > 0$, then $p_{ij}(t) > 0$ for $t > s$; the difficult thing is to prove the inequality for small t . Thus the result asserts that 'what can ever happen can happen arbitrarily quickly'. This has a full extension to processes which may not be autonomous, in the form of the following theorem [5], which reduces to Ornstein's in the autonomous case:

Suppose that $p_{ij}(a,b) > 0$, and let $\varepsilon > 0$. *Then there is a dissection*

$$a = t(0) < t(1) < t(2) < \ldots < t(2n+1) = b$$

and states $k(0) = i$, $k(1), k(2), \ldots, k(n+1) = j$ *such that*

$$p_{k(r-1)k(r)}(t(2r-1),t(2r)) > 0,$$

and

$$\sum \{t(2r) - t(2r-1)\} < \varepsilon \ .$$

In other words, there are trajectories with positive probability which are constant except on a (non-random) set of arbitrarily small measure.

References

[1] J.W. COHEN (1969), *The Single Server Queue* (North-Holland).

[2] G.S. GOODMAN, An intrinsic time for non-stationary Markov chains, *Z. Wahrscheinlichkeitstheorie* **16** (1970) 165-180.

[3] J.F.C. KINGMAN, Geometrical aspects of the theory of non-homogeneous Markov chains, *Math. Proc. Camb. Phil. Soc.* **77** (1975) 171-183.

[4] J.F.C. KINGMAN, The first Erlang century – and the next, *Queueing Syst.* **63** (2009) 3-12.

[5] J.F.C. KINGMAN, Forbidden transitions in Markovian systems, *Proc. Lond. Math. Soc.* **105** (2012) 730-756.